University Library

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Shao, Y;Wang, QJ;Schepen, A;Ryu, D

Title:
Introducing long-term trends into subseasonal temperature forecasts through trend-aware postprocessing

Date:
2022-07-01

Citation:
Shao, Y., Wang, Q. J., Schepen, A. & Ryu, D. (2022). Introducing long-term trends into subseasonal temperature forecasts through trend-aware postprocessing. International Journal of Climatology, 42 (9), pp.4972-4988. https://doi.org/10.1002/joc.7515.

Persistent Link:
https://hdl.handle.net/11343/310825

Shao Yawen (Orcid ID: 0000-0002-9938-669X)
Schepen Andrew (Orcid ID: 0000-0002-6372-735X)

Introducing Long-term Trends into Sub-seasonal Temperature Forecasts through Trend-aware Post-processing

Yawen Shao[a*], Q. J. Wang[a], Andrew Schepen[b], Dongryeol Ryu[a]

a. Department of Infrastructure Engineering, The University of Melbourne, Parkville 3010, Australia

b. CSIRO Land and Water, Dutton Park 4102, Australia

* Corresponding author: Department of Infrastructure Engineering, The University of Melbourne, Parkville 3010, Australia.

E-mail address: yawens@student.unimelb.edu.au

Keywords: sub-seasonal climate forecasting, temperature trend, ensemble forecast calibration, forecast verification

Funding information: this work is funded by an ARC Linkage Project (LP170100922)

Author Manuscript

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/joc.7515

This article is protected by copyright. All rights reserved.

Author Manuscript

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 Australia 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.

Author Manuscript

1 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 Centers 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, 2017; 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.

Author Manuscript

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; Zhao et al., 2019). These existing methods have greatly enhanced forecast performance, but they rarely aim to eliminate the 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.

Previous works proposed a robust trend-aware post-processing methodology for resolving the trend mismatch issue in seasonal climate forecasts (Shao et al., 2021a, 2021b). 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 study, 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 study, 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 re-forecasts of daily minimum and maximum temperatures across the Australian continent produced by the ECMWF extended-range forecasting system.

Author Manuscript

2 Study Data

2.1 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.

2.2 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.

Author Manuscript

3 Methods

3.1 Alignment of daily forecasts and observations

We determine daily Tmax and Tmin data from 6-hourly gridded temperature forecasts, and ensure the daily forecasts are properly aligned with daily observations. In Australia, Tmax and Tmin in the 24 hours are recorded at 9 am local time. 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 for data alignment. Tmax/Tmin forecasts retrieved at 06 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, 06, 12, 18 UTC on the 3rd and 00 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 eastern Australia is 10 am 3rd to 10 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 h across Australia.

3.2 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.

Author Manuscript

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,

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

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. (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 the 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 the corresponding fitted model to validate all the omitted data points. After repeating the cross validation runs 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 and will be handled by the calibration method described in Section 3.3.

Author Manuscript

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 2020, we train the calibration model using the first week raw re-forecast anomalies issued from all February initialisation dates over the period from 2001 to 2019, and corresponding observation anomalies falling between 1991 and 2019. 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.

3.3 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 (Shao et al., 2021a) 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 mean 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,

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

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).

Author Manuscript

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,

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

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

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

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 centered at zero and have an empirically estimated variance, formulated as,

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

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 Shao et al. (2021c) 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 (Shao et al., 2021a, 2021b), 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

Author Manuscript

each grid cell. To determine $m_i^{"}$ for individual cases, we select all the values from the cells within a 7-cell by 7-cell region centered 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 75th 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 Shao et al. (2021b).

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 $\mathbf{D} = \left\{\left(y_1^{'}(t), y_2^{'}(t)\right), t = 1, 2, \ldots, T\right\}$. 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,

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

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 Shao et al. (2021a) and Shao et al. (2021b).

The values of missing variables are sampled from the conditional distribution,

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

where

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

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

(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 $\boldsymbol{\theta}$

$\boldsymbol{\theta}$, we sample a trend-embedded calibrated forecast $y_2^{'}(t^*)$

$y_1^{'}(t^*)$. That is, we set the predictand as having a missing value and formulate a Gibbs sampler to sample a new calibrated forecast value $z_2(t^*)$ based on the conditional

Author Manuscript

distribution of the predictand given the predictor $z_1(t^*)$ as formulated in Eq. (7) – (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 study, 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 Shao et al. (2021a) and Shao et al. (2021b).

3.4 Forecast verification

The cross-validated BJP-ti calibrated ensemble forecast anomalies are verified against raw and BJP calibrated ensemble 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 trends (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 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 (CRPS) (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,

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

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 the reference forecasts to obtain the continuous ranked probability skill score (CRPSS) for each cell. The reference forecasts are leave-one-year-out cross-validated climatology ensemble forecasts produced using the BJP model, based on the marginal distribution of the predictand, which is observation anomaly in this work (Wang et al., 2019). The CRPSS is given as,

$$\mathrm{CRPSS} = \frac{\mathrm{CRPS}_{\mathrm{ref}} - \mathrm{CRPS}}{\mathrm{CRPS}_{\mathrm{ref}}} \times 100(\%) \tag{11}$$

A 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 (PITs) (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,

$$\pi(t) = F(t, y_o(t)) \tag{12}$$

The PIT score has the form,

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

where $\pi(k)$ is the k^{th} ranked PIT value $\pi(t)$ and $\frac{k}{n+1}$ 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.

Author Manuscript

Author Manuscript

4 Result

4.1 Trends in observation 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 1 and Figure 2, while the week-4 trends are presented in Figure 3 and Figure 4 respectively. The trend maps for week-2 and week-3 variables are shown in the Supporting Information (Figure S1-S4).

4.1.1 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 1). Strong warming trends at the rate of larger than 1.2 °C per decade are also statistically significant at 5% level (Figure S5) 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 1) generally exhibit different spatial patterns of directions, magnitudes, and statistical significance (Figure S6). 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.

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 2), with many of the increasing trends statistically significant at 5% level (Figure S7 and Figure S8). 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%

Author Manuscript

level (Figure S8). 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.

Consistent trends in raw and BJP calibrated week-1 forecasts (third and fourth columns in Figure 1 and Figure 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 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 1 and Figure 2). For example, BJP-ti calibrated forecasts reproduce strong observed warming trends apparent 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, 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). 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. 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), 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 Tmax forecasts as shown in Figure 6 in Section 4.2.1. 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.

