

This file is part of the following work:

Schepen, Andrew David (2019) *Harnessing seasonal GCM forecasts for crop yield forecasting through multivariate forecast post-processing methods*. PhD Thesis, James Cook University.

Access to this file is available from:

<https://doi.org/10.25903/5df83e7211d59>

Copyright © 2019 Andrew David Schepen.

The author has certified to JCU that they have made a reasonable effort to gain permission and acknowledge the owners of any third party copyright material included in this document. If you believe that this is not the case, please email

researchonline@jcu.edu.au

**Harnessing seasonal GCM forecasts for crop yield forecasting
through multivariate forecast post-processing methods**

Thesis submitted by

Andrew David SCHEPEN BSc

in June 2019

for the degree of Doctor of Philosophy

in the College of Science and Engineering

James Cook University

Copyright Statement

Every reasonable effort has been made to gain permission and acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

Acknowledgements

I graciously acknowledge my advisors, Associate Professor Yvette Everingham and Professor QJ Wang, for their guidance, encouragement and support throughout my research project. QJ played a pivotal role in setting me on the PhD path and instilling the belief that I could succeed at such a monumental task. I have immensely enjoyed working with Yvette who helped me focus and make sense of my PhD vision, and guided me to the end. Thanks also to Madoc Sheehan for supporting my candidature as an advisory panel member.

I am grateful to my CSIRO leaders, David Robertson, Francis Chiew and Warwick McDonald for ensuring I had the time and support to complete my studies whilst continuing to work at CSIRO.

Finally, a special acknowledgement to my family; my wife Emily, and my three wonderful children, Elliot, Zoe and Oscar, for your unwavering understanding and support, and for making sure I maintained a healthy work/life balance.

Statement of the Contribution of Others

Associate Professor Yvette Everingham (James Cook University) and Professor Q.J. Wang (University of Melbourne) provided advice on my research objectives and candidature supervision. Associate Professor Yvette Everingham provided expert oversight of climate forecasting, statistical analysis and crop modelling aspects. Professor Q.J. Wang provided expert oversight of Bayesian statistical modelling, hydro-climate modelling, ensemble forecasting and forecast verification aspects.

The European Centre for Medium-range Weather forecasts provided access to an online database from which I retrieved seasonal reforecasts that were used in chapters 2–4. Similarly, the Queensland Government provided access to an online database of SILO weather observations from which I retrieved the observed data used in chapters 2–4.

The Commonwealth Scientific and Industrial Research Organisation (CSIRO) supplied high-performance computing infrastructure and cloud storage for data. High-performance computing was necessary to complete the continental scale climate forecasting study reported in Chapter 2 and the ensemble crop-model forecasting study reported in Chapter 4.

The Agricultural Production Systems Simulator (APSIM) software was provided by the APSIM initiative. Jody Biggs from CSIRO Agriculture and Food supplied an example APSIM-Sugar simulation file for the Tully region that was adapted for use in Chapter 4. Justin Sexton from James Cook University assisted with testing and tweaking the APSIM-sugar model to ensure the outputs were in the expected range.

James Cook University contributed approximately \$5000 of funding that was used to attend a 2-day training course on APSIM and a 5-day course on C/C++ programming (travel included).

Figure 1 was supplied by the Queensland Alliance for Agriculture and Food Innovation and is used with permission from Andries Potgieter.

Please note: A professional editing service was not used in the preparation of this thesis.

Abstract

Seasonal climate forecasts may be coupled with crop models to provide quantitative forecasts of crop yield, assess sensitivity to farm management decisions and manage risk associated with seasonal climate variability. Today, seasonal climate forecasts are produced by computationally-expensive, physically-based global climate models, which capture large-scale climate patterns well. However, their coarse spatial resolution (typically >50km) means they do not reliably depict daily weather at sub-grid locations, limiting their direct use in crop models. Consequently, operational crop forecasting systems in Australia typically use alternative meteorological forcings such as historical climate analogues based on El Niño - Southern Oscillation phases, which may be less skilful than global climate model forecasts.

An emerging tactic for coupling global climate model forecasts and crop models is to apply quantile-mapping (otherwise known as cumulative distribution function matching) to adjust forecast ensemble members according to the historical distribution of observations. However, quantile mapping assumes the global climate model forecasts are highly skilful and well-behaved (which they are often not). The overly simplistic formulation of quantile-mapping propagates an assortment of model errors. Additionally, quantile-mapping cannot be used for downscaling to multiple sub-grid locations owing to its deterministic nature. Accordingly, an increasing number of studies are reporting negative results arising from coupling global climate model forecasts and crop models using quantile mapping. Hence, the overarching objective of this thesis is to develop more robust, spatially and temporally relevant post-processing methods to harness global climate model forecasts for use in crop models. To this end, I develop a new multivariate forecast post-processing workflow that combines Bayesian parametric methods and non-parametric methods to calibrate and downscale global climate model forecasts for use in crop models.

Forecast calibration means to

- (1) minimise systematic error such as forecast bias,

- (2) ensure forecast uncertainty is reliably conveyed by ensemble spread, and
- (3) ensure forecasts are at least as skilful as climatology.

Downscaling means, depending on the context, either:

- (1) producing a revised forecast with the correct local weather variability at a spatial scale smaller than the GCM grid
- (2) producing a local forecast based on large-scale climate drivers (e.g. sea surface temperature patterns) (this approach is also referred to as bridging), or
- (3) spatial or temporal disaggregation of a forecast.

Crop forecasting models require physically-coherent inputs of rainfall, temperature and solar radiation. Previous research has established the suitability of the Bayesian joint probability modelling approach for calibrating monthly and three-monthly rainfall forecasts from global climate models. The Bayesian joint probability modelling approach has not previously been applied to post-process temperature or solar radiation forecasts or to post-process multivariate forecasts. However, it is formed on the general assumption that the joint distribution of two or more variables can be modelled as a multivariate normal distribution in transformed space. It can theoretically be extended for multivariate forecast post-processing with a relevant transformation for each variable. Thus the first objective of this thesis is to develop and evaluate several strategies for calibrating multivariate global climate model forecasts using the Bayesian joint probability modelling approach. Three strategies are compared: (1) simultaneous calibration of multiple climate variables in a single statistical model, which explicitly models inter-variable dependence via the covariance matrix; (2) univariate calibration coupled with an empirical ensemble reordering method (the Schaake Shuffle) that injects inter-variable dependence from historical data; and (3) quantile-mapping, which borrows inter-variable dependence from the raw forecasts. Applied to Australian seasonal (three-month) forecasts from the European Centre for Medium-range Weather Forecasts System4 model,

univariate calibration paired with the Schaake Shuffle performs best in terms of univariate and multivariate forecast verification metrics. Direct multivariate calibration is the second-best method, with its far superior performance in in-sample testing vanishing in cross-validation, likely because of insufficient data to reliably infer the sizeable covariance matrix. Bayesian joint probability post-processing is confirmed to outperform quantile-mapping. Hence the Bayesian joint probability modelling approach and the Schaake Shuffle should, therefore, be preferred to quantile-mapping as a basis for calibrating GCM forecasts for crop forecasting applications.

Global climate model forecast skill is best captured by post-processing on seasonal time scales.

However, crop models require daily forecast sequences. Also, it is observed that some operational crop forecasting systems run separate crop models for multiple locations within a region and then aggregate the results into a regional forecast. Therefore, spatial forecasts are also needed.

Accordingly, the second objective of this thesis is to develop and evaluate downscaling and disaggregation methods for post-processing global climate model forecasts to higher spatial and temporal resolutions. To this end, I develop an empirical multivariate downscaling method that imparts observed spatial, temporal and inter-variable relationships into disaggregated forecasts whilst completely preserving the joint distribution of forecasts post-processed at coarser spatial and/or temporal scales. Specifically, a Euclidean distance metric is devised to identify a nearest-neighbour in historical observations for each forecast ensemble member. The method of fragments is subsequently applied to simultaneously disaggregate the forecast spatial and temporally. The new method is demonstrated to perform well for downscaling skilful forecasts of rainfall, temperature and solar radiation for six locations in northeast Australia. The climatological distributions of the downscaled forecasts mirror observations and the observed frequency of wet days is also reproduced in forecasts. The new downscaling method is a step towards full integration of calibrated seasonal climate forecasts into crop models and has a significant advantage over quantile-mapping in that it can be applied for multiple sub-grid locations.

The final objective of this thesis is to feed global climate model forecasts, post-processed using the new methods, to a crop decision support system to demonstrate an end-to-end solution for linking global climate model forecasts with a crop model to produce yield forecasts. The first crop forecasting application of the new methods is for sugarcane yield forecasting in Tully. The region is selected because it is a non-irrigated region, and it is thus suitable for assessing the value of climate forecasts. Two sets of post-processed forecasts are produced for the Tully Mill weather station in North-east Queensland. The first set is obtained by applying the Bayesian joint probability modelling approach to calibrate monthly rainfall, temperature and solar radiation forecasts for the grid cell containing Tully. The second set is obtained by using global climate model forecasts of the Niño3.4 climate index (commonly associated with the El Niño Southern Oscillation), also using the Bayesian joint probability modelling approach, to produce local forecasts of monthly rainfall, temperature and solar radiation. In both cases, the monthly forecasts are subjected to the Schaake Shuffle and subsequently downscaled to daily sequences using identical methods. The calibration and bridging forecasts are used to drive a sugarcane crop model to generate long-lead forecasts of biomass in north-eastern Australia from 1982-2016. A rigorous probabilistic assessment of forecast attributes suggests that the calibration forecasts provide the most skilful forecasts overall although the bridging forecasts give more skilful yield forecasts at certain times. The biomass forecasts are unbiased and reliable for short to long lead times, suggesting that the new downscaling methods are effective.

My end-to-end solution for linking global climate model forecasts and crop models enables quantitative modelling and risk management at the farm level. It has the potential to improve farm productivity and profitability through better decisions. Future research should investigate the value of the post-processing methods for a wide range of crops.

Table of Contents

1.	Thesis introduction	18
1.1.	Preamble	18
1.2.	Weather, Climate and Agriculture	19
1.3.	Agro-climatic decision support tools.....	20
1.4.	Ensemble seasonal global climate model (GCM) forecasts.....	24
1.5.	Current methods for pairing GCM forecasts and crop models.....	26
1.6.	Advancing GCM forecast post-processing for agriculture.....	28
1.7.	Thesis objectives	31
1.8.	Thesis structure and publications	32
2.	Calibrating multivariate seasonal forecasts from GCMs.....	36
2.1.	Preamble	36
2.2.	Introduction.....	37
2.3.	Methods	42
2.3.1.	Multivariate calibration strategies	42
2.3.2.	Marginal transformation	43
2.3.3.	Multivariate BJP calibration (MBJP)	45
2.3.4.	Univariate BJP calibration plus Schaake Shuffle (UBJP+SS)	47
2.3.5.	Transformed Quantile-Mapping (TQM).....	48
2.4.	Application and verification.....	50
2.4.1.	Study data	50
2.4.2.	Univariate and multivariate probabilistic forecast verification	51

2.5.	Results and discussion.....	54
2.5.1.	Bias, reliability and skill of individual variables.....	54
2.5.2.	Overall performance of multivariate forecasts.....	57
2.5.3.	Diagnosing factors that affect performance.....	64
2.5.4.	Extension opportunities.....	71
2.5.5.	Conclusions.....	73
3.	Spatial and temporal disaggregation of climate forecasts.....	75
3.1.	Preamble.....	75
3.2.	Introduction.....	76
3.3.	Methods.....	81
3.3.1.	Forecast Calibration – Multivariate Downscaling.....	81
3.3.2.	BJP forecast calibration.....	83
3.3.3.	Schaake Shuffle ensemble reordering.....	84
3.3.4.	Nearest neighbour downscaling.....	86
3.4.	Application and verification.....	90
3.4.1.	Study area and data.....	90
3.4.2.	Forecast verification.....	92
3.4.2.1.	Cross-validation.....	92
3.4.2.2.	Seasonal skill and reliability evaluation.....	92
3.4.2.3.	Validation of downscaled daily sequences.....	93
3.5.	Results.....	94
3.5.1.	Skill and reliability of BJP-calibrated forecasts.....	94

3.5.2.	Distributions of daily values.....	97
3.5.3.	Temporal correlations	99
3.5.4.	Spatial correlations.....	103
3.5.5.	Inter-variable correlations	104
3.6.	Discussion	105
3.7.	Conclusion	107
4.	Ensemble sugarcane yield forecasting.....	110
4.1.	Preamble	110
4.2.	Introduction.....	111
4.3.	Models and data	115
4.3.1.	Case study location and observed weather data	115
4.3.2.	The APSIM-sugar crop model.....	116
4.3.3.	Climate model forecasts	117
4.3.3.1.	Raw GCM forecasts.....	117
4.3.3.2.	Post-processing (calibration and downscaling).....	118
4.4.	Application and verification.....	119
4.4.1.	Experimental configuration	119
4.4.2.	Reference forecasts and cross-validation.....	119
4.4.3.	Verification metrics	119
4.4.3.1.	Bias, and reliability in ensemble spread	120
4.4.3.2.	Probabilistic forecast skill.....	121
4.4.3.3.	Relative shift and dispersion	121

4.5.	Results	122
4.5.1.	Climate forecast verification	122
4.5.2.	Detailed verification for SON biomass forecasts.....	125
4.5.3.	Overall results	127
4.6.	Discussion	131
4.7.	Conclusion	133
5.	Thesis conclusion	135
5.1.	Preamble	135
5.2.	Objective 1 summary	135
5.3.	Objective 2 summary	139
5.4.	Objective 3 summary	142
5.5.	Highlights and implications.....	146
5.6.	Limitations and future directions.....	148
6.	References.....	151
	Appendix A: Ancillary paper 1.....	159
	Appendix B: Ancillary paper 2	160
	Appendix C: Revisions.....	161

List of Figures

Figure 1.1: Wheat crop outlook for Australian wheat-growing regions, issued by the Queensland Alliance for Agriculture and Food Innovation (QAAFI) in July 2018. The colour scale represents the chance of exceeding the long-term median yield as simulated with the Oz-Wheat model. Figure supplied by QAAFI.	23
Figure 1.2 Graphical overview of my original contributions and progression towards meeting the thesis objectives. In brackets is the relevant chapter or appendix of this thesis document.	33
Figure 2.1: Schematic of the three different modelling approaches tested for producing calibrated multivariate forecasts of Tmin, Tmax and rainfall.	42
Figure 2.2: Plots comparing the overall performance of the various sets of forecasts (raw and post-processed) as the proportion of grid cells where certain bias, reliability and skill score values are exceeded. Columns are for the different metrics and rows are for the different climate variables. ...	55
Figure 2.3: Maps of Energy Skill Scores for UBJP+SS forecasts for the period 1981–2016. The skill scores are calculated using historical observations as climatological reference forecast and using leave-one-year-out cross-validation. Positive skill means lower error in the UBJP+SS forecasts compared to the reference. The skill is mapped for each target season for forecasts issued with one-month-lead-time.	58
Figure 2.4: As for Figure 2.3, except for TQM forecasts	59
Figure 2.5: As for Figure 2.3, except for MBJP forecasts	60
Figure 2.6: Maps of Variogram Skill Scores for UBJP+SS forecasts for the period 1981–2016. The skill scores are calculated using historical observations as climatological reference forecast and using leave-one-year-out cross-validation. Positive skill means lower error in the UBJP+SS forecasts compared to the reference. The skill is mapped for each target season for forecasts issued with one-month-lead-time.	61
Figure 2.7: As for Figure 2.6, except for TQM forecasts	62
Figure 2.8: As for Figure 2.6, except for MBJP forecasts	63

Figure 2.9: Summary of multivariate forecast performance across all grid cells and seasons, and a comparison of the results for various post-processing methods. The curves plot the proportion of cases where ES and VS skill score values are exceeded. The multivariate skill scores consider all three climate variables (Tmin, Tmax and rainfall) in their calculation. The VS is more sensitive to the calibration of the dependencies between the variables. 64

Figure 2.10: As for Figure 2.9, except that leave-one-year-out cross-validation has *not* been applied. 65

Figure 2.11: Comparison of multivariate skill scores for UBJP+SS and MBJP forecasts before and after reshuffling with samples generated from a BJP model fitted jointly to *observed* data. The -R suffix indicates reshuffled forecasts. Reshuffling improves the overall skill of UBJP+SS forecasts. 65

Figure 2.12: Variogram skill scores for MBJP-OBS, a BJP model fitted to observed data (no GCM predictors are involved). The skill scores are calculated using historical observations as the reference forecast and using leave-one-year-out cross-validation. The skill scores being predominantly close to 0 indicates very similar skill between model-fitted and pure-observation climatologies..... 67

Figure 2.13: Variogram skill scores for UBJP+SS forecasts after reshuffling with samples generated from the MBJP-OBS model fitted to observed data (see Figure 12). The -R suffix indicates reshuffled forecasts. The skill scores are calculated using UBJP+SS forecasts as the reference and using leave-one-year-out cross-validation. Positive skill means better VS in the UBJP+SS-R forecasts compared to the reference. The skill is mapped for each target season for forecasts issued with one-month-lead-time. The neutral-positive skill suggests the reshuffling is beneficial. 69

Figure 2.14: PBIAS (%) in UBJP+SS rainfall forecasts for the period 1981–2016. PBIAS departs from zero in very dry areas where the magnitude of the absolute bias is likely to be small..... 70

Figure 3.1: Schematic of the forecast-calibration multivariate-downscaling (FCMD) workflow to produce daily, spatially-downscaled ensemble forecast sequences from coarsely gridded monthly GCM forecasts..... 82

Figure 3.2: The location of the six study weather stations in north-eastern Australia (blue dots) and the ECMWF System4 grid cell boundaries (grey dashed lines). 91

Figure 3.3: CRPS skill scores for BJP-calibrated monthly and seasonal forecasts issued in September. Skill scores are calculated against climatological reference forecasts for the years 1981–2016. 95

Figure 3.4: PIT reliability plots for BJP-calibrated monthly and seasonal forecasts issued in September. Different coloured points represent different lead times from 0–5 months for monthly forecasts and 0–3 months for seasonal forecasts. The dashed-grey lines are the Kolmogorov-Smirnov significance bands. 96

Figure 3.5: Boxplots comparing the distribution of daily forecast values with the distribution of daily observed values for observations (O) and forecasts (F) during Sep-Nov. Boxplots show the IQR and the [0.05,0.95] quantile ranges. For rainfall, the distribution is plotted for wet days only (Precip \geq 1mm) and the proportion of wet days is also given (wet%). 98

Figure 3.6: As for Figure 3.5, except for Dec-Feb forecasts and observations 99

Figure 3.7: Average temporal correlations in Silo observations, FCMD forecasts (cal) and raw Sys4 forecasts for the Sep-Nov months. Kendall correlations are calculated for lags of 1-7 days. 101

Figure 3.8: As for Figure 3.7, except for the Dec-Feb months. 102

Figure 3.9: Comparison of average spatial Kendall correlations in Silo observations, FCMD forecasts and raw Sys4 forecasts for the Sep-Nov months. Each marker represents a station pairing (15 combinations). 104

Figure 3.10: Comparison of average inter-variable Kendall correlations in Silo observations, FCMD forecasts and raw Sys4 forecasts for the Sep-Nov months. Each marker represents a single location. 105

Figure 4.1: Map of the Tully sugarcane farming district located in north-eastern Australia and the location of the Tully Mill weather station 116

Figure 4.2 CRPS skill scores for the FCMD and FMBD climate forecasts for forecasts issued in September. Skill scores reflect performance relative to the climatology reference (CR) forecasts for

the period 1981–2015, and positive values indicate superior performance through greater accuracy and/or reliability..... 124

Figure 4.3 As for Figure 4.2 except for forecasts issued at the start of February for the period 1982–2016. Crop growth is simulated using observed meteorological forcing from September to the end of January with forecasts applied from February. 124

Figure 4.4: Boxplots summarising the biomass forecast distributions at the end of November, for forecasts issued at the beginning of September, for each year 1982–2016. The box covers the interquartile range and the whiskers cover the [0.1, 0.9] quantile range. The blue dot is the simulated biomass from observed meteorological data. FCMD and FBMD are GCM-driven forecasts and CR is the climatological reference-driven forecast. Leave-one-year cross-validation has been applied. 126

Figure 4.5: Boxplots of the percentage bias in the biomass forecasts from each meteorological-forcing data set. The red (leftmost) boxplot for each model summarises the bias for all forecast release months and target months (N=78). The blue (rightmost) boxplot in each case is for the harvest forecasts from each release month (N=12). The boxes are the interquartile range with the line across marking the median. The whiskers are the [0.1, 0.9] quantile range. The markers are all other cases. 127

Figure 4.6: As for Figure 2, except for the PIT reliability metric. The scale of the PIT metric is [0, 1]; however, the y-axis is limited to [0.5, 1] for clarity..... 129

Figure 4.7: CRPS skill scores for the FCMD-driven biomass forecasts for each forecast release and target month. Skill scores reflect performance relative to the climatological reference (CR) forecasts for the period 1982–2016, and positive values indicate superior performance through greater accuracy and/or reliability. The crop model is run with observed data from 1 September up to the beginning of the release month, hence the skill scores are available for the target months equal to the release month and beyond. 129

Figure 4.8: As for Figure 4.7, except for FBMD forecasts. 130

Figure 4.9: Left panel: Average shift in the forecast median of GCM-driven biomass forecasts, relative the CR-driven biomass forecasts, for each forecast release month and target month. The average is taken over the 35 verification years 1982–2016. Right Panel: As left panel, except for average relative dispersion, where dispersion is measured as the [0.1, 0.9] quantile range of the forecast. 131

Figure 5.1 Original contributions associated with thesis objective 1 highlighted in blue boxes. 136

Figure 5.2 Original contributions associated with thesis objective 2 highlighted in blue boxes. 139

Figure 5.3 Original contribution associated with thesis objective 3 highlighted in blue boxes..... 143

List of Tables

Table 1: Publications plan and current status.....	34
--	----

1. Thesis introduction

1.1. Preamble

In this chapter, I present the background and motivation for my research. Firstly, I introduce the drivers of seasonal climate variability in Australia. I then review existing agro-climatic decision support tools that help farmers understand climate information in the farming context. Thereafter I establish the degree to which seasonal climate forecasts are integrated into agro-climatic decision support tools. The climate community has largely progressed to forecasting using dynamical global climate models (GCMs) in preference to purely statistical forecasting. However, the agricultural modelling community have not adapted to the change, in part because sophisticated post-processing is needed to prepare GCM forecasts for quantitative use with other physically-based models and, in part, because GCM forecast data has not routinely been made publicly available. Post-processing methods calibrate forecasts to minimise biases and improve probabilistic reliability. Post-processing challenges include spatially and temporally downscaling climate forecasts to provide localised, physically coherent forecasts of rainfall, temperature and solar radiation at the scale required by crop models. I investigate the adequacy of existing GCM forecast post-processing methods for crop forecasting applications. Commonly used methods like simple mean bias correction and quantile-mapping are not fully adequate and crop modellers have found poor results from using these simple techniques. Harnessing dynamical climate model forecasts for crop forecasting applications requires the advancement of post-processing methods. Therefore, I identify research gaps to address the problem and set out a plan to develop and evaluate multivariate post-processing methods as the basis for my research. My thesis structure and publications end the chapter.

1.2. Weather, Climate and Agriculture

Australian weather and climate are associated with a complex and interacting set of seasonal climate drivers including the El Niño-Southern Oscillation (ENSO), the Indian Ocean Dipole, the Madden-Julien Oscillation and the Southern Annular Mode (Marshall et al. 2014b; Risbey et al. 2009; Schepen et al. 2012). The seasonal climate is modulated on longer time scales by slow climatic oscillations (e.g., the Inter-Decadal Pacific Oscillation) (Fita et al. 2017; Meinke et al. 2005; Power et al. 1999) and climate change (Murphy and Timbal 2008; Yeh et al. 2009). The size and positioning of the Australian continent mean a profusion of climates is presented, which may be broadly categorised into tropical, sub-tropical, temperate, grassland and desert zones (Stern et al. 2000).

For agriculture, variability in seasonal climate affects crop outcomes (Hammer et al. 2000; Nicholls 1986; Potgieter et al. 2002; Potgieter et al. 2005b). For example, Dreccer et al. (2018) examined regional relationships between temperature, water stress and crop yield for winter crops such as wheat, barley and chickpea and found regional differences in sensitivity to the type of and timing of weather fluctuations. For example, in the Western Australia cropping belt, higher overnight temperatures before flowering are associated with increased yields for wheat, barley, canola and chickpea. In south-eastern Australia, elevated early-season maximum temperatures are associated with reduced yields. Overall, canola was found to be particularly susceptible to water stress.

Seasonal climate patterns are also associated with long-duration, high-impact events. For example, the widespread and near-decade-long Millennium Drought (2001–2009) had major hydrological, ecological and agricultural impacts in southeast Australia (van Dijk et al. 2013) and required almost five billion dollars in financial assistance to crippled farmers (Howden et al. 2014). Relating to climate variability, the onset of the millennium drought was associated with recurrent El Niño conditions (over many years) and a consistently positive Southern Annular Mode phase (Verdon-Kidd and Kiem 2009). It is understandable then that Australian farmers have a keen interest in weather and climate variability as one aspect of managing farm risk.

To support Australian farmers with seasonal management decisions, the Australian Bureau of Meteorology monitor and forecast a range of climate and water variables. Historical rainfall and temperature data dating back to 1910 are readily available, as are maps of recent climate and water conditions, including measures of soil moisture and evapotranspiration. One- and three-month outlooks of seasonal rainfall, temperature and streamflow are updated at least once per month and communicated publicly. While the resources provide a wealth of information about the state of the climate and the land, it is difficult for farmers to cognitively translate the information into the farm context. Thus farmers often consult a range of information sources to build up a more comprehensive picture of how the evolving climate affects their farming prospects.

1.3. Agro-climatic decision support tools

Decision support tools help farmers to understand and integrate climate, landscape and other farm operation information to make evidence-based decisions and may include a forecasting component (Hammer et al. 2000) Decision support tools may be purely informational (e.g. a dashboard) or may interactively lead farmers through a series of decision steps (Rose et al. 2016). As such, climate and weather analysis tools, crop growth models, agricultural production simulators and water planning tools may all be considered to be valid components of an agro-climatic decision support system.

Farmers are likely to consult a range of decision support tools and their preferred tools will change over time. Well-designed tools are likely to receive greater use by farmers, especially if they are easy to operate and provide locally relevant information. Rose et al. (2016) suggest it is desirable that decision support tools are targeted to the type of farming activity and also that they scale to various sizes of farming operations. Moreover, the level of reliance on decision support tools is highly dependent on the willingness of the farmers to engage with the tools. Hochman and Carberry (2011) argue that it is more desirable for decision support tools to educate farmers' intuition, rather than replace it, and to allow experimentation with options, rather than propose optimal solutions.

Many decision support tools are available to Australian farmers to support climate risk management on farms. An example of a climate-focused decision support tool for agriculture is the CliMate app (Freebairn and McClymont 2012), which is available to registered users on a website and Android and iPhone smartphones. The CliMate app allows farmers to pinpoint a location and then review relevant weather and climate information from a range of sources. For example, rainfall and temperature in the current season can be compared to previous years using data from the SILO database (Jeffrey et al. 2001). Graphical indicators of the El Niño Southern Oscillation are presented, as are various drought and heat stress indicators. A local probability forecast for the chance of above median rainfall and temperature conditions is presented based on statistical relationships between Pacific and Indian Ocean sea surface temperatures and the local climate variables, using the linear discriminant analysis method of Drosdowsky and Chambers (2001). Additional seasonal forecasts from the Bureau of Meteorology and the Queensland Government are presented as continental-scale maps. While the CliMate app provides a wealth of information in a convenient package, it is generally up to the farmer or advisor to interpret it in the context of their farming activity.

To assist farmers to integrate climate information into agricultural domains, tools have been developed to filter climate information through crop models. Yield Prophet (e.g. Hochman et al. 2009) is one such example. Yield Prophet is a decision support system built around the Agricultural Production SIMulator (APSIM; Keating et al. 2003; McCown et al. 1996) for which wheat modules for APSIM were earlier developed (Meinke et al. 1998; Meinke et al. 1997). Yield Prophet specialises in modelling dryland grain crops, mainly wheat and barley. Yield Prophet uses recent weather observations and in-situ soil measurements, combined with historical climate information to predict not only crop yields but also risks including frosts, heat shock and waterlogging. The full version of Yield Prophet is a commercial product and requires tailoring for each customer.

A simplified version of Yield Prophet, called Yield Prophet Lite was released in 2016. Yield Prophet Lite is not connected to a crop model like the full version of Yield Prophet. Instead, Yield Prophet Lite

allows the user to estimate wheat, barley, canola or oat potential yield using the French & Schultz approach (French and Schultz 1984). The French & Shultz method estimates potential yield, i.e., the maximum possible yield under ideal conditions, using estimated rainfall for the growing season and a predefined set of parameters. An extra feature of Yield Prophet Lite is that it categorises potential yield based on historical rainfall deciles and displays the information alongside seasonal rainfall forecast deciles. However, it remains up to the farmer or advisor to reconcile the two pieces of information, which inherently requires an understanding of the skill and limitations of the seasonal forecasts.

In Western Australia, the government produces a seasonal rainfall outlook for southwestern Australia using statistical relationships between sea surface temperature, atmospheric pressure and observed rainfall. The regional forecast presented alongside the Bureau of Meteorology seasonal forecast and with a supporting narrative in a monthly newsletter. Similarly to Yield Prophet Lite, the WA government also uses the French & Schultz formula to estimate potential yield based on historical rainfall deciles, which users may attempt to reconcile with the seasonal forecast information.

Several decision support tools with more advanced climate forecast capability are in operation in Australia. However, these are run at the regional scale and, therefore, do not provide customised farm information. For example, the Queensland Government releases, each month, seasonal rainfall and pasture growth outlooks for Australia. These outlooks are the result of linking the Southern Oscillation Index (SOI) phase scheme seasonal forecasts (Stone et al. 1996) with the Aussie GRASS pasture growth model (Carter et al. 2000). To obtain inputs to the pasture growth model, the current state of the SOI phase (e.g. positive/negative, rising/falling) is used to conditionally sample historical analogues from observed data records. A very similar approach is used by the Queensland Alliance for Agriculture and Food Innovation (QAAFI) to produce their outlooks for wheat and

sorghum yields, which are based on regional-scale models (Potgieter et al. 2002; Potgieter et al. 2005a). A wheat yield outlook example is shown in Figure 1.1.

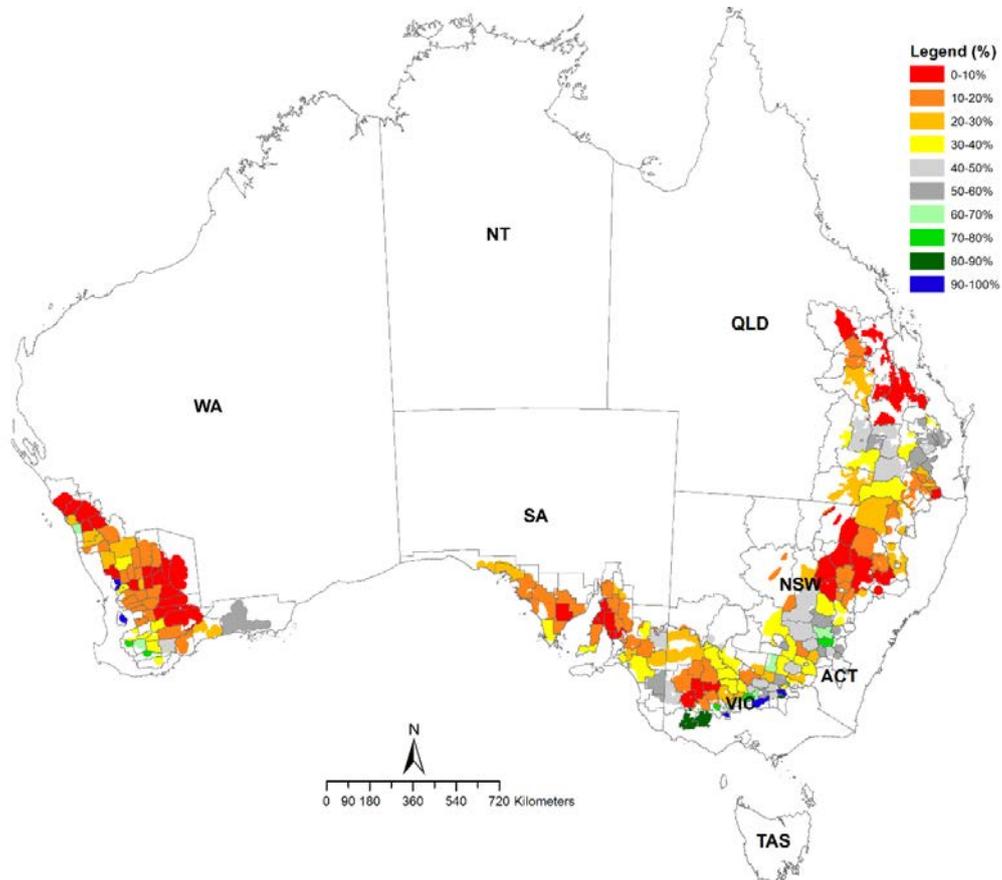


Figure 1.1: Wheat crop outlook for Australian wheat-growing regions, issued by the Queensland Alliance for Agriculture and Food Innovation (QAAFI) in July 2018. The colour scale represents the chance of exceeding the long-term median yield as simulated with the Oz-Wheat model. Figure supplied by QAAFI.

It is noteworthy that the official seasonal climate forecasts from the Australian Bureau of Meteorology (BoM) are not tightly integrated with the forecast agro-climatic decision support tools. The main reason is historic. Up until 2013, the Bureau of Meteorology operated a statistical forecasting system; essentially the same system as currently used in the CliMate app. That statistical forecasting system was only capable of producing forecasts of rainfall and temperature averages over a three month period. For crop forecasting using APSIM or some other crop model, daily

weather sequences and additional variables are required. Hence procedures to preferentially select historical weather sequences, e.g. based on SOI phases, were readily developed and remain in use today. However, the Bureau of Meteorology now generates seasonal climate forecasts using a dynamical global climate model (GCM), which has the potential to be harnessed and tightly integrated into agro-climatic decision support tools.

1.4. Ensemble seasonal global climate model (GCM) forecasts

Many meteorological institutes around the world now develop GCMs for seasonal forecasting. The European Centre for Medium Weather Forecasts (ECMWF) began developing its seasonal forecasting GCM in 1997 (Molteni et al. 2011); the Australian BoM followed suit in 2002, as did The National Centers for Environmental Prediction (NCEP) in the United States in 2004 (Saha et al. 2014). Owing to the complexity of the systems, GCM development has been gradual but most systems could now be considered mature, and many are used to issue public forecasts. The ECWMMF recently upgraded, in late 2017, from System4 to SEAS5. NCEP now operates CFSv2 and the BOM is in the process of replacing POAMA with ACCESS-S (Hudson et al. 2017a), which is a derivative of the UK's GloSea5 model (MacLachlan et al. 2015). In some countries, multiple agencies develop GCMs independently. For example, in the United States, in addition to running CFSv2, NCEP aggregates around 7 competing GCMs into the North American Multi-Model Ensemble (NMME; Kirtman et al. 2014). In Australia, the Commonwealth Scientific and Industrial Organisation (CSIRO) is developing a forecasting model with a multi-year focus, albeit with coverage of seasonal timescales.

The groundswell of support for dynamical models stems from developing evidence that GCMs are beginning to offer skill in excess of longstanding statistical methods (e.g. Barnston et al. 2012). Increases in computing capacity have enabled ensemble forecasting, which provide a range of possible outcomes. Other appealing features of GCM forecasts include the ability to present forecasts in the context of all the relevant climate drivers such as the El Niño Southern Oscillation, the Southern Annular Mode and the Madden-Julien Oscillation, amongst others (Marshall et al.

2014a; Marshall et al. 2014b; Marshall et al. 2011, 2012). Such granularity permits examination of the behaviour of the forecasts within and across seasons (e.g. Hudson et al. 2011; Vitart 2014). That said, GCMs have major limitations. While stunning in their complexity, GCMs are plainly simplified models of the real world, designed mainly to predict global climate patterns. Consequently, they often aren't able to precisely replicate weather and climate at regional or smaller scales, e.g., farm scales or point locations (e.g. Hagedorn et al. 2005; Tian et al. 2014). For localised applications, statistical post-processing is necessary to reduce errors and harness the useful, intrinsic information in GCM forecasts.

As mentioned above, seasonal GCM forecasts are typically issued in the form of ensembles that give a range of possible outcomes. However, the spread in the ensembles, which is supposed to convey forecast uncertainty, is not always reliable (Weisheimer and Palmer 2014). An example of an unreliable forecast is an over-confident one in terms of probabilities. Imagine a forecasting system that often predicts an 80-90% chance of a particular outcome, when in reality there is only a 50-60% chance of that outcome occurring. A farmer acting on the over-confident forecast information over the long term would make the wrong decision more often than they expected or make investments disproportionate to the real risk. While seasonal forecasts are highly uncertain and deserving of scrutiny before using for decision-making, over-confident forecasts are downright misleading and can lead to disillusionment amongst users. A realistic forecast of a 50% chance of a particular outcome is far preferable to an over-confident forecast and is informative in the sense that it helps the farmer understand the true likelihood and range of possible outcomes. By the same token, under-confident forecasts are also problematic and can lead to missed opportunities. Forecast reliability is, therefore, a critical element in helping farmers make optimal decisions that balance risk and opportunity.

1.5. Current methods for pairing GCM forecasts and crop models

Statistical post-processing is required to prepare GCM forecasts for use in quantitative models. The main reasons for forecast post-processing are to produce well-calibrated forecasts and to produce downscaled forecasts. Well-calibrated forecasts have minimal bias, are reliable in ensemble spread and have skill at least as good as climatology (Zhao et al. 2017). Downscaled forecasts have characteristics of a target region at a scale different to the GCM (e.g. a smaller grid cell or a weather station).

For crop forecasting, variables such as rainfall, temperature, solar radiation and evaporation are often needed as inputs to crop models (e.g. Capa-Morocho et al. 2016; Everingham et al. 2016; Han and Ines 2017; Jha et al. 2019). Similar variables are often needed for hydrological applications, e.g. for streamflow forecasting (e.g. Bazile et al. 2017; Lucatero et al. 2018) and so it is not surprising that similar post-processing methods are employed in both fields. Globally, it is very common to use methods such as linear scaling and quantile-mapping to post-process GCM forecasts for hydrological and crop forecasting applications (Bazile et al. 2017; Brown et al. 2018; Crochemore et al. 2016; Ines and Hansen 2006; Jha et al. 2019; Lucatero et al. 2018; Western et al. 2018). As a testament to the popularity of quantile-mapping, the Australian Bureau of Meteorology is developing a post-processing method based on quantile-mapping, which it intends to use to deliver forecasts on a 5km grid for agricultural applications.

Quantile mapping is a reasonably simple concept. Denote the GCM forecast variable of interest (which may be, for example, rainfall or temperature) as y_1 . The corresponding observed variable of interest is y_2 . Quantile mapping relies on knowing the cumulative distribution function for all previous raw forecasts $F_{\text{GCM}}(y_1)$ and the cumulative distribution function of all corresponding observations $F_{\text{OBS}}(y_2)$ (Maraun 2013; Wood et al. 2002; Zhao et al. 2017). A general quantile-mapping function may be expressed as:

$$y_2 = F_{\text{OBS}}^{-1}(F_{\text{GCM}}(y_1)) \quad (1)$$

where F_{OBS}^{-1} is the quantile (inverse) function of F_{OBS} .

In Australia, Brown et al. (2018) forced APSIM wheat models with quantile-mapped forecasts from the Predictive Ocean-Atmosphere Model for Australia (POAMA) to predict yield. The climate forecasts were shown to benefit yield forecast accuracy and narrow the forecast uncertainty range. However, the yield forecasts exhibited a consistent low-yield bias, which was attributed to the shortcomings of quantile-mapping to output post-processed GCM forecast ensembles with realistic autocorrelation structure. Failure to correct both the intensity and frequency of rainfall with quantile-mapping is part of the problem as previously illustrated by Ines et al. (2011). Western et al. (2018) also evaluated quantile-mapped POAMA forecasts, in this instance for plant-available soil water (PASW) forecasts. Whilst PASW forecasts were skilful, most of the skill was attributed to initial soil conditions, and rainfall and potential evapotranspiration forecasts verified worse than climatology. The skill deficiencies were similarly attributed to the limitations of the quantile-mapping to adequately reproduce ensemble forecasts with realistic spatial and temporal variability. Both studies identified the need for more advanced downscaling methods for agricultural applications. Internationally, Jha et al. (2019) post-processed CFSv2 forecasts using quantile-mapping for rice yield forecasting in Nepal. It was found that the GCM-driven forecasts performed poorly with respect to climatology-driven forecasts, partly because CFSv2 did not accurately represent the required intra-seasonal variability. CFSv2 was also found to have poor skill in rainfall forecasts in the target region, with certain events identified to be problematically over- or under-predicted. In accordance with the Australian studies, Jha et al. (2019) noted the inadequacy of quantile-mapping to overcome these problems and concluded that alternative downscaling methods need to be developed for yield prediction.

Prior to the recent efforts to simply apply quantile-mapping to all relevant climate variables, other studies have sought to apply weather generators conditioned on GCM forecasts to produce forecasts with realistic daily variability (e.g. Han et al. 2017; Ines et al. 2011; Semenov and Doblas-Reyes 2007). In some instances, rainfall is treated as the primary variable and temperature and radiation are taken as climatological averages (e.g. Han and Ines 2017; Ines et al. 2011). However, Baigorria et al. (2008), who used a regional climate model rather than a GCM, found that quantile-mapping of all variables (rainfall, temperature and radiation) performed better for crop yield forecasting than post-processing rainfall and using climatological averages for the other variables.

A small number of studies have explored methods other than quantile-mapping and weather generators for using seasonal GCM forecasts for yield forecasting. For example, Peng et al. (2018) combined CFSv2 seasonal climate forecasts with satellite data to predict maize yield using statistical model, finding a small benefit from using the GCM forecasts, although the modelling was largely deterministic in nature. A climate resampling approach whereby historical meteorological sequences are selected proportionally to tercile forecasts of below normal, near-normal and above-normal rainfall has also been developed (Capa-Morocho et al. 2016; Han and Ines 2017; Han et al. 2017). The name of the approach in the literature is FResampler1. In FResampler1, sequences of rainfall, temperature and solar radiation are selected simultaneously to preserve the inter-variable relationships. Han et al. (2017) established the benefits of using FResampler1 for rice crop decision support (e.g. fertilising, planting dates) in the Philippines. A drawback of the FResampler1 approach is that it relies solely on rainfall prediction skill, which is often low (e.g. Jha et al. 2019), and it assumes the forecast ensemble from which the tercile probabilities are derived is reliable, which is far from guaranteed (e.g. Weisheimer and Palmer 2014).

1.6. Advancing GCM forecast post-processing for agriculture

It is evident from the discussion in section 1.5 that quantile-mapping has emerged as a popular post-processing method for GCM forecasts, albeit with underwhelming results in agricultural applications.

The widespread adoption of quantile-mapping can be appreciated by viewing Equation (1), which implies that the implementation of quantile-mapping is flexible, in the sense the CDFs can be constructed using either empirical or parametric methods, and straightforward, in the sense that it requires minimal coding. However, it could be argued that quantile-mapping should be ruled out as a GCM forecast calibration and/or downscaling tool on the basis of its fundamental limitations.

Maraun (2013) pointed out the deficiency of quantile-mapping as a forecast downscaling tool.

Consider downscaling to multiple stations within a grid cell. Through quantile-mapping, each station will be assigned equally-ranked amounts of rainfall, which ignores spatial variability in rainfall amounts. A stochastic element is needed to overcome that problem. Furthermore, Zhao et al. (2017) examined quantile-mapping as a forecast calibration tool for Australian seasonal rainfall forecasts.

