Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique

Puyu Fenga,b, Bin Wangb,⁎, De Li Liub,c,⁎, Cathy Watersd, Dengpan Xiaob,e, Lijie Shi,a,b, Qiang Yuf,a,g,⁎

a School of Life Sciences, Faculty of Science, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia
b NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, Wagga Wagga, NSW 2650, Australia
c Climate Change Research Centre and ARC Centre of Excellence for Climate Extremes, University of New South Wales, Sydney, NSW 2052, Australia
d Australia NSW Department of Primary Industries, Dubbo, NSW 2830, Australia
e Institute of Geographical Sciences, Hebei Academy of Sciences, Shijiazhuang 050011, China
f State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Northwest A&F University, Yangling 712100, Shaanxi, China
g College of Resources and Environment, University of Chinese Academy of Science, Beijing 100049, China

ARTICLE INFO

Keywords:
Wheat yield forecast
Extreme climate events
Remote sensing
APSIM
Random forest

ABSTRACT

Early and reliable seasonal crop yield forecasts are crucial for both farmers and decision-makers. Commonly-used methods for seasonal yield forecasting are based on process-based crop models or statistical regression-based models. Both have limitations, particularly in regard to accounting for growth stage-specific climate extremes (such as drought, heat, and frost). In this study, we firstly developed a hybrid yield forecasting approach by blending of multiple growth stage-specific indicators, i.e. APSIM (a process-based crop model)-simulated biomass, and climate extremes, NDVI (Normalized Difference Vegetation Index), and SPEI (Standardized Precipitation and Evapotranspiration Index) before forecasting dates, using a regression model (random forest or multiple linear regression). Plot-scale wheat yield (2008–2017) in the southeastern Australian wheat belt was dynamically forecasted at the end of several targeted growth stages as the growing season progressed to harvest. Results showed that the forecasting accuracy increased significantly for both systems as forecast time approached harvest time. The forecasting system based on random forest outperformed the forecasting system based on multiple linear regression at each forecasting event. Satisfactory yield forecasts occurred at one month (~35 days) prior to harvest (r = 0.85, LCCC = 0.81, MAPE = 17.6%, RMSE = 0.70 t ha⁻¹, and ROC score = 0.90), and at two months before harvest (r = 0.62, LCCC = 0.53, MAPE = 27.1%, RMSE = 1.01 t ha⁻¹, and ROC score = 0.88). In addition, drought events throughout the growing season were identified as the main factor causing yield losses in the wheat belt during the past decade. With the increasing availability of farming-related data, we expect that the yield forecasting system proposed in our study may be widely extended to other comparable cropping regions to produce sufficiently accurate wheat yield forecasts for stakeholders to develop strategic decisions in their respective roles.

1. Introduction

Seasonal forecasting of crop yield is becoming increasingly important in both developed and developing countries (Basso and Liu, 2018). This is mainly due to the growing demand for maximizing profits in terms of both farm-level outputs and commodities trading. Early and reliable warning information regarding weather and management impacts on crop yield is crucial for stakeholders to make strategic decisions in their respective roles. For crop producers, once crop yield is site-specifically predicted, appropriate farm management practices and security precaution measures (e.g., grain storage) can be determined. For government policy makers and grain marketing agencies, yield forecasting can provide invaluable information for regulating agricultural markets and determining trading strategies. Therefore, many crop yield forecasting approaches have been developed across the world to provide crop yield outlooks (Cai et al., 2019; Chipanshi et al., 2015; Pagani et al., 2017).

At present, commonly-used yield forecasting approaches can be divided into three categories: (1) field surveys, (2) dynamic process-based crop simulation models, and (3) statistical regression-based models.

⁎ Corresponding authors.
E-mail addresses: bin.a.wang@dpi.nsw.gov.au (B. Wang), de.li.liu@dpi.nsw.gov.au (D.L. Liu), yuq@nwafu.edu.cn (Q. Yu).

https://doi.org/10.1016/j.agrformet.2020.107922
Received 27 June 2019; Received in revised form 18 January 2020; Accepted 28 January 2020
0168-1923/ © 2020 Elsevier B.V. All rights reserved.
models. The field survey method is still in use in many operational yield forecasting systems. It is conducted based on the within-season evaluation of crop growth by experienced farmers or farm managers (Nandram et al., 2014). Their evaluation of yield prospects represents farmers’ opinions about the effects of environmental and human factors on the final yield. Numerous crop yield forecasts from farmers can be collected through interviews (e.g., phone interviews) to give a synthetic assessment of yield outlooks for a specified region. However, the field survey method is usually time- and labor-consuming and provides relatively short lead times to inform decisions. Due to these limitations, a large amount of effort has been undertaken to obtain timely and reliable yield forecasts from the other two methods. Crop simulation models are capable of describing key physical and physiological processes by capturing the effects of the complex interactions between crop, soil, weather, and management practices. Thus, they can usually provide satisfactory end-of-season yield forecasts once required input data and parameters are provided. When using crop models for in-season yield forecasting, a major limitation is related to the unknown weather from the forecasting date to the maturity date (Basso and Liu, 2018). Previous studies have used various methods to fill this gap, including historical weather scenarios, seasonal weather forecasts, and climate model outputs, to run crop simulations to maturity date. However, in regions with large inter-annual climate variability, actual weather conditions can vary significantly from the projected weather conditions (Murphy and Timbal, 2008). Thus, great uncertainty can be introduced into the yield forecasting results. Moreover, crop simulation models usually have limited ability to simulate the effects of extreme climate events (ECEs, such as heat wave, frost, and drought). Some of the limitations are related to uncertainties in parameterization and vague descriptions or over-simplifications of certain processes, which can result in inaccurate yield estimations (Eitzinger et al., 2013). For example, the impacts of heat stress in particular are poorly captured in crop models (Barlow et al., 2015). Most crop models simulate the impacts of extreme temperatures on stem carbohydrate accumulation and distribution or leaf senescence, instead of directly modelling damage to reproductive processes or organs (Feng et al., 2019b). These limitations also raise uncertainty regarding the ability of crop models to properly forecast the end-of-season yield.

