Environmental Research Letters

OPEN ACCESS

RECEIVED
13 February 2020

REVISED
28 April 2020

ACCEPTED FOR PUBLICATION
19 June 2020

PUBLISHED
12 August 2020

LETTER

Machine learning-based integration of large-scale climate drivers can improve the forecast of seasonal rainfall probability in Australia

Puyu Feng[1,2], Bin Wang[1,2], De Li Liu[2,3], Fei Ji[4], Xiaoli Niu[2,5], Hongyan Ruan[6], Lijie Shi[2,7] and Qiang Yu[1,7,8]

[1] State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Northwest A&F University, Yangling, Shaanxi 712100, People's Republic of China
[2] NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, Wagga Wagga, NSW 2650, Australia
[3] Climate Change Research Centre and ARC Centre of Excellence for Climate Extremes, University of New South Wales, Sydney, NSW 2052, Australia
[4] Department of Planning, Industry and Environment, Queanbeyan, NSW 2620, Australia
[5] College of Agricultural Equipment Engineering, Henan University of Science and Technology, Luoyang, Henan 471000, People's Republic of China
[6] Technology and Key Laboratory of Beibu Gulf Environment Change and Resources Use Utilization of Ministry of Education, Nanning Normal University, Nanning 530001, People's Republic of China
[7] School of Life Sciences, Faculty of Science, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia
[8] College of Resources and Environment, University of Chinese Academy of Science, Beijing 100049, People's Republic of China

E-mail: yuq@nwafu.edu.cn, bin.a.wang@dpi.nsw.gov.au and de.li.liu@dpi.nsw.gov.au

Keywords: seasonal rainfall forecasting, climate drivers, random forest

Supplementary material for this article is available online

Abstract

Probabilistic seasonal rainfall forecasting is of great importance for stakeholders such as farmers and policymakers to assist in developing risk management strategies and to inform decisions. In practice, there are two kinds of commonly used tools, dynamical models and statistical models, to provide probabilistic seasonal rainfall forecasts. Dynamical models are based on physical processes but are usually expensive to operate and implement, and rely overly on initial conditions. Statistical models are easy to implement but are usually based on simple or linear relationships between observed variables. Recently, machine learning techniques have been widely used in climate projection and perform well in reproducing historical climate. For these reasons, we conducted a case study in Australia by developing a machine learning-based probabilistic seasonal rainfall forecasting model using multiple large-scale climate indices from the Pacific, Indian and Southern Oceans. Rainfall probabilities of exceeding the climatological median for upcoming seasons from 2011 to 2018 were successively forecasted using multiple climate indices of precedent six months. The performance of the model was evaluated by comparing it with an officially used forecasting model, the SOI (Southern Oscillation Index) phase model (SP) operated by Queensland government in Australia. Results indicated that the random forest (RF) model outperformed the SP model in terms of both distinct forecasts and forecasting accuracy. The RF model increased the percentages of distinct forecasts to 64.9% for spring, to 71.5% for summer, to 65.8% for autumn, and to 63.9% for winter, 1.4 ~ 3.2 times of the values from the SP model. Forecasting accuracy was also greatly increased by 28%, 167%, 219%, and 76% for four seasons respectively, compared to the SP model. The proposed rainfall forecasting model is based on readily available data, and we believe it can be easily extended to other regions to provide seasonal rainfall outlooks.

1. Introduction

Rainfall is a natural phenomenon that results from complex global and regional atmospheric processes. Forecasting terrestrial rainfall several months in advance has significant implications for more efficient usage of water resources, *e.g.* agricultural planning (He *et al* 2014). However, accurate and reliable

seasonal rainfall forecasting remains a great challenge for scientific community, which limits the prospective use of natural resources to guide production activities of mankind.

Australia is among the world's largest agricultural exporters (Gunasekera *et al* 2007). For example, Australian wheat commodity contributes roughly 15% of global annual wheat trade (www.aegic.org.au). Thus, Australian agricultural sector is very important to ensure global food supply and security (Qureshi *et al* 2013). However, highly variable interannual seasonal rainfall exerts serious adverse impacts on Australian agricultural productivity (Cobon and Toombs 2013). For example, the drought in 2018 has resulted in yield loss of 53% in eastern Australia compared to the average of past two decades (https://www.agriculture.gov.au/abares). Researchers have developed different seasonal rainfall forecasting tools for decision-makers to deal with high rainfall variability in order to minimize losses in potentially 'bad' seasons and maximize profits in potentially 'good' seasons (Stone *et al* 1996, Mekanik *et al* 2016).

At present, there are two kinds of official seasonal rainfall forecasting programs in Australia, dynamical models and statistical models. Dynamical models are often referred to coupled ocean–atmosphere general circulation models, which are based on the laws of physical processes. The most typical dynamical model in Australia is the ACCESS–S (Australian Community Climate Earth-System Simulator–Seasonal, launched in 2018) developed and run by the Australian Bureau of Meteorology (www.bom.gov.au/climate/ahead/about/model/access.shtml). This model is able to initiate rainfall outlooks for the season ahead as the probability of receiving above median rainfall. Major advantages of dynamical models are that they have the capacity to simulate nonlinear interactions of meteorological processes and are adaptable to climate shift (Schepen *et al* 2012). However, dynamical models are usually expensive to implement and operate, and they are overly dependent on initial conditions. Despite substantial technological advances and research efforts, sophisticated dynamical models are still unable to consistently outperform simple statistical models for forecasting seasonal rainfall (Abbot and Marohasy 2014, Mekanik *et al* 2016).

Statistical models are also extensively used in Australia to issue seasonal rainfall outlooks (He *et al* 2014) with the format of the probability of exceeding the seasonal median (Fawcett and Stone 2010). Statistical models employ empirical relationships between the response variable and various predictor variables to generate forecasts. Therefore, these models depend on the availability of observed data and stationary relationships between the variables (Schepen *et al* 2012). One typical statistical forecasting model in Australia is the Southern Oscillation Index (SOI) phase seasonal rainfall forecasting program (Stone *et al* 1996), which is currently operated by the Queensland Government (https://www.longpaddock.qld.gov.au/seasonal-climate-outlook/rainfall-probabilities/). This program was developed based on the great and lagged impacts of El Nino Southern Oscillation (ENSO) from the Pacific on Australia's climate. Pairs of consecutive monthly SOI values are categorized into five kinds of phases using principal components analysis and cluster analysis. Rainfall probability (exceeding median) of upcoming three months can be quantified based on historical situations with a same SOI phase (Stone *et al* 1996). The SOI phase forecasting program has been widely used by crop producers and pastoral industries to reduce climate-related risks (Cobon and Toombs 2013). However, this program usually has poor performance in western Australia, where the impacts of ENSO are weak due to large spatial distance (Risbey *et al* 2009). Moreover, oceanic activities from the Southern Ocean and the Indian Ocean, *e.g.* Southern Annular Mode (SAM) (Thompson and Wallace 2000) and Indian Ocean Dipole (IOD) (Saji *et al* 1999), also show regulatory effects on the variability of Australian seasonal rainfall (Risbey *et al* 2009). Thus, a forecasting method based solely on ENSO may not be sufficient to be applied in the whole continent or all seasons. In addition, the SOI phase program usually generates probability values of around 50%, however, intermediate probability of exceeding the median is not particularly helpful for making decisions.

