Post-processing Sub-seasonal to Seasonal Climate Forecasts Under Climate Change

Yawen Shao

ORCID: 0000-0002-9938-669X

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

January 2022

Department of Infrastructure Engineering

Faculty of Engineering and Information Technology

The University of Melbourne

Australia

Abstract

For managing the impacts of climate variability and change, climate outlooks on sub-seasonal and seasonal timescales are attracting more interest from climate-sensitive communities, such as water resource management, agriculture, and energy. With a profound knowledge of the sources of climate predictability, modelling techniques are rapidly developed for forecasting future climate conditions. Recent advancements are dynamical global climate models (GCMs), which typically integrate atmosphere, land surface, ocean, and sea ice components to comprehensively simulate earth climate system and output a wide array of climate forecasts. However, GCMs usually suffer from long-standing modelling issues, such as systematic errors and the failure of reproducing the observed trends in seasonal climate forecasts. Statistical post-processing techniques are frequently employed to improve forecast performance. Many commonly used methods are found to be effective at removing biases, maximising forecast skill, and improving forecast reliability in terms of ensemble spread, but they are seldom designed to resolve the trend disparity issue in the post-processed climate forecasts. This issue should not be neglected as global and regional land surface temperature and precipitation have shown discernible temporal trends over recent decades. To address this gap, the overarching objective of this thesis is to develop and demonstrate the merit of a new, trend-aware forecast post-processing method that eliminates the trend disparity between climate forecasts and observations while making the forecasts bias-free, skillful, and reliable.

The first part of this research aims to develop a new statistical post-processing method to embed the observed trend into seasonal temperature forecasts. I extend the capability of a calibration method, the Bayesian joint probability (BJP) modelling approach, by introducing a new trend component into the algorithm. The new model (named BJP-t) is applied to calibrate January mean forecasts of daily maximum temperatures from the SEAS5 seasonal forecasting system, operated by the European Centre for Medium-Range Weather Forecasts (ECMWF) in three test stations in Australia. In these cases, the BJP-t calibrated forecasts are shown to accurately reproduce the observed trends, and are more skillful, more reliable, and sharper than raw and BJP calibrated forecasts.

In the BJP-t model, the trend is entirely inferred from the training data. In practice, given limited available periods of retrospective forecasts for model training, these inferred trends are subject to large sampling errors, and may not reflect true underlying trends in the observations. Accordingly, the second part of my thesis further develops the BJP-t model to account for trend uncertainty. The extended trend-aware forecast post-processing method is applied to SEAS5 seasonal mean minimum and maximum temperature forecasts, and the evaluations are upscaled to the Australian continent. After trend-aware post-processing that deals with trend uncertainty, forecast trends are more robustly inferred than the BJP-t model. Compared to the BJP calibrated forecasts, embedding trends lead to greater forecast accuracy in regions where observed trends are significant or where observed trend direction is wrongly represented in the BJP calibrated forecasts.

The third part of my thesis aims to extend the trend-aware method for post-processing seasonal forecasts of precipitation, which is also a key variable for agriculture and water resource management. Several modifications are made in the model algorithm and evaluation tools to cater for the special features of precipitation amounts, including zero occurrences, highly positive skewness, as well as higher variations and larger uncertainty than temperature variables. I apply this improved trend-aware method to calibrate SEAS5 seasonal precipitation forecasts over Australia. Evaluations show that the trend-aware calibrated forecasts properly reproduce observed trends over the hindcast period of 36 years. In some regions with significant observed trends, skill improvements against the BJP calibrated forecasts are evident by embedding trends into the forecasts. Overall, in most regions, the trend-aware calibrated forecasts outperform raw forecasts with respect to bias, skill, and reliability.

Operational sub-seasonal climate forecasts are produced by GCMs configured not dissimilar to seasonal forecast models, but little attention has been paid to explore the ability of the sub-seasonal forecasting systems to capture the observed trends. The fourth part of my thesis firstly aims to investigate this question. Preliminary results show that the same trend disparity issue exists in the 20-year weekly averaged retrospective temperature forecasts from the ECMWF extended-range forecasting system, particularly beyond the first week. Subsequently, I adapt the trend-aware method to calibrate and correct the trend in sub-seasonal forecasts. I modify the method to embed a 30-year climate trend into the 20-year calibrated forecasts. The embedded trends are therefore robustly representative of long-term climate changes and overcome the problem that trends inferred from a shorter period may be subject to large sampling variability.

Calibration is applied to the ECMWF sub-seasonal minimum and maximum temperature forecasts for Australia with forecast horizons of up to 4 weeks. Results reveal that raw week-1 forecasts exhibit trends consistent with the 20-year observations in many regions while raw week-4 forecasts do not show the trends of the 20-year observations during the hindcast period. After trend-aware post-processing, trends in calibrated week-1 forecasts are roughly aligned with the 20-year observations across Australia, because when raw forecasts are inherently skillful, the trend-aware calibration transfers raw forecast skill and embeds the 20-year apparent observed trends into the calibrated forecasts. For comparison, calibrated week-4 forecasts exhibit the trends of the 30-year observations, because when raw forecasts do not have much skill, the trend-aware calibration reverts the forecasts to the 30-year observed climatology with trends. In general, the trend-aware calibrated forecasts are more reliable than raw forecasts, while being as skillful as or more skillful than raw forecasts for all lead times.

The new trend-aware forecast post-processing method shows robustness for resolving the trend disparity issue for GCM sub-seasonal and seasonal climate forecasts. Wider applications of this method have the potential to deliver quality forecasts and build user confidence in deploying the forecasts for decision-making in a changing climate. Further research will adapt the trend-aware method for other hydrometeorological variables.

Declaration

This is to certify that:

i. This thesis is an original work of the author alone, except where due acknowledgement has been made.
ii. The work has not been submitted previously, in whole or in part, to qualify for any other degree or qualification in any other universities.
iii. The content of the thesis is the outcome of the research which has been carried out during the official PhD candidature.
iv. The thesis contains less than 100,000 words in length, exclusive of tables, figures, references, and supplementary materials.

Yawen Shao

Melbourne, January 2022

Preface

The research works included in this thesis are essentially my own work during my candidature over 2018-2021. The journal articles are co-authored by my supervisors, who offered me support for framing the research questions and editing the manuscript. I established the calibration models, carried out data analysis, produced figures and tables, and wrote the entire manuscript. No part of the work was completed outside of my candidature, no editorial assistance was received from external companies, and no material has been submitted for another qualification.

Chapters 2 to 5 are formatted as journal articles separately and have been published or accepted as below.

Chapter 2 has been published as: Shao, Y., Wang, Q. J., Schepen, A., and Ryu, D. (2021). Embedding trend into seasonal temperature forecasts through statistical calibration of GCM outputs. *International Journal of Climatology*, 41(S1), E1553-E1565. http://doi.org/10.1002/joc.6788.

Chapter 3 has been published as: Shao, Y., Wang, Q. J., Schepen, A., and Ryu, D. (2021). Going with the trend: forecasting seasonal climate conditions under climate change. *Monthly Weather Review*, 149, 2513-2522. http://doi.org/10.1175/MWR-D-20-0318.1.

Chapter 4 has been published as: Shao, Y., Wang, Q. J., Schepen, A., Ryu, D., and Pappenberger, F. (2022). Improved trend-aware post-processing of GCM seasonal precipitation forecasts. *Journal of Hydrometeorology*, 23(1), 25-37. https://doi.org/10.1175/JHM-D-21-0099.1.

Chapter 5 has been accepted as: Shao, Y., Wang, Q. J., Schepen, A., Ryu, D. Introducing Long-term Trends into Sub-seasonal Temperature Forecasts through Trend-aware Post-processing. *International Journal of Climatology*. https://doi.org/10.1002/joc.7515.

Acknowledgements

Firstly, I would like to profoundly acknowledge my supervisors, Professor Q. J. Wang, Associate Professor Dongryeol Ryu, and Dr. Andrew Schepen for their guidance throughout my research journey. Many thanks to Q. J. for valuable ideas and resourceful support to help shape my research thinking. Despite heavy workloads, Q. J. always provides timely responses to my research works and requests. Thanks to Dongryeol for inspiring me to become a researcher. Thanks to Andrew for useful discussions and constructive comments on my research works. Thanks also to Professor Stephan Winter for hosting all my review meetings as my committee chair. My PhD research is fully funded by Melbourne Research Scholarship from The University of Melbourne, and I really appreciate the financial support during my candidature.

To the members in Water Forecasting group and Melbourne Environmental Sensing and Modelling Lab, Dr. Wenyan Wu, Dr. Qichun Yang, Dr. Yating Tang, Dr. Kirsti Hakala, Mr. Pengcheng Zhao, Ms. Yuerong Zhou, Mr. Charles Yang, Ms. Yiliang Du, Ms. Wen Wang, Ms. Leila Forouhar, Dr. Chihchung Chou, Dr. Naveen Joseph, Dr. Shuci Liu, Ms. Jie Jian, Ms. Anne Wang, Mr. Manish Petal, Mr. Arash Parehkar, Ms. Shirui hao, Ms. Huazhen Li, Ms. Zitian Gao, and many others for interesting conversations in daily life and useful discussions in group meetings. Thanks also to many visiting scholars during my candidature, Dr. Lily Deng, Dr. Li Liu, Dr. Yuan Li, Dr. Zixiong Zhang, Dr. Wentao Li, Dr. Huaping Huang, Dr. Yujie Li, Dr. Zongjie Li, Dr. Shuai Xie, Dr. Shan He, Dr. Jiaming Xu, Dr. LiangLiang Tao, and Associate Professor Xiaomeng Song. Thank you for good company and support for my life and research.

I am also grateful to many other staff and PhD colleagues in the Water group. Many thanks to Dr. Danlu Guo, Dr. Conrad Wasko, Dr. Keirnan Fowler for kindly suggestions and experience sharing. Thanks to Ms. Emma Payne, for helping with administrative and logistic arrangements. Thanks also to other PhD candidates, Mr. Suwash Acharya, Ms. Seema Karki, Ms. Xinyang Fan, and many others, for friendship, encouragement, and technical assistance.

During my PhD study, I worked with the AWRA team at the Australian Bureau of Meteorology for a trend analysis project for three months. Thanks to my project supervisors, Dr. Louise Wilson, and Dr. Elisabeth Vogel, for lively discussions in regular meetings, technical assistance for analysing climate data in the supercomputing system and friendly conversations during and after the project. I would also like to acknowledge Dr. Florian Pappenberger from European Centre for Medium-Range Weather Forecasts for detailed comments on my third paper.

Importantly, I am grateful to have supportive parents, who unconditionally encourage me and instil the confidence in my decisions. Special thanks to my dear friends, Ms. Becky Sun, Ms. Hayley Huang, Ms. Amy Liu, and Ms. Crystal Pang, for life suggestions and warm company on this arduous journey, especially during the COVID-19 period. Without their long-standing friendship, I would not have been brave enough to overcome all the difficulties.

Contents

List of Tables

List of Figures

Chapter 1 Thesis Introduction

1.1 Preamble

This chapter introduces my research backgrounds, motivations, objectives, and thesis structure. In Section 1.2, I explore global and regional temporal trends in land surface temperature and precipitation variables during recent decades. These trends associated with climate variability and change have far-reaching impacts on climate-sensitive sectors. To properly manage climate variability and change, information about the future climate conditions on sub-seasonal to seasonal timescales is appealing to the information user community.

In Section 1.3, I describe that with the identification of the predictable sources of the climate system on sub-seasonal to seasonal timescales, the climate community proceeds with the development of statistical and dynamical climate models. With a more advanced understanding of the climate system and improved computational efficiency, dynamical global climate models (GCMs) are being rapidly evolved and now operated by climate centres worldwide. The resulting climate forecasts have been integrated into the application domain to a large extent. Despite decadal evolution, GCMs are still facing significant challenges in delivering quality forecasts, such as model systematic errors, and the failure of representing observed trend information in retrospective seasonal forecasts. Details about the GCM challenges are presented in Section 1.4.

To overcome these challenges, a wide array of statistical techniques has been developed to improve forecast performance. Section 1.5 investigates downscaling, multi-model combination, and statistical post-processing methods ranging from simple bias corrections to complex calibrations. In the literature, a few studies used or modified the simple quantile mapping for post-processing seasonal forecasts of temperature and sea ice concentration with trend adjustment, but these methods are not sufficiently effective for generating skillful and reliable results. On annual-to-decadal timescales, some complex calibration methods were also extended to correct forecast trends. These methods, however, have not been applied to post-process seasonal temperature and precipitation forecasts, and do not deal with uncertainty of inferred parameters. Consequently, they are not fully adequate for post-processing forecasts on shorter prediction range. A robust calibration method is thereby required to eliminate trend disparity between model forecasts and observations while minimising biases, improving forecast skill and reliability.

Through literature review, I identify the research gaps and conceive the plan to address these issues. Detailed research questions and objectives are introduced in Section 1.6. Finally, Section 1.7 outlines thesis structure.

1.2 Changes in the Climate

1.2.1 Trends in land surface air temperature and precipitation

Global averaged land surface air temperature has markedly increased since the 1970s (Hartmann et al., 2013), at the rate of approximately 0.3℃-0.7℃ per 30 years (IPCC, 2018). Regionally, significant temperature trends have also been detected during different time periods in many parts of the world (Ren et al., 2017; Yu et al., 2018). A warming trend dominated most regions across the South American continent for the period 1975-2004, with the strongest trends taking place in central Brazil (de Barros Soares et al., 2017). During 1980-2017, annual mean temperature continuously increased in China, and the patterns of trend shifts, for example monotonic changes or abrupt changes, varied across the country (L. Li et al., 2019). In Iran, there were a widespread warming trend in annual mean temperature, and a stronger warming trend in spring and summer over 1961-2010. In addition, the minimum temperature increased more rapidly than the maximum temperature (Ghasemi, 2015). With resolution gridded temperature dataset, significant warming trends were shown in annual maximum and minimum temperatures across East Africa over 1979-2010 (Gebrechorkos et al., 2019). In New Zealand, both maximum and minimum temperatures at 22 stations also experienced warming trends over 1965-2010, notably from April to August (Caloiero, 2017).

Long-term trends in Australia's temperature are also in accordance with the global warming climate. Australian annual mean temperature exhibited rapid warming trends for the period 1961-2010, with minimum temperature increasing in a slightly faster rate than maximum temperature (Fawcett et al., 2012). In fact, since 1910, the start of the evaluation period, the Australian continent has shown warming trends, and the warming rate has been accelerated since 1950. Consequently, extreme heat events become more frequent while extremely cold days and nights have the propensity to decline over Australia (CSIRO and Australian Government Bureau of Meteorology, 2020).

Changes in precipitation exhibit considerable spatiotemporal variations over land. Global precipitation is energetically constrained to increase at the rate of 2-3% as with the per degree

increase in global mean temperature (Allan et al., 2020). Specifically, precipitation amounts have increased over the mid-latitude land regions of the northern hemisphere since 1901, and such changes are with high confidence after the 1950s (Jia et al., 2019). Elsewhere, while different mixed increasing and decreasing trends have been detected from different observational datasets, these long-term precipitation trends are generally non-significant due to large uncertainty and high natural variability in records (Hartmann et al., 2013). Regionally, station precipitation records were investigated in North Carolina, United States over 1950-2009. Results revealed that annual, spring and summer precipitation amounts showed both increasing and decreasing trends across the state, and only a small number of stations had significant trends at 5% level (Sayemuzzaman and Jha, 2014). In the China- Pakistan Economic Corridor, again, mixed but mostly non-significant trends were evident in winter and annual precipitation records over 1980-2016 (Ullah et al., 2018) while in most parts of Italy, annual total precipitation and the number of wet days experienced drying trends over 1999-2018 (Caporali et al., 2021).

Australian precipitation has exhibited evident long-term trends in the past decades. On a seasonal timescale, increasing precipitation trends were widespread in the north-eastern part in summer and autumn, and the north-eastern part in spring, while decreasing trends were significant over the south-east in autumn over 1956-2005 (Bhend and Whetton, 2015). For the period 1970-2017, Wasko et al. (2021) also pointed out that an increasing trend dominated summer precipitation in northern Australia, and such trend directly led to a positive change in annual precipitation across this region. Furthermore, decreasing winter precipitation was shown to strongly contribute to the annual precipitation trend in the south-west. In more recent decades (i.e., 2000-2019), northern Australia underwent positive changes across all seasons, particularly in the north-west from October to April. In contrast, there was a declining precipitation trend over a large portion of southern Australia between April and October (CSIRO and Australian Government Bureau of Meteorology, 2020).

1.2.2 Impact of Climate Variability and Change

Marked climate trends exhibited in both land surface temperature and precipitation are in part caused by changes of human activities on the hydrological cycle, for example, irrigation and land use changes (Allan et al., 2020; Jia et al., 2019). Climate trends are also partially ascribed to external forcing, such as greenhouse gas (GHG) emissions, aerosols and solar variability (Hartmann et al., 2013). In future conditions, as GHG emissions and aerosols continue to rise,

climate models project that the land surface air temperature will further increase for over 1℃ in many regions by the end of this century (Kumar et al., 2014; Xu and Xu, 2012). Global mean precipitation will exhibit higher spatial variability around the world and increasing GHGs are expected to accelerate precipitation variability in a warmer world (Allan et al., 2020; Collins et al., 2013; Pendergrass et al., 2017). Australian annual mean temperature is projected to increase for over 2℃ by 2090 relative to 1990, while the precipitation is projected to decline in south-western Australia, which is more pronounced in winter and spring, and the changes in northern and eastern Australia are uncertain (Dey et al., 2019; Irving et al., 2012).

Human and environmental systems are sensitive to the spatial and temporal distribution of the climate variables (Lausier and Jain, 2018). Under the climate change condition, variations in the precipitation and temperature are expected to make negative impacts on global food security (Jia et al., 2019; Lobell et al., 2011). For example, global wheat yields may reduce by approximately 4-6% as with per degree temperature increase, and maize productivity may also be constrained as temperature continuously rises. As such, the yield increase benefited from technology and other factors could be substantially offset due to the climate trends in some regions. The changing climate also significantly affects Earth's terrestrial ecosystems, leading to shifted bioclimate zones (Law et al., 2019) as well as changed sizes, locations, abundances and seasonal activities of plants and animals as climate warms (Esquivel-Muelbert et al., 2019; Fadrique et al., 2018).

Australia has a highly variable climate, and the country is particularly vulnerable to the changing climate (CSIRO and Australian Government Bureau of Meteorology, 2020). In the past, the Millennium Drought occurring in south-eastern Australia from 2001 to 2009 has made far-reaching impacts on climate sensitive sectors, such as agriculture, water resource management and ecosystem (Van Dijk et al., 2013). In the future, more extreme high-impact events have been projected to take place, including longer durations of droughts, and more frequent floods and heat waves (Dey et al., 2019). To provide proactive warnings and alleviate potential risks for the coming weeks and months, accurate outlooks of the climate conditions beyond 14 days weather forecasts are in growing expectations within forecaster and user communities under the climate change condition (Mariotti et al., 2020).

1.3 Forecasting Climate Conditions

Estimates of the future average climate conditions extending from two weeks to up to a year sit on sub-seasonal to seasonal timescales, serving as a middle ground between short-term weather

forecasts and long-range decadal predictions to climate change projections (Bruno Soares et al., 2018; Vitart and Robertson, 2019). Seasonal climate forecasting estimates how average conditions may develop over a season and up to a year while sub-seasonal climate forecasting refers to the climate predictions between two weeks and a season ahead. Expanding the predictive capabilities beyond weather forecasting is expected to seamlessly encompass climate predictions from sub-seasonal to decadal time ranges (Merryfield et al., 2020).

1.3.1 Sub-seasonal and seasonal predictability

Before establishing the climate forecasting system, it is crucial to identify the major sources of climate predictability, and the mechanisms associated with climate variations. Here, climate predictability refers to the extent to which the state of the future climate system can be predicted (Krishnamurthy, 2019). The mechanisms summarised below provide a potential predictable signal for forecasting climate conditions on sub-seasonal and seasonal timescales (Stockdale et al., 2010), while the predictability and the attained level of skill are regime dependent and vary substantially in space.

On a seasonal timescale, climate predictability arises from slowly varying forcing on the atmosphere, such as sea surface temperature (SST) anomalies, which exerts remote impacts from various teleconnections (Vitart and Robertson, 2019). A key source of predictability is El Nino-Southern Oscillation (ENSO), which is a coupled ocean-atmosphere mode of climate variability dominating tropical SST anomalies ascribed to the synergy between SSTs and winds (Stockdale et al., 2010). ENSO SST anomalies are highly predictable within a year, especially in early spring and winter. The ability to explicitly predict this large-scale ocean-atmosphere phenomenon directly influences the forecast skill of the climate variables (e.g. seasonal surface air temperature and precipitation) worldwide through global teleconnections (Merryfield et al., 2020).

Apart from ENSO, other internal and external mechanisms also offer the possibility of a predictable signal on seasonal timescales. The tropical SST anomalies across the Atlantic and Indian Oceans could drive a wide array of teleconnections, while a number of ocean climate indices are linked to main patterns of climate variability in the surrounding regions (Doblas-Reyes et al., 2013). One example is Indian Ocean Dipole, which can develop independent variability from ENSO, and be predictable up to 6 months in advance (Saji et al., 1999). Elsewhere, the North Atlantic Oscillation (NAO) is a major source of variability in the midlatitude North Atlantic and Europe in winter (Gong et al., 2002; Merryfield et al., 2020), whose negative phase could result

in skillful predictions (Ferranti et al., 2015). Stratospheric variability and interactions between the troposphere and stratosphere, for example quasi-biennial oscillation (Marshall and Scaife, 2009) and sudden stratospheric warmings (Marshall and Scaife, 2010), are also possible drivers for seasonal predictions. Some land processes, such as soil moisture (Ardilouze et al., 2017) and snow cover (Walsh and Ross, 1988), are associated with seasonal forecast skills in some regions. As documented, soil moisture could significantly contribute to the forecast skill of summer surface temperature with a lead time of up to 2 months in North America, equatorial Africa, and South America (Doblas-Reyes et al., 2013). Despite a non-stationary relationship, snowpack was found to influence large-scale atmospheric circulation patterns, such as NAO (Orsolini et al., 2011).

Seasonal predictability is also modulated by forcing external to the climate system. Variations in atmospheric composition, such as GHG and aerosol concentrations are crucial sources of non-stationary climate system (Doblas-Reyes et al., 2013; Merryfield et al., 2020). Other potential sources include solar variability (Misios et al., 2019), and unusual volcanic eruptions (Ménégoz et al., 2018), which additionally exert impacts on ENSO.

Regarding sub-seasonal predictability, one major global source is Madden-Julian Oscillation (MJO), an organized convective activity (Robertson and Vitart, 2018). With diverse amplitudes and phases, MJO modulates tropical, as well as middle and high latitude phenomena, such as tropical cyclone (C. Zhao et al., 2019) and East Asian summer monsoon (Li et al., 2018). Skillful sub-seasonal forecasts for some extratropical phenomena, such as hail and tornado activities (Baggett et al., 2018), can be produced using certain phases of MJO through tropical-extratropical teleconnections (Lin et al., 2019). Another skill source arises from the tropical and extratropical stratosphere-troposphere interactions (Butler et al., 2019). For instance, in the extratropic, the climatological westerly winds could temporarily turn easterly in the Northern Hemisphere due to Rossby wave breaking. The resulting extreme events are known as sudden stratospheric warmings, which have pronounced impacts on surface temperature and precipitation.

The interactions between land and atmosphere also significantly affect the forecast skill. Examples are soil moisture, soil temperature anomalies, and snow cover (Merryfield et al., 2020). Memory in soil moisture is in close relation to sub-seasonal temperature and precipitation forecasts across some regions and seasons through the changes in the evaporation and surface energy budget (Vitart and Robertson, 2019). Other potential sources of sub-seasonal predictability include atmosphere-ocean interactions, and sea ice over polar and possibly midlatitude regions

(Merryfield et al., 2020). The tropical SST anomalies associated with ENSO, and extratropical SST anomalies are capable of enhancing the skill of sub-seasonal forecasts through teleconnections (DelSole and Tippett, 2016; McKinnon et al., 2016).

1.3.2 Statistical and dynamical forecasting methods

In recognition of the sources of sub-seasonal and seasonal predictability, over time, both statistical and dynamical approaches have emerged and advanced to issue climate outlooks. Statistical forecasting methods establish the empirical relationship between different climate variables, for example SST and precipitation, based on observational data records, and such relationship is then employed for real-time forecasting (Stockdale et al., 2010). Dynamical climate models generally implement comprehensive general circulation modules to represent the chaotic nature of the climate system (Troccoli, 2010). The statistical and dynamical methods are complementary in terms of their advantages and disadvantages.

In statistical forecasting approaches, large-scale atmospheric circulation features, for example SST variability in the tropical Pacific, have been frequently used as the predictor (Doblas-Reyes et al., 2013; Schepen et al., 2012b). Currently, statistical methods are still in wide use, and can produce skillful forecasts in some regions where dynamical models exhibit very limited skill. For example, Tuel and Eltahir (2018) made use of robustly selected SST indices to empirically forecast winter and spring precipitation in Morocco. Results showed that with this method, approximately 35-40% of interannual variability could be predicted. In practice, building a statistical model does not require extensive computing resources, and it is straightforward to apply the consecutively obtained knowledge of climate variability to the model development. However, cautions are needed for the applications of statistical forecasting techniques because the length of available observational data is short and these forecasting techniques have limited capability for predicting unprecedented climate conditions, including long-term changes in the non-stationary climate system (Doblas-Reyes et al., 2013; Stockdale et al., 2010).

Dynamical climate models are designed to accurately represent large-scale climate drivers, notably the relevant predictability mechanisms described in Section 1.3.1, with the aim of skillfully forecasting a broad spectrum of climate variables. Over time, dynamical models have evolved from atmospheric general circulation models to coupled ocean-atmosphere general circulation models, which are now more frequently adopted for providing a detailed global view of the future climate state. The current generation of these global climate models (GCMs)

generally integrate ocean, atmosphere, land surface, and sea ice components, and also account for atmospheric chemistry, such as GHGs, aerosols and ozone (Vitart and Robertson, 2019). GCMs often output an extensive set of retrospective forecasts (abbreviated as re-forecasts) for several objectives: producing relevant products such as anomalies, assessing forecast skill and reliability, and post-processing forecasts against observations to improve forecast performance (Stockdale et al., 2010). Complementary to statistical forecasting methods, dynamical models have the advantage of properly predicting unprecedented climate conditions in a non-linear manner. Furthermore, ensemble techniques are harnessed in the current generation of GCMs to produce an ensemble of climate forecasts using a set of slightly different initial conditions to reflect the uncertainty in them (Schepen and Wang, 2014). However, substantial computing resources are required for running GCMs, and these models always suffer from systematic errors and other challenging issues, even with the use of the latest technology (Stockdale et al., 2010). The challenges in GCMs will be discussed in the later Section 1.4.

With an ongoing development, many climate centres are now operating state-of-the-art climate modelling systems to routinely issue weekly to seasonal climate outlooks to the public. ECMWF hosts and offers extended-range forecasts to the sub-seasonal database of a collaborative Subseasonal-to-Seasonal Prediction project established by the World Climate Research Programme (WCRP) and the World Weather Research Programme (WWRP). Real-time, and re-forecasts from eleven operational climate centres are now stored and updated in this sub-seasonal dataset. The Australian Bureau of Meteorology executes and continuously advances Australian Community Climate Earth-System Simulator–Seasonal (ACCESS-S1) (Hudson et al., 2017; Hudson et al., 2018), a multi-week to seasonal forecasting system that replaced POAMA (Hudson et al., 2013; Marshall et al., 2014) in August 2018. In the United States, the National Centres for Environmental Prediction (NCEP) leads the development of Climate Forecast System, version 2 (CFSv2) (Saha et al., 2014) for operational sub-seasonal and seasonal predictions since March 2011, and operates the North American Multi-Model Ensemble (NMME) (Kirtman et al., 2014), a seasonal forecasting system primarily consisting of seven independent GCMs. In Europe, SEAS5 long-range forecasting system (Johnson et al., 2019) was made operational in November 2017, designed and run by the European Centre for Median-Range Weather Forecasts (ECMWF) for seasonal climate forecasting. There are several highlights for this new development. The seasonal re-forecasts from SEAS5 cover the period of 1981-2016, which is currently the longest hindcast period among the operational GCMs for assessing the past performance. Compared with its predecessor System4, the SEAS5 model implements higher resolutions of the atmosphere and

ocean models and includes a prognostic sea ice model. In other countries, multiple climate agencies also put efforts in operating their own climate forecasting systems, such as UK Met Office, Japan Meteorological Agency (JMA), and Beijing Climate Centre (BCC).

1.3.3 Applications of sub-seasonal and seasonal forecasts

Skillful climate forecasts are valuable for various climate-sensitive sectors because many management decisions in response to climate variability and change are made on sub-seasonal to seasonal forecasting timescales (Bruno Soares et al., 2018; Merryfield et al., 2020). Practically, GCM-based climate forecasts, such as precipitation and temperature forecasts, are often used as the inputs to hydrological prediction systems for ensemble streamflow and soil moisture forecasting (Bennett et al., 2017; W. Li et al., 2019; Shah et al., 2017; Vogel et al., 2021; Yuan and Wood, 2012; Yuan et al., 2015), and further benefit the ongoing research of dynamical drought prediction (Hao et al., 2018), and flood forecasting (White et al., 2015). The accurate prediction of these hydroclimatic extremes has the potential to provide early warnings for proactive risk management. For water resource management, operational reservoir supply systems could be informed by GCM forecasts for desirable decisions (Peñuela et al., 2020; Viel et al., 2016). In the agricultural sector, crop models, such as the Agricultural Production Systems sIMulator (APSIM), have been developed to produce yield forecasts, and their key inputs are climate variables, including temperature, precipitation, solar radiation, and evaporation. In recent years, there is a surge of interest in applying the climate forecasts from the current generation of the GCMs to predict crop production (Brown et al., 2018; Ogutu et al., 2018; Schepen et al., 2020b), and seasonal climate forecasts have shown promise for improving productivity and profitability in the crop industry (An-Vo et al., 2019; Klemm and McPherson, 2017; Parton et al., 2019). With increased forecast skill across ocean, another plausible application field is fisheries management associated with living marine resources (Tommasi et al., 2017). Informative SST forecasts have the potential to increase fishery yields while protecting the fish population from a sudden decline due to the overfishing and environmental changes. Fire predictions (Turco et al., 2018), energy (Troccoli, 2018), food security (Funk et al., 2019) are some other societal sectors that climate forecasts are expected to assist in coping with climate variability and change.

Despite the potential value and benefits of using GCM forecasts in decision-making processes, forecasts alone are not sufficiently informative without the engagement from decision makers (Bruno Soares et al., 2018). For example, the end user may wish to uptake forecast information

on finer spatial scales than the GCM resolution. Under such circumstances, forecasts that are not properly translated to sub-grid or regional locations for user-oriented demands are deemed unusable, even though they are proven highly skillful and reliable on a coarser spatial scale. Indeed, it is still a challenge now to link research, operational forecasting, and end-user needs. More effective communication, interaction and co-evaluation with the forecast user community are needed to allow for a successful incorporation of the forecasts into the decision-making contexts (White et al., 2017).

1.4 Challenges in Climate Modelling

For informative decision-making in the application domain, the GCM-based forecasts are expected to be in good correspondence with the observations and to have high forecast skill. However, GCMs developed for sub-seasonal to seasonal forecasting are typically associated with large systematic errors (Doblas-Reyes et al., 2013; Merryfield et al., 2020). Furthermore, GCM seasonal forecasting systems were reported not to reproduce climate trends in retrospective experiments (Huang et al., 2019; Krakauer, 2019). Multiple reasons contribute to these modelling issues.

1.4.1 Model systematic errors

Given complexities of climate dynamics and restricted computational resources, GCMs are often set up with simplified fluid dynamic equations, and coarse temporal and spatial resolutions with key climate variables numerically solved through parameterisation. Despite decades of evolution, GCMs are still subject to imperfections due to their inability to perfectly model the real world (Doblas-Reyes et al., 2013). Examples for associated systematic errors include incoherent intra-seasonal variability (Shibuya et al., 2021), and a poor representation of the organised tropical convection and cloud processes (Ma et al., 2021; Tan et al., 2018). These long-standing modelling issues could lead to biases in surface temperatures and precipitation, and further degrade forecast skill. In addition, as forecast time increases, model systematic errors begin to develop and result in model drifts from the observed climate associated with the initial conditions to the model biased climate. The mechanisms behind the drifts are difficult to explain because the interactions among parameterized physical processes are non-linear and the origin of the biases may be non-local (Merryfield et al., 2020). Other than these persistent errors, climate models focusing on sub-seasonal or seasonal timescales have their own limitations. One crucial challenge for sub-seasonal

modelling is the prediction of the MJO, which is a key source of sub-seasonal predictability globally and regionally. Current sub-seasonal models generally poorly simulate MJO propagation over the Maritime Continent, such as the Indian Ocean and western Pacific (Lim et al., 2018). For seasonal climate modelling, it remains a challenge to explicitly model the teleconnection patterns between climate variables and relevant climate modes, say ENSO. Current efforts are mainly devoted to making GCMs accurately represent these climate modes, and to alleviating model biases in the ocean variation that could take months or years to develop (Merryfield et al., 2020).

Another challenge leading to model errors is the initialisation issue. Once a GCM is setup, it is a necessary step to realistically initialise all the primary components of the climate system with assimilated observations, including atmosphere, ocean, land surface, and sea ice modules (Doblas-Reyes et al., 2013; Merryfield et al., 2020). The initialisation of the atmospheric component has been advanced for several decades. Sophisticated data assimilation systems have been relatively well established to produce reanalyses and initialise the re-forecasts with high accuracy by assimilating historical observational information and possibly satellite products (Bauer et al., 2015; O'Kane et al., 2019). Meanwhile, available atmospheric observations are dense and generally span over a decade, which is an advantage in implementing forecast post-processing and quality assessment. By contrast, further research is required to precisely initialise the ocean and sea ice (Blanchard-Wrigglesworth et al., 2011; Zampieri et al., 2018), as ocean initial conditions are crucial for skillful predictions of ENSO. As an aisle, one major limitation for ocean initialisation is that the observational dataset generally lasts for less than a decade, posing a challenge in periodically verifying forecast quality (Doblas-Reyes et al., 2013; Hackert et al., 2019). Efforts are also required for the accurate representation of the coupled processes among model components. Conventionally, individual components are initialised separately without coupling, which could lead to initialisation shock arisen from the imbalances in initial states and caused by the limitations of observations and model biases. As a result, initial information is biased, and possibly forecast skill is degraded (Mulholland et al., 2015). To resolve this issue, weakly and strongly coupled methods - new generations of initialisation systems - are being developed and have shown promise for eliminating shocks (Penny and Hamill, 2017). Weakly coupled methods assimilate each component separately within the same coupled model to allow for information exchange across the interfaces, and such methods have been applied operationally. Strongly coupled methods are now in an experimental stage, which simultaneously assimilate multiple model components so that assimilated observational information in different components is interacted (Penny et al., 2019).

The initial conditions of the climate may also include uncertainty that gives rise to biased model results. As mentioned, ensemble techniques have burgeoned to generate a range of independent outcomes with slightly different perturbed initial conditions to convey the uncertainty. However, the ensemble spread of the resulting forecasts is often overconfident, indicating that the forecast range is too narrow to represent the true likelihood of the observations (Weisheimer and Palmer, 2014). Apparently, innovative ensemble generation techniques are required to reliably quantify the uncertainty in the initial conditions and the resulting forecasts.

1.4.2 Trend mismatch issue

Apart from the long-standing model systematic errors, the inability of GCM seasonal forecasting systems to represent historical trends of climate variables is also an unresolved issue that requires attention. The assessment of the forecast trends in multiple works suggested that GCMs poorly represent the historical climate trends in some regions. In earlier works, Boer (2009) found that trends in the ensemble mean forecasts from the second Historical Forecasting Project multi-model two-tier seasonal forecasting system were much weaker than trends in NCEP reanalysis data of the land surface temperature. However, his research might not be sufficiently instructive for the general conclusions because this old generation of the climate model did not integrate land surface and sea ice components. Cai et al. (2009) investigated the surface air temperature trends in the re-forecasts produced by the NCEP CFS version 1 (CFSv1) over 1982-2006 to answer how well the global warming signal was captured in seasonal temperature forecasts. Results showed that with the level of GHG concentrations fixed at the year 1998, the forecast trend over-estimated and under-estimated the warming trend before and after 1988, implying that a realistic representation of the anthropogenic GHGs in the seasonal forecasting system was essential for accurate forecasts. Compared with CFSv1, in CFSv2, the upgraded version of CFSv1, the warming trends of winter temperatures were more realistically simulated in the forecasts across all the continents. However, the increasing trends in India winter temperatures were over-estimated (Nageswararao et al., 2016) and the temperature trends in South America summer were under-estimated (Silva et al., 2014). In another operational climate modelling system, NMME, the warming trends in monthly mean temperature forecasts for both multi-model ensemble means and individual model forecasts were weaker than the trends in observations at all lead times (Krakauer, 2019). More recently, Duan et al. (2021) evaluated the trends in spring temperature forecasts produced from the BCC-CSM1.1 (m) seasonal forecasting model operated by BCC. The significant warming trends in the observations were under-estimated by the model forecasts over 1991-2020 across most parts of

China. While most studies focused on the trend evaluation in temperature forecasts, the ability of the precipitation forecasts to capture the observed changes is also worth investigation. Huang et al. (2019) analysed trends in CFSv2 U.S. seasonal precipitation re-forecasts with 2-month lead time for the period 1958-2017 and demonstrated that the re-forecasts roughly reproduced the observed spring and summer trends during 2000-2017, as well as the observed winter trends in the entire 60-year period. In contrast, the observed precipitation changes were poorly recognised by CFSv2 for spring, summer, and autumn over 1958-1999.

Several factors may contribute to the misrepresentation of the multi-decadal changes in the GCM re-forecasts of temperature and precipitation variables with different forecast lead times. Jia et al. (2013) attempted to explain why the second phase of the Canadian Historical Forecasting Project (HFP2) seasonal forecasting system under-estimated the winter trend in observed surface air temperature across the Eurasian continent. Both initial conditions and SST anomaly forcing used in the HFP2 significantly contributed to the trend underestimation. Trends of the initial conditions are associated with the forecast trends in the first month, while the trends of SST anomaly forcing predominately influence the second and third months. Note that these two trend sources are not completely independent as the initial conditions are indirectly affected by SSTs. Other mechanisms, such as the lack of the GHG forcing, poor representation of the sea ice and snow cover in the HFP2, and some physical causes (Boer, 2009), could partially explain the trend mismatch problem in the HFP2 forecasts. Moreover, trends in SSTs and GHGs may also contribute to the long-term trends in the ensemble mean of 2-month lead precipitation forecasts (Huang et al., 2019). Great efforts have been put to improve the configuration of GHG concentrations and other atmospheric chemicals in current operational GCMs (Johnson et al., 2019; Saha et al., 2014). The modelling of these components is expected to be more realistic in the next generation of the climate model (Meinshausen et al., 2017). Another potential factor causing the trend mismatch issue is the drifts in the seasonal forecasting system. After initialisation, forecasts may fast drift away from the observed state in the first month, which is likely caused by the initialisation errors. Some drifts may slowly evolve to bias the trend of sub-seasonal to seasonal forecasts (Hermanson et al., 2018).

Several studies have revealed that climate trend information is a contributor to the skill of seasonal climate forecasts and enhanced seasonal predictability in many instances (Doblas-Reyes et al., 2013). Globally, the improved skill of surface temperatures is probably ascribed to a better representation of the historical trends in the climate model (Barnston et al., 2010; Boer, 2009;

Doblas-Reyes et al., 2006). In Europe, Weisheimer et al. (2011) showed that the hot summers across Southern Europe were highly predictable partially because the climate model was capable of reproducing the pronounced warming since the 1980s. Jia et al. (2014) demonstrated that the climate trend in the atmosphere enhanced seasonal forecast skill in the Northern Hemisphere, particularly across the Eurasian continent. Duan et al. (2021) revealed that the skill of spring temperature forecasts over China was improved after trend bias was corrected. In contrast, precipitation skill was found to rarely benefit from the explicit representation of the observed trends in the climate model. This is because the precipitation trend is generally weaker than temperature and varies a lot in space and time (Prodhomme et al., 2016). However, updating, and extrapolating climate trends into the forecasts is still possible to achieve more skills in seasonal precipitation forecasts in some regions. Krakauer et al. (2013) improved climatology forecasts of seasonal precipitation over 1995-2012 by updating its climatological probability distribution each year based on the ratio between the slope of climate trends and magnitude of the interannual variability of observations over the preceding years. They then combined the updated climatology forecasts with Climate Prediction Centre (CPC) seasonal forecasts and found the skill of combined precipitation forecasts roughly doubled the skill of sole CPC forecasts. This is not surprising as there was not much skill in CPC precipitation forecasts. This trend extrapolation method, however, is not a standard practice to directly incorporate trend information into the forecasts. More investigations are needed to evaluate the benefits of resolving the trend mismatch issue in climate forecasts.

1.5 Post-processing Sub-seasonal to Seasonal Climate Forecasts

As described in Section 1.4, GCM outputs generally suffer from large systematic errors and fail to properly represent historical trends. Consequently, raw ensemble forecasts are often misleading, and are not suitable for direct delivery to the public or for driving the following quantitative models. Other than an ongoing development of GCMs, several categories of statistical approaches serve as time-saving and cost-effective ways to address GCM challenges (Li et al., 2017; Yang et al., 2020). Commonly used approaches are statistical downscaling, multi-model combination and statistical post-processing methods.

Downscaling is often required to generate local-scale climate forecasts requested by the stakeholders from raw GCM outputs that are in coarse resolution (Stockdale et al., 2010). Spatial and temporal downscaling is mostly achieved by the statistical techniques (Jacob et al., 2020;

Nury et al., 2019; Schepen et al., 2019), such as analogue downscaling (Hwang and Graham, 2013; Shao and Li, 2013), nonhomogeneous hidden Markov model (Pineda and Willems, 2016), and weather generator (Han et al., 2017). Dynamical methods, such as regional climate models, are also readily available for the downscaling purpose, but they are computationally expensive and do not have obvious advantages over statistical methods. This is because statistical downscaling or post-processing may still be needed to remove systematic errors after dynamical downscaling (Xu et al., 2019). Multi-model approaches combine the strength of ensemble forecasts from different sources of forecasting systems to prepare a single set of reliable ensemble forecasts for stakeholders. Pioneering works, which are still in use now, combined models with equal weights or unequal weights determined by multiple linear regression method (Doblas-Reyes et al., 2005; Pegion et al., 2019). Multi-model combined forecasts were found on average more skillful than the forecasts from a single model. A more generalised methodology is Bayesian method (Luo and Wood, 2008; Raftery et al., 1997; Wang et al., 2012a). The Bayesian model averaging technique has been demonstrated effective for merging forecasts from multiple models while producing more skillful and reliable results than the approach with two predictors selected as a priori and the method selecting only one best-performed model (Wang et al., 2012a).

Although downscaling and multi-model combination aim to offer the user with tailored forecasts, both are not designed to generate well-calibrated forecasts that are bias-free, reliable in ensemble spread, as skillful as the climatology forecasts, and have climate trend information embedded (Zhao et al., 2017). In fact, downscaling and multi-model approaches often complement statistical post-processing methods for launching user-oriented forecast products. Statistical post-processing serves as a straightforward tool to align the forecasts with observations, aiming at producing highly performed forecasts. Existing methods are devised with a different level of complexity, ranging from simple bias corrections to full calibrations.

Simple mean corrections (also named linear scaling) and quantile mapping (QM) are two commonly used post-processing techniques to rectify the model biases in the sub-seasonal (Baker et al., 2019; Monhart et al., 2018) and seasonal climate forecasts (Jha et al., 2019; Lucatero et al., 2018; Wang et al., 2019; Wood et al., 2002), credited by their simple concepts and flexible implementations.

With simple mean corrections, an additive correction is applied to temperature and evapotranspiration while a multiplicative correction is employed for precipitation (Wang et al., 2019). The formulas are given as

For temperature and evapotranspiration,

$$O(t^*) = y_1(t^*) + \left[\frac{1}{T}\sum_{t=1}^{T} O(t) - \frac{1}{T}\sum_{t=1}^{T} y_1(t) \right] \tag{1-1}$$

For precipitation,

$$O(t^*) = \frac{\sum_{t=1}^{T} O(t)}{\sum_{t=1}^{T} y_1(t)} y_1(t^*) \tag{1-2}$$

where $O(t)$ is observation, $y_1(t)$ is raw ensemble forecast mean, T is the number of events, $y_1(t^*)$ is a new raw forecast and $O(t^*)$ is a bias-corrected forecast.

Since simple mean corrections are only designed for removing biases in the mean value, the methods cannot always rectify skill deficits of ensemble forecasts, resolve the under-dispersion issue in the ensemble spread, or deal with the trend mismatch issue (Wang et al., 2019). Lucatero et al. (2018) employed linear scaling and QM to post-process daily ensemble forecasts of temperature, precipitation, and reference evapotranspiration ($\mathrm{E_{To}}$) in Denmark from ECMWF System 4 forecasting system. With both post-processing methods, the forecast skill of temperature and $\mathrm{E_{To}}$ was mildly improved at 0-month lead time, while the skill of precipitation and the forecasts at longer lead times barely increased. For comparison, QM performed slightly better than linear scaling in correcting the ensemble spread, and more accurately estimating the number of dry days in a calendar month. Given its simple setup and widespread popularity, QM is being investigated and integrated into the operational or experimental post-processing systems for agricultural and hydrological modelling (Anghileri et al., 2019; Grillakis et al., 2018; Jha et al., 2019; Kolachian and Saghafian, 2019; Schepen et al., 2020c). It is evident from its general formulation that QM performed better predominately in instances when there existed a strong linear relationship between ensemble means and observed data (Lucatero et al., 2018; Zhao et al., 2017),

$$y_2 = F_O^{-1}\left(F_f\left(y_1\right)\right) \tag{1-3}$$

where y_1 is a raw forecast, y_2 is a post-processed forecast, F_f is the cumulative distribution function (CDF) for all raw forecasts in the training period, and F_O^{-1} is the inverse function of F_O that is the CDF for all paired observed data.

Recent efforts have been made to use or extend QM for forecast post-processing with trend adjustment. Duan et al. (2021) proposed a spatial disaggregation and detrended bias correction (SDDBC) method, extended from a spatial disaggregation and bias correction (SDBC) method, to spatially downscale, bias-correct, and trend-correct spring air temperature forecasts from the BCC-CSM1.1 (m) model. In this method, after spatial downscaling, model forecasts and observations were detrended before QM was applied for forecast post-processing, and the observed trends were then added back to the post-processed forecasts. Results showed that after SDDBC post-processing, forecast trends were consistent with the observations, and the forecast skill was higher than raw and SDBC post-processed forecasts over China. Elsewhere, Dirkson et al. (2019) introduced a trend-adjusted quantile mapping method for improving the seasonal forecasts of local sea ice concentrations from the Third Generation Canadian Climate Coupled Global Climate Model (CanCM3) while considering the historical trends of sea ice concentrations over the period 1981-2017. This extended technique was proven effective at rectifying biases and improving forecast reliability. Despite successful applications, both SDDBC technique for temperature post-processing and extended QM method for sea ice concentration post-processing do not overcome the inherent limitation of the QM algorithm as shown in Eq. (1-3). That is, the correlation between raw forecasts and observations is not fully considered, so that spatial, temporal, and inter-variable relationships from the raw forecasts are explicitly inherited that are not always correct (Zhao et al., 2017). Frequently, QM tends to produce forecasts with skill deficiencies and unreliability in terms of ensemble spread (Lucatero et al., 2018; Manzanas et al., 2019).

The drawbacks of the bias correction methods suggest that more sophisticated methods are needed to guarantee that post-processed forecasts are reliable and offer more skill than climatology forecasts. To achieve this objective, full calibration methods have been widely developed for post-processing ensemble climate forecasts in recent years, which in this thesis refer to a collection of statistical post-processing models that explicitly considers the correlation between raw forecasts and corresponding observations in the calibration steps. Examples for such methods available on seasonal timescales are climate conserving recalibration (Hemri et al., 2020; Weigel et al., 2009), the ratio of predictable components (Eade et al., 2014), ensemble regression (Unger et al., 2009), a post-processing method using temporally and spatially smoothed statistics (Kharin et al., 2017), ensemble model output statistics (EMOS) methods, such as non-homogeneous Gaussian regression (Gneiting et al., 2005; Manzanas et al., 2019; Wilks, 2006a), and Bayesian methods (Khajehei and Moradkhani, 2017; Wang and Robertson, 2011; Wang et al., 2019). Manzanas et

al. (2019) comprehensively compared two categories of post-processing methods, simple bias-correction models and ensemble calibration models, for post-processing 1-month lead boreal precipitation and temperature forecasts in winter and summer from four operational seasonal forecasting systems. All the models effectively eliminated large model biases while the ensemble calibration models improved forecast reliability in some regions beyond what was achieved through simple bias corrections. On a sub-seasonal timescale, some of the aforementioned full calibration methods have been employed to post-process sub-seasonal forecasts of multiple climate variables (Li and Jin, 2020; Li et al., 2020; Schepen et al., 2018; Specq and Batté, 2020; van Straaten et al., 2020). Meanwhile, recent works have also proposed sophisticated machine learning techniques to effectively post-process sub-seasonal temperature and precipitation forecasts, such as artificial neural networks (Fan et al., 2021; Scheuerer et al., 2020), and natural gradient boosting (Peng et al., 2020; Scheuerer et al., 2020).

Although full calibration methods have been proven effective for eliminating systematic errors of the GCM forecasting system and making the forecasts skillful and reliable, these methods are not formulated to resolve the issue that the retrospective forecasts from GCM seasonal forecast models do not reproduce observed climate trends. Elsewhere, on annual-to-decadal timescales, some full calibration methods were extended to fix the long-term trend disparity between model forecasts and observations (Kharin et al., 2012; Pasternack et al., 2021). Sansom et al. (2016) proposed a general recalibration framework to calibrate annual forecasts of CanCM4 annual mean near-surface temperature. This framework extended the EMOS method to include the adjustment of linear time-dependent biases in the forecast means. Pasternack et al. (2018) introduced additional parameters into a joint method of non-homogeneous Gaussian regression and climate conserving recalibration to account for lead-time-dependent errors and long-term climate trends in the decadal forecasts of surface temperature from the Max Planck Institute Earth System Model in a low-resolution configuration. In both works of Sansom et al. (2016) and Pasternack et al. (2018), the extended methods outperformed the reference models with respect to forecast skill. Despite effectiveness, these extended full calibration methods have only been applied to post-process annual-to-decadal temperature forecasts and have not been used for correcting the trend in seasonal forecasts of temperature and precipitation, which are frequently demanded variables within the user community. Moreover, these calibration methods ignore the effect of the uncertainty in the inferred parameters. This could become more problematic for sub-seasonal to seasonal forecasts because the period of available forecasts for calibration and validation is

relatively short. Consequently, the uncertainty can be large and unneglectable, particularly in this changing and uncertain climate.