Statistical regression-based models relate crop yields to a number of selected predictors, such as meteorological factors and/or vegetation indices derived from remote sensing data, based on historical observations from a given region. Regression equations are then employed as a function of the inputs from other years in this region, or similar regions, to generate yield forecasts. Generally, statistical regression-based models are usually simple and easy to understand and need fewer parameter settings, thereby making them widely-used around the world. As the observed data are increasing in both quantity and quality in recent years, statistical regression-based models usually present satisfactory performance (Lobell and Asseng, 2017; Mathieu and Aires, 2018), especially under conditions characterized by large year-to-year fluctuations in yields, driven by several factors. However, statistical regression-based models are also not free from problems. Most current statistical regression-based models are based on linear regression models, such as multiple linear regression, which cannot capture the nonlinear relations between the dependent and independent variables. Given that crop yields often show nonlinear response to ECEs (Li et al., 2019; Feng et al., 2019b), linear regression models are likely to perform poorly under conditions with frequent climate extremes. Moreover, the same meteorological or vegetation factors occurring during different growth stages, namely growth stage-specific factors, can have different impacts on crop yield. For example, heat or drought events that occur during the flowering stage are likely to cause greater yield losses than those occurring during vegetative stages (Nielsen et al., 2010; Stratonovich and Semenov, 2015). As statistical regression-based models are usually unable to consider dynamic growth stage changes, they are limited in their ability to disentangle the effects of stage-specific factors.

Given that both crop simulation models and statistical regression-based models have limitations, researchers are attempting to integrate the two types of models in order to achieve complementary advantages. For example, the Crop Growth Modelling System (CGMS) incorporated crop simulation results and linear regression to provide policy makers with in-season regional yield forecasts of the main food crops in Europe (Kogan et al., 2013; Vossen and Rijks, 1995). Pagani et al. (2017) improved CGMS by incorporating agro-meteorological indices (accounting for drought and cold/heat stress) into the linear regression equation and increased the forecasting reliability by 94% in several European countries. The Integrated Canadian Crop Yield Forecaster (ICCYF) is another regional yield forecasting system for producers, grain traders, and policy makers, which integrates remotely sensed vegetation indices, climate, soil, and crop information through a crop simulation model and linear regression algorithms (Chipanshi et al., 2015). As these systems are still based on crop simulation models and traditional linear regression models, they may be limited in modelling yield losses caused by ECEs. Everingham et al. (2016) developed a random forest model (a machine learning method) with climate indicators and APSIM (a biophysical process-based crop model)-simulated biomass as predictors to forecast regional sugarcane (Saccharum officinarum L.) yield and obtained an $R^2$ of 0.67 for the model calibration process. However, they did not include remotely sensed indices as predictors, which might also help increase prediction accuracy. In addition, most yield forecasting studies did not consider growth stage-specific indices (Cai et al., 2019; Kern et al., 2018). Therefore, with the increasing availability of farming-related data, a more comprehensive yield forecasting system that incorporates growth stage-specific climate, remote sensing, soil, and crop information through crop simulation models and advanced regression methods is urgently needed for more accurate yield predictions.

Australia is among the most developed agricultural countries and one of the biggest wheat producers and exporters in the world. Early and reliable wheat yield forecasts for Australia become a critical element in providing guidance to farmers and policy makers. As in the areas of the Oceania region, atmospheric circulation patterns play an influential role on Australia’s weather, resulting in large inter-annual variability. Frequent extreme climate events, such as drought, heat, and frost have caused severe wheat yield losses during the last decades. As such, previous studies in Australia have given more attention to agro-meteorological information than to remotely sensed information when developing yield forecasting models. In this study, we intended to use growth stage-specific information from multiple data sources to make in-season yield forecasts based on the APSIM crop model and the random forest algorithm. The major objectives were to 1) develop a hybrid approach to forecast yield using a biophysical model and machine learning technique at a plot scale, 2) identify the optimum lead time with acceptable accuracy of yield forecasting, and 3) quantify the relative contribution of growth stage-specific predictors for determining end-of-season wheat yield.

2. Materials and methods

2.1. Study area

The study area was the New South Wales (NSW) wheat belt located in southeastern Australia (Fig. 1). The wheat belt is located between the Great Dividing Range and the arid interior of Australia. Topography and climatic conditions vary across the extent of the study area, showing a west-east gradient in both altitude and temperature/precipitation. The western areas consist of plains and the eastern areas are mainly mountains with elevations up to 1100 m. Average (1961–2000) growing season (for winter wheat, May–November) precipitation ranges from 171 mm in the southwest to 763 mm in the southeast, and average growing season temperature ranges from 8.3 °C in the southeast to
have been developed and used extensively in crop yield estimation and forecasting. Daily NDVI data (2008–2017, 500 m spatial resolution) for the study area were obtained from the MODIS/MOD09GA surface reflectance composites hosted by the Google Earth Engine (GEE, https://earthengine.google.com) platform.

2.2.3. Wheat trial data
Wheat variety experiment data (2008–2017) for the 29 sites (Fig. 1) were derived from the GRDC-NVT. The GRDC-NVT is a national project of crop variety testing that assists farmers in making variety decisions. These variety experiments were conducted at plot scale (~1.5 m × ~10 m) at selected trial sites. Three plots at each site were used to determine one observation. Wheat was harvested using well-maintained harvesting equipment at the earliest opportunity after physiological maturity of the plots to minimize grain losses through wind, rain, or pest damage. Other management practices, e.g. sowing, were performed in accordance with local farmers’ practices for a certain location. GRDC-NVT data included wheat yield and management practice information (sowing date, fertilization, etc.) for a variety of cultivars. Soil nutrient condition (including organic carbon, phosphorous, total nitrogen, conductivity, and pH) were also available. We chose the wheat cultivar, Sunvale, to conduct our study, as it was the most widely used cultivar across the study area and also has been parameterized in the APSIM model. The GRDC-NVT trial data were not available for all 10 years at each site, and as such we ultimately were able to use 231 sets of wheat trial data. These relatively recent data from experimental trials exhibited no significant technological trends. Therefore, no de-trending approach was implemented to exclude the effects of factors such as pesticide application, fertilizer practices, and varietal improvement, which were not reproduced by modelling.

