Author/s:
Yang, Q;Wang, QJ;Hakala, K

Title:
Calibrating anomalies improves forecasting of daily reference crop evapotranspiration

Date:
2022-07

Citation:
Yang, Q., Wang, Q. J. & Hakala, K. (2022). Calibrating anomalies improves forecasting of daily reference crop evapotranspiration. Journal of Hydrology, 610, https://doi.org/10.1016/j.jhydrol.2022.128009.

Persistent Link:
https://hdl.handle.net/11343/332619

Calibrating anomalies improves forecasting of daily reference crop evapotranspiration

Qichun Yang[a,*], Quan J Wang[a], and Kirsti Hakala[a]

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

*: Corresponding author

E-mail address: qichun.yang@unimelb.edu.au

Telephone number: +61 411359526

Abstract

Forecasting of short-term reference crop evapotranspiration (ET_o) provides valuable information for hydrological, agricultural, and ecological applications. ET_o forecasts can be derived from weather forecasts of Numerical Weather Prediction (NWP) models, but such raw forecasts need to be calibrated to correct errors and improve reliability. This study calibrates the short-term ET_o forecasts constructed with weather forecasts from the Australian Bureau of Meteorology's Australian Community Climate and Earth-System Simulator G2 version (ACCESS-G2) model, using the recently developed Seasonally Coherent Calibration (SCC) model. The monthly parameterization of the SCC model will not be able to capture the increasing or decreasing trends in ET_o at the submonthly scale, posing a challenge for effective forecast calibration. To address this challenge, we developed a new calibration strategy based on ET_o anomalies and climatological mean. We thoroughly evaluated this strategy at both the continental and weather station scales. Results indicate that calibrating ET_o anomalies improves the correlation coefficient between calibrated forecasts and observations by up to 10%, and increases forecast skill scores by up to 200%, with more significant improvements found at longer lead times (9-day lead time vs. 1-day lead time). Improvements in forecast quality in calibrations across two spatial scales based on different types of observations (gridded observations vs. weather station observations) validate the effectiveness and robustness of the developed strategy. We anticipate that this strategy will be applicable to other calibration models to enhance NWP-based ET_o forecasting.

Keywords: Numerical Weather Prediction, Seasonal Patterns, Submonthly Trends, Climatological Mean

1. Introduction

Evapotranspiration (ET) plays an important role in the terrestrial water cycle (Trenberth et al., 2007). Reliable estimation of ET is essential for modeling hydrological processes, including runoff (Amatya and Skaggs, 2011; Yang et al., 2015), soil moisture dynamics (Jung et al., 2010; Sheffield and Wood, 2008), and groundwater recharge (Condon et al., 2020). To quantify the evaporative demand of the atmosphere, the concept of potential evapotranspiration was developed in the 1940s (Thornthwaite, 1948). Originally, this term referred to the maximum water transfer from land surface to the air, but did not specifically define land surface conditions. During the 1950s -1960s, more specific information about crop types, growing stages, and water availability was added to the original term to refine the definition (Jensen, 1968). In the 1970s, the concept of reference crop evapotranspiration (ET_o) was developed to estimate evaporative water loss from a specific vegetation surface (green grass of 8 to 15 cm tall with sufficient water supply) to support irrigation water management (Doorenbos and Pruitt, 1977). In 1998, the Food and Agriculture Organization (FAO) clearly defined the reference vegetation surface as a "hypothetic crop with an assumed crop height (12 cm) and a fixed surface resistance (70 s/m) and albedo (0.23)" (Allen et al., 1998). Based on this reference vegetation, ET from other crops could be estimated with ET_o and a crop coefficient (*Kc*) (Le Page et al., 2021).

Short-term ET_o forecasting provide valuable information for the planning of water management and farming activities (Wu and Chen, 2013; Xue et al., 2016), particularly in arid and semi-arid regions. Traditionally, ET_o forecasting is mainly based on statistical modeling informed by historical observations. Since ET_o demonstrates a strong annual cycle and autoregressive features, regression models, such as the Wavelet-gaussian process regression and seasonal autoregressive models, have been adopted to predict future ET_o with forecast horizons

ranging from 1 day to 24 months (Ashrafzadeh et al., 2020; Karbasi, 2018). In addition, machine learning techniques, such as Artificial Neural Network and Support Vector Machines, have also been used to predict future ET_o at the daily and monthly scales (Salam and Islam, 2020; Sattari et al., 2021). In addition to these statistical predictions, weather and climate forecasts from Numerical Weather Prediction (NWP) models and General Circulation Models (GCMs), have been increasingly used to produce ET_o forecasts (Fan et al., 2021; Liu et al., 2020). Compared with forecasting based on historical observations (Chauhan and Shrivastava, 2009; Novoa and Tejeda, 2006; Torres et al., 2011), NWP-based ET_o forecasting is less limited by data availability (Cai et al., 2007; Paredes et al., 2018), and thus could be conveniently performed across the daily and weekly scales (Er-Raki et al., 2010; Perera et al., 2014; Tian et al., 2014). However, NWP-based raw ET_o forecasts often inherent significant errors from NWP models and thus could demonstrate systematic inconsistencies with observations (Pelosi et al., 2016; Perera et al., 2016).

Calibration models initially developed for other weather variables (e.g., precipitation or temperature) have been adopted to calibrate raw NWP-based ET_o forecasts (Medina and Tian, 2020; Zhao et al., 2019). Methods based on regression calibration (Medina et al., 2018), affine kernel dressing (Medina and Tian, 2020), model output statistics (Silva et al., 2010), and the Bayesian modeling approach (Zhao et al., 2019) have been proven to be effective in correcting bias and improving skills of ET_o forecasts. However, significant improvements are often limited to short lead times, and skills of calibrated forecasts typically become similar to or even worse than randomly sampled historical observations at long lead times. More sophisticated calibrations are needed to further improve the quality of NWP-based ET_o forecasts.

The Seasonally Coherent Calibration (SCC) model has been developed to produce calibrated precipitation forecasts with seasonal patterns coherent with the observed climatology (Wang et al., 2019; Yang et al., 2021a). The model allows for effective reconstruction of precipitation seasonality in calibrated forecasts, and thus represents a significant advancement in NWP forecast calibration. The SCC model's capability in providing accurate, reliable, and skillful calibrated forecasts has been validated through precipitation forecast calibrations across the continental and site scales (Wang et al., 2019; Yang et al., 2021a; Zhao et al., 2021).

The SCC model's capability should be further enhanced when calibrating ET_o forecasts. This model uses a monthly parameterization, assuming that daily forecasts of the same month have the same mean and standard deviation. This parameterization strategy has been proven to be effective in the calibration of precipitation forecasts (Wang et al., 2019). However, ET_o demonstrates a strong annual cycle and shows increasing trends at the submonthly scale in Spring, and decreasing trends in Autumn (Liu et al., 2017). As a result, monthly parameterization in the original SCC model will not capture the submonthly patterns and thus may not be able to fully utilize the predictability in ET_o forecasts. Instead, it may introduce uncertainties to calibrated ET_o forecasts. For example, for months showing increasing trends at the submonthly scale, the monthly parameterization could lead to overestimation in the first halves of these months, but result in underestimations in the second halves. Similarly, for months with decreasing trends, monthly parameterization would result in negative and positive biases at the beginning and the end of these months, respectively.

To account for the submonthly trends in ET_o, forecast calibration should characterize the strong annual cycle of ET_o at a fine temporal scale. Refining SCC's parameterization from the monthly scale to a submonthly scale (e.g., daily) could be a possible solution, but not feasible,

considering the substantial increases in computation cost and the lack of sufficient data for parameter inference. A valid alternative will be separating the ET_o anomalies in both observations and forecasts from the climatological mean of long-term observations first, and then applying the SCC model to calibrate the derived anomalies (Dabernig et al., 2017). This strategy will allow for building the ET_o seasonal patterns, including the submonthly trends, represented by the climatological mean in the calibrated forecasts while retaining the parsimonious monthly parameterization.

In this study, we expand the SCC model's capability in building forecast climatology coherent with observations to the submontly scale by conducting the calibration based on ET_o anomalies and climatological mean. We hypothesize that this new strategy will further improve forecast quality than applying the SCC model directly to ET_o forecasts. To investigate the effectiveness and robustness of this new calibration strategy, we calibrate ET_o forecasts constructed with weather forecasts from the Australian Bureau of Meteorology's Australian Community Climate and Earth-System Simulator G2 version (ACCESS-G2) model at the continental scale and across 21 weather stations. Results from calibrations based on different observations (gridded observations vs. weather station observations) across the two scales will help test the performance of the strategy rigorously, and validate the feasibility for general application. As a result, this current study primarily focuses on evaluating whether calibrating ET_o anomalies will improve the quality of calibrated forecasts, regardless of the sources of observations used for model fitting. Impacts of data interpolations on gridded observations will be briefly discussed in section 4.4 to help readers understand differences in calibrations across the two spatial scales.

Objectives of this study include i) developing a new calibration strategy to better calibrate NWP-based ET_o forecasts, ii) evaluating the effectiveness and robustness of the new strategy.

2. Method

To thoroughly evaluate the effectiveness and robustness of the strategy, we conducted ET_o forecast calibration at both the continental and weather station scales. We use interpolated gridded datasets as observations for forecast calibration across Australia because gridded data is often the only available observations at large spatial scales (Khajehei et al., 2018; Lucatero et al., 2018; Medina et al., 2018). At the continental scale, our calibration only corrects errors resulting from NWP forecasts, and errors caused by data interpolations will not be corrected. The continental-scale investigation covers a broader range of climate zones and larger areas of ungauged regions, and thus could benefit more forecast users than the calibration at the weather station scale. As a result, we primarily introduce results from the calibration across Australia, and only briefly introduce results of the weather station scale calibration in the main text. Figures summarizing the calibration across 21 weather stations are presented in the Supplementary Material to support findings at the continental scale.

2.1 Data

For the calibration at the continental scale, we used weather variables, including temperature, solar radiation, and vapor pressure, provided by the Australian Water Availability Project (AWAP) (Jones et al., 2014, 2007) to derive a gridded ET_o dataset (treated as observations) for ET_o forecast calibration. The AWAP data has been developed by combining the weather station observations, satellite data, and a water-carbon model (Jone et al., 2009), and has been widely used in hydroclimate investigations in Australia. Gridded wind speed data, developed based on interpolation of site-scale observations (Mcvicar et al., 2008), was also used in the development

158 of the gridded ET_o across Australia. The AWAP data and the gridded wind speed data are daily,
159 with a spatial resolution of 0.5° X 0.5° (~ 5 X 5 km). Based on these datasets, we produced
160 gridded continental-scale ET_o observations for each day during 4/1999-3/2019. For the
161 calibration at the weather station scale, we selected 21 stations
162 (http://www.bom.gov.au/climate/data/stations/) located in different climate zones of Australia
163 (Table S1 and Fig. S1), and compiled the observed temperature, solar radiation, and vapor
164 pressure during 1999-2019 to calculate ET_o observations.

165 Weather forecasts from the Australian Bureau of Meteorology's Australian Community
166 Climate and Earth-System Simulator G2 version (ACCESS-G2) model were compiled to
167 produce raw ET_o forecasts (http://www.bom.gov.au/australia/charts/about/about_access.shtml).
168 The ACCESS-G2 model issues weather forecasts four times every day, at the Coordinated
169 Universal Time (UTC) of 0000 UTC, 0600 UTC, 1200 UTC, and 1800 UTC. The ACCESS-G2
170 forecasts have a spatial resolution of 0.38°(longitude) by 0.25°(latitude). We compiled the
171 model's 3-hourly predictions of temperature, solar radiation, wind speed, and vapor pressure for
172 the future 240 hours to produce raw ET_o forecasts.

173 We modified the grid spacing of the raw ET_o forecasts with a spline interpolation method to
174 match the spatial resolution (~ 5 X 5 km) of the AWAP data (Spath, 1993). We also aggregated
175 the 3-hourly ACCESS-G2 forecasts to the daily scale and matched the timeframe of AWAP data
176 by considering the differences between UTC time and Australian local time. The resultant raw
177 daily ET_o forecasts have a forecast horizon of 9 days. Forecast calibration was performed for the
178 period of 4/2016-3/2019. For the calibration at the continental scale, we used the gridded ET_o
179 observations during 4/1999-3/2019 for model fitting and parameter inference, and data (raw
180 forecasts and gridded ET_o) during 4/2016-3/2019 for forecast generation and evaluation. The

calibration across selected weather stations used the same strategy in selecting observations and forecasts for model fitting and evaluation, to make calibrations across the two scales comparable.

In summary, the study period refers to the 20 years between 4/1999 and 3/2019, and this is also the period of the long-term observed ET_o used for parameter inference. The evaluation period refers to the 3 years between 4/2016 and 3/2019.

2.2 Calculation of ET_o observations and ET_o forecasts

We calculated gridded ET_o, weather station ET_o, and raw ET_o forecasts using the FAO 56 Penman-Monteith equation (Allen, et al., 1998). Since AWAP and weather station observations have all variables needed for calculating ET_o, we adopted the FAO ET_o equation directly to derive the daily gridded ET_o and daily weather station ET_o for the study period.

To generate raw ET_o forecasts, we first combined forecasts of the zonal velocity (u) component and the meridional (v) component from ACCESS-G2 to obtain the wind speed at 10 m. Then we estimated the wind speed at 2 m using the following equation (Allen, et al., 1998):

$$u_2 = u_z \frac{4.87}{\ln(67.8z - 5.42)} \quad (1)$$

where u_2 is the wind speed at 2 m above the ground surface (m/s); u_z is the wind speed at z m above the ground surface (m/s); and z is the height of measurement above the ground surface (m), which is 10 m in this study.

We calculated vapor pressure forecasts based on ACCESS-G2 air pressure and the specific humidity forecasts using the following equation:

$$e_a = \frac{qP}{(\varepsilon + q(1-\varepsilon))} \quad (2)$$

where e_a is actual vapor pressure (kPa); P is air pressure (kPa); ε is a constant (0.622); and q is specific humidity (kg/kg).

The FAO Penman-Monteith equation (Allen, et al., 1998) was used to calculate ET_o observations and forecasts based on several weather variables:

$$ET_O = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T+273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \quad (3)$$

where ET_O is the reference crop evapotranspiration (mm/day); Δ is the slope of the vapor pressure curve (kPa/°C); R_n is net radiation at the crop surface (MJ/m^2/day); G is soil heat flux density (MJ/m^2/day); γ is the psychrometric constant (kPa/°C); T is average air temperature (°C); u_2 is the wind speed at the height of 2 m (m/s); and e_s and e_a are saturated and actual vapor pressure (kPa), respectively.

2.3 Calibration of ET_o forecasts

For the calibration at the continental scale, raw forecasts were calibrated for each lead time and each grid cell separately. In the calibration based on ET_o anomaly, we first estimated the climatological mean of ET_o at the daily scale based on long-term ET_o observations, using a spectral method which is further detailed in section 2.3.1. Then we calculated the ET_o anomalies by removing the derived daily climatological mean from the raw ET_o forecasts and observations. Calibration using the SCC model was then applied to the anomalies. Once we obtained calibrated anomalies, we added them back to the climatological mean to produce calibrated ET_o forecasts. This calibration strategy allows for the sophisticated calibration of day-to-day ET_o variability represented by the anomalies and meanwhile embedding the ET_o annual cycle represented by climatological mean into calibrated forecasts. In addition to the calibration based on ET_o

anomaly, we also applied the SCC model directly to calibrate raw ET_o forecasts. Comparison between these two sets of calibrations will validate the effectiveness of the new calibration strategy. Similarly, two sets of calibrations were also conducted across the 21 weather stations to evaluate the new strategy.