Another popular full calibration method is Bayesian joint probability modelling (BJP) approach (Wang and Robertson, 2011; Wang et al., 2009; Wang et al., 2019), which has been proven effective for post-processing seasonal forecasts of temperatures (Schepen et al., 2020c; Schepen et al., 2016) and precipitation amounts (Schepen and Wang, 2014; Zhao et al., 2017). The BJP method models the relationship between predictor and predictand variables via a joint multivariate normal distribution and employs the Markov chain Monte Carlo sampling method in Bayesian inference to sample multiple sets of model parameters that can represent parameter uncertainty. For a multivariate problem, consider a vector $\mathbf{z}(t)^T = \begin{bmatrix} z_1 & z_2 & \dots & z_d \end{bmatrix}$ that includes d predictors and predictand variables in total. The joint distribution is formulated as

$$\mathbf{z}(t) \sim \mathrm{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \tag{1-4}$$

where $\boldsymbol{\mu}$ is a $d \times 1$ mean vector,

$$\boldsymbol{\mu}^T = \begin{bmatrix} \mu_1 & \mu_2 & \dots & \mu_d \end{bmatrix} \tag{1-5}$$

and $\boldsymbol{\Sigma}$ is a $d \times d$ covariance matrix,

$$\boldsymbol{\Sigma} = \boldsymbol{\sigma}\mathbf{P}\boldsymbol{\sigma}^{\mathbf{T}} \tag{1-6}$$

$\boldsymbol{\sigma}$ is a standard deviation vector,

$$\boldsymbol{\sigma}^T = \begin{bmatrix} \sigma_1 & \sigma_2 & \dots & \sigma_d \end{bmatrix} \tag{1-7}$$

$\mathbf{P}$ is a correlation coefficient matrix,

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,d} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{d,1} & \rho_{d,2} & \cdots & 1 \end{bmatrix} \tag{1-8}$$

As shown in Eq. (1-8), the BJP algorithm embeds the correlation parameters so that even through there is no skill in raw forecasts, the BJP model can reduce the forecasts to climatology (Zhao et al., 2017).

Apart from seasonal temperature and precipitation forecasts, the BJP model has also been applied for post-processing seasonal forecasts of E_{To} (T. Zhao et al., 2019b) and streamflow (Wang et al., 2009). Recent works further demonstrated the value of the BJP model for calibrating sub-seasonal to seasonal forecasts of precipitation (Li et al., 2020; Schepen et al., 2018) and E_{To} (T. Zhao et al., 2019a) at various temporal aggregated scale (i.e. daily and weekly) in different regions. The BJP model has also been employed jointly with other methods for a more robust post-processing of temperature, precipitation, and other climate variables. The calibration, bridging and merging (CBaM) method relies on the BJP model to post-process raw GCM outputs, called calibration, and uses the BJP model to statistically forecast climate variables with large-scale climate indices used as predictors, called bridging. The final component in CBaM is BMA (Wang et al., 2012a), aiming at optimally merging the calibrated and bridged forecasts to make the best use of multiple output fields from the GCM. This Bayesian post-processing technique has been successfully applied to improve the skill and reliability of seasonal temperature and precipitation forecasts over Australia and China (Peng et al., 2014; Schepen and Wang, 2014; Schepen et al., 2016; Schepen et al., 2014), and is being tested for post-processing real-time NMME forecasts in the National Oceanic and Atmospheric Administration's Climate Prediction Centre (Strazzo et al., 2019). In Australia, the CBaM methodology has also been integrated into the "forecast guided stochastic scenarios" (FoGSS), a forecasting system for generating skillful monthly ensemble streamflow forecasts to a 12-month forecast horizon (Bennett et al., 2017). FoGSS uses CBaM post-processed ensemble precipitation forecasts to drive a monthly rainfall-runoff model and employs a three-stage error model to remove biases and correct the over-confident problem of the ensemble streamflow outputs from the hydrological model. Elsewhere, for reliable crop yield forecasting, Schepen (2019) developed a comprehensive tool that combined BJP with other empirical methods to post-process and downscale GCM forecasts of multiple climate variables, and subsequently used the post-processed daily forecasts as the inputs to the crop model. This research connects the GCM forecasts with the impact model to robustly output tailored forecasts for agricultural applications in Australia.

Even with robust applications, again, the initial BJP model was not designed to produce post-processed forecasts with observed trend information embedded. This issue is also not addressed by the recent upgrade of the BJP algorithm by Wang et al. (2019), which implements the Gibbs sampling method to numerically draw samples. The newer BJP algorithm is more simplified in formulations and easier to code than the original version. As with the extended full calibration methods that account for long-term climate trends in annual-to-decadal forecasts, it is expected

that the capability of the BJP algorithm can also be extended to incorporate historical trends into the resulting climate forecasts on shorter timescales.

1.6 Research Questions and Objectives

This thesis aims to develop a new, reliable trend-aware forecast calibration method to eliminate trend disparity between raw GCM forecasts and observations. As a testament, this new method will be validated on forecasts of temperature and precipitation as they are commonly used meteorological variables and are key inputs to drive impact models, such as crop model and hydrological model. The following research questions are proposed to achieve the overarching research objective.

RQ1: How can observed temperature trends be embedded into seasonal temperature forecasts through statistical post-processing?

This research question aims to extend the capacity of the Bayesian joint probability (BJP) modelling approach to introduce the historical trend into seasonal temperature forecasts. The new trend-aware forecast post-processing model is expected to resolve the trend mismatch problem and further enhance forecast performance.

RQ2: Is the trend-aware method applicable for post-processing seasonal temperature forecasts on a continental scale? Can the trend-aware method be improved to better consider trend uncertainty in the Bayesian inference?

The trend-aware model developed in RQ1 infers the trend entirely from the model training data so that the observed trend is accurately embedded into the resulting forecasts. However, given limited data records, the inferred trends are subject to large sampling errors and thus may not realistically represent the true climate trend in the forecasts. To resolve this issue, RQ2 aims to further develop the trend-aware post-processing method to refine the treatment of trend uncertainty and to validate the improved method for seasonal maximum and minimum temperatures across all seasons on a continental application.

RQ3: How can the trend-aware model be adapted to post-process seasonal precipitation forecasts?

Compared with temperature variables, seasonal precipitation amounts have unique characteristics that need to be specially considered in the post-processing steps, such as following a positively

skewed distribution, having zero lower bounds, and being more uncertain and variable. Accordingly, RQ3 aims to improve the trend-aware model algorithm and introduce new evaluation tools to post-process seasonal precipitation forecasts.

RQ4: Are sub-seasonal temperature forecasts capable of reproducing historical trend information? How can the trend-aware model be adapted to post-process sub-seasonal forecasts?

Operational GCM sub-seasonal forecast models are configured similar to seasonal forecasting systems. Since there lacks understanding of how well GCMs represent the historical trend information in sub-seasonal temperature forecasts, the first aim of RQ4 is to investigate whether the trend mismatch problem demonstrated in seasonal forecasts exists in GCM sub-seasonal temperature forecasts for Australia. If not, the trend-aware calibration scheme will then be established to improve forecast quality and to properly incorporate long-term trend information into the sub-seasonal forecasts.

1.7 Thesis Structure and Publications

This thesis includes four main chapters (Chapter 2 - Chapter 5) corresponding to each of the research questions and objectives. Each chapter has been presented as a journal article style, including preamble, abstract, introduction, study data, method, result, discussion, and conclusion parts. Chapter 6 is a conclusion chapter, integrating the main methods and findings, recognising the limitations, raising the extension opportunities, and concluding the thesis.

Chapter 2 Embedding Observed Trend into Seasonal Temperature Forecasts through Statistical Calibration

2.1 Preamble

In Chapter 1, the current generation of GCMs were often found to poorly represent the climate trend in land surface air temperature and precipitation. The existing statistical post-processing approaches, either simple bias-correction or full calibration methods, rarely incorporate the observed trend into the post-processed seasonal climate forecasts by design. Several full calibration methods that correlate raw forecasts and corresponding observations in the calibration steps were established to correct trends in annual-to-decadal forecasts. However, these methods do not account for trend uncertainty in the calibration model and have not been applied to post-process seasonal climate forecasts. As reviewed in Section 1.5, the Bayesian joint probability (BJP) modelling approach is a robust and reliable tool for post-processing seasonal GCM forecasts. The new BJP algorithm implements Gibbs sampling to numerically draw samples, which is much easier to code and is more computationally efficient. This latest model version also does not account for the trend mismatch problem in the resulting forecasts, and this chapter seeks to extend the BJP algorithm to resolve this problem.

Accordingly, Chapter 2 answers RQ1: How can observed temperature trends be embedded into seasonal temperature forecasts through statistical post-processing? I introduce additional trend parameters for the observed data and raw forecasts into the BJP model, and validate the new model on three weather stations for January mean maximum temperature in Australia. The resulting calibrated forecasts are compared with the raw and BJP calibrated forecasts using diagnostic tools to comprehensively examine the forecast trend, bias, skill, reliability, and sharpness.

This chapter has been published in International Journal of Climatology (Impact Factor 3.928). The paper title is ‘Embedding trend into seasonal temperature forecasts through statistical calibration of GCM outputs’ and the authorship is Shao, Y., Wang, Q. J., Schepen, A., and Ryu, D.

2.2 Abstract

Accurate and reliable seasonal climate forecasts are frequently sought by climate-sensitive sectors to support decision-making under climate variability and change. Temperature trend is discernible globally over the past decades, but seasonal forecasts produced by a global climate model (GCM) generally underestimate such trend. Current statistical methods used for calibrating seasonal climate forecasts mostly do not explicitly account for climate trends. Consequently, the calibrated forecasts also fail to capture the observed trend. Solving this problem can enhance user confidence in seasonal climate forecasts. In this study, we extend the capability of the Bayesian joint probability (BJP) modelling approach for statistical calibration of seasonal climate forecasts. A trend component is introduced into the BJP algorithm for embedding the observed trend into calibrated ensemble forecasts. We apply the new model (named BJP-t) to three test stations in Australia. Seasonal forecasts of daily maximum temperatures from the SEAS5 model, operated by the European Centre for Medium-Range Weather Forecasts (ECMWF), are calibrated and evaluated. The BJP-t calibrated ensemble forecasts can reproduce the observed trend, when both raw ensemble forecasts and BJP calibrated ensemble forecasts fail to do so. The BJP-t calibration leads to more accurate, more skillful, more reliable, and sharper forecasts than the BJP calibration.

2.3 Introduction

The global land surface air temperature has exhibited marked temporal trends in recent decades (Hartmann et al., 2013), with accelerated warming since the 1970s (Jia et al., 2019). Regionally, temperature trends have also been identified in many countries in the past century (CSIRO and Australian Government Bureau of Meteorology, 2018; Ghasemi, 2015; L. Li et al., 2019; Yu et al., 2018). The changing near-surface temperature is projected to impact land activities such as agriculture.

Forecast users are seeking more accurate and reliable seasonal forecasts to assist their decision-making for the future in response to climate variability and change (Troccoli, 2010). Ensemble seasonal climate forecasts from global climate models (GCMs) can predict the climate conditions over monthly to seasonal time scales in the form of probability distribution (Schepen and Wang, 2014), thus providing valuable information for climate-sensitive sectors. However, current GCM-based seasonal temperature forecasts generally fail to reproduce the observed temperature trend. For example, Jia and Lin (2013) showed seasonal forecasts from the second phase of the Canadian Historical Forecasting Project (HFP2) considerably underestimated the significant trend of the

surface air temperature in winter on the Eurasian continent. Similarly, Krakauer (2019) noted the warming trend in monthly mean temperature forecasts from the North American Multi-Model Ensemble (NMME) model was weaker than the observed trend.

An accurate representation of the climate trend in the GCM forecasting system can induce additional skill in the forecasts and improve seasonal predictability. Doblas-Reyes et al. (2006) found that the enhanced temperature variability and better forecast quality could be obtained when the climate trend was better represented in seasonal ensemble forecasts in the presence of annually updated greenhouse gas concentrations. Weisheimer et al. (2011) also found that the high predictability of Southern Europe hot summers was partially explained by the ability of the dynamical model to reproduce the warming trend.

For practical use, raw GCM forecasts are normally post-processed to remove bias, to quantify a reliable ensemble spread of the forecasts, and to make the forecasts more skillful than climatology (Barnston et al., 2015; Doblas-Reyes et al., 2013). An example of a comprehensive calibration method is the Bayesian Joint Probability (BJP) modelling approach, which quantifies the relationship between raw forecasts and observations (Wang et al., 2009). Despite successful applications in the statistical calibration of seasonal climate forecasts (Peng et al., 2014; Schepen and Wang, 2014; Schepen et al., 2016; Strazzo et al., 2019; Zhao et al., 2017; T. Zhao et al., 2019b), the Bayesian method, along with other post-processing techniques, is not designed to embed the observed trend into calibrated seasonal forecasts.

In the BJP calibration, when raw GCM forecasts are absent of inherent skill, the observed trend information is not correctly transferred to the calibrated forecasts regardless of whether raw forecasts capture the trend signal or not. Furthermore, when raw ensemble forecasts are skillful, but do not capture the trend well, the forecast trend cannot be corrected by the BJP calibration. To generate forecasts aligning with the changing climate, we aim to embed the observed trend into the BJP calibrated forecasts regardless of how raw forecasts behave. We anticipate the introduction of the trend information will enhance forecast quality and improve confidence in post-processing amongst forecasters and forecast users.

Recent efforts to include the climate trend information in forecast post-processing are concentrated in seasonal-to-decadal time scales. Kharin et al. (2012) introduced a trend-adjustment approach to remove model residual drifts when there exists difference in modelled and observed trends in decadal predictions of annual global mean near-surface temperature.

Elsewhere, Sansom et al. (2016) modified the ensemble model output statistics (EMOS) technique (Gneiting et al., 2005) to address linear time-dependent biases of the forecast ensemble mean in the annual mean near-surface temperature. Pasternack et al. (2018) proposed the Decadal Climate forecast Recalibration Strategy (DeFoReSt) to recalibrate decadal ensemble forecasts of surface temperature. That is, they extended an EMOS-type method to jointly correct the forecast mean error and forecast spread dependent on the lead-time and linear climate trends. On seasonal time scales, a few studies have attempted to incorporate observed trend information into seasonal forecasts of Arctic sea ice (Director et al., 2019; Dirkson et al., 2019; Krikken et al., 2016).

In this study, we extend the capability of the BJP model by introducing linear trend components. This new model (hereafter named the BJP-t model) is applied to three test stations across Australia. We assess the effectiveness of the BJP-t calibration method by comprehensively evaluating and comparing raw ensemble forecasts, the BJP calibrated ensemble forecasts, and the BJP-t calibrated ensemble forecasts of monthly mean daily maximum temperature (Tmax) obtained from SEAS5, a state-of-the-art seasonal forecasting system operated by the European Centre for Medium-Range Weather Forecasts (ECMWF; Johnson et al., 2019).

The rest of the chapter is structured as follows: Section 2.4 introduces SEAS5 forecasts and station data used in this study, and elaborates the BJP-t method, forecast evaluation and verification tools. Section 2.5 presents the results. Section 2.6 further discusses the results. Section 2.7 summarizes the study and points to the main conclusions.

2.4 Data and methods

2.4.1 SEAS5 forecasting data and station data

This study uses ensemble re-forecasts obtained from the ECMWF SEAS5 forecasting system. SEAS5 consists of atmospheric, oceanic, sea-ice and land components. The atmosphere model is the Integrated Forecast System (IFS) atmosphere model cycle 43rl, with horizontal resolution of ~ 36 km. The ocean and cryosphere modules use the Nucleus for European Modelling of the Ocean model (NEMO), and a prognostic sea ice model, the Louvain-la-Neuve sea ice model version 2 (LIM2). The atmosphere initialization of SEAS5 re-forecasts is ERA-Interim. The ocean and sea-ice initial conditions for re-forecasts are supplied by historical reanalyses (ORAS5) from a new operational ocean analysis system, OCEAN5 (Zuo et al., 2018). To represent uncertainty in the initial state, unperturbed and perturbed atmospheric initial conditions, and

stochastic perturbations to the atmospheric models are used for all ensemble members. The ensemble of re-forecasts comprises 25 members and the re-forecasts are initialised on the first of every month from January 1, 1981 to December 1, 2016. Preliminary investigations suggested that 1-month ahead forecasts for monthly averages of Tmax (daily maximum temperature) generally failed to capture the observed trend for all 12 months in Australia (not shown). Here we demonstrate the effectiveness of the BJP-t model through case studies of forecasts for January, which is chosen arbitrarily. That is, we utilise re-forecasts initialised on the first of December each year to obtain the 1-month ahead re-forecasts for January in 1982-2017.

We select three weather stations (Table 2-1) located in different states in Australia (Figure 2-1), where raw and BJP forecasts do not represent the observed trend well. All the stations have observed values for 1982-2017 and have statistically significant trend in observations. The statistical significance is judged from the two-tailed Student's *t* test and Mann-Kendall test (Kendall, 1975; Mann, 1945) at the 5% significance level. Assessment of linear trend often uses the Student's *t* test for significance. This test requires the test statistic to follow a normal distribution. However, our preliminary analysis shows some skewness in the observations. To be prudent, we also apply non-parametric Mann Kendall test to check statistical significance, which does not require the data to be normally distributed. The gridded raw forecasts are paired with the point station data by choosing the value at the nearest SEAS5 grid cell centroid.

2.4.2 Bayesian modelling method

The Bayesian statistical model developed in this study is an extension of the Bayesian joint probability (BJP) modelling approach (Wang and Robertson, 2011; Wang et al., 2009). The BJP model has been widely applied to calibrate seasonal climate forecasts (Schepen and Wang, 2014; Schepen et al., 2016; T. Zhao et al., 2019b). A new BJP algorithm was recently introduced by Wang et al. (2019), which harnesses Gibbs sampling to make the post-processing more computationally efficient. Here, we introduce two trend parameters into the BJP mathematical formulation to develop the BJP-t model, where a Bayesian inference is used to model trend parameters in observations and raw forecasts, so that trend uncertainty is considered. Given limited availability of the data, trend uncertainty can be large and should not be neglected. Therefore, our BJP-t algorithm should be more robust than an approach in which the trend is first removed and later added back after the use of the BJP model. Details of the BJP-t algorithm are given in the following sections.

Table 2-1: Details of the weather stations.

Station Name	Longitude	Latitude	Elevation (m)
Brunette Downs	135.95°E	18.64°S	218
Murray Bridge	139.26°E	35.12°S	33
Wagga Wagga AMO	147.46°E	35.16°S	212

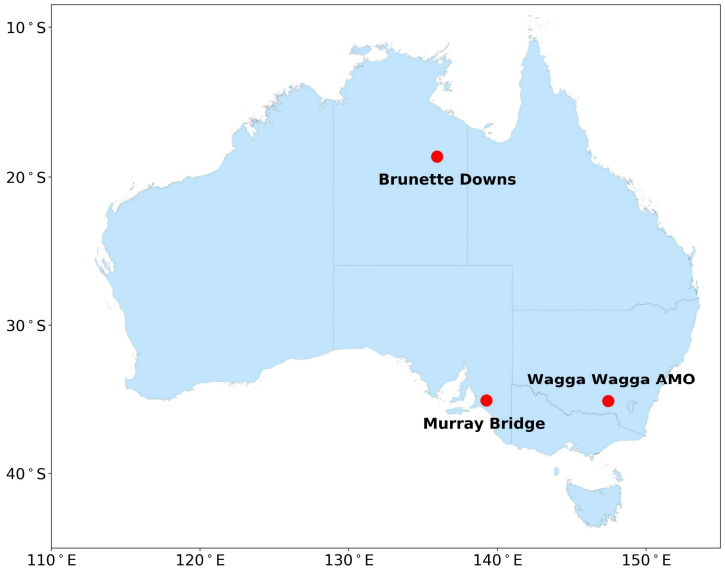

Figure 2-1: Location map of three case stations: Brunette Downs Station in Northern Territory, Murray Bridge Station in South Australia, and Wagga Wagga AMO Station in New South Wales.

2.4.2.1 Model formulation

Consider the ensemble mean of raw temperature forecasts y_1, and corresponding observed data y_2 with n historical data records. Note that information on ensemble spread of the raw forecasts is presently not used in the BJP and BJP-t models. Future work will address this limitation. The modelling of the joint distribution follows the assumption that the marginal distribution of individual variable is normally distributed. From the normalisation test, we find some skewness in the data series (not shown). To achieve normality, we firstly transform the variables by using the Yeo-Johnson transformation (Wang and Robertson, 2011; Wang et al., 2009; Yeo and Johnson, 2000). For the variable y,

$$y^{'} = \begin{cases} \left[(y+1)^{\lambda} - 1\right]/\lambda & \lambda \neq 0, y \geq 0 \\ \log(y+1) & \lambda = 0, y \geq 0 \\ -\left[(-y+1)^{2-\lambda} - 1\right]/(2-\lambda) & \lambda \neq 2, y < 0 \\ -\log(-y+1) & \lambda = 2, y < 0 \end{cases} \tag{2-1}$$

where λ is the transformation parameter. The raw forecasts y_1 and the observations y_2 are transformed to $y_1^{'}$ and $y_2^{'}$ respectively. We apply maximum a posteriori (MAP) estimation method to derive a single best estimate set of transformation parameters for y_1 and y_2 (Schepen et al., 2016).

A continuous bivariate normal distribution is used to formulate the relationship between the predictor z_1 and predictand z_2. The predictor is the detrended transformed raw forecast, and the predictand is the detrended transformed observation, that is

$$z_1(t) = y_1^{'}(t) - \alpha_1(t - t_m) \tag{2-2}$$

$$z_2(t) = y_2^{'}(t) - \alpha_2(t - t_m) \tag{2-3}$$

where the trend parameter α_1 is for raw forecasts, α_2 is for observations, and t is the index of the forecast year, t_m is the index of roughly the middle year (e.g. 1999 in this work) in the training period, $z_1(t)$ is the anomaly from the trendline of the raw forecasts, and $z_2(t)$ is the anomaly from the trendline of the observations.

The bivariate joint model relating z_1 and z_2 is set up as

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2-4}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are mean vector and covariance matrix respectively,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \tag{2-5}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{2-6}$$

where μ_i is a mean, σ_i is a standard deviation, and ρ is a correlation coefficient.

The z_1 and z_2 follow a normal distribution respectively, which can be generalised as

$$[z_i(t)] = \mathrm{N}(\mu_i, \sigma_i^2) \tag{2-7}$$

By combining Eq. (2-2), Eq. (2-3) and Eq. (2-7), the distribution of $y_1^{'}$ and $y_2^{'}$ is given by

$$[y_i^{'}(t)] = \mathrm{N}[\mu_i + \alpha_i(t - t_m), \sigma_i^2] \tag{2-8}$$

Hereafter, we denote the parameter set as $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha_1, \alpha_2\}$.

2.4.2.2 Parameter inference

The model parameters are inferred from the sequence of training data pairs for n years: $\mathbf{D} = \{[y_1^{'}(t), y_2^{'}(t)], t = 1, 2, ..., n\}$. The posterior distribution of the model parameters is

$$p(\boldsymbol{\theta} \mid \mathbf{D}) \propto p(\boldsymbol{\theta}) p(\mathbf{D} \mid \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{t=1}^{n} p(\mathbf{D} \mid \boldsymbol{\theta}) \tag{2-9}$$

where $p(\boldsymbol{\theta})$ is the prior distribution for model parameters, and $p(\mathbf{D} \mid \boldsymbol{\theta})$ is the likelihood function.

We specify the prior for $\boldsymbol{\theta}$ as

$$p(\boldsymbol{\theta}) \propto p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\alpha_1) p(\alpha_2) \tag{2-10}$$

where

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-3/2} \tag{2-11}$$

$$p(\alpha_i) \propto 1 \tag{2-12}$$

The prior for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is non-informative multivariate Jeffreys prior (Gelman et al., 2014), and the prior for α_i is also non-informative.

We derive the conditional posterior distribution for parameters $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ by combining Eq. (2-9) – Eq. (2-12) as

$$[\boldsymbol{\Sigma} \mid \cdot] = \text{Inv- Wishart}_{n-1}(\mathbf{S}) \tag{2-13}$$

$$\left[\boldsymbol{\mu} \mid \cdot\right] = \mathrm{N}\left(\overline{\mathbf{z}}, \boldsymbol{\Sigma} / n\right) \tag{2-14}$$

where

$$\mathbf{S} = \sum_{t=1}^{n}\left[\mathbf{z}(t) - \overline{\mathbf{z}}\right]\left[\mathbf{z}(t) - \overline{\mathbf{z}}\right]^{\mathrm{T}} \tag{2-15}$$

$$\overline{\mathbf{z}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{z}(t) \tag{2-16}$$

The symbol $|\cdot$ refers to the distribution conditioned on all other variables, and Inv-Wishart_{n-1} is the Inverse-Wishart distribution with $n-1$ degrees of freedom.

We can derive the conditional posterior distribution for parameter α_i from Eq. (2-2), Eq. (2-3) and Eq. (2-8) as

$$[\alpha_i \mid \cdot] = \mathrm{N}\left\{\frac{\sum_{t=1}^{n}[y_i^{'}(t) - \mu_i](t - t_m)}{\sum_{t=1}^{n}(t - t_m)^2}, \frac{\sigma_i^2}{\sum_{t=1}^{n}(t - t_m)^2}\right\} \tag{2-17}$$

The conditional distribution of $z_i(t)$ can also be deduced to sample missing values in any variables,

$$[z_i(t) \mid \cdot] = \mathrm{N}[\mu_i^*(t), \Sigma_{i,i}^*] \tag{2-18}$$

where

$$\Sigma_{i,i}^* = \Sigma_{i,i} - \boldsymbol{\Sigma}_{i,(i)} \mid \boldsymbol{\Sigma}_{(i),(i)} \mid^{-1} \boldsymbol{\Sigma}_{(i),i} \tag{2-19}$$

$$\mu_i^*(t) = \mu_i + \boldsymbol{\Sigma}_{i,(i)} \mid \boldsymbol{\Sigma}_{(i),(i)} \mid^{-1} [\mathbf{z}_{(i)}(t) - \boldsymbol{\mu}_{(i)}] \tag{2-20}$$

(i) denotes the index in $\{1,2\}$ except i.

The conditional distribution of $y_i^{'}(t)$ can be derived by combining Eq. (2-2), Eq. (2-3) and Eq. (2-18), as

$$[y_i^{'}(t) \mid \cdot] = \mathrm{N}[\mu_i^*(t) + \alpha_i(t - t_m), \Sigma_{i,i}^*] \tag{2-21}$$

With posterior conditionals for all parameters and missing values, we implement Gibbs sampling to numerically sample multiple sets of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta} \mid \mathbf{D})$ to represent the posterior distribution over parameters $\boldsymbol{\theta}$.

To establish a Gibbs sampler for parameter inference, we firstly set the initial value for α_i and any missing value in $y_i^{'}(t)$. We initialise α_i as 0 and set missing $y_i^{'}(t)$ as $\hat{y}_i^{'}$, the average of non-missing $y_i^{'}(t), t = 1, 2, ..., n$.

For each iteration in Gibbs sampling, we conduct the following steps:

1. Compute $z_i(t)$ from $y_i^{'}(t)$ (see Eqs. (2-2) – (2-3))
2. Sample $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ in sequence (see Eqs. (2-13) – (2-16))
3. If the value of $z_i(t)$ is missing, sample and update $z_i(t)$, and calculate and update $y_i^{'}(t)$ (see Eq. (2-2) and Eq. (2-3), Eqs. (2-18) – (2-21))
4. Sample α_i (see Eq. (2-17))

We note the pseudocode for elaborating the implementation of the Gibbs sampling for the BJP model is illustrated in Wang et al. (2019). The BJP-t model can be coded using the same flow with slight modifications given the above sampling procedures.

2.4.2.3 Prediction use

Once the parameter sets are derived, we can calibrate a newly transformed raw forecast $y_1^{'}(t^*)$ in predictive mode to generate a calibrated (in transformed space) forecast $y_2^{'}(t^*)$. The posterior predictive distribution of $y_2^{'}(t^*)$ is given by

$$f[y_2^{'}(t^*)] = \int p[y_2^{'}(t^*) \mid y_1^{'}(t^*), \boldsymbol{\theta}] p(\boldsymbol{\theta} \mid \mathbf{D}) d\boldsymbol{\theta} \tag{2-22}$$

Here, we sample a calibrated forecast $y_2^{'}(t^*)$ from its posterior predictive distribution by treating the transformed observations as being missing values that can be imputed. The initialisation and implementation of the Gibbs sampler for prediction are similar to the steps illustrated in the pseudocode provided by Wang et al. (2019). Since the new parameter α_i is introduced, in each sampling iteration, we apply the Gibbs sampling procedures to sample and update $z_i(t^*)$, and calculate and update $y_i^{'}(t^*)$ with the newly sampled value if the value of $y_i^{'}(t^*)$ is missing in the original data.

For each bivariate normal parameter set, we derive a single sample of $y_2^{'}(t^*)$. After Gibbs sampling, we produce a collection of calibrated ensemble forecasts of $y_2(t^*)$ by back-transforming each of the sample members to the original space using inverse Yeo-Johnson transformation (see Eq. (2-23)).

$$y = \begin{cases} \sqrt[\lambda]{\lambda y' + 1} - 1 & \lambda \neq 0, y' \geq 0 \\ e^{y'} - 1 & \lambda = 0, y' \geq 0 \\ 1 - \sqrt[2-\lambda]{1 - (2-\lambda) y'} & \lambda \neq 2, y' < 0 \\ 1 - e^{-y'} & \lambda = 2, y' < 0 \end{cases} \tag{2-23}$$

2.4.3 Forecast evaluation

We evaluate and compare the raw, BJP calibrated, and BJP-t calibrated forecasts of monthly average of Tmax from the SEAS5 model for three test stations. The BJP and BJP-t models are evaluated using a leave-one-year-out cross validation, similar to that used in other studies (e.g. Kharin et al., 2017; Dirkson et al., 2019; Schepen et al., 2020c). Before calibration is applied to a historical event, the data pair for that event is hidden from the model inference. The process is repeated for all events. We note that this cross-validation set up is not entirely satisfactory, as it is only effective for the anomaly component and not for the trend component. Although the Bayesian inference used in the BJP and BJP-t models explicitly accounts for uncertainties of the trend parameters, model overfitting is still possible. An alternative method of validation is to leave out short periods of data at the start and end of the full data period for validation, but the results will be subject to large sampling effect and therefore will not be very informative (Dirkson et al., 2019). While Barnston and van den Dool (1993) suggested that leaving out more years could be a more appropriate cross validation strategy, we find the results are not significantly different in both methods, which agrees with the finding from Schepen et al. (2014).

To investigate the climate trend, we fit the linear multi-decadal trend (slope) for the observations and ensemble forecast means using the least squares regression method. Moreover, the uncertainty of the observed trend slope is quantified by the 90% confidence interval, assuming residuals are independently normally distributed (Hartmann et al., 2013). The trend is visualised in a forecast quantile plot with fitted trendlines superimposed. In this plot, forecast quantiles are generated for individual events chronologically and compared with observed values.

We use the root mean square error (RMSE) to measure ensemble-mean forecast accuracy. This metric is calculated as the root mean squared difference between the ensemble forecast mean and corresponding observation, indicating the magnitude of the scale-dependent forecast errors, as

$$\mathrm{RMSE} = \left[\frac{1}{T} \sum_{t=1}^{T} \left(\overline{y^t} - y_{obs}^t \right)^2 \right]^{\frac{1}{2}} \tag{2-24}$$

where $\overline{y^t}$ is the ensemble forecast mean for an event t, and y_{obs}^t is the observation for an event t.

We evaluate the skill of probabilistic forecasts using the continuous ranked probability score (CRPS; Matheson and Winkler, 1976). For time periods $t = 1, 2, ..., T$, CRPS is measured as the difference between the ensemble forecasts and observations (Hersbach, 2000),

$$\mathrm{CRPS} = \frac{1}{T} \sum_{t=1}^{T} \int \left[F\left(y^t\right) - H\left(y^t - y_{obs}^t\right) \right]^2 dy^t \tag{2-25}$$

where F is the forecast cumulative distribution function (CDF) and H is the Heaviside step function which equals 0 if $y^t < y_{obs}^t$ and equals 1 otherwise. CRPS rewards a small spread when the forecast is accurate (Wilks, 2006b). We then convert the CRPS value to a skill score to measure the relative improvement of the model forecasts compared to reference forecasts which are the corresponding leave-one-year-out cross-validated climatology ensemble forecasts generated by the BJP model. Here, the climatology forecasts are generated using the distribution of historical observations as elaborated in Wang et al. (2019). The CRPS skill score is formulated as

$$\mathrm{CRPS_{ss}} = \frac{\mathrm{CRPS_{ref}} - \mathrm{CRPS}}{\mathrm{CRPS_{ref}}} \times 100 \left(\text{unit: } \%\right) \tag{2-26}$$

The CRPS skill score is positively oriented. Perfect forecasts have a maximum skill score of 100 while a score of 0 indicates that the forecasts have no skill and are comparable to reference forecasts. Negative skill scores indicate forecasts are poorer than reference forecasts.

We investigate the reliability of the ensemble forecasts by analysing probability integral transforms (PITs; Wang et al., 2009) of Tmax observations. The PIT value of the observation y_{obs}^t is calculated by $\pi_t = F(y_{obs}^t)$, where F is the CDF constructed from the forecasts. When the probabilistic forecasts are reliable, the collection of PIT values follows a standard uniform distribution. We generate the PIT uniform probability plot to visualise the reliability, where ranked increasing PIT values π_t^* for all the events $t = 1, 2, ..T$ are plotted with the corresponding theoretical quantile of the uniform distribution. Forecasts with perfect reliability follow the 1:1 line. We also calculate the PIT index (named reliability index α in Renard et al., 2010) to

statistically assess the overall tendency for π_t^* to deviate from the 1:1 line in the PIT plot. The PIT index varies between 0 (worst reliability) and 1 (perfect reliability), defined by

$$\text{PIT index} = 1.0 - \frac{2}{T}\sum_{t=1}^{T}\left|\pi_t^* - \frac{t}{T+1}\right| \tag{2-27}$$

We check the sharpness of the ensemble forecasts by numerically calculating the average width of the central 50% (between 0.25 and 0.75 quantile) and 90% (between 0.05 and 0.95 quantile) prediction intervals for all individual events (Gneiting et al., 2007). Narrower interval width indicates sharper probabilistic forecasts.

To determine whether the use of the BJP-t model statistically significantly improves or decreases forecast performance relative to raw and BJP calibrated forecasts, we apply the bootstrap procedure as described in Schepen et al. (2016) to the RMSE, CRPS skill score, PIT index and the average width of the prediction intervals. For each statistical metric, we generate 1000 samples of the estimates for raw and BJP calibrated forecasts and test the significance at the 5% significance level. In this regard, if the metric value of the BJP-t calibrated forecasts is above the 95th percentile or below the 5th percentile of the distribution constructed from 1000 resampled data, we conclude the use of the BJP-t model significantly enhances or worsens the forecast attribute.

2.5 Result

2.5.1 Trend analysis

The calibrated ensemble forecast quantiles and observations are plotted with trendlines of the ensemble forecast means and observations superimposed for the period of 1982-2017 (Figure 2-2). Visually, the general pattern of forecast quantile ranges of the BJP-t calibrated forecasts (right column) is more consistent with the observed trend than the BJP calibrated forecasts (left column). In all cases, the tendency of the 0.5 and 0.8 inter-quantile of the BJP-t calibrated forecasts over time clearly follows the upward or downward observed trend. In contrast, trendlines of the raw forecast means and the BJP calibrated forecast means generally have gentler trend slopes, which are almost horizonal at Brunette Downs Station and Murray Bridge Station. Furthermore, at Brunette Downs Station, the BJP calibrated forecasts tend to go towards the long-term climatological mean, as indicated by less variation of the predictive bands over the time period in

Figure 2-2. This means the BJP model is not capable of modelling the climate variability under the climate change. As a result, the climatology-like ensemble forecasts return low forecast skill (see Section 2.5.2).

Numerically, the multi-decadal linear trends are calculated for the observations and ensemble forecast means at three weather stations (Table 2-2). The trend slopes for both raw and BJP calibrated forecasts fall outside the 90% confidence interval in all cases. Forecast means at Brunette Downs Station even give the inverse trend direction to the observed data. By comparison, we find the trend slope of the BJP-t calibrated forecast means is almost identical to the observed one in all cases. This suggests the BJP-t calibration scheme can effectively embed the climate trend into the calibrated forecasts.

Table 2-2: Fitted linear decadal trend (K/decade) for observed data (with 90% confidence intervals), raw forecast mean, BJP calibrated forecast mean and BJP-t calibrated forecast mean in three cases.

Station Name	Brunette Downs	Murray Bridge	Wagga Wagga AMO
observation	-1.064 ± 0.329	0.860 ± 0.266	0.869 ± 0.313
Raw forecast mean	0.011	0.088	0.234
BJP calibrated forecast mean	0.062	0.029	0.223
BJP-t calibrated forecast mean	-1.078	0.856	0.863

2.5.2 Forecast accuracy and skill

The RMSE values and CRPS skill scores of the ensemble forecasts are presented in Table 2-3. The BJP calibration leads to larger errors in ensemble forecast means than raw forecast means at Murray Bridge Station and Wagga Wagga AMO Station, while the BJP-t calibration reduces the forecast errors in all three cases.

Focusing on the ensemble forecasts, forecast skill for the BJP-t calibrated forecasts is significantly improved at Brunette Downs Station and Murray Bridge Station, as compared with the raw and BJP calibrated forecasts. In these two cases, the negative or neutral skill retained in the BJP calibrated forecasts is turned positive and discernible, with skill score greater than 10%. Since the BJP-t calibration scheme directly extracts the trend information from observations, the BJP-t calibrated ensemble forecasts can sufficiently reproduce the climate trend and extract most of the raw forecast skill. For Wagga Wagga AMO Station, skill improvement is obvious but

insignificant for the BJP-t calibrated forecasts. We suggest this is because raw forecasts have some skills and are better at reproducing the observed trend slope compared with the other two cases. In this regard, the correction of the trend amplitude may not have salient impacts on further skill improvement.

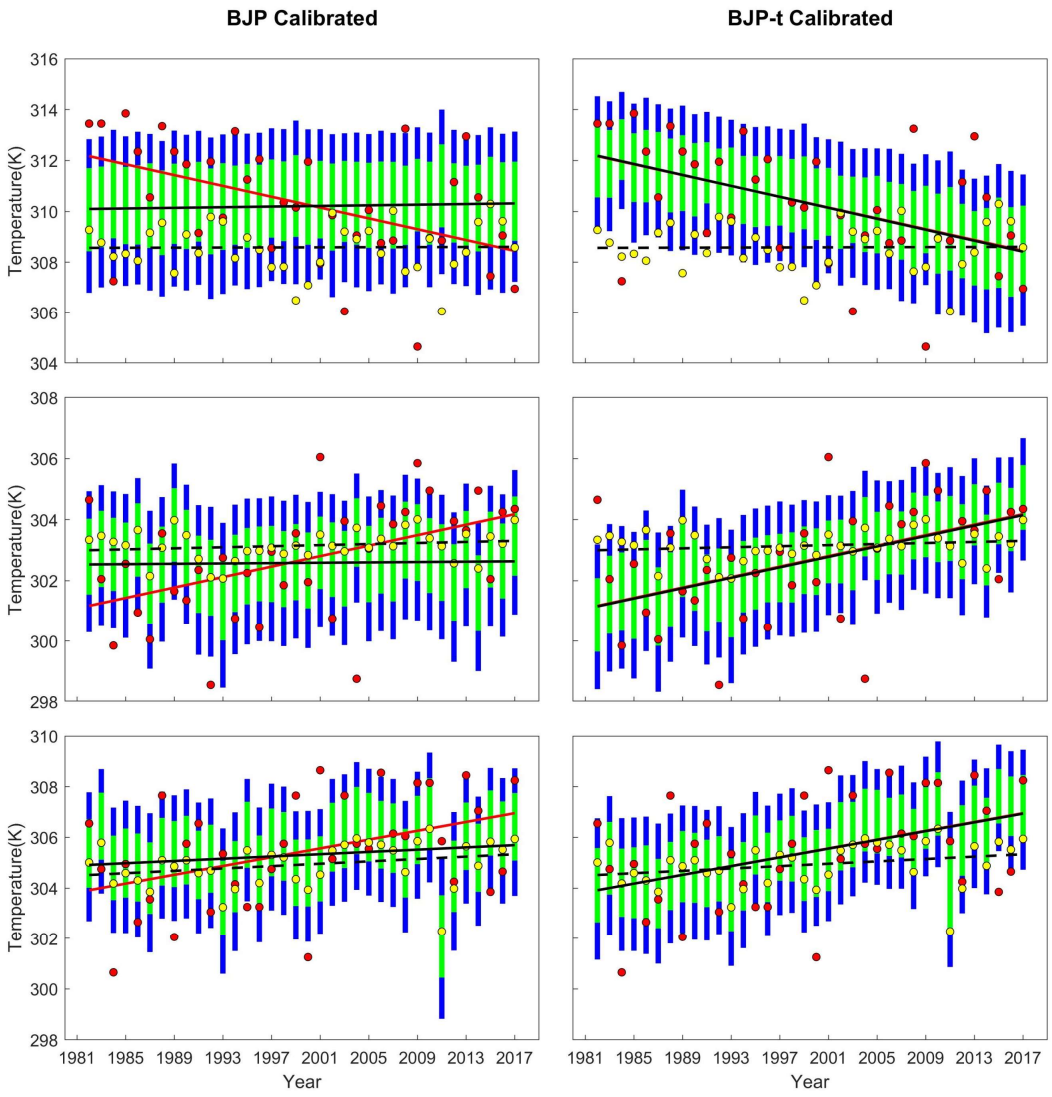

Figure 2-2: Forecast quantiles of cross-validated Tmax forecasts and observed values plotted for BJP calibrated forecasts (left) and BJP-t calibrated forecasts (right). The first row is Brunette Downs Station, the second row is Murray Bridge Station, and the third row is Wagga Wagga AMO Station. The red dots are observed Tmax; yellow dots are raw forecast means; blue vertical lines are forecast [0.10, 0.90] quantile range; green vertical lines are forecast [0.25, 0.75] quantile range. Red line is fitted observed linear trendline; dashed black line is fitted linear trendline of raw ensemble forecast mean; black line is fitted linear trendline of calibrated ensemble forecast mean.

Table 2-3: The RMSE and CRPS skill score for raw forecasts, BJP calibrated forecasts, and BJP-t calibrated forecasts in three stations.

	Station Name	Brunette Downs	Murray Bridge	Wagga Wagga AMO
RMSE (K)	Raw forecast mean	3.04	1.83	2.01
	BJP calibrated forecast mean	2.40	1.90	2.06
	BJP-t calibrated forecast mean	2.15	1.73	1.96
CRPSS (%)	Raw forecast	-45.5	3.30	1.79
	BJP calibrated forecast	-2.28	0.09	4.98
	BJP-t calibrated forecast	10.1*	11.9*	9.32

Note: The symbol * denotes the significant change of the BJP-t results compared to the raw and BJP calibrated forecasts.

2.5.3 Reliability and sharpness

The PIT index and PIT uniform probability plot of the forecasts for three stations are shown in Table 2-4 and Figure 2-3 respectively. Numerically, the BJP-t calibrated forecasts are significantly more reliable than the raw and BJP calibrated forecasts for Brunette Downs Station and Murray Bridge Station. The PIT plots also support this finding. The PIT values for the BJP-t calibrated forecasts are visually closer to the 1:1 diagonal line than the raw and BJP calibrated ensemble forecasts, indicating that the BJP-t calibration could make the forecasts more reliable. At Wagga Wagga AMO Station, the BJP and BJP-t calibrated forecasts are comparably reliable in ensemble spread, and both are more reliable than the raw forecasts.

Probabilistic forecasts that have maximal sharpness and high reliability are hard to produce (Wilks, 2018). The BJP-t calibration appears to maximize the sharpness of forecasts at Brunette Downs Station and Murray Bridge Station, as indicated by more concentrated 50% and 90% prediction distribution intervals in Table 2-4. We note the improvement of the sharpness of BJP-t calibrated forecasts is not at the expense of sacrificing reliability as illustrated above. For Wagga Wagga AMO Station, raw forecasts are found sharper than the BJP and BJP-t calibrated forecasts. Nevertheless, raw ensemble forecasts do not exhibit higher reliability and skill, indicating that they are not corresponding well to the observations.

Table 2-4: The PIT index and average widths of central prediction intervals (50% and 90%) for raw forecasts, BJP calibrated forecasts, and BJP-t calibrated forecasts in three cases.

		Station Name	Brunette Downs	Murray Bridge	Wagga Wagga AMO
PIT index		Raw forecast	0.617	0.902	0.871
		BJP calibrated forecast	0.937	0.929	0.931
		BJP-t calibrated forecast	0.959*	0.958*	0.945
Average widths (K) for the interval	50%	Raw forecast	3.572	2.462	2.537
		BJP calibrated forecast	3.217	2.492	2.712
		BJP-t calibrated forecast	2.892*	2.278*	2.627
	90%	Raw forecast	7.532	6.335	6.098
		BJP calibrated forecast	7.876	6.105	6.604
		BJP-t calibrated forecast	7.167*	5.701*	6.589

Note: The symbol * denotes the significant change of the BJP-t results compared to the raw and BJP calibrated forecasts.

2.6 Discussion

In this study, the BJP-t model infers the linear trend in the transformed space, and the model performance is evaluated in the real space. However, we realise that the modelled linear trendline does not necessarily remain linear when transformed back to the real space. In this regard, the impact of data transformation on the linearity of the trendline has been further investigated. That is, for the ensemble means and observations, we fit a linear trendline in the transformed space, back transform the data points on the trendline to the real space and connect these points. In all cases, the back-transformed trendlines of the BJP-t calibrated forecasts and observations are highly consistent (shown in Supplementary Figure S1-1). Moreover, back-transformed data points are found to be linear-like, with the coefficient of determination close to 1. Besides test cases presented in this chapter, we also conduct the same procedures for broad cases over the Australian continent and derive the same conclusion (not shown). Therefore, for straightforward quantification of the trend, we choose to directly evaluate the linear trend in the real space for this work.

Estimated observed trends for short time periods involve large uncertainty and are sensitive to the start and end year (Hartmann et al., 2013). The computed trends can change for a shorter or longer time period. Over shorter time periods, say a 10-year period, observed trends are subject to

internal variability, such as EI Nino Southern Oscillation. For multi-decadal temperature changes, both internal variability of the climate system and the response to external forcing, such as greenhouse gases, dominate the decadal fluctuations. Since we model the uncertainty in the trend parameter, the sampling effect of the observed trend as well as the uncertainty of the calibrated ensemble forecasts in the real space is reflected in the model sampling procedures. As shown in Figure 2-2, the trend information in the original data can be largely recovered in back transformed predictions, indicated by the consistency of the trend slope between the BJP-t calibrated forecast means and observed data.

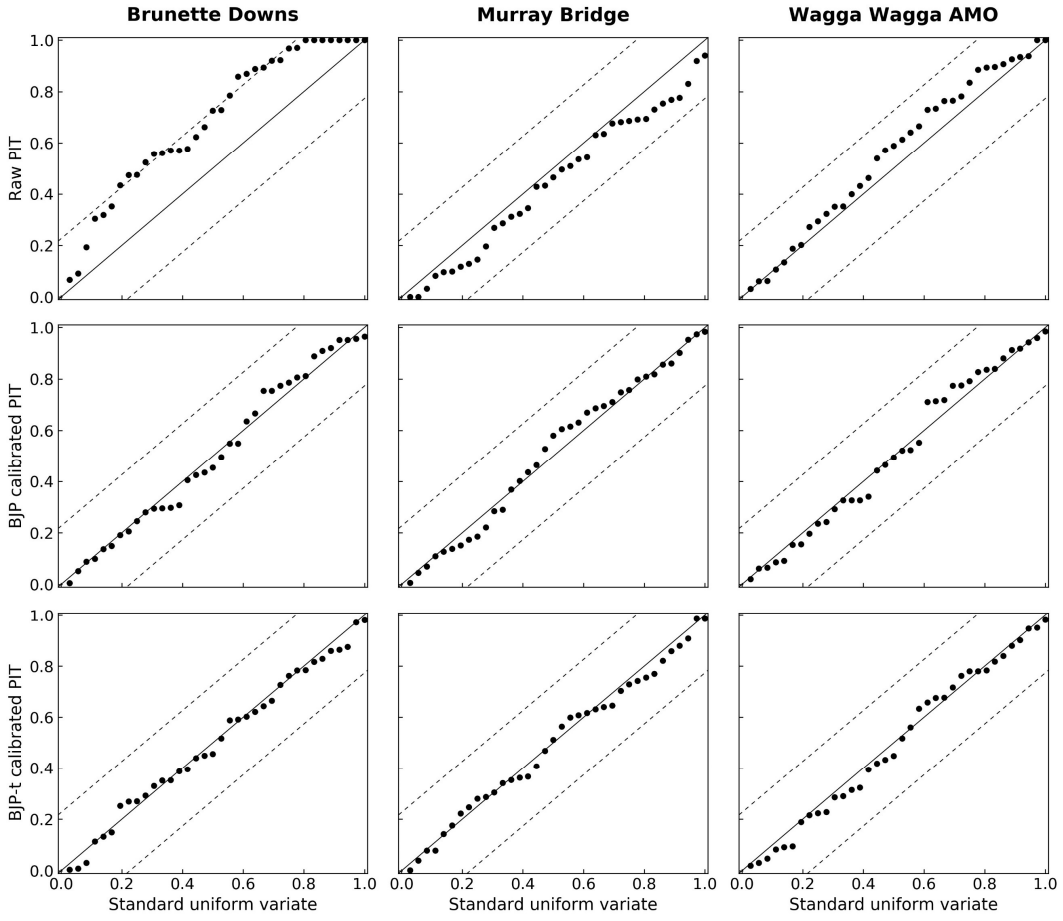

Figure 2-3: PIT uniform probability plot for raw forecasts, BJP calibrated forecasts, and BJP-t calibrated forecasts for three stations. The first column is Brunette Downs Station, the second column is Murray Bridge Station, and the third column is Wagga Wagga AMO Station. Dots are PIT values of observed Tmax; solid line is 1:1 uniform distribution; dashed line is Kolmogorov 5% significance band.

For the calculation of the CRPS skill score, we can also use cross-validated climatology forecasts from the BJP-t model as reference forecasts, which are trend-adjusted. The selection of the reference forecasts depends on the perspective of the forecasters and users. If their starting point is that their (naïve) forecasts have already accounted for trends, trend-adjusted climatology forecasts should be used. If not, climatology forecasts without trend adjustment are probably more appropriate. In this study, our starting point is the forecasts without properly accounting for trends. Therefore, the climatology forecasts from the BJP-model are used, which are not trend-adjusted. We also evaluate RMSEP-based skill score (Wang and Robertson, 2011), which measures the root mean squared error in probability between forecast means and observations. However, the results are not included because they are highly consistent with CRPS skill score results.

We have demonstrated that the BJP-t calibration scheme is effective at introducing the observed trend into calibrated Tmax ensemble forecasts for selected cases. This new model has the potential to be applied to continental scales and different lead times, such as to Australian minimum and maximum temperatures. Preliminary results suggest that in broader cases, when there is not trend in the training data, the BJP-t model defaults to BJP on average. However, the trend parameters are subject to uncertainty in the Bayesian inference and tend to make the forecast spread wider than using just the BJP model. Future research will attempt to resolve this issue.

The BJP-t model can also be potentially adapted to other important meteorological variables, such as precipitation. Embedding the trend into the calibrated rainfall forecasts has several barriers to overcome, including large amount of zero values, complex local precipitation patterns (Schepen et al., 2018), and undetectable trends in the rainfall series subject to high interannual variations (Hartmann et al., 2013). To tailor the BJP-t model for calibrating seasonal rainfall forecasts, we may need to account for zero values in the sampling of the trend parameters.

2.7 Conclusion

Climate-sensitive industries, such as water, energy, and agriculture sectors, often require skillful and reliable climate forecasts to inform decision-making and to develop management plans in the changing climate. Raw GCM seasonal climate forecasts generally underestimate the observed climate trend, which make them less informative for forecast users. Sophisticated statistical calibration methods are often designed for reducing biases and improving reliability in raw forecasts, but seldom for reproducing the climate trend in calibrated ensemble forecasts. As a result, the calibrated forecasts also fail to capture the observed trend.

The BJP-t model proposed in this chapter introduces trend components into the original BJP algorithm. This new model can extract most of the raw forecast skill while transferring the observed climate trend into calibrated ensemble forecasts. Results show that the BJP-t calibration method is effective at embedding the climate trend into calibrated ensemble forecasts while producing forecasts of overall better performance. In selected test stations, when raw and BJP calibrated forecasts have low skills, the BJP-t calibration results in skillful forecasts. The BJP-t model does not significantly improve the forecast performance where raw and BJP calibrated forecasts are already skillful and reliable. The increased or similar sharpness of the BJP-t calibrated forecasts is also found without sacrificing reliability.