2.2.4. Soil hydraulic properties
Detailed soil hydraulic properties for the 29 sites were acquired from the APSis database (http://www.asris.csiro.au) (Balaghi et al., 2008). There are more than 800 soil profiles available for Australian agricultural areas in this database. The majority of those soils are parameterized for wheat modelling. Soils that were identified to be geographically closest to our study sites were ultimately selected (Table 1). Using close APSis soil profiles as APSIM input is a common practice used in many other wheat modelling studies in Australia (Innes et al., 2015; Western et al., 2018). The same soil was selected for Coolah (site 6) and Merriwa (site 14).

2.3. Modelling methodology
2.3.1. APSIM simulations
In the present study, the APSIM (Agricultural Production System sIMulator, http://www.apsim.info/) crop model version 7.7 was implemented to simulate the dynamic changes of wheat phenology and biomass. As the APSIM crop model was developed by Australian institutions, it has been well calibrated in many wheat production areas throughout the Australian wheat belt. The APSIM simulations were set up strictly based on GRDC-NVT trials data (sowing date, variety, fertilization practice, hydraulic properties, and soil nutrient status) at the 29 sites. As stated previously, the Sunvale wheat variety is readily available in the APSIM variety database. The dynamic output of biomass and phenology information were then used to feed the statistical models.

2.3.2. Regression models
We used two regression models, multiple linear regression (MLR) and random forest (RF), to compare their performance in forecasting wheat yield. MLR is a widely used approach to explore the linear relation between the dependent and independent variables. Compared to ordinary least-squares method, MLR can be viewed as an extension that includes more than one predictor variable. It is easy to understand and
use, but usually limited in ability to disentangle the nonlinear relations between the dependent and independent variables.

RF is a popular machine learning method for various regression and classification purposes. In brief, RF is a nonparametric approach that builds numerous independent decision trees and assembles them together to gain a more accurate and stable prediction (Breiman, 2001). RF is capable of modelling nonlinear and hierarchical relationships between the response and the predictor variables. Our previous studies demonstrated the better performance of RF for exploring the nonlinear processes of the ARID index have been described by Woli et al. (2012). The eT \textsubscript{o} can be calculated based on the Priestley and Taylor (1972) method because it is assumed to be proportional to the potential evapotranspiration. Detailed characteristics and calculation processes of the ARID index have been described by Woli et al. (2012). The values of ARID range from 0 to 1. As values greater than 0.6 are considered to indicate severe water deficit, we selected 0.6 as the threshold to assess drought conditions.

We also included a metrological drought index, Standardized Precipitation and Evapotranspiration Index (SPEI) as a predictor in order to consider the impact of long-term drought on wheat yield. It is usually used to assess drought at monthly, seasonal, or longer time scales, rather than daily as the ARID index does. SPEI characterizes drought through standardizing the difference between precipitation

### Table 1
Basic information of the 29 study sites in New South Wales, Australia, including location, soil name (details at http://www.asris.csiro.au/), growing season (May-November) rainfall (GSR, mm), growing season temperature (GST, °C), number of years of yield data available (NY), and wheat yield (t ha \(^{-1}\)) range for recorded years (2008–2017).