In recent years, machine learning algorithms have gradually received wide attention in both classification and regression tasks with the development of artificial intelligence (Aguasca-Colomo *et al* 2019, Scher and Messori 2019). Machine learning algorithms are capable of investigating hierarchical and nonlinear relationships between the response variable and predictor variables based on ensemble learning approaches (Shalev-Shwartz and Ben-David 2014). In seasonal rainfall forecasting, predictor variables may comprise various preceding large-scale climate signals. For example, Hartmann *et al* (2008) used artificial neural network to forecast summer rainfall in the Yangtze River basin using large-scale climate indices including SOI and the Scandinavia pattern. Kashid and Maity (2012) used genetic programming to predict Indian Summer Monsoon Rainfall using large-scale climate signals from both tropical Indian Ocean and tropical Pacific Ocean. However, machine learning-based forecasting methods have rarely been used for forecasting seasonal rainfall probability in Australia (Abbot and Marohasy 2014).

The present study employed a machine learning method with multiple large-scale climate indices aiming for developing a skillful and robust seasonal rainfall forecasting technique. We took the rainfall forecasting results obtained from the SOI phase

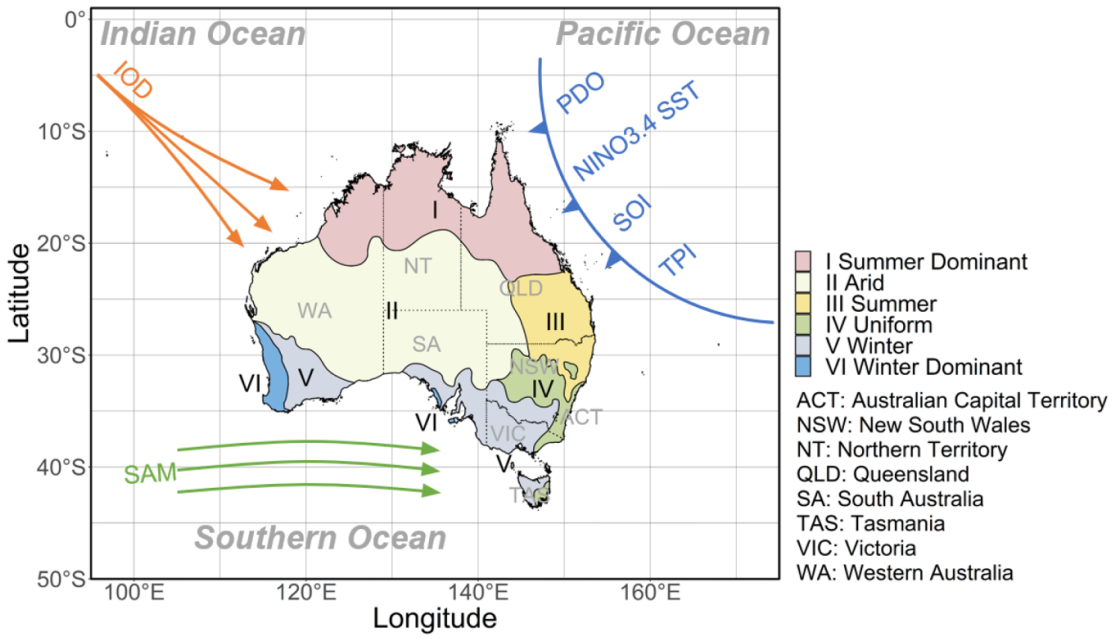

Figure 1. Map of climate zones of Australia and influential large-scale climate indices from its surrounding oceans. IOD: Indian ocean dipole; SAM: southern annular mode; NINO3.4 SST: NINO3.4 sea surface temperature; PDO: Pacific decadal oscillation; SOI: southern oscillation index; TPI: the interdecadal Pacific oscillation tripole index.

forecasting program as the benchmark to compare whether our proposed machine learning method can better predict rainfall probability in Australia.

2. Materials and methods

2.1. Study area

The study area covers the whole Australian continent, with latitude ranging from 10°S to 44°S and longitude ranging from 112°E to 154°E (figure 1). Due to the large geographical size of the country, Australia has a wide variety of climates which have been classified into six distinct climate zones (figure 1) based on seasonal rainfall (https://www.bom.gov.au). The northern and northeastern zones of Australia have a more tropical influenced climate, with humid and hot austral summers (Dec-Feb) and dry and warm austral winters (Jun-Aug). The southern coastal zones have a Mediterranean-like climate, dry and hot during summers and wet and mild during winters. In addition, central interior areas are dominant by a desert climate, mostly governed by sinking air of the subtropical high-pressure belt (Turney *et al* 2007).

2.2. Climate data

2.2.1. Rainfall data.

Long-term (1889–2018) historical monthly rainfall data for 0.05° × 0.05° grids (figure 2) across Australia were obtained from Scientific Information for Land Owners (SILO, www.longpaddock.qld.gov.au/silo/), which is currently hosted by the Queensland government. SILO rainfall dataset is constructed based on observational records provided by the Australian Bureau of Meteorology. Missing data in observational time series and gridded rainfall data are both derived using ordinary kriging interpolation technique. SILO rainfall dataset is readily available for climate applications and has been well tested in many climate-related studies (Boer *et al* 2016, Williamson *et al* 2016, Wang *et al* 2018). In our study, 0.05° × 0.05° spatial resolution was too high and might result in too much unnecessary computational load. Thus, we firstly reduced the spatial resolution and obtained 2780 grids (0.5° × 0.5°) across Australia. Rainfall medians of four seasons for 2780 grids based on data in 1889–2018 are presented in figure 2.