They found quantile-mapping is sufficient only when there is a pre-existing strong correlation between forecasts and observations, that is when the model is skilful. While quantile-mapping effectively minimises bias, it cannot guarantee reliability nor ensure that forecasts are better than climatology (the property of forecasts being at least as skilful as climatology is sometimes termed “coherence” (e.g. Zhao et al. 2017)). As evidenced by equation (1), quantile mapping is a one-to-one operation that does not consider the actual relationship between forecasts and observations, which may be weak, or inverse. Where multivariate forecasts are being post-processed, this is further problematic, since quantile-mapping lacks any model of covariance in general. It inherits spatial, temporal and inter-variable relationships from the raw GCM output, which may or may not be realistic. It is plausible that the fundamental limitations of quantile mapping are what manifested in the findings of Brown et al. (2018), Western et al. (2018) and Jha et al. (2019). A more comprehensive approach is evidently required to harness GCM forecasts for agricultural applications.

In their study on seasonal rainfall post-processing, Zhao et al. (2017) compared quantile-mapping with the Bayesian joint probability modelling approach (Schepen and Wang 2013; Wang and

Robertson 2011; Wang et al. 2009). BJP embeds a parametric multivariate model within a Bayesian framework for model inference and prediction. I do not go into the Bayesian components here (this will be shown in later chapters). For now, for comparison with quantile mapping, I note that the predictive equation for a bivariate normal model, which can be embedded within BJP, assuming normally distributed variables is:

$$y_2 \sim N\left(\mu_{y_2} + \rho \frac{\sigma_{y_2}}{\sigma_{y_1}} (y_1 - \mu_{y_1}), \sigma_{y_1}^2 (1 - \rho^2)\right) \quad (2)$$

where μ_{y_1} and μ_{y_2} are the modelled means of the GCM forecast and observed variables, respectively; σ_{y_1} and σ_{y_2} are the corresponding standard deviations, and ρ is the correlation between the GCM forecast and observed variables. Without going into detail about the estimation of the model parameters or the special handling of non-normal variables, it is visible that the BJP model involves not only the marginal distributions of the variables (as with quantile mapping) but also the correlation between forecasts and observations. Thus BJP has a potential advantage in that the post-processed forecast ensemble spread is modulated by the raw GCM skill. In the worst case that there is no correlation between GCM forecasts and observations, BJP reverts to the climatological distribution of observations. Consequently, Zhao et al. (2017) found that BJP outperformed quantile mapping, in particular by producing forecasts that were more reliable in ensemble spread and by eliminating cases where performance was worse than climatology.

In Australia, significant progress has been made on post-processing GCM rainfall forecasts using BJP for the purpose of streamflow forecasting. Bennett et al. (2016) and (Bennett et al. 2017a) developed the “Forecast-Guided Stochastic Scenarios” (FoGSS) system for forecasting monthly streamflow on seasonal time scales. FoGSS uses BJP post-processed rainfall forecasts from a GCM to drive a monthly water partition and balance model (WAPABA; Wang et al. 2011). The rainfall forecasts are post-processed using BJP to calibrate and downscale GCM forecasts from local and

large scale fields such as sea surface temperatures (Schepen and Wang 2013, 2014; Schepen et al. 2014).

A limitation of BJP, as a parametric post-processing method, is that it becomes unwieldy to apply for multiple lead times or locations. Empirical ensemble reordering methods have become popular in recent times to establish realistic patterns in ensemble forecasts from template data (Bellier et al. 2017; Clark et al. 2004; Schefzik 2016b; Scheuerer et al. 2017; Wu et al. 2018). For example, when spatial, temporal and inter-variable correlation structures are borrowed from historical data, the method is broadly known as the Schaake Shuffle (Clark et al. 2004).

In FoGSS, the Schaake Shuffle is used to connect up ensemble members for monthly rainfall forecasts that have been post-processed independently. The combination of BJP and the Schaake Shuffle means FoGSS is able to generate reliable long-range streamflow forecasts that are skilful where possible, but typically no worse than climatology, a property sometimes termed “coherence” (Zhao et al. 2017). It is quite possible then, that similar approaches could be beneficial for agricultural applications.

1.7. Thesis objectives

The motivation of this thesis is to develop, investigate and evaluate methods to combine the Bayesian joint probability modelling approach and empirical methods to build new tools that can robustly post-process GCM forecasts of temperature, radiation and other variables that are important crop model inputs.

Post-processing an array of variables to provide inputs to crop models poses several new and difficult challenges. Chiefly, forecasts are required at a daily time step and the forecasts must have the correct spatial, temporal and inter-variable correlation structures within an ensemble structure. I propose to resolve these challenges and systematically link seasonal GCM climate forecasts to crop models. My research can be broken down into three main objectives:

Objective 1 - Develop and rigorously evaluate BJP-based methods to calibrate monthly and seasonal GCM forecasts of climate variables needed by crop models

Objective 2 -Develop and evaluate methods for downscaling forecasts to high spatial and temporal resolution as needed by crop models

Objective 3 -Assess the utility of post-processed forecasts as inputs to crop models, and evaluate the performance of crop yield forecasts.

Completion of objectives 1 to 3 will demonstrate an end-to-end solution that is potentially generally applicable to routinely link seasonal global climate model forecasts and crop models.

1.8. Thesis structure and publications

The thesis hereafter is presented as a “three-paper thesis” with three main data chapters, each being written in the style of a journal article addressing one of the objectives. Ancillary work is presented as two additional papers in the appendices, giving a total of five papers. Each chapter and appendix encompasses the relevant literature review and discussion. The target journals and current status of publications are discussed in the preamble to each chapter and listed in Table 1 below. A conclusion chapter synthesises the main findings and sets future directions while rounding out the main body of the thesis.

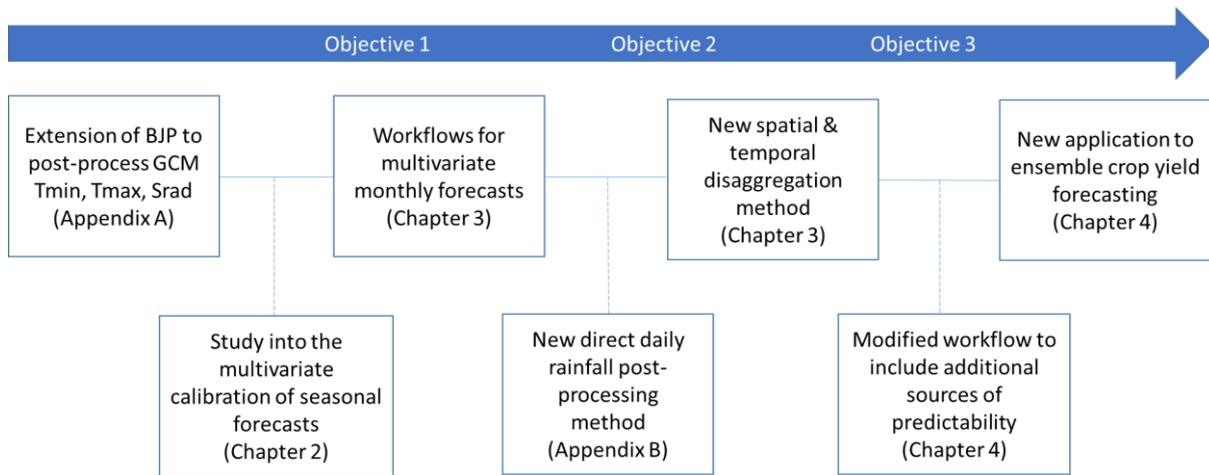


Figure 1.2 Graphical overview of my original contributions and progression towards meeting the thesis objectives. In brackets is the relevant chapter or appendix of this thesis document.

Table 1: Publications plan and current status

Title (relevant chapter / appendix)	Authorship	Journal	Status
Calibration, Bridging, and Merging to Improve GCM Seasonal Temperature Forecasts in Australia (Appendix A)	Schepen, A., Wang, Q.J. Everingham, Y.	Monthly Weather Review (IF = 3.4)	Published (2016)
On the joint calibration of multivariate seasonal climate forecasts from GCMs (Chapter 2)	Schepen, A., Everingham, Y. Wang, Q.J.	Monthly Weather Review (IF = 3.4)	Peer-reviewed and in revision
Coupling forecast calibration and data-driven downscaling for generating reliable, high-resolution, multivariate seasonal climate forecast ensembles at multiple sites (Chapter 3)	Schepen, A., Everingham, Y. Wang, Q.J.	International Journal of Climatology (IF = 3.6)	Peer-reviewed and in revision
A Bayesian modelling method for post-processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments (Appendix B)	Schepen, A. Zhao, T. Wang, Q. J. Robertson, D. E.	Hydrology and Earth Systems Science (IF = 4.3)	Published (2018)
Sugarcane crop yield forecasting by pairing downscaled multivariate dynamical climate model forecasts and a process-based crop model (Chapter 4)	Schepen, A., Everingham, Y. Wang, Q.J.	Agricultural and Forest Meteorology (IF = 4.5)	Internally reviewed in CSIRO and nearing submission

2. Calibrating multivariate seasonal forecasts from GCMs

2.1. Preamble

In Chapter 1, I established that GCM forecasts must be post-processed to use them as inputs to crop models. I explained that the Bayesian joint probability modelling approach is a versatile and reliable tool for post-processing seasonal GCM rainfall forecasts. Moreover, I argued that to prepare GCM forecasts for use in crop models it is vital to use full-calibration methods like BJP instead of simple bias-correction methods like quantile-mapping. Quantile-mapping, while popular, has several clear shortcomings as a forecast calibration and downscaling tool. While BJP has been widely applied to calibrate seasonal rainfall forecasts for hydrological applications, it has not been applied to post-process multivariate climate forecasts. In fact, there appears to be little understanding of the best way to approach multivariate GCM forecast calibration at seasonal time scales.

In this chapter, I address Objective 1: Develop and rigorously evaluate methods to post-process multivariate seasonal climate forecasts from GCMs. While the initial motivation was simply to extend BJP to post-process seasonal forecasts of minimum daily temperature, maximum daily temperature and solar radiation, my research expanded to more comprehensively investigate several candidate approaches to multivariate post-processing.

I develop and compare three approaches for post-processing multivariate seasonal GCM forecasts: (1) Direct multivariate post-processing using BJP; (2) univariate BJP post-processing with inter-variable relationships restored through empirical ensemble reordering; and (3) a novel implementation of quantile-mapping developed to provide consistent comparisons with BJP.

I run continental-scale experiments to test the post-processing methods across Australia and in all seasons. For the most part, I apply field-standard methods for forecast verification. However, the inclusion of multivariate scores, namely the Energy Score and the recently-developed Variogram

Score, adds a unique perspective and allows a deeper understanding of the relative performance of the methods, particularly around the modelling of inter-variable relationships.

The findings of this chapter are directly applied next in Chapter 3, which develops spatially and temporally downscaled forecast sequences from calibrated multivariate seasonal forecasts.

The main body of this chapter has been formatted as a journal article and submitted to *Monthly Weather Review* (Impact Factor 3.043) with the title “On the joint calibration of multivariate seasonal climate forecasts from GCMs” and with authorship: Schepen, A., Y. Everingham and Q.J. Wang. Although the paper is co-authored by my supervisors, I confirm the work is essentially my own. I conducted all of the experiments, analysed the results and wrote all of the paper, including preparing all of the figures. Everingham and Wang helped form the research questions and provided editorial support.

Additionally, prior to expanding this study to examine multivariate post-processing in detail, I examined the extension of BJP to average daily minimum temperature and average daily maximum temperature forecasts. This alone led to an article published in *Monthly Weather Review* with the citation:

Schepen, A., Q.J. Wang, and Y. Everingham, 2016: Calibration, Bridging, and Merging to Improve GCM Seasonal Temperature Forecasts in Australia. *Monthly Weather Review*, 144, 2421–2441.

The paper is also attached as Appendix A.

2.2. Introduction

Seasonal forecasts of climate variables are in high demand around the globe for informing decision-making in climate-sensitive industries and for water resources management. These days, global climate model forecasting systems (GCMs) are widely used for seasonal forecasting, in part because they generate a detailed global view of the climate state and, in part, because they output a broad

spectrum of climate variables of importance to sectors including hydrology, agriculture and public health. Many different GCMs have been developed internationally, with differences in component models (i.e. ocean, atmosphere, land surface and sea-ice), data assimilation strategies, ensemble generation schemes, scales, dynamics and physics leading to systems with vastly different biases and forecasting skill (e.g. Kim et al. 2012; Pegion et al. 2017). Even at the global scale, GCMs differ to some degree in their characterisation of dominant climate patterns such as ENSO (Barnston and Tippett 2013; Shi et al. 2012). Moreover, at the local scale, GCMs vary in their representations of key climate variables (e.g. rainfall and temperature) and associations with seasonal climate drivers (Kim et al. 2012; Lim et al. 2009; White et al. 2013; Zhao and Hendon 2009). Consequently, individual GCMs present nuanced outlooks around broader climate patterns.

For local decision-making and risk-taking on the basis of GCM forecasts, raw GCM forecasts require statistical post-processing to rectify model biases, reduce skill deficits and to improve overall reliability (e.g. Feddersen et al. 1999; Gneiting et al. 2005; Weisheimer and Palmer 2014; Zhao et al. 2017). GCM forecast ensemble spread typically is too narrow and doesn't vary appropriately from one forecast to the next (Barnston et al. 2015; Weisheimer and Palmer 2014). Moreover, where quantitative modelling is to be undertaken using GCM outputs, it is vital that ensemble members have a physically coherent structure across the relevant variables and, depending on the application, in space and time as well. Scheuerer and Hamill (2015) give the perfunctory example of snow-melt in spring being dependent on both rainfall and temperature, suggesting the joint distribution of rainfall and temperature is, therefore, an important consideration. Regression-based calibration and other forms of statistical post-processing are often only practical to apply to individual locations, time periods and variables (e.g. Doblas-Reyes et al. 2005). More problematically, GCM-modelled relationships between these dimensions are easily lost in post-processing where random sampling from statistical distributions occurs, requiring reestablishment of covariance structures through non-parametric ensemble reordering techniques such as ensemble copula coupling (Scheffzik et al. 2013) or the Schaake Shuffle (Clark et al. 2004). For example, Luo and Wood (2008) and Yuan and Wood

(2012) injected the spatiotemporal covariance from observations into rainfall and temperature forecasts generated by a Bayesian linear-regression technique, giving forecasts suitable for use in hydrological applications.

Elsewhere, the Bayesian joint probability modelling approach (BJP; Wang and Robertson 2011; Wang et al. 2009) has been applied to calibrate seasonal GCM forecasts in Australia (Hawthorne et al. 2013; Schepen and Wang 2013), China (Peng et al. 2014) and the United States (Strazzo et al. 2018). Rather than being a typical regression, BJP is designed to model the full joint distribution of any number of predictor and predictand climate variables after allowing for the independent transformation of the marginal distributions (hereafter, marginals). Post-processed ensemble members are obtained through a sequence of conditional sampling of the joint distribution and back-transformation. Various studies have found that BJP produces reliable probabilistic forecasts that capture inherent GCM skill; however, these studies have been limited to a univariate configuration (in the sense of dealing with a single variable). For example, BJP-calibrated seasonal forecasts of rainfall have been subjected to the Schaake shuffle and used to generate reliable long-range ensemble streamflow forecasts (Bennett et al. 2016; Bennett et al. 2017a). However, very little attention appears to have been given to the multivariate calibration of seasonal climate forecasts, which is essential for more complex applications such as agricultural crop-modelling, which requires coherent forecasts of rainfall, temperature and solar radiation.

In contrast to seasonal forecasting, the joint post-processing of weather variables in short-term (NWP) forecasting has become a topic of increasing interest in recent years. Several studies have investigated the bivariate calibration of the u and v components of wind vectors (McLean Sloughter et al. 2012; Pinson 2012; Schuhen et al. 2012) and the joint calibration of temperature and wind speed forecasts (Baran and Möller 2015, 2017; Schefzik 2016a). In particular, Baran and Möller (2015) introduced a Bayesian model averaging methodology and, later (Baran and Möller 2017), an ensemble model output statistics (EMOS) methodology for temperature/wind speed calibration,

both relying on a truncated bivariate normal construction. Earlier, Möller et al. (2013) presented a more general methodology that first calibrates the marginals independently, thereafter constructing the inter-variable dependence structure using Gaussian copulas. Baran and Möller (2017) concluded that all three aforementioned methods (EMOS, BMA and copula-reconstruction) yielded similar reliability and accuracy improvements over raw temperature/wind speed forecasts, and, therefore, they advocated for the bivariate EMOS approach for efficiency reasons.

Schefzik (2016a) surmised there are two broad approaches to multivariate post-processing of weather forecasts. The first is univariate post-processing followed by non-parametric ensemble recording methods to establish spatial, temporal and inter-variable correlation structures. The second is fully parametric post-processing, which is usually tailored for low-dimensional settings. Consequently, Schefzik (2016a) proposed a hybrid post-processing approach that jointly post-processes related variables in low-dimensional settings and thereafter applies an ensemble reordering method with a multivariate ranking to obtain final aggregated, post-processed forecasts for higher-dimensional spaces (e.g. across different locations or lead times). Similarly to earlier studies, the focus was on the truncated-bivariate-normal model for temperature and wind speed.

In this study, I investigate the merits of post-processing multivariate seasonal climate forecasts using several parametric and non-parametric methods. I propose a comparison of (1) directly post-processing multiple climate variables simultaneously using one BJP model; (2) post-processing each variable with a univariate BJP model and subsequently restoring the inter-variable correlations via the Schaake Shuffle; and (3) a quantile-mapping approach for a further reference. It is anticipated that testing these three different strategies will expose the numerous trade-offs that exist between the efficiency and dimensionality of parametric approaches, and the amenity of historical data to fit the parametric model and/or provide realistic covariance structures. While it has been suggested that parametric approaches are quite suitable for low-dimensional forecast calibration problems (Schefzik 2016a; Vannitsem et al. 2018), a priori, I do not suspect which approach will perform better

for seasonal forecast calibration. Direct multivariate calibration may be challenged by the number of parameters relative to a small number of data points available (typically 20-40 for seasonal post-processing). Indeed, Doblas-Reyes et al. (2005) found difficulties establishing robust regression coefficients when using multiple regression for combining multiple seasonal forecasts. That said, studies using BJP for hydrology have successfully exploited its ability to model multiple predictands for forecasting streamflow at multiple sites (Wang and Robertson 2011; Wang et al. 2009) and for multiple months ahead (Zhao et al. 2016), situations where the covariances are likely to be well-structured.

In this study, I target one-month-lead-time forecasts of seasonal (three-month-average) rainfall, minimum temperature and maximum temperature for Australia. These variables are core products in seasonal forecast services globally. For now, my remit is restricted to modelling of inter-variable correlations — models are developed for each month and grid point individually. Forecast skill and reliability are assessed using ECMWF System4 hindcasts from 1981–2016, establishing separate models for each start month from January to December, and with a forecast lead time of 1 month. Forecast skill is quantified as the improvement over a seasonally-dependent climatology reference formed from observations. As another reference for the performance of BJP-calibration, I develop a novel version of quantile-mapping that is consistent with BJP in terms of modelling the marginals. Quantile mapping adjusts the location and ensemble spread of the GCM forecasts but simply transfers information about inter-variable relationships from the raw model output into the observation space; thus it doesn't involve a correction based on the correlation between forecasts and observations, but it has the benefit of fewer parameters. Hereafter I present the modelling and verification methods, followed by a continental scale study, results, discussion and conclusions.

2.3. Methods

2.3.1. Multivariate calibration strategies

Before getting into the detailed methods, I introduce the three general approaches that are developed and tested in this study for multivariate calibration of Tmin, Tmax and rainfall:

- (1) Simultaneous calibration of all climate variables in one BJP model; termed multivariate BJP (MBJP)
- (2) Independent BJP calibration for each variable followed by restoration of inter-variable correlations via the Schaake Shuffle ensemble reordering method; termed univariate BJP plus Schaake Shuffle (UBJP+SS)
- (3) Quantile mapping of transformed variables (TQM)

The workflow for each of these three approaches is shown in Figure 2.1.

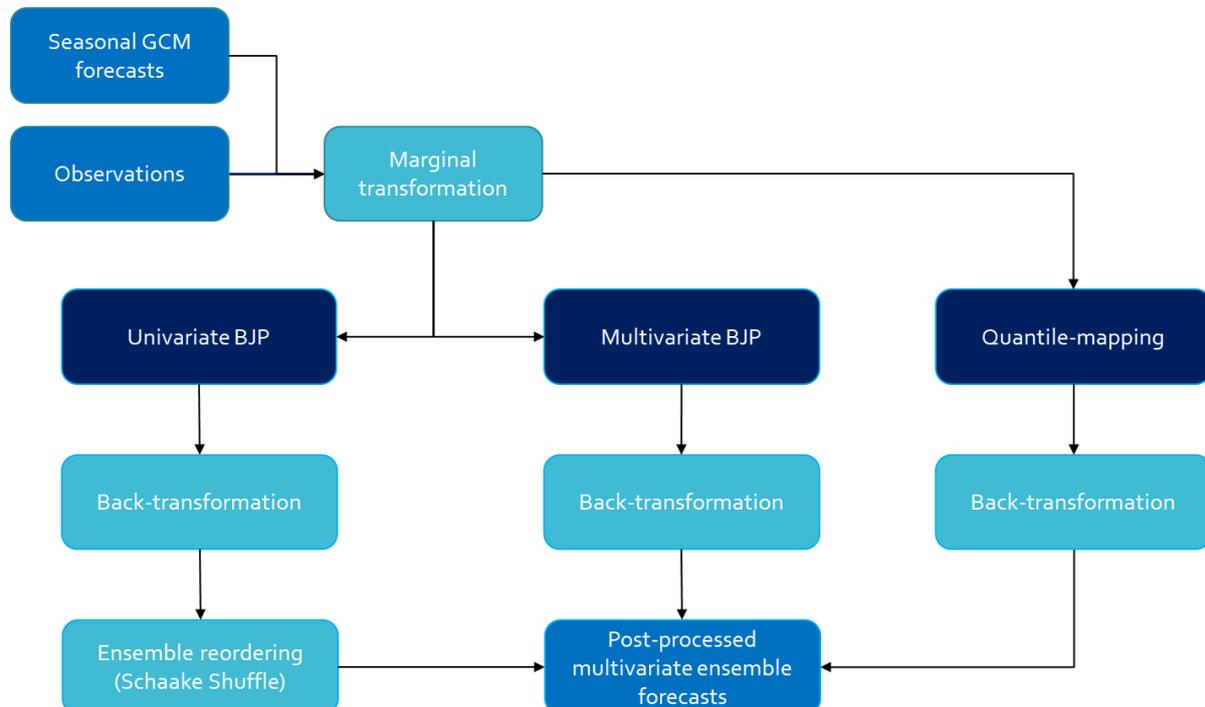


Figure 2.1: Schematic of the three different modelling approaches tested for producing calibrated multivariate forecasts of Tmin, Tmax and rainfall.

2.3.2. Marginal transformation

The three post-processing methods are constructed with the working assumption that the marginal distributions are able to be modelled as normal distributions after being subjected to variance-stabilising transformations. The assumption is patently reasonable for variables like temperature, except that the normal distribution has infinite support and, therefore, the tails may not represent extremes precisely. For rainfall, which ostensibly has a mixed discrete-continuous distribution, the way forward is not immediately obvious. Nevertheless, the ability to model its distribution using a transformed-normal is highly desirable because it allows post-processing of rainfall in the same framework as temperature. The solution adopted here is to treat rainfall data as being left-censored. That is, rainfall data with a value of 0, or some other minimum measurable amount, are assumed to have a true value of less than or equal to that amount, with the precise value unknown. Standard statistical methods are available for the normal distribution and censored data and, therefore, it is possible to use variance-stabilizing transformations for all variables in BJP.

The degree, or the “strength”, of the transformation required to achieve normality, depends on several factors including the range, scale and skewness of the data. I employ two flexible variance-stabilizing transformations in this work. The reason for using two different transformations is because I use the log-sinh transformation (Wang et al. 2012b) for rainfall, which was developed specifically for hydrological variables. All other variables use the Yeo-Johnson transformation (Yeo and Johnson 2000). While temperature is often modelled using a normal distribution, which suggests no transformation is required, preliminary investigations revealed statistically-significant skewness in temperature distributions in some regions and seasons in Australia (not shown) and, therefore, I allow for transformation if needed. The flexibility of the variance-stabilising transformations effectively allows for little or no transformation if need be.

Temperature variables are transformed by the single parameter Yeo-Johnson transformation (Yeo and Johnson 2000):

$$\psi_{\lambda}(y) = \begin{cases} ((y+1)^{\lambda} - 1) / \lambda & \lambda \neq 0, y \geq 0 \\ \log(y+1) & \lambda = 0, y \geq 0 \\ -[(-y+1)^{2-\lambda} - 1] / (2-\lambda) & \lambda \neq 2, y < 0 \\ -\log(-y+1) & \lambda = 2, y < 0 \end{cases} \quad (3)$$

The Yeo-Johnson transformation is highly flexible and can be used to transform both positively and negatively skewed data. It incorporates a range of useful transformations, including the log, square root and inverse transformations and embeds the historically popular Box-Cox transformation (Box and Cox 1964). The main distinction of the Yeo-Johnson transformation from the Box-Cox transformation is that it permits zero and negative values and is, therefore, more readily useful for modelling data presented as anomalies. In this study, transformations are established by using Bayesian maximum a posteriori (MAP) estimation of λ for the posterior probability of (λ, μ, σ) where μ and σ are the normal distribution mean and standard deviation parameters. The full details of the Bayesian estimation procedure, including specification of the prior distributions, is given in Appendix A.

As mentioned, rainfall is transformed by a two-parameter log-sinh transform (Wang et al. 2012b):

$$\psi_{\varepsilon, \lambda}(y) = \frac{1}{\lambda} \log(\sinh(\varepsilon + \lambda y)) \quad (4)$$

where ε and λ are transformation parameters. The log-sinh transformation was developed by Wang et al. (2012b) to handle the pattern of errors in hydrological predictions, which is exemplified by rapid growth at smaller magnitudes followed by tapering to nearly nil growth at higher magnitudes. The log-sinh transformation has been widely applied to transform rainfall and streamflow data in statistical modelling of hydrological data (e.g. Bennett et al. 2016; Giudice et al. 2013; Robertson et al. 2013). MAP estimation of ε and λ is carried out for the posterior probability of $(\varepsilon, \lambda, \mu, \sigma^2)$ using the same type of procedure as for the Yeo-Johnson transformation.

2.3.3. Multivariate BJP calibration (MBJP)

Multivariate BJP calibration is when several different climate variables are calibrated jointly in the one model, with covariance explicitly modelled. The BJP modelling approach uses a multivariate normal distribution to model the relationship between the transformed predictor and predictand variables (hereafter referred to as predictors and predictands). The collection of d transformed predictors and predictands form a column vector $\mathbf{z}^T = [z_1 \ z_2 \ \dots \ z_d]$. Once the marginals have been transformed using a variance-stabilizing transformation, it is assumed that the joint distribution is multivariate normal:

$$\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (5)$$

where $\boldsymbol{\mu}$ is the mean vector

$$\boldsymbol{\mu}^T = [\mu_1 \ \mu_2 \ \dots \ \mu_d] \quad (6)$$

$\boldsymbol{\Sigma}$ is the covariance matrix

$$\boldsymbol{\Sigma} = \mathbf{D}(\boldsymbol{\sigma}) \times \mathbf{P} \times \mathbf{D}(\boldsymbol{\sigma}) \quad (7)$$

$\mathbf{D}(\boldsymbol{\sigma})$ is a diagonal matrix from the standard deviation vector.

$$\boldsymbol{\sigma}^T = [\sigma_1 \ \sigma_2 \ \dots \ \sigma_d] \quad (8)$$

and \mathbf{P} is the symmetric correlation matrix

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,d} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{d,1} & \rho_{d,2} & \cdots & 1 \end{bmatrix} \quad (9)$$

giving a total of $2d + d(d-1)/2$ parameters in addition to the transformation parameters. Previous descriptions of BJP in the literature detail an inference method based on a Metropolis sampler

(Wang and Robertson 2011; Wang et al. 2009). Here, I use a more efficient Gibbs sampler to infer $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Wang et al., 2019; manuscript in review at Environmental Modelling and Software). The following uninformative prior is specified to complete the Bayesian formulation

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(d+1)/2} \quad (10)$$

Beyond the description included here, BJP includes treatments to allow inference in the presence of missing values and censored data. These treatments are described by Wang and Robertson (2011).

To use BJP as a forecasting tool, the multivariate normal distribution is conditioned on the predictors. For a single set of parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, consider the transformed predictors \mathbf{z}_1 and predictands \mathbf{z}_2 organized as

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} \quad (11)$$

and the mean vector and covariance matrix correspondingly partitioned like so:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad (12)$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \quad (13)$$

The conditional distribution of the predictands given the predictors is also a multivariate normal distribution:

$$\mathbf{z}_2 | \mathbf{z}_1 \sim N(\boldsymbol{\mu}', \boldsymbol{\Sigma}') \quad (14)$$

where

$$\boldsymbol{\mu}' = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} [\mathbf{z}_1 - \boldsymbol{\mu}_1] \quad (15)$$

$$\Sigma' = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \quad (16)$$

Forecast values are sampled from the distribution given by Equation (14) and back-transformed to the original space. Gibbs sampling is used to obtain one sample from $\mathbf{z}_2|\mathbf{z}_1$ for M different sets of parameters, thus generating an ensemble of size M .

2.3.4. Univariate BJP calibration plus Schaake Shuffle (UBJP+SS)

Univariate BJP calibration is when there is only one climate variable under consideration (although there are technically two variables in the model: the BJP predictor and the BJP predictand). To establish coherent multivariate forecasts after applying univariate BJP to each variable, I apply the Schaake Shuffle ensemble reordering method (Clark et al. 2004). The Schaake Shuffle imposes the rank correlation structure of randomly-selected historical observations into forecasts. I describe the essential steps of the procedure here. For a given forecast time period (e.g. month), consider an ensemble forecast of size M denoted by

$$\mathbf{X} = (x_1, x_2, \dots, x_M) \quad (17)$$

that can be sorted to obtain

$$\boldsymbol{\chi} = (x_{(1)}, x_{(2)}, \dots, x_{(M)}) \quad x_{(1)} \leq x_{(2)} \dots \leq x_{(M)} \quad (18)$$

Consider also a vector of observations from the historical record for the same time period (e.g. the same month in other years), also of size M

$$\mathbf{Y} = (y_1, y_2, \dots, y_M) \quad (19)$$

that can be sorted to obtain

$$\boldsymbol{\gamma} = (y_{(1)}, y_{(2)}, \dots, y_{(M)}) \quad y_{(1)} \leq y_{(2)} \dots \leq y_{(M)} \quad (20)$$

Furthermore, let rank be a function that determines the position of a value from γ in the original unsorted vector \mathbf{Y} . The shuffled forecast ensemble is constructed as

$$\mathbf{X}_{SS} = (x_{ss,1}, \dots, x_{ss,M}) \quad (21)$$

where $x_{ss,q} = x_{(n)}$ and $q = \text{rank}(\mathbf{Y}, y_{(n)})$ $n = 1, \dots, M$. In this study, to construct \mathbf{Y} , I start with historical daily data offset -30, -15, 0, 15 and 30 days from the start of the seasonal forecast and omitting any data overlapping the forecast. Daily data is aggregated to match the number of days in each season forecast to be shuffled and is subsequently used in the reordering.

2.3.5. Transformed Quantile-Mapping (TQM)

Quantile mapping is a popular method for bias-correcting climate model outputs in impacts studies. It has no model of covariance. Instead, it relies on the inter-variable correlations in the GCM being approximately correct, and, therefore, it isn't a full calibration method (Maraun 2013; Zhao et al. 2017). However, it is a method currently supported by the Australian Bureau of Meteorology and being investigated in agricultural applications of seasonal forecasts (e.g. Brown et al. 2018; Western et al. 2018), and, therefore, it is a useful method for comparison purposes.

Quantile-mapping comes in many forms, which boil down to two main types: empirical quantile-mapping and parametric quantile-mapping. In this study, I develop a new, parametric quantile-mapping methodology using the fitted log-sinh or Yeo-Johnson transformed normal distributions from section 2.3.2 to represent the marginal distributions. Hence I call it transformed quantile-mapping (TQM). Accordingly, the TQM and BJP methodologies model the marginals of each variable in an entirely consistent way, meaning that the results of BJP and QM post-processing are more comparable than if I used another QM implementation.

TQM is described as follows in two parts:

- (1) Model the marginal distributions of the forecasts and observations

- a. Collect all the historical forecast ensemble members
- b. Fit a transformed-normal distribution to the forecasts using either the log-sinh or Yeo-Johnson transformation. Save the estimated normal distribution parameters μ_F and σ_F and the transformation τ_F .
- c. Collect all the observations corresponding to the forecasts from step (a). There will be fewer observation data points than forecast data points because the forecasts are ensembles.
- d. Fit a transformed-normal distribution to the observations using either the Log-Sinh or Yeo-Johnson transformation. Save the estimated normal distribution parameters μ_O and σ_O and the transformation τ_O .

(2) Post-process a new ensemble forecast

- a. Transform the i^{th} ensemble member $y_{F,i}$ to $z_{F,i} = \tau_F(y_{F,i})$
- b. Convert $z_{F,i}$ to a dimensionless z-score: $z_{F,i}^* = (z_{F,i} - \mu_F) / \sigma_F$
- c. Invert $z_{F,i}^*$ using μ_O and σ_O to get $z_{O,i} = (z_{F,i}^* \times \sigma_O) + \mu_O$
- d. Back transform $z_{O,i}$ to $y_{O,i} = \tau_O^{-1}(z_{O,i})$
- e. Repeat steps (a)-(d) for all ensemble members, $k = 1, \dots, M$

The procedure is a fully parametric implementation of quantile-mapping. It differs substantially from any other implementation in the literature because it makes use of the log-sinh and Yeo-Johnson transformations that were developed for use with BJP. In addition, the new method handles the mixed discrete-continuous nature of variables like rainfall using a censored data approach, which is quite different to the more common split-model approach, whereby intensity and frequency are modelled using separate distributions (e.g. Volosciuk et al. 2017).

2.4. Application and verification

2.4.1. Study data

I now evaluate the multivariate post-processing of GCM seasonal forecasts of rainfall, minimum temperature maximum temperature for Australia. These three variables form the basis for seasonal outlooks in Australia and are routinely assessed (e.g. Hudson et al. 2011; Marshall et al. 2014a; Marshall et al. 2014b). Australia is currently switching to a new GCM (ACCESS-S; Hudson et al. 2017b) and doesn't yet have long hindcasts available for verification and calibration studies. In this study, GCM forecasts are obtained from the ECMWF System4 (Sys4) seasonal forecast system, which has been widely evaluated globally.

Sys4 is a coupled system of ocean, atmosphere and land-surface models with sea-ice concentration conditionally resampled from climatology. It implements the NEMO (Nucleus for European Modelling of the Ocean) v3.0 ocean model at a 1-degree resolution in the extratropics. It implements the IFS (integrated forecast system) cycle 36r4 atmospheric model with an approximate horizontal resolution of 80 km. The H-TESEL (Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land) land surface model is integrated into IFS.

Hindcasts are available from 1981–2010 with each model run initialised on the 1st of each month and enduring for 7 months. The hindcast data set is augmented by an archive of real-time forecasts from 2011–2016. In hindcast mode, the ensemble generation scheme outputs 15 ensemble members. In forecast mode, the ensemble size increases to 51. Throughout this study, I make use of the first 15 ensemble members for all years 2011–2016. Hindcasts and archived real-time forecasts are treated as equivalent. All members are treated as statistically exchangeable.

Gridded observed data come from the Silo patched-point database (Jeffrey et al. 2001). Silo data is constructed from Bureau of Meteorology observational records and has been infilled to create a temporally-complete record for all locations. I take the Silo data to be tantamount to observations,

noting that the data quality is dependent on the degree of quality control in Silo processing, the amount of processing, and the density and quality of the original observations. Silo data are available on a 0.05-degree (approximately 5km) grid. I regrid the Silo observations to match the Sys4 data at 0.75-degree (~80km horizontally) resolution.

In this study, I choose to focus on three-month-average forecasts, with a lead-time of 1 month. These types of forecasts represent a true seasonal outlook beyond the current information available about the weather. BJP models are established separately for 12 overlapping seasons from Jan-Feb-Mar (JFM) to Dec-Jan-Feb (DJF). With this configuration, there are 35 data points available to fit each calibration model at each grid cell.

As a preview to the inter-variable relationships in seasonal observations, I calculate the absolute Kendall correlation for all grid cells and months. Between Tmin and Tmax, the median Kendall correlation is 0.34 and the 90th percentile is 0.58. Between Tmax and rainfall (which tend to be negatively correlated), these values are 0.35 and 0.55. For Tmin and rainfall, the result is 0.18 and 0.4. These preliminary results suggest it is prudent to handle inter-variable dependencies in seasonal forecast post-processing of rainfall and temperature.

2.4.2. Univariate and multivariate probabilistic forecast verification

I first apply univariate bias and reliability scores to check the consistency of forecasts and observations for the individual variables. I then apply two multivariate probabilistic scores to assess the overall skill and performance for all variables. In general, quality seasonal forecasts will have little or no bias, be reliable in terms of ensemble spread and supply skill in excess of a climatological reference forecast. All of these aspects of forecast quality are verified here using a leave-one-year-out cross-validation approach.

Forecast bias is recognised as the long-term mean error between forecasts means and observations.

For a single variable, I calculate the percentage bias,

$$\text{PBIAS} = \frac{\sum_{t=1}^T (\bar{x}_t - y_t)}{\sum_{t=1}^T (y_t)} \times 100 \quad (\%) \quad (22)$$

where \bar{x}_t is the forecast ensemble mean for event t , and y_t is the corresponding observation.

Positive PBIAS indicates systematic over-forecasting whereas negative PBIAS indicates systematic under-forecasting.

Reliability is the property of statistical consistency between probabilistic forecasts and observations.

A reliable forecasting system will accurately estimate the likelihood of an event. Reliability is checked by analysing the distribution of probability integral transformations or PIT values (Gneiting et al. 2007). The PIT for a forecast CDF (F_t) and paired observation (y_t) is defined by

$$\pi_t = F_t(y_t) \quad (23)$$

In the case that $y_t = 0$, a pseudo-PIT value is sampled from a uniform distribution with a range $[0, \pi_t]$ (Wang and Robertson 2011) and this value then supplants the original π_t . If a forecasting system is reliable and the forecasts are continuous, then the PIT values for a set of forecasts follow a standard uniform distribution. Hence, I quantitate reliability using a score that measures the deviation of the PIT values from the theoretical standard uniform values (Renard et al. 2010)

$$\text{REL}_{\text{PIT}} = 1.0 - \frac{2}{T} \sum_{i=1}^T \left| \pi_{(i)} - \frac{i}{T+1} \right| \quad (24)$$

where $\pi_{(i)}$ is the i^{th} ranked PIT value. REL_{PIT} ranges from 0 (worst reliability) to 1 (perfect reliability). Visualisation of REL_{PIT} and its interpretation in the context of PIT uniform probability plots are given by Renard et al. (2010).

The overall skill and performance evaluation of the multivariate forecasts is done using multivariate scores, namely the energy score (ES; Gneiting and Raftery 2007) and the variogram score (VS; Scheuerer and Hamill 2015). For an M ensemble member forecast for N components and multivariate observations \mathbf{y} :

$$\text{ES} = \frac{1}{M} \sum_{k=1}^M \|\mathbf{x}_k - \mathbf{y}\| - \frac{1}{2M^2} \sum_{k=1}^M \sum_{l=k}^M \|\mathbf{x}_k - \mathbf{x}_l\| \quad (25)$$

where \mathbf{x}_k is the forecast for ensemble member k and $\|\cdot\|$ denotes a Euclidean-norm. In a single dimension, the ES reduces to the widely-used continuous ranked probability score (CRPS) for single-variable verification.

The ES is an effective measure for determining the aggregate skill of many individual components, however, it is rather insensitive to the miscalibration of dependencies between components (Scheuerer and Hamill 2015). The VS can be much more sensitive to such miscalibration. Using the same notations as for the ES, the VS based on variograms of order p can be estimated for an ensemble forecast by:

$$\text{VS} = \sum_{i=1}^N \sum_{j=1}^N w_{ij} \left(|y_i - y_j|^p - \frac{1}{M} \sum_{k=1}^M |x_{k,i} - x_{k,j}|^p \right)^2 \quad (26)$$

where w_{ij} are weights to promote/demote certain pairs in the calculation of the VS. For example, in the spatial case, it can be used to up-weight proximate pairs and down-weight distant pairs. Here I set $w_{ij} = 1$ to consider all pairings of variables equally; and $p = 0.5$ as commonly used.

The calculate ES and VS will be calculated for variables with different units, which makes the results more challenging to interpret than, for example, applications to one variable across space and/or time. To make the comparison more meaningful, I make the variables dimensionless before calculating the scores. Rainfall is standardised by the dividing by the mean of observations.

Temperature variables are standardised by a z-score transform.

For ES and VS I calculate a skill score (SS) where S is the average score of the post-processed forecasts over a set of events and S_{ref} is the average score over the same events for a climatological reference set of forecasts.

$$SS = \frac{\overline{S_{\text{ref}}} - S}{\overline{S_{\text{ref}}}} \times 100 \quad (\%) \quad (27)$$

Reference forecasts are leave-one-year-out observation data for the same period as the forecasts.

2.5. Results and discussion

2.5.1. Bias, reliability and skill of individual variables

The percentage bias (PBIAS), reliability score (REL_{PIT}) and CRPS skill score metrics are summarised for each variable (Tmin, Tmax and rainfall), for raw forecasts (RSYS4), and for each set of post-processed forecasts (UBJP+SS, MBJP and TQM) (Figure 2.2). The summaries plot the proportion of cases where a score value is exceeded and are constructed after pooling the scores for all grid cells and seasons.

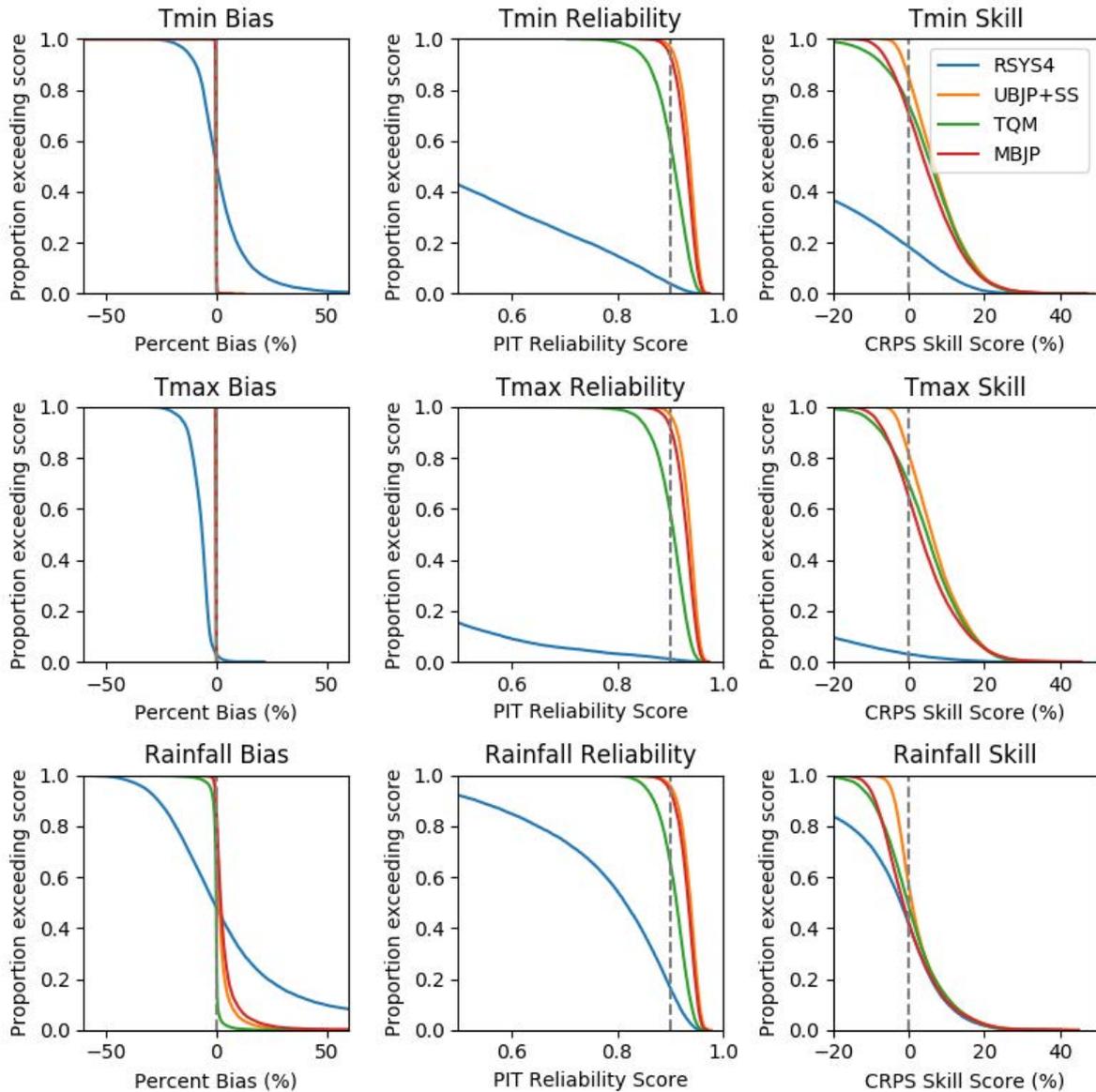


Figure 2.2: Plots comparing the overall performance of the various sets of forecasts (raw and post-processed) as the proportion of grid cells where certain bias, reliability and skill score values are exceeded. Columns are for the different metrics and rows are for the different climate variables.