Author Manuscript

4.1.2 Week-4 observed and forecast trends

Week-4 observations also exhibit warming and cooling trends, but the trend behaviors are distinct from the week-1 observed trends over two evaluation periods (first and second column in Figure 3 and Figure 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 (Figure S5). 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 (Figure S7). 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.

Compared to week-1 forecasts, trends in both raw and BJP calibrated forecasts (third and fourth column in Figure 3 and Figure 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 3 and Figure 4). The possible reason can again be explained using the case example shown in Figure 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 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 CRPSS 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 the climatology forecasts over 2000-2019 but have lower variability than BJP-ti calibrated week-1 forecasts.

Author Manuscript

4.2 Skill scores for model forecasts

The CRPSS 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 6 and Figure 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 Supporting Information (Figure S9 and Figure S10). In addition, we apply a bootstrap method (Schepen et al., 2016; Shao et al., 2021a) to check whether the BJP-ti calibration significantly improves or worsens the CRPSS compared to BJP at 5% significance level for all lead times. Results of the score significance are also presented in the Supporting Information (Figure S11 and Figure S12).

4.2.1 Skill of week-1 forecasts

Raw week-1 forecasts (first column of both panels in Figure 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.

With BJP-ti calibration (second column of both panels in Figure 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 the BJP calibrated forecasts (third column of both panels in Figure 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 1 and Figure 2). For example, statistically significant skill gains higher than 10% are seen along the coastal areas in the east for Tmin in November (Figure S11). 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 2 and Figure S12).

The BJP-ti calibration leads to slight and not statistically significant skill declines (i.e., score difference smaller than 5%) relative to BJP (Figure S11 and Figure S12), particularly in the regions where trends in the BJP calibrated forecasts are highly consistent with the 20-year observations (Figure 1 and Figure 2). Examples are parts of northern Australia for Tmin and central Australia for Tmax, both in May.

4.2.2 Skill of week-4 forecasts

For both Tmin and Tmax, skill scores of raw week-4 forecasts (first column of both panels in Figure 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 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 the 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 the BJP calibrated forecasts (third column of both panels in Figure 7). For example, statistically significant skill score increases (Figure S11 and S12) are evident in some clusters of western Australia in November for both Tmin and Tmax. Despite in different magnitude, in these regions, the 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.

Skill declines by embedding the 30-year observed trends when the trends in the 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

Author Manuscript

Author Manuscript

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.

4.2.3 Overall skill of forecasts

Results of the averaged CRPSS of the 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 8. As with the score maps shown in Figure 6 and Figure 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.

Using the BJP-ti calibration is effective at improving forecast skill compared to raw forecasts (second column in both panels of Figure 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.

4.3 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 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

Author Manuscript

forecasts are already highly reliable in ensemble spread, and the reliability could not be further improved by BJP-ti post-processing.

5 Discussion

In this study, 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 1-4, 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 a 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 CRPSS 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.

Author Manuscript

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 when there are only 19 raw forecast points from a single date 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.

Shao et al. (2021c) extended and applied the trend-aware model 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 an 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.

Author Manuscript

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 community and decision makers are expected to obtain calibrated real-time forecasts in a timely manner.

6 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 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.

7 Acknowledgement

This study is supported by an ARC Linkage Project (LP170100922) funded by the Australian Research Council. We thank the European Centre for Medium-Range Weather Forecasts (ECMWF) for supplying sub-seasonal data archived on the MARS system. The AWAP datasets are freely available on the website of Australian Government Bureau of Meteorology (http://www.bom.gov.au/jsp/awap/temp/). We thank two anonymous reviewers for constructive comments.

8 Supporting Information

Figure S1: Decadal trends for Tmin observations over 2000-2019 and 1990-2019, raw, BJP and BJP-ti calibrated week-2 forecasts over 2000-2019 for all initialisation dates within February, May, August, and November separately.
Figure S2: The same as Figure S1, but for Tmax.
Figure S3: The same as Figure S1, but for week-3 forecasts.
Figure S4: The same as Figure S3, but for Tmax.
Figure S5: 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 at 1% and 5% significance level using two-tailed student t-test.
Figure S6: The same as Figure S5, but for Tmin observations over 1990-2019.
Figure S7: The same as Figure S5, but for Tmax observations.
Figure S8: The same as Figure S6, but for Tmax observations.
Figure S9: CRPSS 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.

Author Manuscript

Figure S10: The same as Figure S9, but for week-3 forecasts.

Figure S11: Statistical significance of the improvement or worsening of the CRPSS of the BJP-ti calibrated forecasts compared to BJP for Tmin at 5% significance level.

Figure S12: The same as Figure S11, but for Tmax.

9 Reference

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

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, and Rubin DB. 2014. *Bayesian data analysis* (Third edition ed.): CRC press.

Hartmann DL, Klein Tank AMG, Rusticucci M, Alexander LV, Brönnimann S, Charabi YAR, . . . Zhai P. 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.

Hersbach H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting, 15*(5), 559-570.

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 HH, Lim EP, Liu GQ, Luo JJ., . . . Zhou XB. 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: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.

Hudson D, Marshall AG, Yin Y, Alves O, and Hendon HH. 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

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

Krakauer NY. 2017. Temperature trends and prediction skill in NMME seasonal forecasts. *Climate Dynamics*, 1-13. doi:https://doi.org/10.1007/s00382-017-3657-2

Author Manuscript

Author Manuscript

Li Y, Wu Z, He H, Wang QJ, 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

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:10.1007/s00382-013-2016-1

Matheson JE, and Winkler RL. 1976. Scoring Rules for Continuous Probability Distributions. *Management Science, 22*(10), 1087-1096. doi:10.1287/mnsc.22.10.1087

Merryfield WJ, Baehr J, Batté L, Becker EJ, Butler AH, Coelho CA, . . . Domeisen DI. 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

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

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: https://doi.org/10.3390/atmos11080823

Renard B, Kavetski D, Kuczera G, Thyer M, and Franks SW. 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research, 46*. doi:10.1029/2009wr008328

Saha S, Moorthi S, Wu X, Wang J, Nadiga S, Tripp P, . . . 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

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

Schepen, A., Zhao TTG, Wang QJ, and Robertson DE. 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:10.5194/hess-22-1615-2018

Scheuerer M, Switanek MB, Worsnop RP, and Hamill TM. 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

Author Manuscript

Shao Y, Wang QJ, 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/10.1175/10.1002/joc.6788

Shao Y, Wang QJ, Schepen A, and Ryu D. 2021b. Going with the trend: forecasting seasonal climate conditions under climate change. *Monthly Weather Review*, *149*, 2513-2522. doi: https://doi.org/10.1175/MWR-D-20-0318.1

Shao Y, Wang QJ, Schepen A, Ryu D, and Pappenberger F. 2021c. Improved trend-aware post-processing of GCM seasonal precipitation forecasts. *Journal of Hydrometeorology*. published-online. doi: https://doi.org/10.1175/JHM-D-21-0099.1.