<table>
<thead>
<tr>
<th>ID</th>
<th>Site</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Soil name</th>
<th>GSR</th>
<th>GST</th>
<th>NY</th>
<th>Yield range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Beckom</td>
<td>147.0</td>
<td>−34.2</td>
<td>Kandosol</td>
<td>298</td>
<td>12.9</td>
<td>9</td>
<td>1.1–4.8</td>
</tr>
<tr>
<td>2</td>
<td>Bellata</td>
<td>149.6</td>
<td>−29.9</td>
<td>Vertosol</td>
<td>307</td>
<td>15.9</td>
<td>9</td>
<td>1.4–5.3</td>
</tr>
<tr>
<td>3</td>
<td>Bullarab</td>
<td>149.3</td>
<td>−29.5</td>
<td>Vertosol</td>
<td>252</td>
<td>17.1</td>
<td>7</td>
<td>1.7–5.4</td>
</tr>
<tr>
<td>4</td>
<td>Canowindra</td>
<td>148.7</td>
<td>−33.5</td>
<td>Kandosol</td>
<td>342</td>
<td>12.6</td>
<td>10</td>
<td>1.3–6.0</td>
</tr>
<tr>
<td>5</td>
<td>Condobolin</td>
<td>147.2</td>
<td>−33.1</td>
<td>Sodosol</td>
<td>290</td>
<td>12.3</td>
<td>7</td>
<td>1.1–4.6</td>
</tr>
<tr>
<td>6</td>
<td>Coolah</td>
<td>149.7</td>
<td>−31.7</td>
<td>Vertosol</td>
<td>507</td>
<td>10.9</td>
<td>9</td>
<td>3.9–4.9</td>
</tr>
<tr>
<td>7</td>
<td>Coonamble</td>
<td>148.6</td>
<td>−31.0</td>
<td>Sodosol</td>
<td>267</td>
<td>15.4</td>
<td>10</td>
<td>1.4–5.5</td>
</tr>
<tr>
<td>8</td>
<td>Galong</td>
<td>146.5</td>
<td>−34.6</td>
<td>Kandosol</td>
<td>392</td>
<td>10.9</td>
<td>9</td>
<td>2.1–5.5</td>
</tr>
<tr>
<td>9</td>
<td>Gerogery</td>
<td>147.0</td>
<td>−35.9</td>
<td>Sodosol</td>
<td>386</td>
<td>12.2</td>
<td>10</td>
<td>3.1–6.9</td>
</tr>
<tr>
<td>10</td>
<td>Gilgandra</td>
<td>148.7</td>
<td>−31.6</td>
<td>Sodosol</td>
<td>313</td>
<td>14.3</td>
<td>10</td>
<td>1.7–4.8</td>
</tr>
<tr>
<td>11</td>
<td>Goonumbla</td>
<td>148.0</td>
<td>−32.9</td>
<td>Sodosol</td>
<td>331</td>
<td>13.6</td>
<td>9</td>
<td>1.3–5.6</td>
</tr>
<tr>
<td>12</td>
<td>Lockhart</td>
<td>146.8</td>
<td>−35.2</td>
<td>Sodosol</td>
<td>312</td>
<td>12.9</td>
<td>9</td>
<td>1.0–5.9</td>
</tr>
<tr>
<td>13</td>
<td>Mayrung</td>
<td>145.3</td>
<td>−35.5</td>
<td>Sodosol</td>
<td>251</td>
<td>13.1</td>
<td>8</td>
<td>4.4–7.4</td>
</tr>
<tr>
<td>14</td>
<td>Merriwa</td>
<td>150.3</td>
<td>−32.1</td>
<td>Vertosol</td>
<td>308</td>
<td>13.3</td>
<td>10</td>
<td>2.0–5.2</td>
</tr>
<tr>
<td>15</td>
<td>Merriwagga</td>
<td>145.7</td>
<td>−33.8</td>
<td>Kandosol</td>
<td>233</td>
<td>14.2</td>
<td>7</td>
<td>1.6–4.9</td>
</tr>
<tr>
<td>16</td>
<td>Narrimine</td>
<td>148.5</td>
<td>−32.1</td>
<td>Sodosol</td>
<td>289</td>
<td>14.5</td>
<td>1</td>
<td>5.2</td>
</tr>
<tr>
<td>17</td>
<td>North Star</td>
<td>150.4</td>
<td>−28.9</td>
<td>Vertosol</td>
<td>270</td>
<td>16.3</td>
<td>9</td>
<td>2.6–5.4</td>
</tr>
<tr>
<td>18</td>
<td>Nyngan</td>
<td>147.1</td>
<td>−31.6</td>
<td>Vertosol</td>
<td>253</td>
<td>15.5</td>
<td>7</td>
<td>1.0–4.7</td>
</tr>
<tr>
<td>19</td>
<td>Oaklands</td>
<td>146.1</td>
<td>−35.5</td>
<td>Sodosol</td>
<td>306</td>
<td>12.8</td>
<td>10</td>
<td>1.4–5.0</td>
</tr>
<tr>
<td>20</td>
<td>Quandialla</td>
<td>147.8</td>
<td>−34.0</td>
<td>Sodosol</td>
<td>317</td>
<td>13.1</td>
<td>9</td>
<td>1.7–5.0</td>
</tr>
<tr>
<td>21</td>
<td>Spring Ridge</td>
<td>150.2</td>
<td>−31.4</td>
<td>Vertosol</td>
<td>314</td>
<td>14.4</td>
<td>10</td>
<td>2.7–6.0</td>
</tr>
<tr>
<td>22</td>
<td>Temora</td>
<td>147.7</td>
<td>−34.3</td>
<td>Chromosol</td>
<td>305</td>
<td>12.2</td>
<td>8</td>
<td>2.0–5.3</td>
</tr>
<tr>
<td>23</td>
<td>Trangie</td>
<td>148.0</td>
<td>−32.0</td>
<td>Sodosol</td>
<td>273</td>
<td>14.7</td>
<td>9</td>
<td>1.2–5.5</td>
</tr>
<tr>
<td>24</td>
<td>Tullloona</td>
<td>150.1</td>
<td>−29.0</td>
<td>Vertosol</td>
<td>263</td>
<td>16.7</td>
<td>8</td>
<td>2.3–4.4</td>
</tr>
<tr>
<td>25</td>
<td>Wagga wagga</td>
<td>147.3</td>
<td>−35.0</td>
<td>Vertosol</td>
<td>364</td>
<td>12.1</td>
<td>7</td>
<td>2.2–5.2</td>
</tr>
<tr>
<td>26</td>
<td>Walgett</td>
<td>148.6</td>
<td>−30.2</td>
<td>Vertosol</td>
<td>239</td>
<td>16.3</td>
<td>6</td>
<td>0.9–5.2</td>
</tr>
<tr>
<td>27</td>
<td>Willbriggie</td>
<td>146.0</td>
<td>−34.4</td>
<td>Sodosol</td>
<td>249</td>
<td>13.7</td>
<td>3</td>
<td>4.5–6.1</td>
</tr>
<tr>
<td>28</td>
<td>Wongarbon</td>
<td>148.7</td>
<td>−32.3</td>
<td>Kandosol</td>
<td>347</td>
<td>13.5</td>
<td>9</td>
<td>1.3–5.1</td>
</tr>
<tr>
<td>29</td>
<td>Yenda</td>
<td>146.3</td>
<td>−34.3</td>
<td>Sodosol</td>
<td>265</td>
<td>13.8</td>
<td>2</td>
<td>3.7–4.4</td>
</tr>
</tbody>
</table>
and potential evapotranspiration (Vicente-Serrano et al., 2010). Thus, both precipitation and temperature, the two most relevant factors associated with drought, are considered in the calculation of SPEI. As a standardized index, values around 0 denote normal conditions, while values < −1 and > 1 indicate dry and wet conditions, respectively. In addition, the SPEI is developed in a way that can monitor drought at timescales from short (1 month) to long (24 months) periods. 1- to 6-month timescale SPEI values are usually used for meteorological and agricultural drought monitoring, while longer timescale SPEI values are more suitable for hydrological drought monitoring. Thus, in our study, we chose 1-, 3-, and 6-month timescale SPEI as explanatory predictors in the regression models to analyze drought impacts on crop yield. SPEI is usually calculated based on calendar month. Forecasting events in our study were triggered at the end of each growth category, which might fall on various dates. SPEI values calculated based on calendar month were therefore not likely to exactly cover the previous periods. Therefore, we defined 30 days, 90 days, and 180 days backward from the forecasting date as 1-, 3-, and 6-month timescales respectively, which enabled the SPEI calculation to consider all precipitation and temperature information over a particular timescale. Average NDVI, number of days with heat, frost, or ARID, and SPEI were calculated or counted for a specific crop growth stage generated from the APSIM simulations.