Regarding bias (Figure 2.2, left column), RSYS4 forecasts are (as expected) biased for all three climate variables: Tmin, Tmax and rainfall. RSYS4 Tmax forecasts have a propensity to be negatively-biased, although the bias magnitude is normally less than 10%. RSYS4 Tmin forecasts can be either positively- or negatively-biased with magnitudes greater than 10% in approximately 30% of cases.

RSYS4 rainfall forecasts are biased positively and negatively in approximately equal measure with magnitudes exceeding 25% not uncommon.

Post-processing substantially reduces PBIAS for all three climate variables. For Tmin and Tmax, bias is reduced to near-zero regardless of the post-processing method. For rainfall, some biases remain after post-processing with UBJP+SS and MBJP, which is mainly a problem in very dry grid cells where small absolute biases manifest as a large percentage bias; further discussion is given in section 2.5.3. For UBJP+SS and MBJP, the median bias for rainfall is around 2-3%, although it can exceed 10%; MBJP performing slightly worse for bias correcting rainfall than UBJP+SS. TQM effectively reduces the bias to near-zero in nearly all rainfall cases.

Regarding reliability (Figure 2.2, middle column), a grey, dashed vertical line is plotted at $REL_{PIT} = 0.9$ as a guiding threshold for highly reliable forecasts. On PIT uniform probability plot (e.g. Renard et al. 2010; Wang et al. 2009), the points would line up closely along the 1:1 line. RSYS4 forecasts of all three climate variables are frequently unreliable, which is in accordance with the observed biases.

Post-processing substantially improves the reliability of the forecasts by reducing bias and improving ensemble spread. The UBJP+SS and MBJP forecasts are almost always highly reliable. TQM forecasts are also frequently highly-reliable, although they are overall less reliable than the BJP forecasts.

Regarding skill (Figure 2.2, right column), a grey, dashed line is plotted at a CRPS skill score value of 0.0 to indicate the skill of the climatology reference forecasts. Skill is positive for the post-processed forecasts in the majority of cases; however, Tmin and Tmax forecasts are overall more skilful than rainfall forecasts. Out of the different post-processing models, UBJP+SS produces the most skilful forecasts with the median CRPS skill score being higher than every other model for every climate variable, even if only by a small margin. UBJP+SS skill scores are rarely negative and when they are, they are not worse than about -5 to -10%, which can be attributable to cross-validation effects. The MBJP model produces forecasts that are overall less skilful than UBJP+SS and occasionally negative

to about -20%, suggesting overfitting may occur; further investigation is given in section 2.5.3. TQM skill is overall better MBJP but worse than UBJP+SS; TQM is sometimes seen to produce skill scores that are considerably negative, particularly for Tmin; however, unlike with MBJP, overfitting is unlikely to be the problem. More likely, it is the inability of TQM to return negatively-skilful forecasts to climatology.

2.5.2. Overall performance of multivariate forecasts

Geographical maps of the energy score (ES) skill scores for the multivariate (Tmin, Tmax, rainfall) forecasts are shown for each season and for each post-processing method in Figure 2.3, Figure 2.4 and Figure 2.5, respectively. Maps of the variogram score (VS) skill scores for each season are shown for the UBJP+SS, MBJP and TQM post-processing methods in Figure 2.6, Figure 2.7 and Figure 2.8, respectively. Summaries of the ES and VS skill scores for all grid cells are shown in Figure 2.9.

The ES has not been widely used to make inter-variable comparisons. As a first check for the instructiveness of the ES skill score in this setting, I visually compare the ES and CRPS skill score maps (not shown), and I confirm that features of CRPS skill maps for individual variables are noticeable in the ES skill maps and that a sensible conjugation occurs. For example, for UBJP+SS forecasts, Tmin and Tmax CRPS skill scores are moderately positive across northern Australia, whereas rainfall CRPS skill scores are neutral. The corresponding ES skill scores are weakly-to-moderately positive. As a second example, for TQM forecasts, all three variables have neutral skill in the southeast of the Australian mainland, a result that translates into the corresponding ES skill score maps.

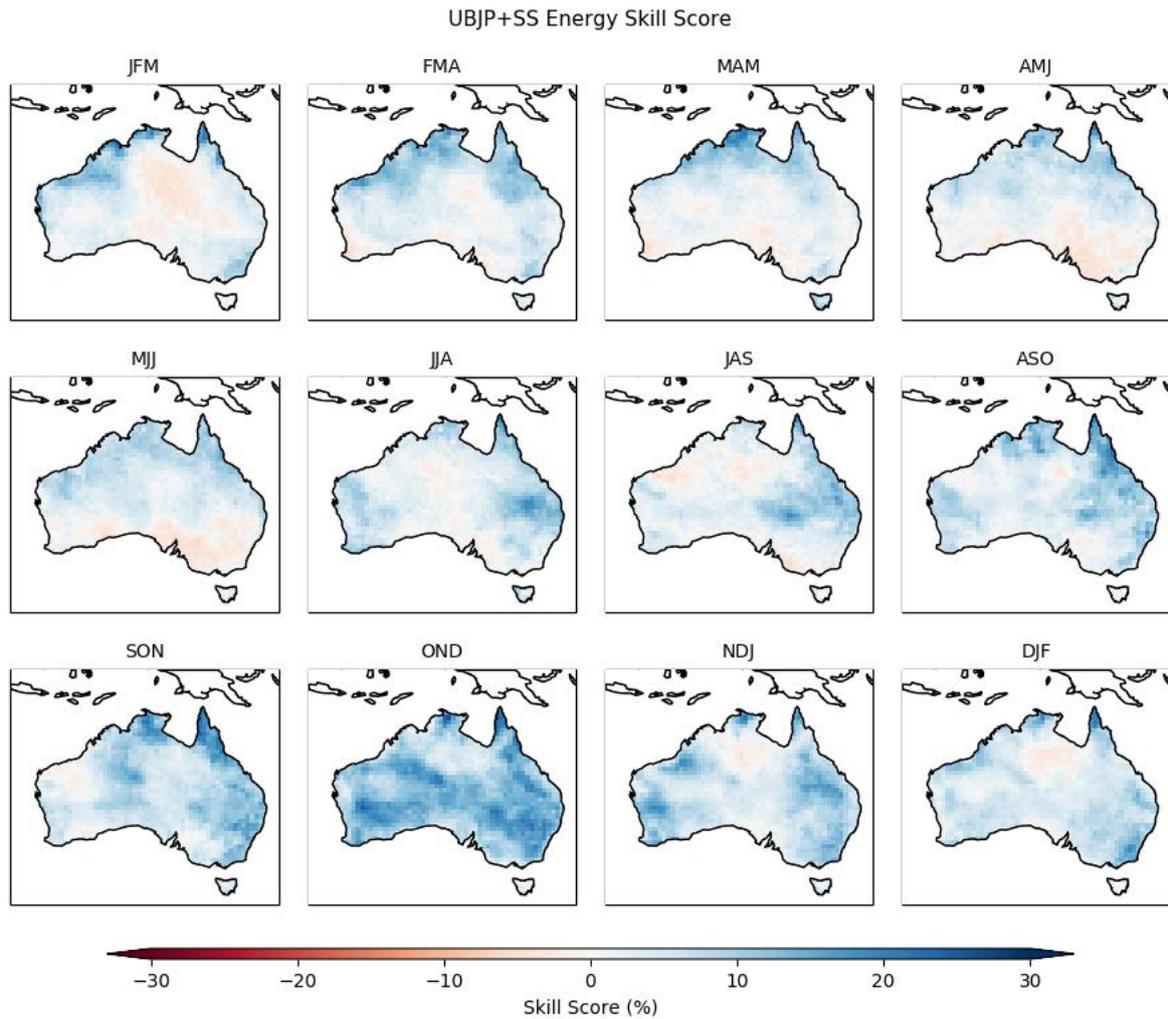


Figure 2.3: Maps of Energy Skill Scores for UBJP+SS forecasts for the period 1981–2016. The skill scores are calculated using historical observations as climatological reference forecast and using leave-one-year-out cross-validation. Positive skill means lower error in the UBJP+SS forecasts compared to the reference. The skill is mapped for each target season for forecasts issued with one-month-lead-time.

Overall, ES skill scores are low (<20%), which is understandable given the well-known low–moderate skill of seasonal forecasts, especially with one-month lead time. Moreover, forecasts of Tmin, Tmax and rainfall are not always similarly skilful across regions and seasons, and ES skill scores will be modulated accordingly. The results for the ES skill scores conform to the findings for the single-variable evaluation. That is, UBJP+SS produces the overall most skilful forecasts. TQM and MBJP forecasts have lower overall skill, albeit with a similar pattern to UBJP+SS.

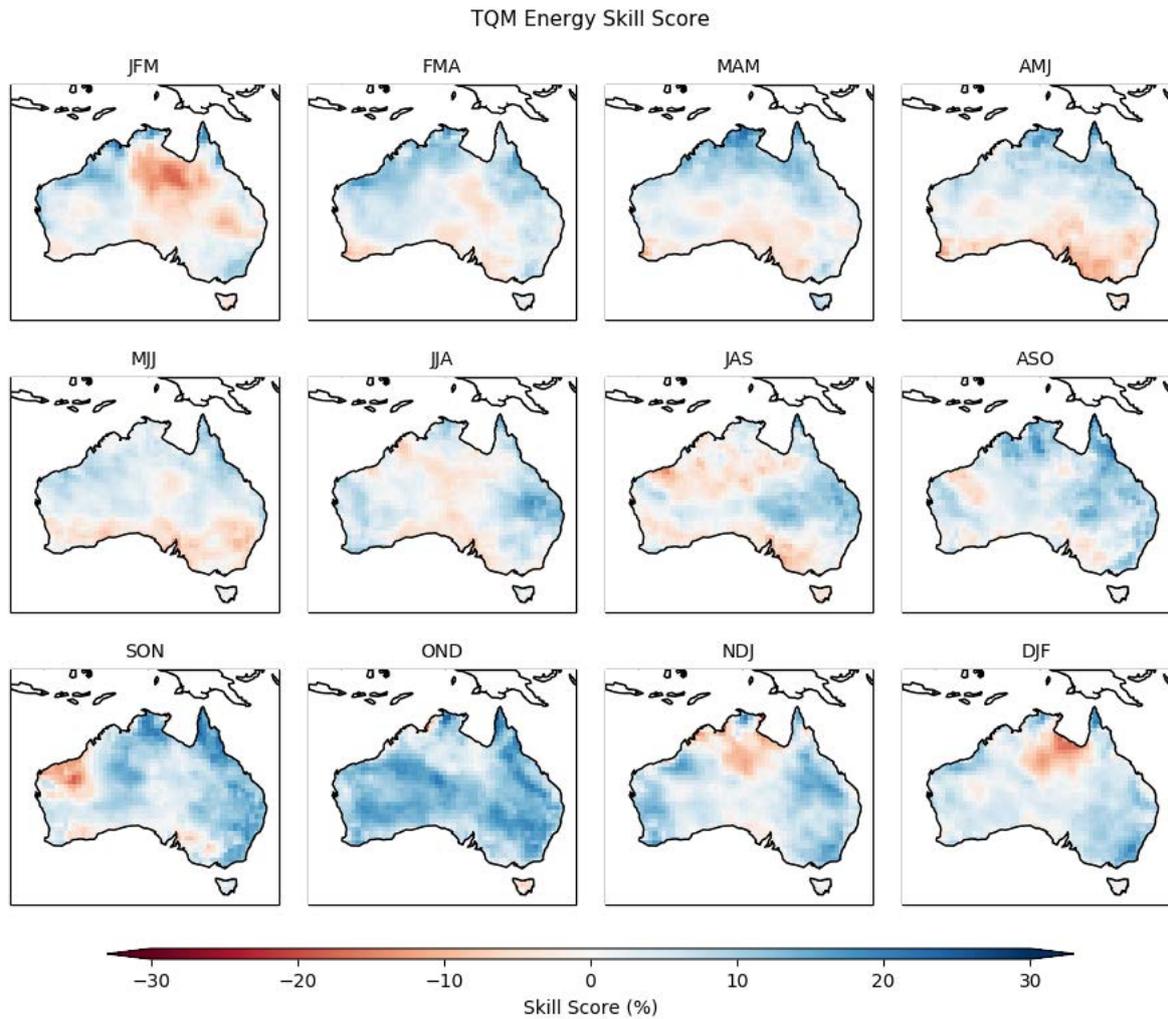


Figure 2.4: As for Figure 2.3, except for TQM forecasts

The maps for the VS skill scores give some unique insights. Overall the VS skill scores are lower than the ES skill scores and are more frequently negative. I interpret the VS skill score maps as highlighting areas where there are remaining weaknesses in the inter-variable dependence structure in the forecasts. For TQM, the inter-variable relationships are largely inherited from the raw model output, and, therefore, it is expected that regions and seasons will have problems with inter-variable correlations. Indeed, negative VS skill is observed for TQM forecasts in various regions across all seasons. However, I expect that either direct modelling of inter-variable relationships in MBJP or ensemble reordering in UBJP+SS will mean that inter-variable correlations are appropriate. However,

the results indicate that there are some deficiencies with both BJP approaches that require further exploration (see section 2.5.3)

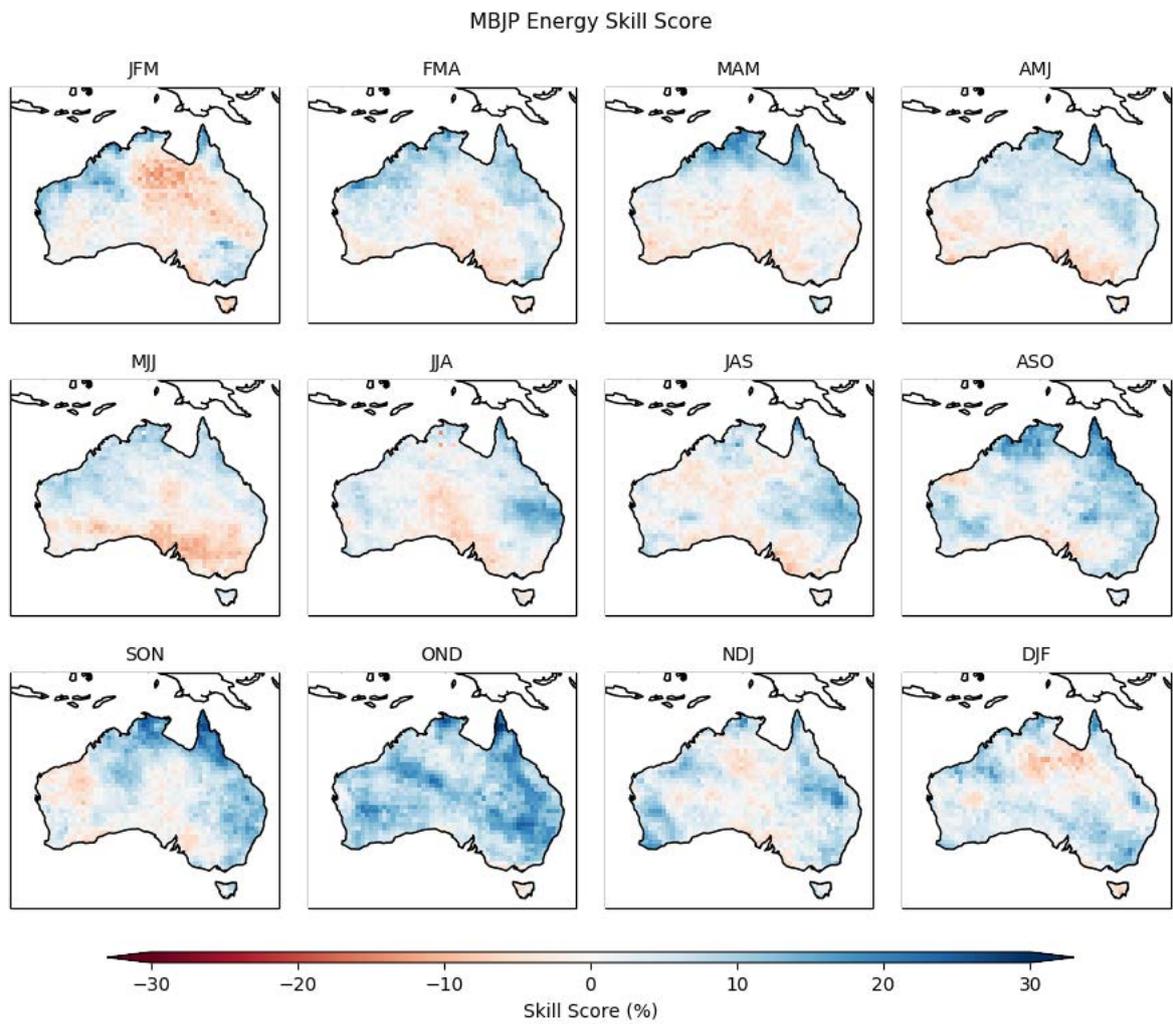


Figure 2.5: As for Figure 2.3, except for MBJP forecasts

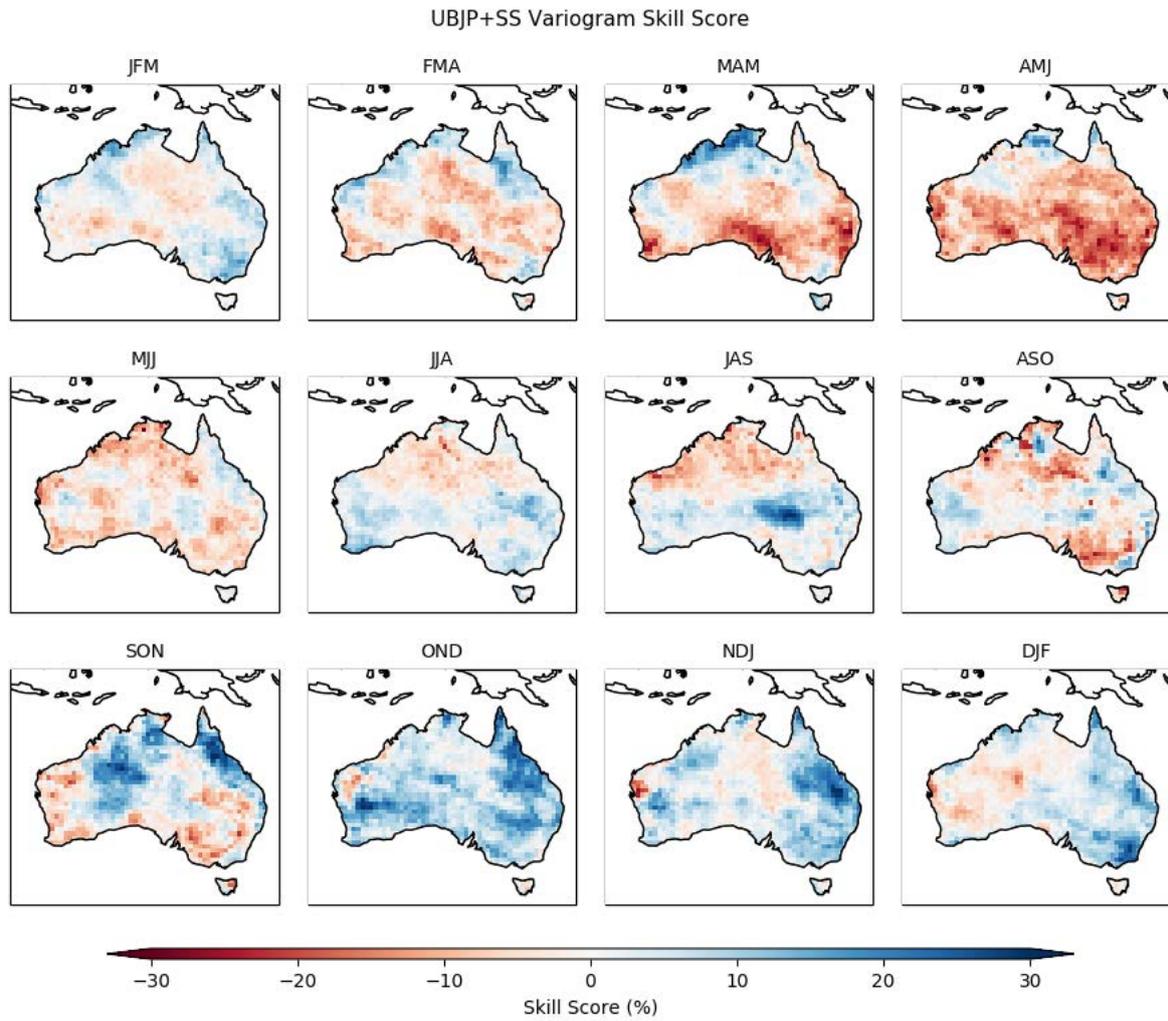


Figure 2.6: Maps of Variogram Skill Scores for UBJP+SS forecasts for the period 1981–2016. The skill scores are calculated using historical observations as climatological reference forecast and using leave-one-year-out cross-validation. Positive skill means lower error in the UBJP+SS forecasts compared to the reference. The skill is mapped for each target season for forecasts issued with one-month-lead-time.

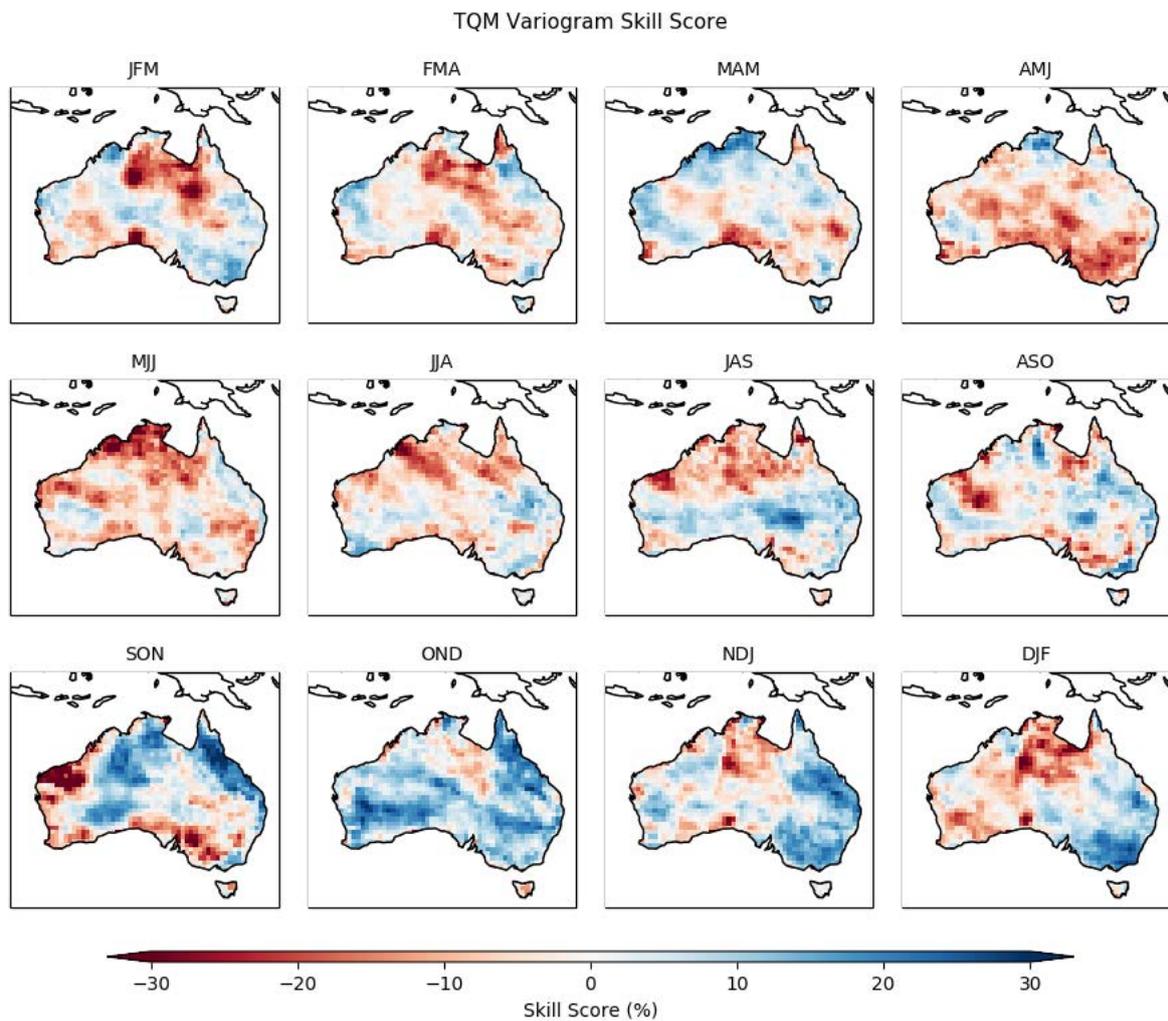


Figure 2.7: As for Figure 2.6, except for TQM forecasts

ES and VS skill score summaries are produced by plotting the proportion of cases where a range of skill score thresholds are exceeded, for all post-processing models (Figure 2.9). The results for all grid cells are pooled for the overall analysis. The skill score summaries support the impression given by comparing the previous skill score maps. The UBJP+SS forecasts exhibit the best overall performance in terms of the energy score, particularly by having fewer low or negative skill scores. MBJP and TQM perform similarly in terms of the energy score, although MBJP has marginally better performance in terms of filtering out negative skill. In terms of the variogram score, the performance of MBJP and UBJP+SS is similar with TQM performing overall worse. The results for the variogram score suggest

that the calibration methods that model or enforce observed correlation structures perform better overall.

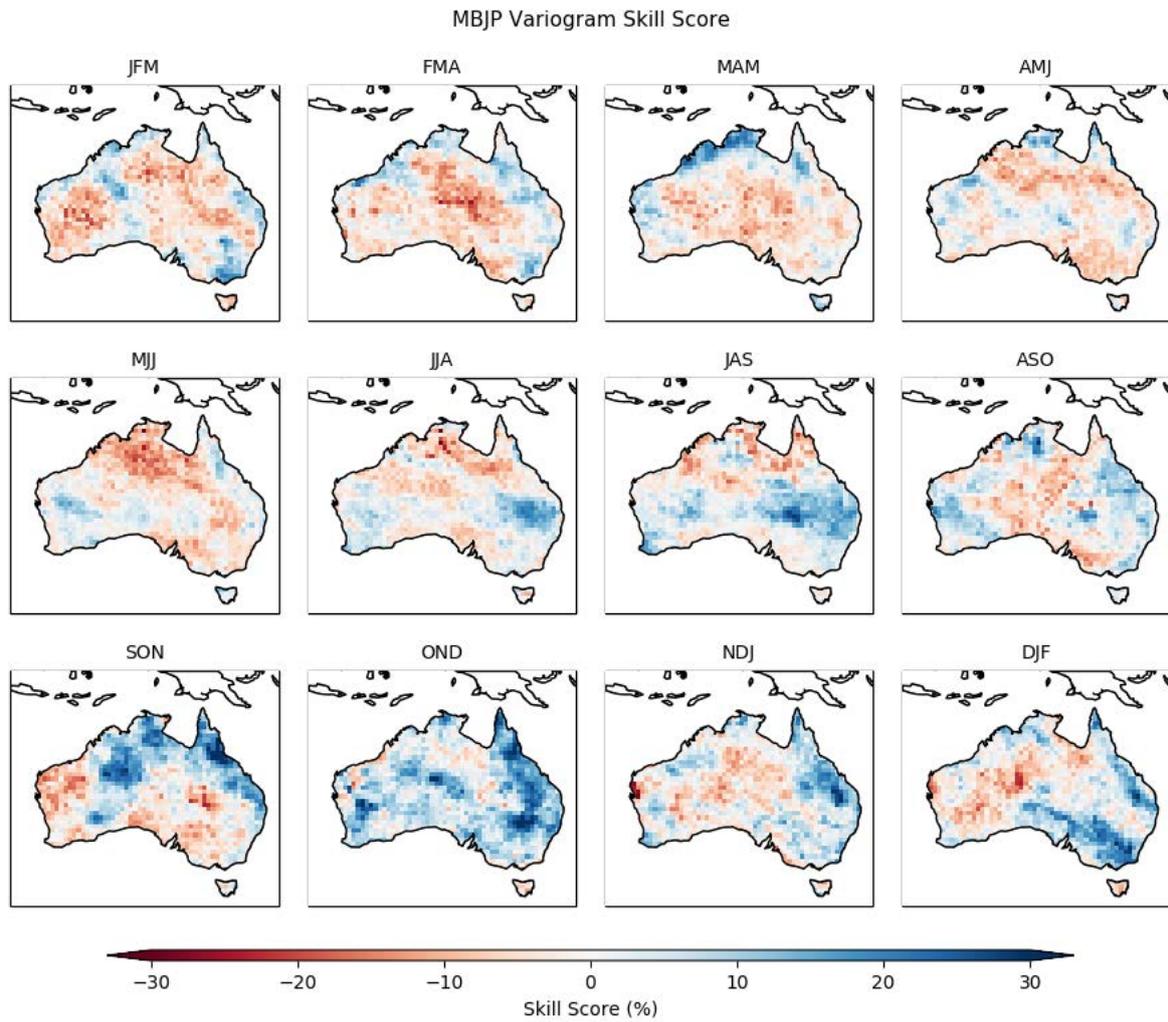


Figure 2.8: As for Figure 2.6, except for MBJP forecasts

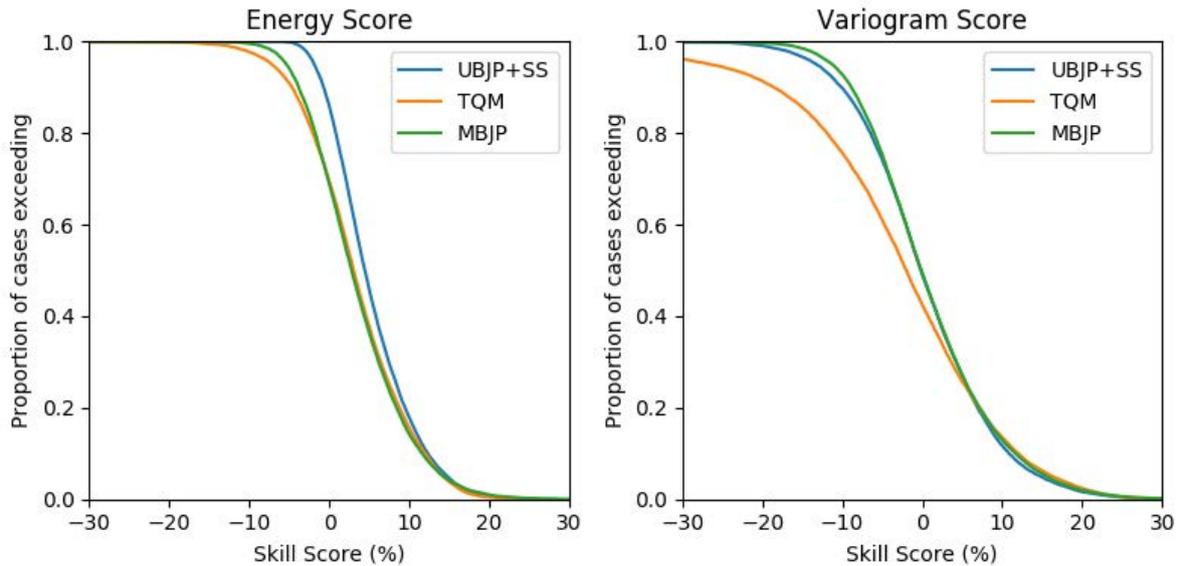


Figure 2.9: Summary of multivariate forecast performance across all grid cells and seasons, and a comparison of the results for various post-processing methods. The curves plot the proportion of cases where ES and VS skill score values are exceeded. The multivariate skill scores consider all three climate variables (Tmin, Tmax and rainfall) in their calculation. The VS is more sensitive to the calibration of the dependencies between the variables.

2.5.3. Diagnosing factors that affect performance

The worse overall performance of MBJP relative to UBJP+SS could be surprising, except that the evaluation is being done within a cross-validation framework and MBJP has more parameters (see section 2.3.3); therefore, overfitting is a real risk. To test whether overfitting is indeed the main problem causing lower performance of MBJP forecasts, I test the effect of removing cross-validation on the results.

The ES and VS skill score summaries for all grid cells are reproduced after redoing the experiments without cross-validation applied (Figure 2.10). I refer to these results as in-sample results whereas the main results are out-of-sample. It is clear that UBJP+SS and MBJP provide increasingly better in-sample predictive performance, while the TQM results are largely unchanged. While the in-sample predictive performance is boosted by artificial skill, the results hint that more sophisticated calibration approaches could be beneficial where sufficient data exists. However, it appears in the

current study that there is insufficient data to robustly infer the MBJP model parameters and realise a predictive performance benefit over UBJP+SS and TQM for calibrating independent (out-of-sample) forecasts.

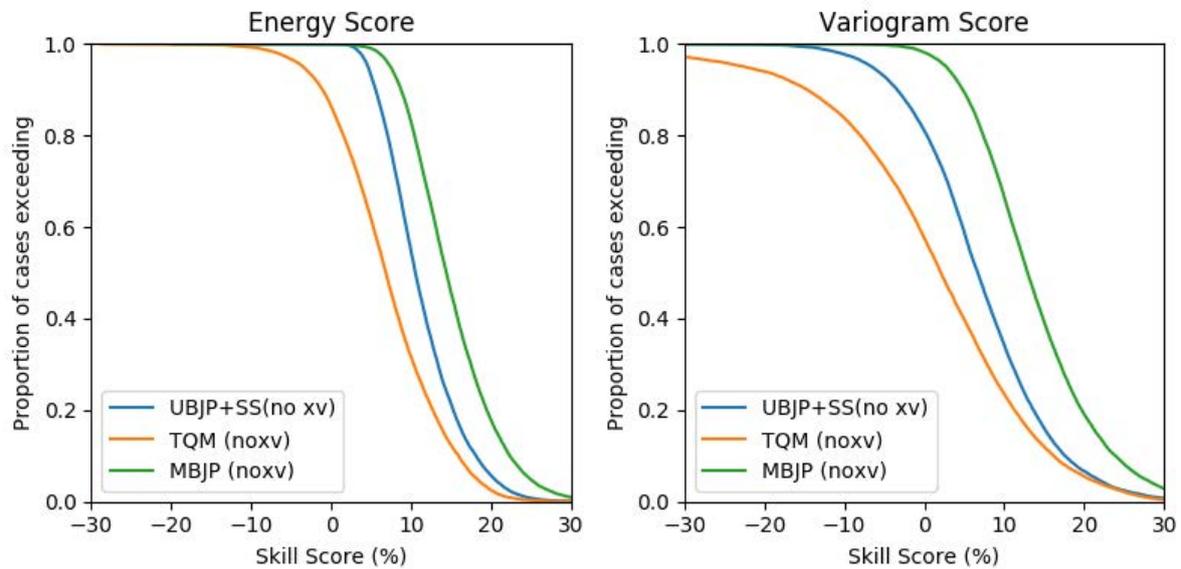


Figure 2.10: As for Figure 2.9, except that leave-one-year-out cross-validation has *not* been applied.

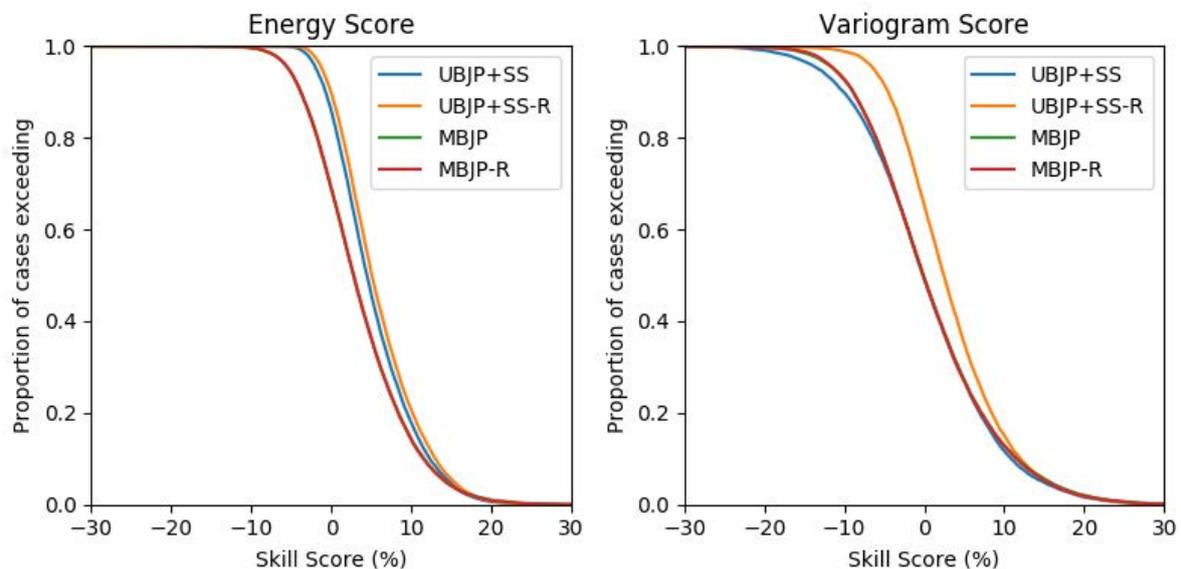


Figure 2.11: Comparison of multivariate skill scores for UBJP+SS and MBJP forecasts before and after reshuffling with samples generated from a BJP model fitted jointly to *observed* data. The -R suffix indicates reshuffled forecasts. Reshuffling improves the overall skill of UBJP+SS forecasts.

The VS skill maps for UBJP+SS show widespread negative skill in MAM and AMJ. The only real plausible explanation is that the Schaake Shuffle did not restore an inter-variable correlation structure that is consistent with the observations. The sliding window that was applied to obtain enough historical data samples to apply the Schaake Shuffle may be problematic for those seasons. To further diagnose the problem with UBJP+SS, I develop a BJP model on observed data, only, for all three climate variables (i.e. with no GCM input), and evaluate the VS skill scores of generated samples. For clarity in what follows, I refer to this BJP model as MBJP-OBS. Leave-one-year-out cross-validation was applied in the same way as for forecast post-processing. It is found that the mean VS skill score for MBJP-OBS is very close to zero when verified against observations across almost all grid cells and seasons (Figure 2.12). Therefore, BJP has no problem generating samples with realistic inter-variable correlations *per se*. Subsequently, I re-apply the Schaake Shuffle to the UBJP+SS forecasts, except I use the MBJP-OBS ensemble members in place of observations. The reshuffled UBJP+SS forecasts are called UBJP+SS-R. For the sake of understanding, I also reshuffle MBJP forecasts using MBJP-OBS samples. The reshuffled MBJP forecasts are called MBJP-R.

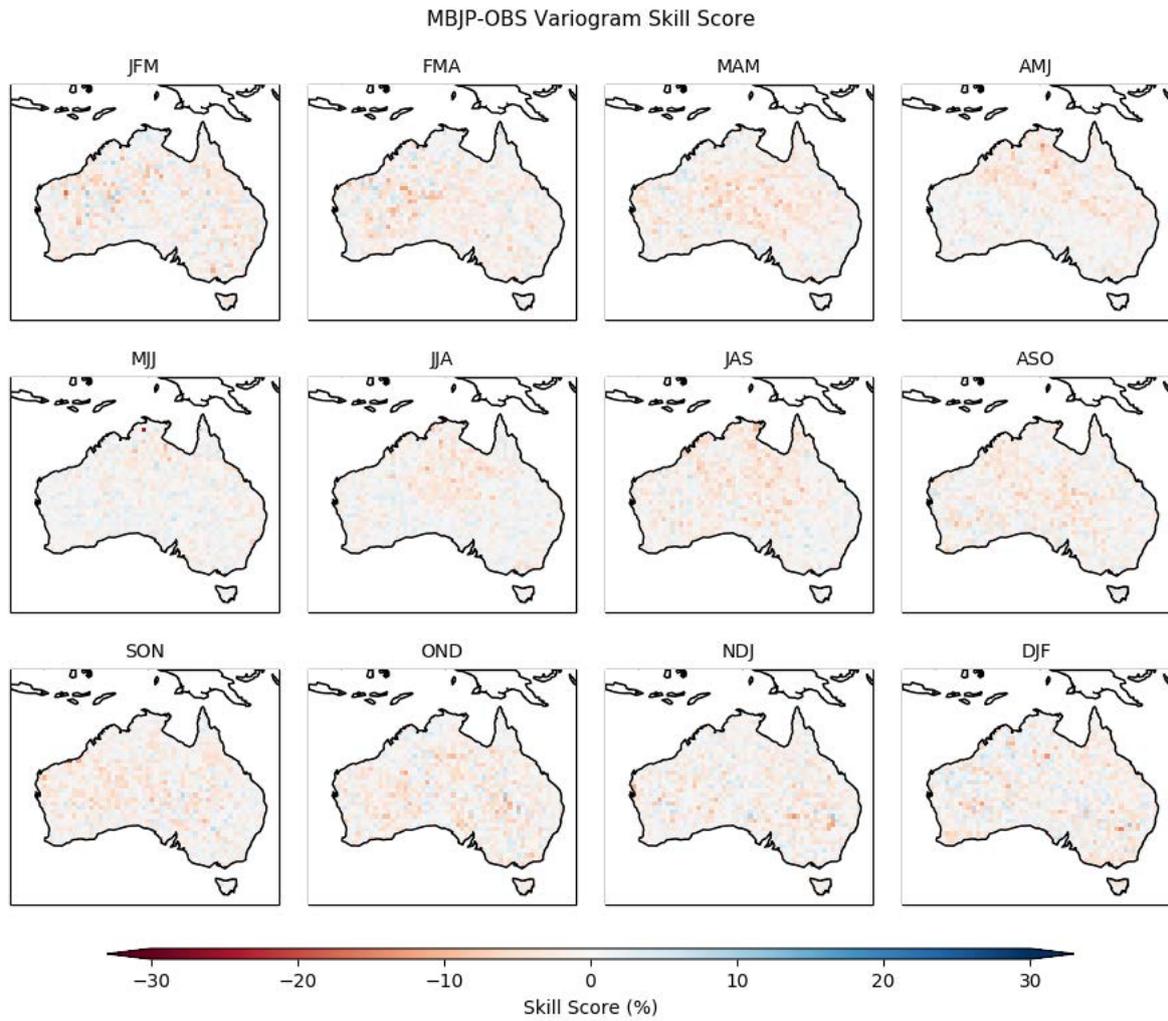


Figure 2.12: Variogram skill scores for MBJP-OBS, a BJP model fitted to observed data (no GCM predictors are involved). The skill scores are calculated using historical observations as the reference forecast and using leave-one-year-out cross-validation. The skill scores being predominantly close to 0 indicates very similar skill between model-fitted and pure-observation climatologies.