From a broader perspective, we anticipate the forecast users and the public would benefit from the BJP-t calibrated ensemble forecasts. This work will be applied to the continental scale and tailored for other meteorological variables in the future.

Chapter 3 Improving the Trend-aware Post-processing Method to Post-process Seasonal Temperature Forecasts

3.1 Preamble

Chapter 2 developed a trend-aware forecast post-processing model, BJP-t, to calibrate seasonal maximum temperature forecasts. This method was modified from the Bayesian joint probability (BJP) modelling approach, which by design, does not explicitly incorporate the observed trend into calibrated forecasts. In test weather stations, the BJP-t calibrated forecasts were found to accurately reproduce the observed trends while being more skillful, more reliable, and sharper than the BJP calibrated forecasts.

When applying the BJP-t model to more cases, I found when there was no trend in the training data, the BJP-t model defaulted to BJP on average but resulted in wider forecast ensemble spreads than the BJP calibrated forecasts. This is because the inferred trends were subject to large sampling errors due to limited data records. As a result, trends entirely inferred from the available data may not explicitly represent the true underlying trends. In addition, the ability of raw SEAS5 temperature forecasts to capture the climate trend and the effectiveness of the trend-aware method when applied on a spatial continental scale remain untested.

For these challenges, Chapter 3 answers RQ2: Is the trend-aware method applicable for post-processing seasonal temperature forecasts on a continental scale? Can the trend-aware method be improved to better consider trend uncertainty in the Bayesian inference? In this Bayesian method, I refine the treatment of trend uncertainty by using priors for the trend parameters. I present and compare two models: BJP-t with non-informative priors, and BJP-ti with informative priors. The BJP-ti model uses regional information to construct priors, where trends are inferred from data but with a degree of moderation. Detailed analyses are conducted on seasonal minimum and maximum temperatures across Australia.

This chapter has been published in Monthly Weather Review (Impact Factor 3.435). The paper title is ‘Going with the trend: forecasting seasonal climate conditions under climate change’ and the authorship is Shao, Y., Wang, Q. J., Schepen, A., and Ryu, D.

3.2 Abstract

For managing climate variability and adapting to climate change, seasonal forecasts are widely produced to inform decision-making. However, seasonal forecasts from global climate models are found to poorly reproduce temperature trends in observations. Furthermore, this problem is not addressed by existing forecast post-processing methods that are needed to remedy biases and uncertainties in model forecasts. The inability of the forecasts to reproduce the trends severely undermines user confidence in the forecasts. In our previous work, we proposed a new statistical post-processing model that counteracted departures in trends of model forecasts from observations. Here, we further extend this trend-aware forecast post-processing methodology to carefully treat trend uncertainty associated with the sampling variability due to limited data records. This new methodology is validated on forecasting seasonal averages of daily maximum and minimum temperatures for Australia based on the SEAS5 climate model of the European Centre for Medium-Range Weather Forecasts. The resulting post-processed forecasts are shown to have proper trends embedded, leading to greater accuracy in regions with significant trends. The application of this new forecast post-processing is expected to boost user confidence in seasonal climate forecasts.

3.3 Introduction

Global surface temperatures have increased since the pre-industrial period (Hartmann et al., 2013) and warming trends have accelerated in recent decades (Jia et al., 2019). To help assess the impact of human activities on the earth system, now and into the future, global climate models (GCMs) have been developed to examine historical climate trends and to make climate projections for decades and centuries ahead (Flato et al., 2013). In parallel, an alternative class of GCMs have been developed expressly for seasonal climate forecasting (Troccoli, 2018; Troccoli et al., 2008). The main difference between seasonal forecasting models and climate projection models is that seasonal forecasting models are repeatedly initialised to estimates of the current oceanic and atmospheric conditions through data assimilation at run time (Doblas-Reyes et al., 2013), while climate projection models typically emphasize the response of the climate system to external forcing such as greenhouse gas emissions and volcanic eruptions (Kirtman et al., 2013). Filling the gap between seasonal forecasting and climate projections is the emerging field of decadal prediction (Kushnir et al., 2019), for which assimilation of subsurface ocean data is essential.

The relatively independent development of the different classes of GCMs means that climatic trends may not be represented consistently across them. Moreover, in seasonal forecasting, reproducing trends is rarely a priority compared to, say, reproducing El Nino-Southern Oscillation cycles (Troccoli, 2010). In fact, there is evidence that current operational seasonal forecasting GCMs poorly reproduce the observed temperature trend in forecasts (Krakauer, 2019; Shao et al., 2021a). As an example of this, the summer mean daily minimum temperature in Melbourne, Australia (Moorabbin Airport station), has increased at a rate of 0.52 ℃ per decade from 1981 to 2016, but seasonal forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF; Johnson et al., 2019) consistently underestimate the observed trend by 0.31 ℃ per decade.

Applications of seasonal climate forecasts require forecast calibration to reduce forecast bias and to quantify forecast uncertainty (Gneiting et al., 2005; Zhao et al., 2017). Various forecast calibration approaches, such as the Bayesian joint probability (BJP) modelling approach (Wang and Robertson, 2011; Wang et al., 2019), are by design incapable of fixing trend disparities. For the Melbourne example above, the trend in BJP calibrated forecasts is still underestimated by 0.30 ℃ per decade. For informative decision-making, seasonal forecasts are expected to simulate natural variability, which has a tangled relationship with long-term trends caused by climate change (Doblas-Reyes et al., 2006; Strazzo et al., 2019; Weisheimer et al., 2011). In this regard, embedding the climate trend into seasonal forecasts could potentially improve forecast performance, particularly in regions with strong temporal trends.

Several studies dealt with the trend problem in forecasts at seasonal or decadal timescales, for example, in seasonal forecasts of sea ice concentration (Dirkson et al., 2019), and decadal predictions of annual global mean temperature (Kharin et al., 2012; Sansom et al., 2016). In Chapter 2, we introduced a new statistical post-processing model to embed observed trends into seasonal forecasts of temperature and validated the model on three test sites in Australia. In this method, trends are entirely inferred from the model training data. However, given limited data records, the inferred trends can be subject to large sampling errors. In this study, we further develop the methodology to account for trend uncertainty and demonstrate the value of this new trend-aware forecast calibration methodology on a continental spatial scale, which is the predominant focus of seasonal forecasting services globally (Strazzo et al., 2019). Detailed analyses are conducted on observations and forecasts of minimum and maximum temperatures

for Australia. By means of this study, we advocate the importance of the seasonal forecasts to properly represent trends caused by climate change.

3.4 Study Data

3.4.1 SEAS5 forecasts of temperature

Gridded daily ensemble forecasts come from ECMWF's SEAS5 seasonal forecasting system (Johnson et al., 2019). It is a fully coupled general circulation model, composed of land, atmosphere, and ocean components with a sea-ice model LIM2 embedded. The ECMWF's IFS (integrated forecast system) atmosphere model cycle 43r1 is implemented for the atmosphere component with a horizontal resolution of 36 km. The MACC (Monitoring Atmospheric Composition and Climate) reanalysis (Inness et al., 2013) is used to calculate scaled seasonally varying climatology for greenhouse gas radiative forcing. Such forcing could capture the long-term trend in greenhouse gas emissions as used in CMIP5 historical greenhouse gases over 1981-2000 and CMIP5 RCP 3-PD since 2000. Tropospheric sulfate aerosol uses CMIP5 climatology that varies decadally. The NEMO (Nucleus for European Modelling of the Ocean) v3.4.1 model is established for the ocean component at 0.25° horizontal resolution. Details of model initialisation schemes and ensemble generation techniques can be found in Johnson et al. (2019).

SEAS5 hindcasts are available from January 1981 to December 2016 with 25 ensemble members initialised on the 1st of each calendar month and running for 7 months. In this study, target variables are seasonal averages of daily minimum temperature (Tmin) and maximum temperature (Tmax) at 1-month lead time. Here, a 1-month lead seasonal forecast is defined as the forecast for a rolling season beginning one month after the initial date. In this regard, target forecasts are from February-April (FMA) 1981 to January-March (JFM) 2017 for 12 overlapping 3-month seasons.

3.4.2 Observed temperature

Gridded seasonal Tmin and Tmax observed data are obtained from the high-quality AWAP (Australian Water Availability Project) climate dataset (Jones et al., 2009). The AWAP data are originally on a 0.05° grid and are regridded to match SEAS5 data at 0.4° by employing a bilinear interpolation method. In this work, the evaluation period is from February 1981 to March 2017, which is aligned with SEAS5 hindcast data. Seasonal Tmin and Tmax data are derived by averaging monthly mean values over three sequential months.

3.5 Methods

3.5.1 A trend-aware forecast post-processing methodology

Here we introduce a trend-aware methodology, modified from the Bayesian joint probability (BJP) modelling approach (Wang et al., 2019) to handle the difference in trends between the observations and raw forecasts. The methodology builds on Chapter 2, with further refinement in treatment of trend uncertainty. That is, in this Bayesian method, we construct priors for trend parameters based on regional information.

Before formulating the post-processing model, we check the normality of temperature variables and find both raw forecasts and observations are mildly different from the normal distribution (not shown). To be prudent, we still employ the single-parameter Yeo-Johnson transformation (Yeo and Johnson, 2000) to transform the potentially nonnormal temperature variables for all grid cells to fulfil the model assumption of a joint normal distribution. Raw forecast ensemble mean y_1 and observation y_2 are thereby transformed to $y_1^{'}$ and $y_2^{'}$ given the transformation parameter values λ_1 and λ_2. The transformation parameters are obtained separately for y_1 and y_2 by using a Bayesian maximum a posteriori (MAP) method (Schepen et al., 2016). We then assume a joint distribution of detrended (as detailed below) transformed predictor z_1 (raw forecast mean) and detrended transformed predictand z_2 (observed data), given as

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim \mathrm{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \tag{3-1}$$

where $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\Sigma}$ is the covariance matrix.

Individual points $z_1(t)$ and $z_2(t)$ for each of n forecast years, $t = 1, 2, \ldots, n$, also known as the anomaly from the trendline of the variables, are derived by

$$z_1(t) = y_1^{'}(t) - \alpha_1(t - t_m) \tag{3-2}$$

$$z_2(t) = y_2^{'}(t) - \alpha_2(t - t_m) \tag{3-3}$$

where α_1 and α_2 are trend parameters for predictors and predictands respectively, and t_m is approximately the middle point (e.g., 19 in this work) in the evaluation time period. We denote the parameter sets for the model inference as $\boldsymbol{\theta} = \left\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha_1, \alpha_2\right\}$.

In model inference mode, model parameters are inferred from the sequence of training data pairs for n years: $\mathbf{D} = \left\{[y_1^{'}(t), y_2^{'}(t)], t = 1, 2, \ldots n\right\}$. The posterior distribution of the model parameters is

$$p(\boldsymbol{\theta} \mid \mathbf{D}) \propto p(\boldsymbol{\theta}) p(\mathbf{D} \mid \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{t=1}^{n} p(\mathbf{D} \mid \boldsymbol{\theta}) \tag{3-4}$$

where $p(\boldsymbol{\theta})$ is the prior distribution for model parameters and $p(\mathbf{D} \mid \boldsymbol{\theta})$ is the likelihood function. A Gibbs sampler is utilised to repeatedly sample parameters from the conditional posterior distributions of the model parameters (Shao et al., 2021a; Wang et al., 2019).

In the predictive mode, we obtain a trend-embedded calibrated forecast $y_2^{'}(t^*)$ given a new transformed predictor $y_1^{'}(t^*)$ using each of the parameter sets $\boldsymbol{\theta}$. Here, a pragmatic approach is used to adjust extremely small or large $z_1^{'}(t^*)$ values that occur in prediction (Wang et al., 2019). In this study, we set extreme thresholds as 0.001 and 0.999 in non-exceedance probability according to the marginal distribution of z_1. With the back transformation of all forecast values, we derive a collection of the calibrated forecasts $y_2(t^*)$ to represent forecast uncertainty.

Since we introduce new parameters corresponding to trends in raw forecasts and observations, we are required to specify priors for these new parameters. Here, we present two methods for the prior specification, leading to two models: BJP-t and BJP-ti. The BJP-t model was introduced in Chapter 2. In this model, the trend is entirely inferred from the model training data. Practically, this is achieved by using a uniform (non-informative) prior for the trend parameters. The BJP-ti model is a variation of the BJP-t model. In the BJP-ti model, the trend is also inferred from data, but with a degree of moderation, achieved by using zero-centred normal (informative) priors, in favour of no or weak trends in calibrated forecasts. The rationale for using the informative priors is that trends inferred from data that cover only a period of 36 years are subject to sampling errors and may not accurately represent the true underlying trends. The informative prior distributions $p(\alpha_i)$, $i = 1, 2$, are written as

$$p(\alpha_i) \propto N(0, m_i^2) \tag{3-5}$$

We specify m_i as $m_i = \delta_i \times m_i^{''}$, where δ_i is the MAP estimate of standard deviation of $y_i^{'}$ found at the data transformation step. Here, δ_i acts as a scaling factor to account for the possible effect of the transformation on the trend parameter α_i. Theroretically, it is more reasonable to use a

common value for $m_i^{"}$ than m_i for all seasons and locations. In specifying the prior, we do not consider the uncertainty of δ_i. Based on the method of Gibbs sampling, the conditional posterior distribution of parameter α_i can be derived as

$$\left[\alpha_i \mid \cdot\right] = \mathrm{N}\left\{\frac{m_i^2 \sum_{t=1}^{n} [y_i^{'}(t) - \mu_i](t - t_m)}{m_i^2 \sum_{t=1}^{n} (t - t_m)^2 + \sigma_i^2}, \frac{m_i^2 \sigma_i^2}{m_i^2 \sum_{t=1}^{n} (t - t_m)^2 + \sigma_i^2}\right\} \tag{3-6}$$

where μ_i is a mean and σ_i is a standard deviation for detrended transformed variables z_i.

The selection of $m_i^{"}$ is based on the following experiment. We calculate and record the ensemble median of the BJP-t estimated α_i in model inference mode for each grid cell and divide the trend slope by δ_i. We repeat the process for all 12 overlapping seasons in Australia. In total, we derive the trend value/ δ_i for 4627 (total number of grid cells) ×12 (number of seasons) times, separately for raw forecasts and observations and separately for Tmin and Tmax. For consistent applications of the BJP-ti model, $m_i^{"}$ is set as a fixed value for all the seasons and locations in Australia. Thus, we pool the trend slopes/ δ_i to summarize the results. For Tmin, 95% of raw forecasts and observations have absolute trend slopes/ δ_i value less than 0.07 and 0.05 respectively. Thus, we set $m_i^{"}$ as 0.05 for raw forecasts and as 0.03 for observations. For Tmax, 95% of raw forecasts and observations have absolute trend less than 0.06 and 0.05 respectively. Thus, we set $m_i^{"}$ as 0.04 for raw forecasts and as 0.03 for observations. This zero-centred normal prior has the effect of slightly moderating the trends in data. The Gibbs sampling implementation for the trend-aware models follows the pseudocode of the BJP-t (Shao et al., 2021a) and BJP model (Wang et al., 2019).

3.5.2 Forecast evaluation

We evaluate and compare raw and post-processed seasonal ensemble forecasts based on their ability to capture observed trends and to be skillful, sharp, and reliable. The evaluation of the ensemble forecasts is conducted at each of 4627 grid cells and for each season separately over Australia under a leave-one-year-out cross validation setup. Each historical event is validated using the model trained by the remaining data points. We note that this cross validation is only appropriate for evaluating the anomaly component rather than the trend component in the trend-aware calibration (Shao et al., 2021a). For the trend component, the results from the cross validation are similar to those of the model fitting (Shao et al., 2021a). This is the limitation of our leave-one-year-out cross validation when it comes to trend evaluation. Furthermore, with this

validation method, the resulting forecast skill may not be extended to the real-time forecasts because the trend in individual cross-validated forecast is inferred from the remaining forecast events that may include future information while the complete trend information from the past evaluation period will be embedded into the real-time forecasts. A proper cross validation could be carried out by just leaving out short data periods at the start or end years of the data records. However, the validation results could be subject to large sampling errors because of the limited number of events used in the validation. If there is a longer data period, say 50 years, this alternative may be more ideal for model validation. For this study, since it is not possible to circumvent the limitation under the cross-validation setup, we need to be cautious when interpreting results.

We estimate decadal trends as the slope of a linear regression (Hartmann et al., 2013), measuring how well the ensemble forecasts reproduce the observed trend. The two-tailed t test is used as trend significance test, and the significance is judged within 99% and 95% confidence intervals.

We use forecast skill to evaluate the accuracy of ensemble forecasts by characterizing the difference between the probabilistic forecasts and observed data. Here, we calculate the continuous ranked probability score (CRPS; Matheson and Winkler, 1976; Hersbach, 2000) for individual events, and then compare the averaged CRPS of the forecasts with the averaged CRPS of reference forecasts over the forecast period to derive the CRPS skill score. The leave-one-year-out cross-validated climatology ensemble forecasts from the BJP model (Wang et al., 2019) are used as the reference forecasts in the score calculation for the BJP, BJP-t, and BJP-ti calibrated forecasts. For each grid cell, the CRPS at each point t is given as

$$\mathrm{CRPS}(t) = \int \left\{ F(t,y) - H\left[y - y_o(t) \right] \right\}^2 dy \tag{3-7}$$

where $F(t,y)$ is the cumulative distribution function (CDF) constructed from the ensemble forecasts, $y_o(t)$ is the observed value; and H is the Heaviside step function which equals 0 if $y < y_o(t)$ and equals 1 otherwise. The CRPS skill score is calculated as

$$\mathrm{CRPS}_{\text{skill score}} = \frac{\overline{\mathrm{CRPS}_{\text{ref}}} - \overline{\mathrm{CRPS}}}{\overline{\mathrm{CRPS}_{\text{ref}}}} \times 100 \quad (\text{unit: \%}) \tag{3-8}$$

The CRPS skill score is positively oriented. The higher the CRPS skill score is, the more skillful the forecasts are. A score of -5 to 5 indicates that the forecasts do not have much skill, and ensemble forecasts performs similarly to climatology forecasts (Schepen et al., 2016).

Sharpness refers to the ability of the ensemble forecasts to predict extreme events, which is a property of forecast only (Gneiting et al., 2007). Reliable ensemble forecasts with maximal sharpness are more desirable to distinguish from climatology forecasts. We check the sharpness of the ensemble forecasts by numerically calculating the average width of the central 50% [0.25, 0.75], 80% [0.1, 0.9] and 90% [0.05, 0.95] prediction intervals for all individual events (Gneiting et al., 2007). The narrower interval width indicates sharper probabilistic forecasts.

For individual cells, we set the averaged interval width of the BJP calibrated forecasts as a baseline and compare the width of trend-aware forecasts against it. The ratio of the interval widths is termed sharpness ratio. A ratio lower than 1 means the trend-aware calibration results in sharper ensemble forecasts. Then the sharpness ratios are pooled across Australia in all seasons, and boxplots are used to visualise the overall forecast performance.

Reliability is an indication of the statistical consistency between ensemble forecasts and observations (Wang et al., 2009). Here, we calculate the probability integral transforms (PITs; Gneiting et al., 2007) of the observations, and then derive the PIT scores to quantitatively measure the deviation of the PIT values from the corresponding uniform quantiles for individual grid cell (Renard et al., 2010). For an observational event $y_o(t)$ and corresponding forecast CDF $F(t, y)$, the PIT value $\pi(t)$ is defined by

$$\pi(t) = F\left[t, y = y_o(t)\right] \tag{3-9}$$

For a reliable forecasting system, the collection of PIT values follows a standard uniform distribution. The PIT score is calculated by combining Eqs. (23a) and (23b) in Renard et al. (2010), which is given as

$$\text{PIT score} = 1.0 - \frac{2}{n}\sum_{j=1}^{n}\left|\pi(j) - \frac{j}{n+1}\right| \tag{3-10}$$

where $\pi(j)$ is the j^{th} ranked PIT value $\pi(t)$ and $j/(n+1)$ is the j^{th} theoretical $\pi(j)$ value. Within the range from 0 to 1, the higher the PIT score is, the more reliable the forecasts are. As with the sharpness ratio, we pool the PIT scores for all cells to summarise the forecast performance in boxplots.

3.6 Results and Discussions

3.6.1 Trends not captured in existing forecasts

To demonstrate the limitations of existing forecasts regarding trend reproduction, we evaluate and compare linear decadal trends in observations and "current generation" of forecasts, including both raw forecasts and forecasts calibrated using existing BJP technology (Wang et al., 2019). For brevity, we limit our results in this section to the four main seasons (Figure 3-1 and Figure 3-2). The full geographic distributions of observed and forecast trends for all 12 overlapping seasons are given in Supplementary Figure S2-1 and Figure S2-2.

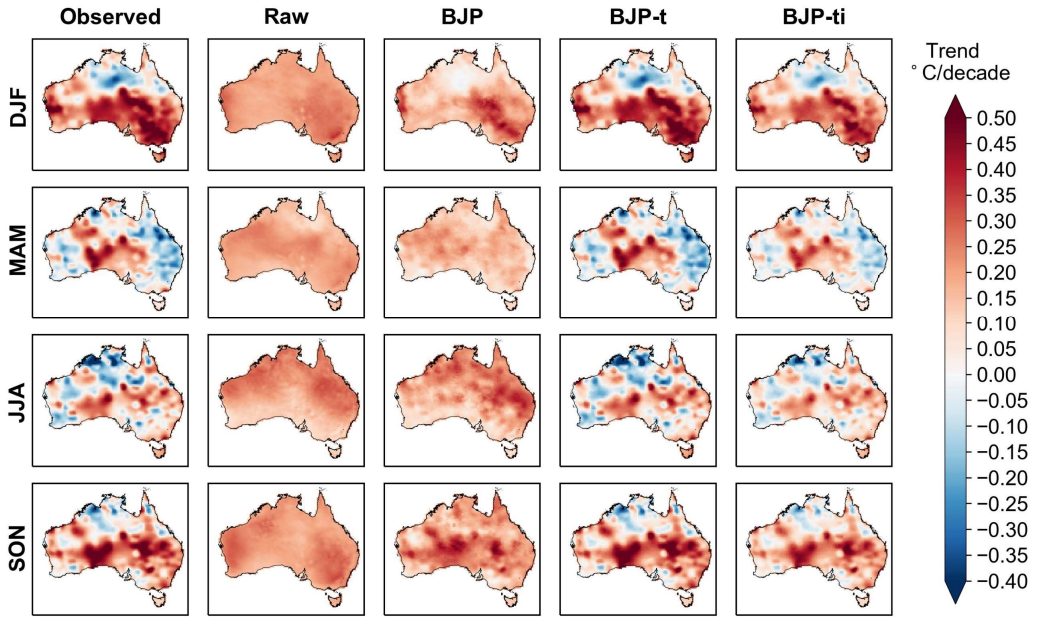

Figure 3-1: Linear decadal trends of seasonal averages of Tmin for observations, raw, BJP, BJP-t, and BJP-ti calibrated forecasts for four seasons at 1-month lead time from MAM 1981 to DJF 2016.

Observed temperatures show noticeable trends over the 36-yr period 1981-2016 (first column in Figure 3-1 and Figure 3-2). Both warming and cooling trends are widespread in Tmin, whereas warming trends are predominant in Tmax, except for northern Australia in DJF (hereafter seasons are abbreviated with initial letters of three consecutive months). Strong, and often statistically significant, trends can be visualised in regional clusters for all seasons across Australia. For example, in Tmin, there is a strong and significant warming in southern Australia in DJF while a significant cooling (at the 5% significance level) is found elsewhere throughout seasons

(Supplementary Figure S2-3). For Tmax, there is a strong and significant warming trend in south-eastern Australia in SON and DJF (Supplementary Figure S2-3).

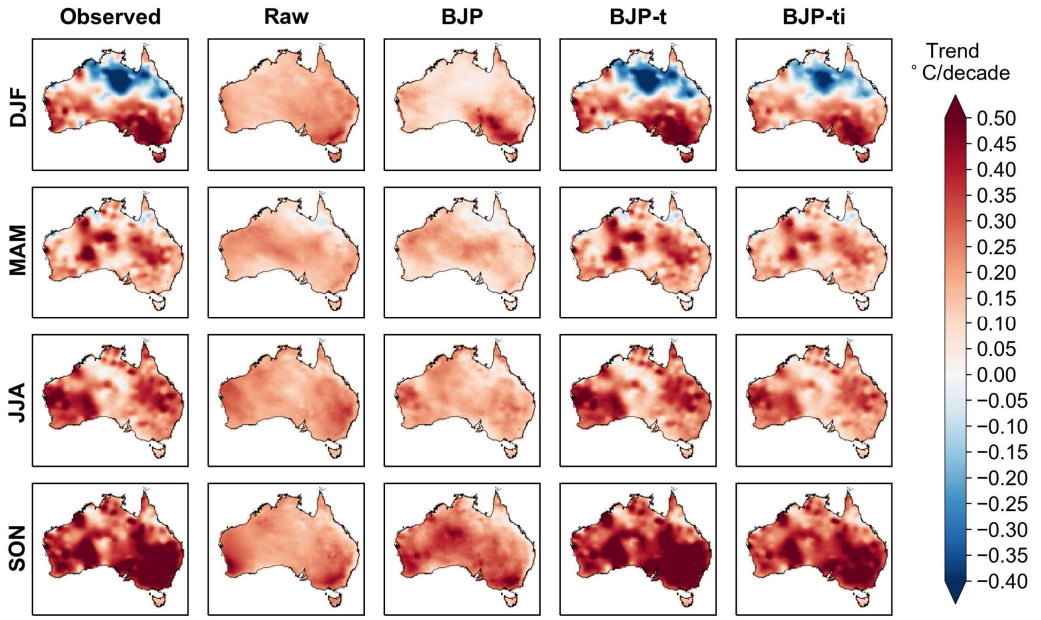

Figure 3-2: As in Figure 3-1, but for Tmax.

Raw GCM forecasts are shown to widely misrepresent the observed trend across Australia. Warming trends without much spatial variability dominate all seasons for raw Tmin forecasts (second column in Figure 3-1). Increasing trends are also detectable for raw Tmax forecasts, but the slopes are flatter than observed trends in most regions (second column in Figure 3-2). It is not clear what contributes to the failure of the SEAS5 model in reproducing the observed trend. Potential factors include unrealistic modelling of greenhouse gas emission (Jia and Lin, 2013), and the propensity of the forecasting system to drift back to a biased state after being initialised to observed datasets (Hermanson et al., 2018). However, key contributors remain unknown, and further investigations are required.

Similarly, the BJP calibrated forecasts misrepresent the observed trend across most of the Australian continent (third column in Figure 3-1 and Figure 3-2). However, greater spatial variability is exhibited in the trends of BJP post-processed forecasts than in the trends of raw forecasts. The trend appears to be a little more consistent with the observed in some regions, such as in eastern Australia for Tmin and south-eastern Australia for Tmax in DJF.

3.6.2 Embedding observed trends into ensemble forecasts

To better reproduce observed climatic trends in seasonal forecasts, we have proposed two trend-aware models: BJP-t and BJP-ti to explicitly model trends in raw forecasts and observations. The BJP-t calibrated forecasts appear to accurately reproduce the observed trends (fourth column in Figure 3-1 and Figure 3-2). In comparison, trends in BJP-ti calibrated forecasts (fifth column in Figure 3-1 and Figure 3-2) are milder than observed data, but the trend signal is still well reproduced. Returning to the Melbourne example given in the Introduction (Section 3.3), the BJP-t and BJP-ti calibrated forecasts result in trends of 0.53 ℃ per decade and 0.45 ℃ per decade, which are more closely aligned with the observed trend of 0.52 ℃ per decade than the BJP calibrated forecasts. There is the question then: which model is more robust and should be recommended for use in the forecast communities? We will answer this question after evaluating a range of other forecast attributes.

First, we evaluate the forecast skill and plot the skill difference between the BJP-ti and BJP calibrated forecasts (Figure 3-3). For both Tmin and Tmax, the BJP-ti calibration leads to widespread skill gain (colored in dark blue) and little skill loss (colored in dark red) compared to the BJP calibration in all seasons over Australia. Considerable skill improvement is achieved in the regions where the BJP calibrated forecasts wrongly represent the trend direction, or where the observed trend is significant at 5% (Supplementary Figure S2-3). Visual examination shows that some regions and seasons benefit more after the BJP-ti calibration. For Tmin forecasts, the skill gain is pronounced in most seasons because the underlying trend is not satisfactorily reproduced by the BJP calibration in broader cases (Supplementary Figure S2-1). Examples are northern Australia from JJA to DJF and parts of southern Australia from OND to FMA. Visible skill gain for Tmax is limited compared to that for Tmin, which mainly concentrates in northern and western Australia from OND to DFJ. In all seasons, the BJP-ti calibration slightly degrades forecast skill when the BJP calibrated forecasts already capture the observed trend. Some seasons are dominated by the skill loss in most regions, for example, MAM for Tmax. This is further confirmed in the numerical summary of skill scores of the calibrated forecasts (Figure 3-4), expressed by the percentage of the grid cells within each of the score ranges. Compared with the BJP model (Figure 3-4a), the BJP-ti calibrated forecasts (Figure 3-4c) produce more cells with positive skill scores in many seasons for Tmin. This is more evident for the score higher than 5%, where the forecasts are deemed more skillful than climatology. The improvement in skill scores larger than 5% is most pronounced for JFM, JJA, and JAS. For Tmax forecasts, improvements

afforded by the BJP-ti calibration are confined to OND and NDJ, while forecast skill is comparable to the BJP calibrated forecasts for other seasons.

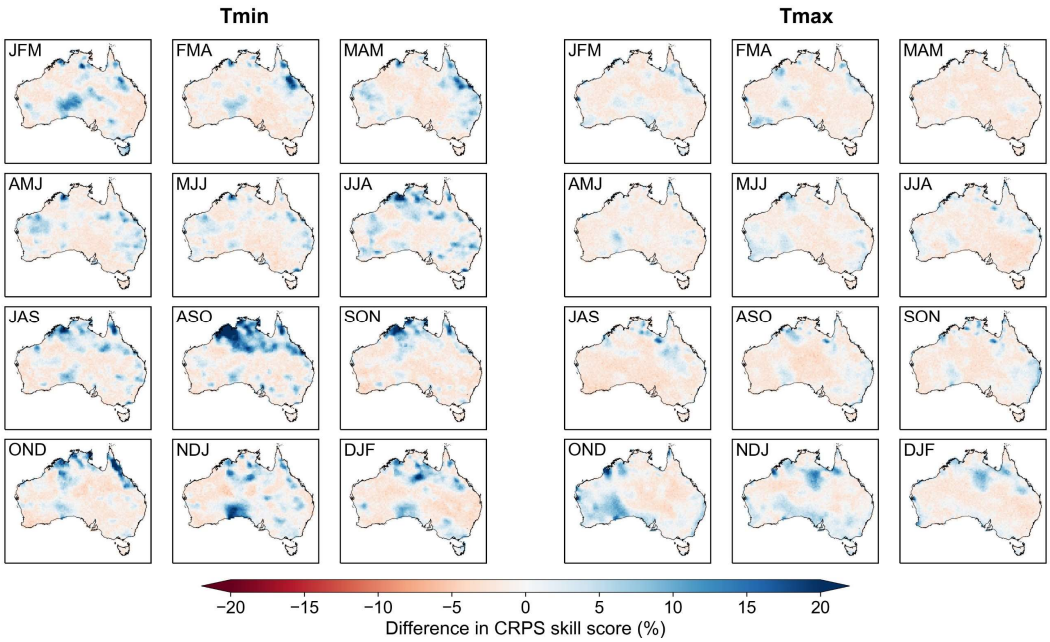

Figure 3-3: CRPS skill score difference between BJP-ti and BJP calibrated forecasts of seasonal averages of Tmin (left) and Tmax (right) at 1-month lead time. The skill score is calculated using leave-one-year-out cross-validated climatology ensemble forecasts from the BJP model as the reference forecasts.

When we examine the numerical summary of the skill scores for BJP-ti and BJP-t calibrated forecasts (Figure 3-4), the BJP-t calibration (Figure 3-4b) is shown to yield a lower percentage of grid cells with positive skills than the BJP-ti calibration (Figure 3-4c). This is also evident from the skill difference between the BJP-ti and BJP-t model (Supplementary Figure S2-4). Visually, the BJP-ti calibrated forecasts are found more skillful (colored in blue) than the BJP-t calibrated forecasts in most regions across all seasons. Recall that the BJP-t model applies non-informative priors for trend components, which could incur errors when trends are entirely inferred from available data series, because of the sampling variability of the observed trends. In contrast, a degree of moderation applied in the BJP-ti model can have a positive effect on the performance of out-of-sample calibrated forecasts. This statement is further supported by an assessment of forecast sharpness. Overall, we find that BJP-ti model (Supplementary Figure S2-5c and Figure S2-5d) produces sharper forecasts than BJP-t model (Supplementary Figure S2-5a and Figure S2-5b), and the ensemble spread of the BJP-ti calibrated forecasts is closer to that of the BJP calibrated forecasts. The final CRPS skill scores of the BJP-ti calibrated forecasts are shown in

Figure 3-5. Generally, positive skill dominates most areas and seasons for both Tmin and Tmax with 1-month forecast lead time. As visualized in Figure 3-4c, after the BJP-ti calibration, for Tmin forecasts, the percentage of cells with positive skill scores is the highest for JJA, NDJ and DJF. For Tmax forecasts, the largest proportion is found in SON, OND and NDJ, where less than 5% of the cells have negative skills. Skillful or at least climatology-like forecasts are produced across all seasons, demonstrating the effectiveness of the BJP-ti model as a post-processing tool. Furthermore, the spatial pattern of the skill score (Figure 3-5) reveals that positive skills are prevalent in the areas with strong warming or cooling trends, for example, northern Australia in ASO for Tmin and western Australia in OND for Tmax. In the regions with skillful forecasts but weak trends, it is expected that the teleconnection patterns, such as ENSO and Indian Ocean dipole, are well modelled in SEAS5 and represent nearly all the underlying forecast skill (Schepen et al., 2016; Wang et al., 2019).

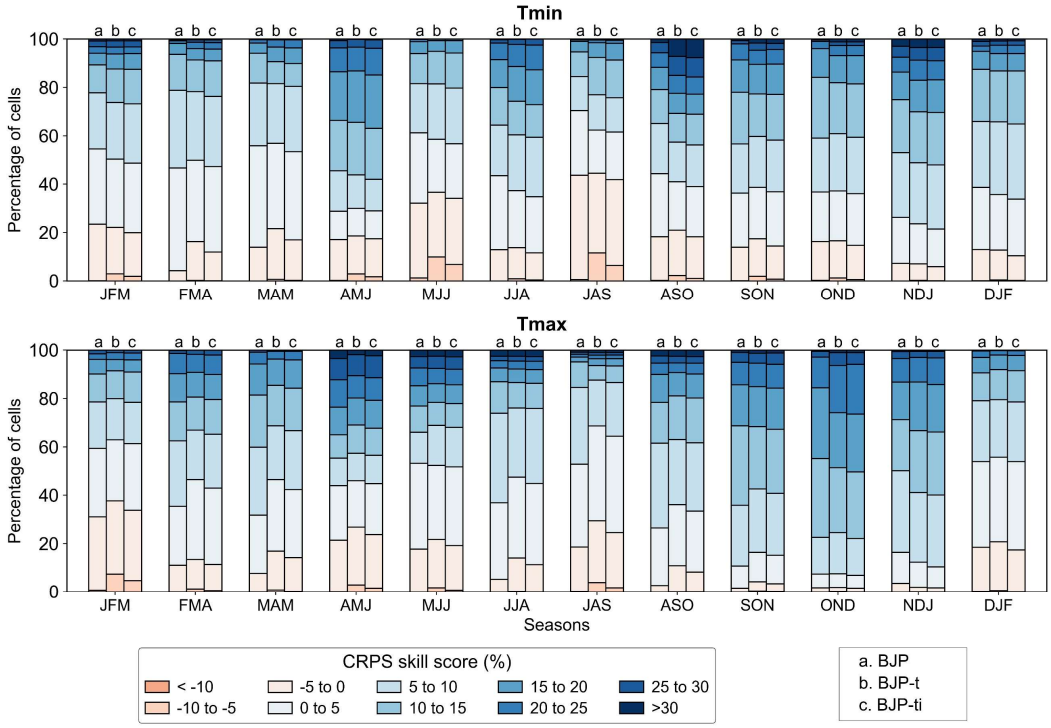

Figure 3-4: Percentage of the grid cells where the CRPS skill score lies in a range of values for BJP, BJP-t, and BJP-ti calibrated of seasonal averages of (top) Tmin and (bottom) Tmax at 1-month lead time. The skill score is calculated using leave-one-year-out cross-validated climatology ensemble forecasts from the BJP model as the reference forecasts.

Besides forecast skill, forecasts are also examined for reliability in ensemble spread to represent forecast uncertainty. Pooled PIT score results show that overall, the BJP-t and BJP-ti models

produce more statistically reliable ensemble forecasts than the BJP calibration, while the difference between BJP-ti and BJP-t calibrated forecasts is minor (Supplementary Figure S2-6). This finding is supported by the averaged values of pooled PIT scores for the BJP, BJP-t, and BJP-ti calibrated forecasts, which are 0.9377, 0.9384 and 0.9383 for Tmin, and 0.9365, 0.9376 and 0.9376 for Tmax, respectively.

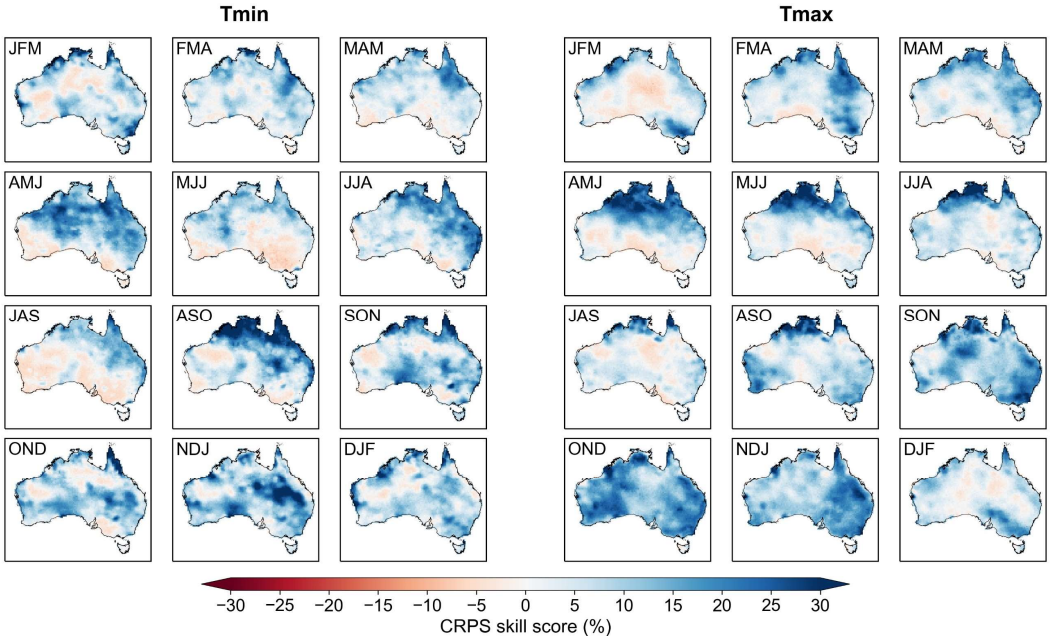

Figure 3-5: CRPS skill score of BJP-ti calibrated forecasts for seasonal averages of (left) Tmin and (right) Tmax at 1-month lead time. The skill score is calculated using leave-one-year-out cross-validated climatology ensemble forecasts from the BJP model as the reference forecasts.

3.6.3 Discussions

In this study, we use a zero-centred normal prior for trend parameters in the BJP-ti model. It is also possible to use a nonzero location parameter in the priors. Such approach supports a belief that the average trend across Australia over 1981-2016 is 'real' while our prior starts with a belief that such average trend could have come from sampling variability. While it is a valid alternative to employ the normal prior with a nonzero location parameter, we find it make little difference to the results in practice.

In the trend maps of observed temperatures (first column in Figure 3-1 and Figure 3-2), there appears to be some patchiness. In the sparsely gauged regions, such as in central Australia, the patchiness could have been caused by the trends in individual stations because the trend in one

station could be broadcast to surrounding grid cells without weather stations. Having said that, the scale of the patchiness tends to be larger than the distances between individual stations, indicating that the observed trends in adjacent stations may have similar magnitude or direction. Otherwise, there would have been a lot more randomness in the trend maps with individual stations all go randomly in trends.

3.7 Conclusions

Seasonal climate forecasts capable of capturing climatic trends could engender user confidence in deploying the forecasts for adapting to changing climate. However, raw seasonal forecasts generated from climate models were reported not to reproduce the observed trend (Krakauer, 2019). Furthermore, this issue is not addressed by most existing forecast post-processing methods that are needed to remedy biases and uncertainties in model forecasts. Here, we confirm that raw and BJP calibrated SEAS5 seasonal forecasts of minimum and maximum temperatures do not properly reproduce the observed trend for Australia. To resolve this issue, in our previous research, we introduced a new trend-aware forecast calibration model, BJP-t, to incorporate observed trends into seasonal temperature forecasts. With this method, the inferred trends are subject to large sampling errors because of limited data records. In this study, we further develop the method to account for trend uncertainty, in which regional information is used to construct priors for the trend parameters. We present and compare two models, BJP-t with non-informative priors for the trend parameters, and BJP-ti with informative priors for the trend parameters.

We find that the calibrated forecasts from the BJP-t model accurately reproduce the observed trend, while the calibrated forecasts from the BJP-ti model adopt a slightly moderated trend. Compared to the BJP model that does not explicitly incorporate any trend, both trend-aware models lead to marked skill improvements when the BJP calibrated forecasts wrongly represent the trend direction, or when the observed trend is statistically significant across all seasons. Elsewhere, the BJP-t model can result in a slight skill degradation because extra trend parameters are introduced into the algorithm and the trends are entirely inferred from the training data, which give rise to a larger spread in the calibrated forecasts, especially when the BJP calibrated forecasts already reproduce the observed trend. For comparison, the BJP-ti model leads to less skill degradation by using priors to moderate the trend parameters, in recognition that the observed trend is subject to sampling errors and therefore needs not to be exactly reproduced. The model

generally produces slightly sharper forecasts than the BJP-t model. Overall, we recommend the use of the BJP-ti model.

Our trend-aware forecast calibration methodology should be suitable for post-processing seasonal temperature forecasts from other GCMs (e.g., Hudson et al., 2017; Hudson et al., 2018), and forecasts of other meteorological or hydrological variables that exhibit significant trends in the past decades such as precipitation (Lausier and Jain, 2018; Polade et al., 2017) and evapotranspiration (Jung et al., 2010; Peng et al., 2017). In future work, we will also investigate incorporating the spread information of the raw forecasts into the methodology and explore the use of the methodology for post-processing sub-seasonal climate forecasts, which serve as a middle ground for seamless prediction between weather and climate forecasts (Robertson et al., 2015).

Chapter 4 Adapting the Trend-aware Post-processing Method to Post-process Seasonal Precipitation Forecasts

4.1 Preamble

In Chapter 3, I further extended the trend-aware forecast post-processing method to refine the treatment of trend uncertainty associated with the sampling variability and demonstrated the value of the improved method for the continental application. Results revealed that with careful treatment of trend uncertainty, the trend-aware calibrated forecasts had proper trends embedded. Compared to the BJP calibrated forecasts, the trend-aware calibration led to more skillful forecasts when observed trends are significant or when the BJP calibrated forecasts misrepresented the observed trend direction.

While Chapter 2 and Chapter 3 have showed that the trend-aware method is highly effective for post-processing seasonal temperature forecasts, the method has not been applied to forecast other meteorological variables, such as seasonal precipitation. In fact, statistical post-processing of precipitation forecasts is more challenging due to the special characteristics of the precipitation amounts, including being lower bounded at zero, being positively skewed and being more variable in space and time, and inherently more uncertain than temperature variables.

Accordingly, Chapter 4 answers RQ3: How can the trend-aware model be adapted to post-process seasonal precipitation forecasts? Here, I introduce new formulation and evaluation tools to account for the abovementioned precipitation features. The more advanced version is applied to post-process SEAS5 forecasts of seasonal precipitation for the Australian continent and selected cases.

This chapter has been published in Journal of Hydrometeorology (Impact Factor 3.891). The paper is titled as 'Improved trend-aware post-processing of GCM seasonal precipitation forecasts' and the authorship is Shao, Y., Wang, Q. J., Schepen, A., Ryu, D and Pappenberger, F.

4.2 Abstract

Climate trends have been observed over the recent decades in many parts of the world, but current global climate models (GCMs) for seasonal climate forecasting often fail to capture these trends. As a result, model forecasts may be biased above or below the trendline. In our previous research, we developed a trend-aware forecast post-processing method to overcome this problem. The method was demonstrated to be effective for embedding observed trends into seasonal temperature forecasts. In this study, we further develop the method for post-processing GCM seasonal precipitation forecasts. We introduce new formulation and evaluation features to cater for special characteristics of precipitation amounts, such as having a zero lower bound and highly positive skewness. We apply the improved method to calibrate ECMWF SEAS5 forecasts of seasonal precipitation for Australia. Our evaluation shows that the calibrated forecasts reproduce observed trends over the hindcast period of 36 years. In some regions where observed trends are statistically significant, forecast skill is greatly improved by embedding trends into the forecasts. In most regions, the calibrated forecasts outperform the raw forecasts in terms of bias, skill, and reliability. Wider applications of the new trend-aware post-processing method are expected to boost user confidence in seasonal precipitation forecasts.

4.3 Introduction

Skillful seasonal climate forecasts are valuable for managing climate variability and change (An-Vo et al., 2019; Pechlivanidis et al., 2020). Global climate models (GCMs) are commonly employed to produce seasonal climate forecasts (Johnson et al., 2019; Kirtman et al., 2014; Saha et al., 2014). Typically, GCMs are run to generate retrospective forecasts (re-forecasts) for a historical period of two to four decades. These re-forecasts are mainly used to evaluate forecast performance, to produce tailored products such as anomalies, and to establish calibration models for new forecasts. One issue that has been identified is the inability of the GCM re-forecasts to capture observed climate trends (Cai et al., 2009; Krakauer, 2019; Shao et al., 2021a; Shin and Huang, 2019). This inability lowers seasonal climate forecast skill and reliability and, importantly, undermines user confidence in using the forecasts (Barnston et al., 2010; Livezey and Timofeyeva, 2008).

Precipitation is a climate variable of crucial importance to climate-sensitive sectors, such as agriculture and water resource management. In recent decades, precipitation has exhibited both increasing and decreasing trends around the world (Hartmann et al., 2013). For example, mixed

state-wide trends have been observed in seasonal and annual precipitation variables in North Carolina in the United States for 1950-2009 (Sayemuzzaman and Jha, 2014). In Australia, south-western and south-eastern parts have a declining trend in April to October while most of northern Australia have increased precipitation across all seasons since 1970s, particularly during northern wet season from October to April (Bhend and Whetton, 2015; CSIRO and Australian Government Bureau of Meteorology, 2020; Wasko et al., 2021). Despite notable changes in precipitation, the observed trends are often sensitive to the evaluation periods, which may substantially vary in time (Hartmann et al., 2013) and their associated uncertainty needs to be carefully interpreted.

Previous research has explored how the climatic trend is represented in seasonal precipitation forecasts produced by global climate models (GCMs). Huang et al. (2019) compared trends in observations and 2-month ahead U.S. seasonal precipitation re-forecasts from a modified version of the CFSv2 model for 1958-78, 1979-99, and 2000-2017 separately. They found that the re-forecasts roughly reproduced the observed trends in winter over the full 60-year period and in spring and summer since the 2000s. However, the re-forecasts failed to capture the observed trends in spring, summer, and autumn during 1958-1978 and 1979-1999. In this study, we will demonstrate that precipitation re-forecasts of ECMWF SEAS5 (Johnson et al., 2019) mismatch trends seen in observations over parts of Australia in some seasons.

While an ultimate solution to the trend mismatch problem lies in further improving the GCMs, there is a practical approach that can yield more immediate benefits, that is, observed climate trends can be embedded into forecasts through statistical post-processing of GCM raw forecasts. Post-processing has, in the past, aimed at removing biases and improving skill and reliability of forecasts. Methods for addressing the trend issue in seasonal forecasts are beginning to emerge (Dirkson et al., 2019; Krikken et al., 2016). Most recently, a trend-aware method was developed to embed trends as well as achieving other aims of post-processing (Shao et al., 2021a, 2021b). Building on a Bayesian joint probability (BJP) modelling approach (Wang and Robertson, 2011; Wang et al., 2009; Wang et al., 2019), this method explicitly models trends in both observations and GCM forecasts. This method has been shown to be effective for post-processing seasonal temperature forecasts in Chapter 2 and Chapter 3.

Before employing the trend-aware method to post-process forecasts of seasonal accumulated precipitation, we need to give careful attention to the following special characteristics of the precipitation data: 1) precipitation records have a natural lower bound of zero occurrence, which is not compatible with the use of continuous bivariate normal distribution in the trend-aware

model (Shao et al., 2021a); 2) precipitation amounts can be strongly positively skewed; 3) precipitation records and trends are often associated with large uncertainties (Hartmann et al., 2013) and trend magnitude varies widely across regions (Kumar et al., 2013), as a result of underlying physical processes (Rowell, 2012) that lead to a lower spatial and temporal auto-correlation of the precipitation than temperature.

In this chapter, we extend the trend-aware method for post-processing GCM forecasts of precipitation. New formulations and evaluation features are introduced to account for the above characteristics of the precipitation amount. We evaluate the improved method on ECMWF SEAS5 seasonal forecasts of precipitation (i.e., Total precipitation with ECMWF parameter ID 228) for the Australian continent.

The remainder of the chapter is organized as follows. Section 4.4 introduces datasets of SEAS5 forecasts and observations. Section 4.5 describes the trend-aware method and forecast verification metrics while Appendix S3 supplements the trend-aware algorithm. Section 4.6 presents the findings. Section 4.7 discusses the results and extension opportunities and concludes the chapter.

4.4 Study Data

In this study, gridded daily precipitation forecasts are derived from ECMWF SEAS5 seasonal forecasting system (Johnson et al., 2019). This global climate model is composed of atmosphere, land, ocean, and sea-ice components. It uses IFS (integrated forecast system) atmosphere model cycle 43r1 with horizontal resolution of ~36 km and integrates HTESSEL (Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land) land surface model into IFS. It implements the NEMO (Nucleus for European Modelling of the Ocean) v3.4.1 model at 0.25° resolution. The atmosphere component of SEAS5 hindcasts is initialized from the ERA Interim, while the initial conditions for the land-surface component are provided from a more recent version of HTESSEL (cycle 43r1) that has been run offline for the hindcast period. The initial conditions for ocean and sea-ice components are generated by the historical reanalyses (ORAS5) from an operational ocean analysis system, OCEAN5. For ensemble generation schemes, perturbations are applied to atmosphere initial conditions, while perturbations to the assimilated observations and the surface forcing fields are used for ocean initial conditions. Furthermore, both stochastically perturbed physical tendency scheme and stochastic kinetic energy backscatter scheme are used to perturb atmospheric model in the generation of all ensemble members. Greenhouse gas (GHG) radiative forcing implemented in SEAS5 utilizes seasonally varying climatology from the Monitoring

Atmospheric Composition and Climate reanalysis (Inness et al., 2013). Such climatology is scaled to capture the trend in GHG emissions as used in CMIP5 GHGs during 1981–2000 and CMIP5 RCP 3-PD from 2000 onwards. A more detailed description of SEAS5 is available in Stockdale (2021) and Johnson et al. (2019).