2.2.2. Large-scale climate indices.

The inter-annual variability of Australia's seasonal rainfall is regulated by climatic activities of three surrounding oceans, the Pacific, Indian, and Southern Oceans (Risbey *et al* 2009). These climatic activities include sea surface air pressure fluctuation, sea surface temperature (SST) fluctuation, atmospheric circulation (*e.g.* Walker and Hadley cells), etc. They exert complex impacts on Australia's rainfall and most impacts remain elusive. In the past decades, a number of large-scale climate indices have been introduced to describe various aspects of oceanic activities. For example, SOI is calculated based on the sea surface air pressure differences between Tahiti and Darwin, which is one of the key climate indices that measure the strength of ENSO-related events in the Pacific Ocean. We collected 6 influential and commonly used large-scale climate indices as potential predictors. A brief description of each involved index

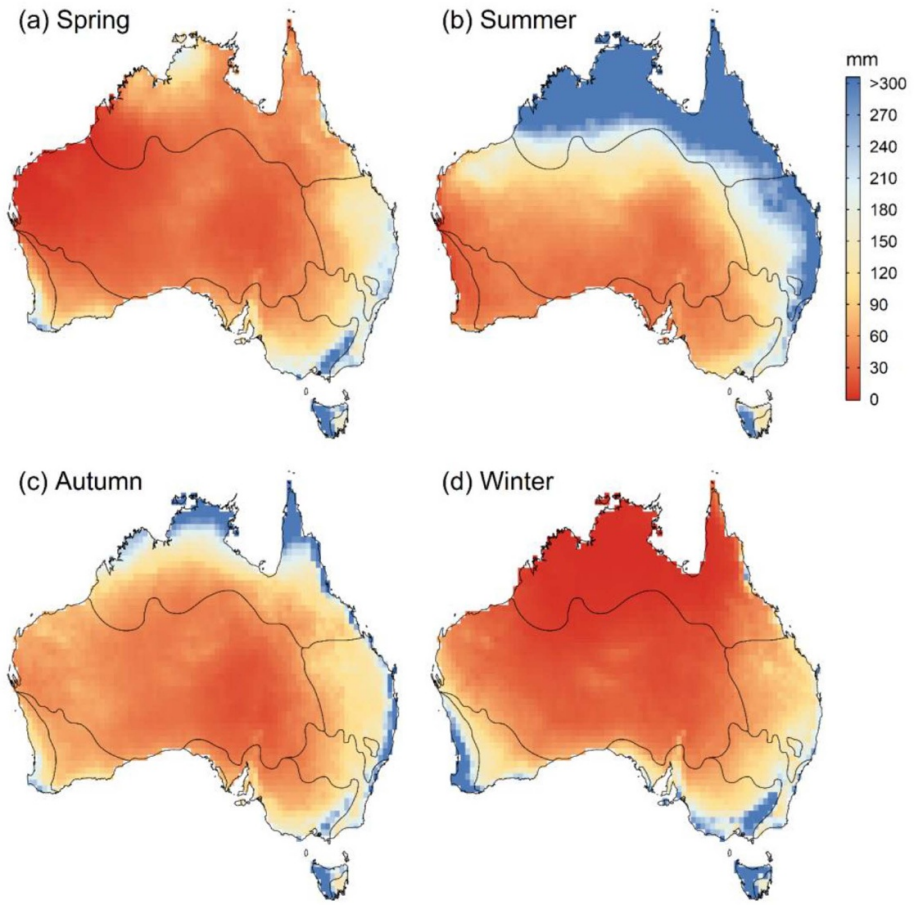

Figure 2. Rainfall medians of four seasons for 2780 grids (0.5° × 0.5°) across Australia based on rainfall data in 1889–2018.

is presented in table 1. Monthly series from 1889 to 2018 for the six indices except SAM were directly obtained from Earth System Research Laboratory (ESRL, https://www.esrl.noaa.gov/). While for SAM, we re-calculated it using the generation code and Hadley Centre Sea Level Pressure dataset from ESRL.

2.3. SOI phase model

The SOI phase (SP) seasonal rainfall forecasting model can provide probabilistic forecasts of rainfall exceeding the median for upcoming three months across Australia (Stone *et al* 1996). The SP model is theoretically based on prognostic features of SOI on rainfall conditions of upcoming few months in Australia. Pairs of consecutive monthly SOI values are categorized into five kinds of phases (consistently negative, consistently positive, rapidly falling, rapidly rising, and consistently near zero) using principal components analysis and cluster analysis. Rainfall probability (exceeding median) of upcoming few months can be quantified based on historical situations with a same SOI phase. In practice, the SP model is explicitly adopted for quantifying rainfall probabilities of upcoming three months with 'zero' lead time (*e.g.* SOI values for April and May are used to predict June-August rainfall). It can be expressed as follows:

$$P_{I,j} = \frac{\sum_{i=1889}^{I-1} N_{SR_{i,j},\,SP_{i,j}}}{\sum_{i=1889}^{I-1} N_{SP_{i,j}}} \times 100\% \tag{1}$$

$$N_{SR_{i,j},\,SP_{i,j}} = \begin{cases} 1, \text{ if } SR_{i,j} > \text{the median } \& \; SP_{i,j} = SP_{I,j} \\ 0, \text{otherwise} \end{cases} \tag{2}$$

$$N_{SP_{i,j}} = \begin{cases} 1, \text{if } SP_{i,j} = SP_{I,j} \\ 0, \text{othewise} \end{cases} \tag{3}$$

where j and I represent a target three-month period j of a target year I. $\boldsymbol{P}_{I,j}$ denotes forecasted rainfall probability of exceeding the climatological median for year I period j. SR and SP are seasonal rainfall and SOI phase respectively. Thus, 60 maps of Australia (five SOI phases × twelve 3-month rainfall periods) are created showing the probability of exceeding the climatological rainfall median for any 3-month period following each SOI phase. These maps are updated progressively monthly as more data are available and can be involved to calculate probabilities.

Table 1. Six large-scale climate indices used in this study.

