University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Zhao, P;Wang, QJ;Wu, W;Yang, Q

Title:
Extending a joint probability modelling approach for post-processing ensemble precipitation forecasts from numerical weather prediction models

Date:
2022-02-01

Citation:
Zhao, P., Wang, Q. J., Wu, W. & Yang, Q. (2022). Extending a joint probability modelling approach for post-processing ensemble precipitation forecasts from numerical weather prediction models. Journal of Hydrology, 605, https://doi.org/10.1016/j.jhydrol.2021.127285.

Persistent Link:
https://hdl.handle.net/11343/354212

Extending a joint probability modelling approach for post-processing ensemble precipitation forecasts from numerical weather prediction models

Pengcheng Zhao*, Quan J. Wang, Wenyan Wu, Qichun Yang

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

*: Corresponding author

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

ABSTRACT

Statistical post-processing has been widely employed to correct bias and dispersion errors in raw ensemble precipitation forecasts from numerical weather prediction models. One prominent post-processing scheme is to establish a joint probability model by fitting a bivariate distribution of raw forecasts and corresponding observations. However, current joint probability models only incorporate ensemble mean as the predictor, and ensemble spread is not considered. This is a major disadvantage of joint probability models as ensemble spread can be informative for forecast uncertainty. In this paper, we propose a two-step calibration approach to combine the strengths of joint probability models and the useful information included in the ensemble spread. In the first step, we take the seasonally coherent calibration (SCC) model as an example of joint probability models to calibrate the ensemble mean. As SCC for precipitation forecasts involves transformations for data normalization and special treatments of zero values, we explore three different ways to estimate ensemble mean values when establishing the SCC model. In the second step, we re-calibrate the ensemble forecasts produced in the first step to incorporate ensemble spread information from the raw forecasts. The performance of this two-step calibration is evaluated using ensemble precipitation forecasts from the Australian Bureau of Meteorology. We find that forecasts calibrated using the two-step calibration have better skills than SCC calibrated forecasts, especially for heavy precipitation events. Strengths of joint probability models and raw ensemble spread information are well utilized in the proposed two-step calibration approach.

Keywords: numerical weather prediction, ensemble precipitation forecasts, statistical post-processing, joint probability model, ensemble spread

1. Introduction

Short-term precipitation forecasting is routinely performed using numerical weather prediction (NWP) models in meteorological centers around the world (Harper *et al.*, 2007). These models are designed to simulate dynamic physics of the atmosphere forward in time to predict future weather conditions. Traditional deterministic NWP models produce single-value forecasts, hence cannot capture the chaotic nature of the atmosphere (Lorenz, 1963; Epstein, 1969). To account for the uncertainty of weather forecasts, ensemble prediction systems have been developed to generate multiple ensemble members from varying initial conditions or parametrizations of NWP models (Ehrendorfer, 1997; Buizza *et al.*, 1999; Palmer *et al.*, 2007). The superiority of ensemble forecasts compared to deterministic forecasts has been widely demonstrated in the literature (Richardson, 2000; Atger, 2001; Rodwell, 2006; Vokoun and Hanel, 2018; Zhao *et al.*, 2020).

Raw ensemble forecasts, however, often have bias (Gneiting *et al.*, 2007; Li *et al.*, 2017) and dispersion errors (Buizza *et al.*, 2005; Buizza, 2018). It has therefore been a common practice to apply statistical post-processing methods to improve raw forecasts (Verkade *et al.*, 2013; Yang *et al.*, 2020; Saminathan *et al.*, 2021). In this paper, we focus on the post-processing of univariate forecasts (i.e., point-by-point post-processing), without considering the spatial-temporal correlation in the forecasts.

A variety of post-processing methods have been developed. Particularly, the methods that are able to produce calibrated forecasts in the form of full predictive probability distributions are popular for forecast users (Brocker and Smith, 2008). Such post-processing methods allow the sampling of ensemble members of any desired size from well-calibrated predictive distributions, as well as the derivation of some frequently used prediction statistics, such as forecast quantiles and probabilities of threshold exceedance.