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

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*, 146, 2654-2670. doi:10.1002/qj.3810

Vigaud N, Tippett MK, Yuan J, Robertson AW, 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: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, . . . 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 AW. 2019. Introduction: Why Sub-seasonal to seasonal prediction (S2S)? In *Sub-seasonal to seasonal prediction* (pp. 3-15): Elsevier.

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

Wang QJ, and Robertson DE. 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 QJ, Robertson DE., and Chiew FHS. 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 QJ, Shao YW, Song Y, Schepen A, Robertson DE, 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

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, 2263-2273. doi: https://doi.org/10.1175/BAMS-D-15-00267.1

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

Zhao T, Wang QJ, and Schepen A. 2019. 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

Author Manuscript

Author Manuscript

Figure captions

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

Figure 2: The same as Figure 1, but for Tmax.

Figure 3: The same as Figure 1, but for week-4 forecasts.

Figure 4: The same as Figure 3, but for Tmax.

Figure 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.

Figure 6: CRPSS 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.

Figure 7: The same as Figure 5, but for week-4 forecasts.

Figure 8: Averaged CRPSS 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.

Figure 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.

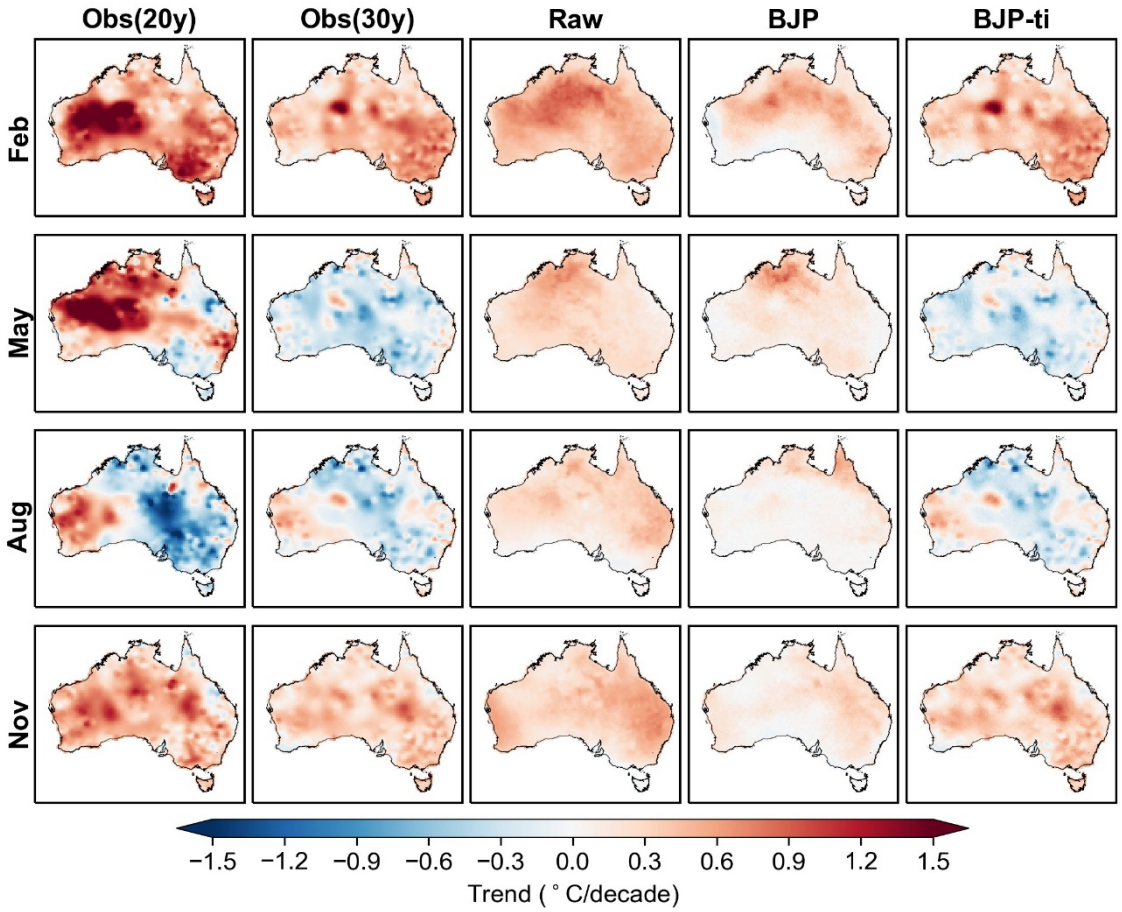
Author Manuscript

Introducing Long-term Trends into Sub-seasonal Temperature Forecasts through Trend-aware Post-processing

Yawen Shao[a*], Q. J. Wang[a], Andrew Schepen[b], Dongryeol Ryu[a]

a. Department of Infrastructure Engineering, The University of Melbourne, Parkville 3010, Australia

b. CSIRO Land and Water, Dutton Park 4102, Australia



This work identifies that 20-year ECMWF retrospective forecasts of sub-seasonal temperatures often fail to reproduce the observed trends. Subsequently, we extend a trend-aware forecast post-processing model, BJP-ti, to calibrate and introduce the 30-year climate trends into the sub-seasonal forecasts. In this figure, week-4 BJP-ti calibrated forecasts of minimum temperatures now exhibit the trends of the 30-year observations in most regions. This extended trend-aware

method has the potential to deliver high-quality sub-seasonal forecasts and support decision-making in a changing climate.