Name	Abbreviation	Description	Ocean	Key reference
Indian Ocean Dipole	IOD	A sea surface temperature dipole between the western and eastern tropical Indian Ocean	Indian	(Saji *et al* 1999)
Southern Annular Mode	SAM	Pressure dipole between the Antarctic and Southern Hemisphere midlatitudes	Southern	(Thompson and Wallace 2000)
Nino3.4 sea surface temperature	NINO3.4	Mean SST over the Nino3.4 region (5°N–5°S, 120°–170°W)	Pacific	(Kaplan *et al* 1998)
Pacific Decadal Oscillation	PDO	A long-lived ENSO-like pattern of Pacific climate variability	Pacific	(Mantua and Hare 2002)
Southern Oscillation Index	SOI	An indication of the development and intensity of El Niño or La Niña events	Pacific	(Horel and Wallace 1981)
Tripole Index	TPI	A robust and stable representation of the Interdecadal Pacific Oscillation phenomenon	Pacific	(Henley *et al* 2015)

2.4. Random forest model

The machine learning algorithm used in this study is random forest (RF), a tree-based ensemble learning algorithm (Breiman 2001). An ensemble method is an algorithm that obtains averaged results from multiple learning models. In the case of RF, it first builds a forest of decision trees, in which each tree is independently created based on randomized subsets of input predictors generated from a bootstrap aggregating procedure (Heung *et al* 2014). All trees in the forest grow to maximum size without pruning and the average of the outputs from all trees is regarded as the final outcome (Cutler *et al* 2007). RF is capable of effectively reducing the variance in comparison with other tree-based models because of the application of the bootstrap aggregating procedure. RF can be used to build predictive models for classification purposes and can also estimate probability for each class.

RF can well process nonlinear and hierarchical relationships between the response and predictor variables and is not sensitive to the problem of multicollinearity among predictors (Breiman 2001, Li *et al* 2015). It can obtain useful information from multiple data sources and has been widely applied to address real-world problems in various fields including remote sensing (Belgiu and Drăguţ 2016), image processing (Alexander *et al* 2014) and ecology (Fox *et al* 2017). However, RF is rarely used for rainfall forecasts in practice. Therefore, we tested the ability of RF in forecasting rainfall using large-scale climate indices. The output of the RF model was used to compare with that of the SP model.

2.5. Model development

We aimed to build a machine learning-based forecasting model with a same feature as the SP model, forecasting rainfall probabilities. Thus, we first built RF classification models in which the response variable was a binary variable that reflected whether rainfall exceeded the climatological median for an upcoming season. We focused on four natural seasons instead of any three-month period in this study. While for predictor variables, we adopted six large-scale climate indices (table 1) over six months (6×6 variables) as input variables for the classification models. We obtained the probability for each class instead of classification results from the RF classification models, as the probability values could be directly compared with the output from the SP model. Thus, the RF model for each season can be expressed as the following form:

$$P_{I,j} = \begin{cases} RF_{I,spring}\left(\mathrm{IOD}_{Mar_I\sim Aug_I}, \mathrm{SAM}_{Mar_I\sim Aug_I}, \mathrm{NINO3.4}_{Mar_I\sim Aug_I}, \mathrm{PDO}_{Mar_I\sim Aug_I}, \mathrm{SOI}_{Mar_I\sim Aug_I}, \mathrm{TPI}_{Mar_I\sim Aug_I}\right) \\ RF_{I,summer}\left(\mathrm{IOD}_{Jun_I\sim Nov_I}, \mathrm{SAM}_{Jun_I\sim Nov_I}, \mathrm{NINO3.4}_{Jun_I\sim Nov_I}, \mathrm{PDO}_{Jun_I\sim Nov_I}, \mathrm{SOI}_{Jun_I\sim Nov_I}, \mathrm{TPI}_{Jun_I\sim Nov_I}\right) \\ RF_{I,autumn}\left(\mathrm{IOD}_{Sep_{I-1}\sim Feb_I}, \mathrm{SAM}_{Sep_{I-1}\sim Feb_I}, \mathrm{NINO3.4}_{Sep_{I-1}\sim Feb_I}, \mathrm{PDO}_{Sep_{I-1}\sim Feb_I}, \mathrm{SOI}_{Sep_{I-1}\sim Feb_I}, \mathrm{TPI}_{Sep_{I-1}\sim Feb_I}\right) \\ RF_{I,winter}\left(\mathrm{IOD}_{Dec_{I-1}\sim May_I}, \mathrm{SAM}_{Dec_{I-1}\sim May_I}, \mathrm{NINO3.4}_{Dec_{I-1}\sim May_I}, \mathrm{PDO}_{Dec_{I-1}\sim May_I}, \mathrm{SOI}_{Dec_{I-1}\sim May_I}, \mathrm{TPI}_{Dec_{I-1}\sim May_I}\right) \end{cases} \quad (4)$$

where $\boldsymbol{P}_{I,j}$ was forecasted probability of rainfall exceeding the climatological median, which was comparable with the output of the SP model.

The SP model followed a concept that all preceding available data are used in forecasting rainfall probability of the upcoming three months. That is, when forecasting rainfall probability of the season j in year I, all available rainfall conditions of season j with a same SOI phase from 1889 to year I-1, were used to calculate the probability. The RF model also followed this approach in our study. All available rainfall conditions of season j and precedent indices from 1889 to year I-1 were used to build the RF model and then generated the forecast for the season j in year I. For example, when forecasting the probability of rainfall exceeding the median for 2011 spring (Sep-Nov), we first built a classification model based on training data including spring rainfall and six large-scale climate indices over six months (Mar-Aug) (36 predictors in total) from 1889 to 2010. Whether or not spring rainfall exceeded the median was adopted as the target variable, and 36 climate variables were adopted as predictors. Then, this classification model was used to forecast whether 2011 spring rainfall exceeded the median. We obtained the probability of 2011 spring rainfall exceeded the median as the final output. Therefore, the output of the RF model can be compared with the SP model. We successively built models and generated forecasts for four seasons of recent eight years (I = 2011, 2002, …, 2018) for each grid. The 'caret' package sourced in the R software was adopted to build RF models. Default values were used for the parameters of the RF model, as RF was not sensitive to parameter settings and default parameters can usually provide satisfactory results (Duro *et al* 2012, Immitzer *et al* 2012).

2.6. Model performance evaluation

We aimed to forecast the probability of rainfall exceeding the climatological median. If the forecasted value is around 50%, we still do not have confidence to say whether the upcoming three months tend to be wetter or drier. In other words, forecasts of ~50% probability cannot provide any instructive information for decision makers to take actions. To compare the accuracy of the two models, we assumed forecasts with $\boldsymbol{P}_{I,j} > 60\%$ as the case that forecasted rainfall was expected to exceed the climatological median. Alternatively, forecasts with $\boldsymbol{P}_{I,j} < 40\%$ meant the case that forecasted rainfall was expected to be lower than the median. We defined forecasts with 40%–60% probabilities as indistinct forecasts, while forecasts with probabilities more than 60% or less than 40% were distinct forecasts. We compared the percentage values of distinct forecasts for each grid and each season from the RF model and the SP model respectively.