Two post-processing schemes are commonly used to produce calibrated probability distributions. The first scheme is based on distributional regressions, such as ensemble model output statistics (EMOS) (Gneiting *et al.*, 2005; Scheuerer, 2014; Baran and Lerch, 2015), ensemble dressing (Fortin *et al.*, 2006), Bayesian model averaging (BMA) (Raftery *et al.*, 2005; Sloughter *et al.*, 2007; Schmeits and Kok, 2010), and machine learning (ML) based regression methods (Dogulu *et al.*, 2015). This scheme typically makes use of statistics of raw ensemble forecasts, usually ensemble mean and ensemble spread, to establish probability distributions.

The second scheme is based on joint probability models, such as the Bayesian Joint Probability (BJP) model (Robertson *et al.*, 2013; Shrestha *et al.*, 2015; Cattoën *et al.*, 2020; Li *et al.*, 2020b; Li *et al.*, 2020c) and the meta-Gaussian distribution (MGD) model (Schaake *et al.*, 2007; Krzysztofowicz and Evans, 2008; Wu *et al.*, 2011; Li *et al.*, 2019). Joint probability models are often established by fitting a bivariate distribution of raw forecasts and corresponding observations. Conditional distributions of observations can be derived for calibration of new forecasts. Although joint probability models were originally designed to post-process single-value (i.e., deterministic) forecasts, they can also be applied to post-processing ensemble mean of ensemble forecasts.

A couple of advantages can be stated about joint probability models. First, they have been shown to be superior to distributional regression models when only ensemble mean is used as a predictor (Li et al., 2019a). One possible reason is that regression relationship in distributional regressions becomes weak when raw precipitation forecasts approach zero (Gebetsberger *et al.*, 2017). Second, because joint probability models take only the ensemble mean as the model predictor, they allow more flexibility in mathematical treatment to incorporate other complexities. For example, the seasonally coherent calibration (SCC) model (Wang *et al.*, 2019b; Zhao *et al.*, 2020; Yang *et al.*, 2021a) based on the joint probability scheme has been developed to produce calibrated forecasts

that are coherent in climatology, including seasonality, consistent with long-term observations even when archived NWP forecasts are limited.

Nevertheless, there is a major disadvantage of joint probability models. None of the current joint probability models have incorporated the ensemble spread, which is commonly assumed to be correlated with forecast uncertainty information (Toth *et al.*, 2001; Scherrer *et al.*, 2004; Grimit and Mass, 2007; Hopson, 2014). By contrast, many distributional regression models have gained additional forecast skills due to the incorporation of ensemble spread (Veenhuis, 2013; Messner *et al.*, 2014b; Scheuerer and Hamill, 2015). In this context, it will be highly valuable to explore a way to enable joint probability models to incorporate the ensemble spread.

In this paper, we develop a two-step calibration approach to integrate the additional forecast information included in the ensemble spread into joint probability models. For demonstration purposes, we employ the SCC model as an example of joint probability models and select precipitation as our target weather variable. In the first step, we apply SCC to calibrate ensemble mean of precipitation forecasts. As SCC for precipitation forecasts involves transformations for data normalization and special treatments of zero values, we explore a number of ways to estimate ensemble mean values when establishing the SCC model. In the second step, we re-calibrate the SCC calibrated forecasts from the first step to make use of raw ensemble spread information. With such an approach, the use of ensemble mean and ensemble spread is separated into two steps, allowing the SCC model to be deployed in its original form (i.e., taking ensemble mean as the only predictor).

The rest of this paper is organized as follows. Section 2 describes the NWP ensemble precipitation forecasts, observed data, and geographic area used for this study. Section 3 presents how we develop the two-step calibration approach with SCC and introduces evaluation metrics. Section 4

illustrates forecast verification results. Section 5 discusses advantages and possible limitations of the two-step calibration approach. Finally, main conclusions are presented in Section 6.