2.3.1 Calculation of daily climatological mean

We adopted the method developed by Narapusetty et al. (2009) to derive the climatological mean:

$$y_{cm}(t) = a_0 + \sum_{j=1}^{H}[a_j \cos(w_j \mathrm{t}) + b_j \sin(w_j \mathrm{t})] \quad (4)$$

where $y_{cm}(t)$ is the climatological mean at the daily scale; *H* is the number of harmonics. Here we use *H*=4 following Narapusetty et al. (2009); a_0, a_j, and b_j are coefficients, estimated through minimizing the mean squared differences between climatological mean and observations; $w_j = 2\pi j/P$; and *P* is days in one year.

To estimate a_0, a_j, and b_j, we converted the above equation to a matrix function:

$$y_{cm} = Xz \quad (5)$$

where *X* is a matrix of sinusoidal values; and *z* is a vector for a_0, a_j, and b_j. The solution for the z vector is:

$$z = (X^T X)^{-1} X^T \mathrm{y} \quad (6)$$

where *y* is the input data. In this study, the daily climatological mean is derived based on the long-term daily ET_o observations.

2.3.2 Calibrating ET_o anomalies

We subtracted the climatological mean from the raw ET_o forecasts and ET_o observations to obtain anomalies, which were then calibrated with the SCC model. The SCC model has been introduced in detail in our continental- and site-scale calibrations of precipitation forecasts (Wang et al., 2019; Yang et al., 2021a). Key steps of calibration with SCC include i) building up a joint probability model to link raw forecasts and observations, ii) characterizing seasonal patterns of the calibrated variable based on the long-term observations, iii) obtaining key parameters for short-archived raw forecasts through reparameterization, and iv) producing calibrated forecasts.

In the calibration of ET_o forecasts, we first transformed the anomalies of raw forecasts and observations using the Yeo-Johnson transformation method (Yeo and Johnson, 2000):

$$\hat{x} = \begin{cases} (\lambda x + 1)^{\frac{1}{\lambda}} - 1, & (x \geq 0, \lambda \neq 0) \\ exp(x) - 1, & (x \geq 0, \lambda = 0) \\ -(\lambda - 2)x + 1)^{\frac{1}{2-\lambda}} + 1, & (x < 0, \lambda \neq 2) \\ -exp(-x) + 1, & (x < 0, \lambda = 2) \end{cases} \quad (7)$$

where λ is a transformation parameter; x refers to anomalies of daily ET_o observations or raw forecasts (mm/day); and $\hat{x}$ is the transformed x.

We then constructed a bivariate normal (BN) distribution for the transformed anomalies based on the assumption that the transformed anomaly of ET_o forecasts ($f(t)$) and transformed anomaly of observed ET_o ($o(t)$) are drawn from this BN distribution:

$$[f(t), o(t)] \sim \mathrm{BN}(f(t), o(t) | \mu_f(\mathrm{m}(t)), \sigma_f^2(\mathrm{m}(t)), \mu_o(\mathrm{m}(t)), \sigma_o^2(\mathrm{m}(t)), \rho(\mathrm{m}(t))) \quad (8)$$

where $\mathrm{m}(t)$ returns the month k (k=1 to 12) of daily forecasts or observations of day t; $\mu_f(\mathrm{m}(t))$ and $\sigma_f(m(t))$ refer to the marginal distribution's mean and standard deviation of $f(t)$ in month $\mathrm{m}(t)$, respectively; $\mu_o(\mathrm{m}(t))$ and $\sigma_o(m(t))$ are the mean and standard deviation of the marginal distribution of $o(t)$ in month $\mathrm{m}(t)$; and $\rho(\mathrm{m}(t))$ is the correlation between $f(t)$ and $o(t)$ in month $\mathrm{m}(t)$.

We directly estimated $\mu_o(\mathrm{m}(t))$ and $\sigma_o(m(t))$ using the transformed anomalies of the 20-year ET_o observations, based on a maximum likelihood optimization. However, for the relatively short-archived (3-year) raw ACCESS-G2 forecasts, estimating ($\mu_f(\mathrm{m}(t))$) and standard deviation ($\sigma_f(m(t))$) with the available data was subject to significant sampling errors. We addressed this challenge by estimating the parameters using following linear regressions:

$$\mu_f(k) = a + b\mu_o(k) \tag{9}$$

$$\sigma_f(k) = c + d\sigma_o(k) \tag{10}$$

$$\rho(k) = r \tag{11}$$

where *k* is month of the year (k=1 to 12); *a*, *b*, *c*, and *d* are parameters characterizing the linear relationships; ρ (*k*) denotes the correlation coefficient between raw forecasts (or anomalies when the calibration is based on ET_o anomaly) and observations for each month; *r* is the correlation coefficient between raw forecasts and observations in the transformed space. We set parameters *b*, *c*, *d,* and *r* to be larger than zero in parameter inference based on the maximum likelihood method to ensure that calculated μ_f and σ_f values are physically meaningful (Wang et al., 2019).

Once we obtained all the parameters of the BN distribution (equation 8), we constructed a conditional distribution for observations ($o(t)$) when a new raw forecast ($f(t)$) is provided.

Next, we randomly drew 100 samples from the conditional distribution of $o(t)$ to produce an ensemble, which was treated as the calibrated ensemble forecasts for the corresponding raw forecast. After that, we back-transformed the calibrated anomalies to their original space. Finally, the calibrated anomalies were added back to the climatological mean to generate calibrated ensemble ET_o forecasts.

2.4 Evaluation of raw and calibrated ET_o forecasts

To evaluate the performance of the calibration, we adopted a leave-one-month-out cross-validation strategy. Specifically, we kept one of the 36 months during 4/2016-3/2019 and the same month in the 20-year ET_o observations unselected and used the remaining months for parameter inference. We then treated raw forecasts of the unselected month as new raw forecasts, which were calibrated using parameters optimized in the previous step. We repeated this process until each of the 36 months was processed, and calibrated forecasts were generated for the same period.

We examined forecast accuracy, reliability, temporal variability, skills, and errors to evaluate the performance of the calibration strategy. Our evaluation of forecast bias, correlation coefficient, and skills was primarily conducted at the seasonal scale. Note that in Australia, Summer lasts from December to next February; Autumn lasts from March to May; Winter is from June to August, and Spring is from September to November. We evaluated forecast reliability over the 3-year evaluation period to check the overall consistency between statistical distributions of forecasts and observations. The evaluation of bias, correlation, skills, and errors during the 3-year period is presented in the Supplementary Material. Details about the calculation of these metrics are introduced as follows.

2.4.1 Bias

We first evaluated the forecast accuracy by comparing the average ET_o in the raw and calibrated forecasts relative to observations using the following equation:

$$Bias = \frac{\frac{1}{T}\sum_{t=1}^{T}(x(t)-y(t))}{\frac{1}{T}\sum_{t=1}^{T}y(t)} \times 100 \quad (12)$$

where *Bias* refers to the bias in ET_o forecasts (%); T is the total days during the period for evaluation; $x(t)$ is raw forecast or ensemble mean of calibrated forecasts (mm/day); and $y(t)$ is the corresponding ET_o observation of the same period (mm/day).

2.4.2 Reliability

We first evaluated the reliability of the calibrated forecasts using the reliability diagram for the calibration at the continental scale. This diagram measures the consistency between forecast probability and observed frequency. The reliability diagram has been widely used to evaluate the overall reliability of forecasts (Hartmann et al., 2002; Pelosi et al., 2016). To create the diagram, we first calculated the probabilities of calibrated forecasts exceeding ET_o thresholds of 4, 8, and 10 mm/day. Next, the probabilities of calibrated forecasts of all the grid cells across Australia, all the nine lead times, and all days during the study period were pooled together. We further divided the pooled probabilities into ten categories and plotted them against the corresponding frequency of ET_o observations. If the calibrated forecasts are perfectly reliable, their probability against observed frequency would show a straight line along the diagonal. A curve above the diagonal indicates underestimations in calibrated forecasts, and a curve below the diagonal suggests overestimations.

Reliable ensemble spread is a critical feature of high-quality ensemble forecasts. To further evaluate the overall consistencies between forecasts and observations in statistical distributions, we calculated the probability integral transform (PIT) value of calibrated forecasts using the following equation:

$$\pi(t) = F\big(t, x = y(t)\big) \tag{13}$$

where $F(t,x)$ is the cumulative density function of the ensemble forecasts, and $y(t)$ is the observation. For reliable forecasts, $\pi(t)$ follows a uniform distribution. We use the alpha (α) index to summarize the reliability in each grid cell with the following equation to check the forecast reliability across Australia (Renard et al., 2010):

$$\alpha = 1 - \frac{2}{T}\sum_{t=1}^{T}\left|\pi^{*}(t) - \frac{t}{t+1}\right| \tag{14}$$

where π^* (t) is the sorted π(t), t=1, 2, …, T in ascending order, and T is the total number of days during the evaluation period. The α index measures the total deviation of calibrated forecasts from the corresponding uniform quantile. Perfectly reliable forecasts should have an α-index of 1, and forecasts with no reliability would have an α-index of 0.

2.4.3 Temporal variability

We used the Pearson correlation coefficient (r) between daily forecasts and ET_o observations to evaluate whether they are consistent in temporal variability. For the calibrated ensemble forecasts, we used the ensemble mean to calculate r:

$$r = \frac{\sum_{t=1}^{n}(x(t)-\bar{x})(y(t)-\bar{y})}{\sqrt{\sum_{t=1}^{n}(x(t)-\bar{x})^2}\sqrt{\sum_{t=1}^{n}(y(t)-\bar{y})^2}} \tag{15}$$

where $x(t)$ is raw forecast or the ensemble mean of calibrated forecasts of ET_o (mm/day), $\bar{x}$ is the average of $x(t)$ (mm/day); $y(t)$ is the corresponding ET_o observation (mm/day) of the same period; and $\bar{y}$ is the average of $y(t)$ (mm/day).

2.4.4 Skills of the raw and calibrated forecasts

We used the continuous ranked probability score (CRPS) to measure skills of the raw and calibrated forecasts (Grimit et al., 2006). Specifically, CRPS is calculated with the following equation:

$$CRPS(t) = \int \{F(t,x) - H(x - y(t))\}^2 dx \quad (16)$$

$$\overline{CRPS} = \frac{1}{T}\sum_{t=1}^{T} CRPS(t) \quad (17)$$

where $F(t,x)$ is the cumulative density function of an ensemble forecast, and $y(t)$ is the observation at time t; H is the Heaviside step function ($H = 1$ if $x - y(t) \geq 0$ and $H = 0$ otherwise); and the overbar represents averaging across the T days. For deterministic raw forecasts, CRPS is reduced to absolute errors.

We further calculated the CRPS skill score ($CRPS_{SS}$) using the following equation:

$$CRPS_{SS} = \frac{CRPS_{reference} - CRPS_{forecasts}}{CRPS_{reference}} \times 100 \quad (18)$$

where $CRPS_{reference}$ is the CRPS value of climatology forecasts; $CRPS_{forecasts}$ refers to CRPS value of raw or calibrated forecasts. Positive CRPS skill scores indicate better skills than the climatology forecasts, and vice versa. To make the CRPS skill score of calibrated forecasts of the two calibrations (with vs. without anomaly) comparable, we used the climatology forecasts from the calibration applying SCC directly to ET_o forecasts to calculate $CRPS_{SS}$.

2.4.5 Root mean square error

We further evaluated errors in raw and calculated forecasts using the Root Mean Square Error (RMSE). For the calibrated ensemble forecasts, we used the ensemble mean of the 100 ensemble members to calculate RMSE:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(x(t)-y(t))^2}{T}} \quad (19)$$

where *RMSE* refers to the root mean square error ($\mathrm{mm/day}$); T is the total days during the period for evaluation; $x(t)$ is raw forecast or ensemble mean of calibrated forecasts ($\mathrm{mm/day}$); and $y(t)$ is the corresponding ET_o observation ($\mathrm{mm/day}$).

3. Results

3.1 Seasonal averages of long-term ET_o observations

[Fig. 1]

According to the gridded observations, ET_o demonstrates strong seasonality and spatial variability (Fig. 1). Summer has the highest ET_o among all seasons, with ET_o larger than 9 mm/day covering most parts of central and western Australia. Spring has the second-highest ET_o, with ET_o larger than 7 mm/day located in northern and central Australia, but shows lower ET_o (less than 5 mm/day) in Victoria and Tasmania. ET_o decreases to less than 7 mm/day in most parts of Australia in Autumn. In Winter, ET_o further decreases to less than 5 mm/day in central and southern Australia. In Autumn, Spring, and Winter, ET_o generally decreases with increasing latitude. In summer, inland regions of western and central Australia demonstrate higher ET_o than the coastal margins of eastern, southern, and northern Australia.

3.2 Bias in raw and calibrated forecasts

[Fig. 2]

Raw ET_o forecasts demonstrate significant biases in most parts of Australia across all seasons (Fig. 2). In northern and central parts of Australia, raw forecasts generally overpredict ET_o by 10%, or even by 15% in Autumn and Winter in northern Australia. Positive biases of 5-10% also occur in central parts of Australia in Summer, and northern parts of Australia in Spring. Negative biases only occur in southern Australia in Winter. The spatial patterns of biases in the raw ET_o forecasts are consistent across all nine lead times.

Biases in raw ET_o forecasts are substantially reduced through the calibration. For all seasons, calibrated forecasts demonstrate biases ranging from -5% to 5% across Australia. The remaining biases in calibrated forecasts reflect the deviation of the ET_o during the evaluation period from the long-term climatology of ET_o observations. The comparison of bias in raw and calibrated ET_o forecasts during the whole evaluation period further demonstrates effectiveness of the calibration in bias correction (Fig. S2).

3.3 Reliability of calibrated forecasts

[Fig. 3]

The reliability diagram indicates high consistency between forecast probabilities and observed frequencies (Fig. 3). The plots of forecast probabilities against frequencies of observations based on three thresholds (4, 8, and 10 mm/day) are all distributed along the diagonal, suggesting the high reliability of calibrated ET_o forecasts.

[Fig. 4]

The calibrated forecasts demonstrate reliable uncertainty spread across all the nine lead times (Fig. 4), with most parts of Australia showing high α-index values (> 0.95). A critical advantage of the calibration with the SCC model is to convert the deterministic raw forecasts to calibrated ensemble forecasts. In this study, we generated 100 ensemble members for each raw forecast and quantified the uncertainty range of forecasts with the ensemble spread. The high α-index indicates reasonable representations of ET_o uncertainties by calibrated forecasts. The high reliability also confirms that the parameters characterizing statistical features of calibrated forecasts are reasonably inferred.

3.4 Correlation coefficients between forecasts and observations

[Fig. 5]

Calibrated ET_o forecasts and observations demonstrate high correlations across all seasons (Fig. 5). Autumn shows higher *r* than other seasons, with *r* values larger than 0.85 covering most parts of Australia at 1-day lead time. In other seasons, *r* values are generally larger than 0.80 at the 1-day lead time. For all seasons, the *r* values decrease with increasing lead time. At 9-day lead time, *r* values in Summer are lower than 0.4 in most parts of Australia. In contrast, they remain larger than 0.5 in most parts of Australia in the other three seasons at this lead time. Evaluations also demonstrate high correlations between calibrated ET_o forecasts and observations (Fig. S3).

[Fig. 6]

The calibration significantly improves the representation of temporal variability of ET_o by forecasts. Compared with the raw forecasts, calibrated forecasts demonstrate higher correlations with ET_o observations in different seasons, particularly at long lead times (Fig. 6). Specifically, in Autumn, Winter, and Spring, improvements in *r* by over 10% are found in the coastal regions

of northwestern Australia at 1-day lead time. Summer does not show significant improvements in *r* at 1-day lead time. At longer lead times, improvements in *r* cover larger areas for all seasons. At 9-day lead time, increases in *r* by over 20% are found in northern Australia in Autumn and Spring. Most parts of Australia show increases by more than 20% in *r* in Winter. Similar improvements are found in western and southeastern Australia, and coastal margins of Northern Territory in Summer at 9-day lead time. Over the three evaluation years, the calibration results in more significant increases in *r* in northwestern Australia than in other regions. The increases are also larger at longer lead times (Fig. S4).