2.3.5. Modelling framework

The overall framework presented in Fig. 2 shows the order of procedures used in this study for in-season wheat yield forecasting model development and evaluation. APSIM simulated biomass and stage-specific climate and remote sensing information were used as predictors for the MLR and RF models to provide in-season yield forecasts. In detail, we first set up APSIM simulations based on information from GRDC-NVT trials (soil, variety, and management) and ran simulations dynamically according to known climate data. Known climate data were the data until the time point that we were forecasting from. A forecasting event would then be triggered at the completion of each growth category. In this study, six forecasting events (S, SG, T, SE, BAF, and M) were triggered successively. APSIM simulated biomass, SPEI, stage-specific ECEs, and NDVI, would be extracted or calculated from APSIM output, climate, and remote sensing information, except at S in which only SPEI was available. These factors were used as input for the forecasting models (RF and MLR) to forecast end-of-season yield.

As wheat phenology progressed, more and more factors would be gradually added to the forecasting models, which could result in increased computation times and over-fitted model due to correlated factors and the “curse of dimensionality” (Rodriguez-Galiano et al., 2012). Thus, we conducted a feature selection procedure before feeding the forecasting models. In the present study, a nonlinear method, genetic algorithm (GA) (Mitchell, 1998), was used to select the most relevant factors. GA is a search heuristic for function optimization based on Charles Darwin’s theory of natural evolution. For feature selection, a number of subsets from input predictor variables were first sampled as candidate solutions. Their corresponding fitness values (root mean square error, RMSE) were calculated to evaluate the quality of a subset. The subsets with the lowest RMSE were randomly combined to generate offsprings. In this process, two major genetic operators (mutation and crossover) were used, which could substantially affect the fitness value. Mutation takes effect through randomly removing or adding features in a subset. Crossover operates by producing a new subset through combining different features from a pair of subsets. Offspring replace the old generation based on the criterion that the new generation was better than the old one. This evolutionary process was repeated again and again until the termination of the search procedure. Many generations were then generated, which was likely to create better and better subsets. In the present study, we applied the GA method through the ‘caret’ package (Kuhn, 2008) coded in R software. Three main parameters were set according to Wang et al. (2018), i.e. mutation rate of 0.1, crossover rate of 0.8, and population size of 50. Selected predictors for each forecasting event based on the GA method are shown in Table 2.

Blending of climate and remote sensing indicators using statistical methods is a common practice in yield forecasting studies. However, most studies calculate indicators at monthly or longer time scale (Cai et al., 2019; Kern et al., 2018), which might be limited in considering stage-specific effects of some indicators. We firstly developed a wheat yield forecasting system which dynamically incorporated growth stage-specific ECEs, SPEI, NDVI, and APSIM-simulated biomass into the RF model. The MLR model was used as a benchmark. Given that crop yields often show nonlinear response to ECEs (Li et al., 2019), RF is expected to show good performance in exploring nonlinear relationships. The RF model was implemented with the ‘randomForest’ package (Liang and Wiener, 2002) based in the R software. The parameters in the RF model were set with mtry (the number of randomly selected predictor variables at each node) = the number of predictor variables divided by 3 (rounded down) and ntree (the number of trees to grow in the forest) = 500. The relative importance of variables was estimated using the “%IncMSE” metric in the RF model.

2.3.6. Model performance assessment

The 10-year observed wheat yield data from the GRDC-NVT trials were used for comparisons with modeled yields. In doing so, a leave-one-year-out (a whole year of data for all sites was left out) cross validation method was applied to the 231 data sets. Both deterministic and probabilistic measurements were adopted to evaluate model performance. The deterministic measurements included Pearson’s correlation coefficient (r), Lin’s concordance correlation coefficient (LCCC) (Lin, 1989), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). These measurements were defined as follows:

\[ r = \frac{\sum_{i=1}^{n} (O_i - \bar{O})(F_i - \bar{F})}{\sqrt{\sum_{i=1}^{n} (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^{n} (F_i - \bar{F})^2}} \]

\[ LCCC = \frac{2r_{O,F} - r^2}{\sigma_O^2 + \sigma_F^2 + (\bar{O} - \bar{F})^2} \]

\[ MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{F_i - O_i}{O_i} \right| \]

\[ RMSE = \sqrt{\frac{\sum_{i=1}^{n} (F_i - O_i)^2}{n}} \]

where \( F_i \) and \( O_i \) represent forecasted and observed values, \( \bar{F} \) and \( \bar{O} \) denote the mean forecasted and observed values, \( \sigma_O \) and \( \sigma_F \) are the variances of observed and forecasted values, \( n \) indicates the number of the samples. RMSE and MAPE indicate the average magnitude of the errors in a set of forecasts. \( r \) measures the strength of the linear relationship between observations and forecasts. LCCC denotes the degree to which forecasted and observed values follow the 1:1 line through the origin. Model forecasts become increasingly accurate as RMSE and MAPE approach 0 and \( r \) and LCCC approach 1.

We also applied a probabilistic forecasting performance measurement: receiver operating characteristic (ROC) score (Fawcett, 2006). The ROC curve was generated by plotting the true positive rate against the false positive rate across various cut-off settings. The ROC score is defined as the area under the curve, which ranges from 0 to 1. Generally, forecasting models with higher scores are considered more skillful. The ROC analysis is not sensitive to forecasting biases, which is said to be helpful for finding potential skill without considering biases similar to the correlation coefficient (Jha et al., 2019). A detailed description of the ROC analysis can be found in Fawcett (2006). ROC score is a useful measurement of model performance for classification tasks. In the present study, observed yields were first categorized into three terciles: below normal (0% to 32%), normal (33% to 66%), and
above normal (67% to 100%). Thus, yield forecast probabilities for each category could be calculated. ROC score was calculated for forecasts of the RF and MLR models for each forecasting event using the R package ‘pROC’ (Robin et al., 2011).