The ES and VS skill score summaries for all grid cells are presented for the original UBJP+SS and MBJP forecasts and the reshuffled UBJP+SS-R and MBJP-R forecasts in Figure 2.11. Reshuffling the UBJP+SS forecasts has little impact on the energy skill scores, which conforms with the understanding that the energy score is relatively insensitive to the specification of the inter-variable correlations. In contrast, the variogram skill scores are much improved, which suggests the inter-variable correlation structure borrowed from the MBJP-OBS samples is more realistic than original inter-variable correlations instilled by the Schaake Shuffle. Further examination of the spatial and seasonal distribution of the improvements (Figure 2.13) is undertaken by plotting variogram skill scores with UBJP+SS-R as the test model and UBJP+SS as the reference model. Widespread improvements are seen in MAM, AMJ and ASO in particular: the same regions where most of the negative skill is observed in Figure 2.6. The forecasts are not made worse by reshuffling. In contrast to the results for reshuffling UBJP+SS forecasts, reshuffling MBJP forecasts appears to have little effect.

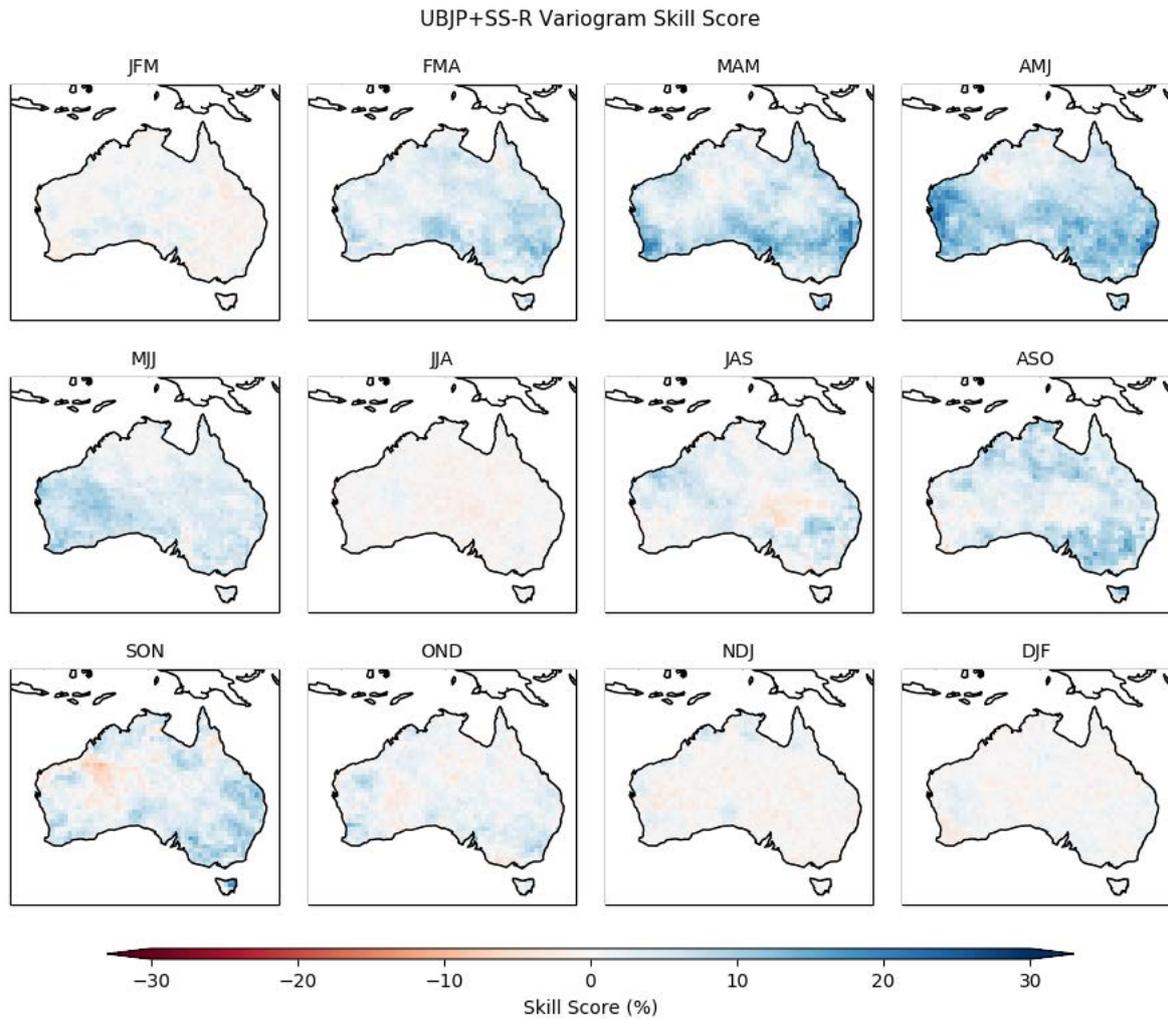


Figure 2.13: Variogram skill scores for UBJP+SS forecasts after reshuffling with samples generated from the MBJP-OBS model fitted to observed data (see Figure 12). The $-R$ suffix indicates reshuffled forecasts. The skill scores are calculated using UBJP+SS forecasts as the reference and using leave-one-year-out cross-validation. Positive skill means better VS in the UBJP+SS-R forecasts compared to the reference. The skill is mapped for each target season for forecasts issued with one-month-lead-time. The neutral-positive skill suggests the reshuffling is beneficial.

Figure 2.2 shows that positive biases in the range of 5-10% can sometimes arise in UBJP+SS and MBJP post-processed rainfall forecasts. Tmin and Tmax forecasts are unaffected. Mapping of the seasonal and spatial distribution of the biases in BJP-based forecasts (Figure 2.14) reveals that these biases are by-and-large contained to very dry grid cells, particularly in northern Australia during the seasons MJJ–JAS when monthly rainfall totals are mostly near-zero. In such cases, a small absolute bias can manifest as a large percentage bias. Moreover, BJP adds parameter uncertainty, which I

suspect can lead to some extreme values being generated in the back-transformation procedure, causing noticeably higher means in very dry grid cells. Although not shown in the main results, I find that BJP models fitted to observed data (i.e. MBJP-OBS) generate samples with same biases, so it is not strictly a problem related to the calibration of GCM forecasts. Further research, outside the scope of this study, is needed to improve BJP modelling and the coupling of transformations and BJP in very dry grid cells where distributions can be highly skewed.

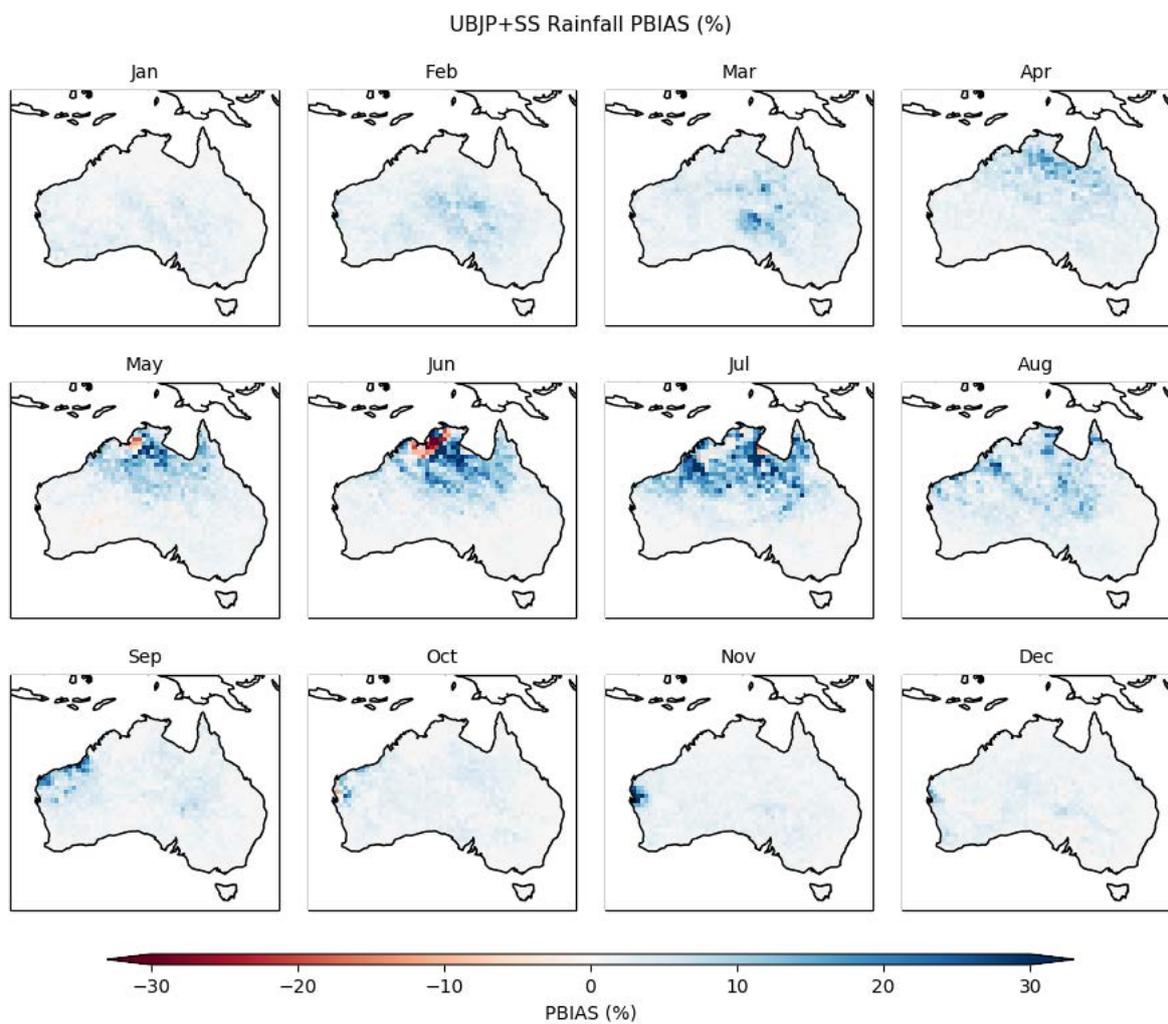


Figure 2.14: PBIAS (%) in UBJP+SS rainfall forecasts for the period 1981–2016. PBIAS departs from zero in very dry areas where the magnitude of the absolute bias is likely to be small.

2.5.4. Extension opportunities

In this study, I only considered post-processing of variables at the local scale. An alternative approach that remains untested, which may add skill while reducing overfitting, is to set up single predictor – multiple predictands models where the predictor represents a relevant large-scale climate feature (i.e. an ENSO climate index). Furthermore, multiple forecasts may be combined using Bayesian model averaging or another combination method to improve skill in different regions and seasons (e.g. Schepen et al. 2014; Wang et al. 2012a).

The results show that flexible modelling of Tmin, Tmax and rainfall marginal distributions permits multivariate post-processing using joint probability models and alternative implementations of extant methods like quantile-mapping. While I used the flexible Yeo-Johnson transformation and hydrologically-specific Log-sinh transformation, any appropriate normalising transformation could be substituted into the workflows (e.g. a Box-Cox transformation). I expect that the strategies employed here could be tested more widely, including to other variables including pressure, wind speed, solar radiation and evaporation, enabling a broader understanding of multivariate forecasting skill in applications beyond agriculture and natural resources management, including in the energy, mining and insurance industries.

In section 2.5.3, it was found that the choice of the unconditional Schaake Shuffle using a window of starting dates led to subpar forecast performance in terms of the variogram score; which can be related to the misspecification of inter-variable correlations. Scheuerer et al. (2017) detected improved results after applying a variation of the Schaake Shuffle in which the dependence template was constructed by the preferential selection of dates; such that the chosen sequences were more representative of the forecast distribution. Such a method could improve the results of UBJP+SS in certain seasons. [As an aside, Scheuerer et al. (2017) also remarked on the enhanced possibility of variogram skill scores being negative compared to the energy score due to it offering less reward for correctly predicting magnitude, a feature that I see in these results.] Other studies have highlighted

the partial ineffectiveness of the Schaake Shuffle (Verkade et al. 2013) or proposed selective variants that yield improvements. For example, Bellier et al. (2017) evaluated analogue-based methods for selecting Schaake Shuffle dates and found it outperformed the unconditional Schaake Shuffle for short-term rainfall forecasts, especially in impact on subsequent streamflow forecasts. Wu et al. (2018) point out how ties in data ranks can impact on the effectiveness of rank reordering schemes, which will be pertinent in daily or sub-daily studies, however, I expect it would only have a very minor impact in this seasonal study (e.g. multiple zeros in rainfall records may occur in exceptionally dry areas). Evidence is building around the shortcomings in ensemble reordering methods and thus further work is needed to identify the most efficient and effective options to use these to restore multivariate dependence structures.

Overall, the results in this study point to plenty of challenges to address in integrating robust low-dimensional post-processing approaches in high-dimensional application domains (e.g. multiple variables, sub-catchments, lead-times and so forth). Figure 2.11 demonstrates that the best result in this study was actually achieved by a hybrid method of calibrating each variable individually and then shuffling the ensemble members based on a joint probability model of the observations. I don't consider this the most practical option to pursue. Rather, there may be gains made by establishing models of covariance that require fewer parameters, particularly in combination with other dimension reduction techniques. For the foreseeable future, both parametric calibration and empirical ensemble reordering methods are going to play a role in seasonal forecast post-processing, while much more research is needed to find balanced solutions that improve multivariate forecasting skill for independent predictions.

In this study, I have addressed only seasonal (three-month) forecasts. However, many operational models that could receive climate forecast information, e.g. hydrological and agricultural models, require data at daily time steps and at sub-grid locations. More research is needed to spatially and temporally downscale multivariate seasonal climate forecasts.

2.5.5. Conclusions

GCM forecasts are increasingly in demand to support the expansion of digital agriculture and other natural resource management initiatives, which require coherent multivariate seasonal climate forecasts. Raw GCM forecasts are readily available, however, they require calibration to remove biases and reliably quantify forecast uncertainty. While multivariate post-processing has been considered previously in the very specific problem of short-term temperature and wind-speed forecasting, very little attention has been paid to the multivariate calibration of seasonal GCM outputs. Usually, any bias-correction or calibration in seasonal forecasting is done on variables independently. In this study, I develop and test three strategies for calibrating multivariate forecasts of Tmin, Tmax and rainfall, finding each approach has unique strengths and weaknesses.

UBJP+SS applies a univariate BJP calibration to each variable and subsequently establishes the inter-variable correlation structure from observations using the Schaake Shuffle. The UBJP+SS approach performs best in terms of univariate skill and reliability scores and multivariate skill scores. This is despite the unconditional sampling of historical trajectories for the Schaake Shuffle being suboptimal in some instances.

MBJP simultaneously calibrates each variable by modelling the full joint distribution of all relevant predictor and predictand variables. In in-sample testing MBJP presents itself as the far superior approach; however, in cross-validation with out-of-sample testing, MBJP generally performs worse than UBJP+SS, apparently due to the lack of sufficient data to robustly infer the more numerous model parameters. That said, MBJP may remain feasible for problems with more data available.

TQM is a quantile-mapping approach that uses the same marginal transformations as BJP. I find that while it offers substantial improvements over raw forecasts and has fewer parameters, its fundamental weakness of not modelling correlations between forecasts and observations or between variables means that it performs overall the worst in terms of univariate and multivariate verification metrics.

Continued research efforts are likely to optimise the calibration of seasonal forecasts for complex application domains requiring multivariate climate inputs. I suggest that further research should be investigated in robust modelling of covariances, dimension reduction techniques and resolution of emerging challenges in ensemble reordering techniques (including handling ties and more efficient construction of conditional dependence templates).

3. Spatial and temporal disaggregation of climate forecasts

3.1. Preamble

In Chapter 2, I compared several strategies for calibrating multivariate seasonal climate forecasts. The best-performing strategy was to first calibrate each variable in a univariate setting and then to restore inter-variable relationships from a historical data template using the Schaake Shuffle, an empirical method for ensemble reordering.

The next challenge is to generate multi-site, daily forecasts as needed by crop models. There are two general approaches that can be taken here. One is to calibrate daily forecasts directly. Another is to make use of forecasts calibrated at coarse resolution and to disaggregate them to obtain higher-resolution forecasts that have the right spatial, temporal and inter-variable correlation structures.

Accordingly, I address Objective 2 in this chapter, which is to “Develop and evaluate methods for downscaling forecasts to high spatial and temporal resolution as needed by crop models”. To establish the best way forward, I examined what has been done previously to provide daily weather inputs for agricultural models. Initially, weather generators, simple mathematical models for generating synthetic time series of weather quantities, were appealing as a means to produce daily values. However, initial development efforts indicated the weather-generator solution would become intractable in the time available due to difficulties simulating coherent forecasts of multiple variables at multiple sites. I, therefore, sought a more pragmatic solution that is novel yet effective.

Previous studies have coupled BJP-calibrated monthly rainfall forecasts with a hydrological model and subsequently generated skilful and reliable ensemble streamflow forecasts. I start out following a similar approach here and introduce new components to generate daily, multivariate meteorological forecasts at multiple sites. The key objective is to preserve the distribution of forecasts at a coarse resolution and to provide ensembles with the correct statistics at fine

resolutions. I run a regional-scale experiment in the Burdekin sugarcane region to rigorously test the downscaling method. The findings of this chapter are directly applied next in Chapter 4, which produces sugarcane crop yield forecasts using a crop model that runs on a daily time step.

The contents of this chapter have been formatted as a journal article and submitted to The International Journal of Climatology (Impact Factor 3.609) with the title “Coupling forecast calibration and data-driven downscaling for generating reliable, high-resolution, multivariate seasonal climate forecast ensembles at multiple sites” and with authorship: Schepen, A., Y. Everingham, and Q.J. Wang. Although the paper is co-authored by my supervisors, I confirm the work is essentially my own. I developed the methodology, conducted all of the experiments, analysed the results and wrote all of the paper, including preparing all of the figures. Everingham and Wang helped form the research questions and provided editorial support.

Additionally, in parallel to the work presented in this chapter, I led a more collaborative project to investigate the alternative approach of calibrating daily forecasts directly, albeit focusing on rainfall forecasts only. The results of this study are published in Hydrology and Earth System Sciences (Impact Factor 4.426) with the citation:

Schepen, A., Zhao, T., Wang, Q. J., & Robertson, D. E. (2018). A Bayesian modelling method for post-processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments. *Hydrology and Earth System Sciences*, 22(2), 1615-1628.

The published paper is attached as Appendix B.

3.2. Introduction

Seasonal forecasts have high potential value to agricultural and water resources management industries (Feldman and Ingram 2009; Hansen 2005; Meza et al. 2008). Long-range ensemble climate forecasts extending six or more months ahead are now routinely generated by scientific and governmental agencies using global climate models (GCMs). Publicly available forecasts are normally

only issued at coarse spatial and temporal scales, which are sufficient for casual followers of weather and climate. However, to support proactive decision-making in water resources management and agriculture, GCM forecasts need to be translated to the decision maker's domain through various post-processing techniques, including bias-correction, calibration and/or downscaling (e.g. Manzanas et al. 2018; Maraun 2013; Schepen and Wang 2014; Tian et al. 2014; Wood et al. 2002; Zhao et al. 2017).

Translation to the application domain can involve both spatial and temporal downscaling. For justification, consider crop models, which are usually established for sub-grid (e.g., farm) locations and whose response can be quite sensitive to the distribution of weather within a season (Brown et al. 2018; Hansen 2005). In some applications, downscaled forecasts are required for locations/catchments spanning several grid cells. For example, Potgieter et al. (2005a) developed a regional-scale sorghum forecasting system that uses climate analogue inputs at multiple locations. Regional-scale systems can support broad outlooks for industry (e.g. for crop insurance risk), but they do require some care in their development, as Hansen and Jones (2000) point out, because when crop response is modelled at multiple locations, aggregates can be biased if spatial heterogeneity is not adequately modelled. Consequently, for GCM forecasts to provide inputs to both local and regional-scale systems, it is critical to develop spatial and temporal downscaling methods that contain the meaningful forecast information from the climate model while preserving the statistical properties of the relevant historical daily sequences (Hansen and Jones 2000). Arguably then, GCM forecasts should be calibrated at an appropriate broad scale of interest before being spatially and temporally downscaled. Calibration of forecasts typically involves using a statistical model to adjust raw forecasts so that they have little bias, are reliable, and are generally more skilful than climatology (Schepen et al. 2016; Zhao et al. 2017)

Coarse-scale GCM adjustment followed by downscaling is a methodology often used in water resources assessments and forecasting, which may not be surprising, since regional-scale or

distributed hydrological modelling requires correctly allocating basin-wide rainfall to sub-catchments to realistically simulate flows. In line with this reasoning, Wood et al. (2002) developed a method to bias-correct coarsely-gridded, monthly climate model outputs of rainfall and temperature, before spatially and temporally downscaling them for use in the VIC semi-distributed hydrological model. The approach has since become widely known as BCSD (bias-correction spatial-disaggregation). Gutmann et al. (2014) reviewed various downscaling methods, albeit for downscaling reanalysis rainfall, and found BCSD still amongst the best performers, especially for the reproduction of extremes and wet day fractions (the proportion of days on which rain is observed). In BCSD, the spatial downscaling is a simple interpolation scheme. The temporal downscaling uses randomly or semi-randomly resampled historical data, e.g., conditioned on wet and dry periods, as a template to derive daily sequences. This data-driven approach to downscaling is appealing because the daily sequences are consistent with both the large-scale forecasts and the local scale observations. Several advances upon BCSD have been proposed over the years, albeit mostly in the context of climate impacts studies. Several studies have proposed alternatives to simple interpolation for spatial downscaling. One suggestion is to match the bias-corrected fields to stochastically-generated, spatially-correlated samples in BCSA (bias-correction spatial-analogues) (Hwang and Graham 2013); another is to use multivariate analogues to downscale multiple variables simultaneously from large scale patterns in MACA (multivariate adaptive constructed analogues) (Abatzoglou and Brown 2012). From a forecasting viewpoint, a major drawback of all the discussed methods (BCSD, SDBC, BCSA and MACA) is that they use quantile-mapping for bias-correction, which is known to be deficient as forecast calibration tool (Brown et al. 2018; Ines and Hansen 2006; Ines et al. 2011; Zhao et al. 2017 and Chapter 2 of this thesis) and as a downscaling tool (Maraun 2013); largely because it assumes a perfect correspondence between forecasts and observations, regardless of scale. The prevalence of quantile-mapping in downscaling has grown out from climate impact studies, where it can be effective as a bias correction tool. Forecasting, however, is quite a different problem with high uncertainty and the added dimension of lead time. Indeed, studies making use of quantile-mapped

GCM forecasts in Australia have found that they perform poorly in agricultural applications (Brown et al. 2018; Western et al. 2018), presumably due to misspecification of temporal, spatial and inter-variable relationships; echoing earlier findings of Ines and Hansen (2006).

As to the complexity of post-processing approaches in active use, simple linear scaling is sometimes seen (e.g. Bazile et al. 2017; Crochemore et al. 2016). Quantile mapping remains popular, despite the aforementioned weaknesses (e.g. Brown et al. 2018; Crochemore et al. 2016; Wetterhall et al. 2015), presumably because it is easy to apply and, under the right conditions, it can yield substantial improvements over raw forecasts. Other studies have investigated stochastic weather generation. After Ines and Hansen (2006) found that quantile-mapping was unable to correct fundamental problems with their GCM's rainfall temporal correlation structure, Ines et al. (2011) developed a weather generator to redistribute monthly rainfall across days in a more realistic fashion (quantile-mapping was still used initially). For each ensemble member, the weather generator was used to repeatedly generate daily sequences until the monthly rainfall was within 5% of the target, after which a linear adjustment was made to create a perfect match. Such a redistribution resulted in daily rainfall frequencies and intensities that were more realistic for the monthly total rainfall. A limitation of the Ines et al. (2011) approach is that it only sources predictability from GCM rainfall, taking temperature and radiation to be the long term means conditioned on the forecast month and the forecasted occurrence of daily rainfall. For broad applicability, downscaling tools need to be capable of handling a wide range of variables such as rainfall, temperature, solar radiation and evaporation and manage their associated inter-variable relationships, possibly across multiple sites.

Recent efforts to advance weather generators to downscale forecasts for multiple sites and to use multiple sources of predictability do not seem to have delivered a generally applicable method. Chen et al. (2017) developed a multi-site weather-generator to generate daily precipitation and maximum and minimum temperatures from GCM outputs. While the method could derive multi-site sequences simultaneously for either rainfall or temperature, each variable was still generated independently.

More recently, Verdin et al. (2018) developed a parametric stochastic weather generator to output downscaled daily weather sequences based on tercile seasonal climate forecasts from the International Research Institute for Climate and Society (IRI). The approach is flexible enough to incorporate any predictors and, therefore, could theoretically be applied to downscale GCM forecasts. However, the approach is still limited to rainfall as the primary variable, with temperature modelled conditionally on the occurrence of rainfall. It appears that stochastic weather-generators require consolidation of ideas and further development to progress as a general-purpose GCM downscaling tool.

In light of the limitations of existing approaches, I propose to couple a rigorous forecast calibration at coarse-scales with a relatively simple multivariate, data-driven downscaling procedure that simultaneously generates correlated daily sequences for multiple sites. With regards to the calibration of coarse-scale forecasts, I will apply the Bayesian joint probability modelling approach (BJP) (Wang and Robertson 2011; Wang et al. 2009), which has been extensively applied to calibrate GCM outputs of rainfall and temperature (Hawthorne et al. 2013; Schepen et al. 2014; Schepen et al. 2016; Strazzo et al. 2018). Here, I calibrate monthly, spatially-aggregated forecasts. Temporal and inter-variable relationships from historical data will be injected into these monthly forecasts using the Schaake Shuffle (Clark et al. 2004; Vrac and Friederichs 2015). I expect these first steps to work well because BJP-calibrated monthly rainfall forecasts have previously been shown to generate skilful and reliable seasonal streamflow forecasts up to 12 months ahead using a monthly hydrological model (Bennett et al. 2016; Bennett et al. 2017b). If BJP generates monthly forecasts with appropriate ensemble spread, then the aim of the downscaling is to preserve the monthly forecast distributions while producing realistic daily sequences with correct weather patterns and extremes. To this end, I combine features of nearest neighbour resampling (Buishand and Brandsma 2001) and the method of fragments (Li et al. 2018; Srikanthan and McMahon 2001; Westra et al. 2012) to downscale each ensemble member individually with a suitable historical pattern.

In this study, I present the new methodology, given an acronym for the sake of clarity — FCMD (forecast-calibration multivariate-downscaling) — along with results of an application to ECMWF System4 forecasts in north-eastern Australia. The calibrated coarse-scale forecasts will be checked for skill and reliability. The daily, spatially-downscaled forecasts will be checked for correct temporal, spatial and inter-variable relationships by comparing with observations and raw forecasts. I structure the remainder of the chapter as methods, application and verification, results, discussion and conclusions.

3.3. Methods

3.3.1. Forecast Calibration – Multivariate Downscaling

I give a broad overview of the Forecast Calibration – Multivariate Downscaling (FCMD) workflow here and in Figure 3.1, with detailed information and references given in the subsequent sections.

The essential FCMD steps are:

- (1) For a given set of downscaling target locations, e.g., a cluster of weather stations or an array of fine resolution grid cell centroids, GCM forecasts are first interpolated to the set of target locations. Nearest neighbour interpolation is used in this study. Observations and forecasts are then aggregated spatially over all of the locations and in time to the desired resolution. Observations and forecasts are averaged to monthly data in this study. The averaging to larger-spatial scales is intended to reduce noise and to improve the chances of capturing useful GCM climate signals.
- (2) The aggregated forecasts are calibrated with the Bayesian joint probability modelling approach (BJP; Wang and Robertson 2011; Wang et al. 2009) to reduce bias, produce reliable forecast uncertainty estimates and return forecasts to climatology where skill is worse than climatology. I establish a separate BJP model for each climate variable and lead time. Hence if there are N climate variables and L lead times to be calibrated, $N \times L$ BJP

models are established. $M = 200$ ensemble members are generated by each model. The Schaake Shuffle (Clark et al. 2004) with a historical data template is then used to connect up the ensemble members across the climate variables and lead times so that the forecast ensembles inherit temporal and inter-variable correlation structures from observations.

- (3) Each monthly ensemble member is downscaled (disaggregated) to daily sequences and to each target location using a pattern obtained from a historical nearest neighbour (or more accurately, the most similar occurrence). All variables are used in the nearest neighbour search (after standardisation) and the same pattern is applied to all variables. I allow nearest neighbours to be found within a three-month window of the target month.

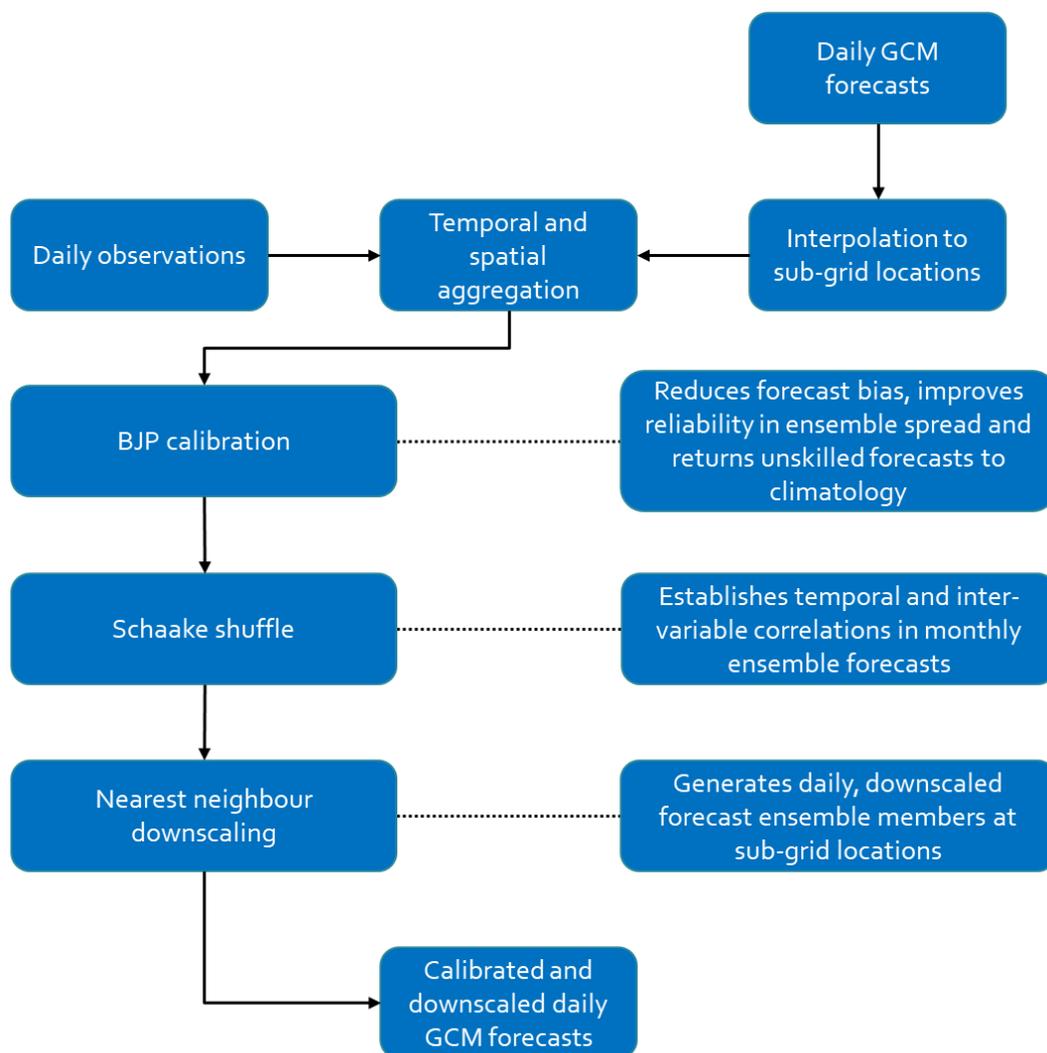


Figure 3.1: Schematic of the forecast-calibration multivariate-downscaling (FCMD) workflow to produce daily, spatially-downscaled ensemble forecast sequences from coarsely gridded monthly GCM forecasts

3.3.2. BJP forecast calibration

Spatially and temporally aggregated forecasts are calibrated using the Bayesian joint probability modelling approach (BJP) (Wang and Robertson 2011; Wang et al. 2009). BJP embeds a model of transformed predictor and predictand variables as a multivariate normal distribution where the transformations allow efficient handling of non-normal variables. It is a general statistical prediction model that can be referred to as a forecast calibration model, a downscaling model or purely as a statistical forecasting model. In this study, BJP predictors are raw monthly GCM forecasts and BJP predictands are monthly observations. Although BJP is described in Chapter 2, a briefer description is given here for completeness.

Denote a generic normalizing transformation function ψ with parameters $\mathbf{\Delta}$. The random vector of transformed predictor and predictand variables is $\mathbf{y} = (\psi_1(x_1), \psi_2(x_2), \dots, \psi_N(x_N))$ where x_i is an original untransformed variable. It is assumed that the joint distribution of a transformed set of variables is multivariate normal, i.e.

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (28)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the means and covariance matrix parameters from the multivariate normal distribution.

For rainfall, ψ is a two-parameter log-sinh transformation (Wang et al. 2012b). For other variables, ψ is a single-parameter Yeo-Johnson transformation (Yeo and Johnson 2000). I estimate a single “best” set of transformation parameters for ψ_i using a Bayesian maximum a posteriori (MAP) solution using the methodology described by Schepen et al. (2016). In contrast to the point estimation of the transformation parameters, the inference of the multivariate normal parameters allows for parameter uncertainty. A Gibbs sampler is used to obtain M samples of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. In predictive mode, the Gibbs sampler is used to obtain a single sample of $\mathbf{y}_{\text{ptand}} | \mathbf{y}_{\text{ptor}}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m$ for each

of $m = 1, 2, \dots, M$ where \mathbf{y}_{ptor} and $\mathbf{y}_{\text{ptand}}$ are the random vectors of BJP predictors and predictands, respectively. The collection of the samples $(\mathbf{y}_{\text{ptand},1}, \dots, \mathbf{y}_{\text{ptand},M})$ constitutes an ensemble forecast $\mathbf{Y}_{\text{ptand}}$ from a BJP model and follows a multivariate normal distribution in transformed space. The forecasts in $\mathbf{Y}_{\text{ptand}}$ for each transformed predictand variable are back-transformed to the original space using the appropriate inverse transformation ψ_i^{-1} to obtain $\mathbf{X}_{\text{ptand}}$, a set of ensemble forecasts in the original (untransformed) space.

Special treatments are involved where a predictor or predictand variable has a physical lower bound, e.g., for rainfall, the lower bound is zero. In the BJP modelling approach, such variables are handled by treating the data as left-censored (Wang and Robertson 2011). The censoring treatment applies to transformation parameter estimation, BJP parameter inference and forecasting.

3.3.3. Schaake Shuffle ensemble reordering

Natural weather patterns indicate that forecast ensemble members traversing time should have a reasonable degree of autocorrelation. For example, it is expected that either warm or cool temperatures can persist for many days and that temperatures exhibit a degree of autocorrelation in general, including at the monthly to seasonal time scales of interest in this study. Similarly, rainfall is expected to have patterns of wet and dry periods. When BJP is applied separately to different variables and months ahead (lead times), the concatenated ensembles will not automatically contain realistic temporal or inter-variable patterns across months by virtue of having little or no correlation between ensemble members with the same indexing position in the ensemble. A further post-processing step is needed to restore realistic covariance in the post-processed ensemble forecasts. In this study, I choose to use the Schaake Shuffle (Clark et al. 2004), previously used in Chapter 2 to reconstruct inter-variable correlations, to additionally impose realistic correlation structure in the monthly BJP forecasts. The fundamental description previously given in Chapter 2 is repeated here

as the description is short and additional commentary is needed about its use in the time series forecasting context. For a given forecast time period (e.g. month), consider an ensemble forecast of size M denoted by

$$\mathbf{X} = (x_1, x_2, \dots, x_M) \quad (29)$$

that can be sorted to obtain

$$\boldsymbol{\chi} = (x_{(1)}, x_{(2)}, \dots, x_{(M)}) \quad x_{(1)} \leq x_{(2)} \dots \leq x_{(M)} \quad (30)$$

Consider also a vector of observations from the historical record for the same time period (e.g. the same month in other years), also of size M

$$\mathbf{Y} = (y_1, y_2, \dots, y_M) \quad (31)$$

that can be sorted to obtain

$$\boldsymbol{\gamma} = (y_{(1)}, y_{(2)}, \dots, y_{(M)}) \quad y_{(1)} \leq y_{(2)} \dots \leq y_{(M)} \quad (32)$$

Furthermore, let rank be a function that determines the position of a value from $\boldsymbol{\gamma}$ in the original unsorted vector \mathbf{Y} . The shuffled forecast ensemble is constructed as

$$\mathbf{X}_{SS} = (x_{ss,1}, \dots, x_{ss,M}) \quad (33)$$

where $x_{ss,q} = x_{(n)}$ and $q = \text{rank}(\mathbf{Y}, y_{(n)})$ $n = 1, \dots, M$. The key to the Schaake Shuffle is the selection of dates used to construct \mathbf{Y} . For the first lead time, the dates may be selected randomly from the historical record. Clark et al. (2004) selected dates from within a 7-day window either side of the target time period. To shuffle subsequent time steps, the dates used to shuffle the first time step are incremented one time step at a time. When the historical observations have high temporal autocorrelation, the result is that the temporal rank correlation structure in labelled ensemble members will be similar to the template data over successive lead times. In this study, to construct

Y , offsets of -30, -15, 0, 15 and 30 days are used to select the first time step of the Schaake Shuffle. Daily data is subsequently aggregated to match the number of days in each monthly forecast to be shuffled.

As alluded to earlier, a neat property of the Schaake Shuffle is that if the same dates are used to construct the template for all forecast variables for a given forecast initialisation date, then both the inter-variable correlations and temporal correlations are reconstructed simultaneously. (The method can also reconstruct spatial correlations in the same way.) Thus the Schaake Shuffle is very useful for producing physically coherent forecasts of multiple variables that are initially calibrated independently using a method like BJP. The Schaake Shuffle in high-dimensional applications that would quickly become intractable with fully parametric methods.

Although part of the beauty of the Schaake Shuffle lies in its simplicity, it does have several drawbacks, and, therefore, many modifications have been proposed in recent times. Suggestions include using historical GCM ensembles as the dependence template (Schefzik et al. 2013) or preferentially selecting start dates using analogues or other similarity criteria (Schefzik 2016b; Scheuerer et al. 2017; Wu et al. 2018). I have opted to apply the Schaake Shuffle in its original form. One argument for not conditioning the Schaake Shuffle is that the forecast ensembles in seasonal forecasting (in contrast to something like weather forecasting) often represent a wide range of outcomes and, therefore, it is necessary to include a wide range of observation trajectories in the Schaake Shuffle.

3.3.4. Nearest neighbour downscaling

Each BJP calibrated forecast contains M ensemble members, each as a sequence of L monthly forecasts. To downscale the forecast to an ensemble of daily sequences at each station, I adapt the method of fragments (MOF), which has mainly been applied for temporal rainfall disaggregation, including from annual-to-monthly to daily-to-sub-daily (e.g. Li et al. 2018; Pui et al. 2012; Srikanthan

and McMahon 2001; Westra et al. 2012). In some instances, it has been applied to disaggregate temperature (Wójcik and Buishand 2003). My motivation for using MOF is not dissimilar to the motivation for Ines et al. (2011) to use a stochastic weather generator to redistribute GCM rainfall more realistically within a month. That is, it is a way to obtain sequences of daily events that have appropriate magnitudes and frequencies of occurrence for the monthly value.

To my knowledge, MOF hasn't been previously applied to downscale ensemble climate forecasts and, therefore, some modifications are required. Here, I apply the method of fragments to downscale all variables simultaneously. Each month is downscaled separately, which raises the possibility of edge effects across month boundaries. I do not expect that the forecast sequences will endure dramatic edge effects since the monthly forecasts have already been shuffled with the Schaake Shuffle. Because it is normally applied as a way to disaggregate observed data (e.g. to infill missing higher-resolution data), MOF usually involves randomly selecting one of k -nearest neighbours according to some probability (e.g. Pui et al. 2012). But because the forecast ensembles generated by BJP are reasonably large ($M = 200$ in this study), representing a continuous density forecast, I take a simpler approach and use the single nearest neighbour as a downscaling template. Choosing the closest nearest neighbour limits any rescaling required to preserve the monthly forecasts. As an (extreme) illustrative example, consider an exceptionally wet monthly rainfall forecast being redistributed with a patently dry sequence; the resulting daily sequence could have unrealistically high rainfall on too few days. Such a problem is noted by Gutmann et al. (2014) in application of the bias-correction statistical-disaggregation (BCSD) approach, requiring limitation of daily values to 150% of previously observed maximums, and a redistribution of remaining rainfall within the month, even when template data is selected conditionally on wet and dry days. I don't limit the daily rainfall in this study; rather, I assume that choosing the closest nearest neighbour will keep extreme events in check. However, it remains prudent to check for outliers in the results, which could still be caused by other components in the Forecast Calibration – Multivariate Downscaling workflow (e.g. data transformation).

A nearest neighbour is found by similarity measured using the Euclidean distance for standardised variables. Given N variables of a point forecast f and corresponding observation set o the Euclidean distance between all the points is

$$d_{\text{Euclid}} = \sqrt{\sum_{i=1}^N (f_i - o_i)^2} \quad (34)$$

A historical nearest neighbour is found by identifying a historical date for which d_{Euclid} is the smallest. To avoid the potential problem of variables on different scales skewing the search towards a particular variable when using the Euclidean distance, the variables are standardised to dimensionless quantities as suggested by Buishand and Brandsma (2001). How much of an effect this has will depend on the chosen units of the variables. For example, if monthly rainfall in millimetres is being compared with monthly average daily minimum and maximum temperatures in degrees Celsius, the benefit of standardisation is likely to be significant because of the possible large variation in rainfall with temperature. However, if rainfall is cast as a daily rate, then standardisation will have a more subtle effect.

The standardisation here is done on spatially and temporally aggregated data, suggesting that seasonality ought to be considered as a factor. In particular, temperature variations and rainfall frequency and intensity can be vary substantially throughout the year. Therefore, it is deemed unreasonable to search for neighbours from seasons distinct from the forecast period and I only permit neighbours to be found within three months of the forecast target month.

In this study, the standardisation step depends on the type of climate variable. A temperature or solar radiation variable x_i is standardised to \tilde{x}_i by

$$\tilde{x}_i = \frac{x_i - \bar{x}_{\text{seas}}}{s_{\text{seas}}} \quad (35)$$

where \bar{x}_{seas} is the mean of data available for the search, and s_{seas} is the corresponding standard deviation. A rainfall variable x_i is standardised to \tilde{x}_i by

$$\tilde{x}_i = \frac{\tilde{x}_i}{\bar{x}_{seas}} \quad (36)$$

As mentioned, the method of fragments has mainly been applied to rainfall. I, therefore, detail the procedure for finding a suitable nearest neighbour that can be applied to downscale a set of diverse climate variables simultaneously. Additionally, to my knowledge the method of fragments hasn't been previously applied as a spatial downscaling tool, so I also present that straightforward extension here.

For the purposes of the description relevant to this study, it is assumed that variables are already transformed to eliminate mixes of positive and negative values and thus keep the description the same for all variables. For example, temperature data are ensured to be in Kelvin rather than degrees Celsius. Accordingly, fragment weights ω for a location s , day d and variable i are calculated from the observation data set as follows:

$$\omega_{sdi} = o_{sdi} / \bar{o}_i \quad (37)$$

where o_{sdi} are the individual daily observed values and \bar{o}_i is the average spatially over K locations in the region and temporally over T_{agg} preceding days.