We also used a commonly used metrics, accuracy (AC), to compare the performance of the RF model and the SP model.

$$\mathrm{AC} = \frac{\text{true positive} + \text{true negative}}{\text{total number of forecasts}} \tag{5}$$

where true positive means the number of records with observed rainfall exceeding the median and forecasted $\boldsymbol{P}_{I,j} > 60\%$; true negative means the number of records with observed rainfall below the median and forecasted $\boldsymbol{P}_{I,j} < 40\%$; total number of forecasts is 8 in our study. In general, forecasts become increasingly accurate as AC approach 1. We compared the AC values for each grid and each season between the RF model and the SP model.

3. Results

3.1. Percentage of distinct forecasts

Distributions of all forecasted rainfall probabilities (n = 2780 grids × 8 years) for four seasons are presented in figure 3. Distributions for each climate zone separately are shown in figure S1 (available online at stacks.iop.org/ERL/15/084051/mmedia). Forecasted probabilities by the SP model were distributed mainly around 50%, particularly for autumn (figure 3(c)). Conversely for the RF model, forecasted probabilities were less centralized with more forecasts located on either side. Thus, the RF model had more distinct forecasts compared to the SP model. The results of percentages of distinct forecasts (figure 3) illustrated that the RF model increased the percentage value to 64.9% for spring, to 71.5% for summer, to 65.8% for autumn, and to 63.9% for winter, which are 1.4 ~ 3.2 times as large as the SP model. Therefore, the RF model could provide more instructive forecasts than the SP model.

3.2. Forecasting accuracy

In general, the RF model performed better than the SP model for all the four seasons in terms of the AC values (figure 4). The RF achieved an AC of >0.3 in most grids for all seasons (figures 4(a1)–(d1)). The SP model only had acceptable performance for spring, but had poor performance for the other three seasons in most grids, which might be due to more indistinct forecasts for the three seasons (figure 3(c)). We also calculated forecasting accuracy for each class (Fig. S3 and S4) and the results also illustrated better performance of the RF model in most grids for all seasons, compared to the SP model.

3.3. Relative contributions of climate drivers to rainfall forecasts

We also obtained the list of importance values of input variables from the RF model to give a preliminary overview of relative contributions of climate drivers to rainfall forecasts. For each climate zone, relative contributions of climate drivers to rainfall forecasts were aggregated values based on the outputs of RF models at all grids located in that zone.

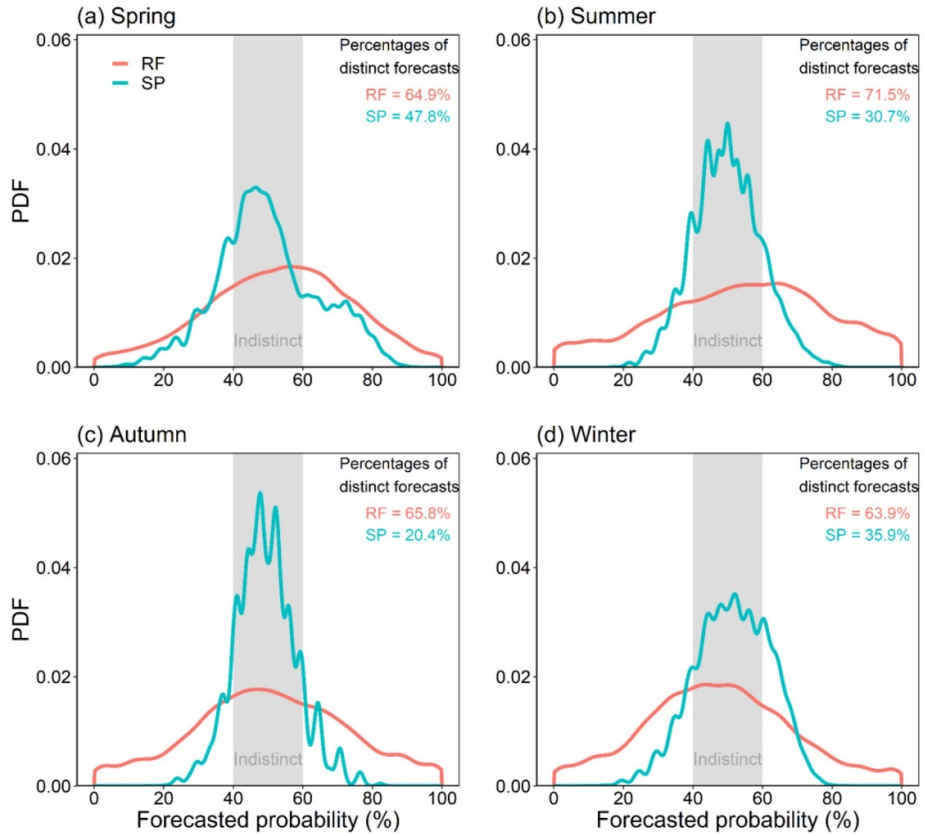

Figure 3. Distributions of all forecasted rainfall probabilities (n = 2780 grids × 8 years) for each season based on probability density function (PDF). Shaded areas indicate indistinct forecasts. Percentages of distinct forecasts for the RF model and the SP model are given in each subplot.

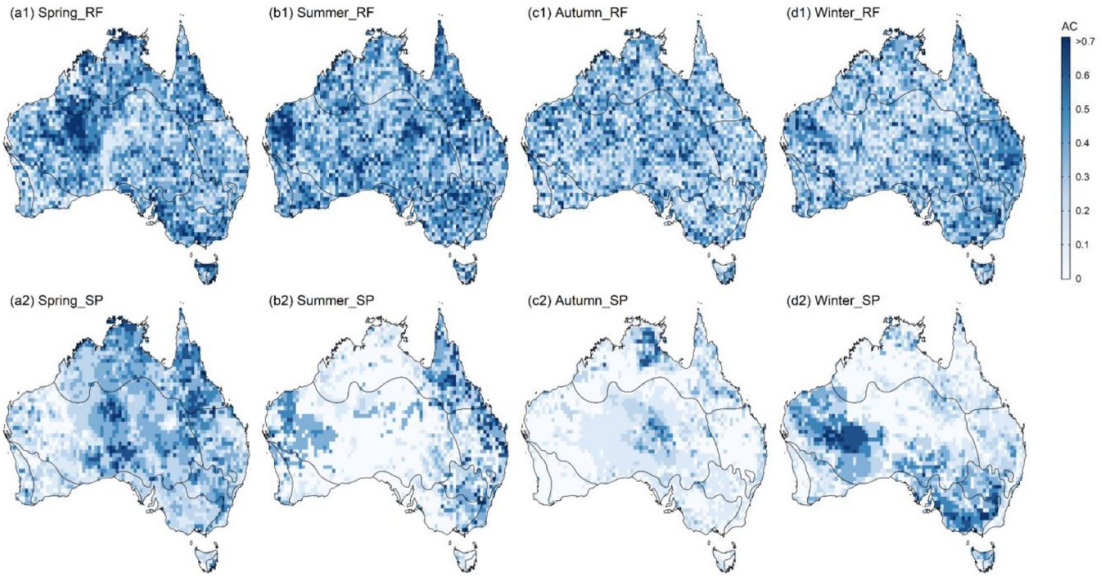

Figure 4. Accuracy values (AC) for validation period (2011–2018) for four seasons across 2780 grids (0.5° × 0.5°) in Australia based on the RF model and the SP model.