3. Results

3.1. Model performance

In this study, APSIM dynamic output, SPEI, stage-specific ECEs, and NDVI were included as predictors into MLR and RF models to make pre-harvest yield forecasts. The performance of both regression models was evaluated by a leave-one-year-out cross validation procedure. Fig. 3 presents the observed and forecasted yields at the 29 study sites from 2008 to 2017 in the NSW wheat belt. Generally, observed wheat yield varied greatly from year to year in the study area, with a low of 2.5 t ha⁻¹ in 2009 and a high of 5 t ha⁻¹ in 2016. The temporal variations of the observed wheat yield were successfully captured by both the MLR and RF models, and the accuracy usually improved as growth stage progressed. The RF model tended to better predict observed yields than the MLR model, especially in years with atypical yields, such as 2009 and 2016. Moreover, forecasted yields for each year from the RF model were less variable than from the MLR model over the course of the six growth stage events, meaning that the RF model could provide better forecasts even at earlier growth stages.
two models using the four deterministic metrics (r, LCCC, MAPE, and RMSE) are shown in Fig. 4. In general, the yield forecast accuracy for both models increased with the progress of wheat growth, as demonstrated by the gradual increase in r and LCCC and decrease in MAPE and RMSE. At the first two forecasting events, S and SG, the two models both showed poor performance in forecasting end-of-season yield, with MAPE values around 40% and RMSE above 1.30 t ha\(^{-1}\). From SG to BAF, the forecasting accuracy increased significantly for both models. For the MLR model, r increased from 0.15 to 0.74, LCCC increased from 0.11 to 0.71, MAPE decreased from 38.4% to 21.3%, and RMSE decreased from 1.33 t ha\(^{-1}\) to 0.86 t ha\(^{-1}\). Similarly, for the RF model, r increased from 0.13 to 0.85, LCCC increased from 0.08 to 0.81, MAPE decreased from 38.8% to 17.6%, and RMSE decreased from 1.30 t ha\(^{-1}\) to 0.70 t ha\(^{-1}\). It should be noted that at each growth stage forecasting event, the RF model always had greater accuracy than the MLR model as denoted by the four metrics. In addition, forecasting accuracy increased little for both regression models from BAF to M.

The evaluation of model performance using the probabilistic measurement, ROC score, is shown in Fig. 5. The ROC scores of forecasts at the six different forecasting events are consistent with the previously described results from the deterministic measurements, i.e., the ROC scores of forecasts improved as wheat developed. At S and SG, the ROC scores were marginally greater than 0.5, reflecting poor forecasting performance by both models. However, from T onwards, both models had satisfactory forecasting performance, as denoted by ROC scores > 0.6, with ROC scores increasing rapidly from SG to BAF. As with the

**Table 2**


<table>
<thead>
<tr>
<th>No.</th>
<th>S</th>
<th>SG</th>
<th>T</th>
<th>SE</th>
<th>BAF</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S_SPEI_1</td>
<td>SG_biomass</td>
<td>T_biomass</td>
<td>SE_biomass</td>
<td>BAF_biomass</td>
<td>M_biomass</td>
</tr>
<tr>
<td>2</td>
<td>S_SPEI_3</td>
<td>S_SPEI_1</td>
<td>S_SPEI_6</td>
<td>T_SPEI_1</td>
<td>SE_SPEI_6</td>
<td>BAF_SPEI_1</td>
</tr>
<tr>
<td>3</td>
<td>S_SPEI_6</td>
<td>SG_SPEI_3</td>
<td>SE_SPEI_3</td>
<td>SE_SPEI_6</td>
<td>BAF_ARID</td>
<td>M_SPEI_1</td>
</tr>
<tr>
<td>4</td>
<td>SG_SPEI_3</td>
<td>S_SPEI_1</td>
<td>T_SPEI_1</td>
<td>SE_SPEI_6</td>
<td>BAF_ARID</td>
<td>SG_ARID</td>
</tr>
<tr>
<td>5</td>
<td>SG_ARID</td>
<td>SG_ARID</td>
<td>T_ARID</td>
<td>SE_ARID</td>
<td>BAF_Heat</td>
<td>BAF_ARID</td>
</tr>
<tr>
<td>6</td>
<td>SG_Frost</td>
<td>T_ARID</td>
<td>SE_ARID</td>
<td>BAF_Frost</td>
<td>BAF_Heat</td>
<td>SG_Frost</td>
</tr>
<tr>
<td>7</td>
<td>SG_NDVI</td>
<td>SG_Frost</td>
<td>SE_ARID</td>
<td>BAF_Frost</td>
<td>BAF_Heat</td>
<td>BAF_Frost</td>
</tr>
<tr>
<td>8</td>
<td>SG_Frost</td>
<td>BAF_Frost</td>
<td>BAF_NDVI</td>
<td>BAF_Heat</td>
<td>BAF_ARID</td>
<td>ARID</td>
</tr>
<tr>
<td>9</td>
<td>SG_Frost</td>
<td>BAF_Frost</td>
<td>BAF_NDVI</td>
<td>BAF_Heat</td>
<td>BAF_ARID</td>
<td>ARID</td>
</tr>
<tr>
<td>10</td>
<td>BAF_Frost</td>
<td>BAF_Frost</td>
<td>BAF_NDVI</td>
<td>BAF_Heat</td>
<td>BAF_ARID</td>
<td>ARID</td>
</tr>
</tbody>
</table>

![Fig. 3](image_url)  
**Fig. 3.** Time series of observed and model-forecasted wheat yields based on the six forecasting events from 2008 to 2017. Wheat yields for each year were averaged across the 29 study sites (results for each site can be found in Figure S1 in the supplementary material). Data were generated from the leave-one-year-out cross validation procedure from the two regression models, MLR: multiple linear regression and RF: random forest. Observed and six forecasted wheat yields are shown as gray circles and colored shapes, respectively. OB: observed, S: sowing, SG: end of seedling growth, T: end of tillering, SE: end of stem elongation, BAF: end of flowering, and M: end of milk development.
deterministic measures of model performance, the rate of increase in ROC scores from BAF to M was much slower than at previous growth stages. In general, the RF model had larger ROC scores at all forecasting events. The highest ROC score was 0.9, achieved by the RF model at M.