Author Manuscript

Author Manuscript

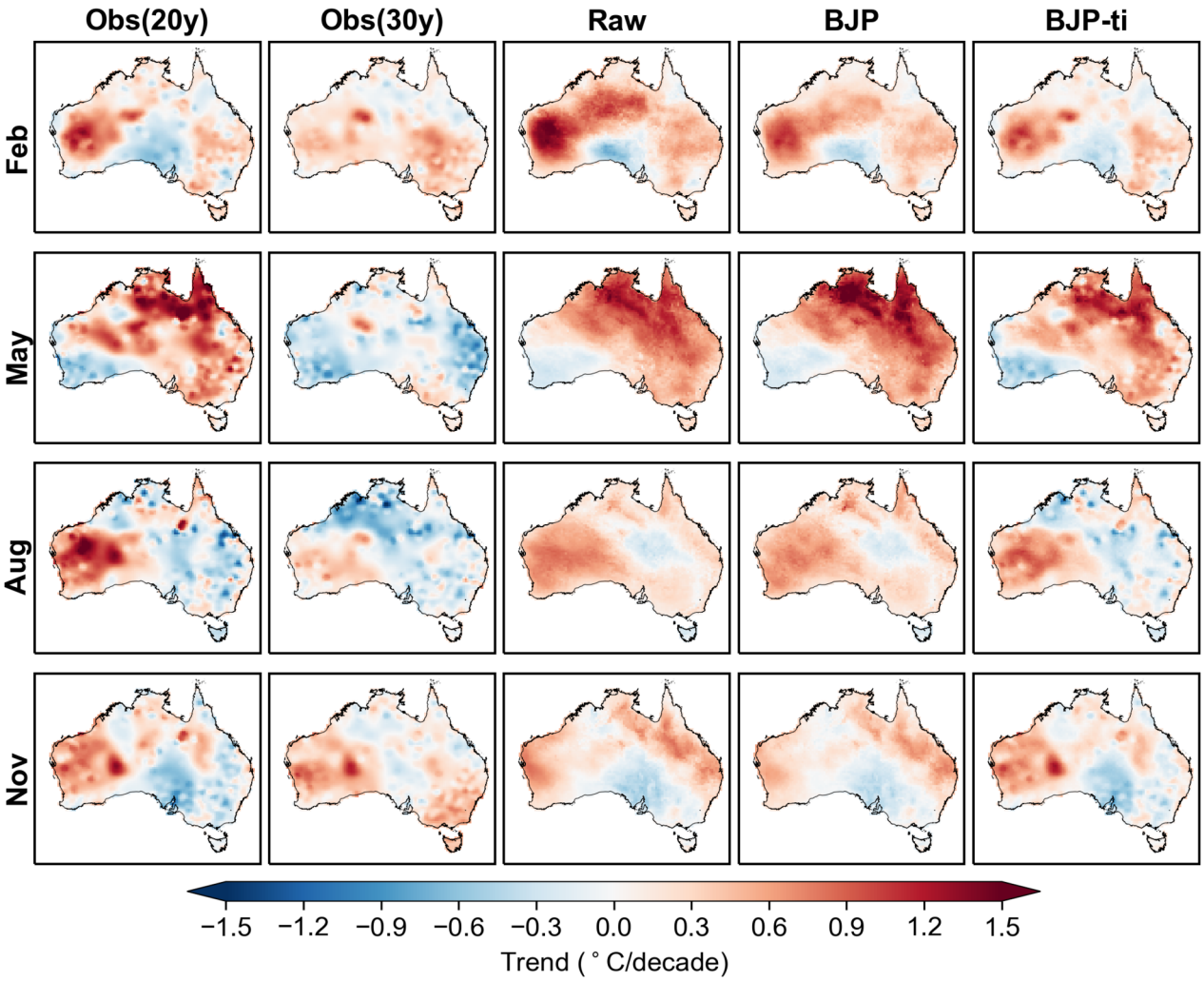


Figure 1.tif

Author Manuscript

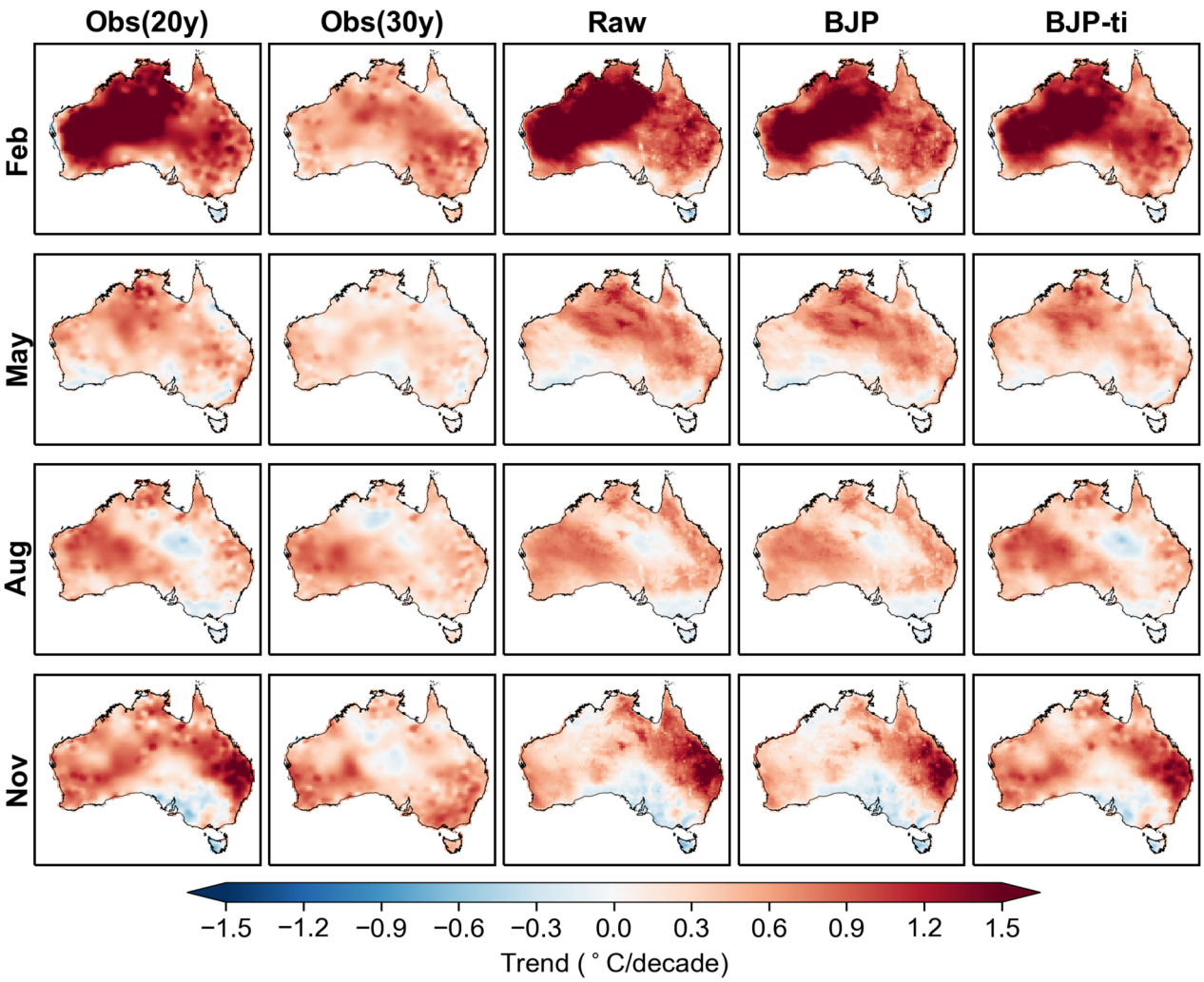