[Fig. 7]

Comparison with the calibration applying the SCC model directly to ET_o forecasts clearly demonstrates the advantage of calibrating ET_o anomalies (Fig. 7). When the calibration was conducted based on ET_o anomaly and climatological mean, the resultant calibrated forecasts demonstrated higher correlations with ET_o observations, particularly at long lead times, than the calibration working with ET_o forecasts directly. Specifically, at the 1-day lead time, increases in *r* by more than 5% are found in coastal regions of northwestern Australia in Winter and Spring (Fig. 7). At 9-day lead time, increases by more than 5% or even 10% are found in large areas of Australia for all seasons. Similar improvements in *r* with the adoption of the new strategy are found when *r* is calculated for the evaluation period (Fig. S5).

3.5 Forecast skills and RMSE

[Fig. 8]

The calibration based on ET_o anomaly substantially improves forecast skills. Raw ET_o forecasts demonstrate skills worse than the climatology forecasts in most parts of Australia

449 across all seasons, even at short lead times (Fig. 8). Specifically, at 1-day lead time, negative
450 skill scores lower than -40 (%) cover most parts of Australia in Autumn and Winter. In Summer
451 and Spring, negative skills in raw forecasts are mainly distributed in large areas of northern and
452 central Australia. Positive CRPS skill scores in raw forecasts at 1-day lead time are only found in
453 coastal areas of South Australia and Queensland in Spring and Summer. In addition, the skills
454 decrease quickly at longer lead times and become worse than the climatology forecasts in most
455 parts of Australia for all seasons at 5-day and 9-day lead times. Calibrated forecasts markedly
456 outperform the raw forecasts in forecast skills across Australia. At 1-day lead time, CRPS skill
457 scores larger than 30 (%) are found in most parts of Australia for all seasons. Although CRPS
458 skill scores of the calibrated forecasts decrease with lead time, they remain above zero, even at 9-
459 day lead time, indicating better performance than the climatology forecasts (Fig. 8 and S6).

460 [Fig. 9]

461 Comparing the CRPS skill score of calibrated forecasts generated based on the new strategy
462 (climatological mean and anomaly) with those from the direct application of the SCC model to
463 ET_o forecasts clearly indicates the advantage of the new calibration strategy (Fig. 9). With the
464 new strategy, the skills of calibrated forecasts are further improved, particularly at long lead
465 times. Summer and Autumn show more significant improvements than the other two seasons. At
466 1-day lead time, the CRPS skill scores are increased by 2-5 (%) in most parts of Australia in
467 Summer and Autumn. At 9-day lead time, improvements increase to 10 (%) or even 15 (%) in
468 eastern Australia in Summer and most parts of the continent in Autumn. Improvements in
469 forecast skills in Winter are not significant at short lead times (1- and 5-day lead times), but
470 reach 5 (%) at 9-day lead time in coastal areas of Queensland and Northern Territory, and
471 northern regions of Western Australia. Improvements in Spring are more significant in northern

regions of Western Australia and Northern Territory, and eastern parts of Australia than in other regions. At 9-day lead time, CRPS skill scores are increased by 5 (%) in these regions. Improvements in forecast skills across the 3-year evaluation period further confirm the advantage of the new calibration strategy. Large areas of eastern Australia show increases in forecast skills by more than 7 (%) at long lead times when the calibration is based on ET_o anomalies (Fig. S7).

Improvements in RMSE of forecasts through the calibrations are consistent with improvements in the CRPS skill score. Calibration with the SCC model and the developed strategy substantially reduces RMSE in ET_o forecasts (Fig. S8). The RMSE in raw forecasts is above 1 mm/day across Australia at 1-day lead time, with high RMSE up to 2 mm/day distributed in central and northern regions of Australia. The RMSE of raw forecasts increases quickly with lead time and reaches 2 mm/day in most parts of the country at 9-day lead time. Calibrated forecasts demonstrate much lower RMSE than raw forecasts. At the 1-day lead time, RMSE in calibrated forecasts is lower than 0.8 mm/day in most parts of Australia, except for inland regions of the continent, where RMSE is slightly above 1 mm/day. At 9-day lead time, RMSE in calibrated forecasts is lower than 1.6 mm/day in most parts of the country. More importantly, calibrating ET_o anomalies leads to lower RMSE in calibrated forecasts than those from the calibration working with ET_o forecasts directly, particularly at long lead times (Fig. S9). For example, at the 9-day lead time, RMSE is reduced by more than 4% in western and southeastern Australia, when the calibration is based on ET_o anomalies.

3.6 Results of the calibration across 21 weather stations

Findings in calibrations based on ground observations at 21 weather stations agree well with the continental-scale calibrations, in the improvements in correlation coefficients, CRPS skill score, and RMSE following the adoption of the developed strategy. Specifically, calibrating ET_o

anomalies increases *r* at 18 of the 21 stations at 5-day lead time, and increases *r* across all stations at 9-day lead time, with improvements reaching up to 3.5% (Fig. S10). Improvements in CRPS skill score further verify the effectiveness of the calibration strategy (Fig. S11). Increases in CRPS skill score by 0.5-5 (%) are achieved with the adoption of the new strategy, relative to the calibration calibrating ET_o forecasts directly. Specifically, higher skill scores are found across 19 of the 21 stations at 5-day lead time, and across all 21 stations at 9-day lead time. Calibrating ET_o anomalies further reduces the RMSE of calibrated forecasts, particularly at long lead times (Fig. S12). Compared with calibrated forecasts at the 1-day and 5-day lead times, which demonstrate both increases and decreases in RMSE, forecasts at the 9-day lead time show that RMSE is reduced at all sites when adopting the new strategy. Reductions in RMSE range from ca. 0 to 3% and reach up to ca. 4% at station 19.

As a result, the calibration based on weather station observations agrees well with the calibration across Australia, regarding the effectiveness of the developed strategy. Improvements measured by the three evaluation metrics (*r*, CRPS skill score, and RMSE) are comparable at the two spatial scales, which both demonstrate marked improvements at long lead times.

4. Discussion

4.1 Effectiveness and robustness of the developed strategy

The agreement between calibrations at the continental scale and across 21 weather stations in the improvements of forecast quality successfully validates the effectiveness and robustness of the developed calibration strategy. Due to data interpolation in the production of gridded temperature, vapor pressure, and wind data, interpolation errors would have been introduced to gridded ET_o observations, resulting in differences from the weather station ET_o (Pelosi et al., 2020a). Different observations used in the calibrations at the two spatial scales must have led to

differences in the climatological means (Equations 4-6), transformation parameters (Equation 7), and key SCC parameters (Equations 8-11). However, despite these differences, the calibrations achieve similar improvements in correlation coefficient, forecast skills, and RMSE, with the adoption of the new calibration strategy. In this study, calibrations at the continental and weather station scales both show improvements in forecast skills when ET_o anomalies rather than the original ET_o forecasts are calibrated. Examination of the strategy across the two scales confirms its strength in enhancing NWP-based ET_o forecasting, and also verifies its robustness for general application.

The effectiveness of the calibration strategy is further confirmed by the comparison with existing ET_o forecasting investigations. Raw ET_o forecasts constructed with NWP weather forecasts often become less skillful than climatology forecasts beyond the first a few days (Perera et al., 2014). Through calibrating ET_o anomalies, we substantially improve the forecast skills and produce more skillful calibrated forecasts than climatology forecasts for all nine lead times. In addition, calibrated forecasts in this study demonstrate comparable skills with those produced using the Bayesian Joint Probability (BJP) model across three weather stations in Australia (Zhao et al., 2019). Specifically, the CRPS skill scores of calibrated forecasts from this study at these stations (39.5 to 45.2 (%)) are close to the skills of ~30-40 (%) in Zhao et al. (2019) at 1-day lead time, although the two studies use different calibration models and NWP forecasts. Thanks to the capability of SCC in calibrating short-archived forecasts, our calibration uses a much shorter period of archived forecasts (three years) for parameter inference than the 23-year forecasts used in Zhao et al. (2019). Compared with ET_o calibrations based on Nonhomogeneous Gaussian Regression (NGR), Affine Kernel Dressing (AKD), and Bayesian Model Averaging (BMA) in the U.S. (with bias of 0.49-1.69%), our calibration results in similar

low bias (0.32-0.95%) in calibrated forecasts at the 1-day lead time (Medina and Tian, 2020). Unlike results from these models, bias in calibrated forecasts of our study does not increase with lead time (Medina and Tian, 2020), indicating the better performance of SCC in correcting bias at long lead times.

4.2 Implications for improving NWP-based ET_o forecasting

This investigation confirms the necessity of improving existing calibration models, including the SCC model, when transferred to calibrate ET_o forecasts. Currently, many calibration models applied to ET_o forecasting were not developed specifically for this variable (Medina and Tian, 2020; Zhao et al., 2019). Most of these models were designed to address common challenges in forecast calibration, and were intended to be variable-independent. However, applying these models, including the SCC model, directly to weather variables with divergent temporal patterns and statistical features, may not fully implement the predictability of the target variables. For example, ET_o demonstrates a much stronger annual cycle and submonthly trends than precipitation (Yang et al., 2021b). Calibrating models working well in precipitation forecast calibration may not necessarily be able to achieve the best performance, if differences between the two weather variables are not considered and accounted for in forecast calibration. Additional skills gained through calibrating ET_o anomalies suggest the developed strategy could be a promising solution for extending the capability of existing calibration models in ET_o forecast calibration.

In this study, we introduce equations and parameters for deriving the climatological mean in detail (Equations 4-6 and associated introductions). Based on such information, future ET_o forecast calibration just needs to add a few more steps to their existing models, to implement the

563 calibration strategy developed in this study. We anticipate that this strategy could be easily
564 applied to future investigations to produce more skillful ET_o forecasts.

565 **4.3 Importance of reconstructing seasonality in weather forecast calibration**

566 Improved forecast skills with the adoption of the developed calibration strategy confirm the
567 necessity of enhancing the representation of ET_o temporal patterns in forecast calibration. It has
568 been challenging to generate skillful calibrated ET_o forecasts at long lead times (Medina et al.,
569 2018; Zhao et al., 2019). Skillful calibrated forecasts better than randomly sampled climatology
570 are often limited to short lead times, usually less than one week. By separating anomaly from the
571 climatological mean, the new strategy allows for calibrating short-term ET_o variability with the
572 sophisticated SCC algorithms and embedding the observed annual cycle into calibrated forecasts
573 simultaneously. Based on this strategy, we produced calibrated forecasts better than the
574 climatology forecasts for all nine lead times, and outperformed the calibration working directly
575 with raw ET_o forecasts, particularly at long lead times. Improvements in forecast skills following
576 embedding the climatological mean in calibrated forecasts clearly show the importance of
577 enhancing the representation of ET_o seasonality for producing skillful calibrated forecasts.

578 This investigation also provides valuable implications for addressing a well-known challenge
579 in NWP-based weather forecasting. Modern NWP models often demonstrate decent performance
580 in forecasting short-term dynamics of the weather system (Bauer et al., 2015). However,
581 processes regulating the annual cycle of weather variables are typically not well resolved in
582 NWP models (Robertson et al., 2014). Lower correlations between raw ET_o forecasts and
583 observations than those of calibrated forecasts confirm limitations of NWP models in modeling
584 slow varying atmospheric processes (Pelosi et al., 2016). Improvements in temporal patterns of
585 calibrated forecasts suggest that building seasonality through statistical calibration could be a

cost-effective solution for dealing with the intrinsic limitations in NWP forecasts. We anticipate that the developed strategy could be transferable for calibrating weather variables demonstrating a strong annual cycle (Cai et al., 2007; Mohan and Arumugam, 1995).

4.4 Future work

Although our calibrations across the two spatial scales confirm the effectiveness and robustness of the developed strategy, additional investigations are needed to further improve ET_o forecast calibration. Like in many regional-scale calibrations (Khajehei et al., 2018; Lucatero et al., 2018; Medina et al., 2018), our calibration adopts gridded observations as the reference to calibrate raw forecasts at the continental scale. Although we achieve significant improvements in forecast quality, it should be noted that the interpolation errors embedded in the gridded observations have been inherited by calibrated forecasts. This type of error could be large in regions where field observations are scarce, and only limited ground observations have been used to produce the gridded data. In these regions, such as the central parts of Australia, interpolation errors could be significant (Pelosi et al., 2020a; Yang et al., 2021a). As a result, calibration based on low-quality observations would have inevitably introduced uncertainties to calibrated forecasts. Existing calibration models often lack the capability of correcting errors resulting from data interpolation. Innovative calibration strategies, such as utilizing remotely sensed data (Pelosi et al., 2020b) or reanalysis data (Su et al., 2021), should be developed to address this critical challenge in future forecast calibrations at large spatial scales.

5. Conclusions

To improve the calibration of NWP-based ET_o forecasts, we enhance the SCC model's capability in building seasonality coherent with observed climatology in calibrated forecasts by calibrating ET_o anomalies. The new calibration strategy allows for the characterization of ET_o

609 temporal patterns at a finer temporal resolution (submonthly scale vs. monthly scale), and thus
610 provides better representations of the strong annual cycle in ET_o, than calibrating ET_o forecasts
611 directly. Evaluation of this new strategy across two spatial scales confirms its effectiveness and
612 robustness in improving ET_o forecast calibration, particularly at long lead times. This
613 investigation highlights the importance of reconstructing seasonality in the calibration of NWP-
614 based ET_o forecasts, and also provides an effective solution.

615 Calibrated ET_o forecasts generated in this study demonstrate skills better than climatology
616 forecasts for all nine lead times. Extension of skillful forecasts to long lead times will enable
617 forecast users to increase preparedness to better manage the hydroclimatic variability. The
618 developed calibration strategy is expected to be applicable to other calibration models to enhance
619 NWP-based ET_o forecasting. We anticipate that this strategy could also be used to calibrate other
620 weather variables showing strong seasonality.

Acknowledgments:

This study has been supported by a collaborative research project (TP707466) between the University of Melbourne and the Australian Bureau of Meteorology and an ARC Linkage Project (LP170100922). Computations of this research were undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government. Computation hours were provided by the National Computational Infrastructure (NCI) LIEF Grant (LE190100021) and facilitated by The University of Melbourne. The authors declare that there is no conflict of interest regarding the publication of this article.

References:

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. FAO Irrigation and drainage paper No.56, Crop evapotranspiration: guidelines for computing crop water requirements.

Amatya, D.M., Skaggs, R.W., 2011. Long-term hydrology and water quality of a drained pine plantation in North Carolina. Am. Soc. Agric. Biol. Eng. 54, 2087–2098.

Ashrafzadeh, A., Ki, O., Aghelpour, P., 2020. Comparative Study of Time Series Models, Support Vector Machines, and GMDH in Forecasting Long-Term Evapotranspiration Rates in Northern Iran. J. Irrig. Drain. Eng. 146, 1–10. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001471

Bauer, P., Thorpe, A., Brunet, G., 2015. The quiet revolution of numerical weather prediction. Nature 525. https://doi.org/10.1038/nature14956

Cai, J., Liu, Y., Lei, T., Pereira, S.L., 2007. Estimating reference evapotranspiration with the FAO Penman – Monteith equation using daily weather forecast messages. Agric. For. Meteorol. 145, 22–35. https://doi.org/10.1016/j.agrformet.2007.04.012

Chauhan, S., Shrivastava, R.K., 2009. Reference evapotranspiration forecasting using different artificial neural networks algorithms. Can. J. Civ. Eng. 36, 1491–1505. https://doi.org/10.1139/L09-074

Condon, L.E., Atchley, A.L., Maxwell, R.M., 2020. Evapotranspiration depletes groundwater under warming over the contiguous United States. Nat. Commun. 11, 1–8. https://doi.org/10.1038/s41467-020-14688-0

Dabernig, M., Mayr, G.J., Messner, J.W., Zeileis, A., 2017. Spatial ensemble post-processing with standardized anomalies. Q. J. R. Meteorol. Soc. 143, 909–916. https://doi.org/10.1002/qj.2975

Doorenbos, J., Pruitt, W., 1977. Guidelines for predicting crop water requirements, Irrigation Drain. Paper No.24. FAO, Tome, Ttaly.