SEAS5 hindcasts are initialized on the first day of every month for 1981-2016 and run for 7 months ahead. The ensemble generation scheme produces 25 ensemble members to represent forecast uncertainty. In this study, we will mainly present results for the 1-month lead forecasts of seasonal precipitation for 12 overlapping seasons from January-March (JFM) to December-February (DJF) in Australia (seasons will be abbreviated as the initial letters of three consecutive months hereafter). Forecasts with 1-month lead time represent the forecasts for a rolling season beginning in 1 month's time. Moreover, seasonal forecasts at all lead times (0-4-month lead) will be investigated for selected cases. As an example, consider the forecasts initialized on the 1st of January, the forecasts aggregated for JFM are with 0-month lead time while the forecasts aggregated for FMA are with 1-month lead time.

Monthly observations of precipitation are derived from the AWAP (Australian Water Availability Project) climate dataset (Jones et al., 2009), and then accumulated over three consecutive months to obtain seasonal precipitation data. The AWAP observations at 0.05° resolution are re-gridded using an area-conservative interpolation method to match the SEAS5 data at 0.4° resolution.

4.5 Methods

4.5.1 Model formulation

We introduce a more advanced version of the trend-aware forecast-calibration method, with adaptations and extensions necessary for post-processing precipitation forecasts. Precipitation variables pose unique challenges because seasonal quantities follow a skewed distribution that is bounded below at zero. Furthermore, precipitation is highly variable in space and time. Precipitation trends are inherently more uncertain and difficult to detect than temperature trends.

4.5.1.1 Data transformation

The forecast calibration model works under the assumption that predictor (raw ensemble forecast means y_1) and predictand (observations y_2) are jointly modelled as a continuous bivariate

normal distribution, and the marginal distributions of $y_i, i = 1,2$ are normal. Since precipitation amount is generally highly skewed, we employ a two-parameter log-sinh transformation scheme to facilitate modelling precipitation data using a normal distribution (Wang et al., 2012b),

$$y_i^{'} = \frac{1}{\lambda_i} \log[\sinh(\varepsilon_i + \lambda_i y_i)] \tag{4-1}$$

where ε_i and λ_i are transformation parameters.

When there are instances of zero precipitation amounts, for example in dry regions and seasons, zero values are treated as left-censored data (see further below).

4.5.1.2 Model specification

After data transformation, we calculate the anomalies $z_i(t)$, $t = 1,2,...,N$, from the trendline of the transformed variables $y_i^{'}$,

$$z_i(t) = y_i^{'}(t) - \alpha_i (t - t_m) \tag{4-2}$$

where t is the event time, α_i is a trend parameter, t_m is chosen to be approximately the time of the middle event in the analysis period. The choice of t_m will only affect the mean of the marginal distribution of z_i but not the final post-processing results. The joint distribution of the detrended transformed predictor z_1 and detrended transformed predictand z_2 is modelled as

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{4-3}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are mean vector and covariance matrix respectively.

We use a Bayesian approach to infer the parameter set $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha_1, \alpha_2\}$. Before inferring the model parameters, their Bayesian prior distributions need to be specified. For parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we use non-informative multivariate Jeffreys priors (Gelman et al., 2014). For trend parameters $\alpha_i, i = 1,2$, Chapter 3 compared two types of prior distribution. The first type was a non-informative uniform prior and the second type was an informative normal distribution prior, centred at zero and with an empirically determined variance. The resulting models were named BJP-t and BJP-ti, respectively. The informative prior in the BJP-ti model was to incorporate information on a broad range of trends observed across Australia and lead to more stable calibrations. The informative prior is of the form

$$p(\alpha_i) \propto N(0, m_i^2) \tag{4-4}$$

where m_i is empirically determined.

Chapter 3 determined m_i for temperature variables using all available data (all seasons of the year and all grid cells across the Australian continent). We suggest that this global approach leads to a weakly determined variance parameter because of vast hetergeneity in the dataset, a problem which may be more detrimental for precipitation applications due to the distinctness of precipitation regimes across tropical, arid and temperate regions, for example. We therefore propose a modification to determine m_i using local data. That is, by prescribing a neighbourhood in terms of grid cells and seasons, m_i is unqiuely determined on a cell-by-cell and season-by-season basis. Details of this new scheme are given in Section 4.8 Appendix.

4.5.1.3 Zero lower bound

Precipitaion has a natural zero lower bound. To allow the use of existing modelling approach, we treat the precipitation variables as left-censored, where zero values are treated as having unknown true values that are equal to or below zero (Wang and Robertson, 2011). When the variable $y_i^{'}(t)$ has a constant censoring threshold of $y_i^{'c}$, the corresponding detrended variable $z_i(t)$ should have a censoring threshold that varies with t,

$$z_i^c(t) = y_i^{'c} - \alpha_i(t - t_m) \tag{4-5}$$

4.5.1.4 Parameter inference

As a first step, the best set of transformation parameters ε_i and λ_i is estimated for each grid cell and for each variable separately by using the method of Bayesian maximum a posteriori (MAP) (Schepen et al., 2020c). These transformation parameters are fixed for the rest of the modelling process.

The parameter set $\boldsymbol{\theta}$, the unknown censored values and any missing values are inferred from $\mathbf{D} = \left\{[y_1^{'}(t), y_2^{'}(t)], t = 1, 2, \ldots, N\right\}$, which is a sequence of training data pairs. We use Bayesian inference with Gibbs sampling to successively obtain samples of the parameters or variables being inferred (Wang et al., 2019).

To begin with the model parameters, the posterior distribution of the model parameter set $\boldsymbol{\theta}$ is given by

$$p(\boldsymbol{\theta} \mid \mathbf{D}) \propto p(\boldsymbol{\theta}) p(\mathbf{D} \mid \boldsymbol{\theta}) \tag{4-6}$$

where $p(\boldsymbol{\theta})$ is the prior distribution for model parameters, and $p(\mathbf{D} \mid \boldsymbol{\theta})$ is the likelihood function. In applying Gibbs sampling, this overall posterior distribution is broken down into a series of conditional distributions for different subsets of parameters. In one iteration, each parameter is sampled from its conditional distribution in turn with the remaining parameters fixed to their current values. This sampling process continues until convergence, that is, the sampled parameters have the same distribution as sampled from the overall posterior distribution. The conditional distributions for the model parameters can be found in Appendix S3.

To deal with missing and censored data of the variables, values are sampled from the conditional distribution,

$$[z_i(t) \mid \cdot] = \mathrm{N}(\mu_i^*(t), \Sigma_{i,i}^*) \tag{4-7}$$

where

$$\Sigma_{i,i}^* = \sigma_i^2 - (\rho \sigma_1 \sigma_2)^2 / \sigma_{(i)}^2 \tag{4-8}$$

$$\mu_i^*(t) = \mu_i + \rho \sigma_1 \sigma_2 / \sigma_{(i)}^2 \times [z_{(i)}(t) - \mu_{(i)}] \tag{4-9}$$

where (i) denotes the index in $\{1, 2\}$ that is not i; σ_1, σ_2 and ρ are the parameters that constitute $\boldsymbol{\Sigma}$; $\mu_{(i)}$ is the parameter that constitutes $\boldsymbol{\mu}$. In case of censored values, the sampling of $z_i(t)$ is restricted to $z_i(t) \le z_i^c(t)$. We note that the frequency of the zero occurrence has an impact on the inference of trend parameters α_i. With a large number of zero values present, say over 20% of the available data, or consecutive zero occurrences present at the start or end of the training data, the magnitude of the inferred trend may be greater than the one inferred from nontreatment of censored values due to the restriction of the sampling of $z_i(t)$.

The sampling of the parameters and the sampling of the missing and censored data of the variables are carried out in sequence, and the whole process is repeated 30,000 times to generate inference chains. The first 5,000 iterations are discarded as burn-in because the early iterations may not be representative of the actual posterior distribution. The implementation and pseudo codes of the Gibbs sampling steps are elaborated in Appendix S3.

4.5.1.5 Model use for prediction

The established model can be used in predictive mode once all the parameter sets $\boldsymbol{\theta}$ are inferred. Given a new transformed predictor value $y_1^{'}(t^*)$, we obtain a calibrated ensemble forecast member $y_2^{'}(t^*)$ corresponding to each set of the model parameters. In each iteration, we treat the predictand as missing in value and use a Gibbs sampler to sample a new calibrated forecast value $z_2(t^*)$ from the conditional distribution of the predictand variable given by Eq. (4-7) – (4-9), and re-trend it to $y_2^{'}(t^*)$. When the variable $y_1^{'}(t^*)$ is of a censored value, the sampling range of the detrended variable $z_1(t^*)$ is restricted to $z_1(t^*) \le z_1^c(t)$. Again, the first 5,000 iterations are discarded as burn-in in predictive mode.

Besides the descriptions above, a pragmatic approach is also used to adjust extremely large $z_1(t^*)$ values that occur in prediction before sampling $z_2(t^*)$. These large, transformed values are considered unrealistic based on the marginal distribution of the transformed raw forecasts. In this study, we set the extreme threshold as 0.999 in the non-exceedance probability based on the marginal distribution of z_1 (Wang et al., 2019).

By back transforming each of sampled $y_2^{'}(t^*)$, and converting the negative value to zero, we derive and save an ensemble of 1,000 calibrated forecast values to represent forecast uncertainty. Readers are referred to Appendix S3 for the complete algorithm and pseudo codes of implementing the trend-aware method.

4.5.2 Forecast verification

In this study, we evaluate and compare the 1,000 ensemble members of the trend-aware BJP-ti calibrated forecasts with the 25 ensemble members of the raw forecasts and the 1,000 ensemble members of the BJP calibrated forecasts. The post-processing models are established separately for each grid cell, each season, and each lead time under a leave-one-year-out cross validation scheme, where each pair of data points within the year left out for validation is omitted from the data series and verified with the calibration model trained by the remaining data. This configuration is only appropriate to validate the anomaly component rather than the trend component. For the latter, the results from the cross validation are similar to that of model fitting, which is an inherent limitation when it comes to trend evaluation during the record period. The cross-validated forecasts over the hindcast period may contain artificial skill because the information from the future period is used to train the calibration model. However, such future information would not be available when real-time forecasts are calibrated for operational use.

Consequently, real-time forecasts may have lower skill than expected after calibration (Risbey et al., 2021). However, given short data records, other alternative validation methods, such as validating the forecasts at the start and end of the full evaluation period, are subject to larger sampling uncertainties because there are not sufficient events to train the model and cover the multi-decadal climate variability (Huang et al., 2019). As a result, such methods are not suitable for the model validation here (Shao et al., 2021a).

We assess the ensemble forecasts via trend testing methods and forecast verification tools. For trend analysis, we use the Theil-Sen approach (Sen, 1968; Theil, 1992) to calculate trend slopes in the observations, raw ensemble forecast medians, and calibrated ensemble forecast medians. This non-parametric trend detection technique does not require the estimated trend to be linear or the time series to conform to a Gaussian distribution (Kumar et al., 2013). In addition, this method can deal with positively skewed distributions and is not sensitive to extreme values (Sayemuzzaman and Jha, 2014).

Given a data sequence $y(t),\ t = 1, 2, ..., N$, the Theil-Sen slope is calculated as

$$\beta = \text{median}\left[\frac{y(b) - y(a)}{b - a}\right] \text{ for all } 1 \le a < b \le N \tag{4-10}$$

Here, Theil-Sen slope β is the median value of the slopes estimated from $N(N-1)/2$ combinations of two data points in the data sequence.

The statistical significance of the trend is checked by the non-parametric two-sided Mann-Kendall test (Kendall, 1975; Mann, 1945), which is a distribution independent method frequently applied for hydroclimatic trend tests (Kumar et al., 2013). Here, we check the trend significance for each grid cell across Australia and summarise the findings based on individual test results. Wilks (2016) pointed out that the global statistical significance results were often overinterpreted when the input data of a global hypothesis test were a collection of the results from individual local hypothesis tests. Although we do not have this problem for all the significance tests in this study, we advise the caution about the interpretation of the collective significance from multiple hypothesis tests.

Forecast skill is evaluated by the continuous ranked probability score (CRPS; Matheson and Winkler, 1976). For an individual event t , the CRPS is defined as

$$\mathrm{CRPS}(t) = \int \left\{ F(t,y) - H\left[y - y_{\mathrm{obs}}(t)\right] \right\}^2 dy \tag{4-11}$$

$$H\left[y - y_{\mathrm{obs}}(t)\right] = \begin{cases} 0 & \text{if} \quad y < y_{\mathrm{obs}}(t) \\ 1 & \text{if} \quad y \geq y_{\mathrm{obs}}(t) \end{cases} \tag{4-12}$$

where H is the Heaviside step function; for an event t, $F(t,y)$ is the cumulative distribution function (CDF) of the ensemble forecasts, and $y_{\mathrm{obs}}(t)$ is the observed value. For each grid cell, the CRPS skill score is calculated as the comparison between the averaged CRPS of the target forecasts and the averaged CRPS of the reference forecasts across total sets of events, given as

$$\mathrm{CRPS}_{\text{skill score}} = \frac{\overline{\mathrm{CRPS}}_{\mathrm{ref}} - \overline{\mathrm{CRPS}}}{\overline{\mathrm{CRPS}}_{\mathrm{ref}}} \times 100\left(\text{unit: \%}\right) \tag{4-13}$$

Here, reference forecasts are leave-one-year-out cross-validated climatology ensemble forecasts generated from the BJP model. A higher value of the resulting CRPS skill score indicates more skillful forecasts.

Forecast bias is measured by the percentage bias, which is the relative error between ensemble forecast means and observations,

$$\mathrm{Bias} = \frac{\sum_{t=1}^{N} \left[\bar{y}(t) - y_{\mathrm{obs}} \right]}{\sum_{t=1}^{N} y_{\mathrm{obs}}} \times 100(\text{unit: \%}) \tag{4-14}$$

where $\bar{y}(t)$ is the ensemble forecast means and y_{obs} is the observation for an event t in each grid cell. The bias value equal to 0 indicates that the ensemble forecast means perfectly correspond to the observations.

Reliability is quantified by a PIT-based score (Renard et al., 2010), measuring the tendency of the PIT (probability integral transform; Gneiting et al., 2007) values to deviate from the corresponding theoretical standard quantiles. Theoretically, a reliable forecasting system has the collection of the PIT values that follow a standard uniform distribution. For each grid cell, the PIT value π_t for an event t is given as

$$\pi_t = F\left[t, y = y_{\mathrm{obs}}(t)\right] \tag{4-15}$$

where $F(t, y)$ is the ensemble forecast CDF and $y_{\text{obs}}(t)$ is the corresponding observations. When $y_{\text{obs}}(t)$ equals to zero, we randomly sample a pseudo-value from a uniform distribution within the range of $[0, \pi_t]$ and replace the zero values (Wang and Robertson, 2011). The final PIT score is calculated as

$$\text{PIT score} = 1.0 - \frac{2}{T}\sum_{t=1}^{N}\left|\pi_{(t)} - \frac{t}{T+1}\right| \tag{4-16}$$

where $\pi_{(t)}$ is the t^{th} ranked PIT value in an increasing order. The PIT score ranges from 0 (worst reliability) to 1 (perfect reliability).

Sharpness is checked by the sharpness ratio, defined as the ratio of the average interval widths between target ensemble forecasts and BJP calibrated ensemble forecasts (as the baseline). Here, for each grid cell, the average interval width is calculated as the average of the central 50% [0.25, 0.75], and 90% [0.05, 0.95] inter-quantile intervals for all individual events (Gneiting et al., 2007). If the resulting ratio is lower than 1, target ensemble forecasts are interpreted as sharper than the BJP calibrated forecasts.

4.6 Result

4.6.1 Trend of observations and forecasts

The geographic distributions of precipitation trends in observations and model forecasts for the period of FMA 1981-JFM 2017 are shown in Figure 4-1. The AWAP observational precipitation dataset is spatially interpolated from the rainfall gauging network across Australia. In central Australia, the network is extremely sparse, resulting in missing daily data in several clusters (see Figure 2 in Brocca et al., 2016). Although this study uses the seasonal precipitation data aggregated from monthly precipitation product, which has been recalibrated with improved data quality. To be prudent, we still focus on interpreting the results in data-rich regions hereafter.

Decadal trends of the observed precipitation are evident across Australia. Precipitation increases at over 20 mm decade^{-1} during the warmer seasons (OND to JFM) across many parts of the continent (first column in Figure 4-1). Strong increasing trends are found in northern Australia during the northern wet seasons (i.e., October to April). Meanwhile, drying trends are dominant during the southern wet seasons (i.e., April to November) in south-western and south-eastern Australia.

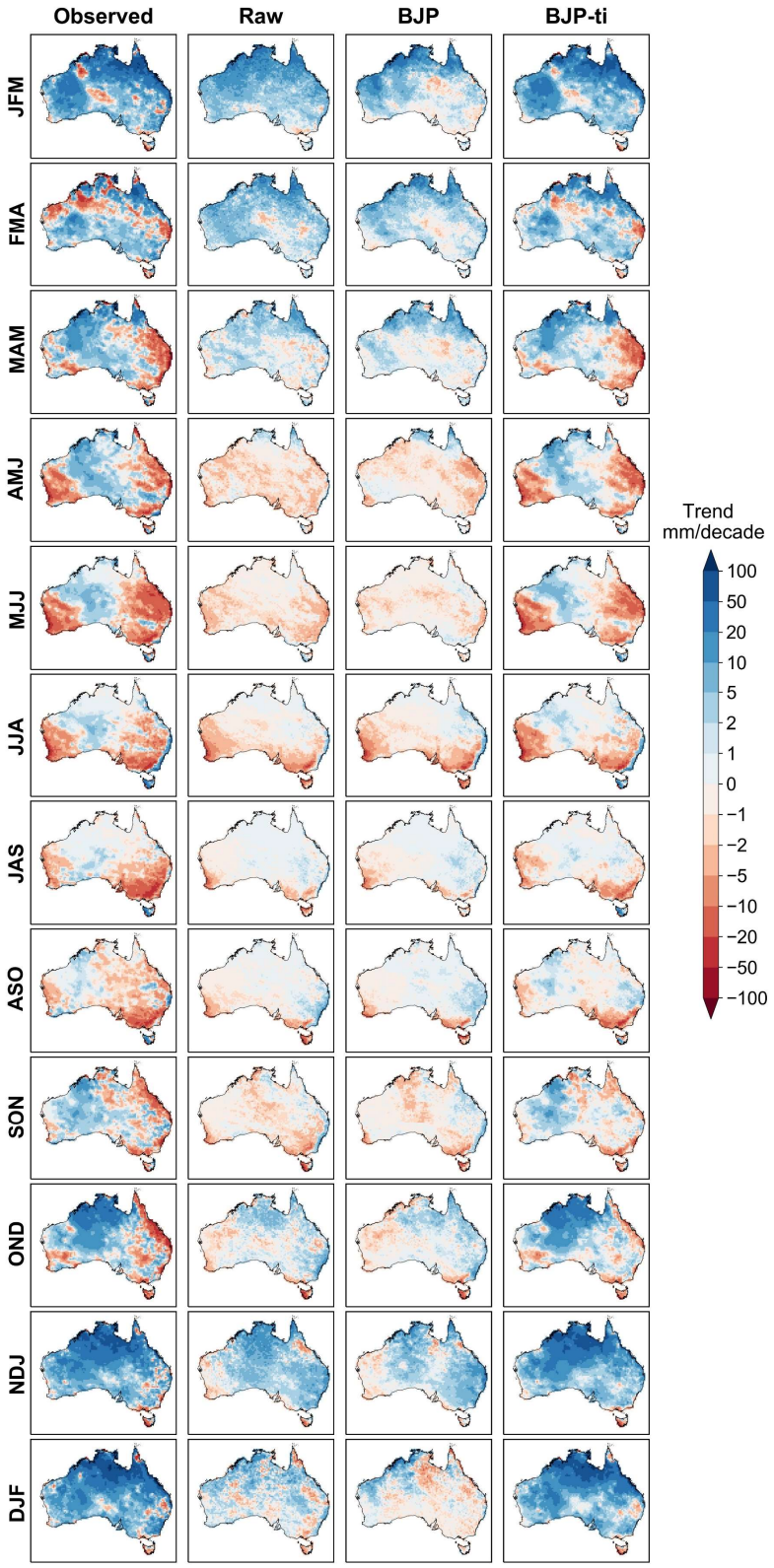

Figure 4-1: Decadal Theil-Sen's slopes for observations, raw, BJP, and BJP-ti calibrated ensemble forecast medians of seasonal precipitation for 12 overlapping seasons with 1-month lead time from FMA 1981 to JFM 2017.

Raw forecasts reproduce the observed precipitation trends over some parts of Australia (second column in Figure 4-1), such as the wetter conditions in the north for the warmer seasons and the drier conditions in the southeast for most seasons. However, these forecast trends are generally weaker than the observed ones. In many other parts, such as western Australia in FMA, trends in raw forecasts significantly mismatch the trends in observations. Likewise, the BJP calibrated forecasts also fail to capture the observed trends in some regions (third column in Figure 4-1). Visually, the spatial patterns of the raw and BJP calibrated forecasts show remarkable consistency in most regions, but different trends are still discernible elsewhere.

By using the trend-aware post-processing model, BJP-ti, the resulting calibrated forecasts are shown to reproduce the observed trends more accurately than raw and BJP calibrated forecasts (Fourth column in Figure 4-1). Strong trends are exactly reproduced in, for example, the wetter conditions in the north from OND to JFM, and the drier conditions in the southwest from AMJ to JJA.

4.6.2 Overall performance of the forecasts

Results of the CRPS skill score maps for the BJP-ti calibrated forecasts are shown in Figure 4-3 (left plot). Skillful forecasts dominate large areas from ASO to NDJ, particularly northern Australia while climatology-like forecasts (i.e., score ranging from -5% to 5% in white) are widespread in some seasons, such as MJJ and DJF. Interestingly, for individual cells when the observed trend is statistically significant at 10% significance level (Figure 4-2), the skill of the BJP-ti forecasts is mostly no worse than climatology. Note that in this study, the collection of the individual test results for the significance of trend is not indicative of the regional significance in Figure 4-2 (Wilks, 2016). For example, it is not valid to say statistically significant trends at 10% significance level are prevailing in northern Australia from OND to DJF.

The skill score difference between BJP-ti and BJP calibrated forecasts is presented in the right plot of Figure 4-3, indicating how forecast skill changes by embedding observed trends into the forecasts through post-processing. Noticeable skill improvement (in darker blue) of the BJP-ti calibrated forecasts predominately occurs in the trend-significant clusters where the BJP calibrated forecasts do not properly represent the observed trends, such as in north-western Australia in SON. We also find that rectifying the trend direction or counteracting large trend difference rarely imparts forecast skill where observed trends are non-significant, such as in parts of eastern Australia in SON and OND. Overall, the BJP-ti calibration leads to detectable skill gain

with the score increase larger than 5% in many trend-significant regions, for example, parts of northern Australia from OND to DJF, and leads to slight skill loss with the score decrease less than 5% relative to the BJP calibration, such as northern Australia in FMA and MAM. Using the BJP-ti calibration also turns the negative skills to positive skills in some regions, such as part of western Australia in AMJ and MJJ, and part of northern Australia in NDJ and DJF (Not shown). We also employ a bootstrap procedure as detailed in Schepen et al. (2016) and Chapter 2 to test whether the BJP-ti calibration significantly improves or worsens the CRPS skill score over the BJP calibrated forecasts at 5% level. Significant skill improvement is largely found in the regions with the score increase larger than 5%, while fewer cases have significant skill worsening (Supplementary Figure S3-1). These findings indicate that the underlying decadal trend is an important contributor to the interannual variability for precipitation, especially in the regions with significant trends, where skill improvement is pronounced by properly incorporating the underlying historical trend into the calibrated forecasts. Other sources also contribute to the high forecast skill, such as the good modelling of the teleconnection between large-scale climate drivers and seasonal precipitation (Wang et al., 2019).

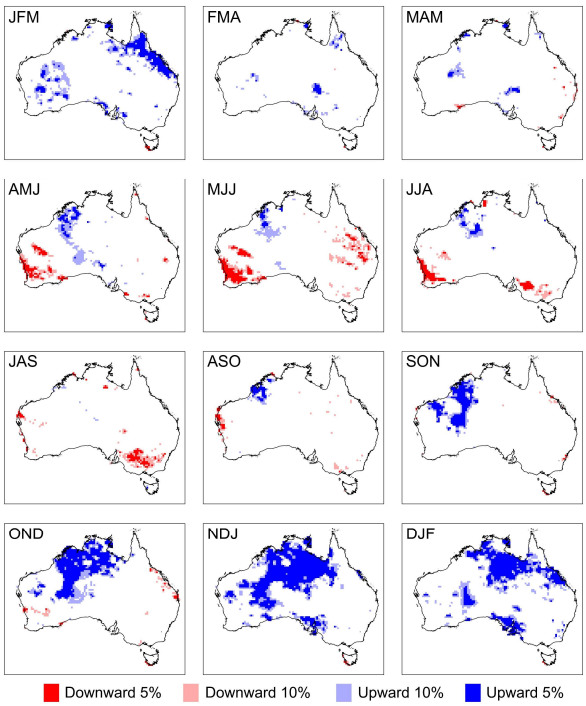

Figure 4-2: Statistical significance of the trend in observations at 5% and 10% significance level for seasonal precipitation using Mann-Kendall test.

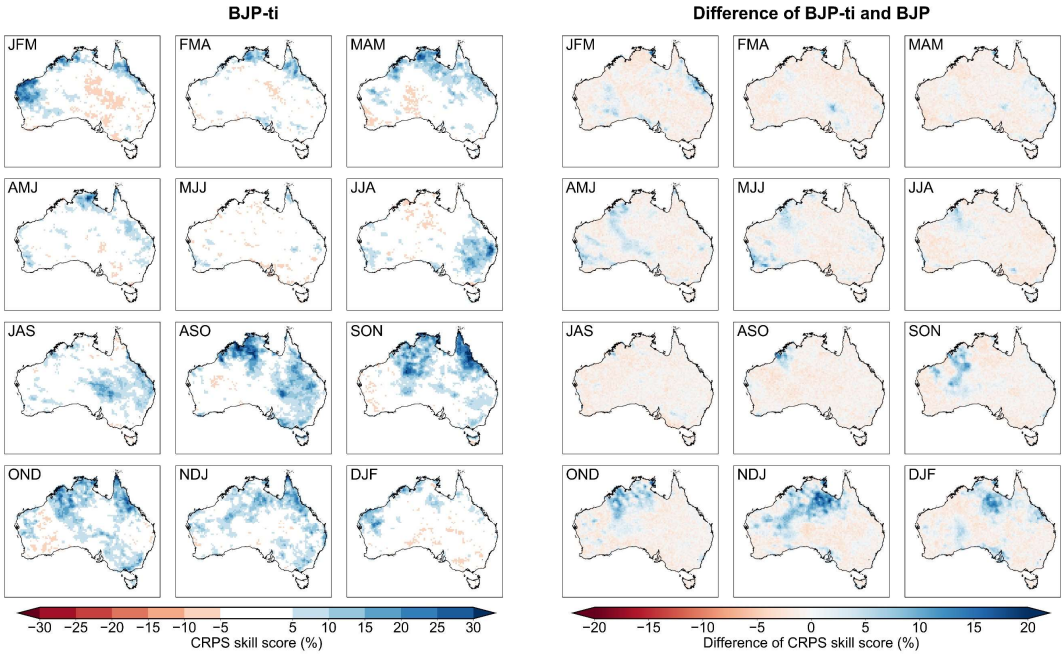

Figure 4-3: CRPS skill score of the BJP-ti calibrated ensemble forecasts (left) and the score difference between the BJP-ti and BJP calibrated ensemble forecasts (right) of seasonal precipitation at 1-month lead time.

We then compare the overall performance for raw, BJP and BJP-ti calibrated ensemble forecasts in terms of the trend difference between the model forecast medians and observations (mm/decade), CRPS skill score (%), percentage bias (%), PIT score, and sharpness ratio. We pool the results of the forecast verification metrics from all the grid cells and all the seasons to plot the proportion of the cells that do not exceed a score value (Figure 4-4). Consistent with the trend results shown in Figure 4-1, there exist substantial trend discrepancies between the raw forecasts, the BJP calibrated forecasts and the observations (Figure 4-4a). For comparison, minor trend difference is seen between the BJP-ti calibrated forecasts and observations, reiterating that the BJP-ti calibration is effective at embedding the observed trends into the resulting forecasts.

For forecast skill (Figure 4-4b), around half of the raw forecasts have negative skill scores while the BJP and BJP-ti calibrated forecasts rarely have scores lower than -5%. Post-processing also increases the proportion of the cases with positive skills. The BJP calibration produces slightly fewer cases with negative skills while the BJP-ti and BJP calibrated forecasts are equally skillful in terms of the positive skill score. For forecast bias (Figure 4-4c), raw forecasts are largely biased, with a higher proportion to be negatively biased. Both post-processing methods effectively reduce the biases in forecast means, where the cumulative lines are closer to the zero-vertical line. The

BJP model slightly outperforms the BJP-ti model in removing the biases in forecast means. After post-processing with BJP and BJP-ti model, there remain more positive biases than negative ones. As discussed in Schepen et al. (2020c), in very dry regions, small biases are shown as large percentage biases. Moreover, the calibration models introduce parameter uncertainty to the resulting forecasts, which may lead to some extreme values and give rise to higher forecast means in these dry regions. Apart from the visual comparison, we also use the Bootstrap method to determine whether using the BJP-ti calibration could result in a greater number of cases with negative skills, and larger magnitude of the bias than the BJP calibration, both at 5% significance level. Results suggest that overall, the BJP-ti model does not lead to significantly more cases with negative skills but does lead to significantly larger magnitude of the biases than the BJP model.

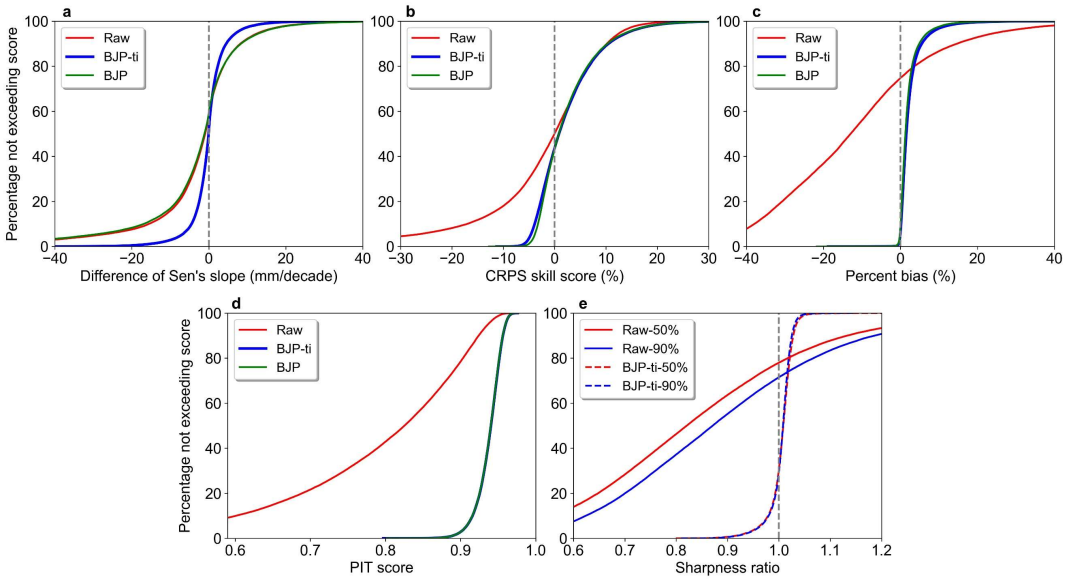

Figure 4-4: Non-exceedance plot comparing the overall performance of the raw, BJP and BJP-ti calibrated ensemble forecasts at 1-month lead time. Note: the blue line is behind the green line for the PIT score plot.

For reliability (Figure 4-4d) and sharpness (Figure 4-4e), although raw ensemble forecasts tend to have narrower widths of the inter-quantile intervals than the BJP calibrated forecasts, they are generally not reliable indicated by the low values of the PIT score. This means that raw forecasts could not accurately estimate the likelihood of the observed events and are thus not informative for forecast users. In contrast, the BJP and BJP-ti calibrated forecasts are comparably reliable (Figure 4-4d), while the BJP-ti calibrated forecasts are slightly less sharp (Figure 4-4e) than the BJP forecasts due to the introduction of additional trend parameters.

We also explore the forecast performance in the region with statistically significant observed trends at 10% level (Supplementary Figure S3-2). Again, the BJP-ti calibrated forecasts are found to represent the observed trend more accurately than both raw and BJP calibrated forecasts. The BJP-ti calibration leads to more skillful forecasts than raw and BJP calibrated forecasts for the trend-significant cases. This finding is consistent with the statements made from Figure 4-2 and Figure 4-3, demonstrating that the BJP-ti calibration leads to apparent skill gains the trend-significant regions. For forecast bias and reliability, the BJP-ti and BJP model are comparably effective at removing biases and making calibrated forecasts more reliable than raw forecasts. As for sharpness, the BJP-ti calibrated forecasts appear to have narrower widths of the inter-quantile intervals than the BJP calibrated forecasts, indicating that the BJP-ti calibration could produce ensemble forecasts with maximal sharpness and high reliability in the regions with statistically significant observed trends.

4.6.3 Forecast performance of selected grid cells

To explore the performance of the BJP-ti model for individual cells and at longer forecast lead times, we select four grid cells (A-D in Figure 4-5) for detailed evaluations. The selection criteria include: 1) observed trends are statistically significant at 5% significance level using Mann-Kendall test over 1981-2016; 2) located in the regions of high rain-gauge density but in different climate zones; 3) the BJP-ti calibration improves the skill of 1-month ahead forecasts in one of the four main seasons.

To investigate how raw ensemble forecasts represent observed precipitation trends, we examine the trends of raw forecast means and each of 25 ensemble members with 1-month lead time for winter (JJA) of cell A, spring (SON) of cell B, autumn (MAM) of cell C and summer (DJF) of cell D (Supplementary Figure S3-3). For a straightforward comparison of trend magnitudes, the trendlines are shown after subtracting the temporal mean of each line, so that all the trendlines meet in the midpoint of the horizontal axis. The trendlines of raw ensemble forecast means generally fail to follow the magnitude of the observed trends, except for cell A. Regarding the trendlines of 25-member raw ensemble forecasts, only a few follow the direction and magnitude of the observed trendlines in cell B and C, while none of the member trendlines appears to be aligned with the magnitude of the observed trendlines in cell A and D.

Again, for these four cases, we plot the BJP and BJP-ti calibrated ensemble forecast quantiles and linear trendlines estimated from the Theil-Sen's slopes and the corresponding intercepts (Figure

4-6). The historical data for cell A and B include occurrences of zero precipitation. As shown in Figure 4-6A and B (right plots), the BJP-ti calibration is capable of concurrently modelling zero and non-zero values of precipitation. Compared to raw and BJP calibrated ensemble forecasts that do not explicitly follow the trend signal (left plots), the inter-quantile ranges ([0.25 0.75] quantile in deep blue and [0.1 0.9] quantile in light blue) of the BJP-ti calibrated forecasts roughly follow the underlying trend (right plots), with the trendline of the forecast medians close to or overlapping the trendline of the observations. This is also evident for cells C and D without zero precipitation values (right plots Figure 4-6C and D), where the BJP-ti calibrated ensemble forecasts appear to model the interannual variability more explicitly than the BJP calibrated ensemble forecasts in these two cases. Overall, the BJP-ti model is capable of properly incorporating the climate trend into the forecasts whilst improving the prediction of the interannual variability.

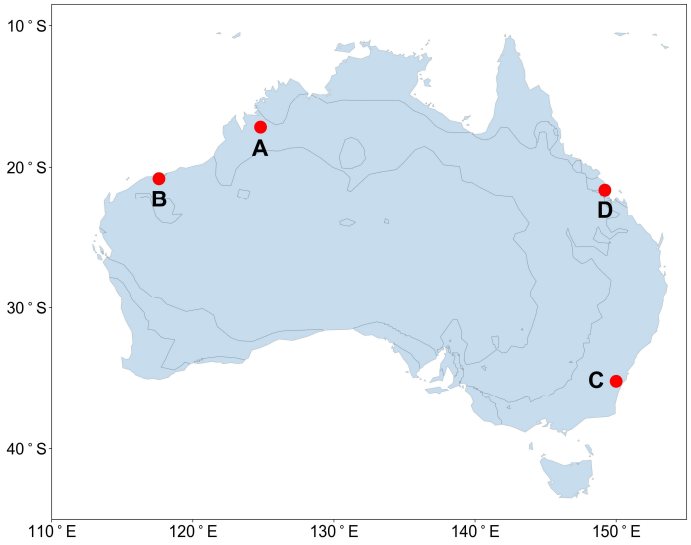

Figure 4-5: Location of selected cases. Contours in grey line show the boundary of major climate zones in Australia (Peel et al., 2007).

We also examine the performance of the BJP-ti calibrated forecasts for all the forecast lead times. For the four target cells, we apply the BJP-ti model to each of the 12 overlapping seasons and each of the forecast lead times separately and explore the forecast skill (Figure 4-7) and the trend slope. In general, the BJP-ti calibrated forecasts with zero- and one-month lead time are more skillful. Most of the resulting forecasts in cells D and A are skillful at all lead times. Negative skill scores are no lower than -10% in all cells, indicating that the BJP-ti calibration is highly effective in producing skillful, and at least climatology-like forecasts for longer lead times.

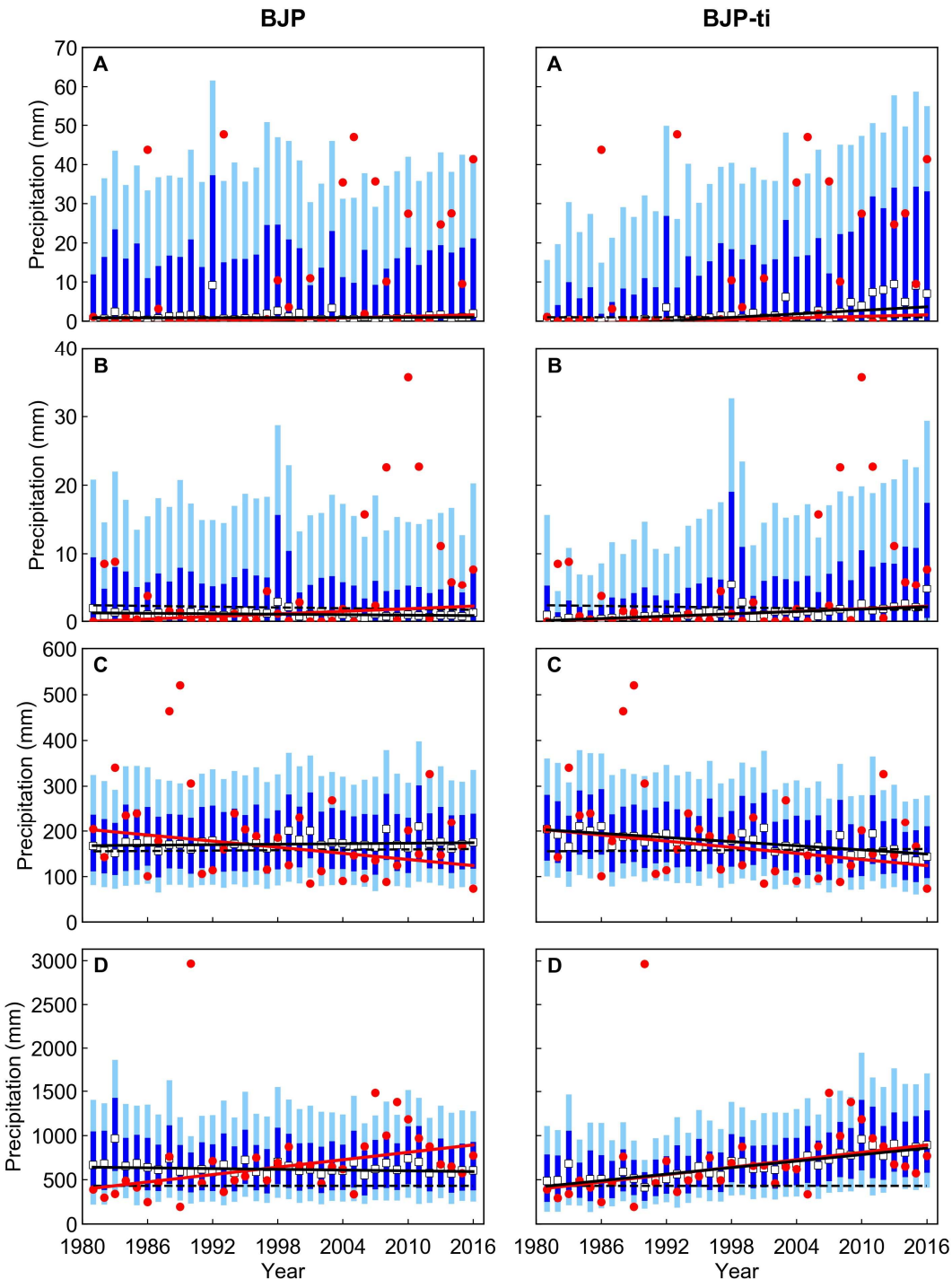

Figure 4-6: Forecast quantile plots for selected cells with 1-month lead time. Red dots are observed data. White squares are calibrated forecast median values. Dashed black lines are trendlines for raw forecast medians. Black lines are trendlines for calibrated forecast medians. Red lines are trendlines for observed data. Light blue strips are [0.1, 0.9] quantile forecasts. Deep blue strips are [0.25, 0.75] quantile forecasts.

Again, trend evaluation demonstrates the ability of the BJP-ti model to properly represent the underlying trend in the calibrated forecasts. As shown in Figure 4-8, trend difference between the BJP-ti calibrated forecasts and observations is generally smaller than 20 mm/decade. In a few other cases, such as NDJ in cell D, although trend difference is still large at all lead times, the trends in the BJP-ti calibrated forecasts are much closer to the observations than the BJP calibrated forecasts that show the trend difference larger than 60 mm (Supplementary Figure S3-4).

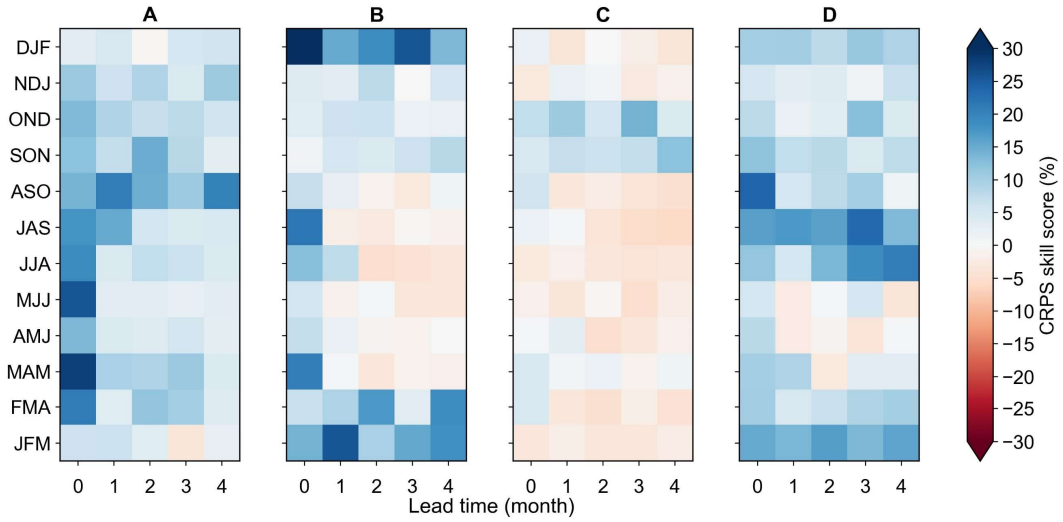

Figure 4-7: CRPS skill score of the BJP-ti calibrated ensemble forecasts of seasonal precipitation for all lead times. Locations of the grid cells are shown in Figure 4-5.

4.7 Discussion and Conclusion

Retrospective forecasts from global climate models (GCM) have often shown inability to reproduce historical climate trends, making the forecasts less informative and undermining user confidence. In this study, we aim to resolve the trend mismatch problem between GCM re-forecasts of seasonal precipitation and observations. The trend-aware forecast post-processing method introduced in the previous work has shown effectiveness for post-processing seasonal temperature forecasts. However, it is not directly applicable to precipitation forecasts due to the unique features of seasonal precipitation amounts, such as following a positive skewed distribution, having zero occurrences, and being more variable and uncertain than temperatures. To overcome these challenges, we make significant improvements to the algorithm for post-processing precipitation forecasts.

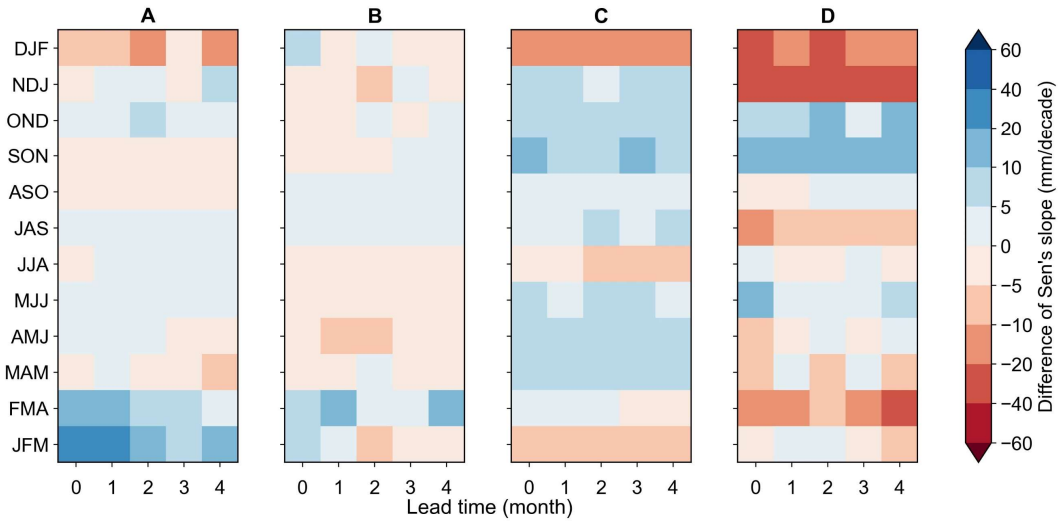

Figure 4-8: Trend difference between the BJP-ti calibrated ensemble forecast medians and observations of seasonal precipitation for all lead times. Locations of the grid cells are shown in Figure 4-5.

This study adopts a refined prior specification scheme for trend parameters. Chapter 3 employs a single prior distribution for trend parameters across all grid cells and all seasons. We found that using this approach could result in weaker trends in the calibrated precipitation forecasts, particularly in the regions where observed trends are strong (not shown). Here we determine the prior distribution cell-by-cell and season-by-season based on the neighbourhood information, which explicitly accounts for the local precipitation regimes on a spatial continental scale. This prior selection approach is applicable for other continental or large-scale studies where there are sufficient data available to specify the prior. When the trend-aware method is to be applied to a site study or a small spatial scale, the settings of informative priors elaborated in this study may not be valid. In these cases, it is more suitable to use the trend-aware BJP-t model (Shao et al., 2021a) with non-informative trend priors or the previous prior selection method that applies a fixed trend prior across all the cells and/or seasons.

To comprehensively assess the trend in precipitation variable, we employ the Theil-Sen approach to quantify the trend slope and the Mann-Kendall test to check the statistical significance of the trend. Referring to Figure 4-1 and Figure 4-2, strong but not statistically significant observed trends are detected in some regions, such as eastern Australia from MAM to MJJ. In fact, the monotonic trend with great magnitude is not necessarily recognised as statistically significant in the Mann Kendall test (Wang et al., 2020). Apart from trend magnitude, other factors can also affect the power of the method. The effectiveness of the method is reduced with limited length of

the data series, large data variance and the existence of the positive autocorrelation in the time series.

In this Chapter, we validate the improved trend-aware method on ECMWF SEAS5 forecasts of seasonal precipitation for Australia over the 36-year period 1981-2016 and compare the trend-aware calibrated forecasts against the raw forecasts and the BJP calibrated forecasts that do not have observed trend information embedded. Results reveal that the trend-aware calibrated forecasts properly capture the observed trends and reproduce the magnitude of strong trends when raw and BJP calibrated forecasts fail to do so. The trend-aware method appears to greatly enhance the forecast skill primarily over the regions where the observed trend is statistically significant, while the method slightly lowers forecast skill elsewhere. Overall, the resulting forecasts substantially outperform raw forecasts and perform comparable to the BJP calibrated forecasts in terms of bias, skill, and reliability. With the use of the method, skillful or at least climatology-like forecasts are produced at all lead times in selected cases.

Our trend-aware post-processing method has shown promise for forecasting seasonal precipitation. The method will be further improved or extended in the subsequent study. First, alternative settings of prior specification scheme for trend parameters are available for testing in future applications. For example, cells located in the same climate zone could share the same prior value for a season to reduce the computational costs. Other forms of the prior distribution are also worth investigating to show local neighbourhood behaviours more explicitly. Other physical and localised properties of the variable, such as mountainous terrain, may also be considered in the prior selection. Second, this work, along with the method developed for post-processing seasonal temperature forecasts in Chapter 2 and Chapter 3, are limited to the single target variable configuration, which models the relationship between a predictor and a predictand climate variable using a continuous bivariate normal distribution. In fact, their predecessor, the BJP method, allows for simultaneously calibrating multiple climate variables in high-dimensional settings (Schepen et al., 2020c; Wang et al., 2019), where any number of the predictors and predictands are jointly modelled. In this respect, the trend-aware post-processing is also extendable for the multivariate forecast calibration of hydrometeorological variables, such as forecasting seasonal streamflow for multiple sites and for months ahead (Wang and Robertson, 2011), and potentially using the calibrated forecasts (e.g. temperatures) to forecast other meteorological variables (e.g. precipitation). Third, future work will test the efficiency of the

trend-aware method for post-processing sub-seasonal climate forecasts and will improve the algorithm or develop a more robust model calibration scheme where necessary.

4.8 Appendix

In Eq. (4-4), m_i is set as $m_i = \delta_i \times m_i^{''}$, where δ_i is the MAP estimate of the standard deviation of $y_i^{'}$ derived at the transformation step. By removing from m_i the scaling factor δ_i, which is affected by the transformation parameter values, $m_i^{''}$ should be more uniform spatially and easier to specify.

To select $m_i^{''}$, we firstly run the trend-aware model with non-informative prior to determine the trend slope from the transformed data only. Specificially, we run the BJP-t model without cross validation setup. Our preliminary analysis shows the final prior values derived from cross validation and non-cross validation experiments do not differ much, so we simply employ the latter setup to reduce computation costs. We record the values of trend parameters α_i sampled in the parameter inference mode. Then we calculate the median value of α_i to represent the trend slope of $y_i^{'}$, and divide the slope by δ_i. This trend slope/δ_i value is calculated and recorded for each grid cell, for each season, for raw forecasts and for observations separately. After achiving all trend slope/δ_i values, the prior parameter $m_i^{''}$ for each case is specified via a temporal and spatial window. That is, for the case of interest, we choose the values of trend slope/δ_i from close-by 49 cells (i.e., cells located within a 7-cell by 7-cell square) and from adjacent 3 seasons (last, this, and next season). In total, the trend slope/δ_i values from 147 cells (49×3) are pooled together to determine the specific prior parameter $m_i^{''}$, estimated as the 75th percentile of the pooled absolute trend slope/δ_i values. Here, when we pool the cells, only land values within the national boundary are considered, which means for the cells along the coastline, there are fewer than 147 cells pooled for specifying the prior parameters. We note that it is also valid to pool cells within a different spatial window, a 5-cell by 5-cell square for example, and the results do not differ much in both settings. The 75th percentile number is specified as a compromise to slightly moderate the inferred trend. In our work, the rationale of applying the zero-centred normal informative prior is that trends inferred from a limited training period are subject to sampling errors and thus may not properly represent the underlying observed trends. If $m_i^{''}$ is determined from a higher percentile (e.g. 90th percentile), the prior may not be useful to moderate the trend because there is a higher chance for the trends to fall in a wider value range. Consequently, the inferred trends are likely to be accurately aligned with the trends of the observations. In contrast,

selecting m_i^* from a relatively lower percentile (e.g. 60th percentile) may overly constrain the inference of the trends, and make the inferred trend closer to zero.