As shown in figure 5, NINO3.4 and TPI, two drivers from the Pacific, had relatively large contributions in most seasons and eastern zones. On the other hand, the contributions from the Indian Ocean (IOD) and the Southern Ocean (SAM) were not neglectable, especially during autumns and winters, ranging from 12% to 18%. Thus, preceding oceanic activities from three surrounding oceans can provide prognostic and useful information for seasonal rainfall forecast in Australia.

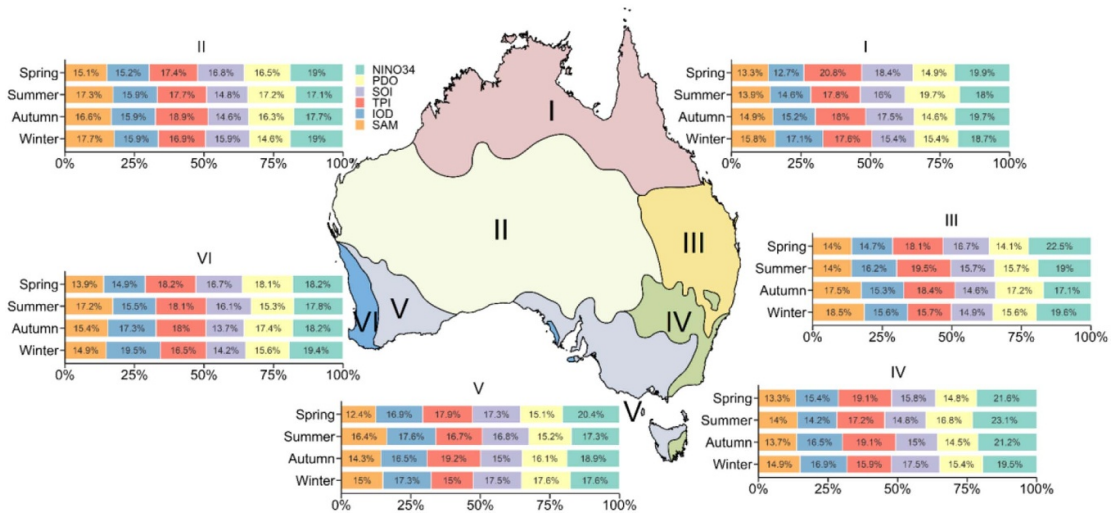

Figure 5. Relative contributions of input predictor variables to seasonal rainfall probabilities in each climate zone as determined from the RF model. The stacked bar plot for a given zone was made using aggregate contribution values from all grids located in that zone.

4. Discussion

4.1. Probabilistic rainfall forecasting

Seasonal rainfall forecasting is one of the available tools to help deal with high rainfall variability and is currently used by nearly half of agricultural producers in decision-making in Australia (Meinke *et al* 1996, Cobon and Toombs 2013). We developed a machine learning-based seasonal rainfall forecasting model, the RF model, based on large-scale climate drivers. The overall performance of the RF model was better compared to the SP model, for rainfall forecasting of four seasons across Australia in terms of both distinct forecasts and forecasting accuracy. This may be because our method considered the impacts of multiple climate drivers rather than the SOI alone. Australian rainfall is regulated by oceanic activities from three surrounding oceans, the Pacific, Indian and Southern Oceans, depending on regions and seasons (Min *et al* 2013). Thus, using only one driver from the Pacific may not provide sufficient information for seasonal rainfall forecasts. Moreover, the ability of disentangling non-linear relationships between the response and predictor variables may also assist the RF model achieve satisfactory performance. For example, Hossain *et al* (2019) demonstrated that non-linear artificial neural network models outperformed multiple linear regression models in forecasting Western Australian spring rainfalls in regards to statistical errors and Pearson correlation. The proposed seasonal rainfall forecasting method is efficient based on readily available data. It can be easily extended to other regions to initiate seasonal rainfall outlooks to enhance the present capabilities of water resource management.

4.2. Relative contributions of climate drivers to seasonal rainfall

Our RF model relied on six large-scale climate indices and illustrated that each index can contribute to 10%–20% of seasonal rainfall forecasts in most climate zones and seasons (figure 5). This is consistent with previous studies which demonstrated that rainfall conditions throughout Australia were generally the result of the synchronization of multiple climate drivers (Cleverly *et al* 2016). Additionally, each climate driver usually accounted for less than 20% of rainfall variability (Risbey *et al* 2009, Gallant *et al* 2012). On the other hand, climate indices from the Pacific, NINO3.4, TPI, PDO and SOI, particularly the first two, show relatively large contributions to rainfall forecasts despite that there are some spatial or seasonal differences. The principal influence on Australian seasonal rainfall is ENSO from the Pacific Ocean and this was well established (Allan 1988, Nicholls *et al* 1997, Wang and Hendon 2007). Nevertheless, the impacts of the Indian Ocean (IOD) and the Southern Ocean (SAM) cannot be ignored in a forecasting model, especially for autumn or winter rainfall (figure 5).

4.3. Future work to improve statistical seasonal rainfall forecasting models

Physics-based dynamical models are normally considered as the mainstream approach by scientific community and are currently used by Australian Bureau of Meteorology to provide official seasonal climate forecasts. However, despite substantial technological advances and research efforts, dynamical models still have similar performance on seasonal

rainfall forecasts in comparison to simple statistical models (Abbot and Marohasy 2014, Cohen *et al* 2019). Moreover, Mekanik *et al* (2016) and Abbot and Marohasy (2014) both demonstrated that machine learning-based statistical models were comparable to the former dynamical model, the Predictive Ocean Atmosphere Model for Australia (POAMA) used by BOM. A review of 27 dynamical models under the Coupled Model Intercomparison Project Phase 5 demonstrated that different models can produce widely divergent rainfall forecasts in Australia (Irving *et al* 2012). Thus, owing to low forecasting skill and/or complexity of dynamical models, statistical models remain the most commonly used methods for seasonal rainfall forecasting in terms of agricultural planning (Meinke *et al* 2007, He *et al* 2014).