Er-Raki, S., Chehbouni, A., Khabba, S., Simonneaux, V., Jarlan, L., Ouldbba, A., Rodriguez, J.C., Allen, R., 2010. Assessment of reference evapotranspiration methods in semi-arid regions: Can weather forecast data be used as alternate of ground meteorological parameters? J. Arid Environ. 74, 1587–1596. https://doi.org/10.1016/j.jaridenv.2010.07.002

Fan, J., Wu, L., Zheng, J., Zhang, F., 2021. Medium-range forecasting of daily reference evapotranspiration across China using numerical weather prediction outputs downscaled by extreme gradient boosting. J. Hydrol. 601.126664. https://doi.org/10.1016/j.jhydrol.2021.126664

Grimit, E.P., Gneiting, T., Berrocal, V.J., Johnson, N.A., 2006. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Q. J. R. Meteorol. Soc. 132, 2925–2942. https://doi.org/10.1256/qj.05.235

Hartmann, H., Pagano, T.C., Sorooshian, S., Bales, R., 2002. Evaluating Seasonal Climate Forecasts from User Perspectives. Bull. Am. Meteorol. Soc. 83, 683–698.

Jensen, M.E., 1968. Water Consumpition by Agricultural Plnat in Kozlowski, T.T., Water Deficits and Plant Growt. Academic Press, New York&London.

Jones, D.A., Wang, W., Fawcett, R., 2014. Australian Water Availability Project Daily Gridded Rainfall [WWW Document]. URL http://www.bom.gov.au/jsp/awap/rain/index.jsp

683 Jones, D.A., Wang, W., Fawcett, R., 2007. Climate Data for the Australian Water Availability Project.
684 Australian Bureau of Meteorology, Melbourne, Australia.

685 Jones,D., Wang, W., Fawcett, R., 2009. High-quality spatial climate data-sets for Australia. Aust.
686 Meteorol. Oceanogr. J.. 58(4).233-248. http://www.bom.gov.au/amm/papers.php?year=2009

687 Jung, M., Reichstein, M., Ciais, P., Seneviratne, S.I., Sheffield, J., Goulden, M.L., Bonan, G., Cescatti,
688 A., Chen, J., de Jeu, R., Dolman, a J., Eugster, W., Gerten, D., Gianelle, D., Gobron, N., Heinke, J.,
689 Kimball, J., Law, B.E., Montagnani, L., Mu, Q., Mueller, B., Oleson, K., Papale, D., Richardson,
690 A.D., Roupsard, O., Running, S., Tomelleri, E., Viovy, N., Weber, U., Williams, C., Wood, E.,
691 Zaehle, S., Zhang, K., 2010. Recent decline in the global land evapotranspiration trend due to
692 limited moisture supply. Nature 467, 951–954. https://doi.org/10.1038/nature09396

693 Karbasi, M., 2018. Forecasting of Multi-Step Ahead Reference Evapotranspiration Using Wavelet-
694 Gaussian Process Regression Model. Water Resour. Manag. 32, 1035–1052.

695 Khajehei, S., Ahmadalipour, A., Moradkhani, H., 2018. An effective post-processing of the North
696 American multi-model ensemble (NMME) precipitation forecasts over the continental US. Clim.
697 Dyn. 51, 457–472. https://doi.org/10.1007/s00382-017-3934-0

698 Le Page, M., Fakir, Y., Jarlan, L., Boone, A., Berjamy, B., Khabba, S., Zribi, M., 2021. Projection of
699 irrigation water demand based on the simulation of synthetic crop coefficients and climate change.
700 Hydrol. Earth Syst. Sci. 25, 637–651. https://doi.org/10.5194/hess-25-637-2021

701 Liu, B., Liu, M., Cui, Y., Shao, D., Mao, Z., Zhang, L., Khan, S., Luo,Y., 2020. Assessing forecasting
702 performance of daily reference evapotranspiration using public weather forecast and numerical
703 weather prediction. J. Hydrol., 590. 125547. https://doi.org/10.1016/j.jhydrol.2020.125547

704 Lucatero, D., Madsen, H., Refsgaard, J.C., Kidmose, J., Jensen, K.H., 2018. On the skill of raw and post-
705 processed ensemble seasonal meteorological forecasts in Denmark. Hydrol. Earth Syst. Sci. 22,
706 6591–6609. https://doi.org/10.5194/hess-22-6591-2018

707 Mcvicar, T.R., Niel, T.G. Van, Li, L.T., Roderick, M.L., Rayner, D.P., Ricciardulli, L., Donohue, R.J.,
708 2008. Wind speed climatology and trends for Australia , 1975 – 2006 : Capturing the stilling
709 phenomenon and comparison with near-surface reanalysis output. Geophys. Res. Lett. 35, 1–6.
710 https://doi.org/10.1029/2008GL035627

711 Medina, H., Tian, D., 2020. Comparison of probabilistic post-processing approaches for improving
712 numerical weather prediction-based daily and weekly reference evapotranspiration forecasts.
713 Hydrol. Earth Syst. Sci. 24, 1011–1030.

714 Medina, H., Tian, D., Srivastava, P., Pelosi, A., Chirico, G.B., 2018. Medium-range reference
715 evapotranspiration forecasts for the contiguous United States based on multi-model numerical
716 weather predictions. J. Hydrol. 562, 502–517. https://doi.org/10.1016/j.jhydrol.2018.05.029

717 Mohan, S., Arumugam, N., 1995. Forecasting weekly reference crop evapotranspiration series. Hydrol.
718 Sci. J. 40, 689–702. https://doi.org/10.1080/02626669509491459

719 Narapusetty, B., Delsole, T., Tippett, M.K., 2009. Optimal estimation of the climatological mean. J. Clim.
720 22, 4845–4859. https://doi.org/10.1175/2009JCLI2944.1

721 Novoa, R.S. a, Tejeda, H.R., 2006. Evaluation of the N2O emissions from N in plant residues as affected
722 by environmental and management factors. Nutr. Cycl. Agroecosystems 75, 29–46.

https://doi.org/10.1007/s10705-006-9009-y

Paredes, P., Fontes, J.C., Azevedo, E.B., Pereira, L.S., 2018. Daily reference crop evapotranspiration in the humid environments of Azores islands using reduced data sets : accuracy of FAO-PM temperature and Hargreaves-Samani methods. Theor. Appl. Climatol. 134, 595–611.

Pelosi, A., Medina, H., Villani, P., D'Urso, G., Chirico, G.B., 2016. Probabilistic forecasting of reference evapotranspiration with a limited area ensemble prediction system. Agric. Water Manag. 178, 106–118. https://doi.org/10.1016/j.agwat.2016.09.015

Pelosi, A., Terribile, F., D'Urso, G., Chirico, G.B., 2020a. Comparison of ERA5-Land and UERRA MESCAN-SURFEX reanalysis data with spatially interpolated weather observations for the regional assessment of reference evapotranspiration. Water 12. https://doi.org/10.3390/W12061669

Pelosi, A., Villani, P., Bolognesi, S.F., Chirico, G.B., D'urso, G., 2020b. Predicting crop evapotranspiration by integrating ground and remote sensors with air temperature forecasts. Sensors 20, 1–18. https://doi.org/10.3390/s20061740

Perera, K.C., Western, A.W., Nawarathna, B., George, B., 2014. Forecasting daily reference evapotranspiration for Australia using numerical weather prediction outputs. Agric. For. Meteorol. 194, 50–63. https://doi.org/10.1016/j.agrformet.2014.03.014

Perera, K.C., Western, A.W., Robertson, R.D., George, B., Nawarathna, B., 2016. Ensemble forecasting of short-term system scale irrigation demands using real-time flow data and numerical weather predictions. Water Resour. Res. 52, 4801–4822. https://doi.org/10.1002/2015WR018532.Received

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resour. Res. 46, 1–22. https://doi.org/10.1029/2009WR008328

Robertson, A., Kumar, A., Pena, M., Vitart, F., 2014. Improving and Promoting Subseasonal to Seasonal Prediction, in: International Conference on Subseasonal to Seasonal Prediction. pp. 49–53. https://doi.org/10.1175/BAMS-D-14-00139.1

Salam, R., Islam, A.R.M.T., 2020. Potential of RT, bagging and RS ensemble learning algorithms for reference evapotranspiration prediction using climatic data-limited humid region in Bangladesh. J. Hydrol. 590, 125241. https://doi.org/10.1016/j.jhydrol.2020.125241

Sattari, M.T., Apaydin, H., Band, S.S., Mosavi, A., Prasad, R., 2021. Comparative analysis of kernel-based versus ANN and deep learning methods in monthly reference evapotranspiration estimation. Hydrol. Earth Syst. Sci. 25, 603–618. https://doi.org/10.5194/hess-25-603-2021

Sheffield, J., Wood, E.F., 2008. Global Trends and Variability in Soil Moisture and Drought Characteristics , 1950 – 2000 , from Observation-Driven Simulations of the Terrestrial Hydrologic Cycle. J. Clim. 21, 432–458. https://doi.org/10.1175/2007JCLI1822.1

Silva, D., Meza, F.J., Varas, E., 2010. Estimating reference evapotranspiration (ETo) using numerical weather forecast data in central Chile. J. Hydrol. 382, 64–71. https://doi.org/10.1016/j.jhydrol.2009.12.018

Spath, H., 1993. Two Dimensional Spline Interpolation algorithms. CRC Press, New York, USA.

Su, C.H., Eizenberg, N., Jakob, D., Fox-Hughes, P., Steinle, P., White, C.J., Franklin, C., 2021. BARRA v1.0: Kilometre-scale downscaling of an Australian regional atmospheric reanalysis over four

763 midlatitude domains. Geosci. Model Dev. 14, 4357–4378. https://doi.org/10.5194/gmd-14-4357-
764 2021

765 Thornthwaite, C.W., 1948. An Approach toward a Rational Cliassification of climate. Geogr. Rev. 38,
766 55–94.

767 Tian, D., Martinez, C.J., 2014. The GEFS-Based Daily Reference Evapotranspiration (ETo) Forecast
768 and Its Implication for Water Management in the Southeastern United States. J. Hydrometeorol. 15,
769 1152–1165. https://doi.org/10.1175/JHM-D-13-0119.1

770 Tian, D., Martinez, C.J., Graham, W.D., 2014. Seasonal Prediction of Regional Reference
771 Evapotranspiration Based on Climate Forecast System Version 2. J. Hydrometeorol. 15, 1166–1188.
772 https://doi.org/10.1175/JHM-D-13-087.1

773 Torres, A.F., Walker, W.R., Mckee, M., 2011. Forecasting daily potential evapotranspiration using
774 machine learning and limited climatic data. Agric. Water Manag. 98, 553–562.
775 https://doi.org/10.1016/j.agwat.2010.10.012

776 Trenberth, K.E., Smith, L., Qian, T., Dai, A., Fasullo, J., 2007. Estimates of the Global Water Budget and
777 Its Annual Cycle Using Observational and Model Data. J. Hydrometeorol. 8, 758–769.
778 https://doi.org/10.1175/JHM600.1

779 Wang, Q.J., Zhao, T., Yang, Q., Robertson, D., 2019. A Seasonally Coherent Calibration (SCC) Model
780 for Postprocessing Numerical Weather Predictions. Mon. Weather Rev. 147, 3633–3647.
781 https://doi.org/10.1175/MWR-D-19-0108.1

782 Wu, Y., Chen, J., 2013. Estimating irrigation water demand using an improved method and optimizing
783 reservoir operation for water supply and hydropower generation: A case study of the Xinfengjiang
784 reservoir in southern China. Agric. Water Manag. 116, 110–121.
785 https://doi.org/10.1016/j.agwat.2012.10.016

786 Xue, J., Gui, D., Zhao, Y., Lei, J., Zeng, F., Feng, X., Mao, D., Shareef, M., 2016. A decision-making
787 framework to model environmental flow requirements in oasis areas using Bayesian networks. J.
788 Hydrol. 540, 1209–1222. https://doi.org/10.1016/j.jhydrol.2016.07.017

789 Yang, Q., Tian, H., Friedrichs, M.A.M., Hopkinson, C.S., Lu, C., Najjar, R.G., 2015. Increased nitrogen
790 export from eastern North America to the Atlantic Ocean due to climatic and anthropogenic changes
791 during 1901-2008. J. Geophys. Res. G Biogeosciences 120, 1046–1068.
792 https://doi.org/10.1002/2014JG002763

793 Yang, Q., Wang, Q.J., Hakala, K., 2021a. Achieving effective calibration of precipitation forecasts over a
794 continental scale. J. Hydrol. Reg. Stud. 35, 100818. https://doi.org/10.1016/j.ejrh.2021.100818

795 Yang, Q., Wang, Q.J., Hakala, K., Tang, Y., 2021b. Bias-correcting input variables enhances forecasting
796 of reference crop evapotranspiration. Hydrol. Earth Syst. Sci., 25, 4773–4788.
797 https://doi.org/10.5194/hess-25-4773-2021

798 Yang, Y., Cui, Y., Bai, K., Luo, T., Dai, J., Wang, W., Luo, Y., 2019. Short-term forecasting of daily
799 reference evapotranspiration using the reduced-set Penman-Monteith model and public weather
800 forecasts. Agric. Water Manag. 211, 70–80. https://doi.org/10.1016/j.agwat.2018.09.036

801 Yeo, I., Johnson, R.A., 2000. A new family of power transformations to improve normality or symmetry.
802 Biometrika 87, 954–959.

Zhao, P., Wang, Q.J., Wu, W., Yang, Q., 2021. Which precipitation forecasts to use? Deterministic versus coarser-resolution ensemble NWP models. Q. J. R. Meteorol. Soc. 147, 900–913.

Zhao, T., Wang, Q.J., Schepen, A., 2019. A Bayesian modelling approach to forecasting short-term reference crop evapotranspiration from GCM outputs. Agric. For. Meteorol. 269–270, 88–101. https://doi.org/10.1016/j.agrformet.2019.02.003

Figures

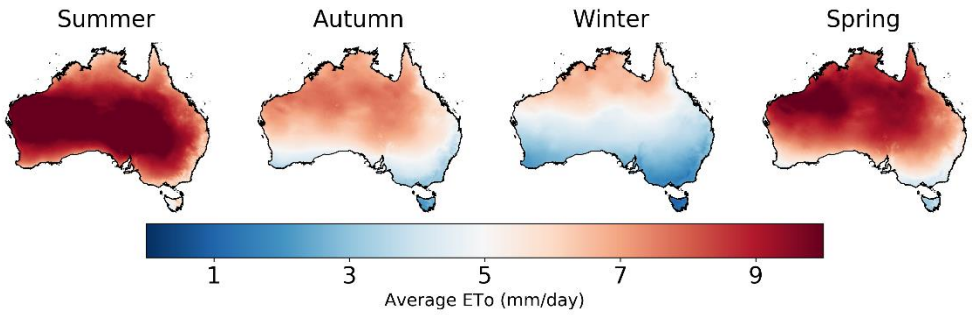


Fig. 1. Seasonal averages of ET_o observations during 4/1999-3/2019 (Summer: December to next February; Autumn: March to May; Winter: June to August; Spring: September to November).