Chapter 5 Introducing Long-term Trends into Sub-seasonal Temperature Forecasts

5.1 Preamble

Target variables in the last three chapters were GCM seasonal climate forecasts, and the trend-aware forecast post-processing method was shown effective for improving forecast skill and reliability while properly embedding the historical trend into seasonal temperature and precipitation forecasts. In recent years, there is a growing interest in understanding and utilising sub-seasonal climate forecasts – a middle ground between short-term weather and long-range seasonal forecasts – for decision-making and long-term planning. Operational GCMs have been implemented by climate centres to produce sub-seasonal climate forecasts. These sub-seasonal forecast models share similar configurations with the GCMs for seasonal forecasting, but whether the trend mismatch issue exists in GCM sub-seasonal climate forecasts have not been explored. The trend-aware forecast post-processing method developed for seasonal forecasts has the potential to be adapted to post-process sub-seasonal forecasts and introduce long-term climate trends into the forecasts.

Accordingly, Chapter 5 answers RQ4: Are sub-seasonal temperature forecasts capable of reproducing historical trend information? How can the trend-aware model be adapted to post-process sub-seasonal forecasts? In this chapter, sub-seasonal minimum and maximum temperature forecasts are extracted from the ECMWF extended-range forecasting system. As a preview, I compare the trend in weekly averaged raw temperature forecasts and paired observations, relative to day-of-year climatology, across Australia over a 20-year retrospective forecast period (2000-2019) and confirm that the trend mismatch issue also occurred in sub-seasonal forecasts. Subsequently, I extend the trend-aware method to calibrate and correct the trend in sub-seasonal forecasts. Since trends estimated from 20-year data records are subject to large sampling variability, I formulate the trend-aware model to embed a longer 30-year climate trend into the forecasts in this work.

The content of this chapter has been accepted for publication in International Journal of Climatology (Impact Factor 3.928). The paper title is ‘Introducing Long-term Trends into Sub-seasonal Temperature Forecasts through Trend-aware Post-processing’, and the authorship is Shao, Y., Wang, Q. J., Schepen, A., and Ryu, D.

5.2 Abstract

Skillful sub-seasonal forecasts are crucial for issuing early warnings of extreme weather events, such as heatwaves and floods. Operational sub-seasonal climate forecasts are often produced by global climate models not dissimilar to seasonal forecast models, which typically fail to reproduce observed temperature trends. In this study, we identify that the same issue exists in the sub-seasonal forecasting system. Subsequently, we adapt a trend-aware forecast post-processing method, previously developed for seasonal forecasts, to calibrate and correct the trend in sub-seasonal forecasts. We modify the method to embed 30-year climate trends into the calibrated forecasts even when the available hindcast period is shorter. The use of 30-year trends is to robustly represent long-term climate changes and overcome the problem that trends inferred from a shorter period may be subject to large sampling variability. Calibration is applied to 20-year ECMWF sub-seasonal forecasts and AWAP observations of Australian minimum and maximum temperatures with forecast horizons of up to 4 weeks. Relative to day-of-year climatology, raw week-1 forecasts reproduce temperature trends of the 20-year observations in many regions while raw week-4 forecasts do not exhibit the 20-year observed trends. After trend-aware post-processing, the behaviour of forecast trends is related to raw forecast skill regarding accuracy. Calibrated week-1 forecasts show apparent trends consistent with the 20-year observations, as the calibration transfers forecast skill and embeds the 20-year observed trends into the forecasts when raw forecasts are inherently skillful. In contrast, calibrated week-4 forecasts exhibit the 30-year observed trends, as the calibration reverts the forecasts to the 30-year observed climatology with trends when raw forecasts have little skill. For both weeks, the trend-aware calibrated forecasts are more reliable, and as skillful as or more skillful than raw forecasts. The extended trend-aware method can be applied to deliver high-quality sub-seasonal forecasts and support decision-making in a changing climate.

5.3 Introduction

Sub-seasonal climate forecasts are attracting growing interest among climate-sensitive sectors because many decisions are made based on future climate conditions from two weeks up to a season ahead (Vitart and Robertson, 2019). Extreme and high-impact meteorological events, such as floods and heat waves, are foreseeable through skillful and reliable sub-seasonal climate forecasts, which are crucial for issuing proactive alerts to vulnerable communities (Merryfield et al., 2020).

In recent years, global climate models (GCMs) have rapidly advanced to output sub-seasonal forecasts of a wide array of climate variables. Operational GCMs could be classified into two types. The first type of GCMs are specifically configured for sub-seasonal climate modelling, such as CFSv2 run by the National Centres for Environmental Prediction (NCEP) for sub-seasonal forecasting (Saha et al., 2014), and the extended-range forecasting system operated by the European Centre for Medium-Range Weather Forecasts (ECMWF, 2021). The second type of GCMs are essentially implemented for seasonal forecasting but are frequently initialised to produce multiple outputs in a calendar month, such as multi-week forecasting systems, POAMA multi-week (M2.4) system (Hudson et al., 2013; Marshall et al., 2014) and its successor ACCESS-S1 (Hudson et al., 2017; Hudson et al., 2018), operated by the Australian Bureau of Meteorology. Even with different configurations, all these sub-seasonal forecasting systems aim to explicitly simulate physical processes, accurately predict large-scale teleconnection patterns, and eventually deliver high-quality sub-seasonal forecasts for practical applications.

Despite recent enhancements, GCMs developed for both sub-seasonal and seasonal forecasting have been encountering some common technical challenges. For example, their model physics is only approximately represented, model components are not accurately initialised, and ensemble generation techniques do not fully account for the uncertainty in the initial conditions. These modelling issues result in model drifts and biases, over-confident ensemble spreads of the forecasts, and degraded forecast skill (Merryfield et al., 2020). As reported in literature, forecast skill horizon for climate variables typically limits to the first 2 weeks (Schepen et al., 2018; Scheuerer et al., 2020; Wang and Robertson, 2019). Another issue already identified for GCM seasonal forecasting systems is their inability to reproduce historical trend information (Huang et al., 2019; Krakauer, 2019; Shao et al., 2021a). Little attention has been paid to whether the same trend issue exists in GCM sub-seasonal forecasting systems. This study will seek to investigate this question.

Given long-standing modelling issues, post-processing is crucial for overcoming these problems while yielding well-calibrated ensemble sub-seasonal climate forecasts. Many studies have formulated statistical post-processing methods for sub-seasonal forecasts with the overarching objective of improving skill and reliability at different spatiotemporal scales (Li et al., 2020; Peng et al., 2020; Schepen et al., 2018; Scheuerer et al., 2020; Vigaud et al., 2020; T. Zhao et al., 2019a). These existing methods have greatly enhanced forecast performance, but they rarely aim to eliminate trend discrepancy between model forecasts and observations. Incorporating the

observed trend information into the post-processed sub-seasonal forecasts has the potential to make the resulting forecasts explicitly reflect the changing climate and more valuable to forecast users.

In Chapters 2-4, a robust trend-aware post-processing methodology was proposed for resolving the trend mismatch issue in seasonal climate forecasts. This method has been shown effective for embedding observed trend information into the forecasts while removing model biases and improving forecast skill and reliability. In this chapter, we extend the trend-aware methodology for the applications on sub-seasonal timescales.

Careful consideration is required for formulating the calibration method to post-process sub-seasonal climate forecasts. Many operational GCM sub-seasonal forecasting systems have relatively short re-forecast periods, say 20 years (Vitart et al., 2017). Apparent trends inferred from such limited periods are subject to large and unrealistic sampling errors (Hartmann et al., 2013). Consequently, the fitted trends may be more representative of sampling variability rather than the underlying trends caused by climate change. Here, we address this challenge by detecting trends from longer observational records, say 30 years' data, and introducing this long-term trend information into the post-processed forecasts. With the use of longer observation periods, the decadal and multi-decadal variability associated with the large-scale climate drivers, such as El Niño–Southern Oscillation and Madden–Julian oscillation, are considered when estimating the underlying changes in the chaotic nature.

In this chapter, we aim to evaluate the capability of GCM sub-seasonal forecasts in capturing the observed trend and to adapt the trend-aware method to post-process sub-seasonal forecasts with long-term climate trend embedded. We evaluate and establish the calibration models for the weekly aggregated retrospective forecasts of daily minimum and maximum temperatures across the Australian continent produced by the ECMWF extended-range forecasting system.

5.4 Study Data

5.4.1 ECMWF sub-seasonal re-forecasts

This study makes use of the retrospective forecasts (hereafter re-forecasts) from the ECMWF extended-range forecasting system. ECMWF re-forecasts are produced 'on the fly'. That is, on every Monday and Thursday, a new set of 11-member ensemble re-forecasts are generated on the

same starting day and month as real-time ensemble forecasts but cover the past 20 years with forecast length of up to 46 days. In this study, we focus on the ensemble re-forecasts associated with the real-time forecasts initialised between 2nd of January and 31st of December 2020. The corresponding sets of re-forecasts thus covered 2nd of January 2000 to 31st of December 2019, giving 2100 date sets (20 years × 105 initialisation dates) for evaluations. This ECMWF global ensemble system integrates atmosphere, ocean, sea ice and land components. The ocean model is NEMO (Nucleus for European Modelling of the Ocean) v3.4.1 with a 0.25° horizontal resolution while the interactive sea-ice model is LIM2 (the Louvain-la-Neuve Sea Ice Model). The land surface component is modelled using HTESSEL (Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land). The horizontal resolution of the atmospheric model degrades from Tco639 (about 16 km) to Tco319 (about 32 km) after first 15 days. Readers are referred to ECMWF (2021) for details on model configurations. In this study, re-forecasts of 6-hourly minimum (Tmin) and maximum (Tmax) temperatures were retrieved from the ECMWF MARS archive system and downloaded at a 0.4° resolution in consideration of computational efficiency, the storage size of the resulting files, and the identification of the grid cell coordinates.

5.4.2 AWAP observations

This study uses daily Tmin and Tmax observations from the AWAP (Australian Water Availability Project) dataset (Jones et al., 2009). The gridded AWAP data have the resolution of 0.05°, and they are upscaled to match forecast resolution at 0.4° resolution using a bilinear interpolation method. We utilise the AWAP records covering a 30-year period 1990-2019, with the last 20 years overlapping the re-forecast period.

5.5 Methods

5.5.1 Alignment of daily forecasts and observations

We determine daily Tmax and Tmin data from 6-hourly gridded temperature forecasts, and ensure daily forecasts are properly aligned with daily observations. In Australia, Tmax and Tmin in the 24 hours are recorded at 9 am local time (e.g., UTC + 8 in Western Australia). On the recording day, Tmax is recorded against the previous day, while Tmin is archived against the recording day. ECMWF forecasts are initialised at midnight UTC, and Australia uses multiple time zones, so the forecasts are not exactly synchronised with the AWAP data.

Take the forecasts initialised at midnight UTC on the 3rd of February and Western Australia as an example. Tmax/Tmin forecasts retrieved at 0600 UTC on the 3rd represent the highest/lowest temperature value from 8 am to 2 pm Australian Western Standard Time. In this case, Tmax/Tmin forecast for Day 1 is determined by getting the maximum/minimum value of the four forecast steps, 0600, 1200, 1800 UTC on the 3rd and 0000 UTC on the 4th from Tmax/Tmin 6-hourly forecasts. In other word, the forecast for Day 1 is searched from 8 am 3rd to 8 am 4th for Western Australia, while the period for parts of eastern Australia is 11 am 3rd to 11 am 4th local time. Subsequently, Tmax forecast for Day 1 is paired with the observation on the 3rd of February, and Tmin forecast for Day 1 is paired with the observation on the 4th of February. In this regard, daily forecasts and daily AWAP observations are aligned with the time discrepancy of approximately 1-2 hr across Australia.

5.5.2 Strategy for model fitting and forecasting

In this study, we establish the calibration models for weekly averages of daily Tmax and Tmin. We pool weekly averaged data for all initialisation dates within each of February, May, August, and November, which are taken as the representative calendar months for the four seasons. With this configuration, some initialisation dates are weeks apart, so that the climatology of both forecasts and observations is likely to change over this period. To remove the seasonality in pooled data, we derive anomalies of daily forecasts and observations relative to the climatology and then aggregate daily anomalies of forecasts and paired observations to weekly averaged anomalies with forecast horizons of up to 4 weeks. Forecasts with 1-week forecast horizon, or termed week-1 forecasts, are defined as the average of the daily forecasts from Day 1 to Day 7, while week-4 forecasts are the average of the daily forecasts from Day 22 to Day 28.

We follow the method of Narapusetty et al. (2009) to calculate the observed 30-year climatological means based on daily temperature observations. For raw forecasts, we estimate the 20-year climatological means for the forecast of each day, from Day 1 to Day 28, separately based on pooled daily raw re-forecast means from all the initialisation dates during 2000-2019 (i.e., 105 dates × 20 re-forecast years to construct time series). Then the climatological means are subtracted from the original values to derive daily anomalies. The climatological mean on a daily scale is formulated as

$$y_{\mathrm{cm}}(t) = a_0 + \sum_{h=1}^{H}\left[a_h \cos(\omega_h t) + b_h \sin(\omega_h t)\right] \tag{5-1}$$

where $y_{\mathrm{cm}}(t)$ is the daily climatological mean, H is the number of annual harmonics, using the default value $H = 4$ as recommended by Narapusetty et al. (2009), parameters a_0, a_h, and b_h are determined by minimising the mean square difference between $y_{\mathrm{cm}}(t)$ and original data, $\omega_h = 2\pi h / P$, and P is the period. We use $P = 365.25$ for both observations and forecasts to account for the leap year in the evaluation period. Note that for each lead day, pooled daily raw forecasts only have 105 data points per year and the remaining dates are regarded as missing. When the climatological mean is calculated, daily raw forecasts need to be aligned with the correct dates while other missing dates are omitted in Eq. (5-1) because the method does not require t to be evenly spaced or to periodically occur during the same phase of the period (Narapusetty et al., 2009).

By pooling the anomalies for multiple dates, we assume that weekly data from one initialisation date to the next is conditionally independent. Calibration models are established for each lead time, each target month, and each grid cell over Australia under a leave-one-year-out cross validation setup. That is, we set aside pairs of data points in each of the 20 re-forecast years, train the remaining data points to fit one calibration model, and use this fitted model to validate all the omitted data points. After this process is repeated for 20 times, all the raw re-forecasts are calibrated. In one cross validation run, we intend to estimate trend parameters from a longer past observation period to ensure the climate trends embedded into the calibrated forecasts more realistically represent the long-term climate change. We fit the calibration model by pairing 19-year anomalies of raw ensemble re-forecast means with 29-year observation anomalies. Observed data are overlapped with forecast data over the 19-year re-forecast period while there are no synchronised forecast data with the first 10-year observed data. In this 10-year period, all the forecast data are treated as having missing values in the model fitting, which is handled by the calibration method to be described in Section 5.5.3.

As an example, consider calibration of week-1 forecast anomalies initialised in February in one cross validation run. To post-process week-1 re-forecast anomalies from 8 initialisation dates, including the 3rd, 6th, 10th, 13th, 17th, 20th, 24th, and 27th of February 2019, we train the calibration model using the first week raw re-forecast anomalies issued from all February initialisation dates over the period from 2000 to 2018, and corresponding observation anomalies falling between 1990 and 2018. In this regard, the sequence of training pairs is composed of 152 data points of raw forecast anomalies (19 years × 8 initialisation dates) and 232 data points of observation anomalies (29 years × 8 initialisation dates) in one cross validation run.

5.5.3 Trend-aware forecast calibration

In this study, we adapt the trend-aware forecast calibration method to post-process weekly averaged sub-seasonal forecasts of temperature variables. The trend-aware model was initially extended from the Bayesian joint probability (BJP) modelling approach (Wang and Robertson, 2011; Wang et al., 2009; Wang et al., 2019) that was demonstrated an effective tool for generating skillful and reliable seasonal forecasts of temperature, precipitation, and streamflow. However, the BJP algorithm is not by design capable of resolving the trend mismatch issue in calibrated forecasts because this method does not incorporate trend components to correct trends in the calibrated forecasts. To overcome this limitation, additional trend parameters are expressly introduced in the trend-aware method.

Here, the trend-aware method formulates the relationship between a predictor y_1 (raw forecast anomaly) and a predictand variable y_2 (observation anomaly). To fulfill the working assumption that the marginal distributions of y_1 and y_2 are normal, we utilise a single-parameter Yeo-Johnson transformation method to normalise temperature variables that are potentially non-normal. In this regard, y_1 and y_2 are transformed to $y_1^{'}$ and $y_2^{'}$ separately,

$$y^{'} = \begin{cases} \left[(y+1)^{\lambda}-1\right]/\lambda & \lambda \neq 0, y \geq 0 \\ \log(y+1) & \lambda = 0, y \geq 0 \\ -\left[(-y+1)^{2-\lambda}-1\right]/(2-\lambda) & \lambda \neq 2, y < 0 \\ -\log(-y+1) & \lambda = 2, y < 0 \end{cases} \tag{5-2}$$

where λ is the transformation parameter optimised for y_1 and y_2 separately using maximum a posteriori (MAP) estimation method (Schepen et al., 2016).

After transformation, we linearly detrend transformed variables $y_i^{'}, i = 1, 2$ to z_i, where the individual anomaly $z_i(t), t = 1, 2, ..., T$ from the trendline of $y_i^{'}$ is calculated as

$$z_i(t) = y_i^{'}(t) - \alpha_i[Y(t) - Y(t_m)] \tag{5-3}$$

where t is a forecast event, t_m is roughly the middle event of the training period, T is the total number of events in the training period, α_i is a trend parameter, Y is a sequence of T time points corresponding to the event time of each individual forecast. In this chapter, time steps in Y are unevenly spaced.

Then detrended transformed predictor z_1 and detrended transformed predictand z_2 are modelled as a continuous bivariate normal distribution, with the form of

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{5-4}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are mean vector and covariance matrix, respectively. The collection of the model parameters to be inferred is denoted as $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha_1, \alpha_2\}$.

Before inferring model parameters, we need to determine their prior distributions. Non-informative multivariate Jeffreys priors (Gelman et al., 2014) are employed for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. In this study, we apply informative normal distribution priors for trend parameters $\alpha_i, i = 1, 2$, with the resulting trend-aware model named BJP-ti (Shao et al., 2021b). This type of prior distributions are centred at zero and have an empirically estimated variance, formulated as

$$p(\alpha_i) \propto N(0, m_i^2) \tag{5-5}$$

We set m_i as $m_i = \delta_i \times m_i^{"}$, where δ_i is the MAP estimate of the standard deviation of $y_i^{'}$ obtained from the variable transformation step. We follow the prior specification scheme elaborated in Chapter 4 to estimate $m_i^{"}$ using spatial and temporal neighbourhood information on a cell-by-cell basis. First, for each cell in Australia, for each of 4 lead times, for each of 12 months, for observation anomalies and for raw forecast anomalies separately, we run the BJP-t model, known as a member of the trend-aware method with non-informative uniform priors for trend parameters, and record the median value of sampled trend parameters α_i in the parameter inference without cross validation. The recorded median value denotes the trend of $y_i^{'}$, and is then divided by δ_i. The value of trend/ δ_i is archived for each grid cell. To determine $m_i^{"}$ for individual case, we select all the values from the cells within a 7-cell by 7-cell region centred at the case, and from the cells in consecutive 3 months (last, this and next month). After pooling the values from all 147 cells together, the value of $m_i^{"}$ is calculated as the 75$^{\text{th}}$ percentile of the absolute trend/ δ_i values. This local searching approach accounts for the distinctness of the temperature regions across Australia and across different months, which is more robust than the strategy of fixing the prior parameter for all evaluation months and all grid cells over Australia as introduced in Chapter 3.

After specifying all the prior distributions for model parameters, we infer the parameter sets $\boldsymbol{\theta}$ and missing values from a sequence of training data pairs $\mathbf{D} = \{[y_1^{'}(t), y_2^{'}(t)], t = 1, 2, \ldots, T\}$. The

Gibbs sampling method is employed to iteratively sample model parameters and missing variables in turn. The parameters sets $\boldsymbol{\theta}$ are sampled from the corresponding conditional posterior distributions deduced from the overall posterior distribution, written as

$$p(\boldsymbol{\theta} \mid \mathbf{D}) \propto p(\boldsymbol{\theta}) p(\mathbf{D} \mid \boldsymbol{\theta}) \tag{5-6}$$

where $p(\boldsymbol{\theta})$ is the prior distribution for model parameters, and $p(\mathbf{D} \mid \boldsymbol{\theta})$ is the likelihood function. The conditional posterior distributions for different subsets of model parameters are elaborated in Chapter 4.

The values of missing variables are sampled from the conditional distribution

$$[z_i(t) \mid \cdot] = \mathrm{N}[\mu_i^*(t), \Sigma_{i,i}^*] \tag{5-7}$$

where

$$\Sigma_{i,i}^* = \sigma_i^2 - (\rho \sigma_1 \sigma_2)^2 / \sigma_{(i)}^2 \tag{5-8}$$

$$\mu_i^*(t) = \mu_i + \rho \sigma_1 \sigma_2 / \sigma_{(i)}^2 \times [z_{(i)}(t) - \mu_{(i)}] \tag{5-9}$$

(i) is the index in $\{1, 2\}$ that is not i; σ_1, σ_2 and ρ are the parameters that constitute $\boldsymbol{\Sigma}$; $\mu_{(i)}$ is the parameter that constitutes $\boldsymbol{\mu}$.

When all the parameter sets $\boldsymbol{\theta}$ are available, we use the model for prediction. For each set of the inferred $\boldsymbol{\theta}$, we sample a trend-embedded calibrated forecast $y_2'(t^*)$ given a new transformed predictor value $y_1'(t^*)$. That is, we set the predictand as having a missing value and formulate a Gibbs sampler to sample a new calibrated forecast $z_2(t^*)$ based on the conditional distribution of the predictand given the predictor $z_1(t^*)$ as formulated in Eqs. (5-7) – (5-9), and re-trend it to $y_2'(t^*)$.

Before sampling $z_2(t^*)$, we use a pragmatic approach to adjust extremely large or small $z_1(t^*)$ temperature values that occur in prediction. In this chapter, we specify the extreme threshold as 0.001 and 0.999 in the non-exceedance probability following the marginal distribution of z_1 (Wang et al., 2019). A collection of calibrated forecast values $y_2'(t^*)$ are back-transformed to the original space $y_2'(t^*)$ to represent forecast uncertainty. Detailed descriptions and implementation of the trend-aware forecast calibration method are provided in Chapter 4.

5.5.4 Forecast verification

The cross-validated BJP-ti calibrated ensemble forecast anomalies are verified against raw and BJP calibrated forecast anomalies with respect to the ability to capture the historical trends, the forecast skill and reliability during the re-forecast period of 2000-2019. All the metrics are calculated for pooled temperature forecast anomalies and observation anomalies from all forecast initialisation dates in one target month for each grid cell and each lead time separately. For brevity, we refer to forecast anomalies as forecasts, and observation anomalies as observations hereafter.

For trend analysis, we adopt the linear regression method to estimate decadal temporal trend (Hartmann et al., 2013) in observations, raw forecast means, BJP and BJP-ti calibrated forecast means. Note that the events in the evaluation period are not evenly spaced in this study, so that the event time needs to vary accordingly. The statistical significance of the trend is checked by the two-tailed t test at 1% and 5% significance level, and the multiple testing problem is considered here. We follow Wilks (2016) to control the false discovery rate (FDR) at level $\alpha_{\mathrm{FDR}} = 0.02$ and 0.1, assuming a strong spatial correlation in the gridded data with $\alpha_{\mathrm{FDR}} = 2\alpha_{\mathrm{global}}$.

To measure forecast skill, we compute the continuous ranked probability score (Hersbach, 2000; Matheson and Winkler, 1976) that characterizes the difference between ensemble forecasts and observations. For each cell, the averaged CRPS value of the forecasts with events $t = 1, 2, ..., n$ is calculated as

$$\mathrm{CRPS} = \frac{1}{n}\sum_{t=1}^{n}\int\left\{F(t,y) - H\left[y - y_o(t)\right]\right\}^2 dy \tag{5-10}$$

where n is the total number of historical events in the analysis period, $F(t,y)$ is the cumulative distribution function (CDF) of the ensemble forecasts of variable y at event t, $y_o(t)$ is the observed value and H is the Heaviside step function that is equal to 0 if $y < y_o(t)$ and equal to 1 otherwise. Then we compare the averaged CRPS value of the model forecasts against the averaged CRPS value of reference forecasts to obtain the CRPS skill score for each cell. The reference forecasts are leave-one-year-out cross-validated climatology ensemble forecasts generated from the BJP model (Wang et al., 2019). The CRPS skill score is given as

$$\mathrm{CRPS}_{\text{skill score}} = \frac{\mathrm{CRPS}_{\mathrm{ref}} - \mathrm{CRPS}}{\mathrm{CRPS}_{\mathrm{ref}}} \times 100\left(\text{unit: \%}\right) \tag{5-11}$$

A CRPS skill score of 100% means that the forecasts perfectly match observations. A skill score closer to zero implies that the forecasts are as accurate as the reference forecasts. A negative skill score suggests that the forecasts perform poorer than the reference forecasts.

To check forecast reliability, for each cell, we firstly compute the probability integral transforms (Wang et al., 2009) for observations, and then calculate the PIT score that quantifies the deviation of the PIT values from the theoretical standard uniform values (Renard et al., 2010). In a perfectly reliable forecasting system, the collection of PIT values follows a standard uniform distribution, where the likelihood of the event is accurately estimated. The PIT value $\pi(t)$ for an observational event $y_o(t)$ and the corresponding forecast CDF $F(t, y)$ is defined as

$$\pi(t) = F\left[t, y_o(t)\right] \tag{5-12}$$

The PIT score has the form

$$\text{PIT score} = 1.0 - \frac{2}{n}\sum_{k=1}^{n}\left|\pi(k) - \frac{k}{n+1}\right| \tag{5-13}$$

where $\pi(k)$ is the k^{th} ranked PIT value $\pi(t)$ and $k/(n+1)$ is the k^{th} theoretical $\pi(k)$ value. The greater PIT score indicates more reliable ensemble forecasts. In this study, for each lead time, we choose to pool the PIT scores for all grid cells, and for all initialisation dates in four months together to summarise the forecast performance with non-exceedance plots.

5.6 Result

5.6.1 Trends in observations and model forecasts

In this section, we examine the spatial and temporal patterns of the linear decadal trends in observations and model forecasts. Here, we mainly present the results for week-1 and week-4 forecasts as the key findings from all-lead-time evaluations can be summarised by exploring these results. The geographic trend patterns of week-1 Tmin and Tmax variables are presented in Figure 5-1 and Figure 5-2, while the week-4 trends are presented in Figure 5-3 and Figure 5-4 respectively. The trend maps for week-2 and week-3 variables are shown in Appendix S4 (Supplementary Figure S4-1 – S4-4).

5.6.1.1 Week-1 observed and forecast trends

For the period 2000-2019, both warming and cooling trends are apparent for week-1 Tmin in four months (first column in Figure 5-1). Strong warming trends at the rate of larger than 1.2 °C per decade are also statistically significant at 5% level (Supplementary Figure S4-5) in some regions, such as part of northern Australia for May and western Australia for August. As such, discernible cooling trends dominate some portions of eastern Australia for August and southern Australia for November. For comparison, observed trends across the 30-year period (1990-2019; the second column in Figure 5-1) generally exhibit different spatial patterns of directions, magnitudes, and statistical significance (Supplementary Figure S4-6). For example, there are weak cooling trends shown in parts of eastern Australia for May over 1990-2019, which are largely reverted to warming trends over 2000-2019. This finding indicates the trends in observed records are highly sensitive to the evaluation periods.

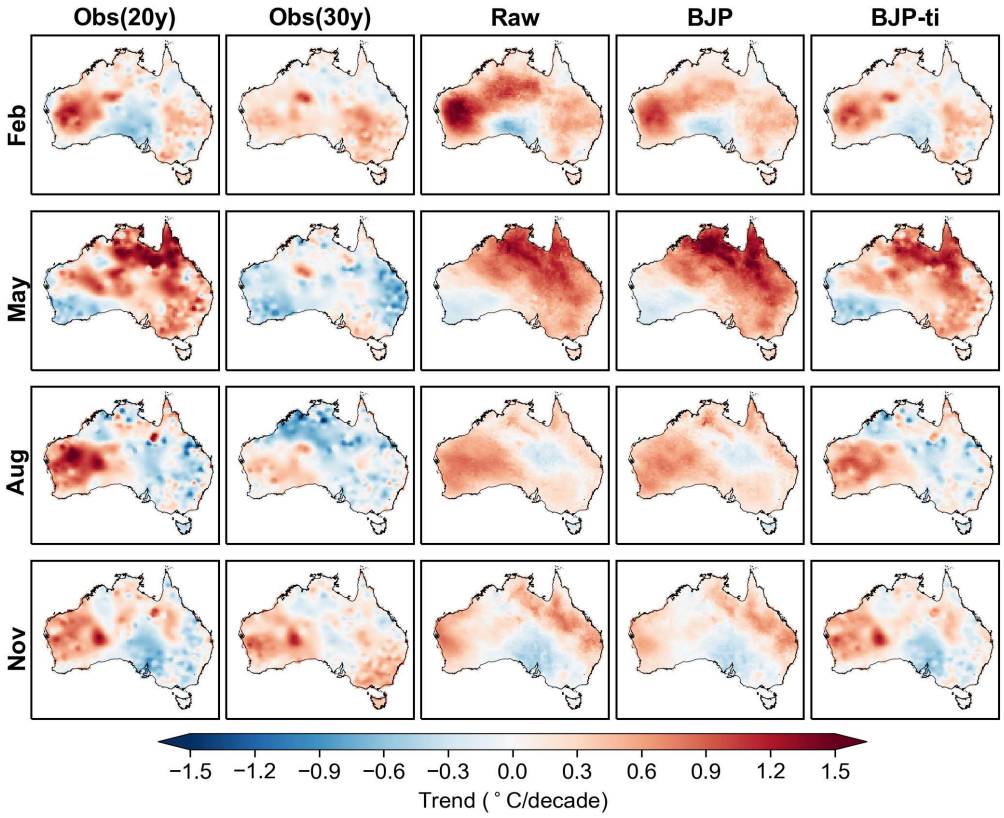

Figure 5-1: Decadal trends for Tmin observations over 2000-2019 and 1990-2019, raw forecasts, BJP calibrated forecasts and BJP-ti calibrated week-1 forecasts over 2000-2019 for all initialisation dates within February, May, August, and November separately.

For Tmax, more warming other than cooling trends are seen in four evaluation months during the 20-year and 30-year observed periods (first and second column in Figure 5-2), with many of the increasing trends statistically significant at 5% level (Supplementary Figure S4-7 and Figure S4-8). Similar to the findings for Tmin, the geographic trend patterns shown in two analysis periods are distinctly different, where the 30-year climate trends are relatively weaker than the apparent trends of the 20-year observations. As an example, in February, extremely strong and statistically significant decadal observed trends (larger than 1.5 °C per decade and significant at 1% level) are observed across western half of the country during 2000-2019, while the longer 30-year trends shown in the same region are mostly lower than 0.6 °C per decade and are widely not significant at 1% level (Supplementary Figure S4-8). This is possibly because sampling variability rather than true climate trends dominate the changes in a short-term period, so that the fitted trend slopes are exceptionally steep.

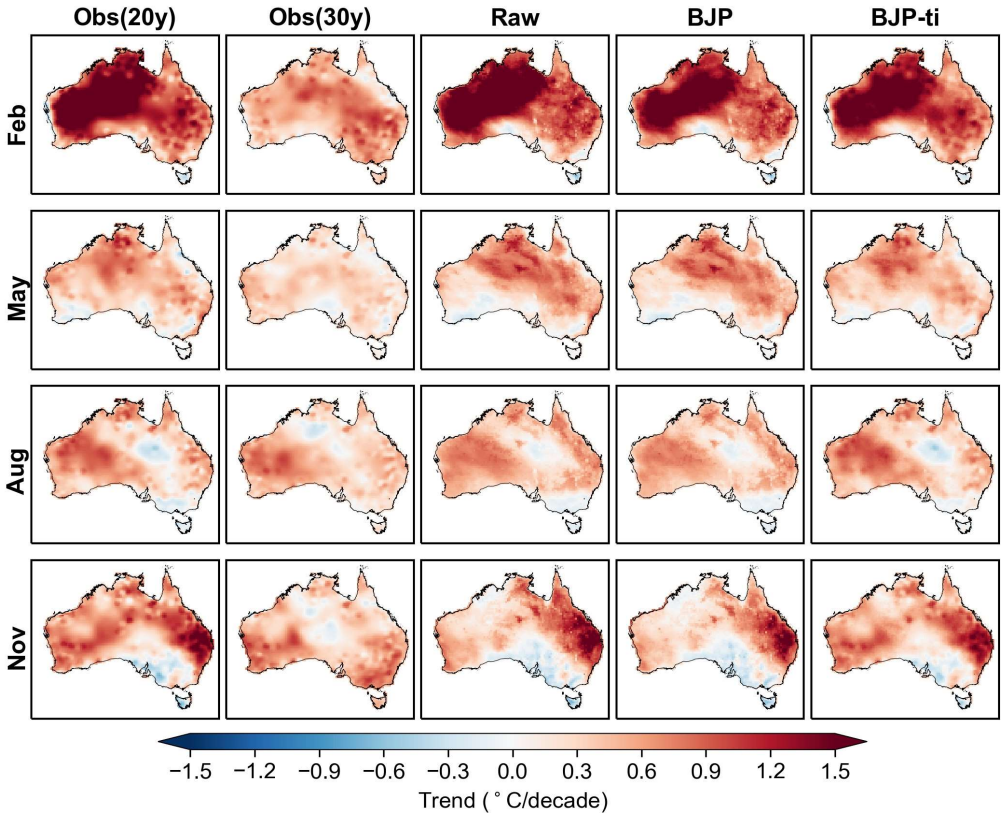

Figure 5-2: As in Figure 5-1, but for Tmax.

Consistent trends in raw and BJP calibrated week-1 forecasts (third and fourth columns in Figure 5-1 and Figure 5-2) are found across all months for both Tmin and Tmax. These forecast trends

widely match the 20-year observations in terms of the trend directions. However, the trend slopes of the model forecasts do not match the apparent 20-year observed trends in some regions, such as part of western Australia for both Tmin and Tmax variables in November.

Interestingly, although the BJP-ti model is constructed to introduce a 30-year observed trend into the calibrated forecasts, the actual forecast trend during the 20-year re-forecast period appears to be roughly aligned with the 20-year observations in all months (fifth columns in Figure 5-1 and Figure 5-2). For example, BJP-ti calibrated forecasts reproduce strong observed warming trends in northern Australia for Tmin in May, and in west-central Australia for Tmax in February over 2000-2019.

To unveil how the trend-aware method works in this study, for a selected cell in February, we plot the BJP-ti calibrated week-1 and week-4 Tmax forecast quantiles and observations, along with trendlines of ensemble forecast means and observations over the calibration period of 1990-2019 (Figure 5-5). The cell is in western Australia (117.2°E, 26°S), and has distinct trend behaviours for the 20-year and 30-year observation period in February. The BJP-ti calibrated forecasts are produced using a leave-one-year-out cross validation setup during the entire 30-year period. Since raw forecasts are only available over 2000-2019, in the model prediction, the predictor values are treated as missing before 2000 and are sampled along with the predictand. The resulting calibrated forecasts are essentially the climatology with 30-year observed trends in the first 10 years. In week-1 (top plot of Figure 5-5), the trendline of the calibrated forecasts over 1990-2019 is roughly consistent with the trendline of the 30-year observations, indicating that the BJP-ti calibration is effective at embedding the observed trend of the entire calibration period into the forecasts. For the 20-year re-forecast period, the trendline of the calibrated forecasts is more aligned with the 20-year observations. Furthermore, the interannual variability of the observations is properly captured by the calibrated ensemble forecasts. This may be associated with the good agreement between raw forecasts and observations, reflected by the high skill score of the raw forecasts as shown in Figure 5-6 in Section 5.6.2.1. Mathematically, when raw forecasts are highly skillful, the BJP-ti model is formulated to transfer raw forecast skill into the calibrated forecasts while embedding the observed trend of the re-forecast period into the forecasts.

5.6.1.2 Week-4 observed and forecast trends

Week-4 observations also exhibit warming and cooling trends, but trend behaviours are distinct from week-1 observed trends over two evaluation periods (first and second column in Figure 5-3

and Figure 5-4) as the observations corresponding to the week-1 and week-4 forecasts are three weeks apart. For Tmin, apparent warming trends are strong and statistically significant warming at 1% level in west-central Australia for February and May while significantly strong, cooling trends are widespread in central Australia for August over 2000-2019 (Supplementary Figure S4-5). For comparison, in west-central Australia, the 30-year observations exhibit weak warming trends for February and weak cooling trends for May and August. Focusing on Tmax trends, over 2000-2019, prominent cooling and regionally significant 20-year trends at 5% level are apparent in central Australia for August (Supplementary Figure S4-7). Warming apparent trends of the 20-year observations dominate most of other regions across all evaluation months, with strong trends at the rate higher than 1.2 °C per decade seen in parts of northern and central Australia for February, May, and November. In contrast, the 30-year observations show weaker trends at the rate of between -0.3 °C and 0.6 °C per decade in all months.

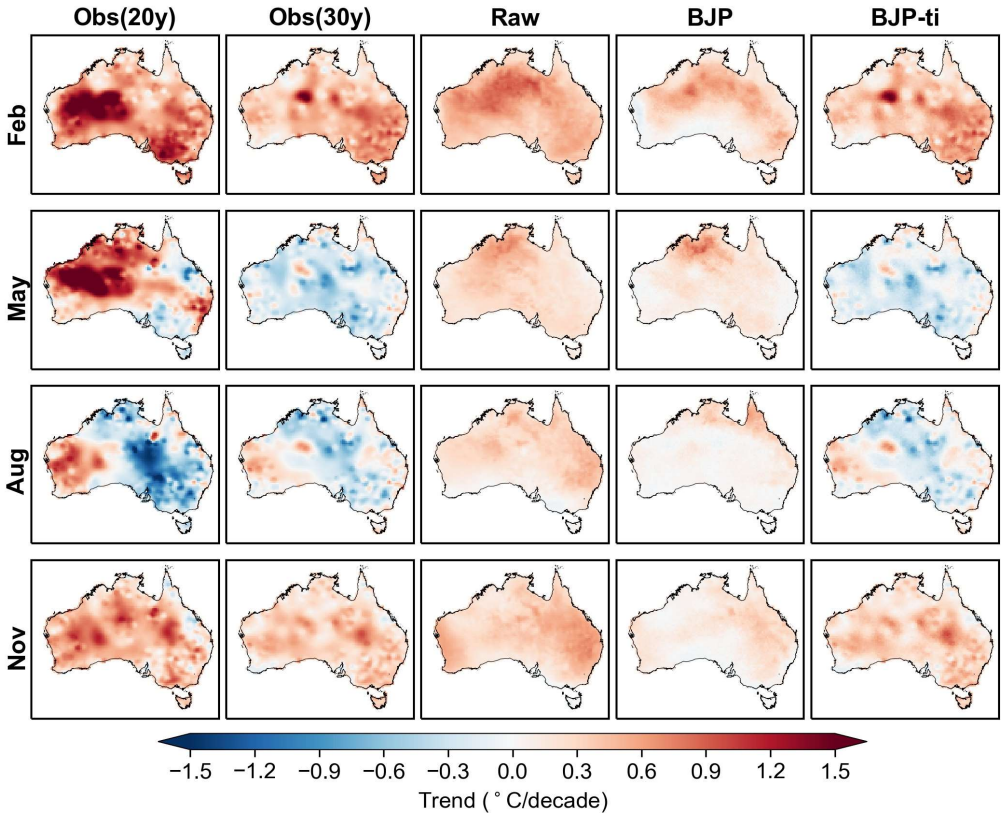

Figure 5-3: As in Figure 5-1, but for week-4 forecasts.

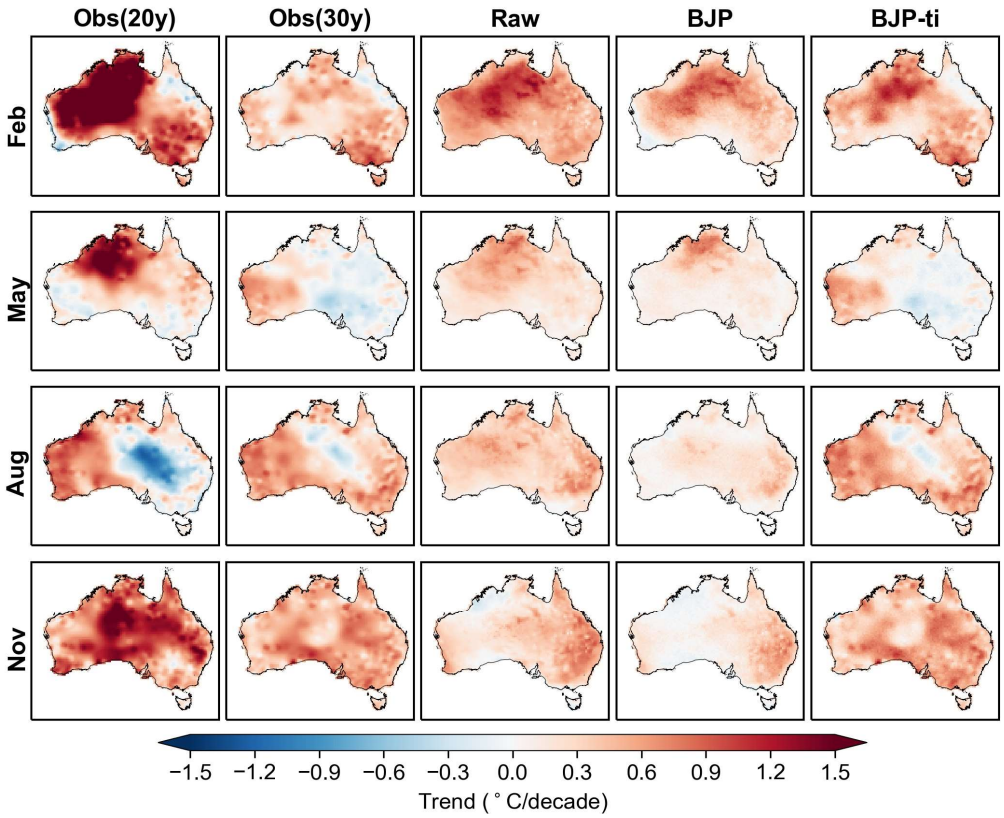

Figure 5-4: As in Figure 5-3, but for Tmax.

Compared to week-1 forecasts, trends in both raw and BJP calibrated forecasts (third and fourth column in Figure 5-3 and Figure 5-4) do not exhibit much spatial variability, and the trends are predominately increasing in most regions. Furthermore, these forecast trends generally under-estimate the apparent trends of the 20-year observations or misrepresent trend directions. After BJP-ti post-processing, the calibrated forecasts show the trend patterns consistent with the 30-year observations for both Tmin and Tmax across all evaluation months over 1990-2019 (fifth column in Figure 5-3 and Figure 5-4). The possible reason can again be explained using the case example shown in Figure 5-5 (bottom). Here, trendlines of the BJP-ti calibrated week-4 forecasts for both 20-year and 30-year periods follow the 30-year observations, possibly because raw forecasts are not in good correspondence with the observations, indicated by low skill score (see Figure 5-7). In this respect, the trend-aware BJP-ti model is formulated to revert the calibrated forecasts to the climatology-like forecasts that have 30-year observed trends embedded for the re-forecast period, which is considered more representative of the underlying trend than what is shown in the 20-year observations. Note that in this case, the CRPS skill score of raw week-4 forecasts is approximately 3.9%, indicating that raw forecasts are not entirely unskillful.

Subsequently, minor forecast skill is transferred into the BJP-ti calibrated forecasts, which show greater interannual variability than climatology forecasts over 2000-2019 but have lower variability than BJP-ti calibrated week-1 forecasts.

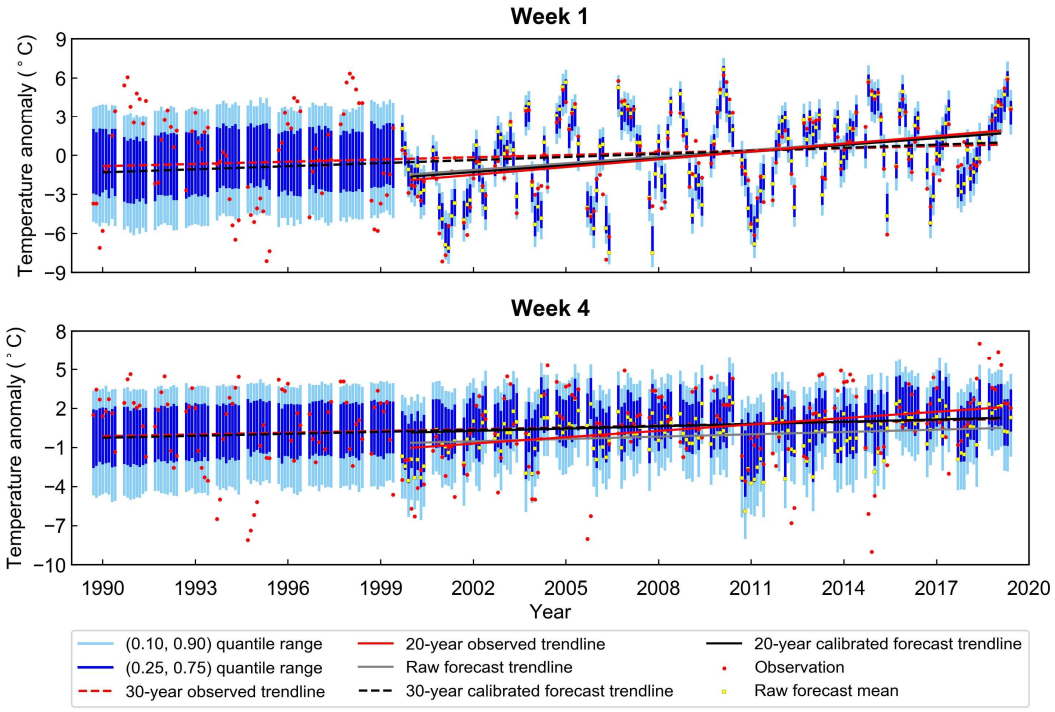

Figure 5-5: Forecast quantiles of BJP-ti calibrated week-1 (top) and week-4 (bottom) Tmax forecasts and observations for a selected cell over 1990-2019. Red squares are 30-year observations, yellow squares are 20-year raw ensemble forecast means, light blue vertical strips are calibrated forecast [0.10, 0.90] quantile range, and dark blue vertical strips are calibrated forecast [0.25, 0.75] quantile range.

5.6.2 Skill scores for model forecasts

The CRPS skill score results of week-1 and week-4 raw forecasts, BJP-ti calibrated forecasts, and score difference of BJP-ti and BJP calibrated forecasts are presented in Figure 5-6 and Figure 5-7. The skill scores are calculated and evaluated over the re-forecast period of 2000-2019. The score difference is explored to show how the skill of the calibrated forecasts changes by embedding observed trends. Results in the first panel are for Tmin forecasts while results in the second panel are for Tmax forecasts. Skill scores for week-2 and week-3 forecasts are shown in the Appendix S4 (Supplementary Figure S4-9 and Figure S4-10). In addition, we apply a bootstrap method illustrated in Chapter 2 to check whether the BJP-ti calibration significantly improves or worsens

the CRPS skill score compared to BJP at 5% significance level for all lead times. Results of the score significance are presented in Appendix S4 (Supplementary Figure S4-11 and Figure S4-12).

5.6.2.1 Skill of week-1 forecasts

Raw week-1 forecasts (first column of both panels in Figure 5-6) generally have positive skill over a large portion for both Tmin and Tmax. Furthermore, Tmax forecasts appear to be more skillful than Tmin forecasts, whose skill scores are above 60% in most regions. The high skill could be explained by the removal of seasonal variation, so that systematic biases in raw forecasts are largely rectified. Elsewhere, raw forecasts still have some pockets of negative skill, particularly in northern Australia in February for Tmin, where the skill score is below -20%. Post-processing is thereby necessary to enhance forecast skill.

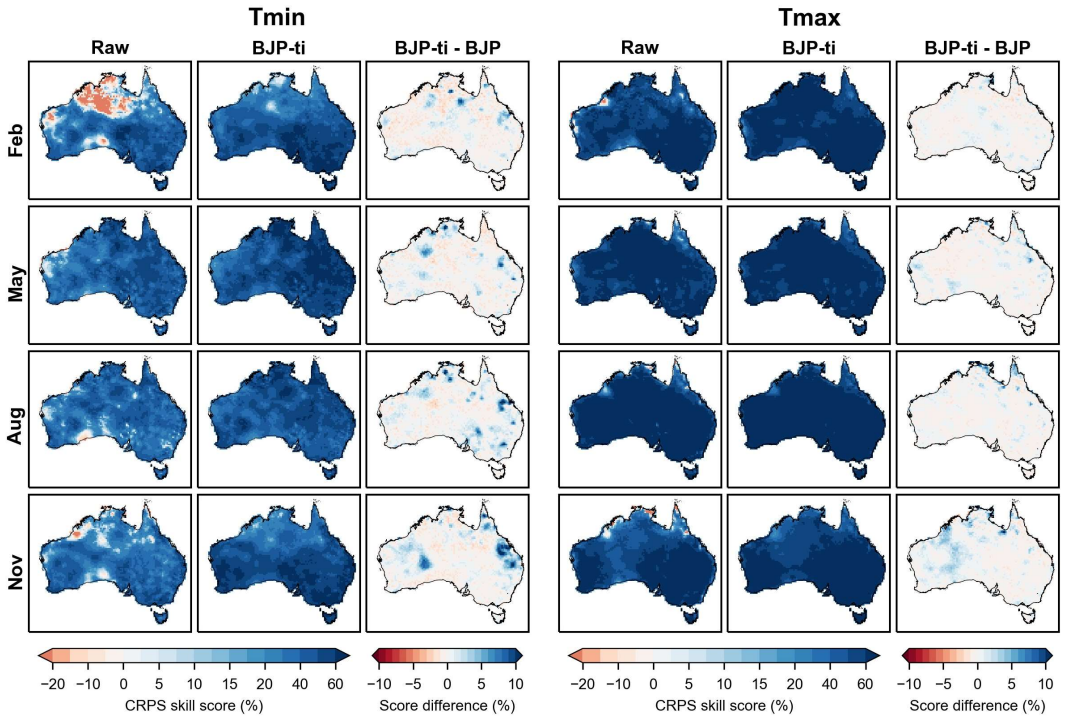

Figure 5-6: CRPS skill scores for Tmin and Tmax week-1 raw forecasts, BJP-ti calibrated forecasts, and score difference between BJP-ti and BJP calibrated forecasts for all initialisation dates within February, May, August, and November over 2000-2019.

With BJP-ti calibration (second column of both panels in Figure 5-6), negative skill pockets are mostly removed. Regions with skillful raw forecasts are generally retained, and skill gains are also evident in some regions. Overall, the skill score of calibrated forecasts is positive across

Australia. For Tmin, very high skill scores (value larger than 60%) dominate parts of southern Australia for February and November, northern and eastern Australia for May, and some clusters for August. For Tmax, highly skillful forecasts are prevailing. Relatively lower forecast skill, ranging between 10% and 40%, is shown in a few areas, particularly in parts of northern Australia for November.