Figure 2.tif

Author Manuscript

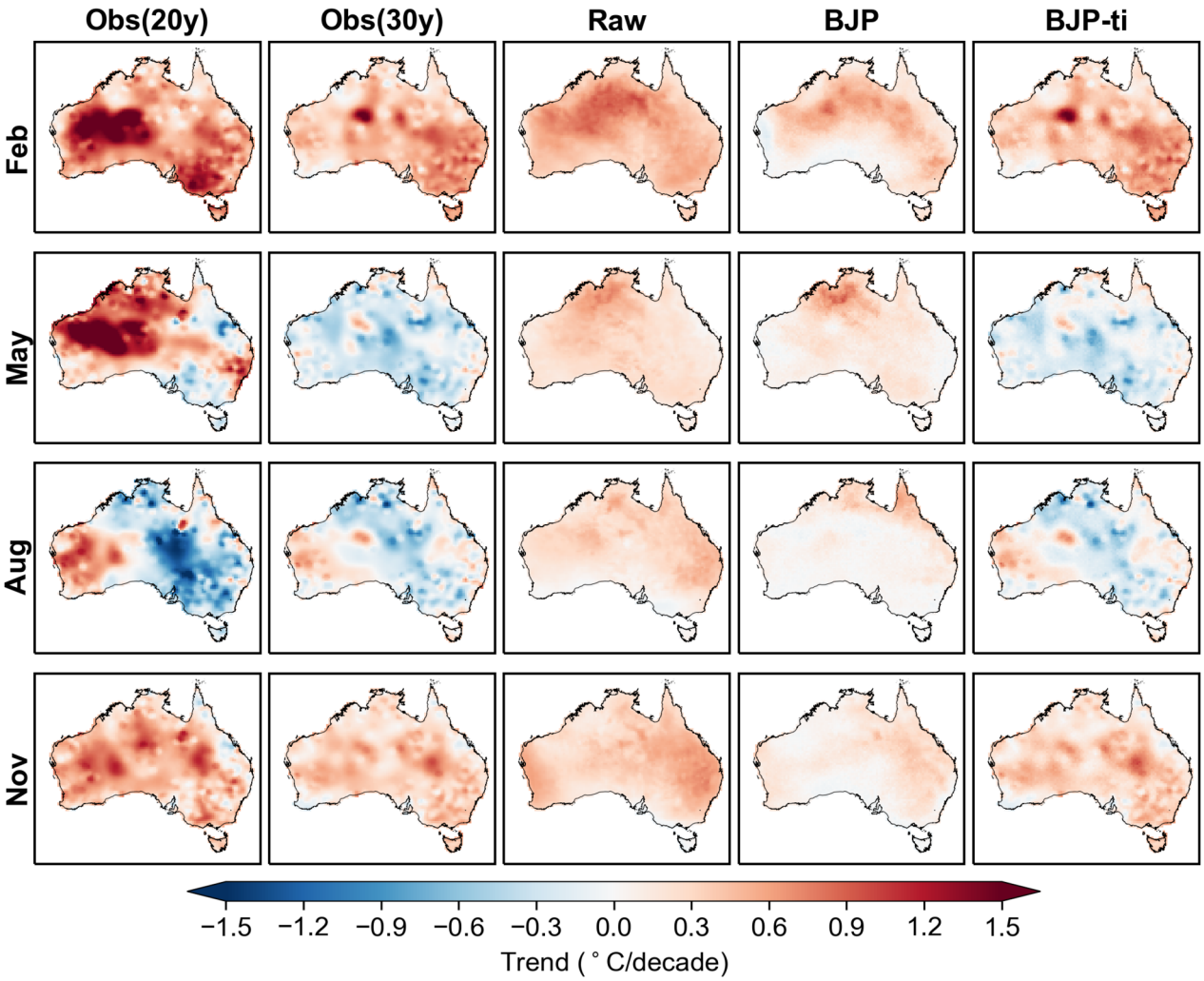


Figure 3.tif

Author Manuscript