Once weights have been calculated, a new forecast for the spatially and temporally averaged target, \bar{f}_i , is downscaled to each location and to daily values as follows:

$$f_{sdi} = \omega_{sdi} \bar{f}_i \quad (38)$$

The downscaled forecast is designed to take on the behaviour of observations because the historical data template preserves temporal, spatial and inter-variable correlation structures (within the aggregation period of length T_{agg}). As with all data-driven methods, a short data record is potentially problematic when searching for nearest-neighbours. Especially with a seasonality restriction, since the same observed daily sequences could be selected repeatedly. To increase the diversity in the selected sequences, given that the sequences are only to be used as a template for downscaling, rolling aggregates of the observed data are allowed in the search, where the rolling aggregates are prepared for periods of 28–31 days, the range of possible number of days in a month. This not only ensures a large number of available samples, but also introduces timing shifts into the forecast ensembles. In the undesirable case that similar forecast ensemble members are downscaled using the same sequence, they will still differ by proportionality. Occasionally, enforcements are needed so that sensible sequences are chosen. For example, if forecast rainfall is a positive amount then it is not reasonable to choose a neighbour that has zero rainfall.

3.4. Application and verification

3.4.1. Study area and data

Six study weather stations are chosen in the Burdekin region of Australia (Figure 3.2) and the forecast period from September to February. The Burdekin region is tropical and experiences a wet season from November to May. The region and period is chosen because it is known that three-month forecasting skill is obtainable from GCMs in the region (Figure 2.3 in Chapter 2) and thus it is a suitable case study to investigate skill at monthly time steps and for longer lead times. Further to the point, targeting a skilful region ensures that the forecasts vary from year-to-year. In regions where no skill exists, BJP calibration will always return forecasts to climatology, which is a less challenging scenario for disaggregation.

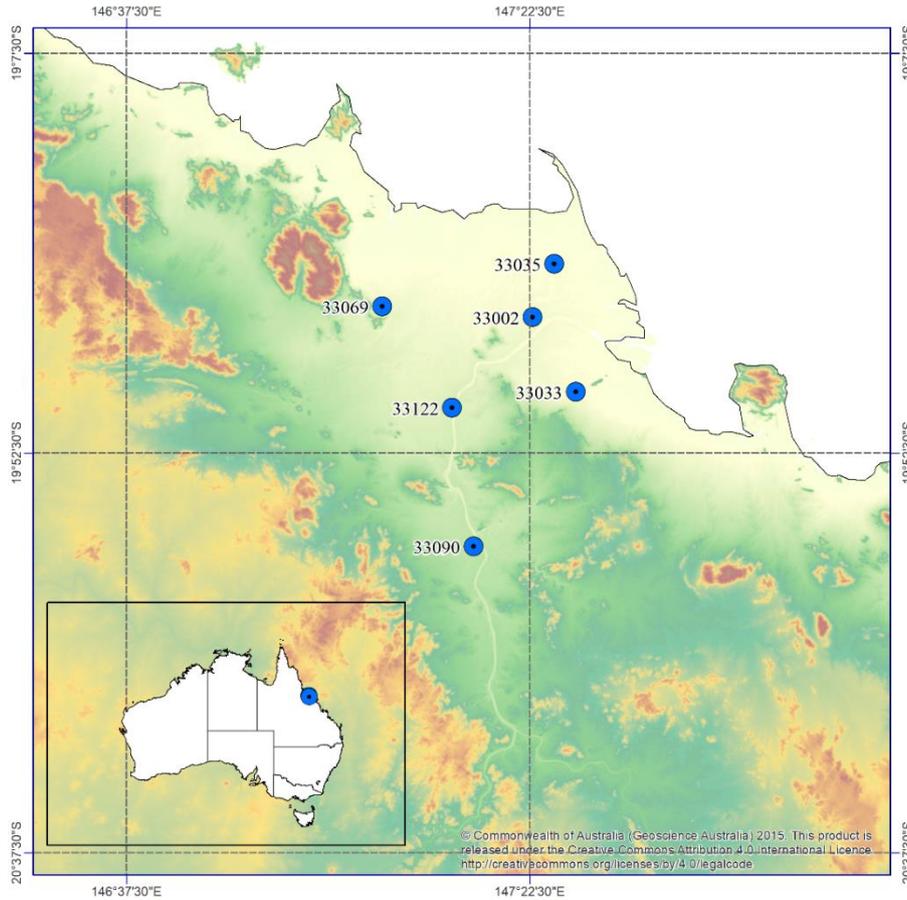


Figure 3.2: The location of the six study weather stations in north-eastern Australia (blue dots) and the ECMWF System4 grid cell boundaries (grey dashed lines).

Observations of rainfall (Precip; mm), daily minimum temperature (Tmin; °C), daily maximum temperature (Tmax; °C) and radiation (Srad; MJ/m²/day) are obtained from the Silo dataset for each of the selected stations. The dataset and variables are commonly used by crop modellers in Australia. The data have a discretised nature with Precip reported to the nearest 0.1 mm, Tmin and Tmax to the nearest 0.5 °C and Srad to the nearest 1.0 MJ/m²/day.

GCM forecasts are obtained from ECMWF's System4 (Sys4; Molteni et al. 2011) seasonal forecast system at a daily time step and on a 0.75 degree (~80km) regular grid for the period 1981–2016. To prepare the data for BJP calibration, GCM forecasts are spatially aggregated by interpolating the GCM forecasts to each point, by nearest neighbour, and then averaging across the six stations. In

other words, the aggregated GCM forecast is weighted by the number of stations within the grid cell. The corresponding Silo observations are simply averaged across the six stations.

3.4.2. Forecast verification

3.4.2.1. Cross-validation

Leave-one-year-out cross-validation is applied to generate and evaluate the study results. That is, for each forecast month, data from the verifying month is omitted from the model fitting part of BJP calibration, plus data in the relevant window is omitted from the Schaake Shuffle and nearest-neighbour downscaling procedures. Leave-one-year-out cross-validation is a standard practice for seasonal climate forecast evaluation where hindcast datasets are relatively short.

3.4.2.2. Seasonal skill and reliability evaluation

Reliability is the property of statistical consistency between forecasts and observations. A reliable forecasting system will accurately estimate the likelihood of an event. Reliability is checked by analysing the distribution of probability integral transformations or PIT values (Gneiting et al. 2007).

The PIT for a forecast CDF (F_t) for event t and paired observation (o_t) is defined by

$$\pi_t = F_t(o_t) \quad (39)$$

If a forecasting system is reliable and the forecasts are continuous, then the PIT values for a set of forecasts will follow a standard uniform distribution. Hence, I check for uniformity visually using the PIT uniform probability plot, otherwise known as predictive Q-Q plot, by plotting sorted PIT values

$\pi_{(i)}$ against theoretical uniform quantiles u_i for $i = 1, \dots, T$ (Renard et al. 2010; Wang and

Robertson 2011). As a more formal test of uniformity I plot Kolmogorov-Smirnov confidence bands

on the PIT uniform probability plot (Laio and Tamea 2007). The test is passed if no PIT values lie

beyond $[u_i - c(\alpha)\sqrt{2/T}, u_i + c(\alpha)\sqrt{2/T}]$ where $c(\alpha) = 1.224$ for a 95% confidence test.

Forecast skill is evaluated by comparing the continuous ranked probability score (CRPS; Matheson and Winkler 1976) for different model forecasts. The CRPS for a given forecast and observation is defined as

$$\text{CRPS}_t = \int [F_t(y) - H(y - o_t)]^2 dy \quad (40)$$

with $H(y - o_t) = \begin{cases} 0 & \text{if } y < o_t \\ 1 & \text{if } y \geq o_t \end{cases}$

where F_t is the forecast CDF for event t ; o_t is the observed value; and H is the Heaviside step function. The CRPS puts weight on both forecast accuracy and reliability in ensemble spread.

The average CRPS for two sets of forecasts is compared to calculate calculating the CRPS skill score.

$$\text{CRPS}_{\text{SS}} = \frac{\overline{\text{CRPS}}_{\text{ref}} - \overline{\text{CRPS}}}{\overline{\text{CRPS}}_{\text{ref}}} \times 100 \quad (\%) \quad (41)$$

where the overbar indicates averaging across a set of events, $\overline{\text{CRPS}}$ is the average CRPS for the focus forecasts and $\overline{\text{CRPS}}_{\text{ref}}$ is the average CRPS for a set of reference forecasts. In this study, reference forecasts are obtained by resampling from marginal distributions fitting to monthly data and disaggregated using the same approach as for forecasts.

3.4.2.3. Validation of downscaled daily sequences

Daily forecasts are validated in terms of having realistic distributions and covariance structures. For each location and variable, the magnitudes of the daily forecast values are visually checked by comparing the [0.25, 0.75] and [0.05, 0.95] quantile ranges of daily forecast values and Silo observations. For rainfall, the frequency of wet days is checked by comparing the proportion of wet days (wet%) where Precip exceeds 1mm. For the examination of quantile ranges and wet day fractions, an average is taken over the M parameter sets so that the same number of data points are used in the comparisons.

Kendall correlations (r), as a measure of rank correspondence, are calculated to evaluate temporal, spatial, and inter-variable relationships. Comparisons are made between the downscaled forecasts, raw Sys4 forecasts and Silo observations to determine any discrepancies in the correlation structures of raw and downscaled forecasts.

For temporal correlations, lags of 1–7 days are evaluated for each station and variable. For spatial correlations and inter-variable correlations, the correlation is calculated on all possible pairwise combinations. Correlations are averaged over T years. For forecasts, a further average is taken over the M ensemble members.

3.5. Results

3.5.1. Skill and reliability of BJP-calibrated forecasts

CRPS skill scores for monthly and seasonal forecasts are shown in Figure 3.3. The seasonal forecasts are obtained from the monthly forecasts by aggregating each ensemble member over three months. All forecasts are issued on the 1st of September. For monthly forecasts, the target months Sep–Feb correspond to lead times of 0–5 months. Similarly, for the seasonal forecasts, the target seasons SON–DJF correspond to forecasts with 0–3 months lead time. The skill scores measure relative reduction of CRPS error relative to the reference forecast (see section 3.4.2.2) over the period 1981–2016.

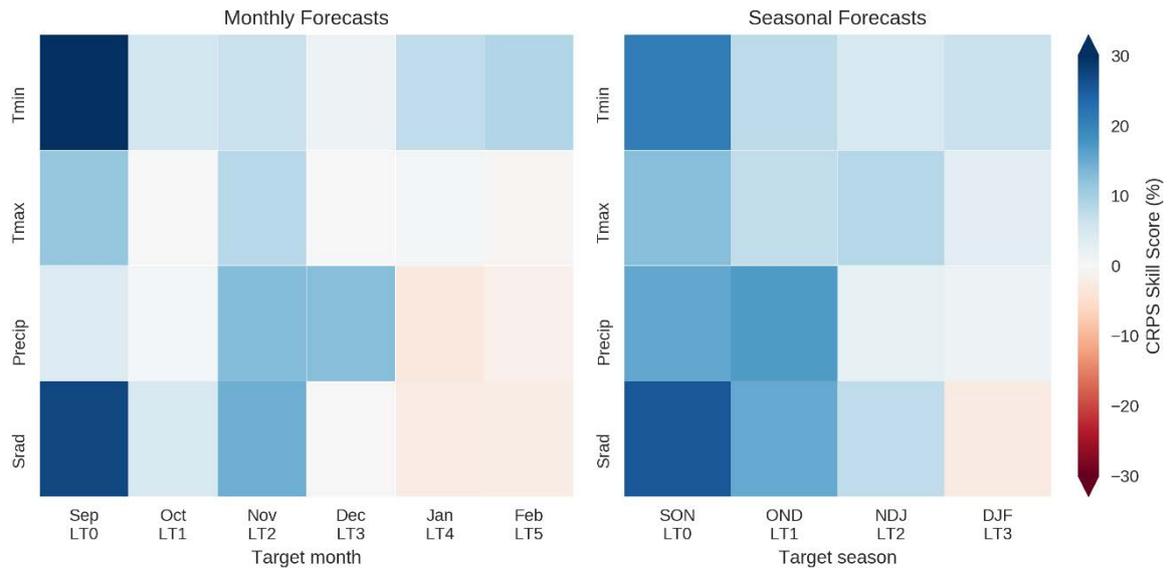


Figure 3.3: CRPS skill scores for BJP-calibrated monthly and seasonal forecasts issued in September. Skill scores are calculated against climatological reference forecasts for the years 1981–2016.

Monthly forecasts of Tmin, Tmax and Srad are skilful for the first month, and more skilful than the corresponding Precip forecasts. Beyond the first month, longer lead-time skill for is evident for Nov forecasts. Longer lead time skill for Nov rainfall forecasts has been previously observed in the study region using other GCMs and observational datasets (Hawthorne et al. 2013). While no study appears to have confirmed it, the skill is possibly related to the timing of the onset of the wet season. Seasonal forecasts of all variables are more skilful than monthly forecasts. Viewing the monthly forecasts as two distinct periods, SON and DJF, illustrates that the first 3 months of the forecast are moderately skilful whereas only a small amount of skill is available in the second 3 months of the forecast. BJP calibration is designed to push forecasts towards climatology where skill is low. However, some small negative skill scores are observed, particularly in monthly forecasts, where there is essentially no skill and because calibration is undertaken in a cross-validation mode.

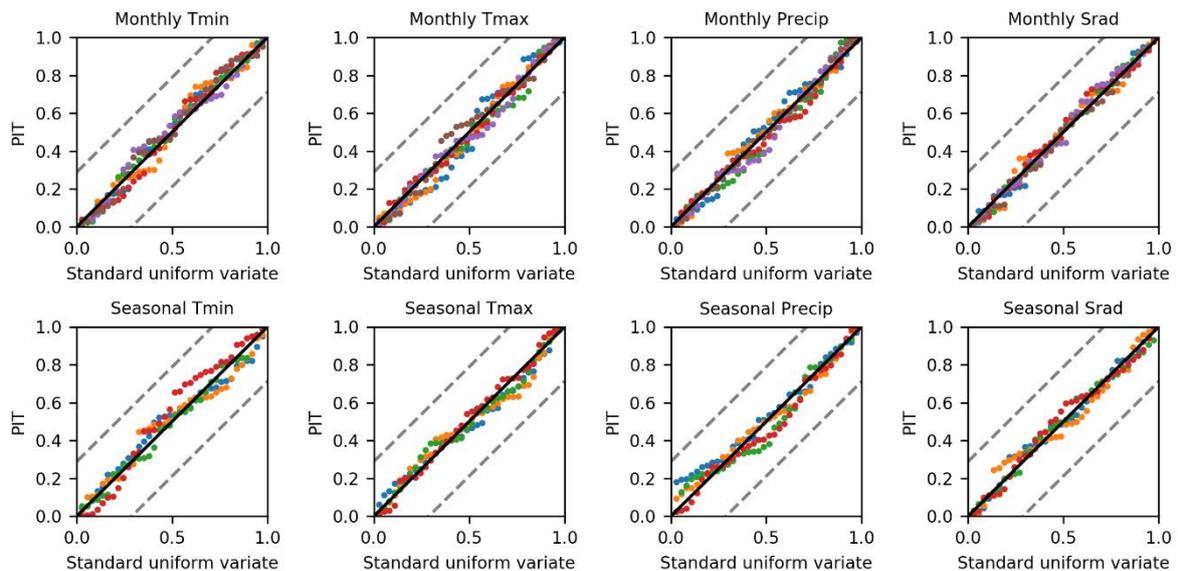


Figure 3.4: PIT reliability plots for BJP-calibrated monthly and seasonal forecasts issued in September. Different coloured points represent different lead times from 0–5 months for monthly forecasts and 0–3 months for seasonal forecasts. The dashed-grey lines are the Kolmogorov-Smirnov significance bands.

PIT uniform probability plots for the monthly and seasonal forecasts are shown in Figure 3.4 to assess reliability. All of the lead times for a given variable are plotted on the same graph using

different colours. It is evident that both the monthly and the seasonal forecasts are overall reliable in terms of ensemble spread. None of the PIT values fall outside the Kolmogorov-Smirnov 95% confidence bands, indicating that the test for uniformity passes, thus providing more formal evidence that the forecasts are statistically reliable.

3.5.2. Distributions of daily values

The distributions of daily calibrated forecast values are compared with Silo observations for the drier months, Sep–Nov (Figure 3.5), and the wetter months, Dec–Feb (Figure 3.6). Stations are presented in columns and variables are presented in rows. For each variable, the all-station median is plotted as a grey dashed line to help discern shifts in the central tendency of the distributions between stations. For rainfall, the median is determined for wet days only, and the proportion of wet days (wet%) is calculated separately.

Overall the distributions of the forecast and observed values have approximately equal quantiles across the range of cumulative probabilities as well as the correct distributional shape. For example, Tmin and Tmax are approximately normally distributed, Precip is positively skewed whereas Srad is negatively skewed. The forecasts and observations appear to follow the same trends between locations. For example, for Precip during Dec–Feb, station 33090 has the highest proportion of wet days yet has the lowest daily rainfall amounts. Distributions of daily rainfall match closely for forecasts and observations, demonstrating that the downscaling method is reliable for generating realistic rainfall values with a realistic frequency of occurrence and intensity on any given day.

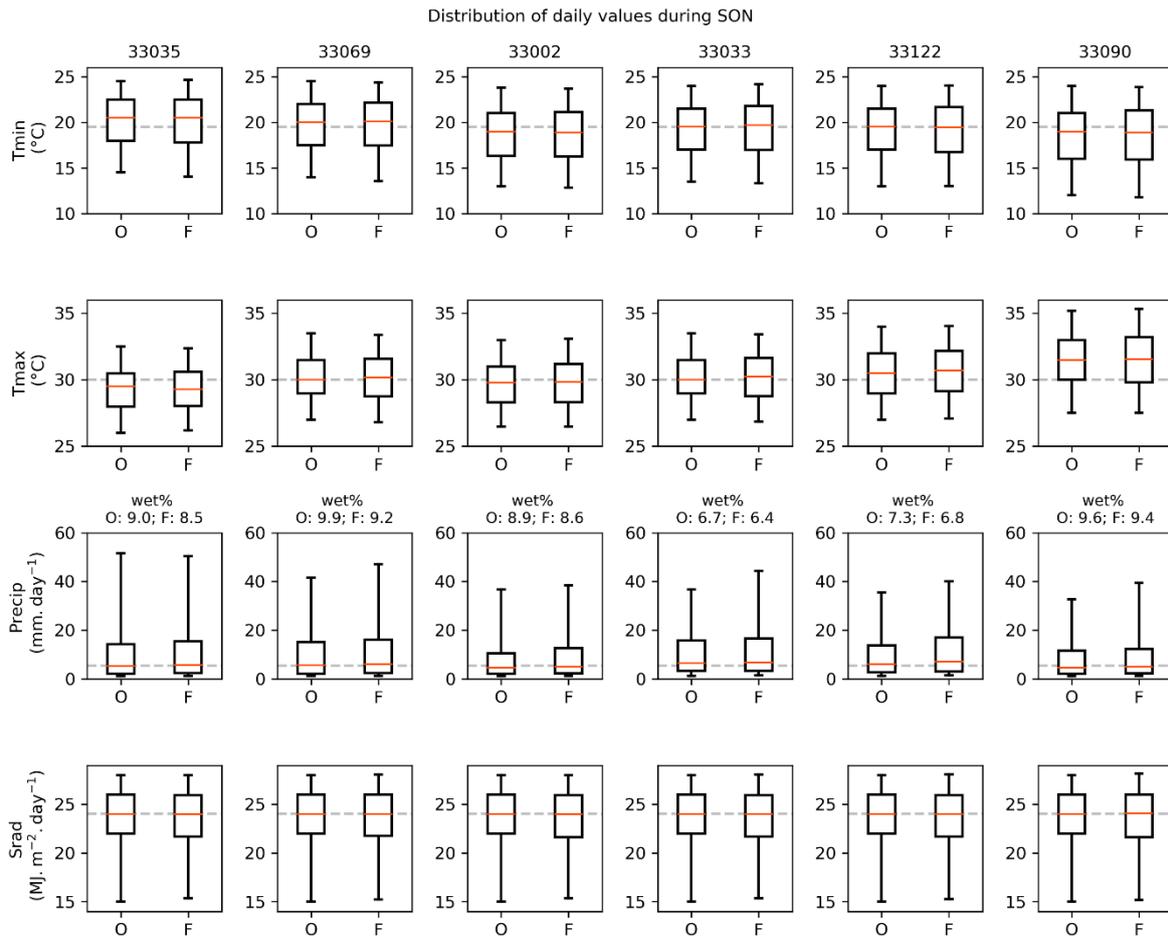


Figure 3.5: Boxplots comparing the distribution of daily forecast values with the distribution of daily observed values for observations (O) and forecasts (F) during Sep-Nov. Boxplots show the IQR and the [0.05,0.95] quantile ranges. For rainfall, the distribution is plotted for wet days only (Precip \geq 1mm) and the proportion of wet days is also given (wet%).

The distributions of daily values appear to accurately reflect the station location in some instances, in both forecasts and observations. For example, compare Tmin and Tmax for the inland station 33090 to the coastal station 33035. The median Tmax at the inland station 33090 is > 1 degree above the all-station median during SON whereas the median Tmax forecast for the station 33035 is below the all-station median. The differences in maximum temperatures can be justified in the context of sea-breezes limiting the maximum daytime temperatures at the coastal location. In contrast, the Tmin values at station 33090 indicate the coldest overnight temperatures are observed there.

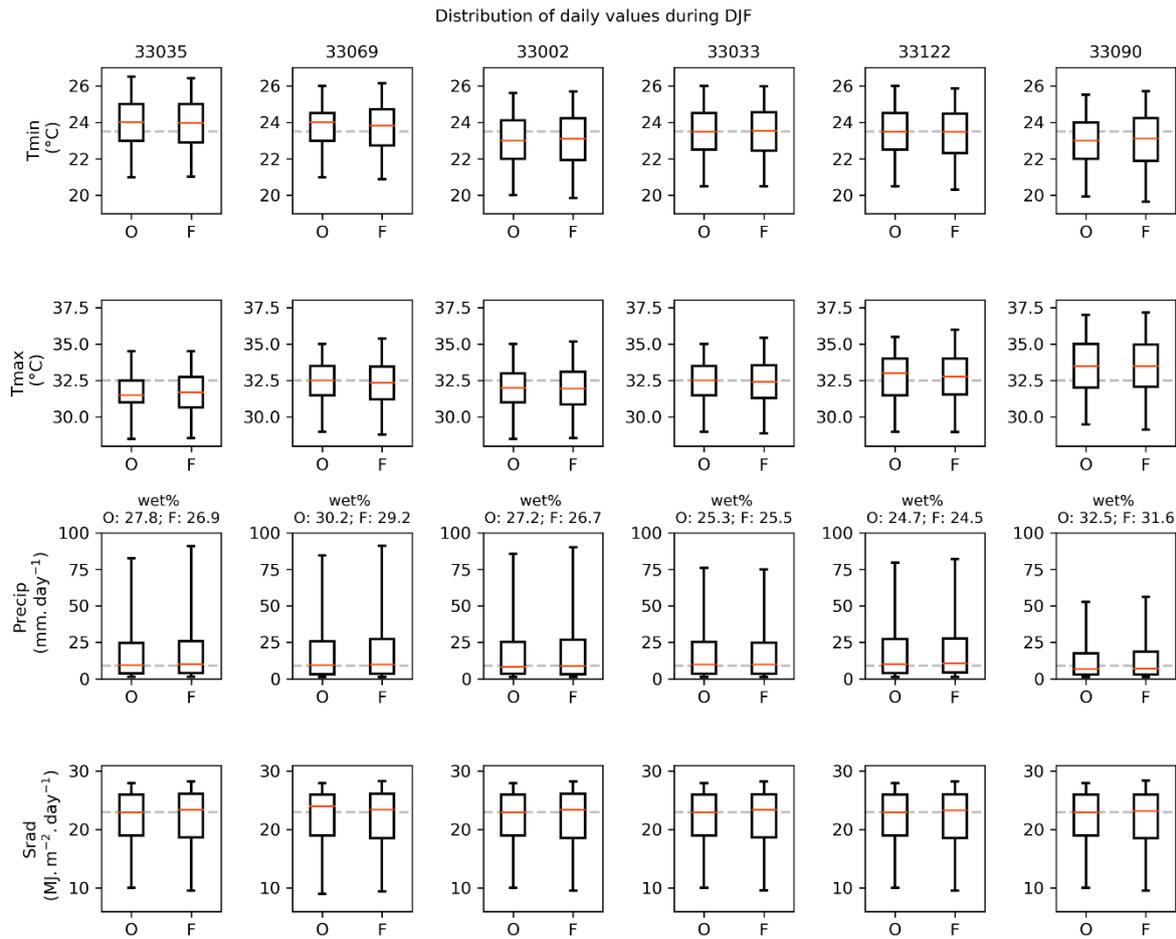


Figure 3.6: As for Figure 3.5, except for Dec-Feb forecasts and observations

3.5.3. Temporal correlations

I compare the average temporal correlations in daily FCMD forecasts with observations and with raw Sys4 forecasts. Average Kendall correlations for lags of 1–7 days are shown for each variable and location. Figure 3.7 is for Sep–Nov and Figure 3.8 is for Dec–Feb. FCMD forecasts, raw forecasts and observations are labelled cal, raw and obs, respectively. In all cases the correlation tends to reduce as the number of days lag increases, however, the rate of decline varies. It is noted that because some of the locations reside within the same grid cell, the correlations in raw GCM forecasts will be the same for those locations (the spatial correlation is 1 – see next section).

Referring to Silo observations, the temporal correlation behaviour varies substantially between variables and time of year. Temporal correlation in Tmin and Tmax is quite strong and persistent during Sep-Nov, reducing from 0.6 to 0.4 as the lag increases from 1–7 days. However, during Dec-Feb, temporal correlations in Tmax and Tmin are generally weaker and less persistent. Similar comments apply to Srad. In contrast, temporal correlations in Precip show some opposite behaviour. 1-day lag correlations in Precip are slightly lower in Sep-Nov (0.2-3) compared to Dec-Feb. Precip temporal correlation tends to 0 at about 3 days lag during Sep-Nov and at about 4-5 days lag during Dec-Feb. In summary, Precip is seen to be more persistent during the wet period and temperature is seen to be more persistent during the drier period.

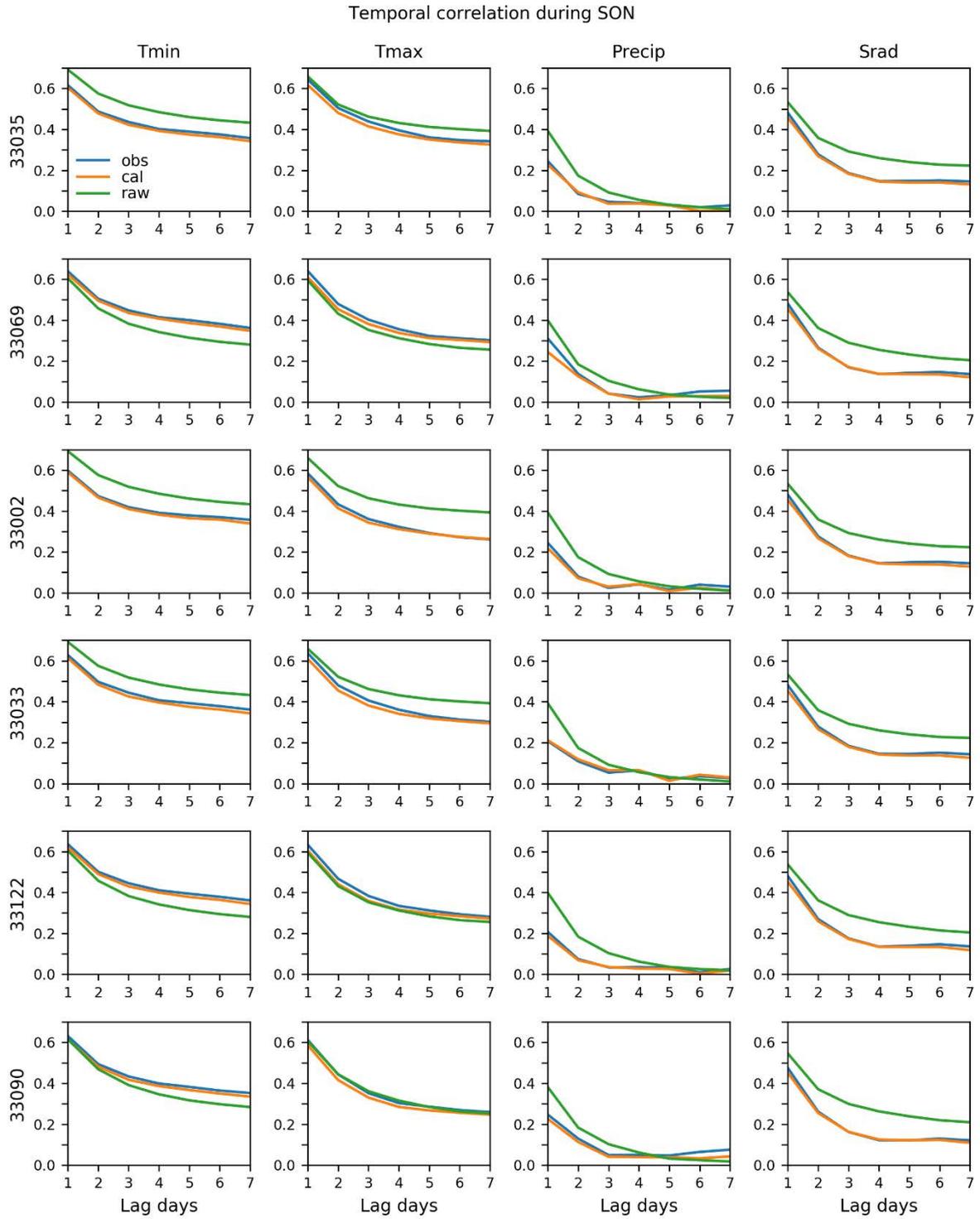


Figure 3.7: Average temporal correlations in Silo observations, FCMD forecasts (cal) and raw Sys4 forecasts for the Sep-Nov months. Kendall correlations are calculated for lags of 1-7 days.

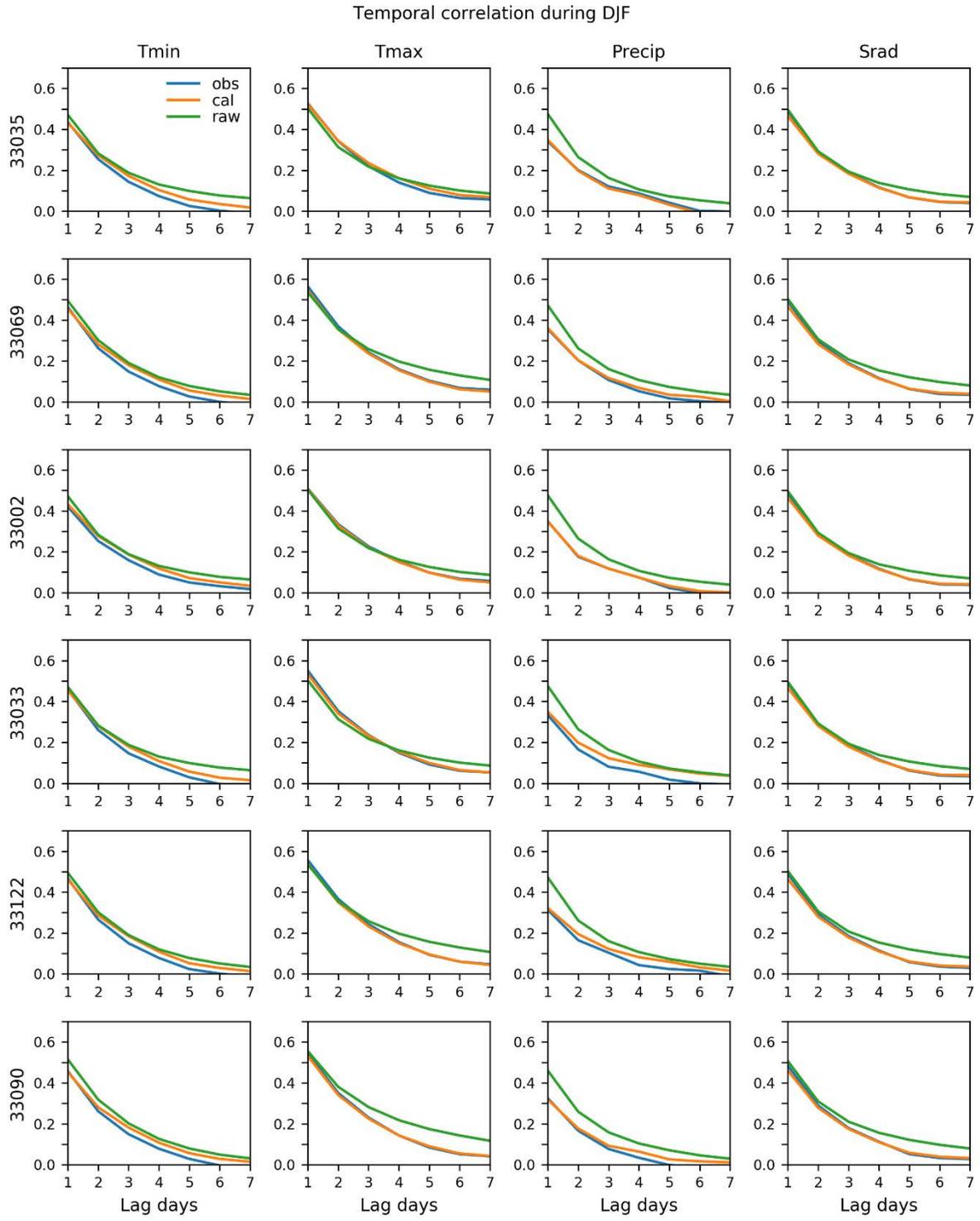


Figure 3.8: As for Figure 3.7, except for the Dec-Feb months.

It is immediately clear from Figure 3.7 (Sep-Nov forecasts) that while temporal correlations in raw GCM forecasts have a similar broad behaviour to Silo observations, the magnitudes of the correlations in raw forecasts tend to be higher, a result that is consistent with previous studies. There are exceptions, however. For Tmin, the raw forecasts have stronger temporal correlations than Silo at some locations and weaker temporal correlations at other locations. Figure 3.8 (Dec-Feb forecasts) shows a much closer match between raw GCM and Silo observations for correlations in Tmin, Tmax and Srad. However, there is a tendency for the correlations to be higher at longer lags, and also the correlation in Precip is still too strong.

The FCMD forecasts exhibit temporal correlations that are much more consistent with Silo observations overall. For all variables and locations the temporal correlations in FCMD forecasts are much improved compared to raw GCM forecasts.

3.5.4. Spatial correlations

Spatial correlations are compared between FCMD forecasts, Silo observations and raw Sys4 forecasts (Figure 3.9). Average Kendall correlations are plotted for each variable and for each pairwise combination of stations (within each plot). I do not necessarily expect that the coarse spatial GCM grid correctly captures the spatial correlation between stations. Indeed, for stations within the same grid cell, all raw Sys4 forecasts will be perfectly correlated when a nearest neighbour interpolation is used to obtain the station values. For Tmin, Tmax and Srad the spatial correlations in raw Sys4 forecasts can be stronger or weaker than in observations. In contrast, for Precip, the spatial correlation is always stronger relative to observations. FCMD downscaling reproduces the observed spatial correlations much more closely. The results are not perfect for Tmin, Tmax and Srad with FCMD spatial correlations being slightly lower than observed.

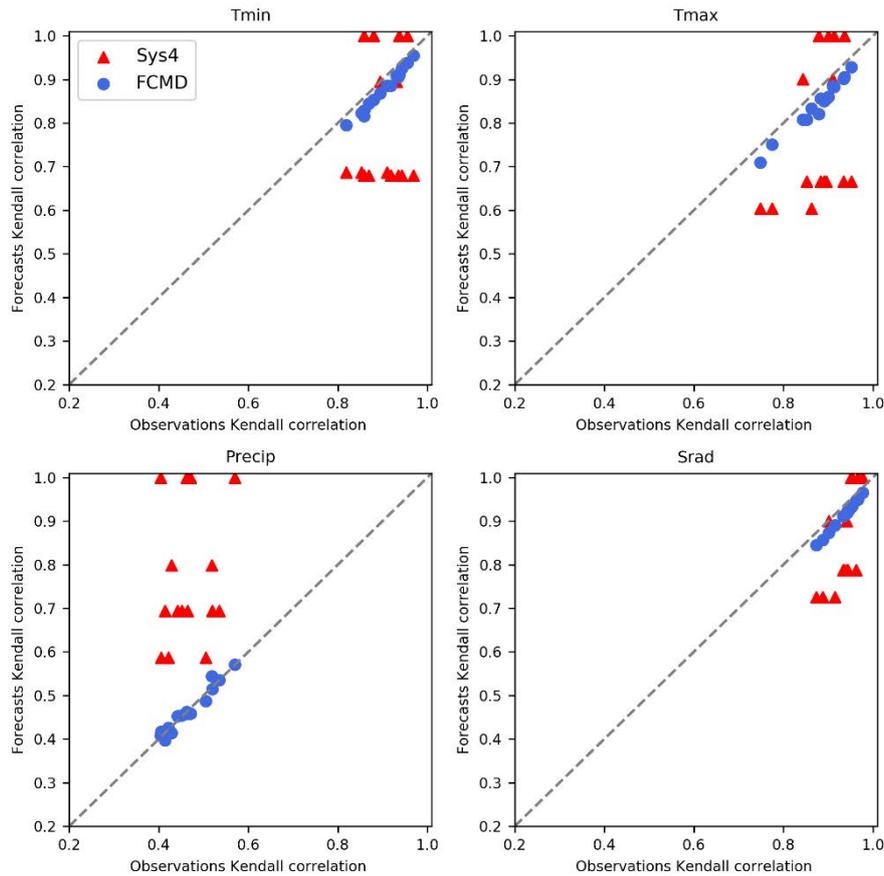


Figure 3.9: Comparison of average spatial Kendall correlations in Silo observations, FCMD forecasts and raw Sys4 forecasts for the Sep-Nov months. Each marker represents a station pairing (15 combinations).

3.5.5. Inter-variable correlations

Inter-variable correlations are compared between FCMD forecasts, Silo observations and raw Sys4 forecasts (Figure 3.10). Average Kendall correlations are plotted for each pairwise combination of variables and for each location (within each plot). As with temporal correlations, the raw Sys4 correlations are identical for locations that are in the same grid cell.

In some instances, the raw Sys4 forecasts have approximately correct inter-variable correlations. For example, the raw Sys4 forecasts capture the little or no correlation between Tmax and Precip and, for some locations, the low correlation between Tmin and Srad. However, in other cases, the raw GCM inter-variable correlation is too strong. For example, at some locations, the correlation between Tmin and Tmax is above 0.6 when the correlation in observations is closer to 0.5. The correlation between Precip-Srad appears to be too strong in general for all locations in the raw Sys4

forecasts. In contrast, the FCMD inter-variable correlations are similar to Silo observations for all pairwise combinations of variables and locations.

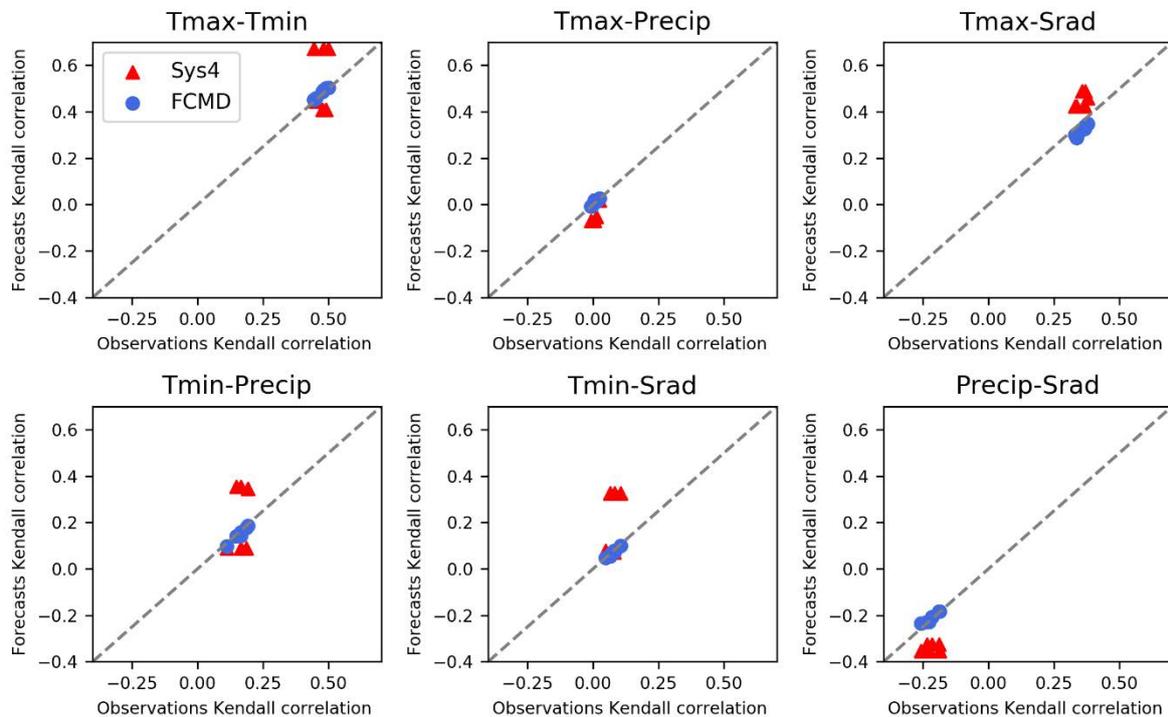


Figure 3.10: Comparison of average inter-variable Kendall correlations in Silo observations, FCMD forecasts and raw Sys4 forecasts for the Sep-Nov months. Each marker represents a single location.

3.6. Discussion

In this study, I compare the correlation structures in raw forecasts with FCMD forecasts. Figure 3.7 and Figure 3.8 largely support previous conclusions that raw GCM forecasts have autocorrelation structures that are too strong when compared to observations (Brown et al. 2018; Ines and Hansen 2006; Ines et al. 2011). However, I also find that the raw Sys4 forecasts sometimes have autocorrelation that is similar to, or weaker than, observations. Additionally, I show that spatial and inter-variable correlations are often wrongly specified in raw GCM forecasts (Figure 3.9 and Figure 3.10). The proposed FCMD method is able to generate downscaled (spatially and temporally) forecast ensemble members that have the correct temporal, spatial and inter-variable correlation

structures. It would be very valuable to determine whether the improved characteristics of FCMD forecasts translate into improved agricultural forecasts and ameliorates problems identified in previous studies that use simpler methods that inherit GCM correlations structures (e.g. quantile-mapping). It is possible that correcting correlations in climate ensembles will still not be enough, and that further post-processing of streamflow or yield predictions will be required.

As described in section 3.3.1, no constraint is placed on the values that can be generated in the nearest neighbour downscaling, thus it is possible that unreasonable values could be generated by the chaining of statistical functions. To check the reasonableness of the downscaled values, particularly extremes, I plotted the value of each downscaled forecast against its nearest neighbour used in the downscaling (not shown). While scaling can occur, it appears that appropriate weather patterns are overwhelmingly selected for the downscaling and, therefore, scaling is limited. Extrapolation occurs within reasonable limits, confirming that filtering of extreme values is not urgently needed, at least not in the current application.

I identify that there may be ways to improve upon the fundamental construction of FCMD. For example, while effective, it is not necessarily optimal to calibrate monthly forecasts independently for each variable with BJP, then inject the multivariate relationships (with the Schaake shuffle) before downscaling. I hypothesise that skill and spread could become quite different through a joint calibration. By way of example, consider the monthly Srad forecasts, which have low or slightly negative skill during Dec–Feb (Figure 3.3). Within the BJP calibration step, these forecasts are returned to an unconditional climatology, whereas it might be more reasonable to return a climatology conditional on the forecasts of the other variables where there is inter-variable correlation (Figure 3.10) and skilful forecasts of the related variables (Figure 3.3). The implication is that the multivariate BJP calibration from chapter 2 has some preferable features in this regard.