Compared to BJP calibrated forecasts (third column of both panels in Figure 5-6), the skill improvement by the BJP-ti calibration takes place in the regions where the trends in BJP calibrated forecasts do not match the 20-year observations (Figure 5-1 and Figure 5-2). For example, statistically significant skill gains higher than 10% are seen along the coastal areas in the east for Tmin in November (Supplementary Figure S4-11). In these regions, slightly decreasing trends are found in the 20-year observations while increasing trends are apparent in both raw and BJP calibrated forecasts. As an additional example, for Tmax, moderate and insignificant skill increases (at approximately 5%) dominate many areas of western Australia in November where the magnitude of the apparent trends in the 20-year observations is underestimated in the BJP calibrated forecasts (Figure 5-2 and Supplementary Figure S4-12).

The BJP-ti calibration leads to slight and not statistically significant skill declines (i.e., score difference smaller than 5%) relative to BJP (Supplementary Figure S4-11 and Figure S4-12), particularly in the regions where trends in the BJP calibrated forecasts are highly consistent with the 20-year observations (Figure 5-1 and Figure 5-2). Examples are parts of northern Australia for Tmin and central Australia for Tmax, both in May.

5.6.2.2 Skill of week-4 forecasts

For both Tmin and Tmax, skill scores of raw week-4 forecasts (first column of both panels in Figure 5-7) are widely below 15% for all evaluation months, except for northern Australia in February for Tmax. Furthermore, negative scores dominate most of the regions in some months, such as August for both variables. Spatially, Tmax forecasts tend to be more skillful than Tmin.

Again, using the BJP-ti model reverts most negatively skilled raw forecasts to climatology-like (skill scores ranging between -5% and 5%) or skillful ensemble forecasts (second column of both panels in Figure 5-7). The positive skill of the raw forecasts is also mostly retained after BJP-ti calibration. However, in some regions, forecast skill could not be further improved, such as in parts of north-western Australia in February for Tmax.

With the BJP-ti calibration, the skill improvement relative to BJP calibrated forecasts dominates the regions where the trends of the 30-year observations are more consistent with the 20-year observations than the trends in BJP calibrated forecasts (third column of both panels in Figure 5-7). For example, statistically significant skill score increases (Supplementary Figure S4-11 and Figure S4-12) are evident in some clusters of western Australia in November for both Tmin and Tmax. Despite in different magnitude, in these regions, observed trends over the 20-year and 30-year evaluation periods are both increasing at the rate higher than 0.3 °C per decade. However, there are almost no trends in BJP calibrated forecasts, with trend slopes close to zero.

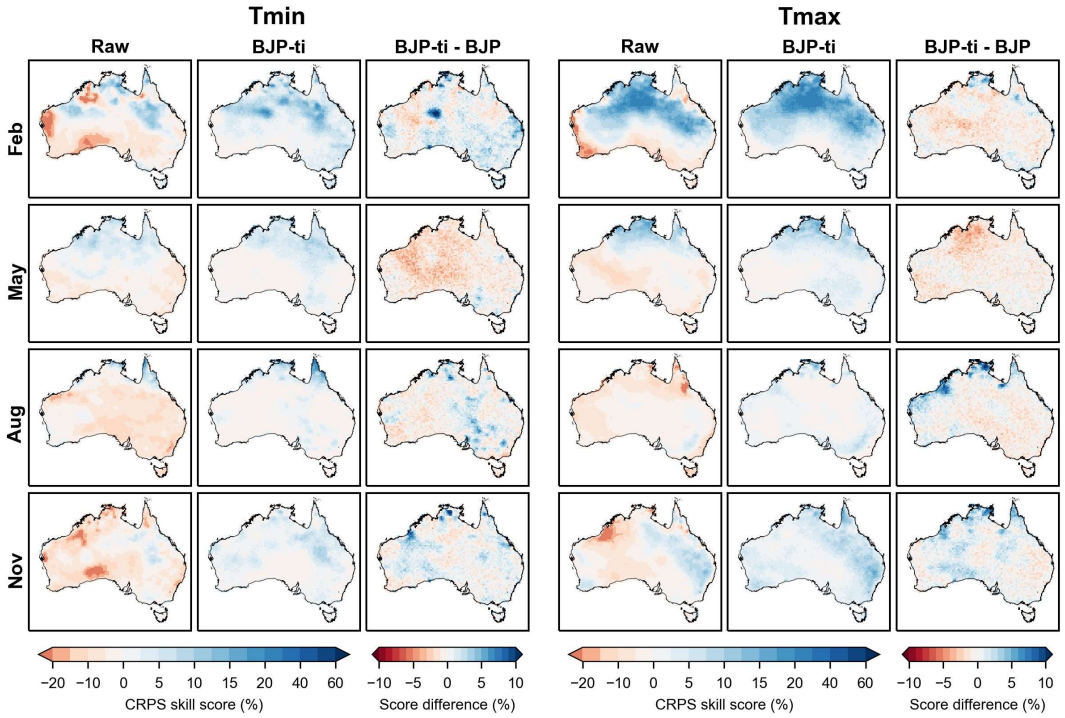

Figure 5-7: As in Figure 5-6, but for week-4 forecasts.

Skill declines by embedding the 30-year observed trends when the trends in BJP calibrated forecasts are already aligned with the 20-year observations, and when the trends of the 30-year observations do not match the 20-year observations. Particularly, the BJP-ti calibration leads to skill loss larger than 5% in the regions where the trend discrepancy between the 30-year observations and the 20-year observations is much larger than that between the BJP calibrated forecasts and the 20-year observations. Examples are central Australia for Tmin and parts of northern Australia for Tmax, both in May. In these cases, the BJP-ti calibrated forecasts exhibit trends that are more consistent with the 30-year observations. Although skill loss is evident in

these cases, the resulting trend-aware forecasts are more representative of the long-term changes in the climate system and are expected to boost user confidence in deploying the forecasts under the climate change condition.

5.6.2.3 Overall skill of forecasts

Results of the averaged CRPS skill scores of raw forecasts, BJP-ti calibrated forecasts and averaged score difference between BJP-ti and BJP calibrated forecasts over four evaluation months with 1-4 weeks forecast horizon are shown in Figure 5-8. As with the score maps shown in Figure 5-6 and Figure 5-7, for both Tmin and Tmax, raw week-1 forecasts are highly skillful (scores larger than 40%) across most of the continent. Relatively high skill score (over 20%) areas are widespread for raw week-2 forecasts. Beyond week-2, the skill score drops to below 10% in a large portion of the continent, with more regions showing negative scores for week-4. For all lead times, the skill of Tmax raw forecasts is generally higher than that of Tmin forecasts.

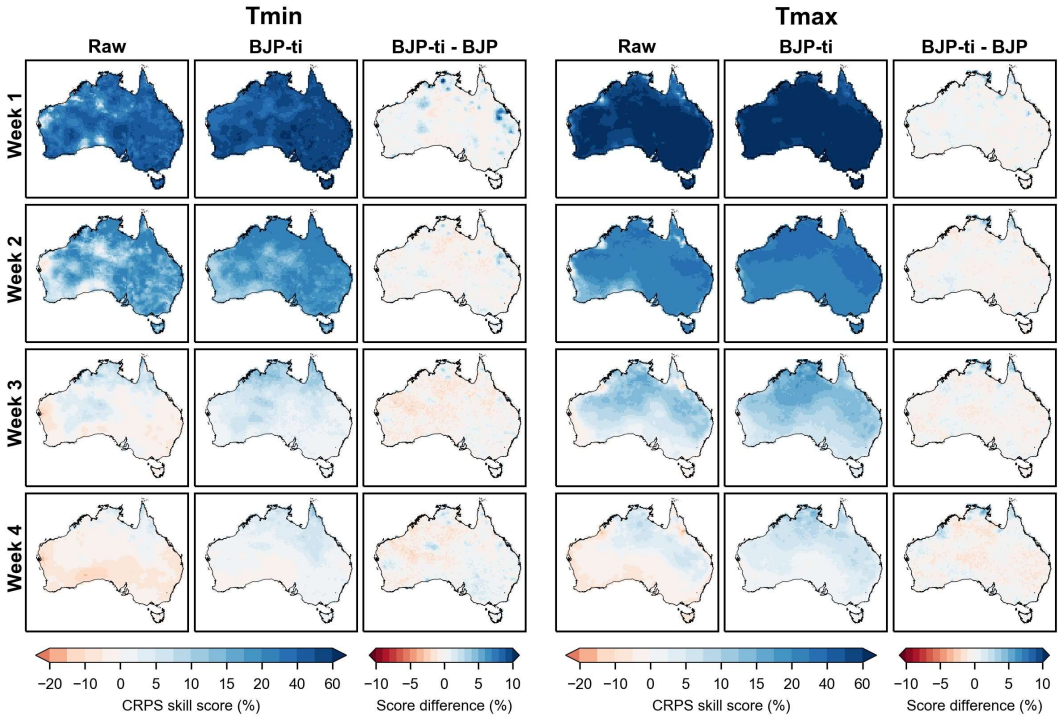

Figure 5-8: Averaged CRPS skill scores for pooled week 1-4 Tmin and Tmax raw forecasts, BJP-ti calibrated forecasts, and the score difference between BJP-ti and BJP calibrated forecasts over 2000-2019. The pooling is conducted over all evaluation months, February, May, August, and November.

Using the BJP-ti calibration is effective at improving forecast compared to raw forecasts (second column in both panels of Figure 5-8). In most regions, the BJP-ti calibrated forecasts are equally skillful as or more skillful than the raw forecasts. However, week-3 and week-4 BJP-ti calibrated forecasts still have widespread low skill (less than 10%), notably for Tmin. Overall, the BJP-ti calibrated forecasts appear to be as comparably skillful as the BJP calibrated forecasts, indicated by minor skill difference (less than 5%) across most parts of the continent.

5.6.3 Reliability

Here, we compare the reliability of raw, BJP calibrated, and BJP-ti calibrated forecasts with respect to pooled PIT scores from all evaluation months for Tmin and Tmax and for each of the lead times (Figure 5-9). Post-processing by both BJP and BJP-ti models leads to much more reliable forecasts than raw forecasts. Furthermore, the BJP-ti calibrated forecasts are generally more reliable than the BJP calibrated forecasts, except for week-1 Tmax. At this lead time, the BJP-ti and BJP calibrated forecasts are comparably reliable, possibly because the BJP calibrated forecasts are already highly reliable in ensemble spread, and the reliability could not be further improved by BJP-ti post-processing.

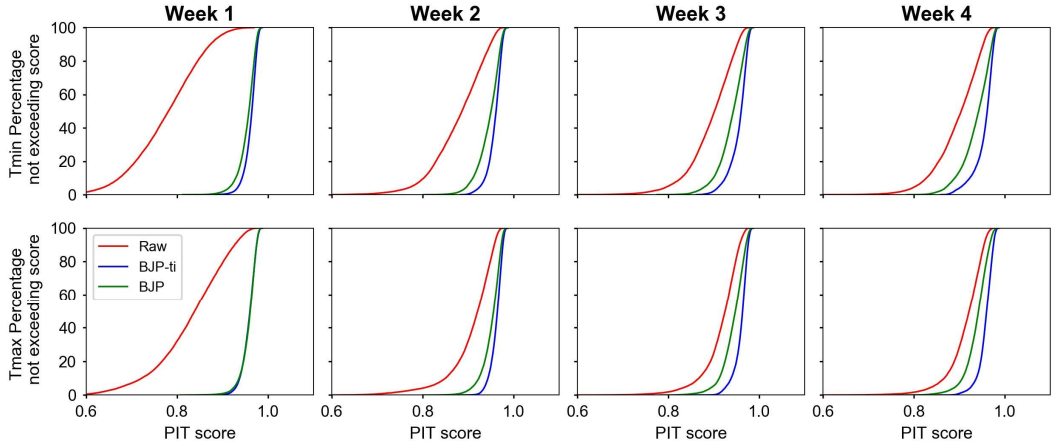

Figure 5-9: Pooled PIT scores for week 1-4 Tmin and Tmax raw, BJP, and BJP-ti calibrated forecasts over 2000-2019. The pooling is conducted over all evaluation months, February, May, August, and November.

5.7 Discussion

In this chapter, the trend-aware BJP-ti model is formulated to introduce the 30-year historical trend into the 20-year calibrated forecasts. In fact, as shown in Figure 5-1 - Figure 5-4, the trend-

aware calibrated forecasts do not necessarily exhibit the trend of the 30-year observations. The trend behaviour of the trend-aware calibrated forecasts is closely related to raw forecast skill. When raw forecasts are skillful, trends in trend-aware calibrated forecasts are roughly aligned with the 20-year observations. When raw forecasts have little skill, forecast trends after trend-aware calibration broadly follow the 30-year observations. As a result, trends shown in trend-aware calibrated forecasts are a mixture of the 20-year and 30-year observed trends. Compared to the BJP model, forecast skill is enhanced by using the trend-aware calibration in the regions where the trends in the trend-aware calibrated forecasts are in better agreement with the trends of the 20-year observations than the BJP calibrated forecasts. In contrast, skill declines by embedding the 30-year observed trends are evident when the trends in the trend-aware calibrated forecasts are less consistent with the 20-year observations than the BJP calibrated forecasts. Compared to raw forecasts, the trend-aware calibration largely retains the positive skill regions while reverting raw forecasts to climatology-like forecasts in the negative skill regions. When raw forecasts are already highly skillful, the trend-aware calibration may not further improve the forecast skill.

Since the trends fitted for the 20-year observations may show more random sampling variability than underlying changes in the past climate, in this study, we estimate trends from a longer 30-year period, which are likely to indicate the decadal changes in temperatures more realistically. The 30-year observations rather than a longer observed period is harnessed here because we seek to align the trend estimation with the number of years used to define the climate normal, which is conventionally calculated as the average value of a 30-year period (World Meteorological Organization, 2017). After trend-aware calibration, the evaluation of the CRPS skill score shows that the resulting forecasts are generally no worse than climatology forecasts with the 30-year climate trends. The long-term climate change is composed of both internal and external changes in the climate system, and trends detected from over 30 years of observations may be valuable for characterising climate change signals. Future work will investigate the merit of utilising prolonged observation periods in increasing the forecast value.

We set up calibration models for each lead time, each grid cell, and for pooled weekly anomaly data from all initialisation dates in a calendar month. The pooling is a more robust choice than the strategy of fitting the calibration model for the data from each of the initialisation date. This is because the model parameters may be poorly estimated given the fact that only 19 raw forecast points are available for model training in one cross-validated run. For comparison, the pooling strategy increases the length of the training data for model fitting to stabilise the inference of the

model parameters. Alternative calibration schemes are also feasible, such as establishing the models for pooled data from all initialisation dates spanning a season (van Straaten et al., 2020) or more than a season (Scheuerer et al., 2020). The results from different pooling schemes may not substantially differ for temperature applications. We note that in these pooling calibration schemes, even after the removal of the seasonality, the variance of derived anomalies may still differ among pooled initialisation dates. Therefore, new inference methods for standardising pooled data should be investigated in the future research. In addition to weekly aggregated data, it is also possible to extend the trend-aware model to handle daily temperature outputs from different GCMs. To do this, future work needs to conceive new strategies on building the daily post-processor that considers computational efficiency and data availability.

In Chapter 4, the trend-aware model was extended and applied to post-process seasonal precipitation forecasts. Technically, the method developed for seasonal precipitations is also applicable for handling sub-seasonal precipitation forecasts as the original BJP model was employed to effectively post-process sub-seasonal to seasonal forecasts of rainfall (Li et al., 2020; Schepen et al., 2018). Since the calibration models are established for daily or weekly aggregated data, there may be insufficient non-zero values in the model inference, leading to poor estimations of model parameters, and further unrealistically large uncertainty in calibrated forecasts. In future work, an effective calibration strategy or algorithm improvement plan will be required to robustly train the post-processing model and make the final inference more stable and efficient.

The skill of sub-seasonal climate forecasts is dominated by drivers of climate variability other than historical trends. Madden–Julian oscillation (MJO), for example, is a recognised global source of sub-seasonal predictability, and many sub-seasonal forecast models are skillful at predicting the MJO between 2 and 4 weeks forecast horizon (Vitart, 2017). It is likely that the skill of sub-seasonal climate forecasts can be enhanced by making use of the large-scale climate features in statistical forecast post-processing (Specq and Batté, 2020), particularly in the regions where modelled teleconnection patterns are poorly represented (Merryfield et al., 2020). Future avenues will seek to predict the climate variables using relevant teleconnection patterns as the predictor when establishing the post-processing model to improve forecast performance.

The trend-aware forecast post-processing method has the potential to be efficiently applied for operational use. In this research, the trend-aware BJP-ti model is coded up in C++ and the compiled C++ packages are called in Python for parameter inference and prediction use. For one cell, it takes less than a minute to generate the calibrated ensemble forecasts for pooled

initialisation dates at one lead time under the cross-validation setup. In addition, with the use of parallel computing, forecast communities and decision makers are expected to obtain calibrated real-time forecasts in a timely manner.

5.8 Conclusion

Sub-seasonal forecasts produced from global climate models (GCMs) could have far-reaching impacts on environmental, social, and economic sectors, as they may provide decision makers with valuable information for advance planning. Little attention has been paid to whether the GCM sub-seasonal forecasting system captures the observed climate trends. In this study, we firstly aim to examine the trends in raw ECMWF sub-seasonal temperature forecasts. Then we extend a trend-aware statistical calibration model, BJP-ti, to correct the trend in sub-seasonal forecasts. We build up a new calibration scheme to introduce a 30-year historical climate trend into the forecasts that have a much shorter 20-year re-forecast period available. Relative to day-of-year climatology, the trend-aware calibrated forecasts are compared with raw forecasts and the forecasts calibrated by the Bayesian joint probability (BJP) modelling approach.

We show that raw and BJP calibrated week-1 forecasts properly reproduce the apparent trend patterns of the 20-year observations in many regions, while trends in raw and BJP calibrated week-4 forecasts do not match the 20-year observations. After trend-aware post-processing, calibrated forecasts exhibit mixed trends of the 20-year and 30-year observations. When raw forecasts are inherently skillful, notably for week-1, trends in trend-aware calibrated forecasts are aligned with the 20-year observations. On the other hand, when raw forecasts have little skill, such as for week-4, trends in the trend-aware calibrated forecasts are largely consistent with the 30-year observations. Overall, week 1-2 trend-aware calibrated forecasts are highly skillful, while the forecasts are comparable with the climatological reference forecasts beyond week-2. In most regions, the calibrated forecasts are more reliable than raw and BJP calibrated forecasts while being as skillful as or more skillful than raw calibrated forecasts for all lead times.

The extended trend-aware forecast post-processing method has the potential to produce high-quality sub-seasonal forecasts and support decision-making. The merit of this method is more than skill improvement. After the forecast trend is corrected, the resulting forecasts should be more valuable for forecast users, especially when raw forecasts are seen consistently lower or higher than the observed values.

Ongoing research is likely to optimise the trend-aware forecast post-processing method for wider applications. We will investigate other feasible calibration schemes, adapt the method for post-

processing sub-seasonal precipitation forecasts and utilise other skill sources for further enhancing the performance of sub-seasonal forecasts.

Chapter 6 Discussions and Conclusions

6.1 Preamble

This final chapter starts with a summary of methods and findings, with emphasis on novel contributions to answering each research question (Section 6.2). Then I discuss limitations of the research and propose extension opportunities for future works (Section 6.3). Finally, Section 6.4 provides highlights and concluding remarks.

6.2 Research Overview and Findings

In the literature review in Chapter 1, an overarching research gap that motivated this thesis has been identified: while discernible climate trends have been observed in many parts of the world, the current generation of the global climate models (GCMs) for seasonal forecasting often fail to represent the observed climate trends in the re-forecasts. Meanwhile, commonly used statistical post-processing methods mostly aim to reduce the model biases and enhance the forecast reliability but are seldom designed to eliminate trend disparity between observations and seasonal forecasts of climate variables. To bridge this research gap, the overarching objective of this thesis is to develop a new, reliable trend-aware forecast post-processing method to introduce the observed trends into post-processed GCM forecasts while eliminating raw forecast biases, redressing skill deficits, and improving reliability in terms of ensemble spreads. Four research questions are proposed and explored to achieve the thesis objective.

RQ1: How can observed temperature trends be embedded into seasonal temperature forecasts through statistical post-processing?

As an initial model development, Chapter 2 extended the capability of the Bayesian joint probability (BJP) modelling approach to explicitly incorporate the observed trend into seasonal temperature forecasts. In this new Bayesian method (named BJP-t), I introduced additional linear trend components into the original BJP algorithm that modelled a bivariate relationship between raw forecasts and observations. A uniform (non-informative) prior was applied for trend parameters to explicitly infer the trend parameters from the training data.

The BJP-t model was tested on three example cases for January mean maximum temperature in Australia over a 36-year hindcast period. Observations were quality-controlled data from the weather stations located in different states. Criteria of case selection included: i) there was no missing observed value for the period 1982-2017, ii) the observed trend was not properly represented in raw and BJP calibrated forecasts, and iii) the observed trend was statistically significant at 5% significance level. Raw forecasts with 1-month lead time were extracted from the SEAS5 seasonal forecasting model, operated by the European Centre for Medium-Range Weather Forecasts (ECMWF). The BJP-t calibrated forecasts were compared against raw and BJP calibrated forecasts. The BJP and BJP-t calibration models were established and evaluated under a leave-one-year-out cross validation setup.

Results showed that the BJP-t calibration accurately embedded the 36-year observed temperature trend into calibrated ensemble forecasts when raw and BJP calibrated ensemble forecasts did not properly capture the observed trend. In two of the three selected cases, the BJP-t calibrated forecasts were significantly more skillful, more reliable, and sharper in the ensemble spread relative to both raw and BJP calibrated forecasts. For the third case, raw and BJP calibrated forecasts were already skillful and reliable, while the BJP-t calibration led to noticeable but not significant skill improvements.

RQ2: Is the trend-aware method applicable for post-processing seasonal temperature forecasts on a continental scale? Can the trend-aware method be improved to better consider trend uncertainty in the Bayesian inference?

Since the BJP-t model was effective for introducing historical trends into calibrated ensemble forecasts in selected cases, the next step was to test its robustness on broader cases. I found that, with the BJP-t calibration, trends entirely inferred from the observations over a limited available period (a 36-year hindcast period) were subject to large sampling errors in the Bayesian inference and might not represent real observed trends, particularly in the cases of no trend. Chapter 3 further developed the trend-aware forecast post-processing method to refine the treatment of trend uncertainty and demonstrated the merit of the improved method for post-processing seasonal temperature forecasts on a spatial continental scale.

I introduced two variants of trend-aware models: BJP-t with non-informative uniform priors for trend parameters and BJP-ti with informative normal distribution priors centred at zero and with empirically determined variance for trend parameters. In the BJP-ti model, a single prior

distribution was employed for each variance parameter. The prior distribution was constructed using trend information in all seasons and locations over Australia for raw forecasts or observations. With this prior configuration, trends were also inferred from the training data, but with a degree of moderation. I applied the calibration models for 1-month-ahead SEAS5 seasonal mean maximum and minimum temperatures forecasts and paired AWAP observations in all the grid cells for 12 overlapping seasons across the Australian continent. The calibration models were established under a leave-one-year-out cross validation setup. Raw and calibrated forecasts were compared and evaluated based on their ability to capture observed trends, and a range of forecast attributes, including skill, reliability, and sharpness.

Results revealed that the BJP-t calibrated forecasts accurately reproduced the observed trends while the trends in BJP-ti calibrated forecasts were slightly weaker than the observed trends. Both BJP-t and BJP-ti models greatly improved the forecast skill when the BJP calibrated forecasts misrepresented the trend direction or where the observed trends were statistically significant at 5% level. When the trends in the BJP calibrated forecasts roughly matched the observed trends, the BJP-t calibration led to a slight skill degradation compared to BJP. The BJP-ti calibration experienced less skill degradation than BJP-t because the inferred trends were moderated to account for large trend uncertainty in the observations. Sharper forecasts were also produced by the BJP-ti calibration compared to the BJP-t model.

RQ3: How can the trend-aware model be adapted to post-process seasonal precipitation forecasts?

Precipitation is another key meteorological variable for climate-sensitive sectors, which has shown marked spatiotemporal changes in many regions. Previous research also documented that GCM seasonal precipitation forecasts poorly captured the climate trends in some regions. Accordingly, Chapter 4 further improved the trend-aware method for post-processing seasonal precipitation forecasts in Australia. Challenges posed for post-processing precipitation variables are that precipitation amounts are lower bounded by zero, follow a positively skewed distribution, and are highly variable and uncertain in space and time.

I introduced new formulations into the advanced version of the trend-aware method to treat occurrences of zero rainfall. A new scheme was proposed in specifying the parameters of the prior distribution for the trend components to account for local precipitation regions. Specifically, this work used the same form of the prior distribution for the BJP-ti model as described in Chapter 3, but here the prior parameter was estimated cell-by-cell based on the temporal and spatial

neighbourhood information rather than share a single prior distribution throughout all cells across Australia and seasons. As with the earlier work, I used SEAS5 forecasts and AWAP observations as study data and compared the BJP-ti calibrated forecasts with the raw and BJP calibrated forecasts. The post-processing models were established for each grid cell and each season separately across Australia at 1-month lead time. For four selected cells, the models were established for all seasons at 0- to 4-month lead times separately. The evaluation of the ensemble forecasts adopted a leave-one-year-out cross validation setup.

The BJP-ti calibrated forecasts were found to properly reproduce the observed trends and to follow the magnitude of the strong trends in all seasons. The BJP-ti calibration achieved higher forecast skill than BJP predominately over the regions where the observed trends were statistically significant at 10% significance level. When pooling the results from all seasons together, the BJP-ti calibrated forecasts were shown to substantially outperform raw forecasts and perform comparably with the BJP calibrated forecasts with respect to forecast bias, skill, and reliability. Results from selected cases revealed that the BJP-ti calibration improved the prediction of the interannual variability and produced skillful or at least climatology-like forecasts for all lead times.

RQ4: Are sub-seasonal temperature forecasts capable of reproducing historical trend information? How can the trend-aware model be adapted to post-process sub-seasonal forecasts?

Current GCM sub-seasonal forecasting systems are configured not dissimilar to seasonal forecast models. While GCM seasonal forecasts were reported to poorly represent climate trends, whether the same issue exists in GCM-based sub-seasonal forecasts had not been investigated yet. Chapter 5 firstly evaluated the ability of GCM sub-seasonal temperature forecasts to capture the observed trend. Subsequently, I adapted the trend-aware forecast post-processing method to embed long-term historical climate trends into sub-seasonal forecasts.

In this part, the 20-year re-forecasts of minimum and maximum temperatures were produced from the ECMWF extended-range forecasting system and paired observations were retrieved from the AWAP dataset. The short re-forecast period means the inferred trend has large sampling variability and may not correctly reflect how past climate changes. Thus, when formulating a trend-aware calibration model, BJP-ti, I allowed a 30-year climate trend to be introduced into sub-seasonal forecasts, which more robustly represented the long-term climate change. Specifically, I made use of 30-year observations over 1990-2019 while raw re-forecasts are from a 20-year period 2000-2019. Anomalies of raw forecasts and observations were obtained to minimise the

effect of seasonality on model fitting. Calibration models were fitted for pooled weekly averaged observation anomalies and weekly averaged forecast anomalies issued from all the initialisation dates for each cell, each lead time (week 1-4) and each of four representative months (February, May, August, and November) across Australia. The calibration models were established under a leave-one-year-out cross validation scheme.

Results showed that relative to day-of-year climatology, raw and BJP calibrated week-1 forecasts roughly reproduced the apparent trends of the 20-year observations in many regions during the re-forecast period. After BJP-ti post-processing, the calibrated forecasts exhibited trends broadly aligned with the 20-year observations. This is because the BJP-ti calibration could explicitly transfer raw forecast skill and embed the apparent 20-year observed trends into the calibrated forecasts when raw forecasts were inherently skillful, notably in week-1. In contrast, raw and BJP calibrated week-4 forecasts did not show apparent trends of the 20-year observations. The BJP-ti calibrated forecasts now exhibited trends of the 30-year observations because the BJP-ti calibration reverted the forecasts to climatology with the 30-year observed trends embedded when raw forecasts did not have much skill. Compared to BJP calibrated forecasts, skill improvement by the BJP-ti calibration dominated the regions where the BJP-ti calibrated forecasts exhibited trends more consistent with the 20-year observations than the BJP calibrated forecasts. When results from four evaluation months were pooled together, BJP-ti calibrated week 1-2 forecasts were highly skillful while the forecasts were comparable with the climatology forecasts beyond week 2. After BJP-ti post-processing, in most regions, the calibrated forecasts were more reliable than raw and BJP calibrated forecasts while being as skillful as or more skillful than raw forecasts at all lead times.

6.3 Limitations and Extension Opportunities

Each of Chapter 2 – Chapter 5 has provided its own discussions. The following paragraphs present an overview of the limitations and future research directions relevant to all the research works in this thesis.

In the trend-aware post-processing method, the trend component is modelled as a linear function for the transformed variables in all cases. In fact, changes in the hydro-meteorological variables could exhibit non-linear behaviour related to evaluation periods and study regions. This is more evident for precipitation, which is highly variable and lacks an easily detectable trend. Many studies identified an abrupt change in precipitation series (Rahmat et al., 2015; Ullah et al., 2018),

indicating that the trend-aware model could be potentially modified to account for such a sudden change. In this regard, whether there exists a step change point in the training data could be identified, and if so, trends will be modelled for the segments before and after the change point separately. With this trend form embedded, the trend-aware method is expected to be more flexible in incorporating temporal trend information into the resulting forecasts.

Another potential modification to the method is to include information on ensemble spreads of the raw forecasts when formulating the post-processing model. In the current version of both BJP and trend-aware model, only ensemble mean of the raw forecasts is utilised, which does not consider the spread-skill relationship as established in another full calibration method, ensemble model output statistics (Gneiting et al., 2005). With the advancement of the ensemble generation technique implemented in the forecasting system, it is highly possible that trend-aware post-processing with spread information will be beneficial for improving the forecast performance.

The leave-one-year-out cross validation method was deployed for validating calibration models throughout the thesis. However, this method has an inherent limitation that it only works for the anomaly component, but not for the trend component, which may lead to model overfitting. An alternative validation method is omitting several years of data at the beginning and end of the entire data records for validation while training the calibration model with the remaining data. However, the re-forecast period of current GCM forecasting systems is generally short, for example 36 years for SEAS5 seasonal forecasts and 20 years for ECMWF sub-seasonal forecasts in this thesis. Consequently, the results may still be affected by large sampling variability using this alternative validation strategy. In the future, if there are longer periods of GCM outputs available, say 50 years, this alternative could be a better choice to validate the trend-aware method.

The trend-aware method infers a linear trendline in the normal space, and such trendline does not necessary remain linear after transformed back to the real space. From preliminary analysis, the trendline of back-transformed data is found to be linear-like for temperature variables while the trendline becomes quadratic-like for precipitation. Therefore, different trend evaluation methods were selected to comprehensively assess the trend slope and its statistical significance for temperature and precipitation in the real space. The linear regression method and two-tailed Student's *t* test were employed for temperature variables while the Theil-Sen slope and Mann-Kendall test were used for precipitation amounts. The use of these evaluation metrics was also aligned with other papers on trend analysis (Hartmann et al., 2013; Kumar et al., 2013; Livada et al., 2019). Despite popular applications, the effectiveness of these conventional methods is subject

to various factors, so they may not always be satisfactory for trend detection. For example, the linear regression is sensitive to outliners and works on the assumption that the data series follows a Gaussian distribution. The power of the Mann-Kendall test could be affected by trend magnitude, sample length, data variance and the existence of the positive autocorrelation in the data series (Wang et al., 2020). Future work will seek more robust trend evaluation tools to assess trends in various variables.

In this thesis, the trend-aware post-processing method was only applied to post-process univariate climate forecasts with only one predictor (raw GCM output) and one predictand (corresponding observation) used to establish the calibration scheme. In practice, there is an opportunity to extend the trend-aware method for multivariate post-processing of hydro-meteorological variables (Schepen et al., 2020c). For example, temperature information could be utilised to post-process precipitation forecasts (Narapusetty et al., 2018) in a multivariate model configuration. Apart from forecast post-processing, it is also feasible to use the trend-aware method for statistical forecasting of hydrometeorological variables. The BJP model, the predecessor of the trend-aware method, has been employed as a statistical model to empirically produce seasonal forecasts of temperature, precipitation, and streamflow. Climate indices are candidate predictors for climate forecasting (Schepen et al., 2016; Schepen et al., 2012a; Schepen et al., 2014) while additional predictors, antecedent streamflow and precipitation, could be adopted for streamflow forecasting (Feikema et al., 2018; Wang and Robertson, 2011).

The trend-aware method has been introduced and applied to SEAS5 seasonal temperature and precipitation forecasts, as well as ECMWF sub-seasonal temperature forecasts. Theoretically, this method is also applicable for post-processing the forecasts from other operational GCM forecasting systems (Hudson et al., 2017; Hudson et al., 2018; Saha et al., 2014), and for post-processing sub-seasonal precipitation forecasts and other hydrometeorological variables on sub-seasonal and seasonal timescales that exhibit marked trends in recent decades, such as evapotranspiration (Dinpashoh et al., 2019; Peng et al., 2017). In future work, I will seek to further extend the trend-aware algorithm and the calibration scheme for a robust post-processing of diverse variables.

Developing reliable tools in post-processing GCM forecasts as well as linking the post-processed forecasts with impact models (e.g. hydrological and crop models) are vital for delivering timely forecasts to practical applications (Schepen et al., 2020a). Integrating the trend-aware forecast-calibration model into the end-to-end forecast workflow may produce more user-oriented

forecasts for water resource management, agriculture, and other climate-sensitive sectors in a changing climate. Future avenue will assess the merit of the trend-aware method and resulting forecasts for end applications.

6.4 Highlights and Concluding Remarks

Prior to this work, it has been documented that seasonal climate forecasts archived from the past and current generations of the GCM seasonal forecasting systems are not capable of reproducing the underlying climate trends in some regions. Several studies ascribed this issue to the unrealistic representation of the greenhouse gas emission (GHG) level. However, recent development has embedded observation-based time-varying GHG components into the climate models for forecasting sub-seasonal and seasonal climate variables, but the trend mismatch problem remains unsolved. Compared to sustained efforts to tackle long-standing modelling issues, statistically embedding historical trends into the GCM outputs is a more straightforward and easier to implement way to overcome the problem.

This research developed a trend-aware forecast calibration method to resolve the trend mismatch issue in climate outputs of GCM sub-seasonal and seasonal forecasting systems. Chapter 3 compared two model variations: BJP-t and BJP-ti. The BJP-t model inferred trends entirely from the available data records by using the non-informative priors for trend parameters. In recognition of trend uncertainty, the BJP-ti model moderated the inferred trends by using the informative priors for trend parameters. The BJP-ti model was shown as a more robust approach through comprehensive evaluations of forecast trends and forecast attributes. In the following Chapter 4 and Chapter 5, the trend-aware BJP-ti model was further improved and demonstrated to be a powerful tool for reducing trend disparity between model forecasts and observations, while making forecasts skillful and reliable. Overall, I recommend to the forecast communities the use of the trend-aware BJP-ti model that treats trend uncertainty.

The benefits of the trend-aware method for real-time forecasts could be significant. By incorporating the trend information inferred from historical climatology, operational sub-seasonal and seasonal climate forecasts will better reflect the changes in the non-stationary climate system. In regions of strong historical climate trends, real-time forecasts after trend-aware calibration are expected to be more skillful in predicting extreme events. The trend-aware calibrated forecasts should be valuable for decision-making and assisting climate-sensitive communities in issuing proactive alerts.

Finally, I would like to advocate an integration of the trend-aware methodology into sub-seasonal and seasonal forecast delivering systems, as well as the development of a rigorous and compatible end-to-end framework for application purposes under climate change.

References

Allan, R. P., Barlow, M., Byrne, M. P., Cherchi, A., Douville, H., Fowler, H. J., Gan, T. Y., Pendergrass, A. G., Rosenfeld, D., Swann, A. L. S., Wilcox, L. J., and Zolina, O. (2020). Advances in understanding large-scale responses of the water cycle to climate change. *Annals of the New Your Academy of Sciences, 1472*(1), 49-75. doi:https://doi.org/10.1111/nyas.14337

An-Vo, D., Mushtaq, S., Reardon-Smith, K., Kouadio, L., Attard, S., Cobon, D., and Stone, R. (2019). Value of seasonal forecasting for sugarcane farm irrigation planning. *European Journal of Agronomy, 104*, 37-48. doi:https://doi.org/10.1016/j.eja.2019.01.005

Anghileri, D., Monhart, S., Zhou, C., Bogner, K., Castelletti, A., Burlando, P., and Zappa, M. (2019). The Value of Subseasonal Hydrometeorological Forecasts to Hydropower Operations: How Much Does Preprocessing Matter? *Water Resources Research, 55*(12), 10159-10178. doi:https://doi.org/10.1029/2019WR025280

Ardilouze, C., Batté, L., Bunzel, F., Decremer, D., Déqué, M., Doblas-Reyes, F. J., Douville, H., Fereday, D., Guemas, V., and MacLachlan, C. (2017). Multi-model assessment of the impact of soil moisture initialization on mid-latitude summer predictability. *Climate Dynamics, 49*(11), 3959-3974. doi:http://doi.org/10.1007/s00382-017-3555-7

Baggett, C. F., Nardi, K. M., Childs, S. J., Zito, S. N., Barnes, E. A., and Maloney, E. D. (2018). Skillful subseasonal forecasts of weekly tornado and hail activity using the Madden-Julian Oscillation. *Journal of Geophysical Research: Atmospheres, 123*(22), 12,661-612,675. doi:https://doi.org/10.1029/2018JD029059

Baker, S. A., Wood, A. W., and Rajagopalan, B. (2019). Developing subseasonal to seasonal climate forecast products for hydrology and water management. *JAWRA Journal of the American Water Resources Association, 55*(4), 1024-1037. doi:https://doi.org/10.1111/1752-1688.12746

Barnston, A. G., Li, S., Mason, S. J., DeWitt, D. G., Goddard, L., and Gong, X. (2010). Verification of the first 11 years of IRI's seasonal climate forecasts. *Journal of Applied Meteorology and Climatology, 49*(3), 493-520. doi:https://doi.org/10.1175/2009JAMC2325.1

Barnston, A. G., Tippett, M. K., van den Dool, H. M., and Unger, D. A. (2015). Toward an Improved Multimodel ENSO Prediction. *Journal of Applied Meteorology and Climatology, 54*(7), 1579-1595. doi:http://doi.org/10.1175/Jamc-D-14-0188.1

Barnston, A. G., and van den Dool, H. M. (1993). A degeneracy in cross-validated skill in regression-based forecasts. *Journal of Climate, 6*(5), 963-977. doi:https://doi.org/10.1175/1520-0442(1993)006<0963:ADICVS>2.0.CO;2

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature, 525*(7567), 47-55. doi:https://doi.org/10.1038/nature14956

Bennett, J. C., Wang, Q. J., Robertson, D. E., Schepen, A., Li, M., and Michael, K. (2017). Assessment of an ensemble seasonal streamflow forecasting system for Australia.

Hydrology and Earth System Sciences, 21(12), 6007-6030. doi:https://doi.org/10.5194/hess-21-6007-2017

Bhend, J., and Whetton, P. (2015). Evaluation of simulated recent climate change in Australia. *Australian Meteorological and Oceanographic Journal, 65*, 4-18.

Blanchard-Wrigglesworth, E., Bitz, C., and Holland, M. (2011). Influence of initial conditions and climate forcing on predicting Arctic sea ice. *Geophysical Research Letters, 38*(18). doi:https://doi.org/10.1029/2011GL048807

Boer, G. J. (2009). Climate trends in a seasonal forecasting system. *Atmosphere-ocean, 47*(2), 123-138. doi:https://doi.org/10.3137/AO1002.2009

Brocca, L., Pellarin, T., Crow, W. T., Ciabatta, L., Massari, C., Ryu, D., Su, C. H., Rüdiger, C., and Kerr, Y. (2016). Rainfall estimation by inverting SMOS soil moisture estimates: A comparison of different methods over Australia. *Journal of Geophysical Research: Atmospheres, 121*(20), 12,062-012,079. doi:https://doi.org/10.1002/2016JD025382

Brown, J. N., Hochman, Z., Holzworth, D., and Horan, H. (2018). Seasonal climate forecasts provide more definitive and accurate crop yield predictions. *Agricultural and Forest Meteorology, 260*, 247-254. doi:https://doi.org/10.1016/j.agrformet.2018.06.001

Bruno Soares, M., Daly, M., and Dessai, S. (2018). Assessing the value of seasonal climate forecasts for decision-making. *Wiley Interdisciplinary Reviews: Climate Change, 9*(4), e523. doi:https://doi.org/10.1002/wcc.523

Butler, A., Charlton-Perez, A., Domeisen, D. I., Garfinkel, C., Gerber, E. P., Hitchcock, P., Karpechko, A. Y., Maycock, A. C., Sigmond, M., and Simpson, I. (2019). Sub-seasonal predictability and the stratosphere. *Sub-seasonal to seasonal prediction*, 223-241. doi:https://doi.org/10.1016/B978-0-12-811714-9.00011-5

Cai, M., Shin, C.-S., Van den Dool, H., Wang, W., Saha, S., and Kumar, A. (2009). The role of long-term trends in seasonal predictions: Implication of global warming in the NCEP CFS. *Weather and Forecasting, 24*(4), 965-973. doi:https://doi.org/10.1175/2009WAF2222231.1

Caloiero, T. (2017). Trend of monthly temperature and daily extreme temperature during 1951-2012 in New Zealand. *Theoretical and Applied Climatology, 129*(1-2), 111-127. doi:http://doi.org/10.1007/s00704-016-1764-3

Caporali, E., Lompi, M., Pacetti, T., Chiarello, V., and Fatichi, S. (2021). A review of studies on observed precipitation trends in Italy. *International Journal of Climatology, 41*, E1-E25. doi:https://doi.org/10.1002/joc.6741

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W. J., Johns, T., and Krinner, G. (2013). Long-term climate change: projections, commitments and irreversibility. In *Climate Change 2013-The Physical Science Basis: Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1029-1136): Cambridge University Press.

CSIRO and Australian Government Bureau of Meteorology. (2018). *State of the Climate 2018*. Retrieved from http://www.bom.gov.au/state-of-the-climate/State-of-the-Climate-2016.pdf.

CSIRO and Australian Government Bureau of Meteorology. (2020). *State of the Climate 2020*. Retrieved from http://www.bom.gov.au/state-of-the-climate/documents/State-of-the-Climate-2020.pdf

de Barros Soares, D., Lee, H., Loikith, P. C., Barkhordarian, A., and Mechoso, C. R. (2017). Can significant trends be detected in surface air temperature and precipitation over South America in recent decades? *International Journal of Climatology, 37*(3), 1483-1493. doi:http://doi.org/10.1002/joc.4792

DelSole, T., and Tippett, M. K. (2016). Forecast comparison based on random walks. *Monthly Weather Review, 144*(2), 615-626. doi:https://doi.org/10.1175/MWR-D-15-0218.1

Dey, R., Lewis, S. C., Arblaster, J. M., and Abram, N. J. (2019). A review of past and projected changes in Australia's rainfall. *Wiley Interdisciplinary Reviews: Climate Change, 10*(3), e577. doi:https://doi.org/10.1002/wcc.577

Dinpashoh, Y., Jahanbakhsh-Asl, S., Rasouli, A., Foroughi, M., and Singh, V. (2019). Impact of climate change on potential evapotranspiration (case study: west and NW of Iran). *Theoretical and Applied Climatology, 136*(1), 185-201. doi:https://doi.org/10.1007/s00704-018-2462-0

Director, H. M., Raftery, A. E., and Bitz, C. M. (2019). Probabilistic Forecasting of the Arctic Sea Ice Edge with Contour Modeling. *arXiv preprint arXiv:1908.09377*.

Dirkson, A., Merryfield, W. J., and Monahan, A. H. (2019). Calibrated probabilistic forecasts of Arctic sea ice concentration. *Journal of Climate, 32*(4), 1251-1271. doi:https://doi.org/10.1175/JCLI-D-18-0224.1

Doblas-Reyes, F. J., Garcia-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R. L. (2013). Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdisciplinary Reviews-Climate Change, 4*(4), 245-268. doi:http://doi.org/10.1002/wcc.217

Doblas-Reyes, F. J., Hagedorn, R., and Palmer, T. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus A: Dynamic Meteorology and Oceanography, 57*(3), 234-252.

Doblas-Reyes, F. J., Hagedorn, R., Palmer, T. N., and Morcrette, J. J. (2006). Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts. *Geophysical Research Letters, 33*(7). doi:http://doi.org/10.1029/2005gl025061

Duan, C., Wang, P., Cao, W., Wang, X., Wu, R., and Cheng, Z. (2021). Improving the Spring Air Temperature Forecast Skills of BCC_CSM1.1 (m) by Spatial Disaggregation and Bias Correction: Importance of Trend Correction. *Atmosphere, 12*(9), 1143. doi:https://doi.org/10.3390/atmos12091143

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., and Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters, 41*(15), 5620-5628. doi:https://doi.org/10.1002/2014GL061146

ECMWF. (2021). ECMWF model description. Retrieved from https://confluence.ecmwf.int/display/S2S/ECMWF+model+description

Esquivel-Muelbert, A., Baker, T. R., Dexter, K. G., Lewis, S. L., Brienen, R. J. W., Feldpausch, T. R., Lloyd, J., Monteagudo-Mendoza, A., Arroyo, L., Álvarez-Dávila, E., Higuchi, N., Marimon, B. S., Marimon-Junior, B. H., Silveira, M., Vilanova, E., Gloor, E., Malhi, Y., Chave, J., Barlow, J., Bonal, D., et al. (2019). Compositional response of Amazon forests to climate change. *Global Change Biology, 25*(1), 39-56. doi:https://doi.org/10.1111/gcb.14413

Fadrique, B., Báez, S., Duque, Á., Malizia, A., Blundo, C., Carilla, J., Osinaga-Acosta, O., Malizia, L., Silman, M., and Farfán-Ríos, W. (2018). Widespread but heterogeneous responses of Andean forests to climate change. *Nature, 564*(7735), 207-212. doi:https://doi.org/10.1038/s41586-018-0715-9

Fan, Y., Krasnopolsky, V., van den Dool, H., Wu, C.-Y., and Gottschalck, J. (2021). Using Artificial Neural Networks to Improve CFS Week 3-4 Precipitation and 2-Meter Air Temperature Forecasts. *Weather and Forecasting*. doi:https://doi.org/10.1175/WAF-D-20-0014.1

Fawcett, R. J. B., Trewin, B. C., Braganza, K., Smalley, R. J., Jovanovic, B., and Jones, D. A. (2012). *On the sensitivity of Australian temperature trends and variability to analysis methods and observation networks*. (050). Melbourne: Centre for Australian Weather and Climate Research Retrieved from http://cawcr.gov.au/technical-reports/CTR_050.pdf.

Feikema, P., Wang, Q., Zhou, S., Shin, D., Robertson, D., Schepen, A., Lerat, J., Bennett, J., Tuteja, N., and Jayasuriya, D. (2018). Service and Research on Seasonal Streamflow Forecasting in Australia. In *Bridging Science and Policy Implication for Managing Climate Extremes* (pp. 157-175): World Scientific.

Ferranti, L., Corti, S., and Janousek, M. (2015). Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society, 141*(688), 916-924. doi:https://doi.org/10.1002/qj.2411

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M. (2013). Evaluation of Climate Models. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex , & P. M. Midgley (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Funk, C., Shukla, S., Thiaw, W. M., Rowland, J., Hoell, A., McNally, A., Husak, G., Novella, N., Budde, M., and Peters-Lidard, C. (2019). Recognizing the famine early warning systems network: Over 30 years of drought early warning science advances and partnerships promoting global food security. *Bulletin of the American Meteorological Society, 100*(6), 1011-1027. doi:https://doi.org/10.1175/BAMS-D-17-0233.1

Gebrechorkos, S. H., Hülsmann, S., and Bernhofer, C. (2019). Long-term trends in rainfall and temperature using high-resolution climate datasets in East Africa. *Scientific Reports, 9*(1), 1-9. doi:https://doi.org/10.1038/s41598-019-47933-8

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis* (Third edition ed.): CRC press.

Ghasemi, A. R. (2015). Changes and trends in maximum, minimum and mean temperature series in Iran. *Atmospheric Science Letters, 16*(3), 366-372. doi:http://doi.org/10.1002/asl2.569

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 69*, 243-268. doi:http://doi.org/10.1111/j.1467-9868.2007.00587.x

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review, 133*(5), 1098-1118. doi:https://doi.org/10.1175/MWR2904.1

Gong, G., Entekhabi, D., and Cohen, J. (2002). A large-ensemble model study of the wintertime AO–NAO and the role of interannual snow perturbations. *Journal of Climate, 15*(23), 3488-3499. doi:https://doi.org/10.1175/1520-0442(2002)015<3488:ALEMSO>2.0.CO;2

Grillakis, M., Koutroulis, A., and Tsanis, I. (2018). Improving Seasonal Forecasts for Basin Scale Hydrological Applications. *Water, 10*(11). doi:https://doi.org/10.3390/w10111593

Hackert, E. C., Kovach, R. M., Busalacchi, A. J., and Ballabrera-Poy, J. (2019). Impact of Aquarius and SMAP satellite sea surface salinity observations on coupled El Niño/Southern Oscillation forecasts. *Journal of Geophysical Research: Oceans, 124*(7), 4546-4556. doi:https://doi.org/10.1029/2019JC015130

Han, E., Ines, A. V., and Baethgen, W. E. (2017). Climate-Agriculture-Modeling and Decision Tool (CAMDT): A software framework for climate risk management in agriculture. *Environmental Modelling & Software, 95*, 102-114. doi:http://doi.org/10.1016/j.envsoft.2017.06.024

Hao, Z., Singh, V. P., and Xia, Y. (2018). Seasonal drought prediction: advances, challenges, and future prospects. *Reviews of Geophysics, 56*(1), 108-141. doi:https://doi.org/10.1002/2016RG000549

Hartmann, D. L., Klein Tank, A. M. G., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y. A. R., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., Soden, B. J., Thorne, P. W., Wild, M., and Zhai, P. M. (2013). Observations: Atmosphere and surface. In *Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Vol. 9781107057999, pp. 159-254): Cambridge University Press.