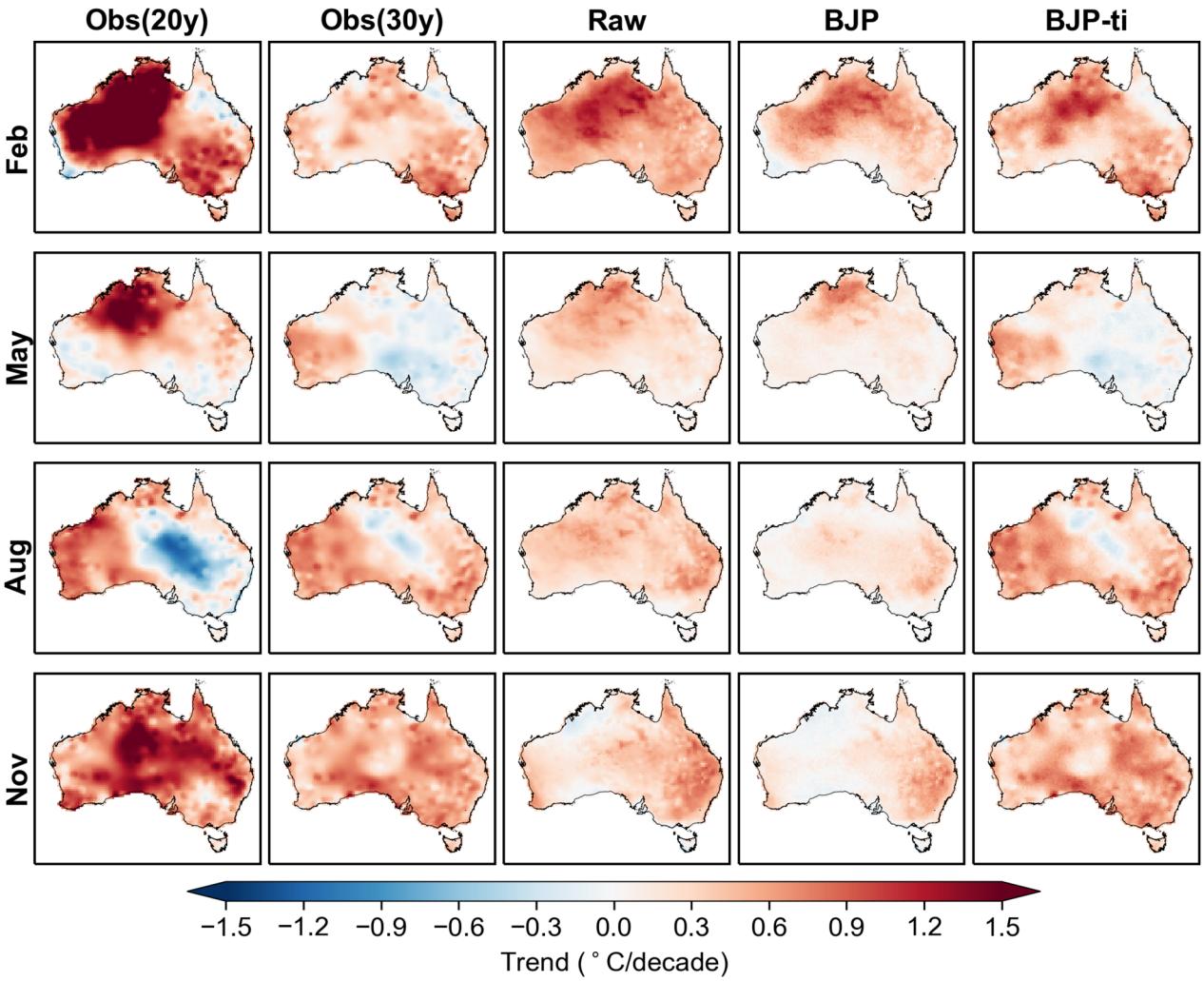


Figure 4.tif

Author Manuscript

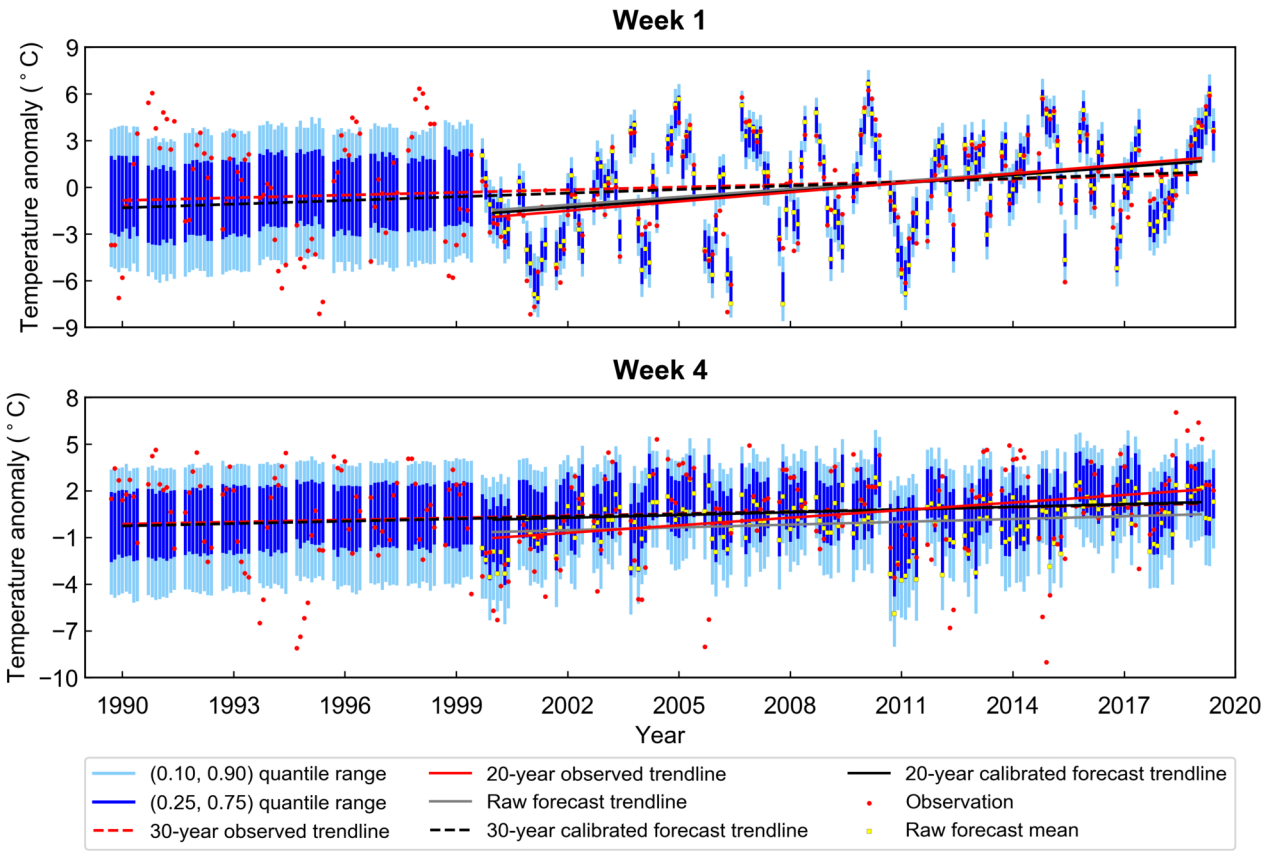


Figure 5.tif

Author Manuscript

