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

Title:
Embedding trend into seasonal temperature forecasts through statistical calibration ofGCMoutputs

Date:
2021-01

Citation:
Shao, Y., Wang, Q. J., Schepen, A. & Ryu, D. (2021). Embedding trend into seasonal temperature forecasts through statistical calibration ofGCMoutputs. International Journal of Climatology, 41 (S1), pp.E1553-E1565. https://doi.org/10.1002/joc.6788.

Persistent Link:
https://hdl.handle.net/11343/276219

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

Embedding trend into seasonal temperature forecasts through statistical calibration of GCM outputs

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: temperature trend, seasonal forecasting, ensemble forecast calibration, forecast verification

Funding information: ARC Linkage Project (LP170100922) funded by the Australian Research Council and industry partners.

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.6788

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 the raw ensemble forecasts and the BJP calibrated ensemble forecasts both fail to do so. The BJP-t calibration leads to more accurate, more skilful, more reliable and sharper forecasts than the BJP calibration.

1 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 (Ghasemi, 2015; Caloiero, 2017; Yu et al., 2018; CSIRO and Australian Government Bureau of Meteorology, 2018; Li et al., 2019). 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 the 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 (2017) 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. (2005) 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 postprocessed to remove bias, to quantify a reliable ensemble spread of the forecasts, and to make the forecasts more skilful than climatology (Doblas-Reyes et al., 2013; Barnston et al., 2015). 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 (Schepen and Wang 2014; Peng et al., 2014; Schepen et al., 2016; Zhao et al., 2017, 2019; Strazzo et al., 2019; Wang et al., 2019), the Bayesian method, along with other post-processing techniques, is not designed to embed the observed trend into calibrated seasonal forecasts.

In 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 skilful, 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 the 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 the 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 (Krikken et al., 2016; Dirkson et al., 2019; Director et al., 2019).

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 paper is structured as follows: Section 2 introduces SEAS5 forecasts and station data used in this study, and elaborates the BJP methods, forecast evaluation and verification tools. Section 3 presents the results. Section 4 further discusses the results. Section 5 summarizes the study and points to the main conclusions.

2 Data and methods

2.1 SEAS5 forecasting data and station data

This study uses ensemble reforecasts 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 approximately 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 reforecasts is ERA-Interim. The ocean and sea-ice initial conditions for reforecasts 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 reforecasts comprises 25 members and the reforecasts are initialised on the 1st of every month from 1 January 1981 to 1 December 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 reforecasts initialised on the 1st of December each year to obtain the reforecasts of January in 1982-2017.

We select three weather stations (Table 1) located in different states in Australia (Figure 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 (Mann, 1945; Kendall, 1975) 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.

2.2 Bayesian modelling method

The Bayesian statistical model developed in this study is an extension of the Bayesian joint probability (BJP) modelling approach (Wang et al., 2009; Wang and Robertson, 2011). The BJP model has been widely applied to calibrate seasonal climate forecasts (Schepen and Wang, 2014; Schepen et al., 2016; Zhao et al., 2019). 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, which uses a Bayesian inference to model trend parameters in observations and raw forecasts, so that the trend uncertainty is considered. Given limited availability of the data, the trend uncertainty can be large and should not be neglected. Therefore, our BJP-t algorithm should be more robust than an approach by 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.

2.2.1 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 (Yeo and Johnson 2000; Wang et al.; 2009; Wang and Robertson, 2011). For the variable y,

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

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 posterior (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,

$$z_1(t) = y_1'(t) - \alpha_1(t - t_m) \quad (2)$$

$$z_2(t) = y_2'(t) - \alpha_2(t - t_m) \quad (3)$$

where the trend parameter α_1 is for raw forecasts, α_2 is for observations, and t is the forecast year, t_m is 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,

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

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

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

and

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

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,

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

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

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

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

2.2.2 Parameter inference

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

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

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,

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

where

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

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

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. (9) – Eq. (12) as,

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

$$[\boldsymbol{\mu} \mid \cdot] = \text{N}(\bar{\mathbf{z}}, \boldsymbol{\Sigma} / n) \tag{14}$$

where

$$\mathbf{S} = \sum_{t=1}^{n} (\mathbf{z}(t) - \bar{\mathbf{z}})(\mathbf{z}(t) - \bar{\mathbf{z}})^{\mathrm{T}} \tag{15}$$

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

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), Eq. (3) and Eq. (8) as,

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

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

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

where

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

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

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

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

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

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 Eq. (2) – (3))
2. Sample $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ in sequence (see Eq. (13) – (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) – (3), Eq. (18) – (21))
4. Sample α_i (see Eq. (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.