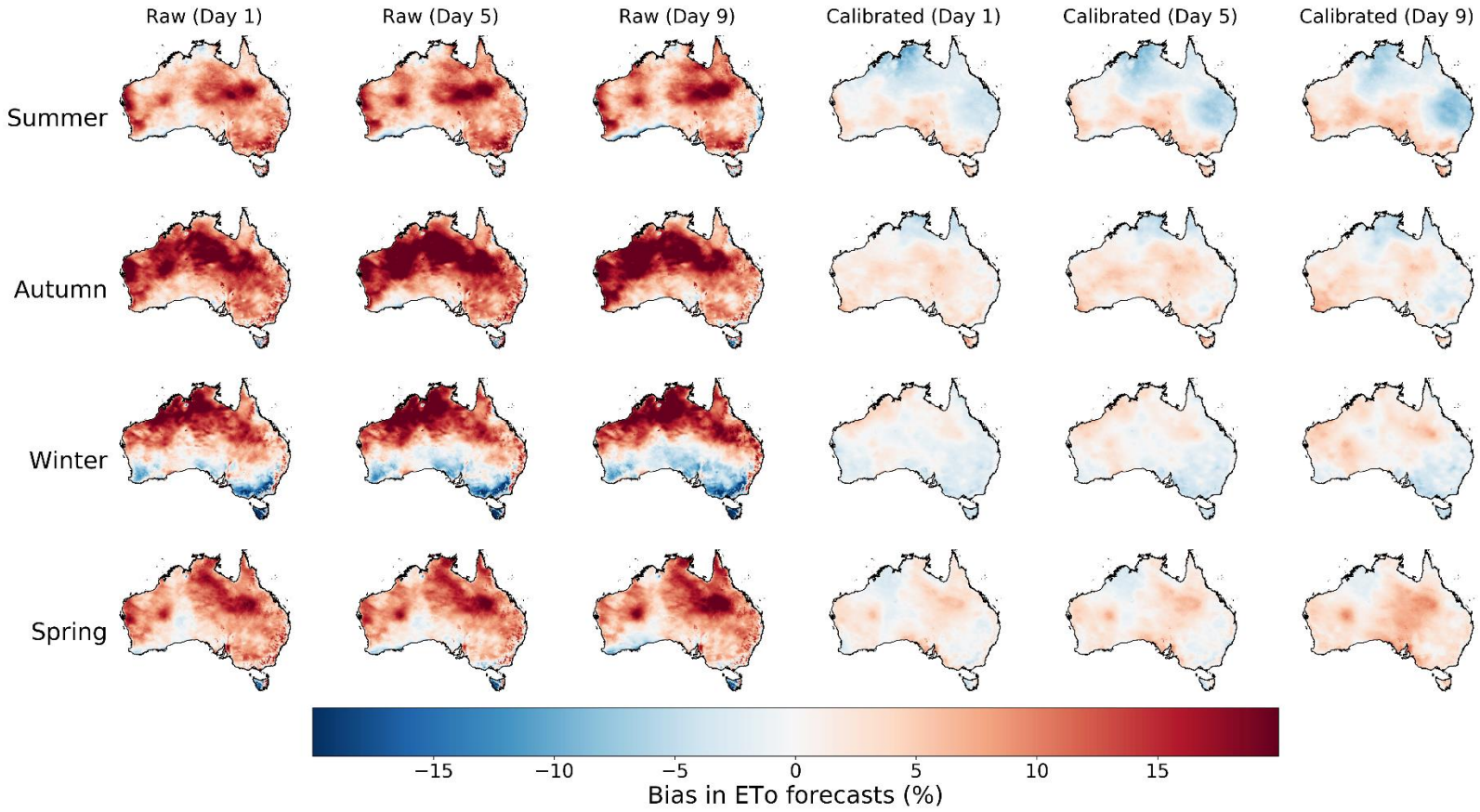


Fig. 2. Bias in raw (three columns on the left) ET_o forecasts and calibrated (three columns on the right) ET_o forecasts (calibration based on ET_o anomaly) in different seasons for 1-day, 5-day, and 9-day lead times.

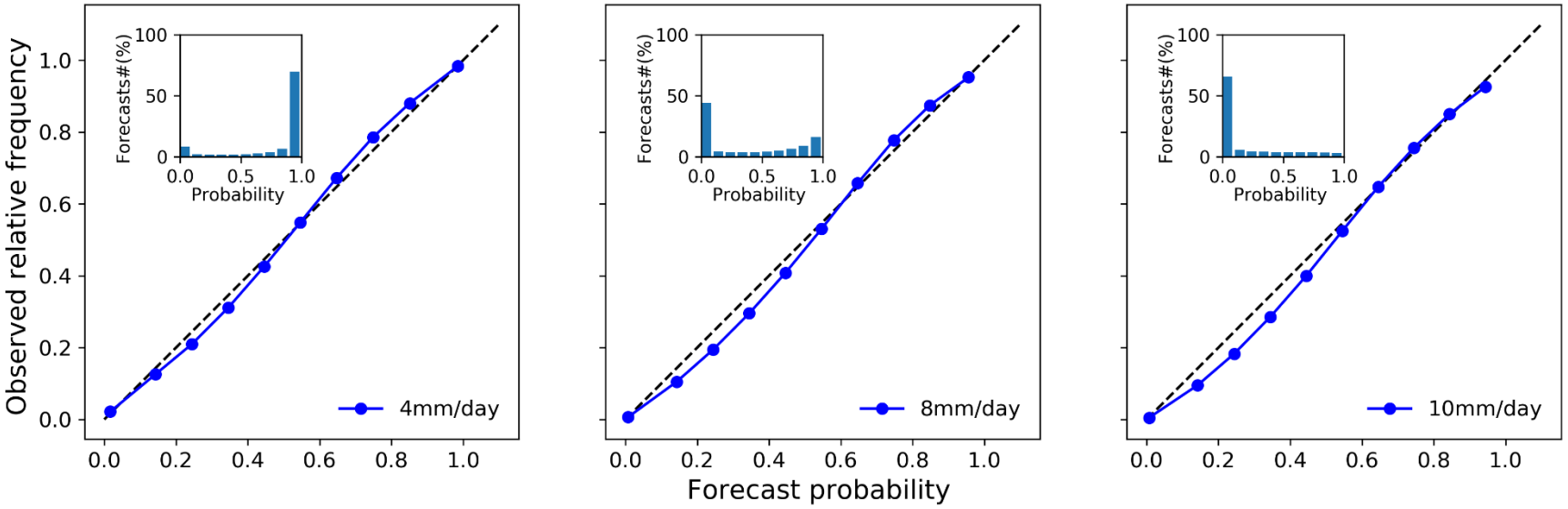
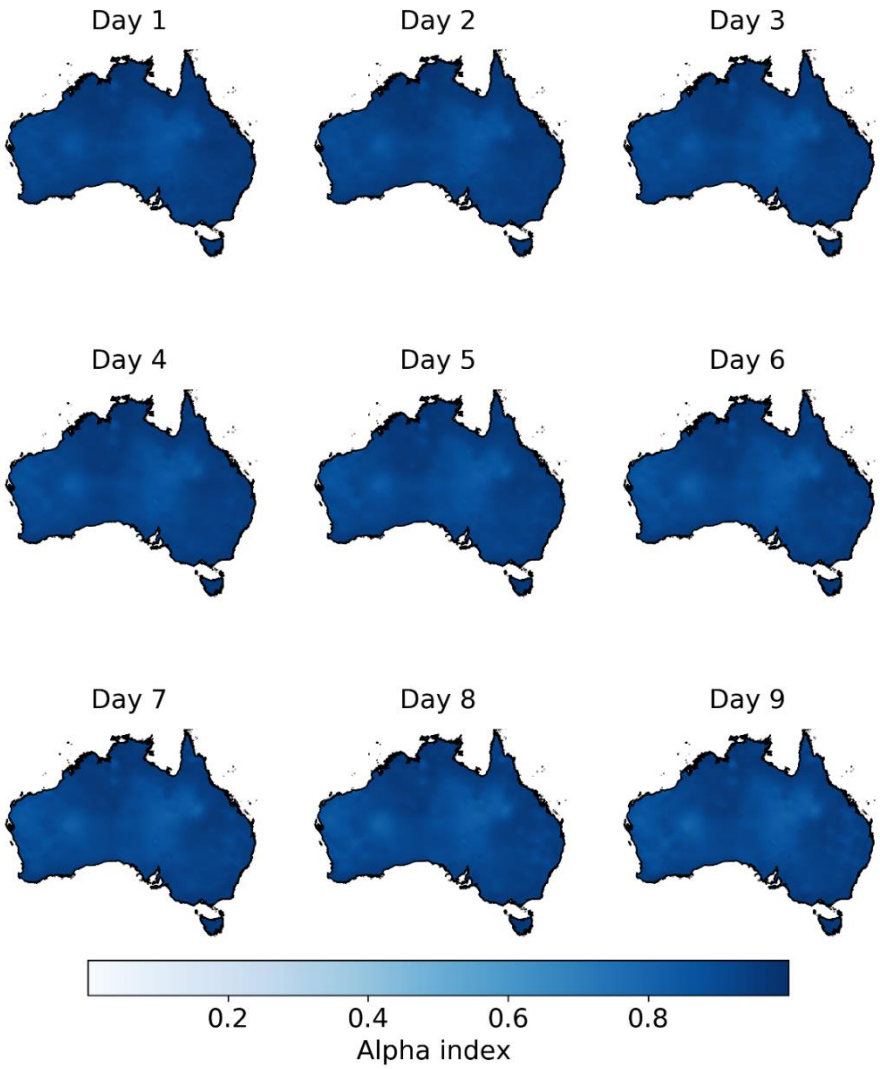


Fig. 3. Reliability diagrams of calibrated ET_o forecasts (calibration based on ET_o anomaly) with thresholds of 4, 8, and 10 mm/day.



829

830 Fig. 4. Alpha index in calibrated ET_o forecasts (calibration based on ET_o anomaly).

831

832

833

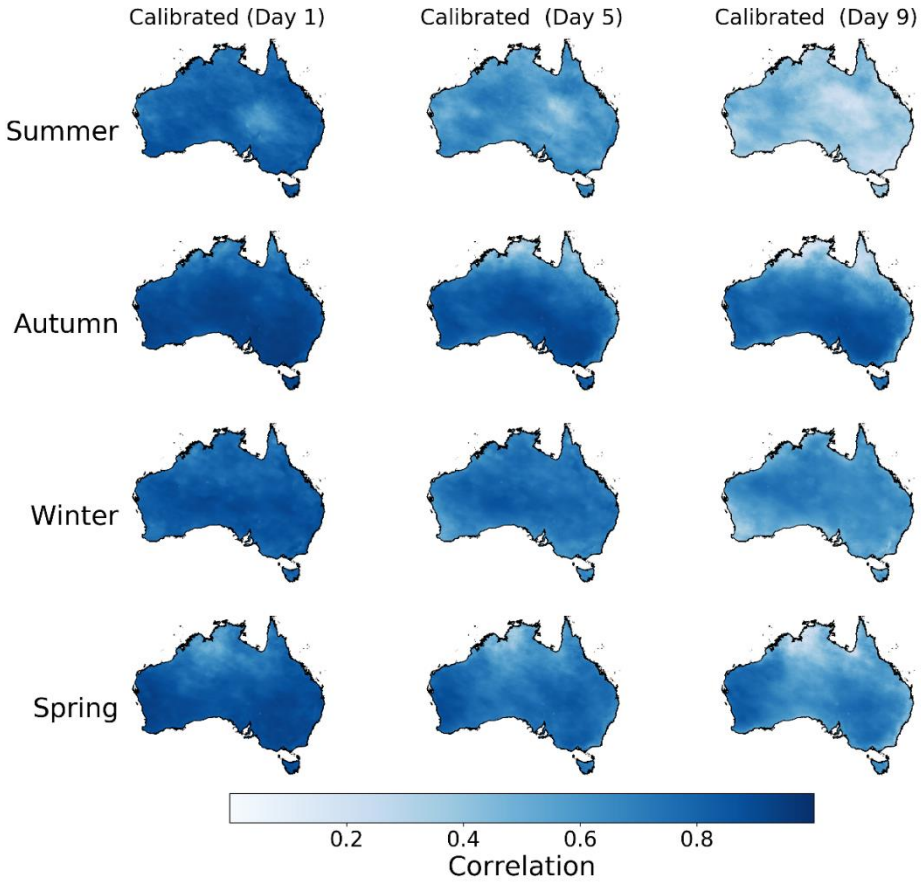
834

835

836

837

838

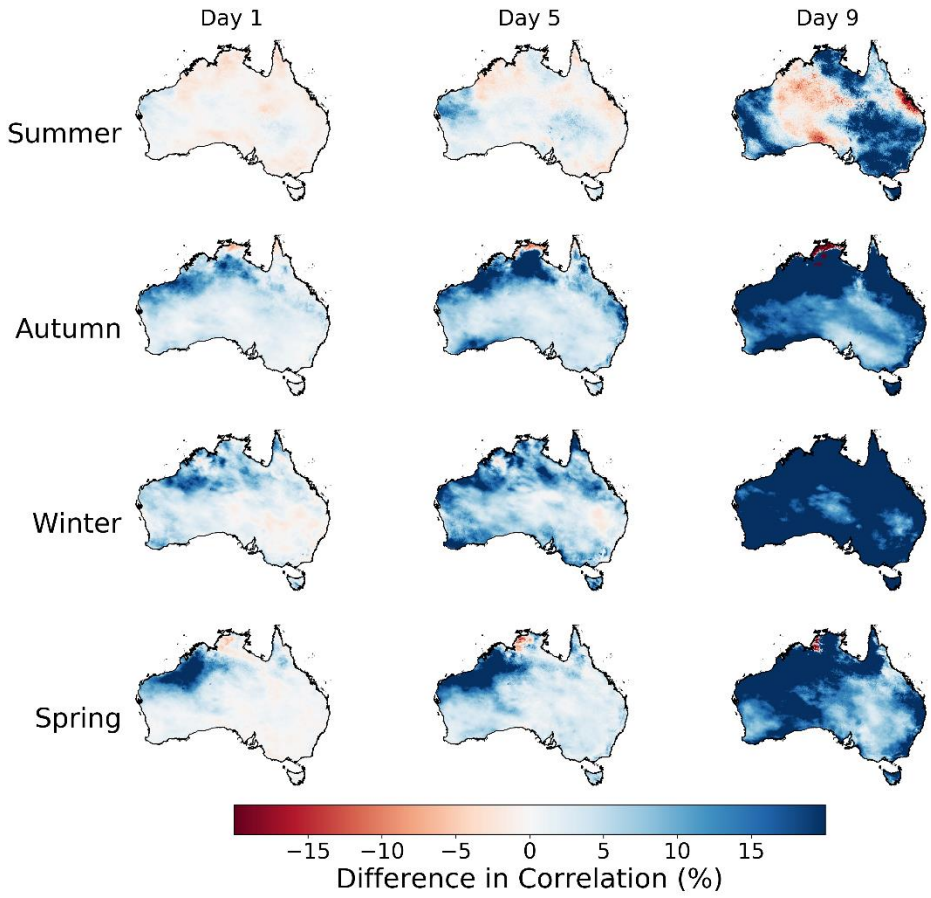


839

840 Fig. 5. Correlation coefficient (r) between calibrated ET_o forecasts (calibration based on ET_o
841 anomaly) and observations in different seasons.

842

843



844

845 Fig. 6. Differences between the correlation coefficient (*r*) of calibrated forecasts (calibration
846 based on ET_o anomaly) and observations, and the *r* of the raw forecasts and ET_o observations in
847 different seasons.