The provision of domain-specific forecasts opens up opportunities to tailor forecasts to specific applications. The BJP calibration component could also be augmented to predict variables that aren't

output by the GCM, e.g. potential evaporation (Zhao et al. 2019). Future work should investigate additional ways to improve skill and provide additional output variables. For example, the BJP calibration component of FCMD can be augmented to include other climate predictors from the GCM, related to climate drivers such as ENSO (e.g. Schepen et al. 2014; Schepen et al. 2016). I will address this idea in chapter 4.

FCMD is designed for seasonal forecasting and, therefore, it is not expected that forecasts from the first weeks would be suitable for inclusion in a hydrological or crop model for short term forecasting. Alternative approaches focusing on calibration of daily or multi-week forecasts directly (Schepen et al. 2018) are arguably needed, especially where forecasts may be highly skilful initially and tied to the GCM initial conditions.

Keeping on with alternative approaches, spatiotemporal relationships may be introduced into the forecasts by explicit/parametric modelling. Such extensions can quickly become intractable due to increases in the number of parameters and model complexity, however, approaches including Bayesian hierarchical models are worthy of investigation, particularly in conjunction with dimensionality-reducing methods. My intuition is that the while development of parameterised approaches is a long-term goal, hybrid post-processing systems involving pragmatic approaches like the data-driven downscaling method in FCMD, will continue to have a role in linking GCM forecasts and applications in the near term.

3.7. Conclusion

The water resources management and agricultural sectors often want to incorporate GCM forecasts into their modelling tools and decision support tools. Seasonal forecasting centres tend to issue coarsely-gridded outlooks with a supporting narrative, which are informative for casual forecast users. However, calibrated and downscaled daily forecast sequences are required for quantitative modelling using hydrological models and crop models.

A significant amount of literature has developed methods for bias-correcting and spatially or temporally downscaling GCM outputs or reanalysis data, mainly for climate impacts studies. In climate impacts studies there is no strict lead time dimensionality and simulations are not always synchronous with observations. In forecasting, skill is highly variable with forecast lead time, as well as with time of year and location. Accordingly, forecasting involves ensembles to represent a range of possible forecast trajectories. Even at coarse model resolutions, it is essential that forecasts are properly calibrated to reduce bias, improve ensemble spread, capture model skill and filter out negative skill, all while improving the representation of temporal, spatial and inter-variable correlations. Downscaling the forecast information to the local domain is a significant further challenge since the calibration of the coarse-scale forecast must be preserved.

Researchers and practitioners alike have recognised partial-calibration methods such as quantile-mapping are deficient for GCM forecast calibration and downscaling, especially for applications in hydrological and crop modelling. Therefore, I developed a new methodology, termed FCMD (forecast calibration – multivariate downscaling) that sought inspiration from current approaches like BCSD (bias correction – spatial disaggregation), and built a workflow to couple full forecast calibration with empirical downscaling approaches. I introduce a modified nearest-neighbour downscaling approach to downscale individual ensemble members temporally (from monthly to daily) and spatially (from regions to points).

FCMD captures GCM skill at aggregated spatial and temporal scales where climate signals tend to be stronger. In this study, the focus was on the direct calibration of rainfall, temperature and solar radiation using BJP, however, FCMD is easily extensible to incorporate prediction of other variables of interest as well as other sources of predictability. FCMD forecasts were calibrated and monthly time scales and were shown to be reliable at both monthly and seasonal time scales.

FCMD downscaled forecasts were demonstrated to have approximately correct temporal, spatial and inter-variable correlations. Indeed, the correlation structures of FCMD forecasts are much

improved compared to raw GCM forecasts and, by extension, forecasts corrected with techniques like mean bias correction and quantile-mapping. The development of calibrated and downscaled forecasts with realistic correlation structures at both monthly and daily time scales through FCMD has the potential to significantly improve the uptake of GCM forecasts in agriculture and water resources management. Work is underway to rigorously evaluate FCMD forecasts for long-range hydrological forecasting and hydrological modelling.

4. Ensemble sugarcane yield forecasting

4.1. Preamble

In Chapter 3, I developed a new methodology to produce calibrated daily forecasts of multiple variables that can be used as inputs to crop models. The final goal of this thesis is to investigate whether properly calibrated forecast ensembles, generated using the methods developed in Chapters 2 and 3, can produce skilful and reliable crop yield forecasts.

In this chapter I address Objective 3: “Assess the utility of post-processed forecasts as inputs to crop models, and evaluate the performance of crop yield forecasts”. The application reported in this chapter is sugarcane yield forecasting in the Tully region of north-east Queensland in Australia. I apply APSIM to generate ensemble biomass forecasts, which are an indicator of total annual yield in the Tully region. As with the previous chapters, I undertake a rigorous probabilistic assessment of the ensemble forecasts.

Previous research has identified the value of long-lead seasonal climate forecasts for the eastern Australia sugar industry, including Tully. I identified Tully as a particularly suitable starting point because sugarcane production in the region is largely reliant on natural rainfall. Other sugarcane producing regions, such as the Burdekin region, for which climate forecasts were explored in Chapter 3, are irrigated. Irrigated regions require a more complex decision framework in the crop model. Investigation of these more complex cases is left to future studies.

The contents of this chapter has been formatted as a journal article and prepared for submission to *Agricultural and Forest Meteorology* (Impact Factor 4.461) with the title “Skilful and reliable forecasts of crop yield using seasonal climate model forecasts” and with authorship: Schepen, A., Y. Everingham and Q.J. Wang. Although the paper is co-authored by my supervisors, I confirm the work is essentially my own. I conducted all of the experiments, analysed the results and wrote all of the paper, including preparing all of the figures. Everingham and Wang helped form the research

questions and provided editorial support. I also acknowledge the provision of an APSIM-sugar model for Tully by Jody Biggs and colleagues at CSIRO.

4.2. Introduction

Prudent use of seasonal climate forecasts has great potential to improve productivity and profitability in agricultural businesses (Klemm and McPherson 2017; Meinke and Stone 2005).

Today, seasonal climate forecasts are commonly produced using outputs from global climate models (GCMs) (Johnson et al. 2018; Kirtman et al. 2014; MacLachlan et al. 2015; Saha et al. 2014). GCM seasonal forecast systems couple physical models of the ocean, atmosphere, land surface and sea-ice. They are very computationally expensive to run and require vast amount of initialisation data. Hence they are typically only run by specialist climate forecast centres. Nevertheless, interest in applying GCM-based climate forecasts in application domains such as agriculture is increasing on the back of evidence that they are beginning to offer skill in excess of longstanding statistical forecasting methods (e.g. Barnston et al. 2012; Rodriguez et al. 2018). However, GCMs are designed mainly to predict global climate patterns, such as the El Niño Southern Oscillation, and their gridded outputs are normally very coarse (for example, 50-100km across), limiting the usefulness of raw model outputs.

The coarse structure of GCMs means they do not reproduce weather and climate statistics for small application domains (e.g. Hagedorn et al. 2005; Tian et al. 2014). However, statistical post-processing may be used to improve GCM forecasts for use in quantitative models such as crop models. The main objectives of forecast post-processing are to produce well-calibrated forecasts and to produce downscaled forecasts. Well-calibrated forecasts have minimal bias, are reliable in ensemble spread and have skill at least as good as climatology (Zhao et al. 2017). The property of forecasts being at least as skilful as climatology has been termed “coherence” (Zhao et al. 2017). Downscaled forecasts have characteristics of a target region at a scale different to the GCM (e.g. a smaller grid cell or a weather station).

For crop forecasting, variables such as rainfall, temperature, solar radiation and evaporation are typically needed as inputs to crop models (e.g. Capa-Morocho et al. 2016; Everingham et al. 2016; Han and Ines 2017; Hansen et al. 2004; Jha et al. 2019). Because multivariate forecast post-processing is by no means straightforward, a method known as quantile-mapping has gained popularity despite its deficiencies for downscaling and calibration being previously reported (Maraun 2013; Zhao et al. 2017). Results from studies applying quantile-mapped GCM forecasts in agricultural applications have been underwhelming. For example, Brown et al. (2018) forced APSIM wheat models with quantile-mapped forecasts from the Predictive Ocean-Atmosphere Model for Australia (POAMA) to predict wheat yield in Australia. The climate forecasts were shown to benefit yield forecast accuracy and narrow the forecast uncertainty range. However, the yield forecasts exhibited a consistent low-yield bias, which was attributed to the inability of quantile-mapping to correct unrealistic autocorrelation structure in rainfall forecasts. Western et al. (2018) also evaluated quantile-mapped POAMA forecasts, in this instance for plant-available soil water (PASW) forecasts. Whilst PASW forecasts were skilful, most of the skill was attributed to initial soil conditions, and rainfall and PET forecasts verified worse than climatology. The skill deficiencies were attributed to the limitations of the quantile-mapping to adequately reproduce ensemble forecasts with realistic spatial and temporal variability. Jha et al. (2019) post-processed CFSv2 forecasts using quantile-mapping for rice yield forecasting in Nepal. It was found that the GCM-driven forecasts performed poorly with respect to climatology-driven forecasts, partly because CFSv2 did not accurately represent the required intra-seasonal variability. CFSv2 was also found to have poor rainfall forecasting skill in the target region, with certain events identified to be problematically over- or under-predicted. Jha et al. (2019) noted the inadequacy of quantile-mapping to overcome these problems and concluded that alternative downscaling methods need to be developed for improved yield prediction.

The poor performance of quantile mapping is directly attributable to its over-simplistic formulation, which cannot guarantee coherence (as defined earlier) nor correct defective covariance structures.

While it can perform well where the raw GCM forecasts are sensible and require only light adjustments to correct bias and/or ensemble spread, widespread application will invariably reveal its shortcomings. For example, in an application to catchment rainfall post-processing, Schepen et al. (2018) (Appendix B of this thesis) identified that while quantile mapping could perform well in some regions, calibration based on the Bayesian joint probability modelling approach (Wang and Robertson 2011; Wang et al. 2009) performed much better overall in terms of skill and reliability over a long period (20+ years) and over a range of catchments in disparate climate zones. Therefore, in this study, we explore an alternative, recently developed technique for multivariate forecast calibration and downscaling (developed in chapter 3 of this thesis), which is based on BJP calibration, as an alternative tool for systematically tailoring climate forecasts for crop model applications. To be clear, the new calibration and downscaling method has not yet been tested in a crop model application.

In the new calibration and downscaling method, climate forecast calibration is undertaken using the Bayesian joint probability modelling approach at a monthly time step. Non-parametric methods are subsequently used to produce realistic daily forecast sequences that fully retain the joint distribution of the calibrated monthly forecasts. The purported benefits of the new approach are that it: (1) embeds a reliable forecast calibration component based on model output statistics; (2) generates a large ensemble to reliably estimate forecast uncertainty; (3) ensures each ensemble member has realistic spatial, temporal and inter-variable covariance; and (4) permits the augmentation of climate forecasts beyond the end of the GCM forecast, which is needed for long-lead time applications.

As proposed in chapter 3, the standard application of the new forecast calibration and downscaling method uses GCM outputs from directly over the target region. However, the Bayesian joint probability modelling component of the methodology is quite general in that it allows the choice of predictors. Therefore, it is also possible to use GCM predictions of remote large-scale climate indices as predictors. The use of GCM climate index forecasts to predict local climate is known as bridging

and has been reported in studies that have found GCMs often have stronger relationships between well-known large-scale climate patterns and observed variables compared to the local GCM forecast of those variables (Hawthorne et al. 2013; Peng et al. 2014; Schepen et al. 2014; Strazzo et al. 2018). Bridging is thus included as a way to demonstrate how sophisticated post-processing methods can make use of a range of GCM outputs and potentially attain more skilful forecasts. Hence two workflows are compared in this study. The standard workflow using local forecasts is referred to as Forecast Calibration – Multivariate Downscaling (FCMD). The alternative workflow using climate indices is referred to as Forecast Bridging – Multivariate Downscaling (FBMD). As the names suggest, the multivariate downscaling part is identical in both workflows.

Sugarcane yield forecasting is the target application in this study. Sugarcane is widely grown in tropical and sub-tropical regions globally, and contributes the bulk of the world's sugar production as well as to other industries such as biofuels. In Australia, sugarcane is grown in coastal regions from northern Queensland to northern New South Wales (Skocaj et al. 2013); a region where steep, complex topography enhances the need to downscale coarse GCM forecasts (Everingham et al. 2015) and where the value of climate information has been well established. For example, Everingham et al. (2003) found that annual sugarcane yield variability in Australia is associated with the Southern Oscillation phase during the late austral spring. El Nino is associated with a higher frequency of below average rainfall years and La Nina is associated with a higher frequency of above average rainfall years. Further to the point, Skocaj and Everingham (2014) reported that in the wet tropics region, excessive winter/spring rainfall negatively impacts on yields. Management of nutrient runoff is also an important consideration for the Australian sugar industry, given the proximity of many farms to the Great Barrier Reef. Accordingly, Kandulu et al. (2018) developed a framework for combining climate and economic variability in nutrient management analyses that could incorporate seasonal climate forecasts in the future.

Seasonal forecasting efforts for the Australian sugar industry have long focused on exploiting knowledge of the El Niño Southern Oscillation state. Targeted long-lead rainfall forecasts have been developed based on the SOI-phases (Everingham et al. 2008) and Niño3.4 indices (Clarke et al. 2010). More recently, An-Vo et al. (2019) linked tercile seasonal climate forecasts (probability of wet, neutral or dry conditions), based ENSO phases, with APSIM to examine sugarcane irrigation planning in the Burdekin region. Resampled historical meteorological sequences related to the tercile probabilities drove the APSIM model to inform an economic analysis. The study found increased profits are possible when seasonal climate forecasts are added into irrigation planning practices. The latter example is also evidence that resampling of historical weather data is sometimes still adopted as an easy downscaling method. An upside to historical data resampling is that the meteorological sequences will have the correct characteristics of observations. However, there is evidence such approaches can perform worse than simply using full historical climatology, such as Rodriguez et al. (2018), who investigated forecasting for sorghum crops.

The foremost objective of this study is to benchmark the skill and reliability of sugarcane yield forecasts in Tully achieved through the FCMD and FBMD approaches. A successful outcome will build the case to trial the new calibration and downscaling methods in other crop forecasting applications. To this end, I undertake a fully probabilistic skill assessment to assess the accuracy and reliability of ensemble forecasts from crop models, similar to what is commonly done in ensemble hydrological forecasting. In the remainder of the chapter, I present the case study details, give an overview of the climate forecast calibration and downscaling methods and describe the relevant ensemble forecast verification methods; followed by detailed results, discussion and conclusions.

4.3. Models and data

4.3.1. Case study location and observed weather data

Observations of rainfall, minimum temperature, maximum temperature and solar radiation are obtained from the Silo dataset (Jeffrey et al. 2001) for the Tully Mill weather station (Silo ID 32042;

latitude -17.9364S, 145.9253E) (Figure 4.1). Silo data is advantageous over raw station data as the record is infilled and interpolated to provide a continuous, high-resolution record from the early 1900s to the present. Tully is located in Australia's Wet Tropics, with average annual rainfall exceeding 4m. Farming activity in the region is thus largely reliant on natural rainfall and irrigation is uncommon.



Figure 4.1: Map of the Tully sugarcane farming district located in north-eastern Australia and the location of the Tully Mill weather station

4.3.2. The APSIM-sugar crop model

I adopt an existing APSIM (version 7.8; Holzworth et al. 2014) sugarcane crop model (APSIM-sugar) previously developed specifically for the Tully region (Kandulu et al. 2018; Thorburn et al. 2011). The model components and parameters remain unaltered in this study. I configure the APSIM-sugar

model to simulate a first ratoon crop grown from September and harvested at the end of August. Starting values for soil water, nitrogen and surface organic matter are reset to the same values after each harvest such that the standard starting conditions are medium soil water, low nitrogen and high organic matter. The cultivar is Q117 grown in Coom soil. Fertiliser (urea) is applied after 10 days at a rate of 150kg/ha. Irrigation is not applied under dryland production rules. Thus the biomass predictions are modified only by the climate inputs.

Simulated yields for each year are obtained by running the crop model with observed meteorological data. The simulated biomass (g/m^2) after 9-12 months growth is correlated with industry-reported yields (tonnes/hectare) with Pearson correlation coefficients in the range of 0.75–0.82. For the purposes of this study, the simulated yields will be treated as the observations for yield forecast verification.

4.3.3. Climate model forecasts

4.3.3.1. Raw GCM forecasts

Raw ECMWF System4 (Sys4; Molteni et al. 2011) historical forecasts of rainfall, minimum temperature, maximum temperature and solar radiation are obtained for the grid cell nearest the Tully Mill for use in the local forecast calibration (FCMD) workflow. I also obtain the raw sea surface temperature forecasts for the Niño3.4 region in the Pacific Ocean (an average over 5°N – 5°S ; 170° – 120°W) for use in the forecast bridging (FBMD) workflow. The Sys4 grid cell over the Tully region is approximately 75km across and lies over both ocean and very steep terrain. Thus the forecasts need post-processing (as described in the next section) to produce ensembles with the correct local variability and extract as much skilful information out of the forecast as possible.

Historical Sys4 forecasts for the next 6 months ahead are obtained for the period 1981–2016. These forecasts are initialised on the 1st day of each month. However, in this study the goal is to produce forecasts 12 months ahead. The augmentation of the forecasts beyond 6 months is explained in the next section on forecast post-processing methods.

4.3.3.2. Post-processing (calibration and downscaling)

The Forecast Calibration – Multivariate Downscaling (FCMD) workflow is described in chapter 3, however, a brief overview is included here. FCMD calibrates and downscales the GCM forecasts with regard to the correlation between raw forecasts and observations and reproduces observed temporal and inter-variable correlations in the ensemble forecasts. FCMD calibrates monthly forecasts using the Bayesian joint probability (BJP) modelling approach (Wang and Robertson 2011; Wang et al. 2009; Wang et al. 2019, to be submitted). Ensemble means of the raw climate forecasts are treated as BJP predictors and the Silo observations are treated as BJP predictands. Realistic inter-variable and month-to-month temporal correlations are established in ensemble members after BJP calibration by using the Schaake Shuffle (Clark et al. 2004), an empirical technique that borrows patterns from template data, usually historical observations. Historical observations are used in this study. A non-parametric disaggregation method downscales the monthly forecasts to daily temporal resolution, preserving inter-variable correlations. The raw GCM forecasts go out 6 months. To extend the FCMD forecasts out to 12 months, the predictors are specified as missing in the BJP models for months 7–12, thus the model essentially reverts to climatological forecasts after month 6. In this application of FCMD, I generated a 200 member ensemble.

In the Forecast Bridging – Multivariate Downscaling (FBMD) workflow, Sys4 forecast Niño3.4 is the sole predictor in the BJP models. Niño3.4 is a large-scale climate index that is commonly used to measure the state of the El-Niño Southern Oscillation. Strongly negative Niño3.4 is indicative of La Niña conditions and strongly positive Niño3.4 is indicative of El Niño conditions. Because SSTs are much more persistent than atmospheric conditions, to extend the forecasts from 7 to 12 months with FBMD, the last SST forecast from month 6 is persisted across months 7–12. Again, a 200 member ensemble is generated.

4.4. Application and verification

4.4.1. Experimental configuration

Yield forecasts are produced for each year from 1982-2016 with the forecast beginning in September of the year prior. At the beginning of September, a 12-month forecast is issued, corresponding to a 12-month growth and harvest cycle. At the beginning of October, an updated forecast is issued, which runs for 11-months to the supposed harvest. In this scenario, the crop model is run with observed data for one month, and then the forecasts are inserted for the remaining months. The sequence continues with forecasts updated each month until August.

4.4.2. Reference forecasts and cross-validation

In addition to FCMD and FBMD forecasts, I also generate a set of climatological reference (CR) forecasts. These are generated by fitting distributions to monthly observed data and downscaling using the multivariate downscaling procedure. It is the same procedure used to augment FCMD forecasts beyond 6 months. They are only forecasts in the sense that they are constrained to the seasonal climatology. There is no input from a GCM. The CR forecasts establish a baseline to help isolate the value added by the GCM forecasts.

Leave-one-year-out cross-validation is applied to generate the FCMD, FBMD and CR forecasts for each year. That is, for each forecast, all data that coincides with the 12 month forecast period is omitted from the fitting of the models, meaning the forecasts are effectively generated out-of-sample for testing purposes. Leave-one-year-out cross-validation is a standard practice for assessment of seasonal climate forecast post-processing methods and their application.

4.4.3. Verification metrics

In this study I verify the key output of the APSIM-sugar crop model, which is biomass. In the subsections that follow, the “observation” is the biomass simulated by APSIM using real weather observations. Biomass is output on a daily basis. It is thus possible to verify forecasts on any day in the 12 month forecast period. I verify the biomass at the end of each month that has forecast input,

giving results in a triangular 12x12 matrix with dimensions of forecast issue month and lead time (months ahead) as per Figure 4.7 in the results. To complete the terminology, the forecast issue month and lead time can be used to calculate the target month.

4.4.3.1. Bias, and reliability in ensemble spread

Bias is calculated as the average mean error. I calculate the relative bias as a percentage:

$$\text{PBIAS} = \frac{\sum_{t=1}^T (\bar{y}_t - o_t)}{\sum_{t=1}^T (o_t)} \times 100 \quad (\%) \quad (42)$$

where \bar{y}_t is the forecast ensemble mean for event t , and o_t is the corresponding observation.

Positive PBIAS indicates systematic over-forecasting whereas negative PBIAS indicates systematic under-forecasting.

Reliable forecasting systems output forecasts with ensemble spreads that are statistically consistent with the distribution of observations. Hence I measure reliability by analysing the uniformity of probability integral transformation (PIT) values for the set of forecast and observations. The PIT for a forecast CDF (F_t) for event t and paired observation (o_t) is

$$\pi_t = F_t(o_t) \quad (43)$$

Plotting sorted PIT values $\pi_{(i)}$ against theoretical uniform quantiles u_i for $i = 1, \dots, T$ yields a uniform probability plot (Renard et al. 2010; Wang and Robertson 2011), which is a visual aid for assessing reliability. The information in PIT uniform probability plots can also be condensed into various metrics (Renard et al. 2010). Given the large number of cases to verify (models, different forecast initialisation and lead times), I calculate the PIT reliability score:

$$\text{REL}_{\text{PIT}} = 1.0 - \frac{2}{T} \sum_{i=1}^T \left| \pi_{(i)} - \frac{i}{T+1} \right| \quad (44)$$

where $\pi_{(i)}$ is the i^{th} ranked PIT value. REL_{PIT} ranges from 0 (worst reliability) to 1 (perfect reliability). The REL_{PIT} quantifies the departure of PIT values from the 1:1 line in a PIT uniform probability plot (where alignment along the 1:1 line indicates perfect reliability).

4.4.3.2. Probabilistic forecast skill

The continuous ranked probability score (CRPS; Matheson and Winkler 1976) is a metric that combines information about the accuracy and reliability of ensemble forecasts. The CRPS for a CDF forecast and corresponding observation for event t is defined as

$$\text{CRPS}_t = \int [F_t(y) - H(y - o_t)]^2 dy \quad (45)$$

with $H(y - o_t) = \begin{cases} 0 & \text{if } y < o_t \\ 1 & \text{if } y \geq o_t \end{cases}$

where F_t is the forecast CDF; o_t is the observed value; and H is the Heaviside step function.

The average CRPS for a set of events is calculated for the test model ($\overline{\text{CRPS}}$) and a reference model ($\overline{\text{CRPS}}_{\text{REF}}$). The relative difference in these average scores are then used to calculate a generalised skill score

$$\text{CRPS}_{\text{SS}} = \frac{\overline{\text{CRPS}}_{\text{ref}} - \overline{\text{CRPS}}}{\overline{\text{CRPS}}_{\text{ref}}} \times 100 \quad (\%) \quad (46)$$

If $\text{CRPS}_{\text{SS}} = 0$, then the models are said to be equally skilful (because the average CRPS score is the same). If $\text{CRPS}_{\text{SS}} = 100$, then the test model is perfect. Negative scores are unbounded and imply worse performance of the test model on average.

4.4.3.3. Relative shift and dispersion

Potgieter et al. (2003) identified reliability, distribution shift and change in dispersion (sharpness) as the three necessary dimensions for probabilistic verification of integrated crop model and climate

forecasts. Therefore, in addition to evaluating the aggregate skill with the CRPS, I analyse the shift and change in the dispersion of the GCM-driven biomass forecasts relative to the climatological-reference-driven forecasts. If the GCM forecasts add value, which may also be seen through the skill metric, then the GCM forecasts should cause a shift in the forecast median and a reduction in dispersion.

For the test and reference forecasts for event t , F_t and $F_{\text{REF},t}$, respectively, the distribution shift for a non-exceedance probability p is defined as

$$S_t(p) = \frac{F_t^{-1}(p) - F_{\text{REF},t}^{-1}(p)}{F_{\text{REF},t}^{-1}(p)} \times 100 \quad (\%) \quad (47)$$

where I choose $p = 0.5$ corresponding to the shift in forecast median. Similarly, the change in dispersion for event t is calculated as

$$D_t(p_1, p_2) = \frac{F_t^{-1}(p_2) - F_t^{-1}(p_1)}{(F_{\text{REF},t}^{-1}(p_2) - (F_{\text{REF},t}^{-1}(p_1)))} \times 100 \quad (\%) \quad (48)$$

where I choose $p_1 = 0.1$ and $p_2 = 0.9$ to measure the change in dispersion of the 10th to 90th percentile range.

4.5. Results

4.5.1. Climate forecast verification

Prior to investigating biomass forecasting skill, examples of climate forecasting skill are presented as a guide to the level of skill available. Skill scores for monthly climate forecasts initialised in September and February are shown in Figure 4.2 and Figure 4.3, respectively. The skill scores show the reduction in error of the GCM-based forecasts (FCMD and FMBD) relative to climatology reference (CR) forecasts. In other words, positive skill scores indicate when the GCM-based forecasts add value over climatology. While climate forecasts are not shown for all initialisation months, it can

be seen later in the biomass forecasting results that the patterns of climate forecasting skill translate reasonably consistently into biomass forecasting skill.

For the September-issued forecasts, FCMD climate forecasts are skilful short lead times, with minimum temperature (Tmin), rainfall (Precip) and solar radiation (Srad) forecasts exhibiting some skill up to three months ahead. The FBMD forecasts have little skill in the first month. However, the skill pattern for forecasts targeting Oct–Feb is very similar to FCMD. The implication is that FCMD is able to source skill for first month from the initial conditions of the local atmosphere, whereas little relationship exists between the first month of Niño3.4 sea surface temperatures and local Tully climate.

There is some slight negative skill even though the BJP calibration in FCMD and FBMD is supposed to return forecasts to climatology and thus ensure coherence. The negative skill is attributable to cross-validation effects. Moreover, compared to the climate forecasting skill for the Burdekin region (Figure 3.3 in chapter 3) the negative skill scores observed here are slightly worse. A plausible explanation that the downscaling here is only to one point, whereas in the Burdekin example, the downscaling was to the average of six locations, which would reduce “noise”. In other words, the downscaling to single locations is more susceptible to spurious correlations in the cross-validation process.

Skill scores for the February-issued forecasts are shown in Figure 4.3. Again the FCMD forecasts show good climate forecasting skill for one month ahead, except for Tmin. Evidence of skill is sparse after the first month. The FBMD forecasts show weak skill in Tmax, Precip and Srad for the first two months. There is also no skill for Tmin in FBMD forecasts. A review of the correlations between ensemble forecast means and observations for the February forecasts (not shown) reveals that the GCM-based forecasts fail to reliably predict a number of very cool Tmin years, limiting the ability to capture forecast skill through BJP post-processing.

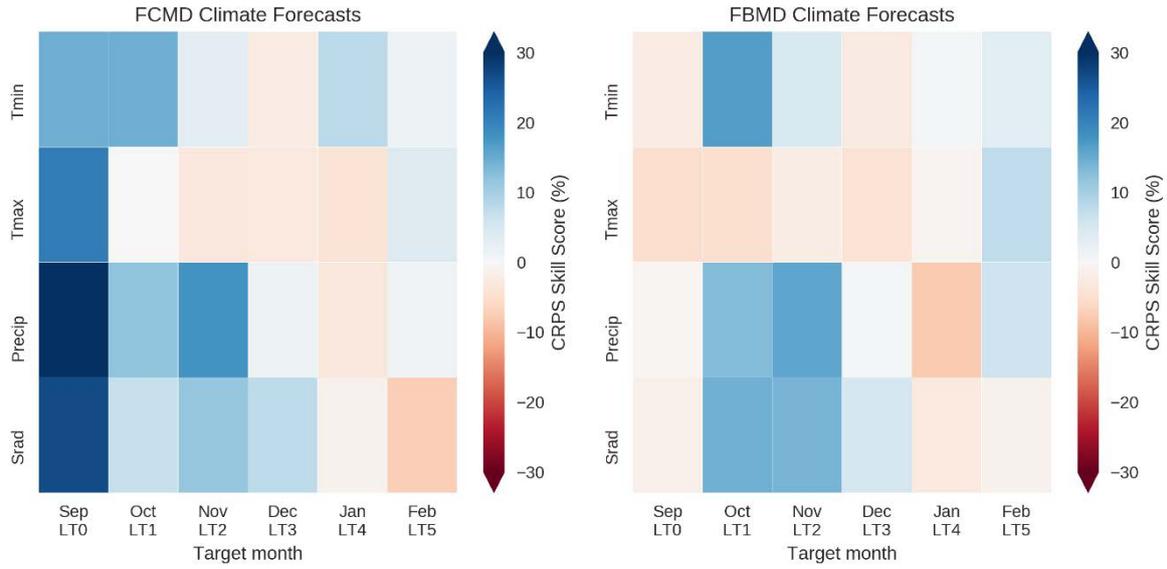


Figure 4.2 CRPS skill scores for the FCMD and FBMD climate forecasts for forecasts issued in September. Skill scores reflect performance relative to the climatology reference (CR) forecasts for the period 1981–2015, and positive values indicate superior performance through greater accuracy and/or reliability.

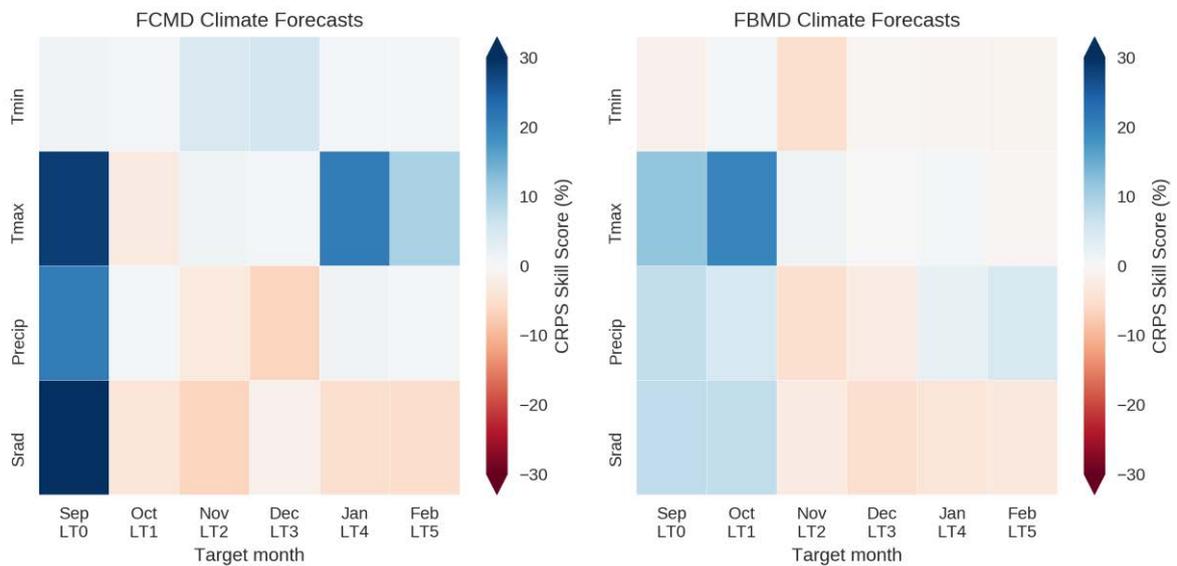


Figure 4.3 As for Figure 4.2 except for forecasts issued at the start of February for the period 1982–2016. Crop growth is simulated using observed meteorological forcing from September to the end of January with forecasts applied from February.

4.5.2. Detailed verification for SON biomass forecasts

Prior to verifying biomass forecasts for all initialisation dates and lead times, I present a detailed verification of the biomass after three-month growth from September to November, which serves the dual purpose of presenting example forecasts. Figure 4.4 presents the biomass forecasts from each model (FCMD, FBMD and CR) for each year from 1982-2016. These forecasts correspond to using inputs Oct, Nov and Dec forecasts for which skill is shown in Figure 4.2. The forecast distributions are presented as boxplots (see caption for details) and the simulated “observed” biomass is plotted as a dot.

Visual inspection shows that the FCMD and FBMD yield forecasts behave in a similar manner, which is consistent with the dominant influence of ENSO in the period. The FCMD and FBMD yield forecasts do exhibit shifts and changes in dispersion relative to the CR forecasts, indicating the GCM forecasts are adding value. In contrast, the CR forecasts remain largely constant from year-to-year, with the minor shifts seen most likely due to cross-validation. It is also noted that the CR forecasts provide a sensible climatology, as may be determined through a visual comparison with observations.

In quantitative terms, the percentage bias in the FCMD and FBMD biomass forecasts are -0.9% and -3.1%, respectively. The CRPS skill scores are 28.2% and 23.9% respectively. The average absolute shifts in the forecast medians away from the CR median are 23.8% and 23.7%. Lastly, the relative widths of the 10th to 90th percentile range are 84.8% and 85.5%. While FCMD biomass forecasts are slightly better in terms of accuracy, the FBMD biomass forecasts “catch-up” remarkably despite having little skill in the first month, which may be associated with FBMD having better forecasting skill for rainfall and solar radiation in Oct and Nov (Figure 4.2). Overall, the shift and dispersion results are nearly identical. Overall, the uncertainty of the seasonal biomass forecasts is quite high, which is commensurate with the generally low skill of seasonal climate forecasts. Nevertheless, a review of the most extreme simulated yields, a high in 2003 and a low in 2011, shows that the

forecast distributions shift in the expected directions. The 2011 event is notable, as record low yields were achieved due to unusually high rainfall. To gain a better understanding of the evolution of forecast skill and how it varies with lead time, I will assess the skill metrics for every release month and month ahead in the next section.

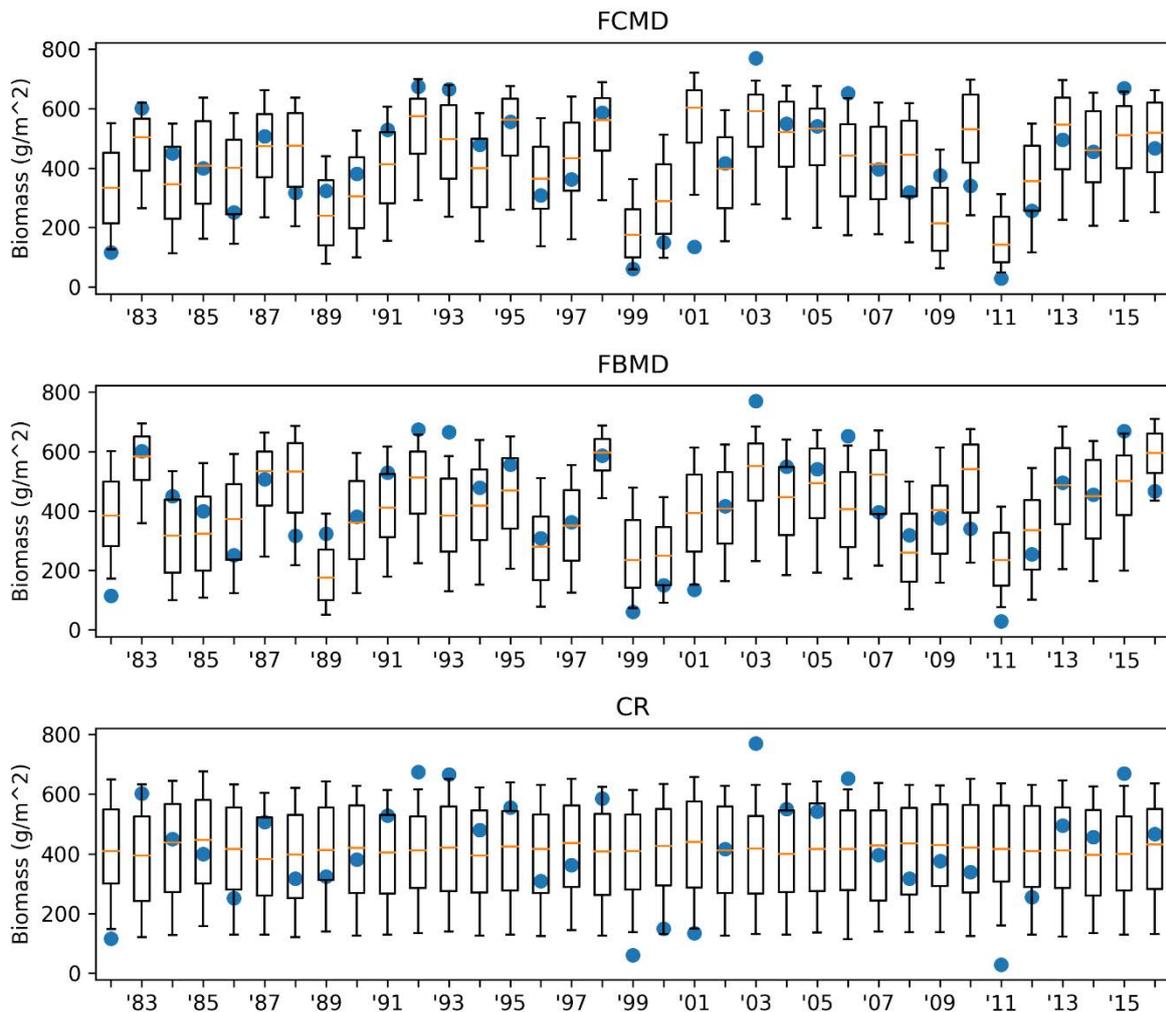


Figure 4.4: Boxplots summarising the biomass forecast distributions at the end of November, for forecasts issued at the beginning of September, for each year 1982–2016. The box covers the interquartile range and the whiskers cover the [0.1, 0.9] quantile range. The blue dot is the simulated biomass from observed meteorological data. FCMD and FBMD are GCM-driven forecasts and CR is the climatological reference-driven forecast. Leave-one-year cross-validation has been applied.

4.5.3. Overall results

Percentage bias in biomass forecasts is summarised in Figure 4.5 for all forecast issue months and lead times (target months). Bias in the biomass forecasts at the time of harvest is typically small and normally within 1%. However, some small negative biases with a magnitude of up to 5% can occur at various points in the forecast. For context, the bias of the FBMD forecasts in Figure 4.4 is -3.1%, which is although one of the “worst” cases, virtually unnoticeable in Figure 4.4. An inspection of bias by forecast initialisation month and lead time (not shown) reveals that the percentage biases are greatest when the crop is very small and the crop model has had limited exposure to observed data. Bias is always very near 0% for a mature crop.

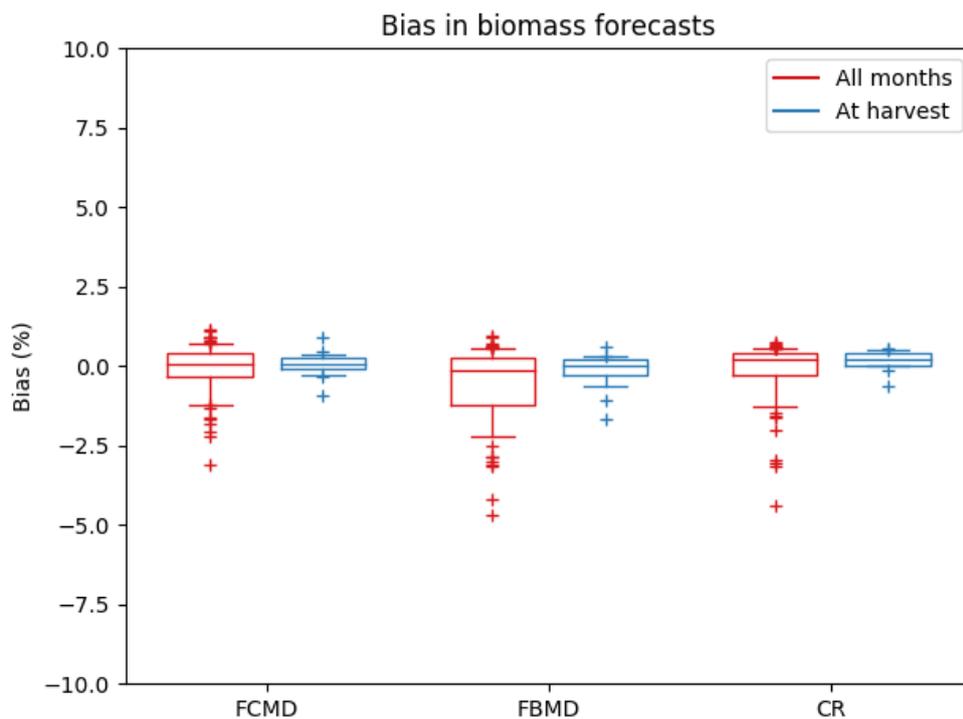


Figure 4.5: Boxplots of the percentage bias in the biomass forecasts from each meteorological-forcing data set. The red (leftmost) boxplot for each model summarises the bias for all forecast release months and target months (N=78). The blue (rightmost) boxplot in each case is for the harvest forecasts from each release month (N=12). The boxes are the interquartile range with the line across marking the median. The whiskers are the [0.1, 0.9] quantile range. The markers are all other cases.

Reliability scores for biomass forecasts are summarised in Figure 4.6 for all forecast issues months and lead times. Biomass forecasts from all three sets of climate forecasts (FCMD, FBMD and CR) are similarly reliable. The biomass forecasts are highly reliable with the PIT_{REL} values normally exceeding 0.9, which indicates that the PIT values closely follow a uniform distribution. Another interpretation is that the forecast ensemble spreads are neither too wide or too narrow.

CRPS skill scores for FCMD biomass forecasts are presented in Figure 4.7 for each forecast initialisation month and lead time. (Skill scores are 0 up to the forecast release month because observed data is used in all cases.) Recall that the skill is measured relative to the CR forecasts, and thus isolates the skill added by the GCM. The skill scores for FCMD biomass are positive in the vast majority of cases, with values typically in the range of 0-30%. FCMD skill is greatest for forecasts issued between September and January. The skill attained at short-lead times persists through to the harvest yield. Beyond February, the FCMD forecasts add less value over climatology. Nevertheless, the skill scores are positive, especially for short-lead times.