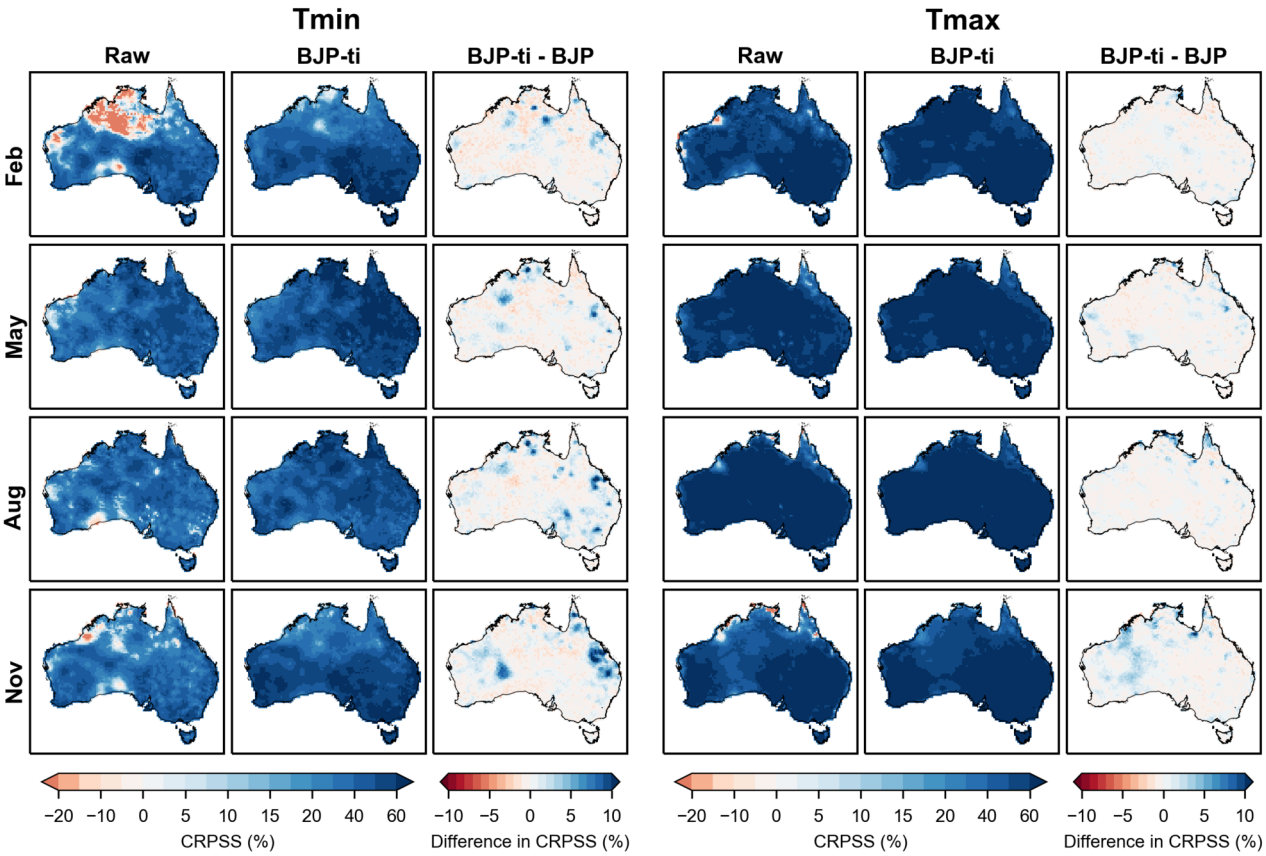


Figure 6.tif

Author Manuscript

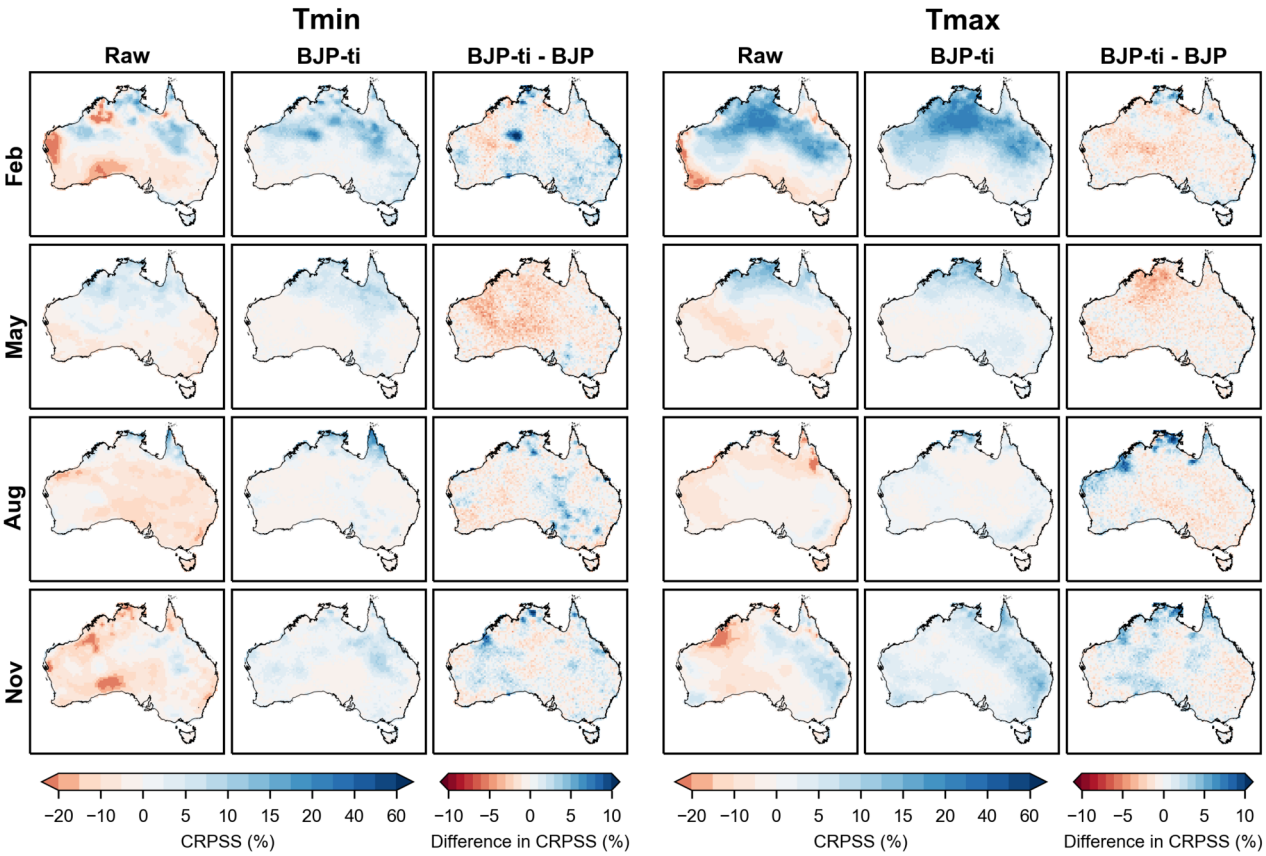


Figure 7.tif

Author Manuscript