3.2. Optimum forecasting event analysis

According to the model performance results, the RF-based forecasting system showed better performance than the MLR-based model for our study area. In general, stakeholders prefer to obtain an accurate yield forecast as early as possible. However, there is a tradeoff between greater accuracy and longer lead time in any yield forecasting system (Basso and Liu, 2018). In our study, model accuracy gradually increased as the growing season progressed towards harvest, but slowed down from BAF to M (Figs. 4 and 5). To determine the optimum forecasting event with respect to the stakeholders’ objective, we analyzed the relative changes in model accuracy as growing season progressed. This was done by normalizing the four model performance measurements from 0 to 100% by the following equation

\[ x_{\text{new}} = \left( \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \right) \times 100\% \]  

where \( x \) represents either \( r, \) LCCC, MAPE, RMSE, or ROC. This normalization procedure allowed a comparison of the magnitude of change for each performance measurement at each of the six forecasting events. As shown in Fig. 6, \( r_n, \) LCCC\( \text{r}_n, \) and ROC\( \text{n} \) score continued to increase from S to M (except for ROC score from S to SG), and MAPE\( \text{n} \) and RMSE\( \text{n} \) continued to decrease from S to M. The greatest increase in
MAPEn decreased by 30.4%, and RMSEn decreased by 46.2%. However, LCCCn increased by 24.4%, LCCCn increased by 33.3%, ROCn score increased by 47.6%. This was expected as biomass is cumulated with crop growth and development.

**3.3. Relative importance of growth selected predictors**

The RF model can provide a list of the relative importance of predictor selection (Were et al., 2015). Importance values of input predictors (selected by GA, Table 2) for each forecasting event are presented in Fig. 7. There was an increasing trend of the importance values of APSIM-simulated biomass from SG to BAF. APSIM-simulated biomass ranked first in the last three forecasting events. This was expected as biomass is cumulated with crop growth and development. While for other indices, drought was the most influential factor affecting wheat yields in the study area, as the first three indices were all drought related indices at each forecasting event. At the first four forecasting events, SPEI ranked first, but at BAF and M, stage-specific drought (indicated by the ARID index) exceeded SPEI. This may because wheat yield was very sensitive to daily time scale drought events at BAF and M. Nevertheless, SPEI is a useful index to reflect drought-induced yield losses, as SPEI was involved in all forecasting events and usually ranked high (Fig. 7). ARID and frost events that occurred during SG were two indices that were consistently selected at the various forecasting events, meaning that ARID and frost from sowing to the end of seedling growth are likely to greatly affect final wheat yields. Heat and frost events during BAF also had great impacts on final yield and were both selected at the BAF and M forecasting events. In contrast, NDVI was relatively unimportant, only being selected at the SG and BAF forecasting events and was not highly ranked at either time.

**4. Discussion**

In this study, we combined a crop simulation model (APSIM) with statistical regression-based models to dynamically forecast wheat yield at several points during the growing season based on growth stage-specific climate and remote sensing indices. The APSIM + RF hybrid model obtained satisfactory results in yield prediction. This was primarily because we succeeded in exploiting the merits of each model. Our models not only took advantage of biophysical processes among crop, soil, management, and climate information, but also made use of machine learning technique to account for climate extremes and remote sensing information. In addition, the machine learning technique used in the study showed an overall advantage over traditional regression methods (Figs. 4 and 5) in exploring the relationships between crop yield and environmental factors. Given the increasing availability of farming-related, climate, and remote sensing data, Keating and Thorburn (2018) introduced the blending of advanced statistical and mechanistic models within crop-environment research fields. The yield forecasting system proposed in our study can be regarded as a feasible approach capable of being extended to other wheat cropping areas in order to gain new insights that will guide agricultural practice and grain marketing.

Another advantage of our RF-based wheat yield forecasting system is that it accounts for wheat growth stage-specific ECEs. Generally, crop growth for a given season is subjected to two types of climatic conditions, i.e., mean climate conditions and climate extremes (Challinor et al., 2007). More mean climate conditions tend to result in high harvestable yield, while climate extremes generally lead to yield loss. Most crop models adequately simulate the effects of mean climate conditions but encounter limitations when estimating yield losses due to climate extremes (Barlow et al., 2015). In APSIM, impacts of drought (water stress) are incorporated by stress functions that can restrain biomass accumulation and leaf expansion, while heat and frost are defined as functions which can lead to leaf senescence (Zheng et al., 2014). However, most of these functions are simple and linear. Additionally, other damaging effects, such as heat- or frost-induced sterility, are not considered in APSIM. Thus, APSIM has some limitations in accurately estimating the effects of climate extremes on yields. While statistical regression-based models are able to determine relationships between yield and variables quantifying climate extremes, the variables selected are usually vague and not stage-specific. The majority of previous research using regression models has extracted variables based on long time periods, typically covering the entire growing season (Pinke and Lövei, 2017; Wang et al., 2015). This approach may result in inaccurate estimations of yield losses as different crop growth stages can have different tolerances for the same extreme event. For example, Baigorria et al. (2007) demonstrated that the
timing and duration of dry periods had different impacts on final maize 
(Zea mays L.) yield. Kern et al. (2018) emphasized the importance of 
shorter timescales for variable calculation and prepared predictors at 
monthly resolution, which explained 67% of the variation in winter 
wheat yields. However, they still did not associate the predictors with 
concrete growth stages such as anthesis or grain filling. In our study, we 
acquired crop phenology information by dynamically running APSIM 
simulations that triggered forecasting events at the end of specific 
growth stages. Indices of climate extremes were then calculated ac-
cording to past growth stages. The APSIM-simulated biomass can be 
viewed as representative of the mean climate condition. Thus, our 
model considered the effects of mean climate conditions and stage-
specific climate extremes simultaneously, thereby resulting in satisfac-
tory yield forecasting results.