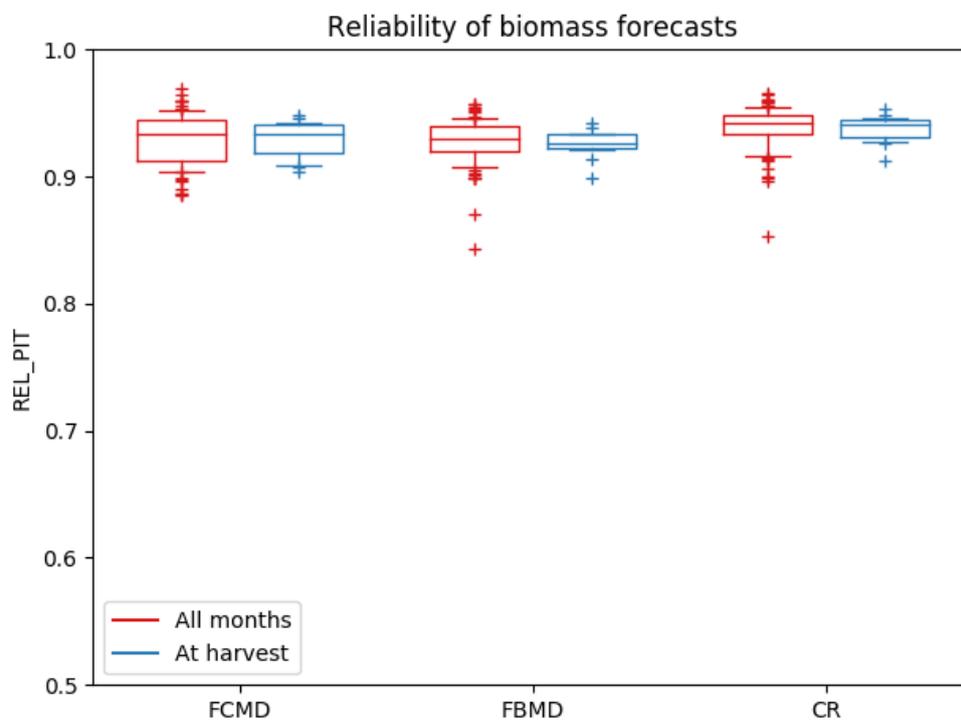


Figure 4.6: As for Figure 2, except for the PIT reliability metric. The scale of the PIT metric is [0, 1]; however, the y-axis is limited to [0.5, 1] for clarity.

The corresponding CRPS skill scores for FBMD forecasts are presented in Figure 4.8. The skill of the FBMD forecasts is overall lower than the FCMD forecasts, especially for forecasts issued from March onwards. However, for certain cases, e.g. for longer-lead times of October-issued forecasts, the skill of FBMD forecasts is higher than FCMD. Further discussion on the skill differences and opportunities are discussed in section 4.6.

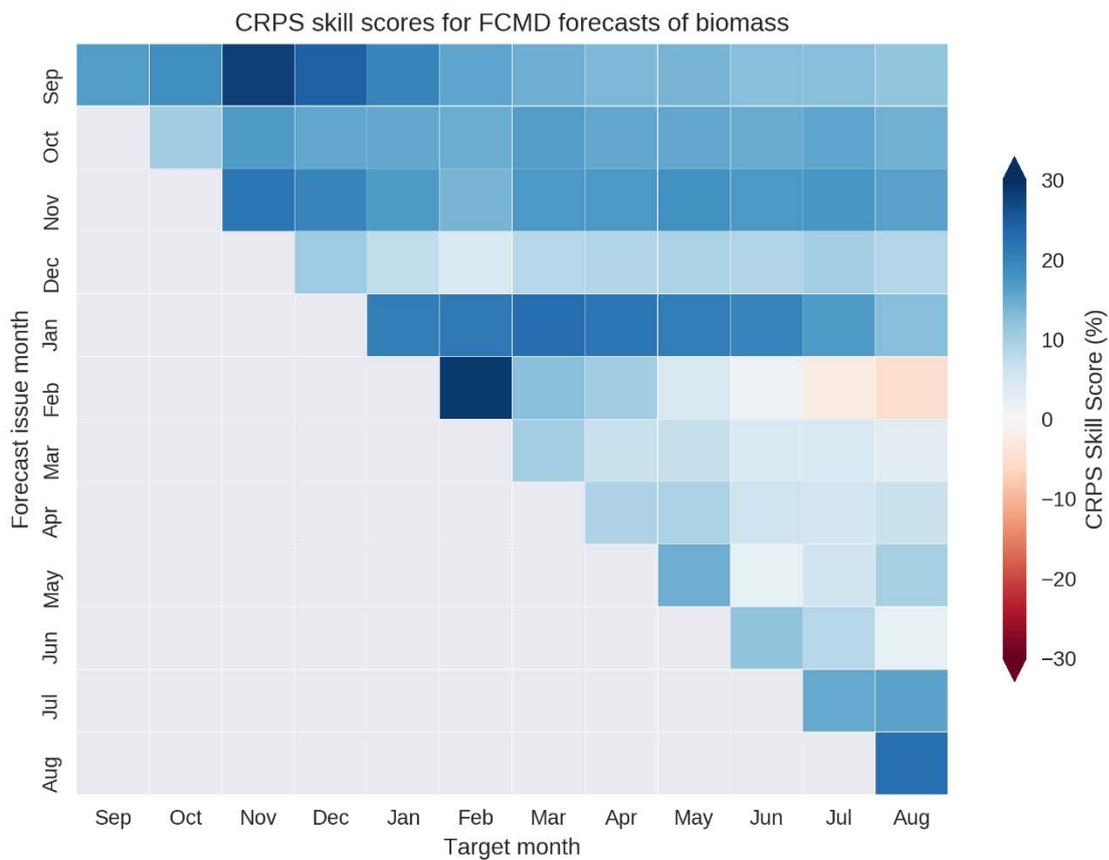


Figure 4.7: CRPS skill scores for the FCMD-driven biomass forecasts for each forecast release and target month. Skill scores reflect performance relative to the climatological reference (CR) forecasts for the period 1982–2016, and positive values indicate superior performance through greater accuracy and/or reliability. The crop model is run with observed data from 1 September up to the beginning of the release month, hence the skill scores are available for the target months equal to the release month and beyond.

The average shift in forecast median and average relative dispersion of the GCM-based forecasts, relative to the CR forecasts, are compared for each forecast initialisation month and target month in Figure 4.9. The average shift in forecast median is similar between the FCMD and FBMD biomass forecasts. The average shift can be up to 25% but is typically less than 10%. Individual forecasts do have shifts of 50% or more (not shown). For a proportion of the samples, the FBMD forecasts exhibit a slightly stronger shift on average, by a few percent. The dispersion of the GCM-based forecasts is on average less than the dispersion of the CR forecasts. The relative dispersion of the GCM-forecasts is typically 90-100% of the dispersion of the CR forecast, however, in some instances it is much lower, with the FCMD forecasts being 75% narrower for a few cases.

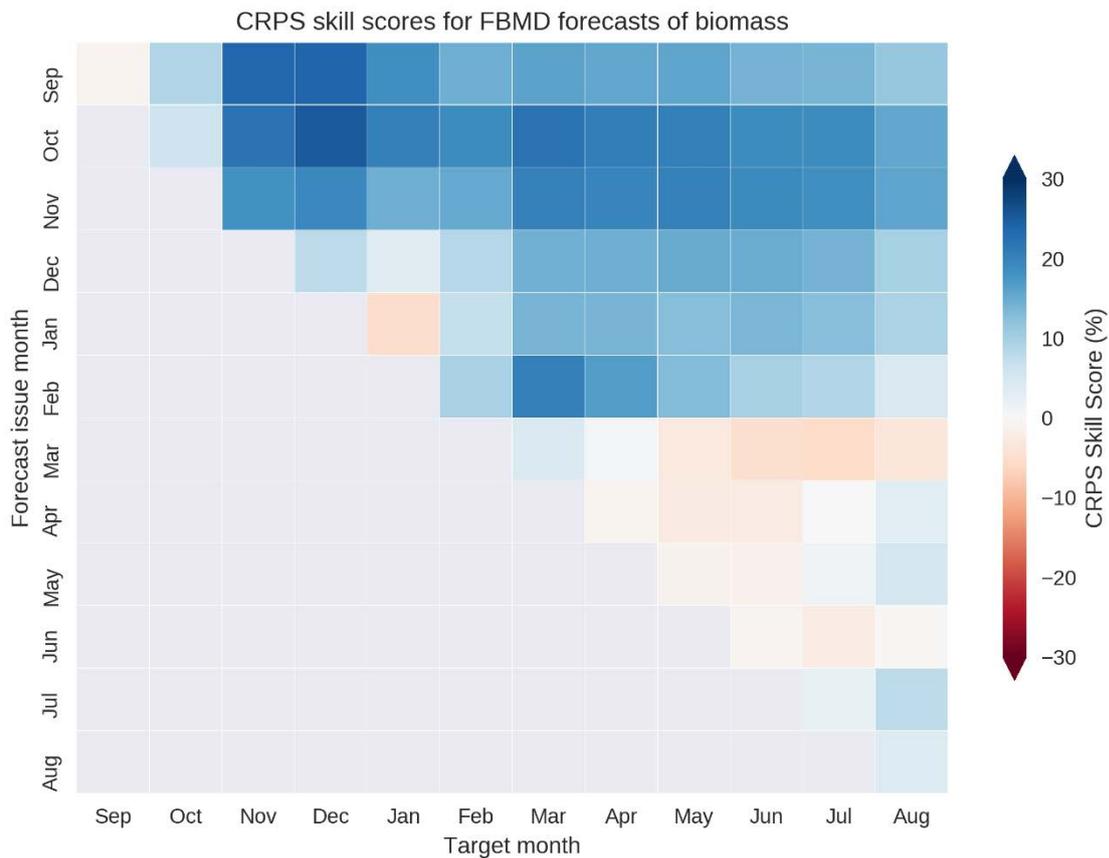


Figure 4.8: As for Figure 4.7, except for FBMD forecasts.

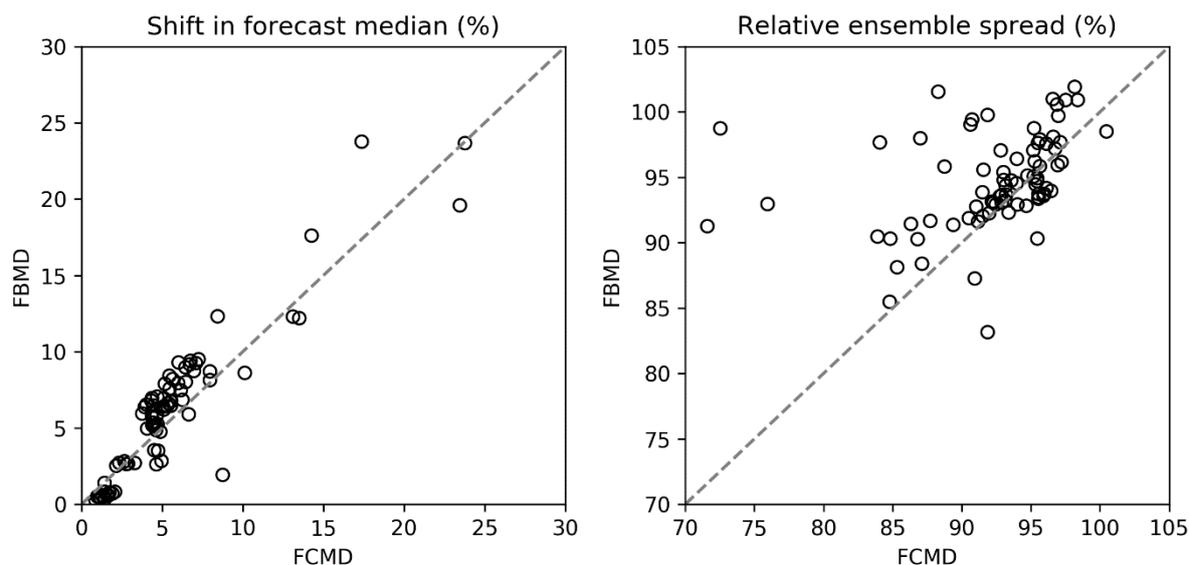


Figure 4.9: Left panel: Average shift in the forecast median of GCM-driven biomass forecasts, relative the CR-driven biomass forecasts, for each forecast release month and target month. The average is taken over the 35 verification years 1982–2016. Right Panel: As left panel, except for average relative dispersion, where dispersion is measured as the [0.1, 0.9] quantile range of the forecast.

4.6. Discussion

Our study examines the viability of more sophisticated forecast calibration and downscaling methodologies (FCMD and FBMD) that may support the uptake of GCM forecasts in seasonal crop forecasting activities. The results of this study show that Tully sugarcane biomass forecasts are essentially unbiased and very reliable from very short to very long lead times. However, it will be very important to validate the methodologies for other crop forecasting applications. Sensitivity to the input climate variables will vary for different crops, providing unique challenges.

The purpose of this study is to evaluate and compare the skill and reliability of different post-processed seasonal climate forecasts in an integrated crop model application. The skill metrics presented isolate the effects of the GCM forecasts. If the biomass forecasts were to be evaluated against a climatology of all historical biomasses, the skill scores would be much higher. In other words, more accurate predictions of biomass are naturally made as the season progresses.

As described in section 4.3.2, management options were held fixed in the APSIM-sugar model. It would be valuable in future work to apply the FCMD forecasts in other APSIM-sugar applications

that include different management options. Appropriate examples are the nutrient management study of Kandulu et al. (2018) or the irrigation scheduling problem studied by An-Vo et al. (2019). On the topic of improving the simulation configuration, I suggest that the crop models could be better “warmed up” in real-time applications by running observed meteorological sequences for a longer period, e.g. 12 months, prior to the first forecast. For this rainfed system, I expect a longer warm-up could slightly improve the overall performance of the long lead time forecasts. Moreover, allowance for real initial conditions in terms of soil water, nitrogen levels, organic matter and type of crop (e.g., cultivar, plant vs ratoon) would give more accurate biomass predictions.

The results show that forecasts based on ENSO only, i.e., FBMD forecasts, have overall lower skill than forecasts based on local raw GCM forecasts (Figure 4.7 and Figure 4.8). This is informative, given that it remains very common to use ENSO-pattern-based forecasts for agricultural applications in Australia, including for the sugar industry. Of particular note is that for forecasts issued from March onwards, the skill for FBMD forecasts largely vanishes, which means the forecasts based on GCM Niño3.4 forecasts do not add value over climatology. The result is consistent with the well-known (boreal) “spring predictability barrier” but, more specifically, the fast decay of ENSO prediction skill previously seen in coupled climate models (Jin et al. 2008). That said, the results show that a local forecast calibration, i.e. FCMD, can produce moderately skilful forecasts during this period, which is promising not only for sugarcane applications but for (austral) winter crops like wheat.

Furthermore, there are certain instances, e.g., forecasts issued in Oct, Nov and Dec, where the skill of FBMD yield forecasts exceeds the skill of FCMD yield forecasts. These results suggest that an optimal combination of the FBMD and FCMD forecasts could provide the best overall forecasts. Merging calibrated and “bridged” forecasts has previously been shown to be effective for seasonal climate forecasts (Peng et al. 2014; Schepen and Wang 2014; Schepen et al. 2014; Schepen et al.

2016; Strazzo et al. 2018). Alternatively, merging could be applied to the biomass forecasts directly, as has been done for seasonal streamflow forecasts (e.g., Schepen and Wang 2015).

As described in section 4.3.2, simulations of biomass using observed meteorological inputs are treated as the true observations for forecast verification purposes. While the simulated biomass is reasonably well correlated with industry-reported yields for the Tully region, further post-processing is required to transform the APSIM biomass forecast distributions to actual regional-yield forecast distributions. That said, if yield forecasting is the last quantitative modelling step required for decision support, then probabilistic ensemble biomass forecasts will remain informative in terms of guidance on the likelihood of above average or below average yields.

Initiatives to further improve full-calibration methods, such as using additional predictors, using all forecast ensemble members, or incorporating trend in the model are currently receiving attention in other studies.

4.7. Conclusion

Seasonal climate forecasts from global climate models can help farmers make decisions with increased confidence, especially by connecting the climate forecasts with crop models and other agricultural decision support tools. Coarsely-gridded GCM forecasts are biased and unsuitable for direct use in crop models. Statistical post-processing of GCM forecasts is needed to harness skill and produce revised forecasts ensembles that are more suitable for use in applications models.

In this study, we investigate forecasting sugarcane biomass in north-eastern Australian. For GCM forecast post-processing, we adapt a newly-developed approach for calibration and downscaling multivariate forecasts. In particular, the calibration and downscaling approach ensures forecasts are coherent (forecasts typically no worse than climatology) and instils realistic temporal and inter-variable correlations in the forecasts, which is crucial for forecast reliability. The post-processing outputs ensemble rainfall, temperature and solar radiation forecasts at a daily time step.

Two variations of the calibration and downscaling workflow were investigated. In the first option, which is largely the original method, localised forecasts of rainfall, temperature and radiation over the target region were calibrated against observed data using a statistical Bayesian joint probability model on a monthly time step. In the second option, the model was adapted to use a large-scale climate pattern (Niño3.4, representing Pacific Ocean Sea Surface Temperature anomalies) to predict the local climate. Both approaches produce skilful monthly forecasts, although the direct forecast calibration tends to be more skilful overall.

The monthly climate forecasts (from both workflows) were disaggregated to a daily time step and used to drive an APSIM-sugar model to predict biomass up to 12 months ahead. A rigorous probabilistic forecast evaluation was undertaken. The APSIM-generated biomass forecasts are virtually unbiased and are skilful for many combinations of forecast initialisation month and lead time. The biomass forecasts produced using the new workflows are also reliable in the statistical sense, allowing for trustworthy forecast probabilities to be derived from the ensemble. In line with the climate-forecast evaluation, the locally-calibrated climate forecasts lead to greater coverage of skill for biomass forecasting across the matrix of issue months and lead times. However, better performance is seen for Niño3.4-based forecasts initialised during the austral spring, especially at longer lead times. Thus, optimal sugarcane biomass forecasting skill in north-eastern Australia is potentially achieved through the combination of locally-calibrated forecasts and forecasts based on large-scale climate patterns.

The results of this study show promise for sophisticated forecast calibration and downscaling workflows to support the connection of seasonal GCM climate forecasts and crop models. However, it will be important to trial the workflows for other crop forecast applications such as grains.

5. Thesis conclusion

5.1. Preamble

In this final chapter of my thesis, I summarise my progress against the three objectives originally set out in Chapter 1. A discussion of results, conclusions and limitations are interweaved in the following objective summaries. Implications and personal learnings are considered in concluding remarks. To finish, I set out possible follow-up research projects in the future directions section.

5.2. Objective 1 summary

Develop and rigorously evaluate BJP-based methods to calibrate monthly and seasonal GCM forecasts of climate variables needed by crop models

The Bayesian joint probability modelling approach (BJP) has previously been leveraged to post-process GCM forecasts of rainfall at GCM grid scales and to downscale forecasts to hydrological catchments. For crop forecasting applications, additional climate variables such as minimum and maximum daily temperatures and solar radiation are needed. Therefore, it was necessary to extend BJP post-processing to coherently post-process forecasts of multiple variables. My original contributions associated with objective 1 are highlighted in Figure 5.1.

As the first step in pursuing objective 1, I investigated multivariate BJP post-processing for seasonal (three-month) forecasts at continental spatial scales. BJP was firstly applied in a univariate sense for calibrating the additional variables needed by crop models. The extension of BJP to new variables required the selection of an appropriate transformation for each variable, to allow the use of a transformed multivariate normal distribution model, and the development of methods to robustly infer the transformation parameters. The transformation step also included defining a censoring threshold for bounded variables as appropriate (e.g. a lower bound of 0 for rainfall, solar radiation and evaporation).

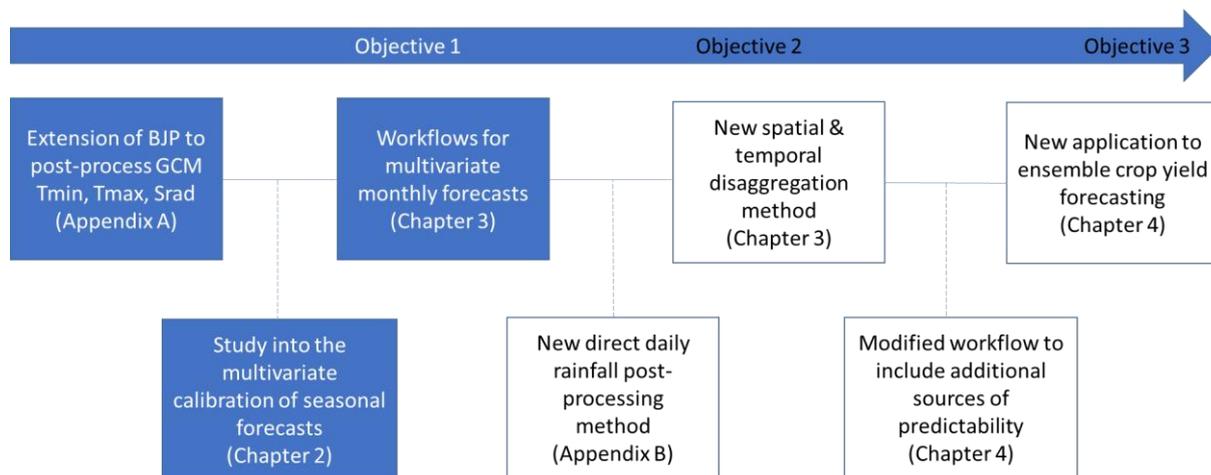


Figure 5.1 Original contributions associated with thesis objective 1 highlighted in blue boxes.

The log-sinh transformation was used to transform rainfall, for which it is specially designed. The Yeo-Johnson transformation was adopted for all other variables. A procedure for the reliable Bayesian maximum a posteriori estimation of Yeo-Johnson transformation parameters was devised. The procedure requires first rescaling the data to a common range, allowing the specification of general prior distributions for the transformation parameters regardless of the distribution of the original variable. (A similar procedure was developed for the log-sinh transformation in work outside of this thesis.) Preliminary results of BJP post-processing for minimum and maximum temperature forecasts were published in the paper attached as Appendix A.

After establishing BJP model for univariate forecast calibration, I investigated the suitability of BJP for simultaneous calibration of rainfall, minimum temperature and maximum temperature (Chapter 2). In low-dimensional settings, for example, for a small number of variables at one grid point and for one time period, BJP has the capability to post-process multivariate forecasts using a single model because the covariance matrix acts as an explicit model of inter-variable covariance. However, in the typical framework where seasonal forecasts are calibrated for each initialisation month separately, the number of data points available, 35 in cross-validation this case, is small. Therefore, I suspected that overfitting (i.e., fitting noise in the data) was a risk, which could reduce the performance for

independent predictions. Based on this understanding, I developed three strategies for calibrating seasonal GCM forecasts of rainfall, minimum temperature and maximum temperature: (1) simultaneous calibration of all the variables using BJP; (2) univariate BJP calibration coupled with empirical ensemble reordering (the Schaake Shuffle), which restores multivariate dependence from observations; and (3) transformation-based quantile-mapping, which inherits multivariate dependence from the GCM fields.

The third option (transformation-based quantile-mapping) is included because, despite its theoretical drawbacks, quantile-mapping remains a popular choice for post-processing GCM forecasts; a fact that virtually demands that it be included in any comparison. I note that the transformation-based quantile-mapping is a novel contribution arising from this thesis. I developed it so that the marginal distributions of each climate variable is modelled in the same way as BJP, enabling fairer comparisons than if a third-party package was used.

The three forecast calibration strategies were applied to Australian seasonal forecasts from the ECMWF System4 model at 0.75x0.75-degree (~80km) resolution across the whole continent, for all 12 overlapping seasons JFM, FMA,...,DJF, for the years 1981-2016. I applied a leave-one-year-out cross-validation procedure to ensure that the forecast performance was not overly inflated by “artificial skill”, which arises when the same data is used for model training and testing. Computer code was assembled in a mix of C/C++ and Python to obtain a balance between execution speed and ease of operation. Linux-based cluster computing was required to complete the continental-scale experiments.

A suite of univariate and multivariate forecast verifications metrics were applied to compare the performance of the post-processed forecasts. Multivariate forecast verification metrics, such as the energy score and variogram score, have been applied at weather time scales, however, they usually consider the same variable in space and/or time. In my research, the application of the energy score and the variogram score required the development of a standardisation strategy, to make the errors

of forecasts for variables with different units more similar, thus making the interpretation meaningful. I found the strengths and weakness of the energy and variogram scores were in accordance with previous studies. The energy score appears to be a good aggregate measure of the calibration of the marginal distributions and the variogram score adds unique information by granting deeper insight into the quality of the calibration with respect to the multivariate dependencies.

Univariate BJP calibration paired with Schaake Shuffle empirical ensemble reordering (abbreviated to UBJP+SS) is found to perform best in terms of univariate and multivariate forecast verification metrics for calibrating Sys4 forecasts of rainfall, average minimum daily temperature and average daily maximum temperature. However, the performance of empirical ensemble reordering using the Schaake Shuffle is dependent on the selection of historical data used to construct the dependence template. The discussion in Chapter 2 refers to studies of conditional implementations of the Schaake Shuffle (preferential selection of dates in other words) that can reportedly improve outcomes. Conditional Schaake Shuffles were not explored in my thesis because the high uncertainty in seasonal forecasts implies sequences representing a wide range of climatological possibilities are required. A time window for selection of Schaake Shuffle dates is already needed in order to obtain enough sequences to shuffle the 200 member ensembles and preferential selection of dates would require widening the search for dates in other ways, e.g. by searching further in space.

Direct multivariate calibration (MBJP) performed as the second-best method even though a priori it may be expected to outperform UBJP+SS by virtue of explicitly modelling correlations. To understand the result, the calibration experiments were repeated without cross-validation. MBJP does perform better when cross-validation is not applied, however, this advantage vanishes once proper cross-validation is applied. This result is evidence the overfitting is a problem with the MBJP model. MBJP may perform better in settings where more data points are available. The continental-scale forecast calibration and verification exercise highlighted the overall better performance of BJP-

based calibrations compared to quantile-mapping, with both UBJP+SS and MBJP outperforming transformation-based quantile-mapping (TQM).

5.3. Objective 2 summary

Develop and evaluate methods for downscaling forecasts to high spatial and temporal resolution as needed by crop models

The results from the work on objective 1 suggest a combination of BJP calibration and empirical ensemble reordering is suitable for post-processing multivariate GCM forecasts at coarse spatial and temporal scales. The follow-up challenge, forming the basis for objective 2, is to extend the post-processing methods to provide daily meteorological inputs for use in crop models.

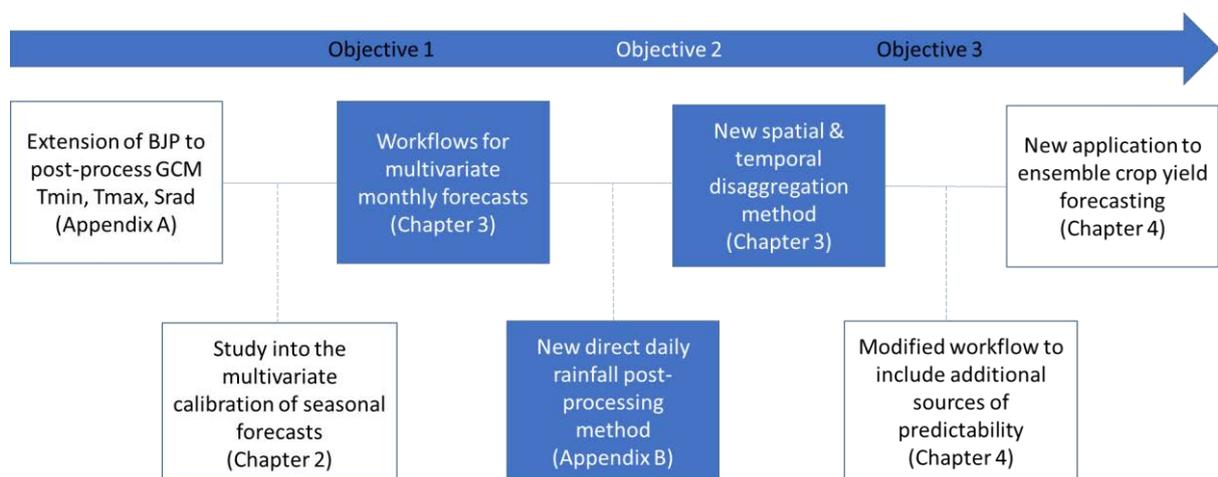


Figure 5.2 Original contributions associated with thesis objective 2 highlighted in blue boxes.

Figure 5.2 highlights my original contributions made in meeting objective 2. The relevant studies and results are reported in chapter 3 and Appendix B. The studies reported in Chapter 3 can be considered in two parts: (1) generation of coherent rainfall, average minimum daily temperature, average maximum daily temperature and solar radiation forecasts at monthly time steps; and (2) simultaneous spatial and temporal downscaling/disaggregation to obtain coherent daily forecast sequences at

multiple sites. Combined, parts (1) and (2) form the new FCMD (Forecast-calibration multivariate-downscaling) methodology.

For part (1), forecast calibration, BJP is applied to calibrate forecasts that have been spatially aggregated in a manner that is entirely consistent with the planned spatial disaggregation. For example, if there are two locations, forecasts are calibrated on derived datasets representing an average over the two locations. Separate BJP models are established for each month ahead. I note that while parametric methods such as BJP may be used to produce correlated forecasts over multiple lead times or at multiple sites, the dimensionality incurred can be too high for parametric modelling to be feasible. Therefore, my solution for monthly forecast calibration is to apply BJP to calibrate forecasts for each variable and month independently with the Schaake Shuffle subsequently applied as a practical means to connect up the ensemble members across variables and over many months ahead. The result is a continuous, multivariate, ensemble forecast at monthly time steps of length equal to the raw GCM forecast.

For part (2), multivariate downscaling, I initially investigated the suitability of a range of existing tools to simulate daily weather conditional on a monthly forecast; weather generators, mainly. However, the prospects of success with weather generators appeared dim in light of the preference to maintain the joint distribution of the calibrated forecasts across lead times. The main limitations of weather generator type approaches are restrictions on the number or type of variables modelled and/or the number of locations that can be modelling simultaneously. In other words, weather generators also become impractical to apply for high-dimension problems.

Ultimately I devised a new empirical multivariate downscaling method, which combines elements of a nearest-neighbour search (to find the most similar forecast in the past at the monthly time scale) and the method of fragments (to simultaneously disaggregate the forecast spatially and temporally). The nearest-neighbour search explored a multidimensional Euclidean-distance space to identify a suitable historical pattern for disaggregation of the multivariate forecast. To ensure a fair balance

amongst the variables in finding a nearest neighbour, the search was conducted with standardised data.

The multivariate downscaling method is applied independently for each month in the forecast. As an empirical method, it projects patterns from historical data onto new forecasts, and thus requires a substantial historical database of historical daily weather data. The data requirements for the multivariate downscaling are more onerous than for the empirical ensemble reordering. To build up more samples for disaggregation, I allowed rolling aggregates of daily data to be used in the search, rather than strictly using calendar month data. However, the search remained restricted to within a month either side of the target month to ensure that the selected weather patterns are seasonally appropriate. The multivariate downscaling method is a pragmatic solution that retains the distribution of the calibrated forecasts at the aggregated spatial and temporal scale and ensures that each ensemble member has daily forecasts with realistic temporal, spatial and inter-variable correlation structures.

In chapter 3, the methods from part (1) and part (2) were combined into the Forecast Calibration – Multivariate Downscaling (FCMD) method. FCMD was tested in the Burdekin region of Australia, producing daily forecasts of rainfall, minimum temperature, maximum temperature and solar radiation for 6 stations intersecting 3 GCM grid cells. The Burdekin region was selected as seasonal forecasts are known to be skilful in the region (e.g., Figure 2.3) and the test is to harness skill rather than produce it. Moreover, calibrated forecasts in a skilful region will vary from year-to-year, which provides a more stringent test of disaggregation methods, compared to an unskilful region, where each forecast would approximate climatology. As with earlier work, I used the ECMWF System4 climate model to obtain raw GCM forecasts. FCMD was applied for forecasts 6-months ahead. Analyses of spatial, temporal and inter-variable (Kendall) correlations demonstrated that FCMD worked as designed, in the sense that it captured skill at seasonal time scales and produced forecasts with daily statistics similar to observations. For rainfall, the relative frequency of wet days was

reproduced with high accuracy, as was the distribution of daily rainfall amounts, which demonstrates that the FCMD method resolves the oft-cited problem of GCMs overestimating rainfall frequency and underestimating rainfall intensity.

Additional work on objective 2 is reported in chapter 4 where FCMD ensemble forecasts are used as inputs to a crop model. The crop forecasting application actually requires forecasts up to 12 months ahead. To achieve this goal, FCMD was run with no GCM input beyond 6 months, effectively reverting to climatological forecasts. The modification seamlessly produced 12 months of forecasts whilst making use of GCM forecast information when it was available.

At the same time, I devised another method for producing downscaled forecasts. Instead of using BJP to calibrate the local forecasts of the four climate variables, the GCM forecast of Niño3.4, a climate index used to monitor the El Niño Southern Oscillation, was used a predictor in the BJP models instead. The modified approach is named FBMD (Forecast Bridging – Multivariate Calibration). Because SSTs are generally more persistent than the atmosphere, the last Niño3.4 forecast was persisted in FBMD to generate the 7–12 month ahead forecasts rather than simply appending climatology as in FCMD. While FCMD produced the most skilful results overall, there is evidence that FBMD can augment forecast skill in some seasons. The FCMD and FBMD methods have several distinguishing features that theoretically make them suitable for general crop forecasting applications. These are: (1) reliable forecast generation using BJP with flexibility to choose a range of predictors and predictands; (2) generation of a large ensemble to reliably estimate forecast uncertainty; (3) production of ensembles with realistic spatial, temporal and inter-variable covariance; and (4) ability to augment GCM forecasts many months ahead.

5.4. Objective 3 summary

Assess the utility of post-processed forecasts as inputs to crop models, and evaluate the performance of crop yield forecasts

The final objective of this thesis was to evaluate the performance of the post-processed GCM forecasts in a crop model application, thus demonstrating an integrated solution for GCM-driven crop forecasting. Figure 5.3 highlights the final original contribution to meet objective 3.

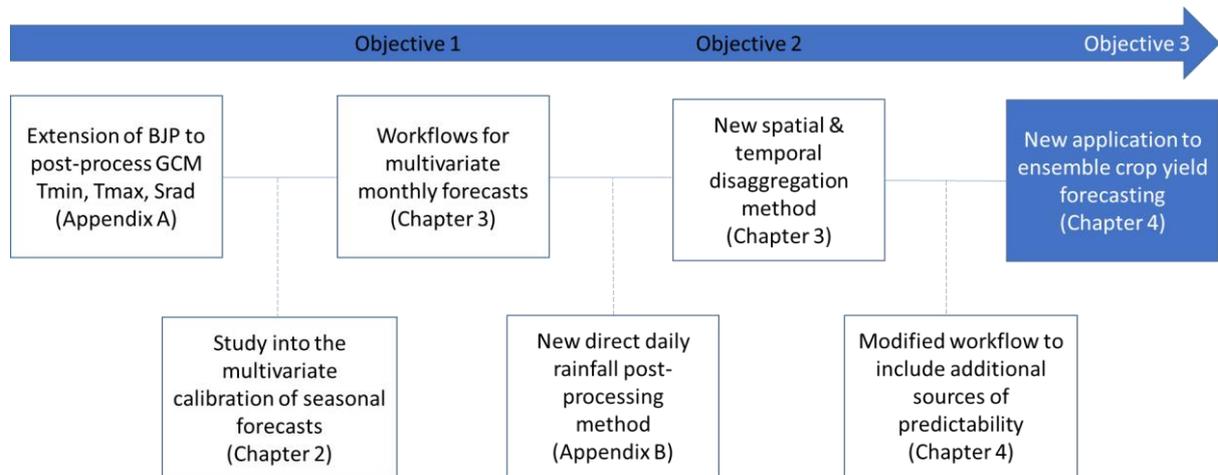


Figure 5.3 Original contribution associated with thesis objective 3 highlighted in blue boxes.

Chapter 4 examines the application of FCMD and FBMD climate forecasts to generate crop model forecasts up to 12 months ahead. Forecasts of daily rainfall, minimum and maximum temperature, and solar radiation were used to drive an APSIM model thus producing ensemble biomass forecasts. The biomass forecasts were evaluated using ensemble verification techniques very similar to those applied to climate forecasts in earlier chapters.

APSIM models have been developed for a wide range of crops and can be extensively configured to execute an array of management options. For this thesis, I selected a single APSIM application for sugarcane yield forecasting in the Tully region of Australia. The choice is motivated by previous research, which has identified the value of long-lead seasonal climate forecasts for the eastern Australia sugar industry, but which has not made use of GCM forecasts. The Tully region is also suitable as starting point for the application of the new GCM forecasts as sugarcane production there is largely without irrigation as an external factor. Downscaled forecasts were already produced

for the Burdekin region in Chapter 3, however, the region is irrigated, and thus requires a more complex decision framework in the crop model, and investigation is left to future studies.

An APSIM-sugar model developed specifically for the Tully region was supplied by CSIRO. I have assumed that the APSIM model is based on expert knowledge and, therefore, I did not attempt to fine-tune the parameters for yield prediction. However, I did ensure that the crop model produced sensible biomass predictions. Actual yields for the Tully Mill were obtained from Sugar Research Australia for each year from 1982-2016. In preliminary experiments, I applied APSIM to simulate biomass using observed meteorological data and found a high correlation between simulated biomass and actual yields, confirming that the model is suitable for yield forecasting.

In my research, the management options in APSIM were held fixed to isolate the benefit of the GCM forecasts. The restriction of management variations also aided the efficiency of the APSIM simulations. The FCMD and FBMD post-processing methods output 200 ensemble members. A large ensemble ensures that the forecast distributions are sufficiently represented and minimises sampling effects on the results, however, APSIM computational demands are directly proportional to ensemble size. APSIM is currently a Windows-only program that uses text files and XML files for input and output. It assumes a fixed meteorological input with changing management options, whereas, for my experiments, the meteorological input was changing, and the management options were fixed. The operation of APSIM thus had to be scripted because the experiments could not be run from the graphical user interface. Even though I have focused on only one application with fixed management options, cluster computing on a Windows platform was necessary to efficiently complete cross-validation exercises in order to rigorously verify the biomass forecast skill. Overall, I perceive that APSIM requires an improved interface to efficiently run it with large climate ensembles.

The sugarcane application experimented with biomass forecasts over a 12 month growth and harvest cycle. It was assumed that the crop was harvested on the same date each year, at the end of

August. Initial conditions were set to the same values each year and the crop was modelled as a first ratoon crop. Whilst these settings do not reflect real world conditions (e.g., the year-to-year plant and harvest dates will differ, as will ratooning strategies), the purpose is to minimise the effects of external factors when comparing the results for different meteorological forecast inputs. In fact, for an operational forecasting system, the use of real initial conditions and other management options may improve the accuracy of the biomass forecasts compared to the results obtained using my experimental system.

Results of cross-validation experiments (chapter 4) showed that the FCMD- and FBMD-driven biomass forecasts are unbiased and reliable whether the target month is for few months ahead or up to 12 months ahead. Bias in harvest yields are close to 0%. Improvement in the skill, i.e. reduction in error, afforded by the GCM-driven biomass forecasts is consistent with the skill of the seasonal climate forecasts. I remark that the biomass forecasts initialised early in the growth period contain a high amount of uncertainty that is almost solely due to future climate uncertainty. As the season progresses, the observed weather is increasingly incorporated into the crop modelling, much improving the accuracy of the biomass predictions.

Rigorous probabilistic forecast verification, as is common in climate and hydrology forecast studies, remains relatively rare in agricultural spheres. Hence, chapter 4 also presents a suite of verification tools that can be applied to benchmark other seasonal climate forecasting approaches in crop forecasting applications. In particular, the reliability of forecasts in ensemble spread is rarely tested in crop forecasting. I have shown how the probability integral transform and its related metrics and diagrams are useful for evaluating reliability in ensemble spread and I suggest it, or a similar measure, always be used alongside other metrics for quantifying forecast bias and error. Regarding forecast error, ensemble verification requires the use of a score such as the continuous ranked probability score (CRPS), which measures error between a forecast distribution and an observation. For understanding the value of forecasts for decision-making, attributes of forecast without regard

to the actual observation, such as distribution shift (measuring deviation from normal) and change in ensemble spread (indicating the change in forecast uncertainty) are valuable additions to the crop forecaster's toolbox. Measures like distribution shift and change in ensemble spread are also more readily communicated to forecast users. Moreover, crop forecasts should be tested for skill levels against a climatological reference forecast, which makes use of historical observed data to drive the crop model.

The results of the sugarcane application are encouraging and demonstrate that using post-processed GCM forecasts is a viable option for seasonal forecasting in Australia. I conclude that the FCMD, FBMD methods, and other GCM forecast post-processing approaches, should continue to be developed and tested in new applications to harness the power of seasonal GCMs for crop yield forecasting.

5.5. Highlights and implications

At the outset of my research, the state of play was that the climate community had shot ahead and begun generating seasonal forecasts using global climate models. Despite the initial allure of daily meteorological outputs, the agricultural modelling community quickly realised they had no reliable means to adopt GCM forecasts for quantitative modelling. Consequently, operational crop forecasting systems have remained in technological limbo, i.e. relying on historical observations, or conditionally resampled observations as "forecasts". I believe my research has provided a mix of strong theoretical reasoning and pragmatic solutions that enable solid progress on this problem. I would thus like to remark on some of the highlights from my research.

A first highlight of this thesis is establishing the benefits of my new post-processing methods over popular existing methods like quantile-mapping that are being widely applied in the search for ways to link GCM forecasts and crop models. Indeed, at the very beginning of my thesis I did not know that a series of papers would emerge in 2018-19 that essentially documented negative results of using the quantile-mapping method in agricultural forecasting applications. It highlights the

timeliness of this thesis, particularly as the Australian Bureau of Meteorology advances an implementation of quantile-mapping that is purported to provide calibrated and downscaled GCM forecasts on a 5km national grid for widespread use in agricultural applications. My point here is that the inherent risks of quantile-mapping are becoming more widely known and forecast providers should not ignore its shortcomings. Reliability is primal in probabilistic forecasting and thus robust calibration methods, such as those presented in this thesis, should be considered as a minimum for forecast calibration. At the very least, simply bias-corrected or quantile-mapped forecasts should never be referred to as calibrated, as this could lead to confusion in the community and reputational damage for forecast providers and scientists. Personally, I believe also that the optimal solutions for agricultural applications are in tailored solutions. The provision of high-resolution gridded data does not necessarily resolve the problem of scale. My new post-processing methods can be customised to each application to ensure a truer calibration and downscaling.

A second highlight of my research is the in-depth evaluation of the multivariate calibration of seasonal forecasts. To my knowledge, the application of multivariate calibration techniques and, especially, the application of multivariate ensemble forecast verification methods has rarely been seen in seasonal forecasting. The desire to look more deeply into the multivariate problem, in part, came from questions or discussions at conferences, where a common question was whether the joint distribution of variables was considered in Bayesian joint probability calibration. In one respect, the results of my experiments were surprising, in that I had a priori assumed the multivariate calibration with BJP would provide superior performance over univariate calibration. Furthermore, the investigation using relatively new multivariate forecasting metrics such as the energy score and the variogram score, or more accurately, the assessment of both, enabled a deeper understanding of multivariate calibration. I would recommend multivariate forecast verification techniques be more routinely considered in seasonal forecasting in the future, particularly in applications where multiple inputs are required, such as hydrology and agriculture.

A final highlight of my research is the development of an end-to-end solution that is flexible enough to be adapted for other applications. My research spanned the development of a workflow that begins with raw climate forecasts and ends with outputs of ensemble biomass forecasts. The expertise required to develop the solution befits a niche that involves climate forecasting, statistical modelling, high-performance computing, crop-modelling and forecast verification. The value of the end-to-end solution is amplified, compared to say new methods in hydrology, because ensemble forecasting and assessment of ensemble forecasts in agriculture is relatively underdeveloped. My research shows that many of the forecast verification techniques applied in climate and hydrology are also useful for verification of crop model forecasts. Moreover, a platform has been developed upon which to reliably investigate the sensitivity of farm productivity and profits to seasonal global climate model forecasts.

5.6. Limitations and future directions

In this thesis, I have made use of parametric and non-parametric post-processing methods. I foresee that the inclusion of empirical methods in post-processing will still be required in the short term. However, the methods can become computationally demanding and managing historical data templates is cumbersome. Future projects should investigate improving the robustness of parametric methods and, if possible, extend the models to explicitly model spatial and temporal components to reduce the reliance on empirical methods.