2.2.3 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

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

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 calibrated ensemble forecast of $y_2(t^*)$ by back-transforming each of the sample members to the original space using inverse Yeo-Johnson transformation (see Eq. 23).

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

2.3 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., 2020). 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 et al. (1993) suggests that leaving out more years can 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 to 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

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

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

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

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 small spread when the forecast is accurate (Wilks, 2006). We then convert the CRPS value to a skill score to measure the relative improvement of the model forecasts compared to reference forecasts which is 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:

$$\mathrm{CRPS}_{\mathrm{ss}} = \frac{CRPS_{ref} - CRPS}{CRPS_{ref}} \times 100 \tag{26}$$

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 have errors 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 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,

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

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 the raw and BJP calibrated forecasts and test the significance at the 5% significance level. In this regard, if the BJP-t calibrated value 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.

3 Result

3.1 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). 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 forecast 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. 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 3.2).

Numerically, the multi-decadal linear trends are calculated for the observations and ensemble forecast means at three weather stations (Table 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.

3.2 Forecast accuracy and skill

The RMSE values and CRPS skill scores of the ensemble forecasts are presented in Table 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 the 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 the raw forecasts have some skills and are better at reproducing the observed trend slope compared to the other two cases. In this regard, the correction of the trend amplitude may not have salient impacts on further skill improvement.

3.3 Reliability and sharpness

The PIT index and PIT uniform probability plot of the forecasts for three stations are shown in Table 4 and Figure 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 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, the raw forecasts are found sharper than the BJP and BJP-t calibrated forecasts. Nevertheless, the raw ensemble forecasts do not exhibit higher reliability and skill, indicating that they are not corresponding well to the observations.

4 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 Figure S1). 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 paper, 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 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 the 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, 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.

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.

5 Conclusion

Climate-sensitive industries, such as water, energy and agriculture sectors, often require skilful 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 paper 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 the raw and BJP calibrated forecasts have low skills, the BJP-t calibration results in skilful forecasts. The BJP-t model does not significantly improve the forecast performance where the raw and BJP calibrated forecasts are already skilful 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. It is expected that this work will be applied to the continental scale and tailored for other meteorological variables in the future.

6 Acknowledgement

We thank Dan Collins and Sarah Strazzo from NOAA Climate Prediction Centre for useful discussion on conceiving the initial idea of this work. This study is linked to an ARC Linkage Project (LP170100922) funded by the Australian Research Council and industry partners. We thank the European Centre for Medium-Range Weather Forecasts (ECMWF) for supplying the

SEAS5 data. The weather station data are available on the website of Australian Government Bureau of Meteorology: http://www.bom.gov.au/climate/data/.

7 Supporting information

Additional supporting information for Figure S1 may be found online in the Supporting Information section.

8 References

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

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:10.1175/Jamc-D-14-0188.1

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:10.1007/s00704-016-1764-3

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.

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

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:10.1002/wcc.217.

Gelman A, Carlin J B, Stern H S, Dunson D B, Vehtari A and Rubin D B. 2014. Bayesian Data Analysis, third ed. CRC press, Boca Raton.

Ghasemi A R. 2015. Changes and trends in maximum, minimum and mean temperature series in Iran. *Atmospheric Science Letters*, 16(3), 366-372. doi: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: 10.1111/j.1467-9868.2007.00587.x.