2. Data

Observed daily precipitation data are obtained from the Australian Water Availability Project's climate datasets (AWAP) (Jones *et al.*, 2009). With a high spatial resolution (0.05° x 0.05°), gridded AWAP data are obtained by interpolating rain gauge observations, which are collected from 0900 h of the previous day to 0900 h of the current day according to Australian local time, including daylight saving. Observations for a period of 30 years from 1 August 1989 to 31 July 2019 are used as reference data for SCC model establishments and for forecast evaluations.

We select ensemble precipitation forecasts from the Australian Community Climate and Earth-System Simulator Global Ensemble 2 (ACCESS-GE2) NWP model for forecast calibration. ACCESS-GE2 is based on the Met Office Global and Regional Ensemble Prediction System (MOGREPS) (Naughton, 2016) and has a horizontal grid spacing of approximately 60 km. Each ACCESS-GE2 ensemble consists of 24 exchangeable members that result from multiple stochastic disturbances to both model physics and initial conditions of the ACCESS-GE2 model. Ensemble forecasts are issued at 1200 UTC on a daily basis for lead times up to 10 days and are available at a 3-hourly temporal resolution. ACCESS-GE2 forecasts of a 3-year period from 1 August 2016 to 31 July 2019 are selected for this study.

To match the forecasts with observed data, we modify the spatial and temporal resolutions of ACCESS-GE2 data according to the AWAP data. We apply bilinear interpolation to re-grid the ACCESS-GE2 forecasts to the spatial resolution of AWAP. And we aggregate the 3-hourly

126 precipitation forecasts to daily values according to the AWAP records. Because AWAP data are
127 produced based on Australian local time which has UTC offsets, we adjust the matching hours of
128 ACCESS-GE2 and AWAP, and finally obtain daily ensemble precipitation forecasts for 9 days
129 ahead.

130 For a comprehensive forecast evaluation, we select 20 sites across a variety of climates in Australia,
131 as shown in Figure 1. The average annual precipitation is calculated based on the 30 years of
132 AWAP data.

133 [Figure 1]

134

135 3. Methods

136 In this section, we first introduce the formulation of the SCC modelling method. We then
137 demonstrate how we develop the two-step calibration approach with SCC. For the first step, three
138 SCC variants are derived based on different ways to estimate ensemble mean values: SCC1 and
139 SCC2 are established by calculating ensemble mean before and after data transformations,
140 respectively, and SCC3 is established with a proposed data augmentation approach for handling
141 (near) zero values of ensemble precipitation data. For the second step, a re-calibration (RC) method
142 is developed to further calibrate the SCC calibrated forecasts using raw ensemble spread
143 information. Details of the two-step calibration approach are illustrated in Figure 2. Finally, we
144 introduce verification methods used for forecast evaluation.

145 [Figure 2]

3.1 SCC model formulation

3.1.1 Log-sinh transformations and a joint probability model

The seasonally coherent calibration (SCC) model was developed based on the joint probability distribution to produce calibrated forecasts that are coherent in climatology, including seasonality, consistent with long-term observations even when archived NWP forecasts are limited (Wang *et al.*, 2019b). When used to post-process precipitation forecasts, raw forecasts and observations need to be normalized before a bivariate normal distribution is applied, as precipitation data are highly skewed. In this study, we employ log-sinh transformations (Wang *et al.*, 2012; Bennett *et al.*, 2014) for normalizing raw precipitation forecasts $x(t)$ ($t = 1,2, \dots, T$), where T is the total number of days in the 3-year period, and corresponding observations $y(t)$ to $f(t)$ and $o(t)$, respectively. The transformed variables $f(t)$ and $o(t)$ are then assumed to follow a bivariate normal distribution:

$$\begin{bmatrix} f(t) \\ o(t) \end{bmatrix} \sim N\left(\begin{bmatrix} u_f[m(t)] \\ u_o[m(t)] \end{bmatrix}, \begin{bmatrix} \sigma_f[m(t)]^2 & \rho[m(t)]\sigma_f[m(t)]\sigma_o[m(t)] \\ \rho[m(t)]\sigma_f[m(t)]\sigma_o[m(t)] & \sigma_o[m(t)]^2 \end{bmatrix} \right) \quad (1)$$