Hemri, S., Bhend, J., Liniger, M. A., Manzanas, R., Siegert, S., Stephenson, D. B., Gutiérrez, J. M., Brookshaw, A., and Doblas-Reyes, F. J. (2020). How to create an operational multi-model of seasonal forecasts? *Climate Dynamics, 55*(5), 1141-1157. doi:https://doi.org/10.1007/s00382-020-05314-2

Hermanson, L., Ren, H.-L., Vellinga, M., Dunstone, N., Hyder, P., Ineson, S., Scaife, A., Smith, D., Thompson, V., and Tian, B. (2018). Different types of drifts in two seasonal forecast

systems and their dependence on ENSO. *Climate Dynamics, 51*(4), 1411-1426. doi:https://doi.org/10.1007/s00382-017-3962-9

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting, 15*(5), 559-570. doi:https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Huang, B., Shin, C.-S., and Kumar, A. (2019). Predictive skill and predictable patterns of the US seasonal precipitation in CFSv2 reforecasts of 60 years (1958–2017). *Journal of Climate, 32*(24), 8603-8637. doi:https://doi.org/10.1175/JCLI-D-19-0230.1

Hudson, D., Alves, O., Hendon, H. H., Lim, E. P., Liu, G. Q., Luo, J. J., MacLachlan, C., Marshall, A. G., Shi, L., Wang, G. M., Wedd, R., Young, G., Zhao, M., and Zhou, X. B. (2017). ACCESS-S1: The new Bureau of Meteorology multi-week to seasonal prediction system. *Journal of Southern Hemisphere Earth Systems Science, 67*(3), 132-159. doi:http://doi.org/10.22499/3.6703.001

Hudson, D., Alves, O., Shi, L., and Young, G. (2018). *Improved skill for regional climate in the ACCESS-based POAMA model*. Retrieved from Sydney, Australia: https://ausveg.com.au/app/uploads/technical-insights/VG13092.pdf

Hudson, D., Marshall, A. G., Yin, Y., Alves, O., and Hendon, H. H. (2013). Improving intraseasonal prediction with a new ensemble generation strategy. *Monthly Weather Review, 141*(12), 4429-4449. doi:https://doi.org/10.1175/MWR-D-13-00059.1

Hwang, S., and Graham, W. D. (2013). Development and comparative evaluation of a stochastic analog method to downscale daily GCM precipitation. *Hydrology and Earth System Sciences, 17*(11), 4481-4502. doi:https://doi.org/10.5194/hess-17-4481-2013

Inness, A., Baier, F., Benedetti, A., Bouarar, I., Chabrillat, S., Clark, H., Clerbaux, C., Coheur, P., Engelen, R., and Errera, Q. (2013). The MACC reanalysis: an 8 yr data set of atmospheric composition. *Atmospheric Chemistry and Physics, 13*(8), 4073-4109. doi:http://doi.org/10.5194/acp-13-4073-2013

IPCC. (2018). *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*. Retrieved from https://www.ipcc.ch/site/assets/uploads/sites/2/2019/06/SR15_Full_Report_High_Res.pdf

Irving, D. B., Whetton, P., and Moise, A. F. (2012). Climate projections for Australia: a first glance at CMIP5. *Australian Meteorological and Oceanographic Journal, 62*(4), 211-225. doi:http://doi.org/10.22499/2.6204.003

Jacob, D., Teichmann, C., Sobolowski, S., Katragkou, E., Anders, I., Belda, M., Benestad, R., Boberg, F., Buonomo, E., and Cardoso, R. M. (2020). Regional climate downscaling over Europe: perspectives from the EURO-CORDEX community. *Regional environmental change, 20*(2), 1-20. doi:https://doi.org/10.1007/s10113-020-01606-9

Jha, P. K., Athanasiadis, P., Gualdi, S., Trabucco, A., Mereu, V., Shelia, V., and Hoogenboom, G. (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. doi:https://doi.org/10.1016/j.agrformet.2018.11.029

Jia, G., Shevliakova, E., Artaxo, P., De Noblet-Ducoudré, N., Houghton, R., House, J., Kitajima, K., Lennard, C., Popp, A., Sirin, A., Sukumar, R., and Verchot, L. (2019). Land–climate interactions. In P. Bernier, J. C. Espinoza, & S. Semenov (Eds.), *Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems*

Jia, X., Lee, J., and Lin, H. (2014). Interdecadal change in the Northern Hemisphere seasonal climate prediction skill: part II. predictability and prediction skill. *Climate Dynamics, 43*, 1611-1630. doi:https://doi.org/10.1007/s00382-014-2084-x

Jia, X. J., and Lin, H. (2013). The Possible Reasons for the Misrepresented Long-Term Climate Trends in the Seasonal Forecasts of HFP2. *Monthly Weather Review, 141*(9), 3154-3169. doi:http://doi.org/10.1175/Mwr-D-12-00302.1

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., and Balsamo, G. (2019). SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development, 12*(3). doi:https://doi.org/10.5194/gmd-12-1087-2019

Jones, D. A., Wang, W., and Fawcett, R. (2009). High-quality spatial climate data-sets for Australia. *Australian Meteorological and Oceanographic Journal, 58*(4), 233-248. doi:http://doi.org/10.22499/2.5804.003

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., Bonan, G., Cescatti, A., Chen, J., and De Jeu, R. (2010). Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature, 467*(7318), 951-954. doi:https://doi.org/10.1038/nature09396

Kendall, M. G. (1975). *Rank Correlation Methods* (4th ed.). London Charles Griffin.

Khajehei, S., and Moradkhani, H. (2017). Towards an improved ensemble precipitation forecast: A probabilistic post-processing approach. *Journal of Hydrology, 546*, 476-489. doi:https://doi.org/10.1016/j.jhydrol.2017.01.026

Kharin, V. V., Boer, G. J., Merryfield, W. J., Scinocca, J. F., and Lee, W. S. (2012). Statistical adjustment of decadal predictions in a changing climate. *Geophysical Research Letters, 39*. doi:http://doi.org/10.1029/2012gl052647

Kharin, V. V., Merryfield, W. J., Boer, G. J., and Lee, W. S. (2017). A Postprocessing Method for Seasonal Forecasts Using Temporally and Spatially Smoothed Statistics. *Monthly Weather Review, 145*(9), 3545-3561. doi:http://doi.org/10.1175/Mwr-D-16-0337.1

Kirtman, B., Power, S. B., Adedoyin, A. J., Boer, G. J., Bojariu, R., Camilloni, I., Doblas-Reyes, F., Fiore, A. M., Kimoto, M., Meehl, G., Prather, M., Sarr, A., Schär, C., Sutton, R., van Oldenborgh, G. J., Vecchi, G., and Wang, H.-J. (2013). Near-term climate change: projections and predictability. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, & P. M. Midgley (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*.

Kirtman, B. P., Min, D., Infanti, J. M., Kinter III, J. L., Paolino, D. A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M. P., and Becker, E. (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*(4), 585-601. doi:https://doi.org/10.1175/BAMS-D-12-00050.1

Klemm, T., and McPherson, R. A. (2017). The development of seasonal climate forecasting for agricultural producers. *Agricultural and Forest Meteorology, 232*, 384-399. doi:https://doi.org/10.1016/j.agrformet.2016.09.005

Kolachian, R., and Saghafian, B. (2019). Deterministic and probabilistic evaluation of raw and post processed sub-seasonal to seasonal precipitation forecasts in different precipitation regimes. *Theoretical and Applied Climatology, 137*(1-2), 1479-1493. doi:https://doi.org/10.1007/s00704-018-2680-5

Krakauer, N. Y. (2019). Temperature trends and prediction skill in NMME seasonal forecasts. *Climate Dynamics, 53*, 7201–7213. doi:https://doi.org/10.1007/s00382-017-3657-2

Krikken, F., Schmeits, M., Vlot, W., Guemas, V., and Hazeleger, W. (2016). Skill improvement of dynamical seasonal Arctic sea ice forecasts. *Geophysical Research Letters, 43*(10), 5124-5132. doi:https://doi.org/10.1002/2016GL068462

Krishnamurthy, V. (2019). Predictability of weather and climate. *Earth and Space Science, 6*(7), 1043-1056. doi:https://doi.org/10.1029/2019EA000586

Kumar, D., Kodra, E., and Ganguly, A. R. (2014). Regional and seasonal intercomparison of CMIP3 and CMIP5 climate model ensembles for temperature and precipitation. *Climate Dynamics, 43*(9-10), 2491-2518. doi:https://doi.org/10.1007/s00382-014-2070-3

Kumar, S., Merwade, V., Kinter III, J. L., and Niyogi, D. (2013). Evaluation of temperature and precipitation trends and long-term persistence in CMIP5 twentieth-century climate simulations. *Journal of Climate, 26*(12), 4168-4185. doi:https://doi.org/10.1175/JCLI-D-12-00259.1

Kushnir, Y., Scaife, A. A., Arritt, R., Balsamo, G., Boer, G., Doblas-Reyes, F., Hawkins, E., Kimoto, M., Kolli, R. K., Kumar, A., Matei, D., Matthes, K., Müller, W. A., O'Kane, T., Perlwitz, J., Power, S., Raphael, M., Shimpo, A., Smith, D., Tuma, M., et al. (2019). Towards operational predictions of the near-term climate. *Nature Climate Change, 9*(2), 94-101. doi:https://doi.org/10.1038/s41558-018-0359-7

Lausier, A. M., and Jain, S. (2018). Overlooked Trends in Observed Global Annual Precipitation Reveal Underestimated Risks. *Scientific Reports, 8*(1), 1-7. doi:https://doi.org/10.1038/s41598-018-34993-5

Law, D. J., Adams, H. D., Breshears, D. D., Cobb, N. S., Bradford, J. B., Zou, C. B., Field, J. P., Gardea, A. A., Williams, A. P., and Huxman, T. E. (2019). Bioclimatic envelopes for individual demographic events driven by extremes: plant mortality from drought and warming. *International Journal of Plant Sciences, 180*(1), 53-62. doi:https://doi.org/10.1086/700702

Li, L., Zhang, Y., Liu, Q., Ding, M., and Mondal, P. P. (2019). Regional differences in shifts of temperature trends across China between 1980 and 2017. *International Journal of Climatology, 39*(3), 1157-1165. doi:https://doi.org/10.1002/joc.5868

Li, M., and Jin, H. (2020). Development of a postprocessing system of daily rainfall forecasts for seasonal crop prediction in Australia. *Theoretical and Applied Climatology, 141*, 1331-1349. doi:http://doi.org/10.1007/s00704-020-03268-3

Li, W., Chen, J., Li, L., Chen, H., Liu, B., Xu, C.-Y., and Li, X. (2019). Evaluation and bias correction of S2S precipitation for hydrological extremes. *Journal of Hydrometeorology, 20*(9), 1887-1906. doi:https://doi.org/10.1175/JHM-D-19-0042.1

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., and Di, Z. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews: Water, 4*(6), e1246. doi:https://doi.org/10.1002/wat2.1246

Li, X., Gollan, G., Greatbatch, R. J., and Lu, R. (2018). Intraseasonal variation of the East Asian summer monsoon associated with the Madden–Julian oscillation. *Atmospheric Science Letters, 19*(4), e794. doi:https://doi.org/10.1002/asl.794

Li, Y., Wu, Z., He, H., Wang, Q. J., Xu, H., and Lu, G. (2020). Post-processing sub-seasonal precipitation forecasts at various spatiotemporal scales across China during boreal summer monsoon. *Journal of Hydrology*, 125742. doi:https://doi.org/10.1016/j.jhydrol.2020.125742

Lim, Y., Son, S.-W., and Kim, D. (2018). MJO prediction skill of the subseasonal-to-seasonal prediction models. *Journal of Climate, 31*(10), 4075-4094. doi:https://doi.org/10.1175/JCLI-D-17-0545.1

Lin, H., Frederiksen, J., Straus, D., and Stan, C. (2019). Tropical-extratropical interactions and teleconnections. In *Sub-seasonal to seasonal prediction* (pp. 143-164): Elsevier.

Livada, I., Synnefa, A., Haddad, S., Paolini, R., Garshasbi, S., Ulpiani, G., Fiorito, F., Vassilakopoulou, K., Osmond, P., and Santamouris, M. (2019). Time series analysis of ambient air-temperature during the period 1970-2016 over Sydney, Australia. *Science of the Total Environment, 648*, 1627-1638. doi:http://doi.org/10.1016/j.scitotenv.2018.08.144

Livezey, R. E., and Timofeyeva, M. M. (2008). The first decade of long-lead US seasonal forecasts: Insights from a skill analysis. *Bulletin of the American Meteorological Society, 89*(6), 843-854. doi:https://doi.org/10.1175/2008BAMS2488.1

Lobell, D. B., Schlenker, W., and Costa-Roberts, J. (2011). Climate trends and global crop production since 1980. *Science, 333*(6042), 616-620. doi:http://doi.org/10.1126/science.1204531

Lucatero, D., Madsen, H., Refsgaard, J. C., Kidmose, J., and Jensen, K. H. (2018). On the skill of raw and post-processed ensemble seasonal meteorological forecasts in Denmark. *Hydrology and Earth System Sciences, 22*(12), 6591-6609. doi:https://doi.org/10.5194/hess-22-6591-2018

Luo, L. F., and Wood, E. F. (2008). Use of Bayesian Merging Techniques in a Multimodel Seasonal Hydrologic Ensemble Prediction System for the Eastern United States. *Journal of Hydrometeorology, 9*(5), 866-884. doi:http://doi.org/10.1175/2008jhm980.1

Ma, H.-Y., Zhou, C., Zhang, Y., Klein, S. A., Zelinka, M. D., Zheng, X., Xie, S., Chen, W.-T., and Wu, C.-M. (2021). A multi-year short-range hindcast experiment for evaluating

climate model moist processes from diurnal to interannual time scales. *Geoscientific Model Development, 14*, 73–90. doi:https://doi.org/10.5194/gmd-14-73-2021

Mann, H. B. (1945). Nonparametric Tests against Trend. *Econometrica, 13*(3), 245-259. doi:http://doi.org/10.2307/1907187

Manzanas, R., Gutiérrez, J. M., Bhend, J., Hemri, S., Doblas-Reyes, F. J., Torralba, V., Penabad, E., and Brookshaw, A. (2019). Bias adjustment and ensemble recalibration methods for seasonal forecasting: A comprehensive intercomparison using the C3S dataset. *Climate Dynamics, 53*(3-4), 1287-1305. doi:https://doi.org/10.1007/s00382-019-04640-4

Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., Collins, D. C., Dirmeyer, P. A., Ferranti, L., Johnson, N. C., and Jones, J. (2020). Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bulletin of the American Meteorological Society, 101*(5), E608-E625. doi:https://doi.org/10.1175/BAMS-D-18-0326.1

Marshall, A., Hudson, D., Wheeler, M., Alves, O., Hendon, H., Pook, M., and Risbey, J. (2014). Intra-seasonal drivers of extreme heat over Australia in observations and POAMA-2. *Climate Dynamics, 43*(7-8), 1915-1937. doi:https://doi.org/10.1007/s00382-013-2016-1

Marshall, A. G., and Scaife, A. A. (2009). Impact of the QBO on surface winter climate. *Journal of Geophysical Research: Atmospheres, 114*(D18). doi:https://doi.org/10.1029/2009JD011737

Marshall, A. G., and Scaife, A. A. (2010). Improved predictability of stratospheric sudden warming events in an atmospheric general circulation model with enhanced stratospheric resolution. *Journal of Geophysical Research: Atmospheres, 115*(D16). doi:http://doi.org/10.1029/2009JD012643

Matheson, J. E., and Winkler, R. L. (1976). Scoring Rules for Continuous Probability Distributions. *Management Science, 22*(10), 1087-1096. doi:http://doi.org/10.1287/mnsc.22.10.1087

McKinnon, K. A., Rhines, A., Tingley, M., and Huybers, P. (2016). Long-lead predictions of eastern United States hot days from Pacific sea surface temperatures. *Nature Geoscience, 9*(5), 389-394. doi:https://doi.org/10.1038/ngeo2687

Meinshausen, M., Vogel, E., Nauels, A., Lorbacher, K., Meinshausen, N., Etheridge, D. M., Fraser, P. J., Montzka, S. A., Rayner, P. J., Trudinger, C. M., Krummel, P. B., Beyerle, U., Canadell, J. G., Daniel, J. S., Enting, I. G., Law, R. M., Lunder, C. R., O'Doherty, S., Prinn, R. G., Reimann, S., et al. (2017). Historical greenhouse gas concentrations for climate modelling (CMIP6). *Geoscientific Model Development, 10*(5), 2057-2116. doi:http://doi.org/10.5194/gmd-10-2057-2017

Ménégoz, M., Bilbao, R., Bellprat, O., Guemas, V., and Doblas-Reyes, F. J. (2018). Forecasting the climate response to volcanic eruptions: Prediction skill related to stratospheric aerosol forcing. *Environmental Research Letters, 13*(6), 064022. doi:https://doi.org/10.1088/1748-9326/aac4db

Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A., Danabasoglu, G., Dirmeyer, P. A., Doblas-Reyes, F. J., and Domeisen, D. I. (2020). Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the American*

Meteorological Society, 101(6), E869-E896. doi:https://doi.org/10.1175/BAMS-D-19-0037.1

Misios, S., Gray, L. J., Knudsen, M. F., Karoff, C., Schmidt, H., and Haigh, J. D. (2019). Slowdown of the Walker circulation at solar cycle maximum. *Proceedings of the National Academy of Sciences, 116*(15), 7186-7191. doi:https://doi.org/10.1073/pnas.1815060116

Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C., and Liniger, M. A. (2018). Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *Journal of Geophysical Research: Atmospheres, 123*(15), 7999-8016. doi:https://doi.org/10.1029/2017JD027923

Mulholland, D. P., Laloyaux, P., Haines, K., and Balmaseda, M. A. (2015). Origin and impact of initialization shocks in coupled atmosphere–ocean forecasts. *Monthly Weather Review, 143*(11), 4631-4644. doi:https://doi.org/10.1175/MWR-D-15-0076.1

Nageswararao, M. M., Mohanty, U. C., Prasad, S. K., Osuri, K. K., and Ramakrishna, S. S. V. S. (2016). Performance evaluation of NCEP climate forecast system for the prediction of winter temperatures over India. *Theoretical and Applied Climatology, 126*(3-4), 437-451. doi:http://doi.org/10.1007/s00704-015-1588-6

Narapusetty, B., Collins, D., Murtugudde, R., Gottschalck, J., and Peters-Lidard, C. (2018). Bias correction to improve the skill of summer precipitation forecasts over contiguous United States by the North American Multi-Model Ensemble system. *Atmospheric Science Letters, 19*, e818. doi:https://doi.org/10.1002/asl.818

Narapusetty, B., DelSole, T., and Tippett, M. K. (2009). Optimal estimation of the climatological mean. *Journal of Climate, 22*(18), 4845-4859. doi:https://doi.org/10.1175/2009JCLI2944.1

Nury, A. H., Sharma, A., Marshall, L., and Mehrotra, R. (2019). Characterising uncertainty in precipitation downscaling using a Bayesian approach. *Advances in Water Resources, 129*, 189-197. doi:https://doi.org/10.1016/j.advwatres.2019.05.018

O'Kane, T. J., Sandery, P. A., Monselesan, D. P., Sakov, P., Chamberlain, M. A., Matear, R. J., Collier, M. A., Squire, D. T., and Stevens, L. (2019). Coupled data assimilation and ensemble initialization with application to multiyear ENSO prediction. *Journal of Climate, 32*(4), 997-1024. doi:https://doi.org/10.1175/JCLI-D-18-0189.1

Ogutu, G. E. O., Franssen, W. H. P., Supit, I., Omondi, P., and Hutjes, R. W. A. (2018). Probabilistic maize yield prediction over East Africa using dynamic ensemble seasonal climate forecasts. *Agricultural and Forest Meteorology, 250*, 243-261. doi:http://doi.org/10.1016/j.agrformet.2017.12.256

Orsolini, Y. J., Kindem, I. T., and Kvamsto, N. G. (2011). On the potential impact of the stratosphere upon seasonal dynamical hindcasts of the North Atlantic Oscillation: a pilot study. *Climate Dynamics, 36*(3-4), 579-588. doi:http://doi.org/10.1007/s00382-009-0705-6

Parton, K. A., Crean, J., and Hayman, P. (2019). The value of seasonal climate forecasts for Australian agriculture. *Agricultural Systems, 174*, 1-10. doi:https://doi.org/10.1016/j.agsy.2019.04.005

Pasternack, A., Bhend, J., Liniger, M. A., Rust, H. W., Muller, W. A., and Ulbrich, U. (2018). Parametric decadal climate forecast recalibration (DeFoReSt 1.0). *Geoscientific Model Development, 11*(1), 351-368. doi:http://doi.org/10.5194/gmd-11-351-2018

Pasternack, A., Grieger, J., Rust, H. W., and Ulbrich, U. (2021). Recalibrating Decadal Climate Predictions–What is an adequate model for the drift? *Geoscientific Model Development Discussions, 14*, 4335–4355. doi:https://doi.org/10.5194/gmd-14-4335-2021

Pechlivanidis, I., Crochemore, L., Rosberg, J., and Bosshard, T. (2020). What are the key drivers controlling the quality of seasonal streamflow forecasts? *Water Resources Research, 56*(6), e2019WR026987. doi:https://doi.org/10.1029/2019WR026987

Peel, M. C., Finlayson, B. L., and McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences, 11*(5), 1633-1644.

Pegion, K., DelSole, T., Becker, E., and Cicerone, T. (2019). Assessing the fidelity of predictability estimates. *Climate Dynamics, 53*(12), 7251-7265. doi:http://doi.org/10.1007/s00382-017-3903-7

Pendergrass, A. G., Knutti, R., Lehner, F., Deser, C., and Sanderson, B. M. (2017). Precipitation variability increases in a warmer climate. *Scientific Reports, 7*(1), 1-9. doi:https://doi.org/10.1038/s41598-017-17966-y

Peng, S., Ding, Y., Wen, Z., Chen, Y., Cao, Y., and Ren, J. (2017). Spatiotemporal change and trend analysis of potential evapotranspiration over the Loess Plateau of China during 2011–2100. *Agricultural and Forest Meteorology, 233*, 183-194. doi:http://doi.org/10.1016/J.AGRFORMET.2016.11.129

Peng, T., Zhi, X., Ji, Y., Ji, L., and Tian, Y. (2020). Prediction Skill of Extended Range 2-m Maximum Air Temperature Probabilistic Forecasts Using Machine Learning Post-Processing Methods. *Atmosphere, 11*(8), 823. doi:http://doi.org/10.3390/atmos11080823

Peng, Z., Wang, Q., Bennett, J. C., Schepen, A., Pappenberger, F., Pokhrel, P., and Wang, Z. (2014). Statistical calibration and bridging of ECMWF System4 outputs for forecasting seasonal precipitation over China. *Journal of Geophysical Research: Atmospheres, 119*(12), 7116-7135. doi:https://doi.org/10.1002/2013JD021162

Penny, S., Bach, E., Bhargava, K., Chang, C. C., Da, C., Sun, L., and Yoshida, T. (2019). Strongly coupled data assimilation in multiscale media: Experiments using a quasi-geostrophic coupled model. *Journal of Advances in Modeling Earth Systems, 11*(6), 1803-1829. doi:https://doi.org/10.1029/2019MS001652

Penny, S. G., and Hamill, T. M. (2017). Coupled data assimilation for integrated earth system analysis and prediction. *Bulletin of the American Meteorological Society, 98*(7), ES169-ES172.

Peñuela, A., Hutton, C., and Pianosi, F. (2020). Assessing the value of seasonal hydrological forecasts for improving water resource management: insights from a pilot application in the UK. *Hydrology and Earth System Sciences, 24*(12), 6059-6073. doi:https://doi.org/10.5194/hess-24-6059-2020

Pineda, L. E., and Willems, P. (2016). Multisite Downscaling of Seasonal Predictions to Daily Rainfall Characteristics over Pacific-Andean River Basins in Ecuador and Peru Using a

Nonhomogeneous Hidden Markov Model. *Journal of Hydrometeorology, 17*(2), 481-498. doi:https://doi.org/10.1175/JHM-D-15-0040.1

Polade, S. D., Gershunov, A., Cayan, D. R., Dettinger, M. D., and Pierce, D. W. (2017). Precipitation in a warming world: Assessing projected hydro-climate changes in California and other Mediterranean climate regions. *Scientific Reports, 7*(1), 1-10. doi:http://doi.org/10.1038/s41598-017-11285-y

Prodhomme, C., Doblas-Reyes, F., Bellprat, O., and Dutra, E. (2016). Impact of land-surface initialization on sub-seasonal to seasonal forecasts over Europe. *Climate Dynamics, 47*(3-4), 919-935. doi:https://doi.org/10.1007/s00382-015-2879-4

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American statistical association, 92*(437), 179-191. doi:https://doi.org/10.1080/01621459.1997.10473615

Rahmat, S. N., Jayasuriya, N., and Bhuiyan, M. A. (2015). Precipitation trends in Victoria, Australia. *Journal of Water and Climate Change, 6*(2), 278-287. doi:https://doi.org/10.2166/wcc.2014.007

Ren, Y.-Y., Ren, G.-Y., Sun, X.-B., Shrestha, A. B., You, Q.-L., Zhan, Y.-J., Rajbhandari, R., Zhang, P.-F., and Wen, K.-M. (2017). Observed changes in surface air temperature and precipitation in the Hindu Kush Himalayan region over the last 100-plus years. *Advances in Climate Change Research, 8*(3), 148-156. doi:https://doi.org/10.1016/j.accre.2017.08.001

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research, 46*. doi:https://doi.org/10.1029/2009WR008328

Risbey, J. S., Squire, D. T., Black, A. S., DelSole, T., Lepore, C., Matear, R. J., Monselesan, D. P., Moore, T. S., Richardson, D., and Schepen, A. (2021). Standard assessments of climate forecast skill can be misleading. *Nature communications, 12*(1), 1-14.

Robertson, A., and Vitart, F. (2018). *Sub-seasonal to seasonal prediction: The gap between weather and climate forecasting*: Elsevier.

Robertson, A. W., Kumar, A., Peña, M., and Vitart, F. (2015). Improving and promoting subseasonal to seasonal prediction. *Bulletin of the American Meteorological Society, 96*(3), ES49-ES53. doi:https://doi.org/10.1175/BAMS-D-14-00139.1

Rowell, D. P. (2012). Sources of uncertainty in future changes in local precipitation. *Climate Dynamics, 39*(7-8), 1929-1950.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y., and Iredell, M. (2014). The NCEP climate forecast system version 2. *Journal of Climate, 27*(6), 2185-2208. doi:https://doi.org/10.1175/JCLI-D-12-00823.1

Saji, N., Goswami, B., Vinayachandran, P., and Yamagata, T. (1999). A dipole mode in the tropical Indian Ocean. *Nature, 401*(6751), 360-363. doi:https://doi.org/10.1038/43854

Sansom, P. G., Ferro, C. A. T., Stephenson, D. B., Goddard, L., and Mason, S. J. (2016). Best Practices for Postprocessing Ensemble Climate Forecasts. Part I: Selecting Appropriate

Recalibration Methods. *Journal of Climate, 29*(20), 7247-7264. doi:http://doi.org/10.1175/Jcli-D-15-0868.1

Sayemuzzaman, M., and Jha, M. K. (2014). Seasonal and annual precipitation time series trend analysis in North Carolina, United States. *Atmospheric Research, 137*, 183-194. doi:https://doi.org/10.1016/j.atmosres.2013.10.012

Schepen, A., Everingham, Y., and Wang, Q. J. (2019). Coupling forecast calibration and data-driven downscaling for generating reliable, high-resolution, multivariate seasonal climate forecast ensembles at multiple sites. *International Journal of Climatology*. doi:http://doi.org/10.1002/joc.6346

Schepen, A., Everingham, Y., and Wang, Q. J. (2020a). Coupling forecast calibration and data-driven downscaling for generating reliable, high-resolution, multivariate seasonal climate forecast ensembles at multiple sites. *International Journal of Climatology, 40*(4), 2479-2496. doi:https://doi.org/10.1002/joc.6346

Schepen, A., Everingham, Y., and Wang, Q. J. (2020b). An improved workflow for calibration and downscaling of GCM climate forecasts for agricultural applications–a case study on prediction of sugarcane yield in Australia. *Agricultural and Forest Meteorology, 291*, 107991. doi:https://doi.org/10.1016/j.agrformet.2020.107991

Schepen, A., Everingham, Y., and Wang, Q. J. (2020c). On the Joint Calibration of Multivariate Seasonal Climate Forecasts from GCMs. *Monthly Weather Review, 148*(1), 437-456. doi:https://doi.org/10.1175/MWR-D-19-0046.1

Schepen, A., and Wang, Q. J. (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. doi:http://doi.org/10.1016/j.jhydrol.2014.03.017

Schepen, A., Wang, Q. J., and Everingham, Y. (2016). Calibration, Bridging, and Merging to Improve GCM Seasonal Temperature Forecasts in Australia. *Monthly Weather Review, 144*(6), 2421-2441. doi:http://doi.org/10.1175/Mwr-D-15-0384.1

Schepen, A., Wang, Q. J., and Robertson, D. (2012a). Evidence for Using Lagged Climate Indices to Forecast Australian Seasonal Rainfall. *Journal of Climate, 25*(4), 1230-1246. doi:http://doi.org/10.1175/Jcli-D-11-00156.1

Schepen, A., Wang, Q. J., and Robertson, D. E. (2012b). Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall. *Journal of Geophysical Research-Atmospheres, 117*. doi:http://doi.org/10.1029/2012jd018011

Schepen, A., Wang, Q. J., and Robertson, D. E. (2014). Seasonal Forecasts of Australian Rainfall through Calibration and Bridging of Coupled GCM Outputs. *Monthly Weather Review, 142*(5), 1758-1770. doi:http://doi.org/10.1175/Mwr-D-13-00248.1

Schepen, A., Zhao, T. T. G., Wang, Q. J., and 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. doi:http://doi.org/10.5194/hess-22-1615-2018

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

Scheuerer, M., Switanek, M. B., Worsnop, R. P., and Hamill, T. M. (2020). Using Artificial Neural Networks for Generating Probabilistic Subseasonal Precipitation Forecasts over California. *Monthly Weather Review, 148*(8), 3489-3506. doi:https://doi.org/10.1175/MWR-D-20-0096.1

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American statistical association, 63*(324), 1379-1389. doi:http://doi.org/10.1080/01621459.1968.10480934

Shah, R., Sahai, A. K., and Mishra, V. (2017). Short to sub-seasonal hydrologic forecast to manage water and agricultural resources in India. *Hydrology and Earth System Sciences, 21*(2), 707-720. doi:https://doi.org/10.5194/hess-21-707-2017

Shao, Q., and Li, M. (2013). An improved statistical analogue downscaling procedure for seasonal precipitation forecast. *Stochastic environmental research and risk assessment, 27*(4), 819-830. doi:https://doi.org/10.1007/s00477-012-0610-0

Shao, Y., Wang, Q. J., Schepen, A., and Ryu, D. (2021a). Embedding trend into seasonal temperature forecasts through statistical calibration of GCM outputs. *International Journal of Climatology, 41*, E1553-E1565. doi:https://doi.org/10.1002/joc.6788

Shao, Y., Wang, Q. J., Schepen, A., and Ryu, D. (2021b). Going with the trend: forecasting seasonal climate conditions under climate change. *Monthly Weather Review*. doi:https://doi.org/10.1175/MWR-D-20-0318.1

Shibuya, R., Nakano, M., Kodama, C., Nasuno, T., Kikuchi, K., Satoh, M., Miura, H., and Miyakawa, T. (2021). Prediction Skill of the Boreal Summer Intra-Seasonal Oscillation in Global Non-hydrostatic Atmospheric Model Simulations with Explicit Cloud Microphysics. *Journal of the Meteorological Society of Japan. Ser. II*. doi:https://doi.org/10.2151/jmsj.2021-046

Shin, C.-S., and Huang, B. (2019). A spurious warming trend in the NMME equatorial Pacific SST hindcasts. *Climate Dynamics, 53*(12), 7287-7303. doi:https://doi.org/10.1007/s00382-017-3777-8

Silva, G. A. M., Dutra, L. M. M., da Rocha, R. P., Ambrizzi, T., and Leiva, E. (2014). Preliminary Analysis on the Global Features of the NCEP CFSv2 Seasonal Hindcasts. *Advances in Meteorology*. doi:http://doi.org/10.1155/2014/695067

Specq, D., and Batté, L. (2020). Improving subseasonal precipitation forecasts through a statistical-dynamical approach: application to the southwest tropical Pacific. *Climate Dynamics*. doi:https://doi.org/10.1007/s00382-020-05355-7

Stockdale, T. N. (2021). *SEAS5 user guide*. Retrieved from https://www.ecmwf.int/node/20150

Stockdale, T. N., Alvesb, O., Boerc, G., Dequed, M., Dinge, Y., Kumarf, A., Kumarg, K., Landmanh, W., Masoni, S., Nobrej, P., Scaifek, A., Tomoakil, O., and Yunm, W. T. (2010). Understanding and Predicting Seasonal-to-Interannual Climate Variability - The Producer Perspective. *Procedia Environmental Sciences, 1*, 55-80. doi:http://doi.org/10.1016/j.proenv.2010.09.006

Strazzo, S., Collins, D. C., Schepen, A., Wang, Q. J., Becker, E., and Jia, L. W. (2019). Application of a Hybrid Statistical-Dynamical System to Seasonal Prediction of North

American Temperature and Precipitation. *Monthly Weather Review, 147*(2), 607-625. doi:http://doi.org/10.1175/Mwr-D-18-0156.1

Tan, J., Oreopoulos, L., Jakob, C., and Jin, D. (2018). Evaluating rainfall errors in global climate models through cloud regimes. *Climate Dynamics, 50*(9), 3301-3314. doi:https://doi.org/10.1007/s00382-017-3806-7

Theil, H. (1992). A rank-invariant method of linear and polynomial regression analysis. In *Henri Theil's contributions to economics and econometrics* (pp. 345-381): Springer.

Tommasi, D., Stock, C. A., Hobday, A. J., Methot, R., Kaplan, I. C., Eveson, J. P., Holsman, K., Miller, T. J., Gaichas, S., and Gehlen, M. (2017). Managing living marine resources in a dynamic environment: the role of seasonal to decadal climate forecasts. *Progress in Oceanography, 152*, 15-49. doi:https://doi.org/10.1016/j.pocean.2016.12.011

Troccoli, A. (2010). Seasonal climate forecasting. *Meteorological Applications, 17*(3), 251-268. doi:http://doi.org/10.1002/met.184

Troccoli, A. (2018). *Weather & Climate Services for the Energy Industry*: Springer Nature.

Troccoli, A., Harrison, M., Anderson, D. L., and Mason, S. J. (2008). *Seasonal climate: forecasting and managing risk* (Vol. 82): Springer Science & Business Media.

Tuel, A., and Eltahir, E. A. (2018). Seasonal precipitation forecast over Morocco. *Water Resources Research, 54*(11), 9118-9130. doi:https://doi.org/10.1029/2018WR022984

Turco, M., Jerez, S., Doblas-Reyes, F. J., AghaKouchak, A., Llasat, M. C., and Provenzale, A. (2018). Skilful forecasting of global fire activity using seasonal climate predictions. *Nature communications, 9*(1), 1-9. doi:https://doi.org/10.1038/s41467-018-05250-0

Ullah, S., You, Q., Ullah, W., and Ali, A. (2018). Observed changes in precipitation in China-Pakistan economic corridor during 1980–2016. *Atmospheric Research, 210*, 1-14. doi:https://doi.org/10.1016/j.atmosres.2018.04.007

Unger, D. A., van den Dool, H., O'Lenic, E., and Collins, D. (2009). Ensemble Regression. *Monthly Weather Review, 137*(7), 2365-2379. doi:http://doi.org/10.1175/2008mwr2605.1

Van Dijk, A. I., Beck, H. E., Crosbie, R. S., de Jeu, R. A., Liu, Y. Y., Podger, G. M., Timbal, B., and Viney, N. R. (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*(2), 1040-1057. doi:https://doi.org/10.1002/wrcr.20123

van Straaten, C., Whan, K., Coumou, D., van den Hurk, B., and Schmeits, M. (2020). The influence of aggregation and statistical post-processing on the subseasonal predictability of European temperatures. *Quarterly Journal of the Royal Meteorological Society*. doi:https://doi.org/10.1002/qj.3810

Viel, C., Beaulant, A.-L., Soubeyroux, J.-M., and Céron, J.-P. (2016). How seasonal forecast could help a decision maker: an example of climate service for water resource management. *Advances in Science and Research, 13*, 51-55. doi:http://doi.org/10.5194/asr-13-51-2016

Vigaud, N., Tippett, M. K., Yuan, J., Robertson, A. W., and Acharya, N. (2020). Spatial Correction of Multimodel Ensemble Subseasonal Precipitation Forecasts over North America Using Local Laplacian Eigenfunctions. *Monthly Weather Review, 148*(2), 523-539. doi:http://doi.org/10.1175/mwr-d-19-0134.1

Vitart, F. (2017). Madden-Julian Oscillation prediction and teleconnections in the S2S database. *Quarterly Journal of the Royal Meteorological Society, 143*(706), 2210-2220. doi:https://doi.org/10.1002/qj.3079

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., and Fuentes, M. (2017). The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society, 98*(1), 163-173. doi:https://doi.org/10.1175/BAMS-D-16-0017.1

Vitart, F., and Robertson, A. W. (2019). Introduction: Why Sub-seasonal to seasonal prediction (S2S)? In *Sub-seasonal to seasonal prediction* (pp. 3-15): Elsevier.

Vogel, E., Lerat, J., Pipunic, R., Frost, A., Donnelly, C., Griffiths, M., Hudson, D., and Loh, S. (2021). Seasonal ensemble forecasts for soil moisture, evapotranspiration and runoff across Australia. *Journal of Hydrology*, 126620. doi:https://doi.org/10.1016/j.jhydrol.2021.126620

Walsh, J. E., and Ross, B. (1988). Sensitivity of 30-day dynamical forecasts to continental snow cover. *Journal of Climate, 1*(7), 739-754. doi:https://doi.org/10.1175/1520-0442(1988)001<0739:SODDFT>2.0.CO;2

Wang, F., Shao, W., Yu, H., Kan, G., He, X., Zhang, D., Ren, M., and Wang, G. (2020). Re-evaluation of the power of the mann-kendall test for detecting monotonic trends in hydrometeorological time series. *Frontiers in Earth Science, 8*, 14. doi:https://doi.org/10.3389/feart.2020.00014

Wang, L., and Robertson, A. W. (2019). Week 3–4 predictability over the United States assessed from two operational ensemble prediction systems. *Climate Dynamics, 52*(9-10), 5861-5875. doi:https://doi.org/10.1007/s00382-018-4484-9

Wang, Q. J., and Robertson, D. E. (2011). Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resources Research, 47*. doi: https://doi.org/10.1029/2010WR009333

Wang, Q. J., Robertson, D. E., and Chiew, F. H. S. (2009). A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resources Research, 45*. doi:https://doi.org/10.1029/2008WR007355

Wang, Q. J., Schepen, A., and Robertson, D. E. (2012a). Merging Seasonal Rainfall Forecasts from Multiple Statistical Models through Bayesian Model Averaging. *Journal of Climate, 25*(16), 5524-5537. doi:http://doi.org/10.1175/Jcli-D-11-00386.1

Wang, Q. J., Shao, Y. W., Song, Y., Schepen, A., Robertson, D. E., Ryu, D., and Pappenberger, F. (2019). An evaluation of ECMWF SEAS5 seasonal climate forecasts for Australia using a new forecast calibration algorithm. *Environmental Modelling & Software, 122*. doi:https://doi.org/10.1016/j.envsoft.2019.104550

Wang, Q. J., Shrestha, D. L., Robertson, D. E., and Pokhrel, P. (2012b). A log-sinh transformation for data normalization and variance stabilization. *Water Resources Research, 48*. doi:http://doi.org/10.1029/2011wr010973

Wasko, C., Shao, Y., Vogel, E., Wilson, L., Wang, Q., Frost, A., and Donnelly, C. (2021). Understanding trends in hydrologic extremes across Australia. *Journal of Hydrology, 593*, 125877. doi:https://doi.org/10.1016/j.jhydrol.2020.125877

Weigel, A. P., Liniger, M. A., and Appenzeller, C. (2009). Seasonal Ensemble Forecasts: Are Recalibrated Single Models Better than Multimodels? *Monthly Weather Review, 137*(4), 1460-1479. doi:http://doi.org/10.1175/2008mwr2773.1

Weisheimer, A., Doblas-Reyes, F. J., Jung, T., and Palmer, T. N. (2011). On the predictability of the extreme summer 2003 over Europe. *Geophysical Research Letters, 38*. doi:http://doi.org/10.1029/2010gl046455

Weisheimer, A., and Palmer, T. N. (2014). On the reliability of seasonal climate forecasts. *Journal of the Royal Society Interface, 11*(96). doi:http://doi.org/10.1098/rsif.2013.1162

White, C., Franks, S., and McEvoy, D. (2015). Using subseasonal-to-seasonal (S2S) extreme rainfall forecasts for extended-range flood prediction in Australia. *IAHS-AISH Proceedings and Reports, 370*, 229-234. doi:https://doi.org/10.5194/piahs-370-229-2015

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J., Lazo, J. K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A. J., and Murray, V. (2017). Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications, 24*(3), 315-325. doi:https://doi.org/10.1002/met.1654

Wilks, D. S. (2006a). Comparison of ensemble-MOS methods in the Lorenz'96 setting. *Meteorological Applications, 13*(3), 243-256. doi:https://doi.org/10.1017/S1350482706002192

Wilks, D. S. (2006b). Forecast Verification. In *Statistical Methods in the Atmospheric Sciences* (2nd ed.). Oxford, UK: Academic Press.

Wilks, D. S. (2016). "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bulletin of the American Meteorological Society, 97*(12), 2263-2273. doi:https://doi.org/10.1175/BAMS-D-15-00267.1

Wilks, D. S. (2018). Enforcing calibration in ensemble postprocessing. *Quarterly Journal of the Royal Meteorological Society, 144*(710), 76-84. doi:http://doi.org/10.1002/qj.3185

Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D. P. (2002). Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research-Atmospheres, 107*(D20), ACL 6-1-ACL 6-15. doi:http://doi.org/10.1029/2001jd000659

World Meteorological Organization. (2017). *WMO guidelines on the calculation of climate normals*. Retrieved from https://library.wmo.int/doc_num.php?explnum_id=4166

Xu, C.-H., and Xu, Y. (2012). The projection of temperature and precipitation over China under RCP scenarios using a CMIP5 multi-model ensemble. *Atmospheric and Oceanic Science Letters, 5*(6), 527-533. doi:https://doi.org/10.1080/16742834.2012.11447042

Xu, Z., Han, Y., and Yang, Z. (2019). Dynamical downscaling of regional climate: A review of methods and limitations. *Science China Earth Sciences, 62*(2), 365-375. doi:https://doi.org/10.1007/s11430-018-9261-5

Yang, C., Yuan, H., and Su, X. (2020). Bias correction of ensemble precipitation forecasts in the improvement of summer streamflow prediction skill. *Journal of Hydrology, 588*, 124955. doi:https://doi.org/10.1016/j.jhydrol.2020.124955

Yeo, I. K., and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika, 87*(4), 954-959. doi:http://doi.org/10.1093/biomet/87.4.954

Yu, L. J., Zhong, S. Y., Heilman, W. E., and Bian, X. D. (2018). Trends in seasonal warm anomalies across the contiguous United States: Contributions from natural climate variability. *Scientific Reports, 8*. doi:https://doi.org/10.1038/s41598-018-21817-9

Yuan, X., and Wood, E. F. (2012). Downscaling precipitation or bias-correcting streamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast. *Water Resources Research, 48*. doi:https://doi.org/10.1029/2012WR012256

Yuan, X., Wood, E. F., and Ma, Z. G. (2015). A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development. *Wiley Interdisciplinary Reviews-Water, 2*(5), 523-536. doi:http://doi.org/10.1002/wat2.1088

Zampieri, L., Goessling, H. F., and Jung, T. (2018). Bright prospects for Arctic sea ice prediction on subseasonal time scales. *Geophysical Research Letters, 45*(18), 9731-9738. doi:https://doi.org/10.1029/2018GL079394

Zhao, C., Ren, H. L., Eade, R., Wu, Y., Wu, J., and MacLachlan, C. (2019). MJO modulation and its ability to predict boreal summer tropical cyclone genesis over the northwest Pacific in Met Office Hadley Centre and Beijing Climate Center seasonal prediction systems. *Quarterly Journal of the Royal Meteorological Society, 145*(720), 1089-1101. doi:https://doi.org/10.1002/qj.3478

Zhao, T., Bennett, J. C., Wang, Q. J., Schepen, A., Wood, A. W., Robertson, D. E., and Ramos, M. H. (2017). How Suitable is Quantile Mapping For Postprocessing GCM Precipitation Forecasts? *Journal of Climate, 30*(9), 3185-3196. doi:http://doi.org/10.1175/Jcli-D-16-0652.1

Zhao, T., Wang, Q. J., and Schepen, A. (2019a). A Bayesian modelling approach to forecasting short-term reference crop evapotranspiration from GCM outputs. *Agricultural and Forest Meteorology, 269*, 88-101. doi:https://doi.org/10.1016/j.agrformet.2019.02.003

Zhao, T., Wang, Q. J., Schepen, A., and Griffiths, M. (2019b). Ensemble forecasting of monthly and seasonal reference crop evapotranspiration based on global climate model outputs. *Agricultural and Forest Meteorology, 264*, 114-124. doi:http://doi.org/10.1016/j.agrformet.2018.10.001

Zuo, H., Balmaseda, M. A., Mogensen, K., and Tietsche, S. (2018). *OCEAN5: the ECMWF Ocean Reanalysis System ORAS5 and its Real-Time analysis component*. Retrieved from https://www.ecmwf.int/sites/default/files/elibrary/2018/18519-ocean5-ecmwf-ocean-renalysis-system-and-its-real-time-analysis-component.pdf

Appendix

S1 Supplementary Material for Chapter 2

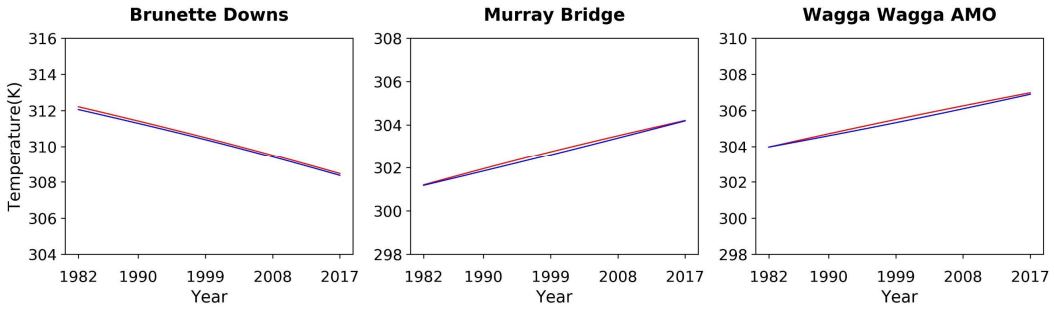

Figure S1-1: Back-transformed trendlines of BJP-t calibrated forecasts (blue lines) and observations (red lines) in three stations.

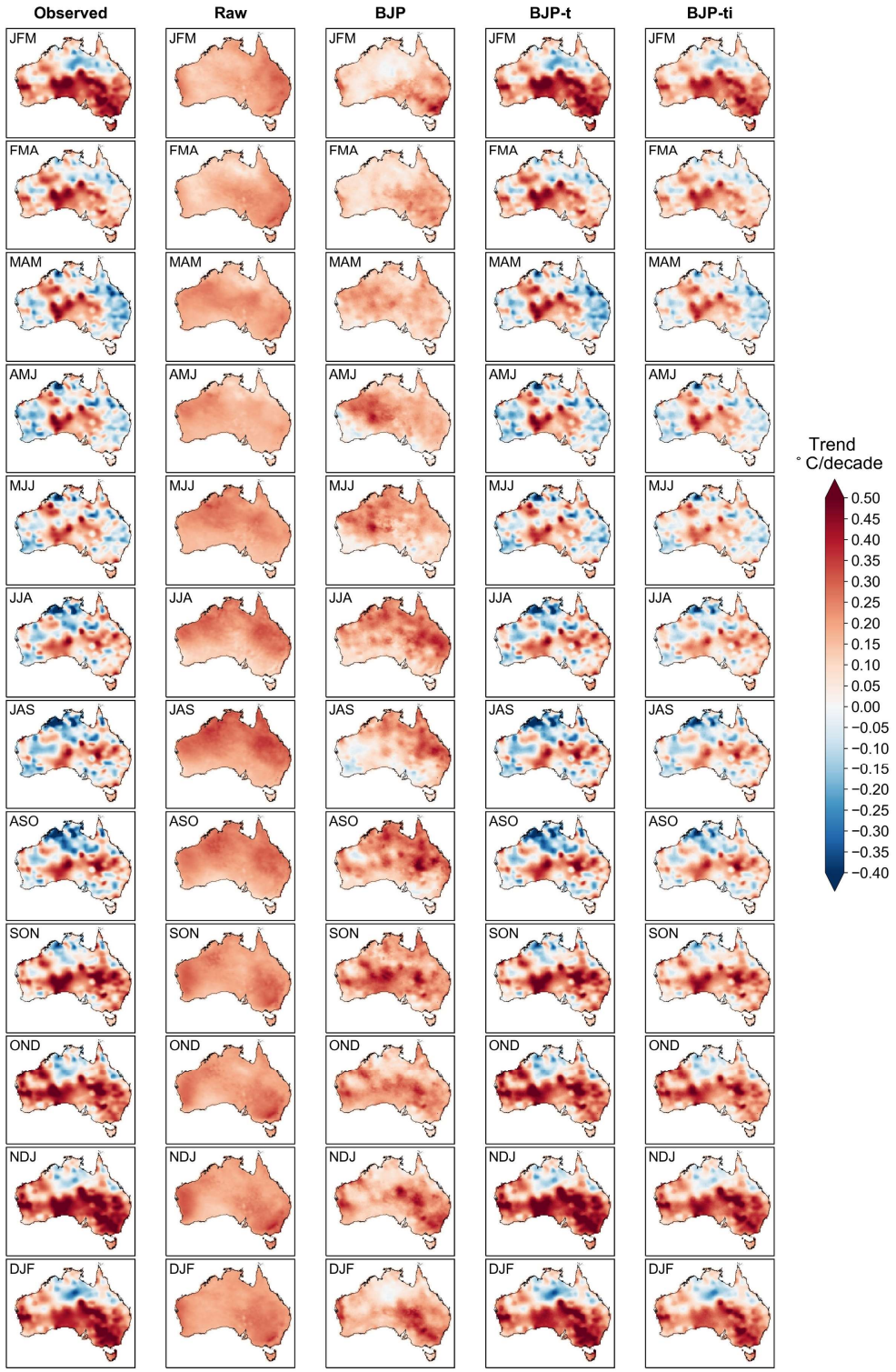

Figure S2-1: Linear decadal trends of seasonal averages of Tmin for observations, raw, BJP, BJP-t, and BJP-ti calibrated forecasts for 12 overlapping seasons at 1-month lead time from 1981 FMA to 2017 JFM.

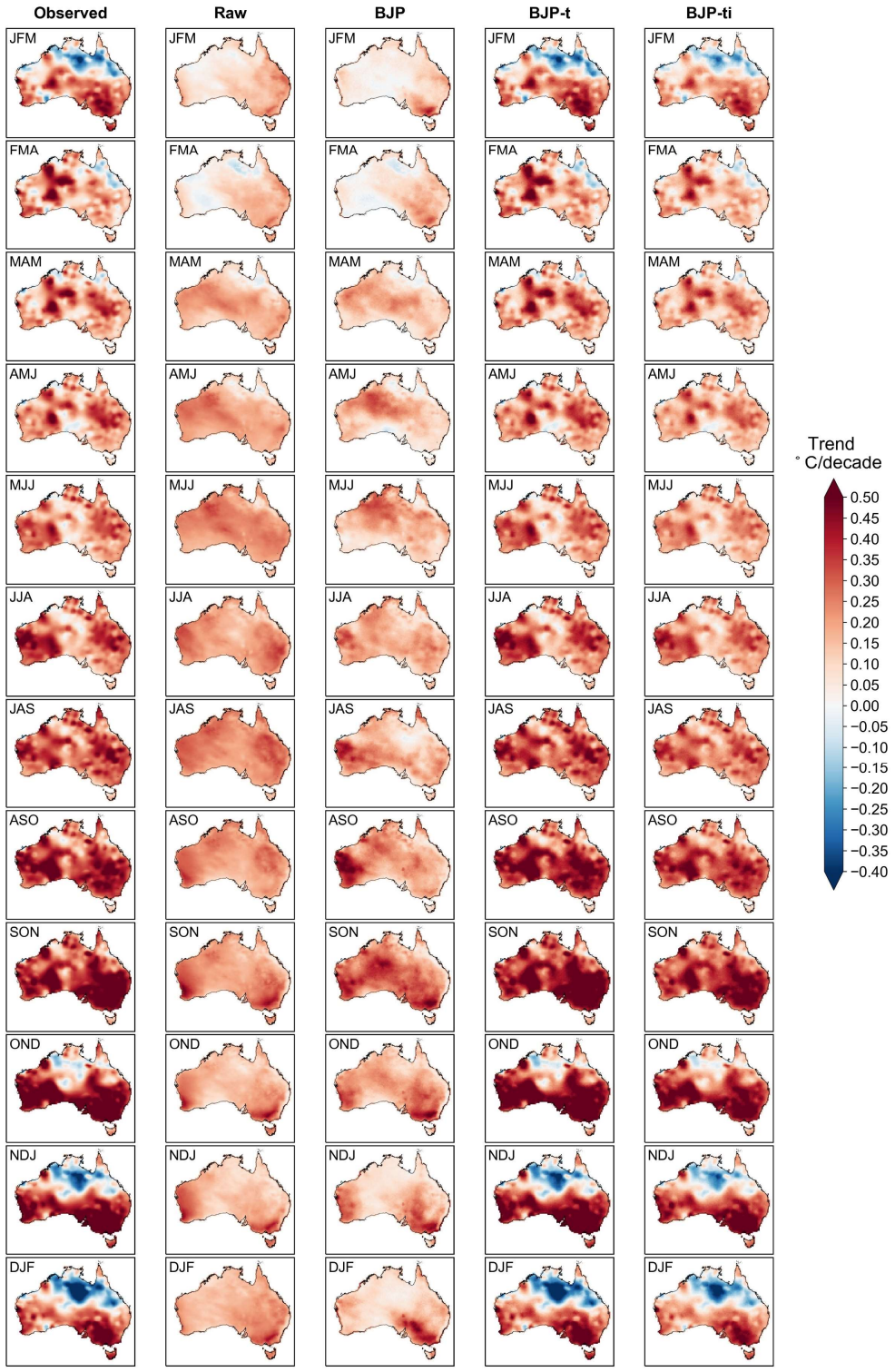

Figure S2-2: As in Figure S2-1, but for Tmax.