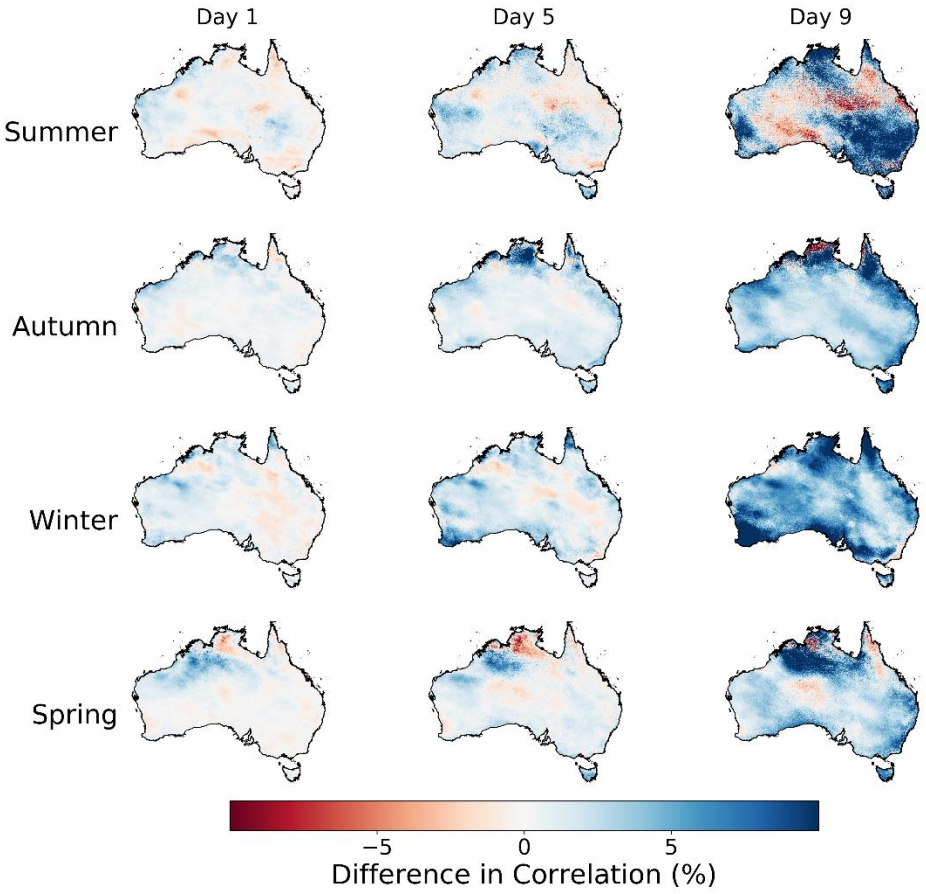
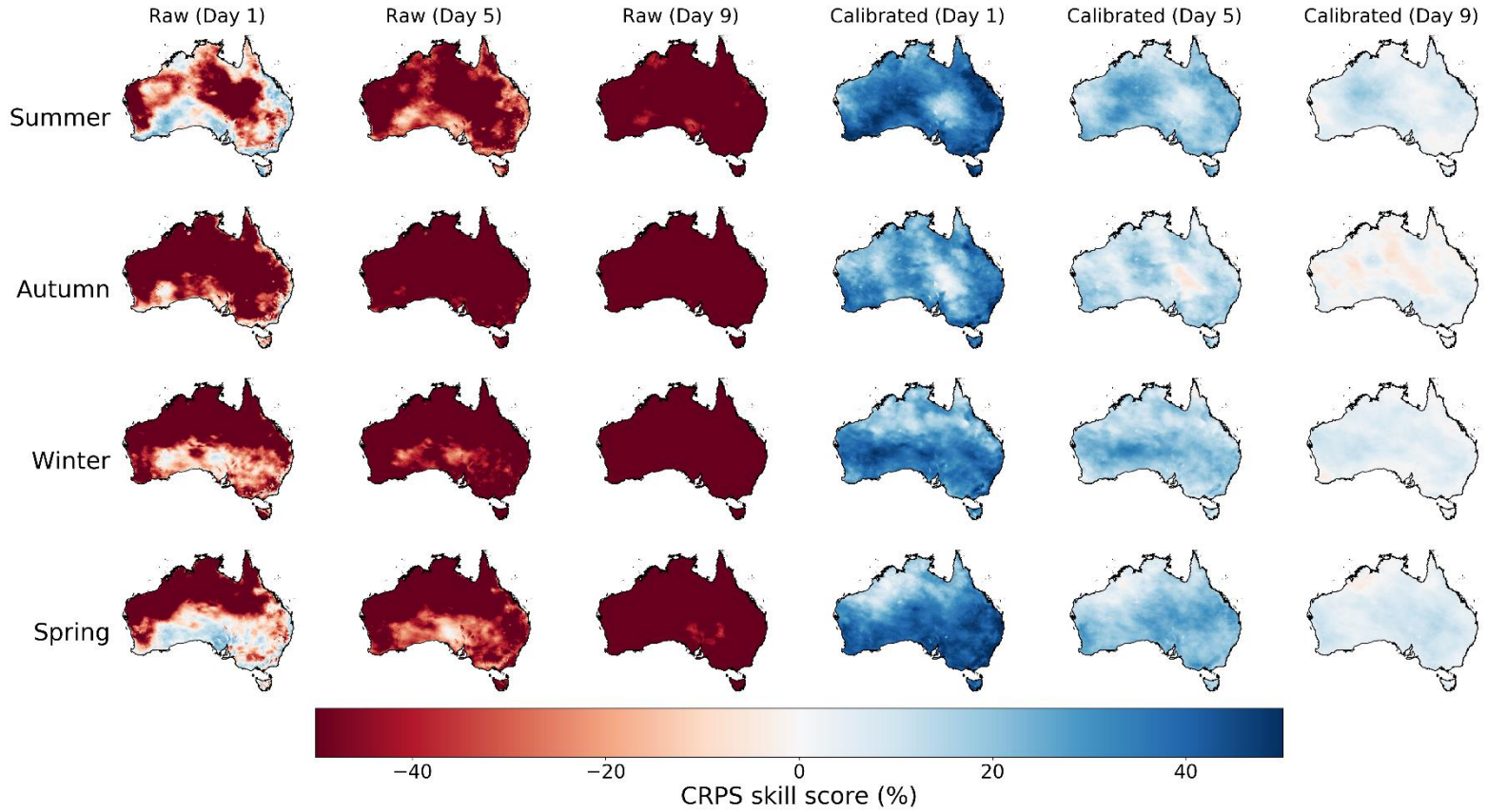


Fig. 7. Differences in the correlation coefficient (r) between calibrated ET_o forecasts from the calibration based on ET_o anomaly and observations, and the r of calibrated forecasts from the calibration working with ET_o forecasts directly and observations.



853

854 Fig. 8. CRPS skill score in raw (three columns on the left) and calibrated (three columns on the right) ET_o forecasts (calibration based
855 on ET_o anomaly) in different seasons.

856

857

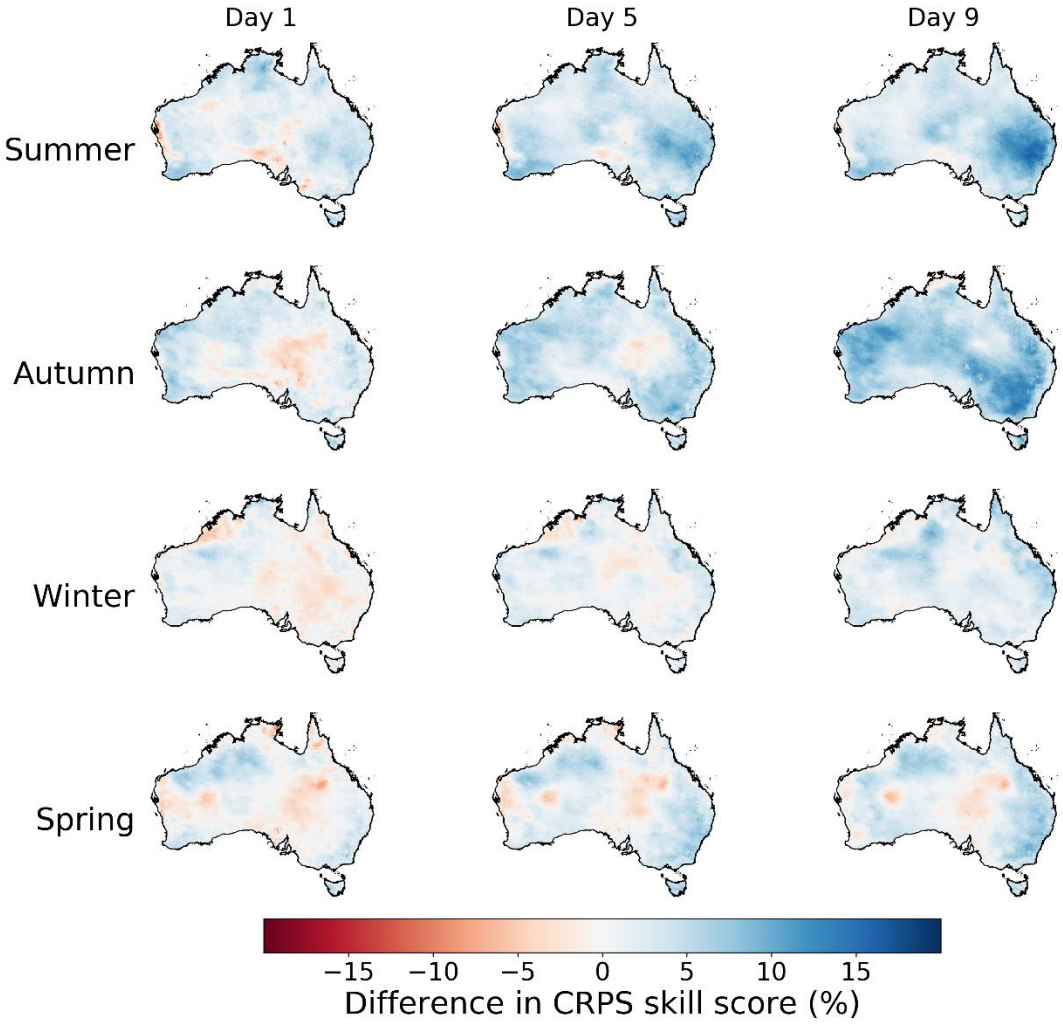


Fig. 9. Differences in CRPS skill scores of calibrated forecasts between the calibration based on anomaly, and the calibration calibrating ET_o forecasts directly in different seasons.

Supplementary Material

Calibrating anomalies improves forecasting of daily reference crop evapotranspiration

Qichun Yang[a,*], Quan J Wang[a], and Kirsti Hakala[a]

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

*: Corresponding author

E-mail address: qichun.yang@unimelb.edu.au

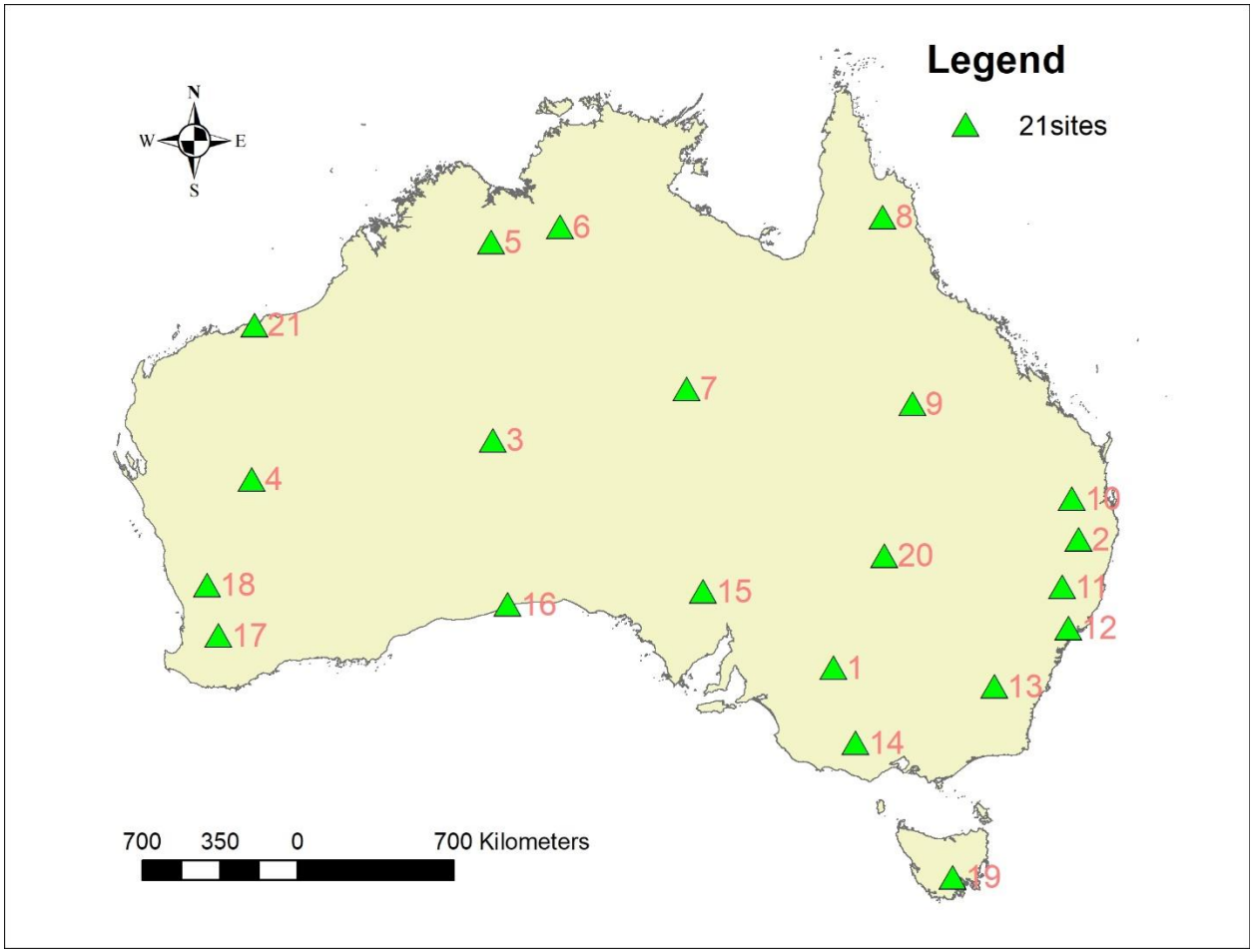
Telephone number: +61 411359526

881 **Table S1. 21 Weather stations selected for ET_o forecast calibration**

Name	Site number	latitude	longitude	ID
MILDURA AIRPORT	76031	-34.24	142.09	1
TENTERFIELD	56032	-29.0479	152.0172	2
GILES METEOROLOGICAL OFFICE	13017	-25.0341	128.301	3
MEEKATHARRA	7045	-26.6136	118.5372	4
WARMUN	2032	-17.0154	128.2174	5
VICTORIA RIVER DOWNS	14825	-16.403	131.0145	6
JERVOIS	15602	-22.9494	136.1442	7
PALMERVILLE	28004	-15.9999	144.0754	8
BARCALDINE POST OFFICE	36007	-23.5544	145.2883	9
OAKEY AERO	41359	-27.4034	151.7413	10
WOOLBROOK (WOOLBROOK ROAD)	55136	-30.9651	151.3505	11
PATERSON (TOCAL AWS)	61250	-32.6296	151.5919	12
BURRINJUCK DAM	73007	-34.9997	148.5984	13
ARARAT PRISON	89085	-37.2769	142.9786	14
WOOMERA AERODROME	16001	-31.1558	136.8054	15
EUCLA	11003	-31.6797	128.8958	16
NARROGIN	10614	-32.9342	117.1797	17
WONGAN HILLS	8137	-30.8917	116.7186	18
BUSHY PARK (BUSHY PARK ESTATES)	95503	-42.7097	146.8983	19
WANAARING POST OFFICE	48079	-29.7029	144.1484	20
PORT HEDLAND AIRPORT	4032	-20.3725	118.6317	21