where $m(t)$ denotes the month for day t, namely $m(t) \in \{1, 2, \dots, 12\}$; $u_f[m(t)]$ and $\sigma_f[m(t)]$ are the mean and standard deviation of the marginal distribution of $f(t)$, and $u_o[m(t)]$ and $\sigma_o[m(t)]$ for $o(t)$; and $\rho[m(t)]$ is the correlation between $f(t)$ and $o(t)$. Because the bivariate normal distribution applies to continuous variables, a censoring technique is used to treat (near) zero precipitation as "censored" data. The term "censored" means these data are treated as below certain thresholds but are not precisely specified (Wang and Robertson, 2011; Messner *et al.*, 2014a; Scheuerer and Hamill, 2015; Zhao *et al.*, 2015; Stauffer *et al.*, 2017). Two thresholds x_c and y_c are applied here for $x(t)$ and $y(t)$ with corresponding thresholds f_c and o_c for better fitting $f(t)$ and $o(t)$ in their corresponding marginal distributions (Wang and Robertson, 2011). Here x_c

and y_c are set to 0.01 and 0.2 mm/day, reflecting the precision of available forecasts and observed data respectively, according to Robertson *et al.* (2013).

3.1.2 Reparameterization of the joint probability model

Post-processing models generally require long-term statistical characteristics of both forecasts and corresponding observations in order to produce forecasts that are coherent with the climatology of observations (Zhao *et al.*, 2017). However, although a long record of observations is often accessible, the archived record of forecasts is commonly short as NWP models are frequently updated. In this context, the SCC model was developed to work with limited NWP data to produce coherent calibrated forecasts. This significant advantage of the SCC model is achieved through the following two procedures.

First, a climatology that is representative of long-term statistical characteristics of observations is brought into the SCC model. In this study, we use 30 years of observed data to estimate the marginal distribution parameters of observations, namely the mean $u_o[m(t)]$ and standard deviation $\sigma_o[m(t)]$ for each month. Second, a reparameterization of the joint probability model is conducted to estimate $u_f[m(t)]$, $\sigma_f[m(t)]$, and $\rho[m(t)]$ for each month:

$$u_f[m(t)] = a + bu_o[m(t)] \tag{2}$$

$$\sigma_f[m(t)] = c + d\sigma_o[m(t)] \tag{3}$$

$$\rho[m(t)] = r \tag{4}$$

where b, c, $d \geq 0$ and $0 \leq r \leq 1$, and these 4 parameters are constrained to avoid nonsensical relationships. It should be noted that a can be either positive or negative, as $u_f[m(t)]$ can be positive or negative in the log-sinh transformed space. Equations (2) and (3) assume that NWP models, if run long enough, can produce the seasonality pattern that is linearly related with the

observed climatology. Equation (4) assumes that the underlying forecast skill of NWP models is constant for all 12 months. These assumptions are supported by a case study from Wang *et al.* (2019b) and may help reduce sampling effects of forecast evaluation results. The reparameterization of the joint probability model replaces 36 parameters ($u_f[m(t)]$, $\sigma_f[m(t)]$, and $\rho[m(t)]$, $m(t) \in \{1,2,\dots,12\}$) with five parameters ($a$, b, c, d, and r), therefore making it feasible to estimate model parameters from a short period of archived NWP forecasts. All of the parameters from the above two procedures can be derived using the method of maximum likelihood and the Nelder-Mead searching approach (Nelder and Mead, 1965). Related likelihood functions are given in Wang *et al.* (2019b).