The Bayesian joint probability modelling approach used in this study assumes stationarity in the relationships between predictors and predictands. Trends in forecasts or observations may violate this assumption. Modifications to the methods are needed to ensure the post-processed forecasts adhere to observed trends.

This thesis was weighted towards the development and evaluation of forecast calibration methods rather than the crop forecasting applications. Hence there was only one crop forecasting application, which was for sugarcane biomass forecasting. The value of FCMD and FBMD forecasts should be

established for other cropping systems in Australia, including wintertime crops such as wheat. I have already commenced work on this task in collaboration with the Queensland Alliance for Agriculture and Food Innovation using the Oz-Wheat regional-scale model. FCMD forecasts are ideal for Oz-Wheat, which requires daily forecasts at many locations within a “shire”. Actually, a small extension is required to product evaporation for use in Oz-Wheat; however, this is easily achieved by including it as another variable in FCMD. There is tremendous scope to apply the FCMD and FBMD methods, or variants thereof, to explore a range of crop forecasting applications across not only yield forecasting, but other applications like understanding how irrigation schedules and nutrient management can be optimised with seasonal climate forecasts; not only within Australia, but in applications globally.

A motivation for this thesis was the good performance of the Bayesian joint probability modelling approach for calibrating monthly GCM rainfall forecasts for monthly hydrological forecasting. Even though the FCMD forecasts have been developed for crop forecasting applications, there is opportunity to go back and apply the methods to drive hydrological models at a daily time step. Outside of Australia, snowmelt is often a big driver of seasonal streamflow, in which case rainfall, minimum temperature and maximum temperature forecasts are required to run coupled snow/hydrological models. I have begun a trial using FCMD forecasts to generate streamflow forecasts for a snowy Canadian catchment in collaboration with researchers at Sherbrooke University. However, the FCMD approach is not necessarily the best approach for hydrological applications. In the preamble to Chapter 3, it was argued that direct daily forecasting is suitable for multi-week forecasting and to take advantage of GCM initial conditions. In the collaboration with researchers at Sherbrooke University, we are exploring the multivariate post-processing of daily rainfall and temperature to compare with FCMD for hydrological forecasting, using an extension of the methods reported in Appendix B. However, as mentioned above in section 5.3, daily forecast post-processing is more suited to multi-week time horizons. Therefore, I suggest that future research

should investigate the optimal combination of direct-daily post-processing and disaggregation approaches.

It is safe to say that global climate models, or dynamical climate models, are now the primary source of seasonal climate forecast information globally. In this research, I have investigated the ECMWF forecasting model. The Australian Bureau of Meteorology is readying the release of version 2 of the Australian Climate Community and Earth System Simulator seasonal forecast system (ACCESS-S2), which will provide long hindcasts, which will be ideal for the development and testing of seasonal climate and crop forecasts in Australia.

The Commonwealth Scientific and Industrial Research Organisation in Australia has recently established the Digiscape Future Science Platform. Digiscape is a digital agriculture initiative that includes connecting seasonal forecasts with crop models. Digiscape is developing an online platform to deliver post-processed seasonal forecasts. A future technical project could integrate the FCMD and FBMD post-processing methods into this system, to deliver tailored forecasts for Australian agricultural applications. Hence, this thesis is a timely addition to Australia's digital agriculture future.

6. References

- Abatzoglou, J. T., and T. J. Brown, 2012: A comparison of statistical downscaling methods suited for wildfire applications. *International Journal of Climatology*, **32**, 772-780.
- An-Vo, D.-A., S. Mushtaq, K. Reardon-Smith, L. Kouadio, S. Attard, D. Cobon, and R. Stone, 2019: Value of seasonal forecasting for sugarcane farm irrigation planning. *European Journal of Agronomy*, **104**, 37-48.
- Baigorria, G. A., J. W. Jones, and J. J. O'Brien, 2008: Potential predictability of crop yield using an ensemble climate forecast by a regional circulation model. *Agricultural and Forest Meteorology*, **148**, 1353-1361.
- Baran, S., and A. Möller, 2015: Joint probabilistic forecasting of wind speed and temperature using Bayesian model averaging. *Environmetrics*, **26**, 120-132.
- , 2017: Bivariate ensemble model output statistics approach for joint forecasting of wind speed and temperature. *Meteorology and Atmospheric Physics*, **129**, 99-112.
- Barnston, A. G., and M. K. Tippett, 2013: Predictions of Nino3. 4 SST in CFSv1 and CFSv2: a diagnostic comparison. *Climate Dynamics*, **41**, 1615-1633.
- Barnston, A. G., M. K. Tippett, H. M. van den Dool, and D. A. Unger, 2015: Toward an Improved Multimodel ENSO Prediction. *Journal of Applied Meteorology and Climatology*, **54**, 1579-1595.
- Barnston, A. G., M. K. Tippett, M. L. L'Heureux, S. Li, and D. G. DeWitt, 2012: Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing? *Bulletin of the American Meteorological Society*, **93**, 631-651.
- Bazile, R., M.-A. Boucher, L. Perreault, and R. Leconte, 2017: Verification of ECMWF System 4 for seasonal hydrological forecasting in a northern climate. *Hydrol. Earth Syst. Sci*, **21**, 5747-5762.
- Bellier, J., G. Bontron, and I. Zin, 2017: Using Meteorological Analogues for Reordering Postprocessed Precipitation Ensembles in Hydrological Forecasting. *Water Resources Research*, **53**, 10085-10107.
- Bennett, J. C., Q. J. Wang, M. Li, D. E. Robertson, and A. Schepen, 2016: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model. *Water Resources Research*, **52**, 8238-8259.
- Bennett, J. C., Q. J. Wang, D. E. Robertson, A. Schepen, M. Li, and K. Michael, 2017a: Assessment of an ensemble seasonal streamflow forecasting system for Australia. *Hydrology and Earth System Sciences*, **21**, 6007-6030.
- , 2017b: Assessment of an ensemble seasonal streamflow forecasting system for Australia. *Hydrol. Earth Syst. Sci. Discuss.*, **In review**.
- Box, G. E., and D. R. Cox, 1964: An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211-252.
- Brown, J. N., Z. Hochman, D. Holzworth, and H. Horan, 2018: Seasonal climate forecasts provide more definitive and accurate crop yield predictions. *Agricultural and Forest Meteorology*, **260**, 247-254.
- Buishand, T. A., and T. Brandsma, 2001: Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. *Water Resources Research*, **37**, 2761-2776.
- Capa-Morocho, M., A. V. Ines, W. E. Baethgen, B. Rodríguez-Fonseca, E. Han, and M. Ruiz-Ramos, 2016: Crop yield outlooks in the Iberian Peninsula: Connecting seasonal climate forecasts with crop simulation models. *Agricultural systems*, **149**, 75-87.
- Carter, J., W. Hall, K. Brook, G. McKeon, K. Day, and C. Paull, 2000: Aussie GRASS: Australian grassland and rangeland assessment by spatial simulation. *Applications of seasonal climate forecasting in agricultural and natural ecosystems*, Springer, 329-349.
- Chen, J., H. Chen, and S. Guo, 2017: Multi-site precipitation downscaling using a stochastic weather generator. *Climate Dynamics*.

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, **5**, 243-262.

Clarke, A. J., S. Van Gorder, and Y. Everingham, 2010: Forecasting Long-Lead Rainfall Probability with Application to Australia's Northeastern Coast. *Journal of Applied Meteorology and Climatology*, **49**, 1443-1453.

Crochemore, L., M. H. Ramos, and F. Pappenberger, 2016: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.*, **20**, 3601-3618.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting — II. Calibration and combination. *Tellus A: Dynamic Meteorology and Oceanography*, **57**, 234-252.

Dreccer, M. F., J. Fainges, J. Whish, F. C. Ogbonnaya, and V. O. Sadras, 2018: Comparison of sensitive stages of wheat, barley, canola, chickpea and field pea to temperature and water stress across Australia. *Agricultural and Forest Meteorology*, **248**, 275-294.

Drosowsky, W., and L. E. Chambers, 2001: Near-global sea surface temperature anomalies as predictors of Australian seasonal rainfall. *Journal of Climate*, **14**, 1677-1687.

Everingham, Y., A. Clarke, and S. Van Gorder, 2008: Long lead rainfall forecasts for the Australian sugar industry. *International Journal of Climatology*, **28**, 111-117.

Everingham, Y., G. Inman-Bamber, J. Sexton, and C. Stokes, 2015: A dual ensemble agroclimate modelling procedure to assess climate change impacts on sugarcane production in Australia. *Agricultural Sciences*, **6**, 870-888.

Everingham, Y., J. Sexton, D. Skocaj, and G. Inman-Bamber, 2016: Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for sustainable development*, **36**, 27.

Fedderson, H., A. Navarra, and M. N. Ward, 1999: Reduction of Model Systematic Error by Statistical Correction for Dynamical Seasonal Predictions. *Journal of Climate*, **12**, 1974-1989.

Feldman, D. L., and H. M. Ingram, 2009: Making science useful to decision makers: climate forecasts, water management, and knowledge networks. *Weather, Climate, and Society*, **1**, 9-21.

Fita, L., J. Evans, D. Argüeso, A. King, and Y. Liu, 2017: Evaluation of the regional climate response in Australia to large-scale climate modes in the historical NARCLIM simulations. *Climate Dynamics*, **49**, 2815-2829.

Freebairn, D., and D. McClymont, 2012: CliMate—a smartphone App for analysing climate data. *16th Australian Agronomy Conference*, Armidale, NSW, Australia, 4 pp.

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

Giudice, D. D., M. Honti, A. Scheidegger, C. Albert, P. Reichert, and J. Rieckermann, 2013: Improving uncertainty estimation in urban hydrological modeling by statistically describing bias. *Hydrology and Earth System Sciences*, **17**, 4209-4225.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359-378.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 243-268.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, **133**, 1098-1118.

Gutmann, E., T. Pruitt, M. P. Clark, L. Brekke, J. R. Arnold, D. A. Raff, and R. M. Rasmussen, 2014: An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resources Research*, **50**, 7167-7186.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting — I. Basic concept. *Tellus A*, **57**, 219-233.

Hammer, G. L., N. Nicholls, and C. Mitchell, 2000: *Applications of seasonal climate forecasting in agricultural and natural ecosystems*. Vol. 21, Springer Science & Business Media.

Han, E., and A. V. Ines, 2017: Downscaling probabilistic seasonal climate forecasts for decision support in agriculture: A comparison of parametric and non-parametric approach. *Climate Risk Management*, **18**, 51-65.

Han, E., A. V. M. Ines, and W. E. Baethgen, 2017: Climate-Agriculture-Modeling and Decision Tool (CAMDT): A software framework for climate risk management in agriculture. *Environmental Modelling & Software*, **95**, 102-114.

Hansen, J. W., 2005: Integrating seasonal climate prediction and agricultural models for insights into agricultural practice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 2037-2047.

Hansen, J. W., and J. W. Jones, 2000: Scaling-up crop models for climate variability applications. *Agricultural Systems*, **65**, 43-72.

Hansen, J. W., A. Potgieter, and M. K. Tippett, 2004: Using a general circulation model to forecast regional wheat yields in northeast Australia. *Agricultural and Forest Meteorology*, **127**, 77-92.

Hawthorne, S., Q. Wang, A. Schepen, and D. Robertson, 2013: Effective use of general circulation model outputs for forecasting monthly rainfalls to long lead times. *Water Resources Research*, **49**, 5427-5436.

Hochman, Z., and P. Carberry, 2011: Emerging consensus on desirable characteristics of tools to support farmers' management of climate risk in Australia. *Agricultural Systems*, **104**, 441-450.

Hochman, Z., and Coauthors, 2009: Re-inventing model-based decision support with Australian dryland farmers. 4. Yield Prophet® helps farmers monitor and manage crops in a variable climate. *Crop and Pasture Science*, **60**, 1057-1070.

Holzworth, D. P., and Coauthors, 2014: APSIM—evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software*, **62**, 327-350.

Howden, M., S. Schroeter, S. Crimp, and I. Hanigan, 2014: The changing roles of science in managing Australian droughts: An agricultural perspective. *Weather and Climate Extremes*, **3**, 80-89.

Hudson, D., O. Alves, H. H. Hendon, and A. G. Marshall, 2011: Bridging the gap between weather and seasonal forecasting: intraseasonal forecasting for Australia. *Quarterly Journal of the Royal Meteorological Society*, **137**, 673-689.

Hudson, D., L. Shi, O. Alves, M. Zhao, H. H. Hendon, and G. Young, 2017a: Performance of ACCESS-S1 for key horticultural regions, 39 pp.

Hudson, D., and Coauthors, 2017b: ACCESS-S1: The new Bureau of Meteorology multi-week to seasonal prediction system. *Journal of Southern Hemisphere Earth Systems Science*, **67**, 132-159.

Hwang, S., and W. D. Graham, 2013: Development and comparative evaluation of a stochastic analog method to downscale daily GCM precipitation. *Hydrology and Earth System Sciences*, **17**, 4481-4502.

Ines, A. V., and J. W. Hansen, 2006: Bias correction of daily GCM rainfall for crop simulation studies. *Agricultural and forest meteorology*, **138**, 44-53.

Ines, A. V. M., J. W. Hansen, and A. W. Robertson, 2011: Enhancing the utility of daily GCM rainfall for crop yield prediction. *International Journal of Climatology*, **31**, 2168-2182.

Jeffrey, S. J., J. O. Carter, K. B. Moodie, and A. R. Beswick, 2001: Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling & Software*, **16**, 309-330.

Jha, P. K., P. Athanasiadis, S. Gualdi, A. Trabucco, V. Mereu, V. Shelia, and G. Hoogenboom, 2019: Using daily data from seasonal forecasts in dynamic crop models for yield prediction: A case study for rice in Nepal's Terai. *Agricultural and Forest Meteorology*, **265**, 349-358.

Jin, E. K., and Coauthors, 2008: Current status of ENSO prediction skill in coupled ocean-atmosphere models. *Climate Dynamics*, **31**, 647-664.

Johnson, S. J., and Coauthors, 2018: SEAS5: The new ECMWF seasonal forecast system. *Geosci. Model Dev. Discuss.*, **2018**, 1-44.

Kandulu, J., P. Thorburn, J. Biggs, and K. Verburg, 2018: Estimating economic and environmental trade-offs of managing nitrogen in Australian sugarcane systems taking agronomic risk into account. *Journal of Environmental Management*, **223**, 264-274.

Keating, B. A., and Coauthors, 2003: An overview of APSIM, a model designed for farming systems simulation. *European journal of agronomy*, **18**, 267-288.

Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter. *Climate Dynamics*, **39**, 2957-2973.

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. *Bulletin of the American Meteorological Society*, **95**, 585-601.

Klemm, T., and R. A. McPherson, 2017: The development of seasonal climate forecasting for agricultural producers. *Agricultural and forest meteorology*, **232**, 384-399.

Laio, F., and S. Tamea, 2007: Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, **11**, 1267-1277.

Li, X., and Coauthors, 2018: Three resampling approaches based on method of fragments for daily-to-subdaily precipitation disaggregation. *International Journal of Climatology*, **38**, e1119-e1138.

Lim, E.-P., H. H. Hendon, D. Hudson, G. Wang, and O. Alves, 2009: Dynamical forecast of inter-El Nino variations of tropical SST and Australian spring rainfall. *Monthly Weather Review*, **137**, 3796-3810.

Lucatero, D., H. Madsen, J. C. Refsgaard, J. Kidmose, and K. H. Jensen, 2018: Seasonal streamflow forecasts in the Ahlergaarde catchment, Denmark: the effect of preprocessing and post-processing on skill and statistical consistency. *Hydrol. Earth Syst. Sci.*, **22**, 3601-3617.

Luo, L., and E. F. Wood, 2008: Use of Bayesian Merging Techniques in a Multimodel Seasonal Hydrologic Ensemble Prediction System for the Eastern United States. *Journal of Hydrometeorology*, **9**, 866-884.

MacLachlan, C., and Coauthors, 2015: Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, **141**, 1072-1084.

Manzanas, R., and Coauthors, 2018: Dynamical and statistical downscaling of seasonal temperature forecasts in Europe: Added value for user applications. *Climate Services*, **9**, 44-56.

Maraun, D., 2013: Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate*, **26**, 2137-2143.

Marshall, A., D. Hudson, H. Hendon, M. Pook, O. Alves, and M. Wheeler, 2014a: Simulation and prediction of blocking in the Australian region and its influence on intra-seasonal rainfall in POAMA-2. *Climate dynamics*, **42**, 3271-3288.

Marshall, A., D. Hudson, M. Wheeler, O. Alves, H. Hendon, M. Pook, and J. Risbey, 2014b: Intra-seasonal drivers of extreme heat over Australia in observations and POAMA-2. *Climate dynamics*, **43**, 1915-1937.

Marshall, A. G., D. Hudson, M. C. Wheeler, H. H. Hendon, and O. Alves, 2011: Assessing the simulation and prediction of rainfall associated with the MJO in the POAMA seasonal forecast system. *Climate dynamics*, **37**, 2129-2141.

—, 2012: Simulation and prediction of the Southern Annular Mode and its influence on Australian intra-seasonal climate in POAMA. *Climate dynamics*, **38**, 2483-2502.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Management science*, **22**, 1087-1096.

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

McLean Slaughter, J., T. Gneiting, and A. E. Raftery, 2012: Probabilistic Wind Vector Forecasting Using Ensembles and Bayesian Model Averaging. *Monthly Weather Review*, **141**, 2107-2119.

Meinke, H., and R. C. Stone, 2005: Seasonal and inter-annual climate forecasting: the new tool for increasing preparedness to climate variability and change in agricultural planning and operations. *Climatic change*, **70**, 221-253.

Meinke, H., G. Hammer, H. Van Keulen, and R. Rabbinge, 1998: Improving wheat simulation capabilities in Australia from a cropping systems perspective III. The integrated wheat model (I_WHEAT). *European Journal of Agronomy*, **8**, 101-116.

Meinke, H., G. L. Hammer, H. van Keulen, R. Rabbinge, and B. A. Keating, 1997: Improving wheat simulation capabilities in Australia from a cropping systems perspective: water and nitrogen effects on spring wheat in a semi-arid environment. *Developments in Crop Science*, Elsevier, 99-112.

Meinke, H., and Coauthors, 2005: Rainfall variability at decadal and longer time scales: signal or noise? *Journal of Climate*, **18**, 89-96.

Meza, F. J., J. W. Hansen, and D. Osgood, 2008: Economic Value of Seasonal Climate Forecasts for Agriculture: Review of Ex-Ante Assessments and Recommendations for Future Research. *Journal of Applied Meteorology and Climatology*, **47**, 1269-1286.

Möller, A., A. Lenkoski, and T. L. Thorarinsdottir, 2013: Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, **139**, 982-991.

Molteni, F., and Coauthors, 2011: *The new ECMWF seasonal forecast system (System 4)*. European Centre for Medium-Range Weather Forecasts.

Murphy, B. F., and B. Timbal, 2008: A review of recent climate variability and climate change in southeastern Australia. *International Journal of Climatology*, **28**, 859-879.

Nicholls, N., 1986: Use of the Southern Oscillation to predict Australian sorghum yield. *Agricultural and Forest Meteorology*, **38**, 9-15.

Pegion, K., T. DelSole, E. Becker, and T. Cicerone, 2017: Assessing the fidelity of predictability estimates. *Climate Dynamics*.

Peng, Z., Q. Wang, J. C. Bennett, A. Schepen, F. Pappenberger, P. Pokhrel, and Z. Wang, 2014: Statistical calibration and bridging of ECMWF System4 outputs for forecasting seasonal precipitation over China. *Journal of Geophysical Research: Atmospheres*, **119**, 7116-7135.

Pinson, P., 2012: Adaptive calibration of (u, v)-wind ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **138**, 1273-1284.

Potgieter, A., G. Hammer, and D. Butler, 2002: Spatial and temporal patterns in Australian wheat yield and their relationship with ENSO. *Australian Journal of Agricultural Research*, **53**, 77-89.

Potgieter, A., G. Hammer, A. Doherty, and P. De Voil, 2005a: A simple regional-scale model for forecasting sorghum yield across North-Eastern Australia. *Agricultural and Forest Meteorology*, **132**, 143-153.

Potgieter, A. B., Y. L. Everingham, and G. L. Hammer, 2003: On measuring quality of a probabilistic commodity forecast for a system that incorporates seasonal climate forecasts. *International Journal of Climatology*, **23**, 1195-1210.

Potgieter, A. B., G. L. Hammer, H. Meinke, R. C. Stone, and L. Goddard, 2005b: Three putative types of El Nino revealed by spatial variability in impact on Australian wheat yield. *Journal of Climate*, **18**, 1566-1574.

Power, S., T. Casey, C. Folland, A. Colman, and V. Mehta, 1999: Inter-decadal modulation of the impact of ENSO on Australia. *Climate Dynamics*, **15**, 319-324.

Pui, A., A. Sharma, R. Mehrotra, B. Sivakumar, and E. Jeremiah, 2012: A comparison of alternatives for daily to sub-daily rainfall disaggregation. *Journal of Hydrology*, **470-471**, 138-157.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks, 2010: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, **46**, W05521.

Risbey, J. S., M. J. Pook, P. C. McIntosh, M. C. Wheeler, and H. H. Hendon, 2009: On the remote drivers of rainfall variability in Australia. *Monthly Weather Review*, **137**, 3233-3253.

Robertson, D. E., D. L. Shrestha, and Q. J. Wang, 2013: Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. *Hydrol. Earth Syst. Sci.*, **17**, 3587-3603.

Rodriguez, D., P. de Voil, D. Hudson, J. Brown, P. Hayman, H. Marrou, and H. Meinke, 2018: Predicting optimum crop designs using crop models and seasonal climate forecasts. *Scientific reports*, **8**, 2231.

Rose, D. C., and Coauthors, 2016: Decision support tools for agriculture: Towards effective design and delivery. *Agricultural systems*, **149**, 165-174.

Saha, S., and Coauthors, 2014: The NCEP climate forecast system version 2. *Journal of Climate*, **27**, 2185-2208.

Schefzik, R., 2016a: Combining parametric low-dimensional ensemble postprocessing with reordering methods. *Quarterly Journal of the Royal Meteorological Society*, **142**, 2463-2477.

—, 2016b: A Similarity-Based Implementation of the Schaake Shuffle. *Monthly Weather Review*, **144**, 1909-1921.

Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling. *Statist. Sci.*, **28**, 616-640.

Schepen, A., and Q. Wang, 2013: Toward accurate and reliable forecasts of Australian seasonal rainfall by calibrating and merging multiple coupled GCMs. *Monthly Weather Review*, **141**, 4554-4563.

—, 2014: Ensemble forecasts of monthly catchment rainfall out to long lead times by post-processing coupled general circulation model output. *Journal of hydrology*, **519**, 2920-2931.

—, 2015: Model averaging methods to merge operational statistical and dynamic seasonal streamflow forecasts in Australia. *Water Resources Research*, **51**, 1797-1812.

Schepen, A., Q. Wang, and D. Robertson, 2012: Evidence for using lagged climate indices to forecast Australian seasonal rainfall. *Journal of Climate*, **25**, 1230-1246.

Schepen, A., Q. Wang, and D. E. Robertson, 2014: Seasonal forecasts of Australian rainfall through calibration and bridging of coupled GCM outputs. *Monthly Weather Review*, **142**, 1758-1770.

Schepen, A., Q. Wang, and Y. Everingham, 2016: Calibration, bridging, and merging to improve GCM seasonal temperature forecasts in Australia. *Monthly Weather Review*, **144**, 2421-2441.

Schepen, A., T. Zhao, Q. J. Wang, and D. E. Robertson, 2018: A Bayesian modelling method for post-processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments. *Hydrology and Earth System Sciences*, **22**, 1615-1628.

Scheuerer, M., and T. M. Hamill, 2015: Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities. *Monthly Weather Review*, **143**, 1321-1334.

Scheuerer, M., T. M. Hamill, B. Whitin, M. He, and A. Henkel, 2017: A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resources Research*, **53**, 3029-3046.

Schuhen, N., T. L. Thorarinsdottir, and T. Gneiting, 2012: Ensemble model output statistics for wind vectors. *Monthly weather review*, **140**, 3204-3219.

Semenov, M. A., and F. J. Doblas-Reyes, 2007: Utility of dynamical seasonal forecasts in predicting crop yield. *Climate Research*, **34**, 71-81.

Shi, L., H. H. Hendon, O. Alves, J.-J. Luo, M. Balmaseda, and D. Anderson, 2012: How predictable is the Indian Ocean dipole? *Monthly Weather Review*, **140**, 3867-3884.

Skocaj, D., and Y. Everingham, 2014: Identifying climate variables having the greatest influence on sugarcane yields in the Tully mill area. Australian Society of Sugar Cane Technologists.

Skocaj, D. M., Y. L. Everingham, and B. L. Schroeder, 2013: Nitrogen management guidelines for sugarcane production in Australia: can these be modified for wet tropical conditions using seasonal climate forecasting? *Springer Science Reviews*, **1**, 51-71.

Srikanthan, R., and T. A. McMahon, 2001: Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences Discussions*, **5**, 653-670.

Stern, H., G. De Hoedt, and J. Ernst, 2000: Objective classification of Australian climates. *Australian Meteorological Magazine*, **49**, 87-96.

Stone, R. C., G. L. Hammer, and T. Marcussen, 1996: Prediction of global rainfall probabilities using phases of the Southern Oscillation Index. *Nature*, **384**, 252-255.

Strazzo, S., D. Collins, C. A. Schepen, Q. J. Wang, E. Becker, and L. Jia, 2018: Application of a hybrid statistical-dynamical system to seasonal prediction of North American temperature and precipitation. *Monthly Weather Review*, In Revision.

Thorburn, P., J. Biggs, S. Attard, and J. Kemei, 2011: Environmental impacts of irrigated sugarcane production: nitrogen lost through runoff and leaching. *Agriculture, ecosystems & environment*, **144**, 1-12.

Tian, D., C. J. Martinez, W. D. Graham, and S. Hwang, 2014: Statistical downscaling multimodel forecasts for seasonal precipitation and surface temperature over the southeastern United States. *Journal of Climate*, **27**, 8384-8411.

van Dijk, A. I. J. M., and Coauthors, 2013: The Millennium Drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society. *Water Resources Research*, **49**, 1040-1057.

Vannitsem, S., D. S. Wilks, and J. Messner, 2018: *Statistical Postprocessing of Ensemble Forecasts*. Elsevier Science.

Verdin, A., B. Rajagopalan, W. Kleiber, G. Podestá, and F. Bert, 2018: A conditional stochastic weather generator for seasonal to multi-decadal simulations. *Journal of Hydrology*, **556**, 835-846.

Verdon-Kidd, D. C., and A. S. Kiem, 2009: Nature and causes of protracted droughts in southeast Australia: Comparison between the Federation, WWII, and Big Dry droughts. *Geophysical Research Letters*, **36**.

Verkade, J. S., J. D. Brown, P. Reggiani, and A. H. Weerts, 2013: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, **501**, 73-91.

Vitart, F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society*, **140**, 1889-1899.

Volosciuk, C. D., D. Maraun, M. Vrac, and M. Widmann, 2017: A combined statistical bias correction and stochastic downscaling method for precipitation. *Hydrology and Earth System Sciences*, **21**, 1693-1719.

Vrac, M., and P. Friederichs, 2015: Multivariate—Intervariable, Spatial, and Temporal—Bias Correction. *Journal of Climate*, **28**, 218-237.

Wang, Q., and D. Robertson, 2011: Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resources Research*, **47**.

Wang, Q., D. Robertson, and F. Chiew, 2009: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resources Research*, **45**.

Wang, Q., A. Schepen, and D. E. Robertson, 2012a: Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging. *Journal of Climate*, **25**, 5524-5537.

Wang, Q., D. Shrestha, D. Robertson, and P. Pokhrel, 2012b: A log-sinh transformation for data normalization and variance stabilization. *Water Resources Research*, **48**, W05514.

Wang, Q., T. Pagano, S. Zhou, H. Hapuarachchi, L. Zhang, and D. Robertson, 2011: Monthly versus daily water balance models in simulating monthly runoff. *Journal of hydrology*, **404**, 166-175.

Weisheimer, A., and T. Palmer, 2014: On the reliability of seasonal climate forecasts. *Journal of The Royal Society Interface*, **11**, 20131162.

Western, A. W., K. B. Dassanayake, K. C. Perera, R. M. Argent, O. Alves, G. Young, and D. Ryu, 2018: An evaluation of a methodology for seasonal soil water forecasting for Australian dry land cropping systems. *Agricultural and Forest Meteorology*, **253**, 161-175.

Westra, S., R. Mehrotra, A. Sharma, and R. Srikanthan, 2012: Continuous rainfall simulation: 1. A regionalized subdaily disaggregation approach. *Water Resources Research*, **48**.

- Wetterhall, F., H. Winsemius, E. Dutra, M. Werner, and E. Pappenberger, 2015: Seasonal predictions of agro-meteorological drought indicators for the Limpopo basin. *Hydrology and Earth System Sciences*, **19**, 2577-2586.
- White, C. J., D. Hudson, and O. Alves, 2013: ENSO, the IOD and the intraseasonal prediction of heat extremes across Australia using POAMA-2. *Climate Dynamics*, **43**, 1791-1810.
- Wójcik, R., and T. A. Buishand, 2003: Simulation of 6-hourly rainfall and temperature by two resampling schemes. *Journal of Hydrology*, **273**, 69-80.
- Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research: Atmospheres*, **107**, ACL 6-1-ACL 6-15.
- Wu, L., Y. Zhang, T. Adams, H. Lee, Y. Liu, and J. Schaake, 2018: Comparative Evaluation of Three Schaake Shuffle Schemes in Post-processing GEFS Precipitation Ensemble Forecasts. *Journal of Hydrometeorology*, **0**, null.
- Yeh, S.-W., J.-S. Kug, B. Dewitte, M.-H. Kwon, B. P. Kirtman, and F.-F. Jin, 2009: El Niño in a changing climate. *Nature*, **461**, 511.
- Yeo, I. K., and R. A. Johnson, 2000: A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954-959.
- Yuan, X., and E. F. Wood, 2012: Downscaling precipitation or bias-correcting streamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast. *Water Resources Research*, **48**.
- Zhao, M., and H. H. Hendon, 2009: Representation and prediction of the Indian Ocean dipole in the POAMA seasonal forecast model. *Quarterly Journal of the Royal Meteorological Society*, **135**, 337-352.
- Zhao, T., A. Schepen, and Q. Wang, 2016: Ensemble forecasting of sub-seasonal to seasonal streamflow by a Bayesian joint probability modelling approach. *Journal of Hydrology*, **541**, 839-849.
- Zhao, T., Q. J. Wang, A. Schepen, and M. Griffiths, 2019: Ensemble forecasting of monthly and seasonal reference crop evapotranspiration based on global climate model outputs. *Agricultural and Forest Meteorology*, **264**, 114-124.
- Zhao, T., J. Bennett, Q. J. Wang, A. Schepen, A. Wood, D. Robertson, and M.-H. Ramos, 2017: How suitable is quantile mapping for post-processing GCM precipitation forecasts? *Journal of Climate*.

Appendix A: Ancillary paper 1

In the first year of my PhD, I used outputs from the Bureau of Meteorology's POAMA global climate model. Early on in my candidature, the Bureau announced that POAMA would be abandoned and a new Australian model, ACCESS-S (now operational), would be derived from the United Kingdom's Met Office's model. For reasons of stability, I made the decision to switch to using the European Centre for Medium-range Weather Forecasting's System4 for the remainder of my research.

System4 has a long archive of reforecasts available for testing and is generally regarded as a world-leading model. Moreover, its performance had not previously been assessed for Australia.

The referenced paper forming this appendix presents results from my initial research using POAMA. It is highly supportive to the thesis as it documents the extension of BJP to post-process seasonal GCM forecasts of minimum and maximum temperature at continental scales in line with objective 1.

Reference:

Schepen, A., Q. Wang, and Y. Everingham, 2016: Calibration, bridging, and merging to improve GCM seasonal temperature forecasts in Australia. *Monthly Weather Review*, 144, 2421-2441.

Appendix B: Ancillary paper 2

When addressing the need for daily meteorological forecasts, the question arises of whether it's feasible to apply BJP calibration to post-processing global climate model forecasts directly on a daily time step. Theoretically and computationally it is feasible. In work ancillary to this thesis, I developed a BJP-based method to directly calibrate daily ACCESS-S catchment rainfall forecasts. Post-processing daily forecasts directly appears highly suitable for multi-week forecasting where it is possible to take advantage of skill afforded by GCM initial conditions. The daily forecast approach may be extended in future research to other variables such as temperature and radiation to support short-term agricultural decision-making such as irrigation scheduling. The study and findings are reported in the referenced paper "A Bayesian modelling method for post-processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments".

Reference:

Schepen, A., T. Zhao, Q. J. Wang, and D. E. Robertson, 2018: A Bayesian modelling method for post processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments. *Hydrology and Earth System Sciences*, 22, 1615-1628.

Appendix C: Revisions

Detailed Response to Examiner Comments by Andrew Schepen				
Page Number of Original Thesis	Examiner Comment	Candidate Response to Comment	Amendments made to Thesis	Page Number in Amended Thesis
Examiner 1				
4	Clarify if the ENSO forecasting systems are only applicable to Australia or also Internationally. If Internationally then other analogue based systems should be mentioned too. I am also not convinced that the ENSO analogues are the only analogue system used in Australia? Please clarify this sentence.	ENSO and analogue forecasting systems are certainly used globally, not only in Australia. However, my comment pertains to the Australian focus. Whilst my understanding is that ENSO analogues are the most common meteorological forcing used for crop modelling in Australia, I appreciate that alternative forcings are available. I am happy to clarify this sentence accordingly.	The sentence has been clarified to state that the focus is for Australia and that other meteorological forcings are available.	4
18	Page 18: “the agricultural modelling community have not adapted to...” Please elaborate more here Other reason include - accuracy and readily availability of GCM forecast - limit number of years available for GCM - accurate temperature, radiation and rainfall in daily	I agree these are useful points to make. I note that the section referred to here is the preamble, and much more detail is given in the body of the chapter. Nevertheless, I am happy to mention these points in the preamble.	The preamble has been modified to mention that: (1) GCM data availability has been limited historically and (2) raw GCM forecasts of meteorological variables requires post-processing at a daily time step for use in crop models.	18

	time steps to drive crop models generally was not available before			
18	<p>I suggest that the candidate should read the book by Hammer, G.L. 2000 and include it as reference to strengthen his case further.</p> <p>- Hammer, G.L., 2000. Applying seasonal climate forecasts in agricultural and natural ecosystems - A synthesis. In: G.L. Hammer, N. Nicholls and C. Mitchell (Editors), Applications of Seasonal Climate Forecasting in Agricultural and Natural Ecosystems – The Australian Experience. Atmospheric and Oceanographic Sciences Library, Kluwer., pp. 453-462.</p>	<p>I certainly read this book, which is a collection of papers from a scientific symposium, in the early stages of my PhD candidature. I am happy to refer to the book to recognise the early efforts in seasonal forecasting for agriculture in Australia.</p>	<p>The Hammer et al. (2000) book is now cited twice in the introduction.</p>	19 and 20
18	<p>Describing the ENSO at global scales could be find here:</p> <p>- Allan, R.J., 2000. El Niño and the Southern Oscillation: Multiscale variability and its impacts on natural ecosystems and society. In: H.F. Diaz and V. Markgraf (Editors), ENSO and climatic variability in the last</p>	<p>ENSO variability on global scales is well understood and a vast body of research is readily available. Allen (2000) is a book chapter that does not seem to be readily available. I believe that the current references in my thesis can guide the interested reader to find additional information about</p>	<p>No change.</p>	

	150 years. . Cambridge Univ. Press., Cambridge, UK., pp. 3 - 55.	ENSO variability on global scales.		
19	<p>Impact of ENSO on crop yields across Australia can be find here:</p> <p>- Nicholls, N., 1986. Use of the Southern Oscillation to predict Australian sorghum yield. Agricultural and Forest Meteorology, 38(1-3): 9-15.</p> <p>- Meinke, H. et al., 2005. Rainfall variability at decadal and longer time scales: signal or noise? . Journal of Climate, 18: 89-96.</p> <p>- Potgieter, A.B., Hammer, G.L., Meinke, H., Stone, R.C. and Goddard, L., 2005. Three putative types of El Nino revealed by spatial variability in impact on Australian wheat yield. Journal of Climate, 18(10): 1566-1574.</p> <p>- Potgieter, A.B., Hammer, G.L. and Butler, D., 2002. Spatial and temporal patterns in Australian wheat yield and their relationship with ENSO. Australian Journal of Agricultural Research, 53: 77-89.</p>	I have reviewed each of these references in the context of the impact of ENSO on crop yields.	<p>Three references added at the suggested place.</p> <p>The reference to Meinke fits better with the reference to decadal oscillations and has been added to the references in the previous paragraph.</p>	19

21	<p>I suggest that there are more examples of applications of APSIM in decision support tool by Hammer et al and Meinke et al which needs to be included.</p> <p>- Meinke, H., Hammer, G.L., van Keulen, H. and Rabbinge, R., 1998. Improving wheat simulation capabilities in Australia from a cropping systems perspective. III. The integrated wheat model (I_WHEAT). European Journal of Agronomy, 8: 101-116.</p> <p>- Meinke, H., Hammer, G.L., van Keulen, H., Rabbinge, R. and Keating, B.A., 1997. Improving wheat simulation capabilities in Australia from a cropping systems perspective. Water and nitrogen effects on spring wheat in a semi-arid environment. . European Journal of Agronomy, 7: 75-88.</p>	<p>I am happy to recognise these earlier efforts to develop wheat prediction capability in APSIM</p>	<p>A sentence has been extended to incorporate the suggested references.</p>	21
26	<p>crop forecasting applications....</p> <p>- Hansen, J.W., Potgieter, A. and Tippet, M.K., 2004. Using a general circulation model to forecast regional wheat yields in northeast Australia. Agricultural and Forest Meteorology, 127(1-2): 77-92.</p>	<p>I added this reference to the introduction to chapter 4 in response to a later comment.</p>	<p>Reference added to chapter 4 introduction.</p>	112

27	PET not define?	Yes, PET had not yet been defined	PET replaced with “potential evapotranspiration”	27
71	(last paragraph): can you elaborate on the likelihood of machine learning approaches like NN or LASSO to assist in downscaling?	I could, but neither of these approaches have direct relevance to the ensemble forecast post-processing methods used in my thesis. Therefore, I believe it is best to avoid discussing them.	No change	
74	How does Hidden Markov chain approaches fit or compare to other approaches in weather generators and thus downscaling the temporal patterns to a station level? See Hammer 2000 book section II chapter 9.	I have read the suggested chapter. For my thesis, a Hidden Markov Model approach would only be used for the purpose of stochastic weather generation. I therefore disregard HMMs for the same reasons as other types of weather generators. In my thesis, I believe it is important to mention that I experimented with weather generators; however, I ultimately took a different path. It is not clear that unpacking the differences between different techniques for weather generation will add value.	No change	
90	It is not clear to me what method of calculating solar radiation has been implemented in this chapter? Please clarify.	Solar radiation has been obtained from the Silo dataset as stated. My comment about calculating solar radiation suggests a reference for people who do not have access to a solar radiation data set and wish to	Two sentences discussing the possibility of missing radiation data are deleted.	91

		calculate it. Since the comment may be misleading, I will delete it.		
93	Kendall correlation and Spearman correlations maybe include small r in brackets (r).	OK	(r) added after Kendall correlation	94
96	Please elaborate on how this region compare to the broad cropping areas of Australia in terms of daily average rainfall? This will help to illuminate the study area and skill in this environment better. Also, will be good to maybe include a table of total seasonal and average daily rainfalls for each station.	The average daily rainfall for two different periods are already shown in Figures 3.5 and 3.6. In fact, the distribution is shown for each meteorological variable, not just rainfall, and for each site. Therefore, it would be redundant to include an additional table with this information. However, I do agree some more information about the climate is warranted and I will add some further descriptive text.	A sentence is added to section 3.4.1 describing the seasonality of rainfall in the Burdekin region.	90
98	Section 3.53.: Please quantify the correlations figures in the text	The correlations are already quantified in paragraph 2.	No change	
	Figures 3.7 and 3.8: Include the abbreviations for the legends in the text	OK	A sentence is added to describe the labels	99
110	Include the work done by Hansen et al 2004: o Hansen, J.W., Potgieter, A. and Tippett, M.K., 2004. Using a general circulation model to forecast regional wheat yields in	OK	Added reference	112

	northeast Australia. Agricultural and Forest Meteorology, 127(1-2): 77-92.			
113	<p>It is not clear to me how ENSO (and nino 3.4) effects the rainfall for the study area? Please clarify. I suggest have a look at the research by Allen 2000: o Allan, R.J., 2000. El Niño and the Southern Oscillation: Multiscale variability and its impacts on natural ecosystems and society. In: H.F. Diaz and V. Markgraf (Editors), ENSO and climatic variability in the last 150 years. . Cambridge Univ. Press., Cambridge, UK., pp. 3 - 55</p>	<p>It was not explicit in the text about the effect of ENSO on rainfall in the region. While such information is widely available, for example through the Bureau of Meteorology, I agree it is useful to be clearer about the ENSO effect on the study area. I have added a sentence in this regard.</p>	<p>A sentence is added to describe that the rainfall odds shift lower / higher in El Nino and La Nina years.</p>	114
114	<p>work cited here by Rodriguez 2018 : which cropping system was this done on?</p>	<p>They studied a sorghum crop. My citation of their work is only concerned with the rainfall component of their study, but I am happy to mention that the crop is sorghum.</p>	<p>A sentence has been modified to indicate that the studied crop is sorghum.</p>	115
120	<p>equally skilful on average Please clarify what you mean by skill? Accuracy, reliability, bias etc..?</p>	<p>By equally skilful, I mean the average error score is the same. I have used CRPS as the error score, which is sensitive to all of accuracy, reliability and bias.</p> <p>I agree the language could be more precise.</p>	<p>The sentence is modified to say that if the models are equally skilful then the average CRPS is the same.</p>	121

123	Figure and text for figure 4.3. Not clear to me which February the forecast is issued relative to the forecasts periods? Please clarify.	It is for February issued forecasts in the year of harvest.	The caption is updated to clarify that the forecast is issued in February for the year of harvest after using observed meteorological forcings from the previous September up to January.	124
130	Management options. Please include a table with the management options used here for running APSIM?	The important details about the APSIM-sugar model are described earlier in section 4.3.2. The only scheduled activities are fertiliser application and harvest. The crop is treated as a ratoon crop, meaning it simply regrows after harvest. There is no irrigation. I can see that some additional detail about the cultivar, soil type and fertiliser application rates may be of interest. I am happy to share the APSIM XML configuration with interested persons if permitted by the original model developers.	Section 4.3.2 is updated with more detail about the APSIM-sugar configuration.	117
131	doubt it that soil water will impact the final simulated biomass here since as I understand from described before this is a fully irrigated crop?	Actually, irrigation is not applied in the model. I will clarify.	It is clarified that the system as rainfed.	132
131	Quantify the degree of shift and dispersion of the down-scaled crop forecasts have gained compared to climatology?	The degree of shift and dispersion relative to climatology is shown in Figure 4.9 and discussed in the text on page 129.	No change	
136	"...at 0.75x0.75-degree resolution..." State the	OK	The approximate cell size is now stated in brackets	51, 91 and 137

	cell size in km in brackets			
139	Please elaborate the scientific premise of weather generators earlier in your thesis	I have referred to weather generators without explaining what they are, so I am happy to include additional detail earlier on as suggested.	Para 2 of the preamble to chapter 3 is updated to describe weather generators, their purpose, and why they are not pursued in my thesis.	75
145	The value of the forecast can be much better communicated to users by using measures like shift and distribution. This should be highlighted.	Agree	A sentence has been added highlighting how measures like shift and dispersion can be readily communicated.	146
147	Implications. It would be good to have a paragraph in text that will state the likely limitations of this study and how this could be solved in future.	The next section (5.6) addresses limitations of the study, including around methodology and data, and suggests some follow-up research.	Section 5.6 has been renamed "Limitations and future directions"	148