In our study, the crop physiological interpretation was easily un-
derstood as the yield predictors were associated with concrete growth 
stages. For these growth stage-specific indices, our results identified 
biomass, which integrated drought effects, as the most important factor 
determining yield in the study area. This result was expected as drought 
is a recurring feature of Australia’s climate (Ummenhofer et al., 2009), 
and has caused severe yield losses during past decades (Feng et al., 
2018). In the present study, we found that dry periods during seedling 
growth are likely to result in carryover effects on wheat growth, while 
dry periods during anthesis and grain filling are major factors de-
termining final yield. This is because insufficient water supply can re-
strain leaf expansion and root growth during vegetative growth stages 
(Chaves et al., 2002), and can reduce mobilization of carbohydrates 
from vegetative organs to grain during reproductive growth stages 
(Royo et al., 2006). Meanwhile, SPEI was identified as a potential 
drought index reflecting the effects of water deficiency on crop yield in 
this study area as it was always chosen by the RF model as a variable 
having high importance. Frost and heat events during reproductive 
growth stages and frost events during seedling growth also had im-
portant impacts on final yield. This is because these stages are more 
sensitive to temperature anomalies (Hlaváčová et al., 2018). In con-
trast, the remotely sensed NDVI did not significantly improve model 
accuracy and was only selected as important by the RF model at the SG 
and BAF forecasting events, and even at those two stages, NDVI was not 
ranked high as an important variable influencing wheat yield (Fig. 7). 
This might be due to the plot-scale data used in the study. NDVI values 
with spatial resolution of 500 m did not have sufficiently fine resolution 
to reflect the vegetation conditions of specific experimental plots. When 
applying this system for yield forecasts at a larger scale, remote sensing 
information is likely to contribute more to model accuracy.

Our proposed yield forecasting system has great potential for 
practical agricultural production systems. In the present study, we 
achieved satisfactory results for plot-level wheat yield forecasting with 
a 35-day lead time (r = 0.85, LCCC = 0.81, MAPE = 17.6%, 
RMSE = 0.70 t ha⁻¹, and ROC score = 0.90) and with a two-month 
lead time (r = 0.62, LCCC = 0.53, MAPE = 27.1%, 
RMSE = 1.01 t ha⁻¹, and ROC score = 0.88). In comparison with other 
yield forecasting studies conducted in Australia, Filipi et al. (2019) used 
multiple data sources (e.g., soil, climate, and remote sensing) and 
random forest algorithm to forecast yield of three winter crops (wheat, 
barley, and canola) in the southern agricultural region of Western 
Australia. They developed one model with crop type as a predictor 
variable rather than three individual models for each crop, and ob-
tained a result of RMSE = 0.36 t ha⁻¹ and LCCC = 0.92 for the late-
season (September) forecast. Cai et al. (2019) compared the perfor-
mance of three machine learning models and one linear regression 
method to forecast wheat yield in Australia using climate and remote 
sensing data, and achieved the optimal prediction performance 
(R² = 0.73) with two-month lead time before harvest. However, re-

gion-level wheat yield records (rather than plot-level wheat yields as 
used in our study) were used in their study. Wheat yields at a smaller 
scale are more difficult to forecast due to the variable conditions even 
within the same region, which tends to require more kinds of data 
sources with finer resolution. Further development of data acquisition 
techniques will allow the acquisition of more detailed farming-related 
data (Filipi et al. 2019), such as agronomic information, soil moisture 
conditions (Peng et al., 2017), and high spatiotemporal remote sensing 
images for a growing season (Zambrano et al., 2018). Our proposed 
method can also be extended to and validated for larger areas to de-
termine crop yield outlooks.

Our forecasting system satisfactorily predicted yield in the NSW 
heat belt. However, this method requires a large amount of data from 
different sources (including soil, climate, crop, and management) to 
drive the APSIM crop model. These data are usually not available for 
larger-scale (e.g., districts and countries) forecasts or in data-poor 
areas. Nevertheless, with societal and economic developments, more
and more areas will have sufficient data available to implement this yield forecasting method. In addition, El Nino Southern Oscillation (ENSO)-related indices are also frequently used indicators for yield forecasting in Australia, but were not considered in our study. Future studies using a similar modelling approach to ours may use additional information (such as ENSO-related indices) and potentially achieve even greater forecasting accuracy.

5. Conclusions

In the present study, we succeeded in developing a wheat yield forecasting system by incorporating multiple data sources such as crop model output, indices of extreme climate, and remote sensing information into two statistical regression-based models. Plot-level wheat yields were dynamically forecasted at the end of targeted growth stages during the growing season progressing to harvest. Stage-specific extreme climate events were fully considered in the system. We found that the machine learning-based model produced more accurate forecasts of wheat yield than the traditional multiple linear regression model. The optimum forecasting events that produced sufficiently accurate yield predictions were those providing one- and two-month lead times. We expect that this forecasting system using crop simulation modelling and a machine learning method specifically addressing the effects of stage-specific climate extremes on crop yield can be used for operational forecasting purposes in Australia and potentially other similar dryland cropping systems around the world. With the further development of information technology and remote sensing technology, our proposed system can be directly extended to region- and country-scale forecasts. This yield-forecasting method will become increasingly important in providing information to mitigate the detrimental effects of climate change on global food supply.

Declaration of Competing Interests

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.

Acknowledgements

This study was jointly supported by the China Scholarship Council (CSC), the International Partnership Program of the Chinese Academy of Sciences (16146KYSB20170013), and the Chinese Academy of Sciences “Light of West China” Program. Facilities for conducting this study were provided by the New South Wales Department of Primary Industries and University of Technology Sydney. Thanks to Grains Research and Development Corporation (GRDC) National Variety Trials (NVT), Australia (http://www.nvtonline.com.au/) for their data support. We are grateful to the three anonymous reviewers whose comments have greatly improved this manuscript.

Supplementary materials


Reference


Nandram, B., Berg, E., Barboza, W., 2014. A hierarchical Bayesian model for forecasting