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

Hartmann D L, Klein Tank A M G, Rusticucci M, Alexander L V, Brönnimann S, Charabi Y A R, . . . 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 Forecasting*, 15, 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Jia G, Shevliakova E, Artaxo P, De Noblet-Ducoudré, N, Houghton R, House J, Kitajima K, Lennard C, Popp A, Sirin A, Sukumar R, Verchot L. 2019. Land–climate interactions. In: *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 [P.R. Shukla, J. Skea, E. Calvo Buendia, V. Masson-Delmotte, H.-O. Pörtner, D.C. Roberts, P. Zhai, R. Slade, S. Connors, R. van Diemen, M. Ferrat, E. Haughey, S. Luz, S. Neogi, M. Pathak, J. Petzold, J. Portugal Pereira, P. Vyas, E. Huntley, K. Kissick, M, Belkacemi, J. Malley, (eds.)]. In press.

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: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, Balsamo G, Keeley S P E, Mogensen K, Zuo H and Monge-Sanz B M. 2019. SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, 12, 1087–1117. https://doi.org/10.5194/gmd-12-1087-2019.

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

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: https://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: https://doi.org/10.1175/MWR-D-16-0337.1

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

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

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*, 1-9. doi: 10.1002/joc.5868

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

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

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:10.5194/gmd-11-351-2018

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, 7116-7135. http://doi.org/10.1002/2013JD021162

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:10.1029/2009wr008328

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:10.1175/Jcli-D-15-0868.1

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:10.1016/j.jhydrol.2014.03.017

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: https://doi.org/10.1175/Mwr-D-13-00248.1

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:10.1175/Mwr-D-15-0384.1

Schepen A, Zhao T, 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 Science*, 22, 1615–1628. https://doi.org/10.5194/hess-22-1615-2018.

Schepen A, Everingham Y and Wang Q J. 2020. 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.

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:10.1175/Mwr-D-18-0156.1

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

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: 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: 10.1029/2008WR007355

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

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:10.1029/2010gl046455

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

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

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: 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: 10.1038/s41598-018-21817-9

Zhao T T G, 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:10.1175/Jcli-D-16-0652.1

Zhao T T G, Wang Q J, Schepen A and Griffiths M. 2019. Ensemble forecasting of monthly and seasonal reference crop evapotranspiration based on global climate model outputs. *Agricultural and Forest Meteorology*, 264, 114-124. doi: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

Figure caption

Table 1: Details of the weather stations.

Table 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.

Table 3: The RMSE and CRPS skill score for raw forecasts, BJP calibrated forecasts, and BJP-t calibrated forecasts in three stations. The symbol * denotes the significant change of the BJP-t results compared to the raw and BJP calibrated forecasts.