We advocate that statistical models should continue to be improved and we believe it will play a key role in seasonal rainfall forecasting. We summarize three aspects that statistical forecasting model might be further improved as follows:

(a) Search for better indices. Large-scale climate indices can provide prognostic information for statistical models. Each index is likely to have its own affected zones and seasons. For example, SOI has been shown to mainly influence eastern Australia and have strong correlations with spring and winter rainfall but weak correlations with summer and autumn rainfall (Cobon and Toombs 2013). IOD was negatively correlated with rainfall from June to October in Western Australia, Victoria, South Australia, and southern New South Wales (Steptoe *et al* 2018). As only six climate indices were included in the proposed RF model, some impacts from certain oceanic activities may be missed in consideration. That might result in relatively poor performance in the RF model for autumn rainfall forecast (figure 4). Moreover, impacts of many oceanic activities on terrestrial rainfall remain undiscovered. It is also possible that there are some influential but undiscovered oceanic activities from surrounding oceans affecting Australian rainfall. Therefore, more efforts should be made to explore potential oceanic activities and their impacts on terrestrial rainfall to improve the performance of statistical models.

(b) Forecast extremes rather than above or below median. Currently, the two official forecasting models, the ACCESS–S model and the SP model, both provide rainfall outlooks as the probability of getting above median rainfall for the seasons ahead. However, the probability of above median can provide limited information in guiding agricultural activities, as decision-makers focus more on extreme conditions (*e.g.* drought or flood) to develop targeted strategies (He *et al* 2014). One example is the 2018 drought in eastern Australia, during which rainfall shortage resulted in a 53% reduction in winter crop production (www.agriculture.gov.au/abares). Categorizing rainfall into more classes, *e.g.* low (0%–33%), median (34%–66%), and high (67%–100%), and forecasting a probability value for each class may be a feasible choice, which may provide more useful information for application sides.

(c) Focus on non-stationary predictor-predictand relationships. Statistical models usually assume stationary relationships between the response and predictor variables (Schepen *et al* 2012). However, the relationships between rainfall and large-scale climate drivers are normally non-stationary and changing with time. For example, the impacts of IOD on Australia's rainfall has enhanced in recent decades and the major driver of several main droughts in 20th century in Australia was attributed to increased positive IOD events rather than ENSO related phenomena (Cai *et al* 2012, Yuan and Yamagata 2015, Nguyen-Huy *et al* 2018). Meanwhile, global warming may also contribute to the alteration of the effects of different climate drivers (Cai *et al* 2015). Some deep learning algorithms (*e.g.* the long short-term memory algorithm) that can dynamically explore predictor-predictand relationships could be introduced to develop statistical forecasting models to achieve better accuracy.

5. Conclusion

Our study developed a seasonal rainfall forecasting model for Australia using machine learning technique and multiple precedent large-scale climate indices. The officially used SOI phase model was adopted as the benchmark. Results indicated that the RF model could provide better forecasts in nearly all climate subregions and four seasons in terms of both the percentage of distinct forecasts and forecasting accuracy, compared to the SP model. However, the proposed model had relatively poor performance for autumn rainfall forecasts, which highlights more efforts to explore oceanic activities occurring in surrounding oceans.

Australia is a major food producer and exporter in the world. Reliable seasonal rainfall forecasting can effectively help stakeholders reduce rainfall shortage-induced yield losses, which is of great importance for both national food supply and global food security. We believe the seasonal rainfall forecasting model developed in our study can provide valuable information for both Australian farmers and policy makers. Moreover, the proposed model could also easily be implemented in other regions as input data are readily available.

Acknowledgments

This study was jointly supported by the Natural Science Foundation of China (No. 41961124006) and the International Partnership Program of the Chinese Academy of Sciences (161461KYSB20170013). The first author acknowledges the China Scholarship Council (CSC) for the financial support for his Ph.D. study. Facilities for conducting this study were provided by the New South Wales Department of Primary Industries. Thanks to Dr Jian Liu of Northwest A&F University for re-calculating SAM. Bernie Dominiak provided comments on an earlier version of this manuscript.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Abbot J and Marohasy J 2014 Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks *Atmos. Res.* **138** 166–78

Aguasca-Colomo R, Castellanos-Nieves D and Méndez M 2019 Comparative analysis of rainfall prediction models using machine learning in islands with complex orography: Tenerife Island *Appl. Sci.* **9** 4931

Alexander D C, Zikic D, Zhang J, Zhang H and Criminisi A 2014 Image quality Transfer via random forest regression: applications in diffusion MRI *Medical Image Computing and Computer-Assisted Intervention. MICCAI 2014* (Lecture Notes in Computer Science vol 8675) ed P Golland, N Hata, C Barillot, J Hornegger and R Howe (Cham: Springer)

Allan R J 1988 El Niño southern oscillation influences in the Australasian region *Prog. Phys. Geogr.* **12** 313–48

Belgiu M and Drăguţ L 2016 Random forest in remote sensing: a review of applications and future directions *ISPRS J. Photogramm. Remote Sens.* **114** 24–31

Boer M M *et al* 2016 Future changes in climatic water balance determine potential for transformational shifts in Australian fire regimes *Environ. Res. Lett.* **11** 065002

Breiman L 2001 Random forest *Mach. Learn.* **45** 5–32

Cai W *et al* 2015 ENSO and greenhouse warming *Nat. Clim. Chang.* **5** 849

Cai W, Van Rensch P, Cowan T and Hendon H H 2012 An asymmetry in the IOD and ENSO teleconnection pathway and its impact on Australian climate *J. Clim.* **25** 6318–29

Cleverly J, Eamus D, Luo Q, Restrepo Coupe N, Kljun N, Ma X, Ewenz C, Li L, Yu Q and Huete A 2016 The importance of interacting climate modes on Australia's contribution to global carbon cycle extremes *Sci. Rep.* **6** 23113

Cobon D H and Toombs N R 2013 Forecasting rainfall based on the southern oscillation index phases at longer lead-times in Australia *Rangeland J.* **35** 373–83

Cohen J, Coumou D, Hwang J, Mackey L, Orenstein P, Totz S and Tziperman E 2019 S2S reboot: an argument for greater inclusion of machine learning in subseasonal to seasonal forecasts *Wiley Interdisc. Rev.: Clim. Change* **10** e00567

Cutler D R, Edwards T C, Beard K H, Cutler A and Hess K T 2007 Random forests for classification in ecology *Ecology* **88** 2783–92