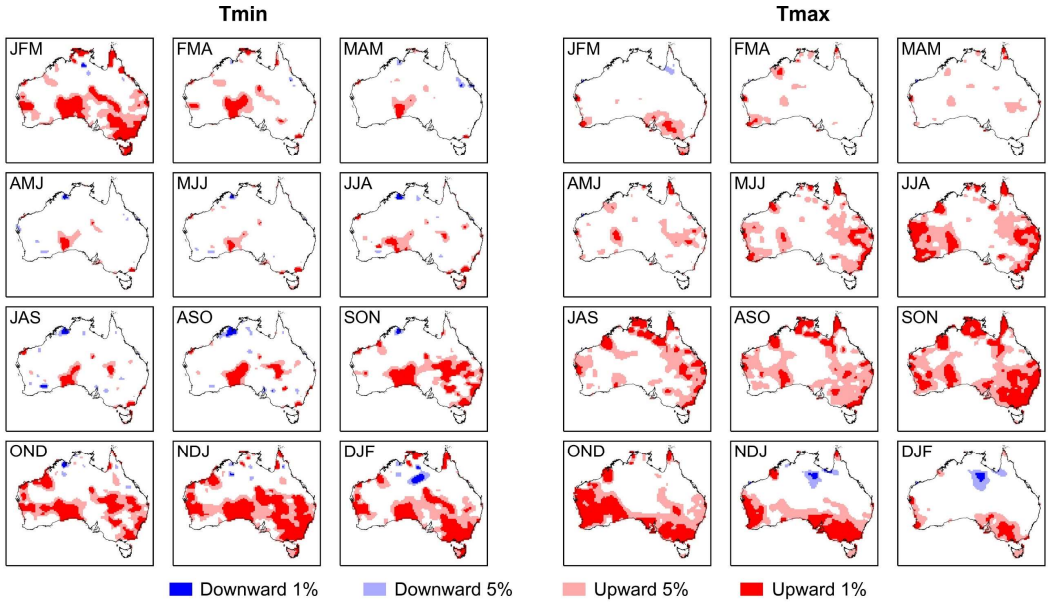

Figure S2-3: Trend significance of observations at 1% and 5% significance level for seasonal averages of (left) Tmin and (right) Tmax at 1-month lead time using t-test.

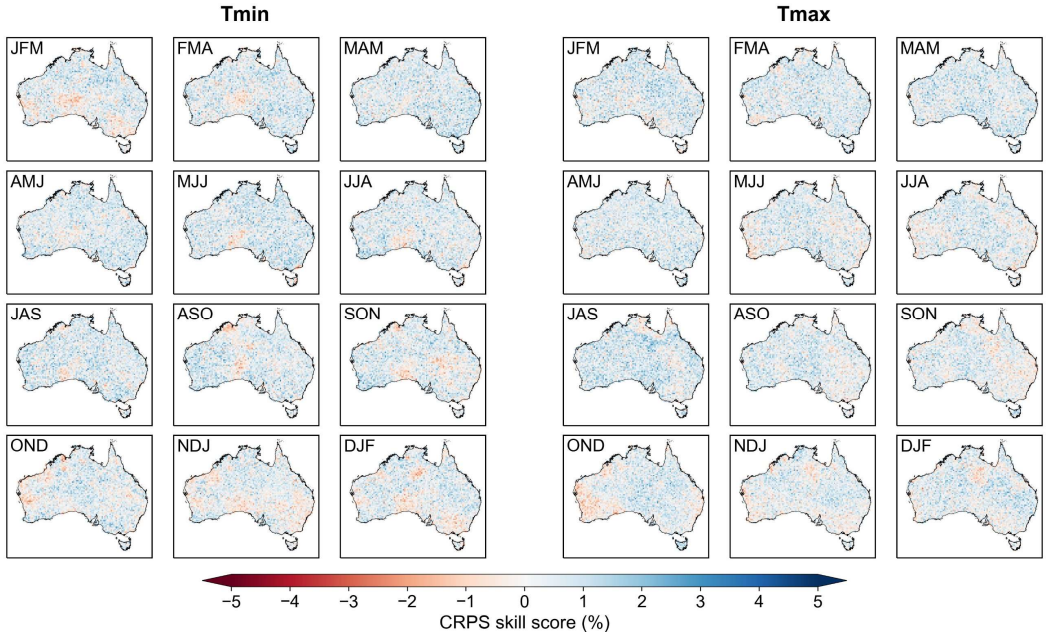

Figure S2-4: CRPS skill score difference between BJP-ti and BJP-t calibrated forecasts of seasonal averages of (left) Tmin and (right) Tmax at 1-month lead time. The skill score is calculated using leave-one-year-out cross-validated climatology ensemble forecasts from the BJP model as the reference forecasts.

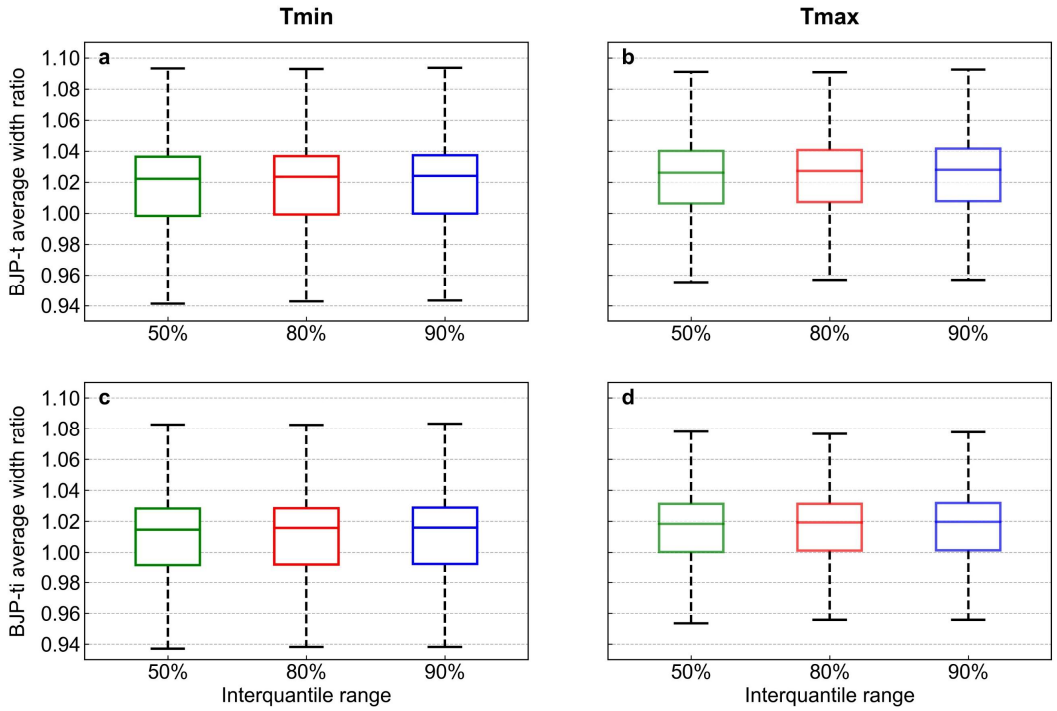

Figure S2-5: Average band width ratio of BJP-t versus BJP calibrated forecasts (a,b), and BJP-ti versus BJP calibrated forecasts (c,d) at different inter-quantile ranges for seasonal averages of Tmin (a,c) and Tmax (b,d) at 1-month lead time.

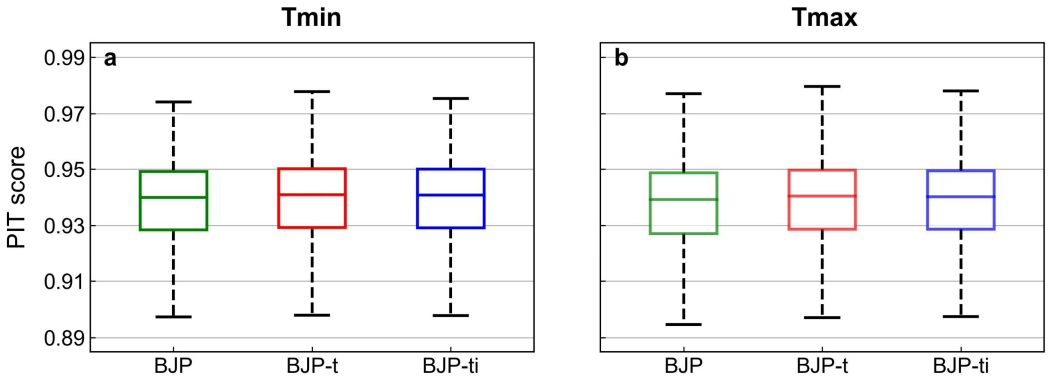

Figure S2-6: PIT scores of BJP, BJP-t, and BJP-ti calibrated for seasonal averages of Tmin (a) and Tmax (b) at 1-month lead time.

S3 Supplementary Material for Chapter 4

Section 4.5.1 introduces the key steps for the more advanced version of the trend-aware model for post-processing seasonal precipitation forecasts. Here, we present the complete math and pseudo code for the algorithm.

S3.1 The model

Consider a predictor ($y_1^{'}$ - raw ensemble forecast mean after transformation) and a predictand ($y_2^{'}$ - observation after transformation) with N historical data records,

$$\mathbf{y}^{'}(t) = \begin{bmatrix} y_1^{'}(t) \\ y_2^{'}(t) \end{bmatrix} \tag{S3-1}$$

where t is the event time ($t = 1, 2, ..., N$). We calculate anomalies $z_i(t), i = 1, 2$ from the linear trendlines of the variables $y_i^{'}$,

$$z_i(t) = y_i^{'}(t) - \alpha_i (t - t_m) \tag{S3-2}$$

where α_i are trend parameters, t_m is approximately the time of the middle event in the analysis period. The joint distribution of z_1 and z_2 is modelled as,

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{S3-3}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are a 2×1 mean vector and a 2×2 covariance matrix,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \tag{S3-4}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{S3-5}$$

where μ_i is a mean, σ_i is a standard deviation, and ρ is a correlation coefficient. The marginal distribution of z_i is,

$$[z_i(t)] = \mathrm{N}(\mu_i, \sigma_i^2) \tag{S3-6}$$

The distribution of $y_i^{'}$ is derived from combining Eq. (S3-2) and Eq. (S3-6),

$$[y_i^{'}(t)] = \mathrm{N}[\mu_i + \alpha_i(t - t_m), \sigma_i^2] \tag{S3-7}$$

Hereafter, the parameter set to be inferred is denoted as $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha_1, \alpha_2\}$. We assume that the variables may have missing values or left censored values.

S3.2 Parameter inference

The posterior distribution of the model parameters is derived based on Bayes' theorem,

$$p(\boldsymbol{\theta} \mid \mathbf{D}) \propto p(\boldsymbol{\theta}) p(\mathbf{D} \mid \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{t=1}^{N} p(\mathbf{D} \mid \boldsymbol{\theta}) \tag{S3-8}$$

where $\mathbf{D} = \left\{[y_1^{'}(t), y_2^{'}(t)], t = 1, 2, ..., N\right\}$, $p(\boldsymbol{\theta})$ is a prior distribution of the model parameters, and $p(\mathbf{D} \mid \boldsymbol{\theta})$ is the likelihood. The prior for $\boldsymbol{\theta}$ is specified as,

$$p(\boldsymbol{\theta}) \propto p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\alpha_1) p(\alpha_2) \tag{S3-9}$$

We apply the non-informative multivariate Jeffreys prior for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Gelman et al. 2014),

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \left|\boldsymbol{\Sigma}\right|^{-3/2} \tag{S3-10}$$

As introduced in Section 4.5.1, the prior for α_i can be specified in the form of a non-informative uniform distribution or an informative normal distribution centered at zero. The resulting models are named BJP-t and BJP-ti respectively. In this work, we mainly present the BJP-ti model, with the informative prior of the form,

$$p(\alpha_i) \propto N(0, m_i^2) \tag{S3-11}$$

where m_i are empirically determined as detailed in Section 4.8.

By combining Eqs. (S3-8) – (S3-11), we can derive the following conditional distributions for the parameters (Gelman et al., 2014),

$$[\boldsymbol{\Sigma} \mid \cdot] = \text{Inv- Wishart}_{N-1}(\mathbf{S}) \tag{S3-12}$$

$$[\boldsymbol{\mu} \mid \cdot] = \mathrm{N}(\overline{\mathbf{z}}, \boldsymbol{\Sigma} / N) \tag{S3-13}$$

where

$$\mathbf{S} = \sum_{t=1}^{N} [\mathbf{z}(t) - \overline{\mathbf{z}}][\mathbf{z}(t) - \overline{\mathbf{z}}]^{\mathrm{T}} \tag{S3-14}$$

$$\overline{\mathbf{z}} = \frac{1}{N} \sum_{t=1}^{N} \mathbf{z}(t) \tag{S3-15}$$

The symbol $|\cdot$ indicates conditional on all other variables, and Inv-Wishart_{N-1} is the Inverse-Wishart distribution with $N-1$ degrees of freedom, and $[\mathbf{z}(t) - \overline{\mathbf{z}}]^{\mathrm{T}}$ is the transpose of $[\mathbf{z}(t) - \overline{\mathbf{z}}]$.

We can derive the conditional distribution for α_i from Eq. (S3-2) and Eq. (S3-7) as,

$$[\alpha_i \,|\cdot] = \mathrm{N}\left\{ \frac{m_i^2 \sum_{t=1}^{N} [y_i^{'}(t) - \mu_i](t - t_m)}{m_i^2 \sum_{t=1}^{N} (t - t_m)^2 + \sigma_i^2}, \frac{m_i^2 \sigma_i^2}{m_i^2 \sum_{t=1}^{N} (t - t_m)^2 + \sigma_i^2} \right\} \tag{S3-16}$$

When $z_i(t)$ has a missing value, its conditional distribution is obtained as,

$$[z_i(t) \,|\cdot] = \mathrm{N}[\mu_i^*(t), \Sigma_{i,i}^*] \tag{S3-17}$$

where

$$\Sigma_{i,i}^* = \sigma_i^2 - (\rho \sigma_1 \sigma_2)^2 / \sigma_{(i)}^2 \tag{S3-18}$$

$$\mu_i^*(t) = \mu_i + \rho \sigma_1 \sigma_2 / \sigma_{(i)}^2 \times [z_{(i)}(t) - \mu_{(i)}] \tag{S3-19}$$

(i) denotes the index in $\{1, 2\}$ that is not i. when $z_i(t)$ has a censored value rather than a missing value, we can also use Eq. (S3-17) to sample a value, but the sampling range of $z_i(t)$ is restricted to $z_i(t) \le z_i^c(t)$,

$$z_i^c(t) = y_i^{'c} - \alpha_i (t - t_m) \tag{S3-20}$$

where $y_i^{'c}$ is the constant censoring threshold of the variable $y_i(t)$.

We use the conditional distributions of Eqs. (S3-12) – (S3-20) to set up the Gibbs sampling to sequentially draw one sample from each of the conditional distributions with the remaining parameters fixed to their current values and repeat many iterations of such sampling. This generates a parameter chain, and the sampling processes continues until convergence. That is, the sampled parameter sets have the same distribution as sampled from the overall joint posterior distribution.

Below are the pseudo codes for implementing the Gibbs sampling for parameter inference.

Setting initial values

If the value of $y_i^{'}(t)$ is censored in the original data, set $y_i^{'}(t) = y_i^{'c}$

If the value of $y_i^{'}(t)$ is missing in the original data, set $y_i^{'}(t) = \hat{y}_i^{'}$, where $\hat{y}_i^{'}$ is the average of non-missing $y_i^{'}(t)$, $t = 1, \ldots, N$

Set $\alpha_i = 0$

Gibbs sampling

Repeat sampling

Compute $z_i(t)$

Compute $\mathbf{S}$ and $\overline{\mathbf{z}}$

Sample $\boldsymbol{\Sigma}$. Check if $\boldsymbol{\Sigma}$ is positive definite. If not, resample $\boldsymbol{\Sigma}$

Sample $\boldsymbol{\mu}$

Randomise the ordering $\{1, 2\}$ to ran_ord( ) by sampling without replacement

Do j=1,2

i=ran_ord(j)

If any of the values of $z_i(t)$ $t = 1, \ldots, N$, is missing or censored in the original data, compute $\rho\sigma_1\sigma_2 / \sigma_{(i)}^2$ and $\Sigma_{i,i}^*$

Do t=1 to N

If the value of $z_i(t)$ is missing in the original data, compute $\mu_i^*(t)$, sample $[z_i(t) \mid \cdot] = \mathrm{N}[\mu_i^*(t), \Sigma_{i,i}^*]$ and update $z_i(t)$

If the value of $z_i(t)$ is censored in the original data, compute $\mu_i^*(t)$, sample $[z_i(t) \mid \cdot] = \mathrm{N}[\mu_i^*(t), \Sigma_{i,i}^*]$, $z_i(t) \le z_i^c(t)$ and update $z_i(t)$

End do

End do

Calculate and update $y_i^{'}(t)$

Sample α_i

End repeat

S3.3 Prediction use

Once all the sampled parameter sets are sampled, we can apply the model for prediction, in which the posterior distribution of the predictand is found based on the predictor information. This

distribution is sampled by treating the predictand as missing in values, and following the Gibbs sampling steps shown in Eqs. (S3-17) – (S3-19). The predictor $z_1(t^*)$ can also be missing or censored and be included in the Gibbs sampling. In case of censored values, the sampling is restricted to the range of $z_1(t^*) \le z_1^c(t)$ as presented in Section S3.2. In addition, before sampling for prediction, a pragmatic approach is used to adjust extremely large $z_1(t^*)$ values that occur in prediction.

Below are the pseudo codes for implementing the Gibbs sampling for prediction use.

Setting initial values

If the value of $y_i'(t)$ is censored in the original data, $y_i'(t) = y_i'^c$

If the value of $y_i'(t)$ is missing in the original data, $y_i'(t) = \mu_i + \alpha_i(t - t_m)$ (value of the first set of parameters)

Gibbs sampling

Repeat sampling

Use $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, α_i already sampled

Compute $z_i(t)$

Given the marginal distribution $\mathrm{N}(z_1 \mid \mu_1, \sigma_1^2)$, check if $\Phi(z_1(t) \mid \mu_1, \sigma_1^2) > P_{\text{extreme}}$ is true. If true, set $z_1(t)$ back to the extreme threshold $z_1(t) = \Phi^{-1}(P_{\text{extreme}} \mid \mu_1, \sigma_1^2)$

Randomise the ordering $\{1, 2\}$ to ran_ord( ) by sampling without replacement

Do j=1,2

i=ran_ord(j)

If the value of $z_i(t)$ is missing or censored in the original data, compute $\mu_i^*(t)$ and $\Sigma_{i,i}^*$

If the value of $z_i(t)$ is missing in the original data, sample $[z_i(t) \mid \cdot] = \mathrm{N}(\mu_i^*(t), \Sigma_{i,i}^*)$ and update $z_i(t)$

If the value of $z_i(t)$ is censored in the original data, sample $[z_i(t) \mid \cdot] = \mathrm{N}(\mu_i^*(t), \Sigma_{i,i}^*)$, $z_i(t) \le z_i^c(t)$ and update $z_i(t)$

End do

Calculate and update $y_i'(t)$

Set $y_1'(t)$ back to its initial value if $z_1(t)$ was set as $\Phi^{-1}(P_{\text{extreme}} \mid \mu_1, \sigma_1^2)$

End repeat

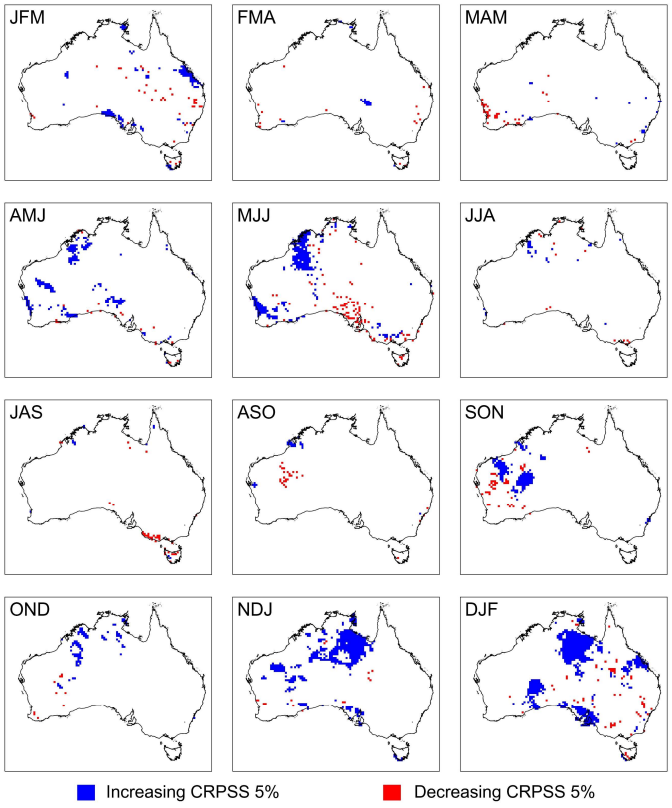

Figure S3-1: Statistical significance of the improvement or worsening of the CRPS skill score of the BJP-ti calibrated forecasts compared to BJP at 5% significance level.

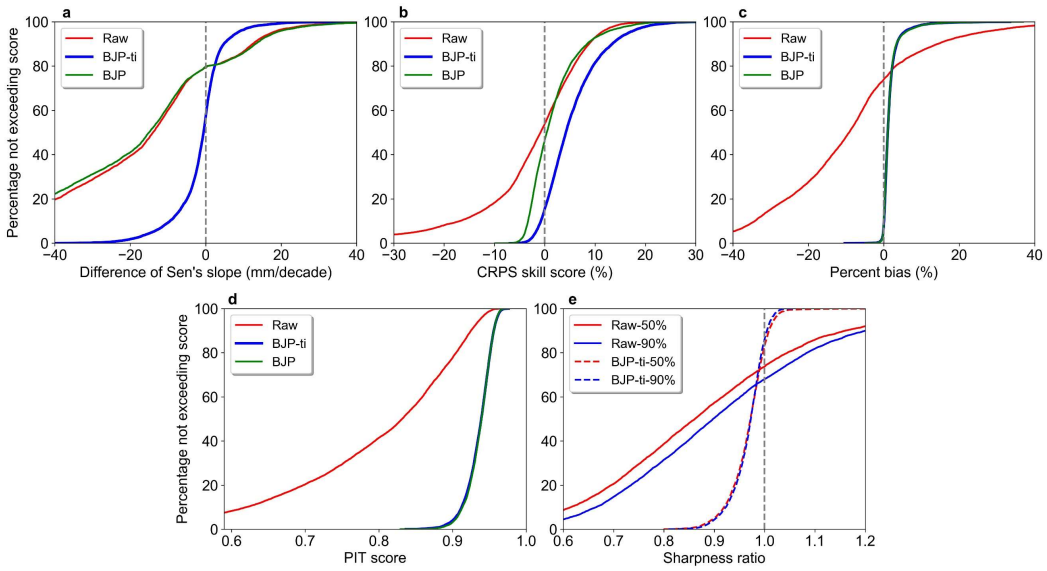

Figure S3-2: Non-exceedance plot comparing the overall performance of the raw, BJP and BJP-ti calibrated forecasts at 1-month lead time in the region with statistically significant observed trends at 10% level.

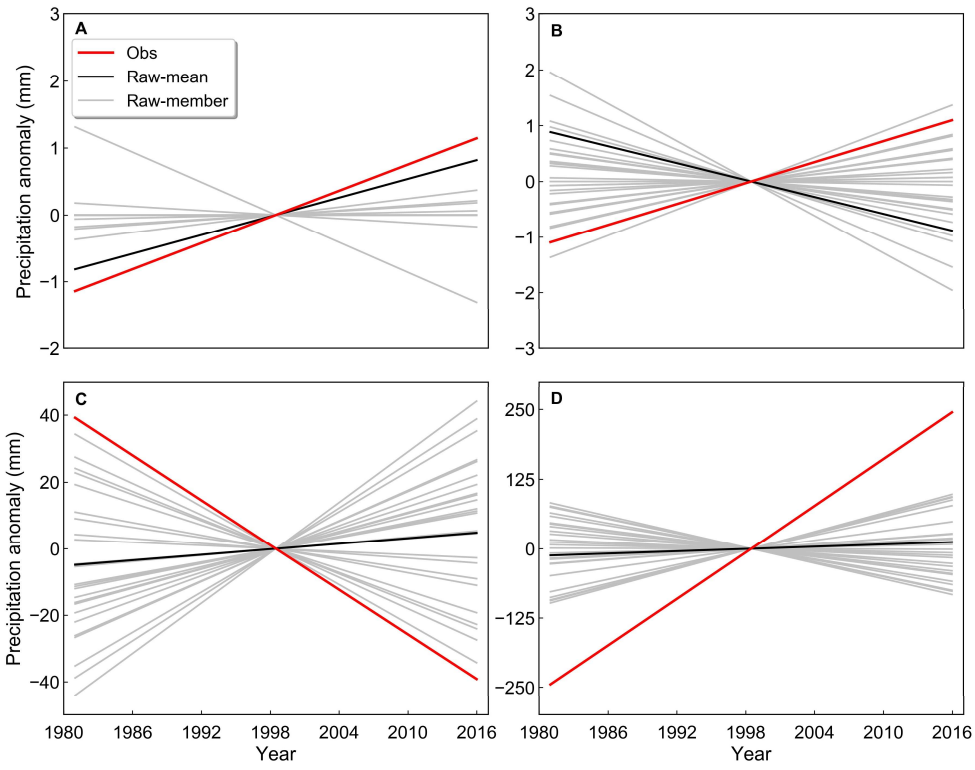

Figure S3-3: The trendlines of observations, raw ensemble forecast means, each member of raw ensemble forecasts relative to the temporal mean of each trendline for four selected cells. Locations of the grid cells are shown in Figure 4-5.

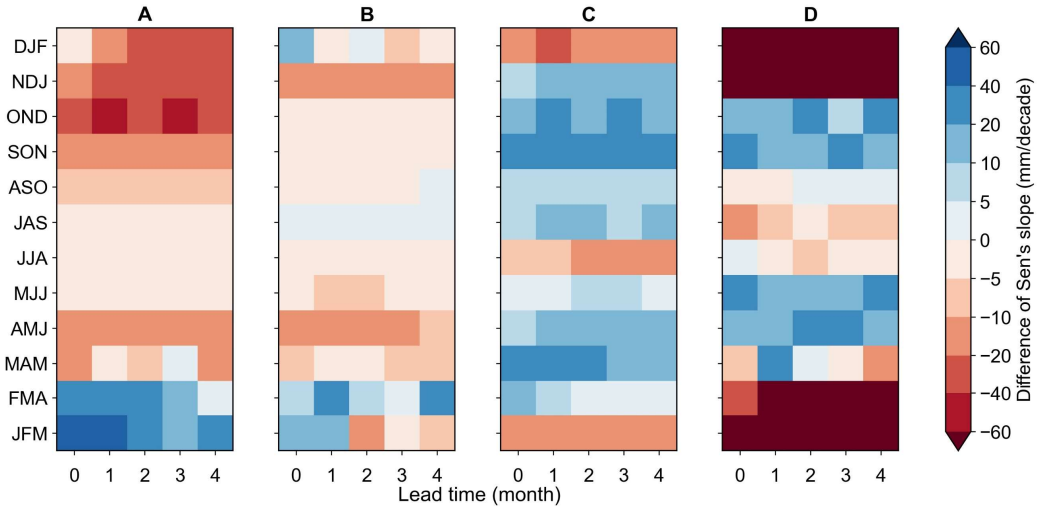

Figure S3-4: Trend difference between the BJP calibrated ensemble forecast medians and observations of seasonal precipitation for all lead times. Locations of the grid cells are shown in Figure 4-5.

S4. Supplementary Material for Chapter 5

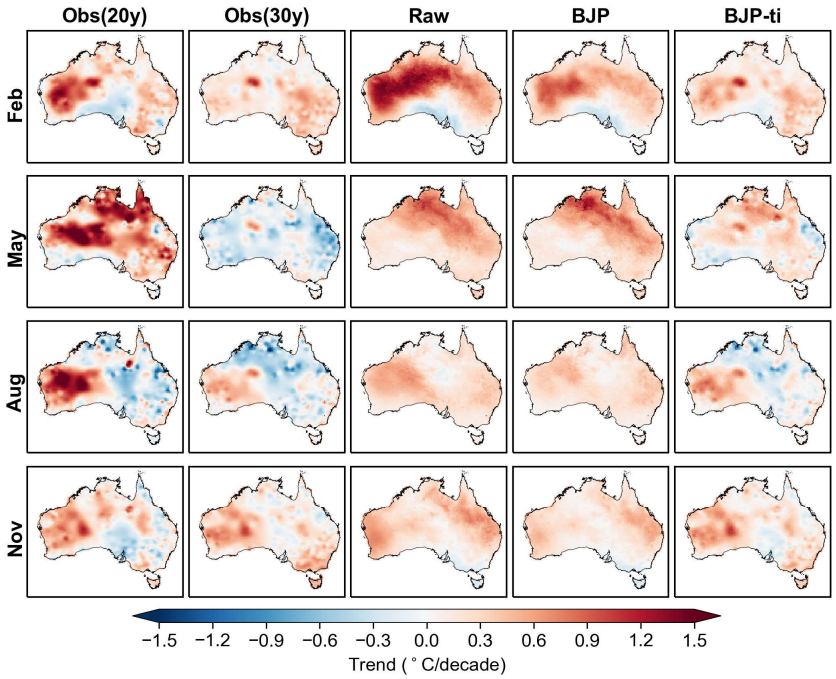

Figure S4-1: Decadal trends for Tmin observations over 2000-2019 and 1990-2019, raw forecasts, BJP calibrated forecasts, and BJP-ti calibrated week-2 forecasts over 2000-2019 for all initialisation dates within February, May, August, and November separately.

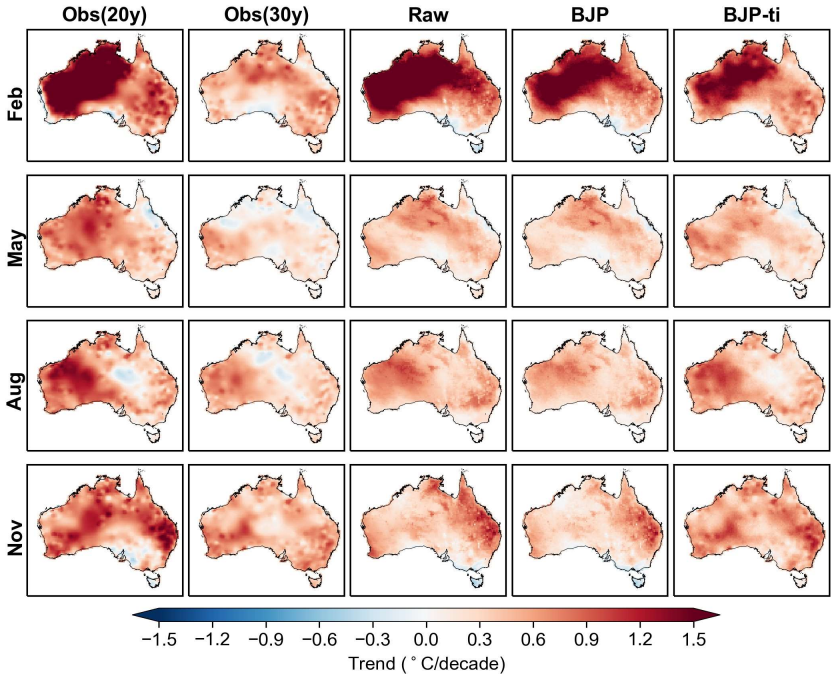

Figure S4-2: As in Figure S4-1, but for Tmax.

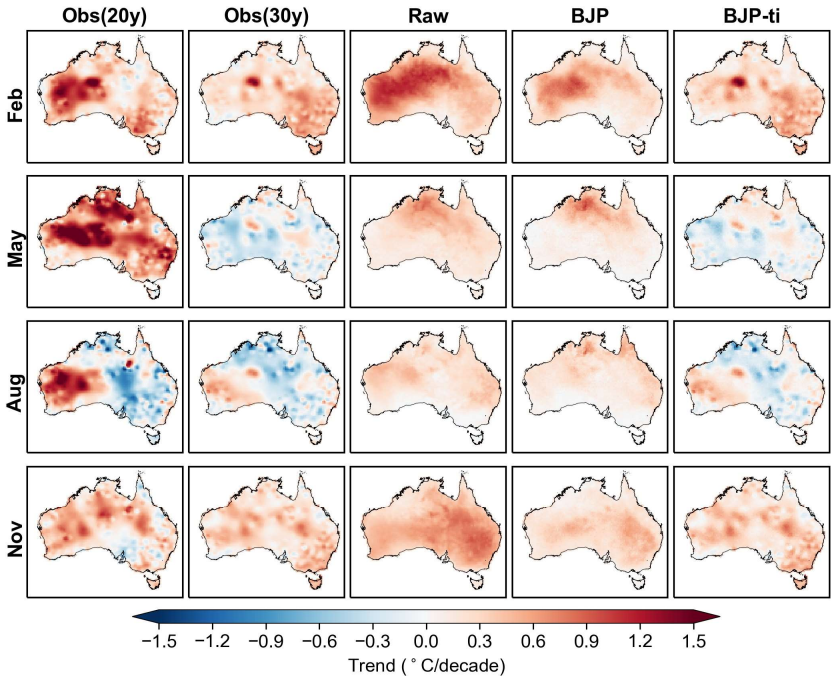

Figure S4-3: As in Figure S4-1, but for week-3 forecasts.

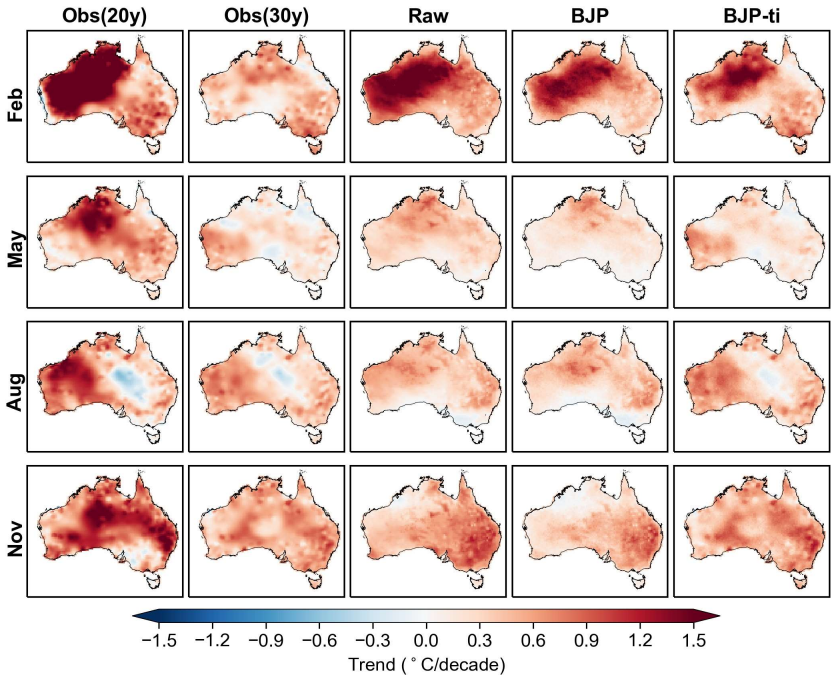

Figure S4-4: As in Figure S4-3, but for Tmax.

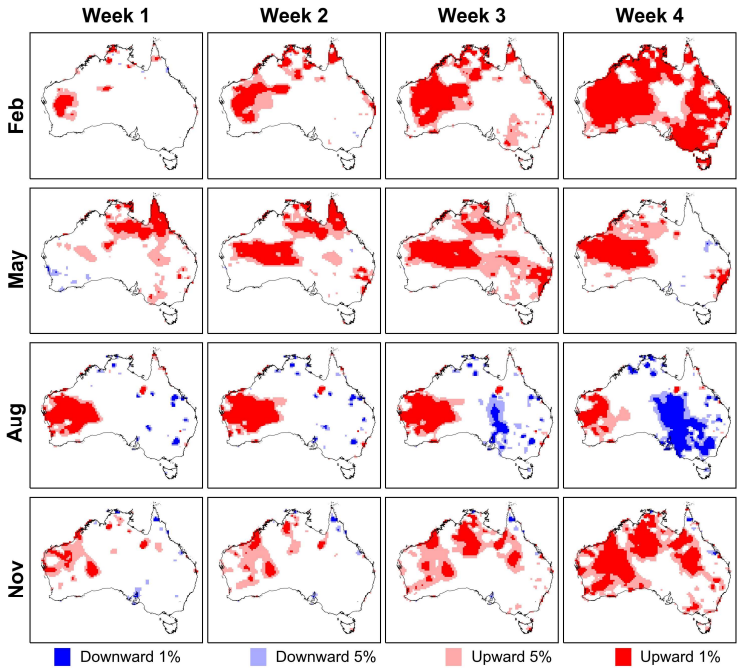

Figure S4-5: Statistical significance of the trend in week 1-4 Tmin observations for all initialisation dates within February, May, August, and November separately over 2000-2019 using two-tailed student's t-test that accounts for the false discovery rate.

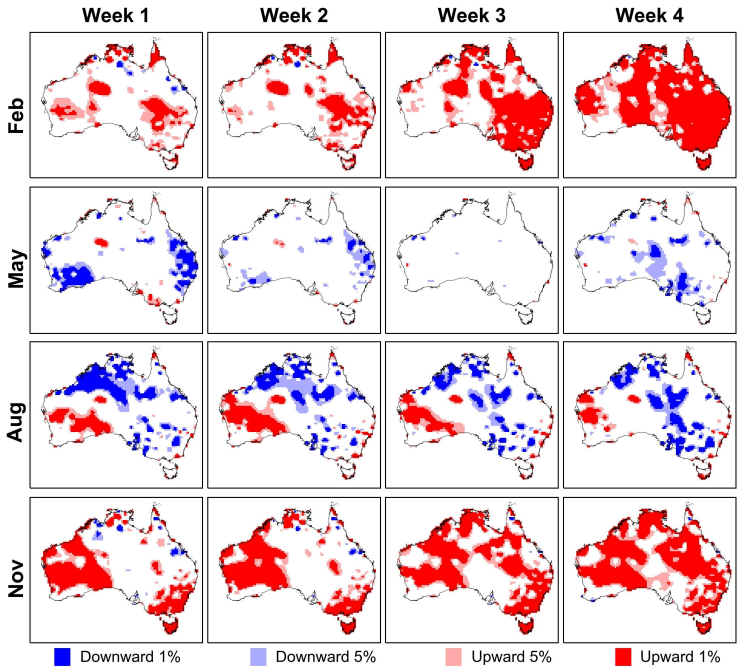

Figure S4-6: As in Figure S4-5, but for Tmin observations over 1990-2019.

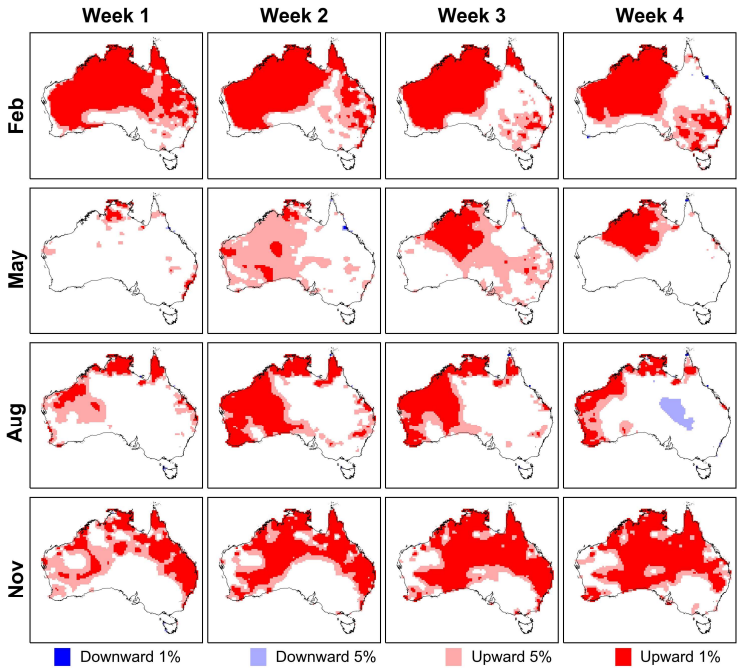

Figure S4-7: As in Figure S4-5, but for Tmax observations.

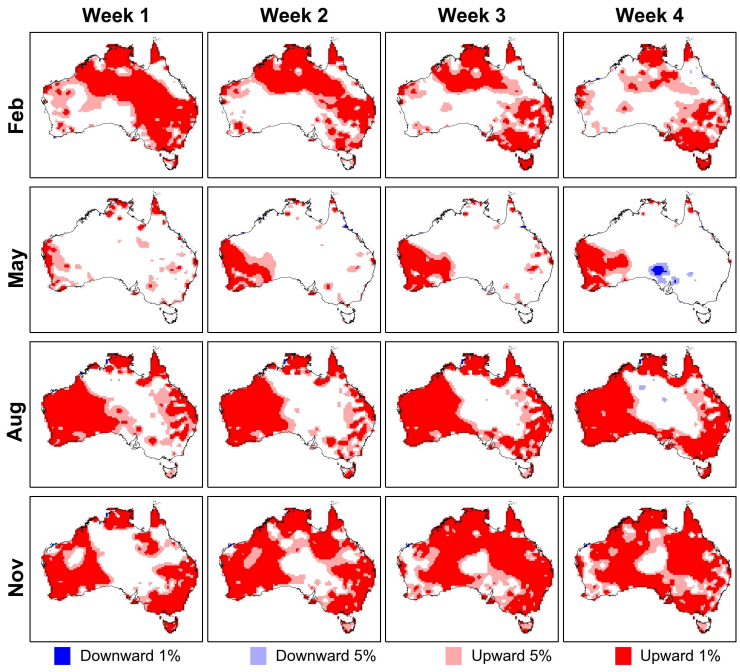

Figure S4-8: As in Figure S4-6, but for Tmax observations.

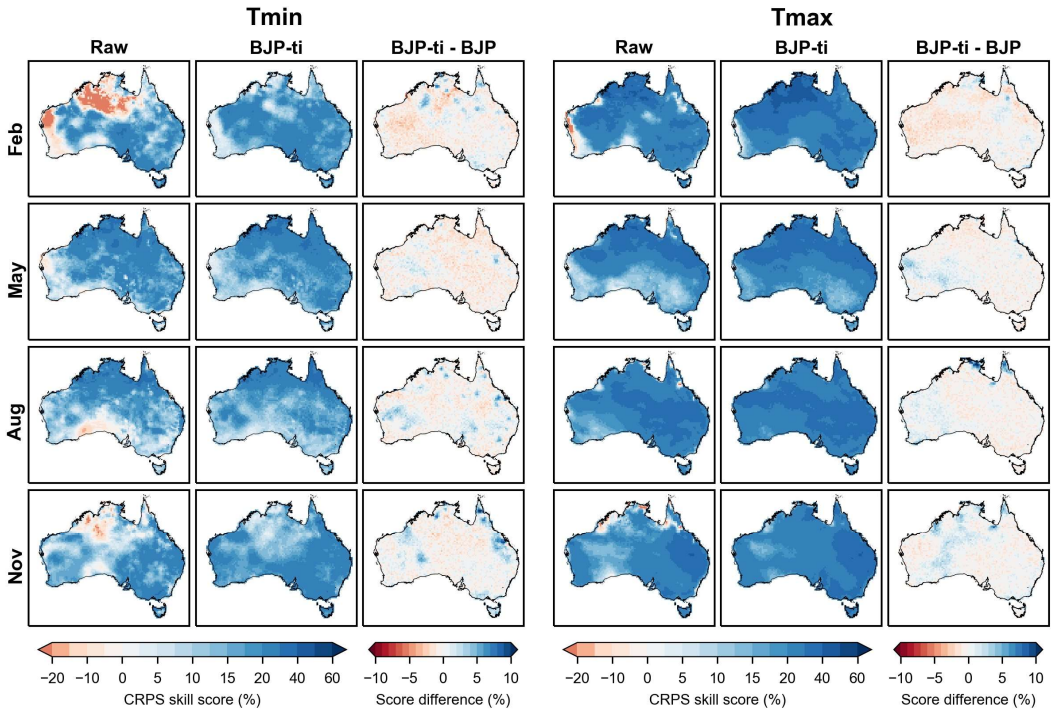

Figure S4-9: CRPS skill scores for Tmin and Tmax week-2 raw forecasts, BJP-ti calibrated forecasts, and the score difference between BJP-ti and BJP calibrated forecasts for all initialisation dates within February, May, August, and November separately over 2000-2019.

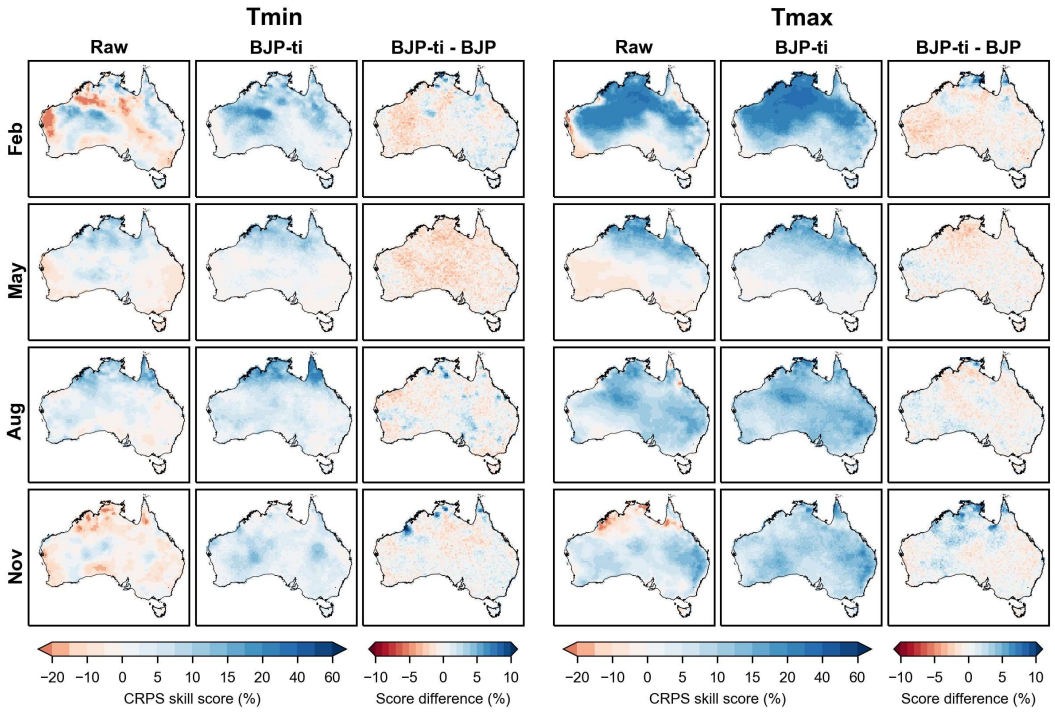

Figure S4-10: As in Figure S4-9, but for week-3 forecasts.

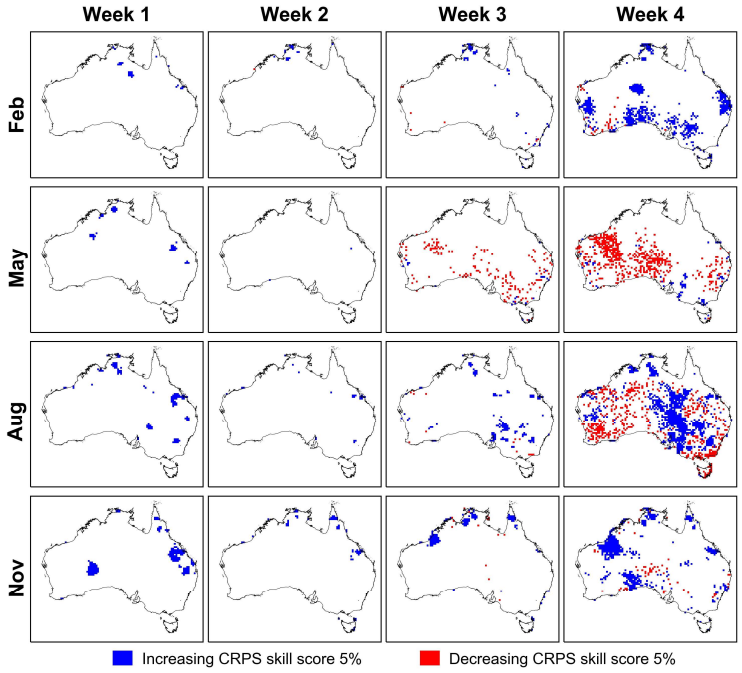

Figure S4-11: Statistical significance of the improvement or worsening of the CRPS skill score of the BJP-ti calibrated forecasts compared to BJP for Tmin at 5% significance level.

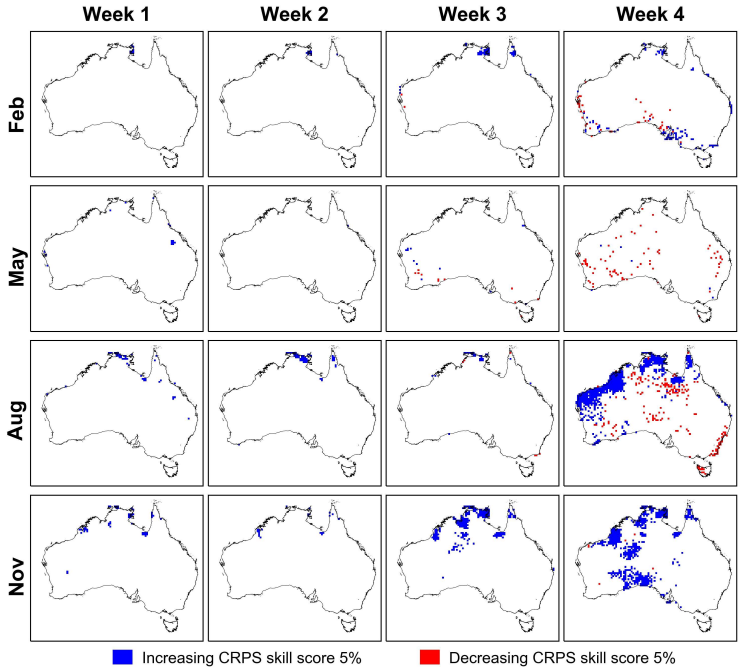

Figure S4-12: As in Figure S4-11, but for Tmax.