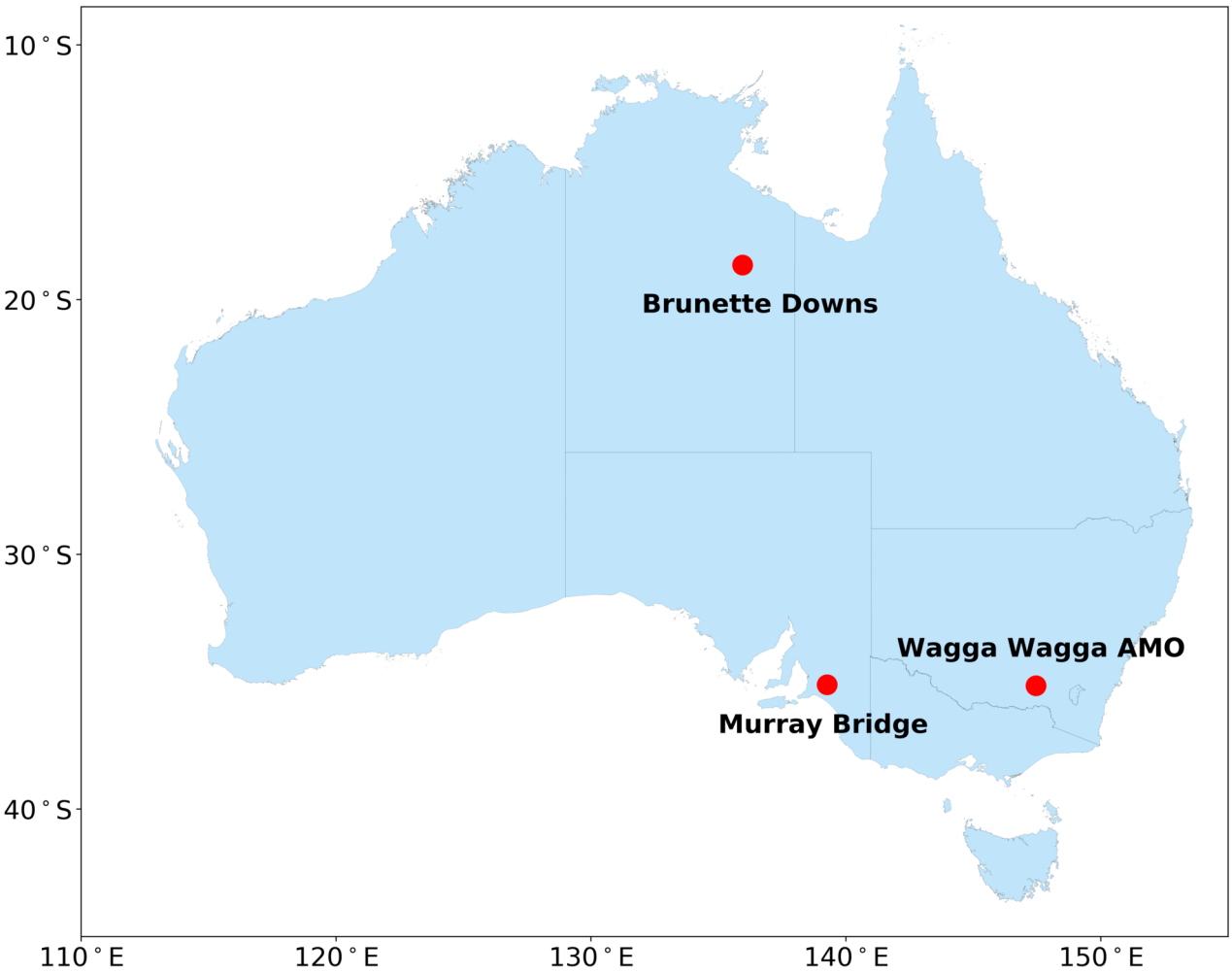
Table 4: The PIT index and average widths of central prediction intervals (50% and 90%) for raw forecasts, BJP calibrated forecasts, and BJP-t calibrated forecasts in three cases. The symbol * denotes the significant change of the BJP-t results compared to the raw and BJP calibrated forecasts.

Figure 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.

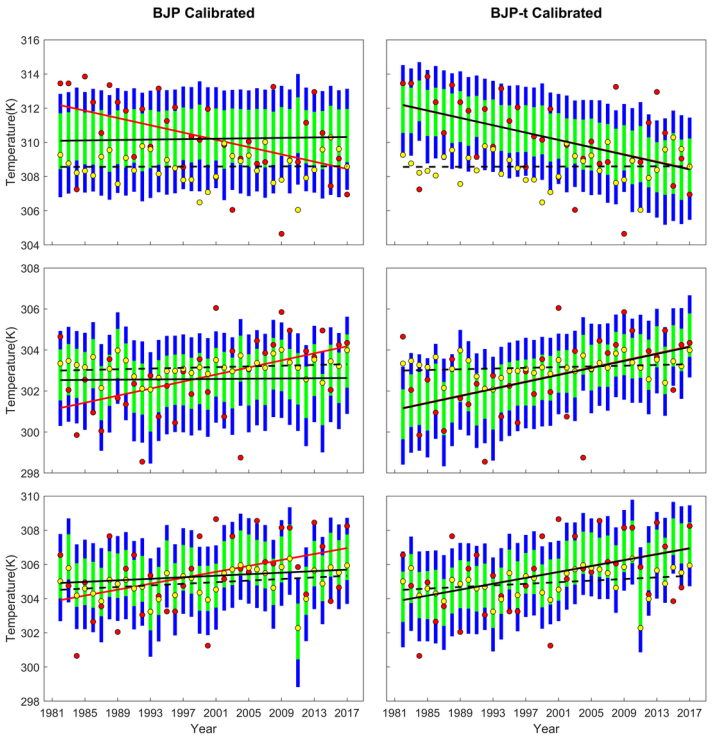
Figure 2: Forecast quantiles of cross-validated Tmax forecasts and observed values plotted for BJP calibrated forecasts (left) and BJP-t calibrated forecasts (right). The first row is Brunette Downs Station, the second row is Murray Bridge Station, and the third row is Wagga Wagga AMO Station. The red 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.