882

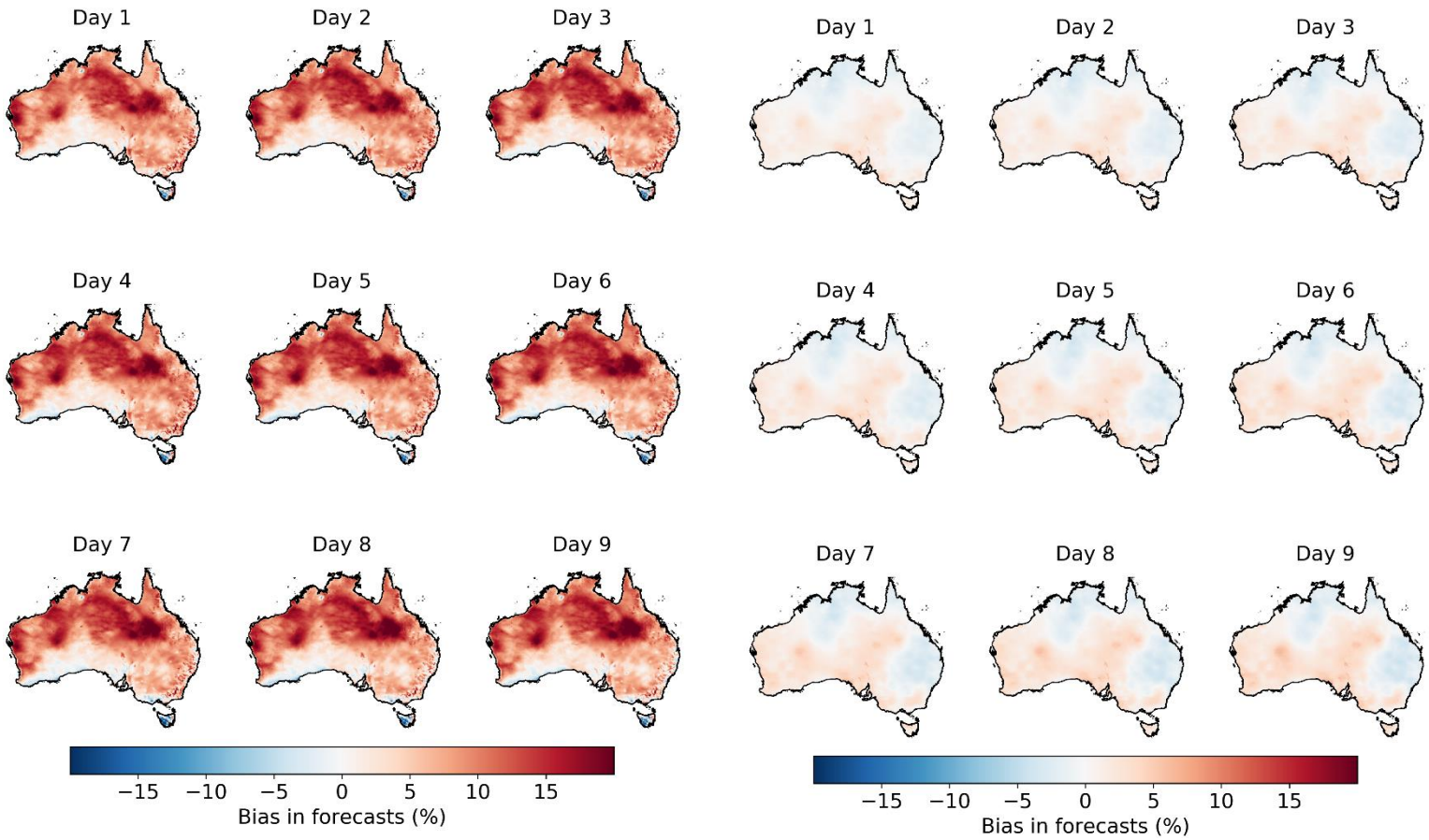
883



884

885 Fig. S1. Locations of 21 weather stations.

886



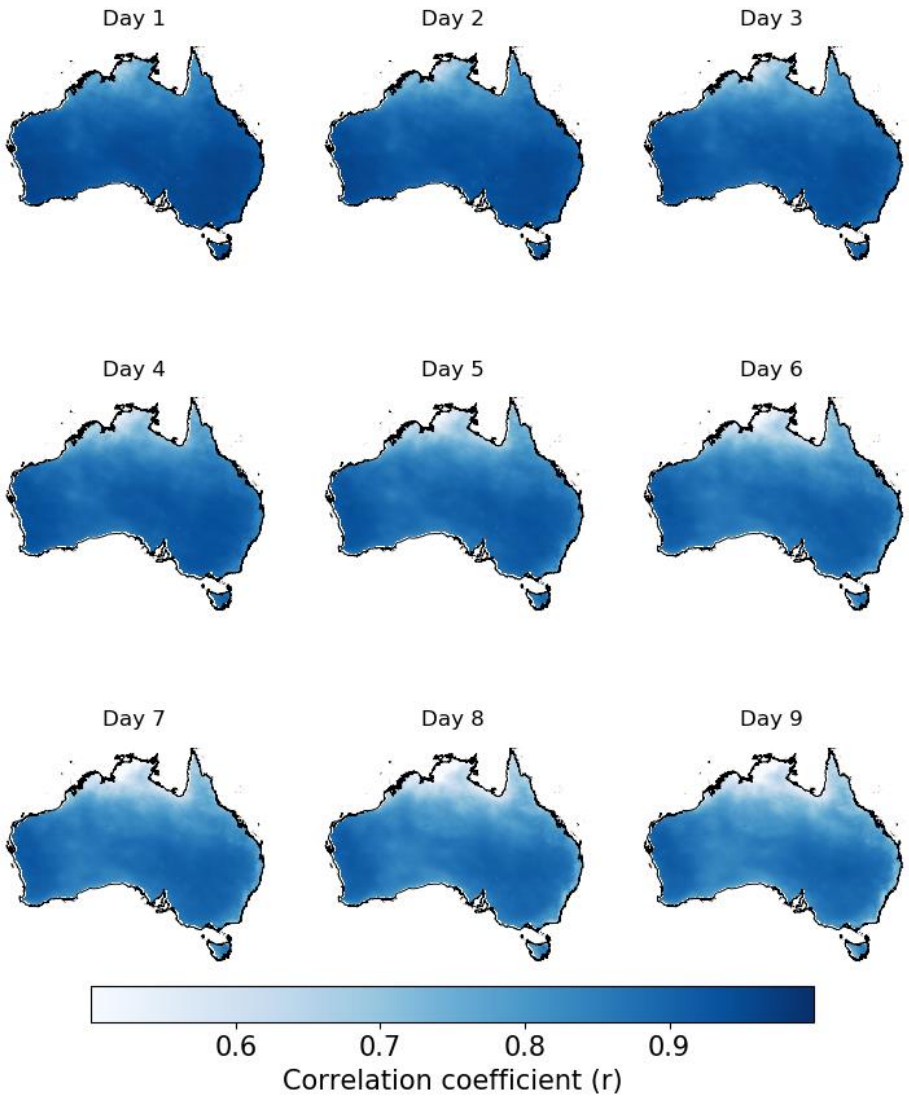
887

888

889 Fig. S2. Bias in raw (three panels on the left) and calibrated (three panels on the right) ET_o forecasts.

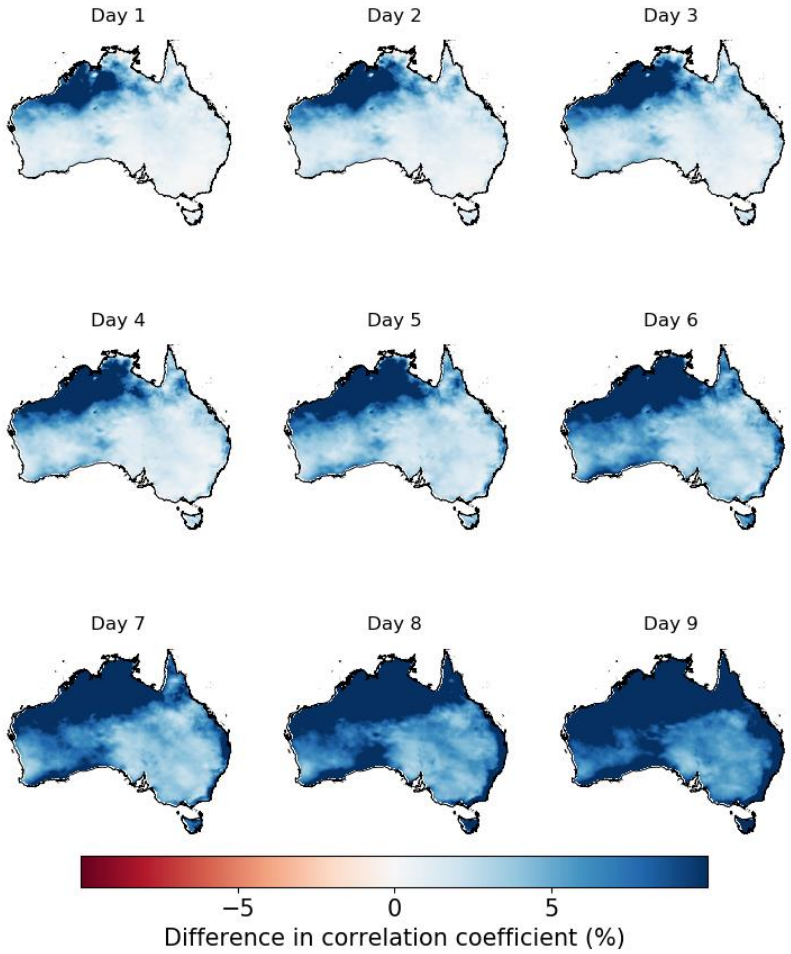
890

891



892

893 Fig. S3. Correlation coefficient (*r*) between calibrated ET_o forecasts (calibration based on ET_o anomaly) and ET_o observations.



894

895 Fig. S4. Differences in the correlation coefficient (r) between calibrated forecasts (calibration based on ET_o anomaly) and
896 observations, and r between the raw forecasts and observations.

897

898

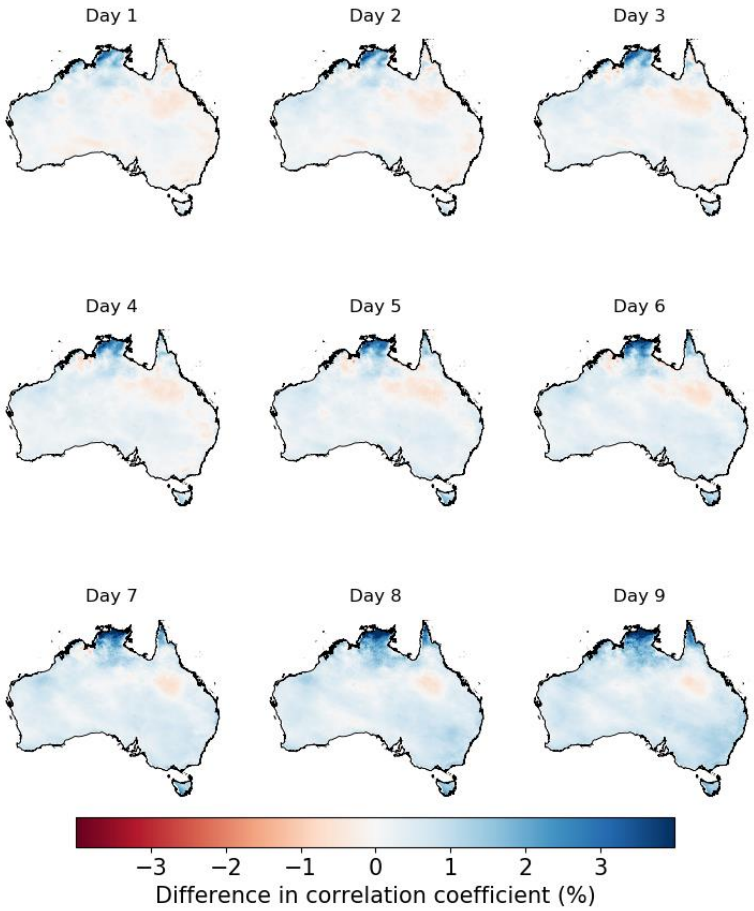
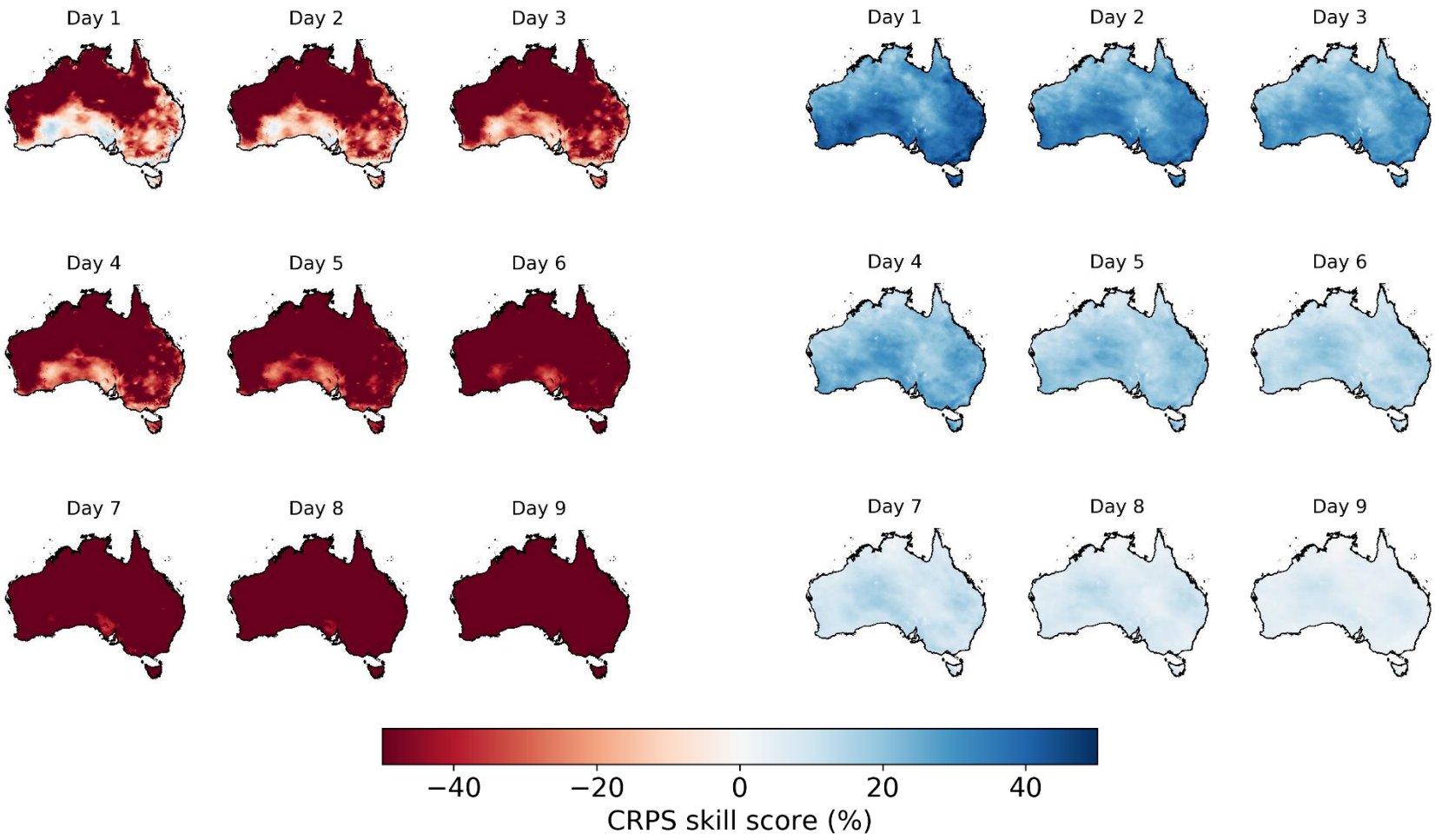


Fig. S5. Differences in the correlation coefficient (r) between calibrated forecasts (calibration based on ET_o anomaly) and ET_o observations vs. those from the calibration calibrating ET_o forecasts directly.