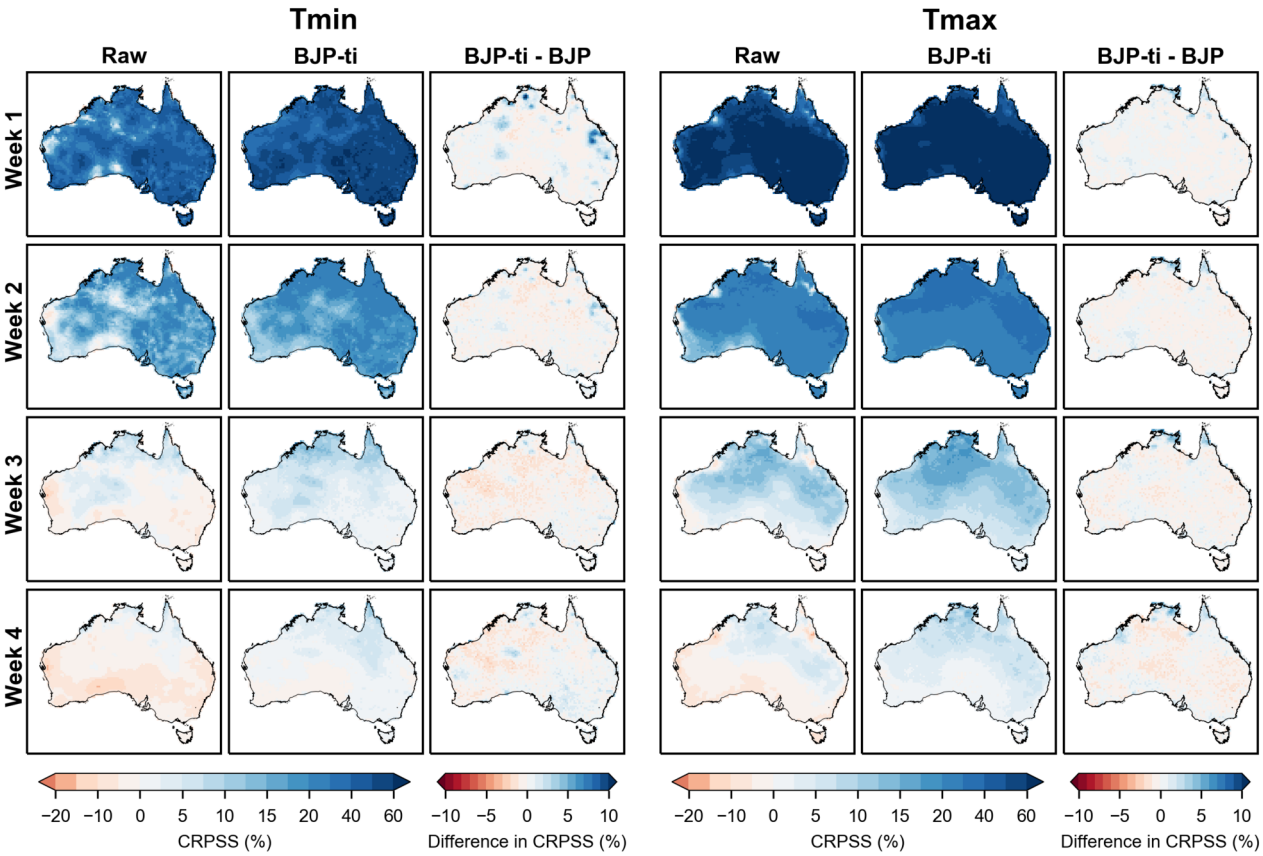


Figure 8.tif

Author Manuscript

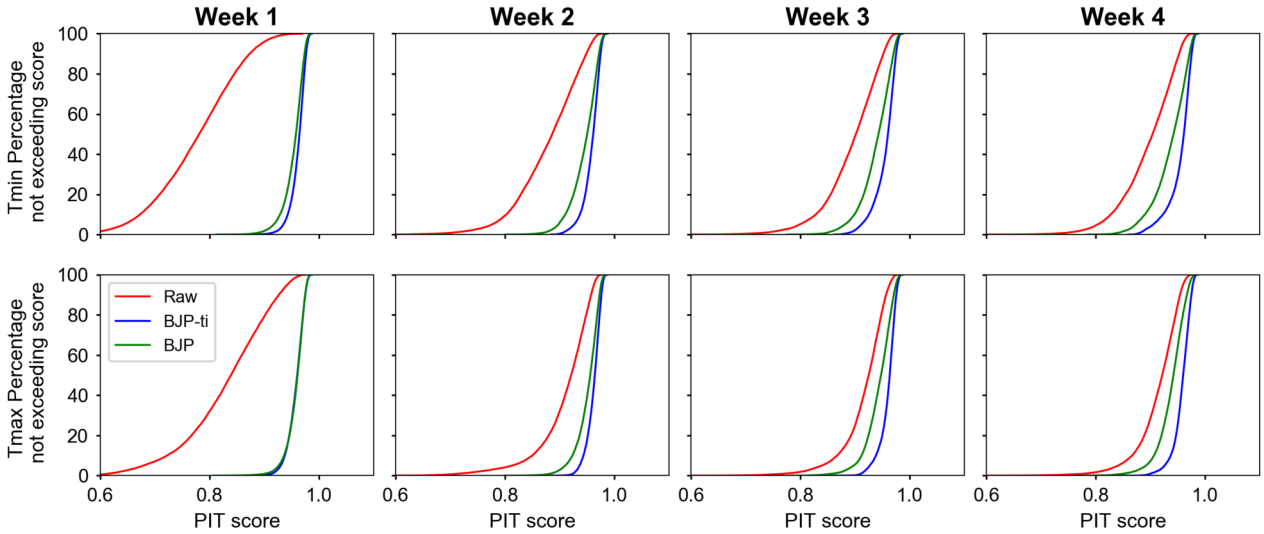


Figure 9.tif