Figure 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

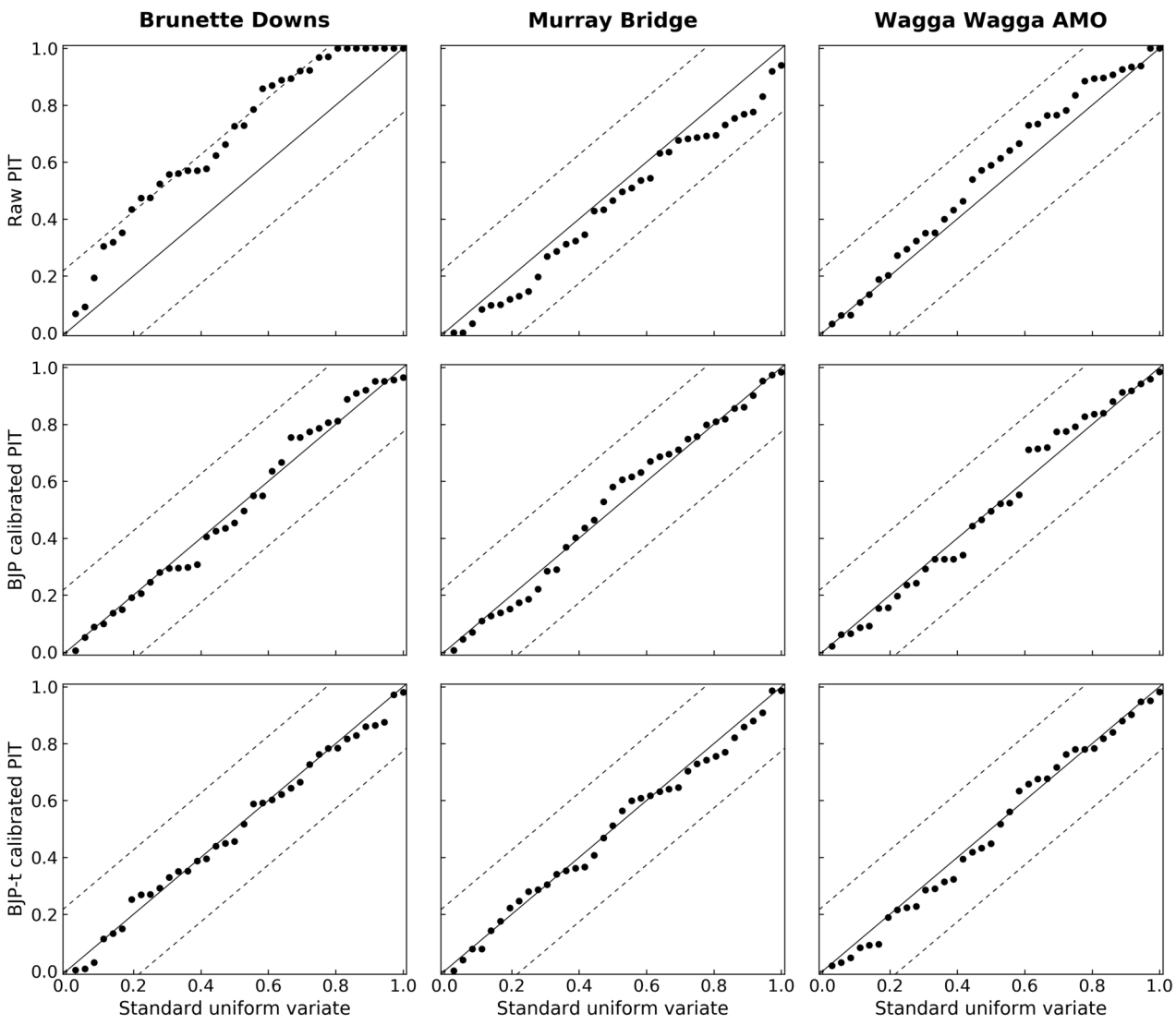
PIT values of observed Tmax; solid line is 1:1 uniform distribution; dashed line is Kolmogorov 5% significance band.