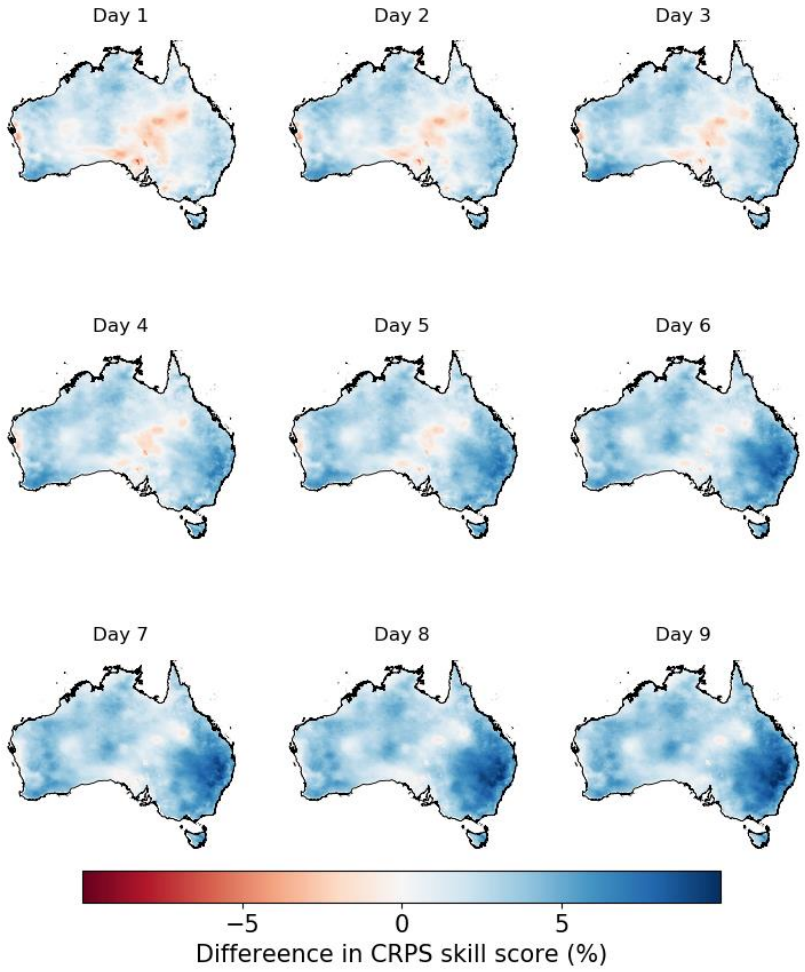
902

903

904 Fig. S6. CRPS skill score in raw (left three panels) and calibrated (right three panels) forecasts (calibration based on ET_o anomaly)
905 across Australia.

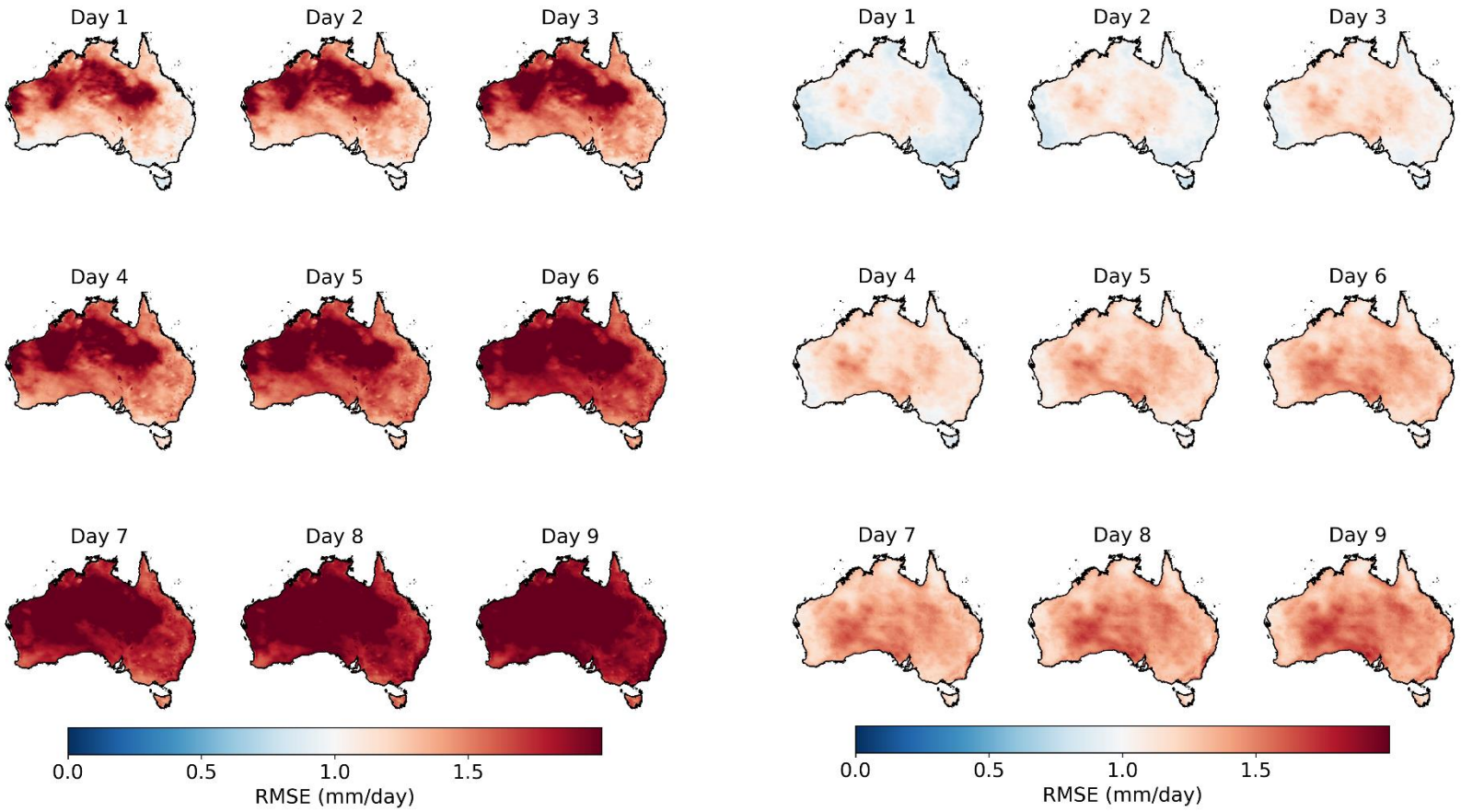
906

907



908

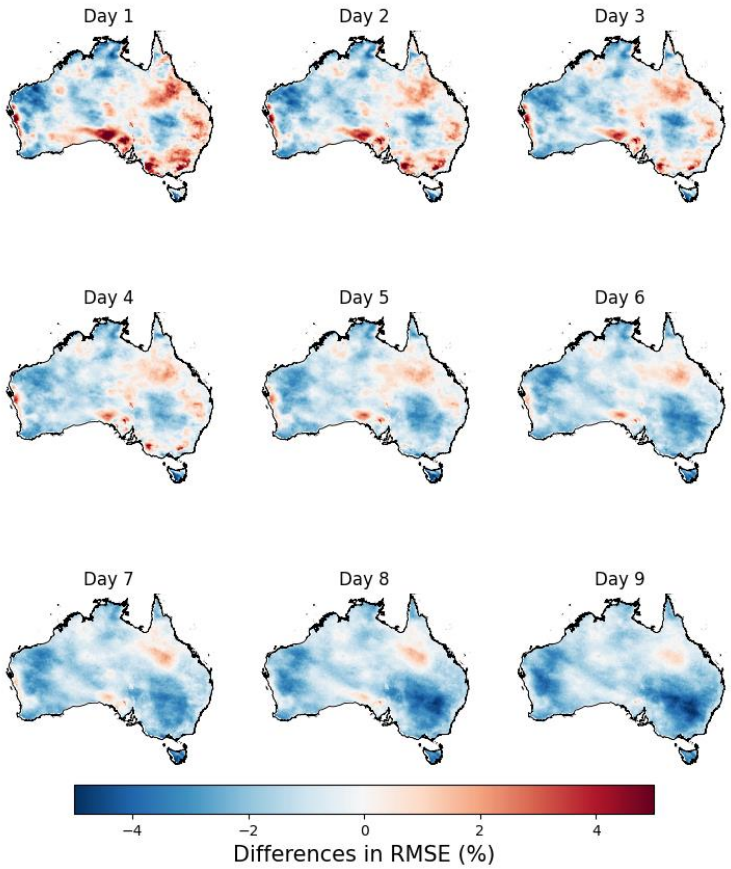
909 Fig. S7. Differences in CRPS skill scores of calibrated forecasts between the calibration based on anomaly, and the calibration based
910 on the original ET_o forecasts.



911

912 Fig. S8. RMSE in raw (three panels on the left) and calibrated (three panels on the right) ET_o forecasts (calibration based on ET_o
913 anomaly).

914



915

916 Fig. S9. Differences in RMSE between calibrated forecasts based on anomaly and those based on original ET_o forecasts.

917

918

919

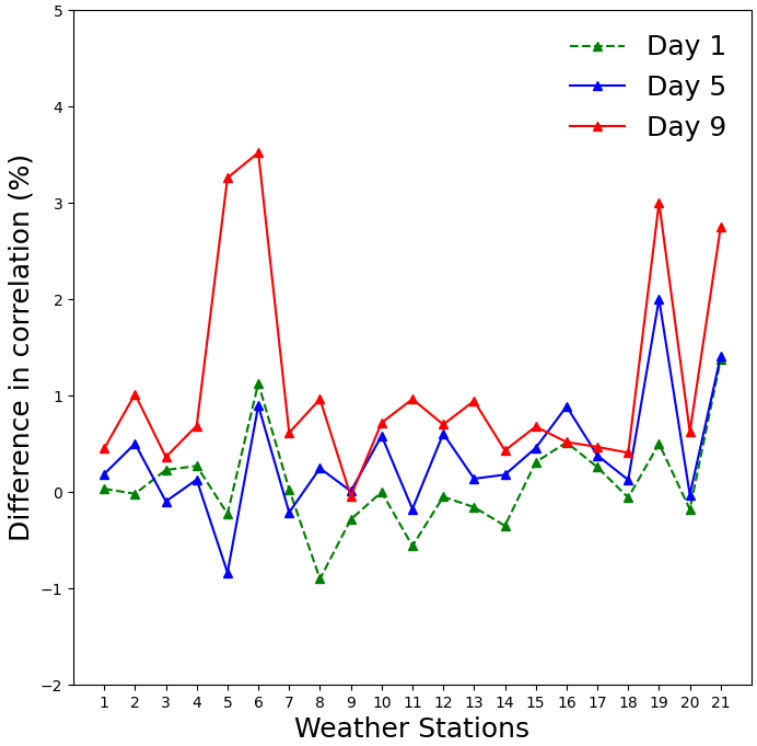
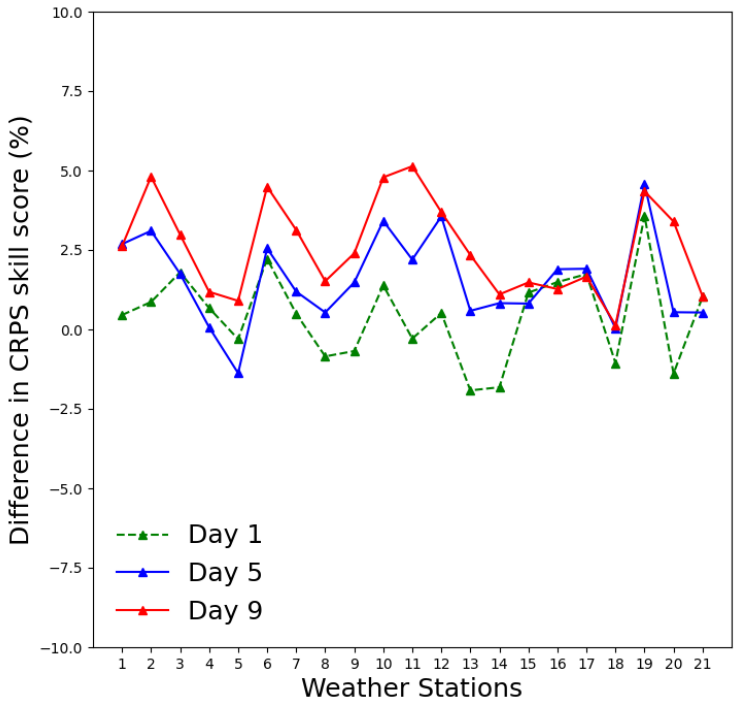


Fig. S10. Differences in correlation coefficients (*r*) between calibrated forecasts from the calibration based on ET_o anomaly and observation, and those between calibrated forecasts from the calibration calibrating ET_o forecasts and observations across 21 weather stations.



927

928 Fig. S11. Differences in CRPS skill scores of calibrated forecasts from the calibration based on ET_o anomaly vs. calibrated forecasts
929 from the calibration calibrating ET_o forecasts directly across 21 weather stations.

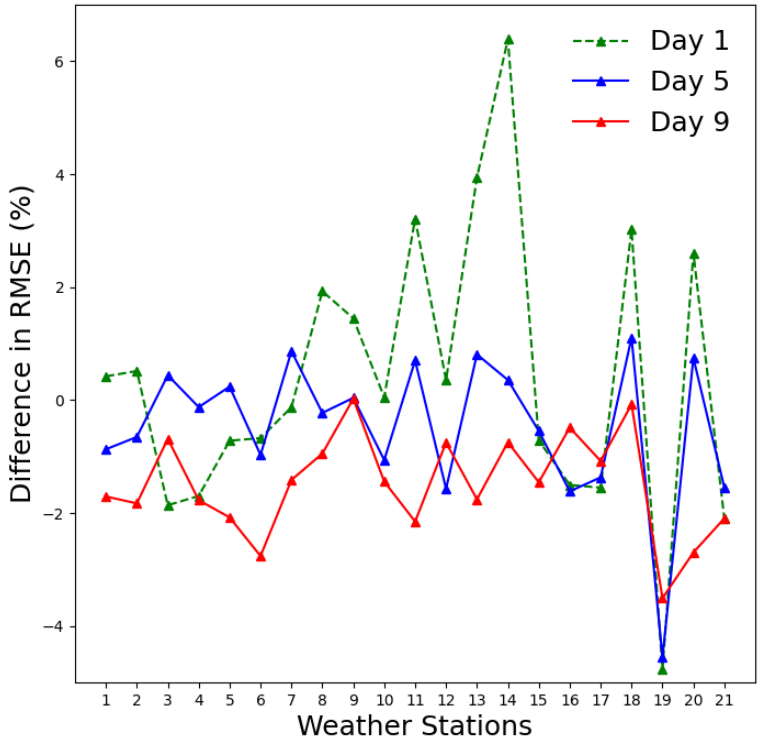


Fig. S12. Differences in RMSE of calibrated forecasts from the calibration based on ET_o anomaly vs. the calibration calibrating ET_o forecasts directly across 21 weather stations.