Duro D C, Franklin S E and Dubé M G 2012 A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery *Remote Sens. Environ.* **118** 259–72

Fawcett R and Stone R 2010 A comparison of two seasonal rainfall forecasting systems for Australia *Aust. Meteorol. Oceanogr. J.* **60** 15–24

Fox E W, Hill R A, Leibowitz S G, Olsen A R, Thornbrugh D J and Weber M H 2017 Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology *Environ. Monit. Assess.* **189** 316

Gallant A, Kiem A, Verdon-Kidd D, Stone R and Karoly D 2012 Understanding hydroclimate processes in the Murray-Darling basin for natural resources management *Hydrol. Earth Syst. Sc.* **16** 2049–68

Gunasekera D, Kim Y, Tulloh C and Ford M 2007 Climate change—impacts on Australian agriculture *Aust. Commod.: Forecasts Iss.* **14** 657–76

Hartmann H, Becker S and King L 2008 Predicting summer rainfall in the Yangtze River basin with neural networks *Int. J. Climatol.* A **28** 925–36

He X, Guan H, Zhang X and Simmons C T 2014 A wavelet-based multiple linear regression model for forecasting monthly rainfall *Int. J. Climatol.* **34** 1898–912

Henley B J, Gergis J, Karoly D J, Power S, Kennedy J and Folland C K 2015 A tripole index for the interdecadal Pacific oscillation *Clim. Dynam.* **45** 3077–90

Heung B, Bulmer C E and Schmidt M G 2014 Predictive soil parent material mapping at a regional-scale: a random forest approach *Geoderma* **214** 141–54

Horel J D and Wallace J M 1981 Planetary-scale atmospheric phenomena associated with the southern oscillation *Mon. Weather Rev.* **109** 813–29

Hossain I, Rasel H M, Imteaz M A and Mekanik F 2020 Long-term seasonal rainfall forecasting using linear and non-linear modelling approaches: a case study for Western Australia *Meteorol. Atmos. Phys.* **132** 131–41

Immitzer M, Atzberger C and Koukal T 2012 Tree species classification with random forest using very high spatial resolution 8-band WorldView-2 satellite data *Remote Sens.* **4** 2661–93

Irving D B, Whetton P and Moise A F 2012 Climate projections for Australia: a first glance at CMIP5 *Aust. Meteorol. Oceanogr. J.* **62** 211–25

Kaplan A, Cane M A, Kushnir Y, Clement A C, Blumenthal M B and Rajagopalan B 1998 Analyses of global sea surface temperature 1856–1991 *J. Geophys. Res. Oceans* **103** 18567–89

Kashid S S and Maity R 2012 Prediction of monthly rainfall on homogeneous monsoon regions of India based on large scale circulation patterns using genetic programming *J. Hydrol.* **454** 26–41

Li X, Zhai T, Jiao Y and Wang G 2015 Using Bayesian hierarchical models and random forest algorithm for habitat use studies: a case of nest site selection of the crested ibis at regional scales *PeerJ PrePrints* **3** e871v1

Mantua N J and Hare S R 2002 The Pacific decadal oscillation *J. Oceanogr.* **58** 35–44

Meinke H, Sivakumar M, Motha R P and Nelson R 2007 Preface: climate predictions for better agricultural risk management *Aust. J. Agric. Res.* **58** 935–8

Meinke H, Stone R C and Hammer G L 1996 SOI phases and climatic risk to peanut production: a case study for northern Australia *Int. J. Climatol.* A **16** 783–9

Mekanik F, Imteaz M and Talei A 2016 Seasonal rainfall forecasting by adaptive network-based fuzzy inference system (ANFIS) using large scale climate signals *Clim. Dynam.* **46** 3097–111

Min S K, Cai W and Whetton P 2013 Influence of climate variability on seasonal extremes over Australia *J. Geophys. Res.: Atmos.* **118** 643–54

Nguyen-Huy T, Deo R C, Mushtaq S, An-Vo D-A and Khan S 2018 Modeling the joint influence of multiple synoptic-scale, climate mode indices on Australian wheat yield using a vine copula-based approach *Eur. J. Agron.* **98** 65–81

Nicholls N, Drosdowsky W and Lavery B 1997 Australian rainfall variability and change *Weather* **52** 66–72

Qureshi M E, Hanjra M A and Ward J 2013 Impact of water scarcity in Australia on global food security in an era of climate change *Food Policy* **38** 136–45

Risbey J S, Pook M J, Mcintosh P C, Wheeler M C and Hendon H H 2009 On the remote drivers of rainfall variability in Australia *Mon. Weather Rev.* **137** 3233–53

Saji N, Goswami B, Vinayachandran P and Yamagata T 1999 A dipole mode in the tropical Indian Ocean *Nature* **401** 360

Schepen A, Wang Q and Robertson D E 2012 Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall *J. Geophys. Res. Atmos.* **117**

Scher S and Messori G 2019 Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground *Geosci. Model Dev.* **12** 2797–809

Shalev-Shwartz S and Ben-David S 2014 *Understanding Machine Learning: From Theory to Algorithms* (Cambridge: Cambridge University Press)

Steptoe H, Jones S and Fox H 2018 Correlations between extreme atmospheric hazards and global teleconnections: implications for multihazard resilience *Rev. Geophys.* **56** 50–78

Stone R C, Hammer G L and Marcussen T 1996 Prediction of global rainfall probabilities using phases of the southern oscillation index *Nature* **384** 252

Thompson D W and Wallace J M 2000 Annular modes in the extratropical circulation. Part I: month-to-month variability *J. Clim.* **13** 1000–16

Turney C S, Scourse J, Rodbell D and Caseldine C 2007 Quaternary climatic, environmental and archaeological change in Australasia *J. Quat. Sci.* **22** 421–2

Wang B, Zheng L, Liu D L, Ji F, Clark A and Yu Q 2018 Using multi-model ensembles of CMIP5 global climate models to reproduce observed monthly rainfall and temperature with machine learning methods in Australia *Int. J. Climatol.* **38** 4891–902

Wang G and Hendon H H 2007 Sensitivity of Australian rainfall to inter–El Niño variations *J. Clim.* **20** 4211–26

Williamson G J, Prior L D, Jolly W M, Cochrane M A, Murphy B P and Bowman D M J S 2016 Measurement of inter-and intra-annual variability of landscape fire activity at a continental scale: the Australian case *Environ. Res. Lett.* **11** 035003

Yuan C and Yamagata T 2015 Impacts of IOD, ENSO and ENSO Modoki on the Australian winter wheat yields in recent decades *Sci. Rep.* **5** 17252