JOC_6788_Figure_1.tif



JOC_6788_Figure_2.tif



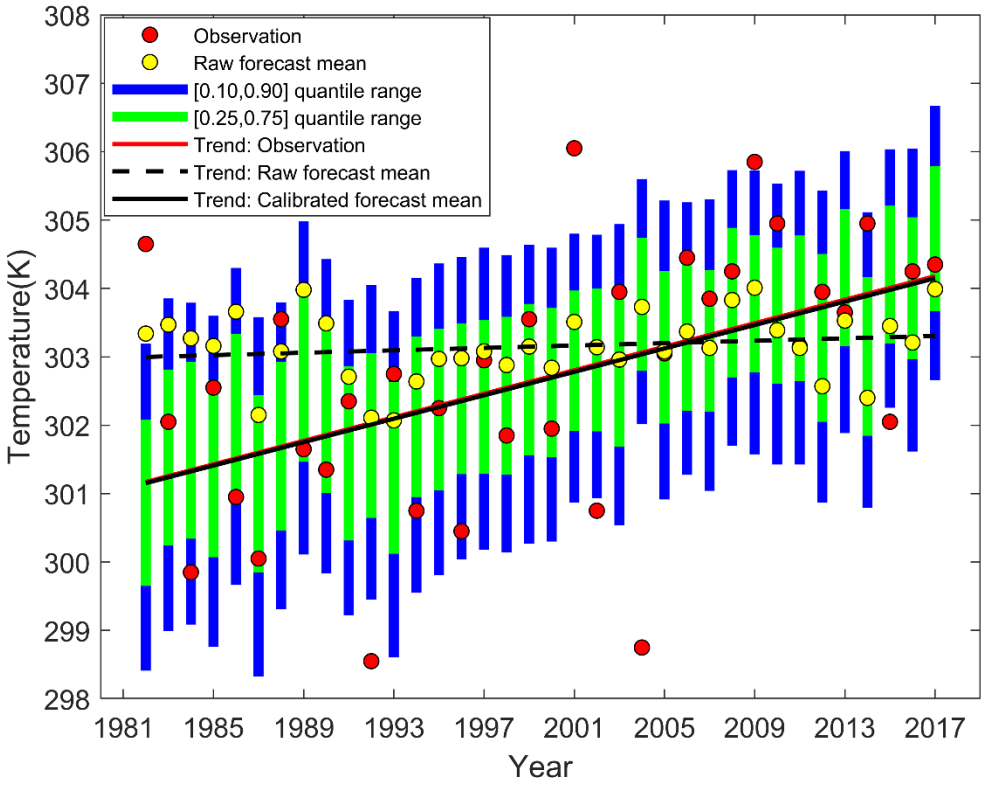
JOC_6788_Figure_3.tif

Embedding trend into seasonal temperature forecasts through statistical calibration of GCM outputs

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



Accurate and reliable seasonal temperature forecasts are often sought by industries and governments for managing climate variability and change. The land surface temperature has exhibited marked temporal trends, but such trend is generally underestimated by raw forecasts (from the global climate model) or post-processed forecasts. This work resolves this problem by introducing trend components into a statistical model. As visualised in the image, the new model is able to accurately embed the observed temperature trend into calibrated ensemble forecasts.

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

Table 1: Details of the weather stations.

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

Table 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
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

Table 3: The RMSE and CRPS skill score for raw forecasts, BJP calibrated forecasts, and BJP-t calibrated forecasts in three stations. The symbol * denotes the significant change of the BJP-t results compared to the raw and BJP calibrated forecasts.

		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

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