3.1.3 Calibration of new forecasts

After obtaining all the parameters of the SCC model, we can apply SCC to calibrate new forecasts. Given a new forecast $x(t)$, a conditional distribution of $o(t)$ can be derived based on $f(t)$:

$$[o(t)|f(t)] \sim \mathbf{N}\{o(t)|\tilde{u}_o(t), \tilde{\sigma}_o^2(\mathrm{t})\} \tag{5}$$

where

$$\tilde{u}_o(t) = u_o[m(t)] + \rho[m(t)]\frac{\sigma_o[m(t)]}{\sigma_f[m(t)]}\{f(t) - u_f[m(t)]\} \tag{6}$$

$$\tilde{\sigma}_o^2(\mathrm{t}) = \{1 - \rho^2[m(t)]\}\sigma_o^2[m(t)] \tag{7}$$

We sample an ensemble of values from the conditional distribution to represent the forecast probability distribution. When $f(t) > f_c$, an ensemble $o(t,n)$ ($n = 1,2,\dots,N$) of any size N (100 for this study) can be sampled from the conditional distribution. When $f(t) \leq f_c$, a random value $f'(t)$ from the marginal distribution $\mathbf{N}\{f(t)|u_f[m(t)], \sigma_f^2[m(t)]\}$ is first sampled in the range of $[-\infty, f_c]$; then a sample can be drawn from the conditional distribution $[o(t)|f'(t)]$ to give an ensemble member. These two steps are repeated N times to form the ensemble $o(t,n)$. Finally, the

sampled ensemble members $o(t,n)$ can be transformed back by an inverse of the log-sinh transformation to give a calibrated ensemble forecast $y(t,n)$ in original scale. When there are negative ensemble members in $y(t,n)$, we set them to zero.

3.2 The proposed two-step calibration approach

3.2.1 Step One: Calibration of ensemble mean using SCC

The SCC model described above was originally developed to calibrate deterministic forecasts. When it comes to ensemble forecasts, ensemble mean can be calculated to provide the "deterministic" value. As SCC for precipitation forecasts involves data transformations and the censoring setup for (near) zero precipitation values, there are different ways to obtain the ensemble mean. Here we establish three SCC models, each corresponding to a particular way of estimating the ensemble mean values. The three models are named SCC1, SCC2, and SCC3 for our study.

The SCC1 model is established by calculating ensemble mean before data transformations. Denoting raw ensemble forecasts as $x(t,k)$ ($k = 1,2,\dots,K$), where K is the ensemble size (24 for ACCESS-GE2 ensembles), we calculate the ensemble mean prior to the data transformations, as in previous studies (Scheuerer and Hamill, 2015; Li *et al.*, 2019):

$$x_m(t) = \frac{1}{K}\sum_{k=1}^{K} x(t,k) \tag{8}$$

then $x_m(t)$ can replace the aforementioned $x(t)$ as the raw precipitation forecast series to establish the SCC1 model. The rest of the setup of the SCC1 model follows the same process as the original SCC model.

The SCC2 model is established by calculating ensemble mean after data transformations. We transform $x(t,k)$ with the log-sinh transformation to get $f(t,k)$ ($k = 1,2,\dots,K$). The

transformation involves all of the ensemble members (without any censoring). We can calculate the ensemble mean as below:

$$f_m(t) = \frac{1}{K}\sum_{k=1}^{K} f(t,k) \quad (9)$$

then $f_m(t)$ can replace the aforementioned $f(t)$ (Equations 1, 5, and 6) as the transformed forecast series to establish the SCC2 model. Likewise, the rest of the SCC2 model setup remains the same.

The SCC3 model is established with a proposed data augmentation approach for censored ensemble forecasts. The ensemble mean calculation becomes problematic when an ensemble contains censored ensemble members, because these members are only known to be below or equal to the censoring threshold x_c or f_c. It would be sensible to assign exact values to these censored data for calculating the ensemble mean. Here we propose a hierarchical model to simulate distributions of transformed ensemble forecasts. For any ensemble $f(t,k)$ ($k = 1,2,\dots,K$), ensemble members are assumed to follow a normal distribution:

$$f(t,k) \sim \mathbf{N}(f_u(t), \sigma^2) \quad (10)$$

where $f_u(t)$ ($t = 1,2,\dots,T$) and σ are the mean and standard deviation of the normal distribution, respectively. For each ensemble, $f_u(t)$ is a latent variable, and needs to be inferred from another normal distribution:

$$f_u(t) \sim \mathbf{N}(u_u, \sigma_u^2) \quad (11)$$

where u_u and σ_u are the mean and standard deviation of the normal distribution, respectively. To infer the parameters σ, u_u, and σ_u, the latent variable $f_u(t)$, and censored ensemble members, we apply Gibbs sampling (Gelfand, 2000) to iteratively sample these variables from their corresponding conditional distributions until the hierarchical model reaches a statistically steady

condition. Conditional distributions of these parameters and variables can be derived from the following likelihood function L:

$$L = \prod_{t=1}^{T} \prod_{k=1}^{K} \mathbf{N}(f(t,k)|f_u(t), \sigma^2) \times \mathbf{N}(f_u(t)|u_u, \sigma_u^2) \quad (12)$$

For example, in any sampling iteration, given $f_u(t)$ and σ, if there are censored members in an ensemble, we can draw random samples from the conditional distribution $\mathbf{N}(f(t,k)|f_u(t), \sigma^2)$ in the range of $[-\infty, f_c]$ to give the augmented values. After the augmentation of censored ensemble members, we obtain a new series of ensemble forecasts $f'(t,k)$ ($k = 1,2,\dots,K$). Then the ensemble mean can be calculated as below:

$$f'_m(t) = \frac{1}{K} \sum_{k=1}^{K} f'(t,k) \quad (13)$$

Similarly, $f'_m(t)$ can be used to replace the aforementioned $f(t)$ (Equations 1, 5, and 6) to establish the SCC3 model.

3.2.2 Step Two: Re-calibration of SCC calibrated forecasts using ensemble spread

To maximise forecast skills from post-processing raw ensemble forecasts, ensemble spread should also be incorporated into post-processing models. We select SCC3 out of the three SCC models established in the first step to demonstrate the re-calibration (RC) method, given that SCC3 is the most sophisticated in a sensible way and might perform best in the post-processing. We refer this integrated model as SCC3-RC in this study.

Generally, ensemble spread is represented using the standard deviation among ensemble members. Raw ensemble standard deviation $\sigma_{raw}(t)$ and SCC3 calibrated ensemble standard deviation $\sigma_{cali}(t)$ in transformed space can be calculated as:

$$\sigma_{raw}(t) = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} [f'(t,k) - f'_m(t)]^2} \quad (14)$$

$$\sigma_{cali}(t) = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}[o(t,n) - o_m(t)]^2} \quad (15)$$

where

$$o_m(t) = \frac{1}{N}\sum_{n=1}^{N} o(t,n) \quad (16)$$

is the ensemble mean of SCC3 calibrated ensemble $o(t,n)$ $(n = 1,2,\ldots,N)$. RC is formulated to adjust each ensemble member of $o(t,n)$ individually:

$$o_{RC}(t,n) = \lambda_1 \times o_m(t) + (o(t,n) - o_m(t)) \times (\lambda_2 + \lambda_3 \frac{\sigma_{raw}(t)}{\sigma_{cali}(t)}) \quad (17)$$

where $o_{RC}(t,n)$ is the re-calibrated ensemble member; λ_1, λ_2, and λ_3 are the RC parameters. Specifically, λ_1 represents the modification of the calibrated ensemble mean $o_m(t)$; λ_2 and λ_3 represent the contribution weights of $\sigma_{cali}(t)$ and $\sigma_{raw}(t)$, respectively. A larger λ_3 would indicate that more ensemble spread information from the raw forecasts is utilized. RC is only applied to calibrated ensemble forecasts whose corresponding raw ensemble mean $f'_m(t)$ is larger than f_c. This is because their raw ensemble spread is relatively more trustworthy than those calculated mostly based on augmented ensemble members.

We use a continuous ranked probability score (CRPS) minimization method (Gneiting *et al.*, 2005; Van Schaeybroeck and Vannitsem, 2015) to optimize λ_1, λ_2, and λ_3. Details of the CRPS are shown in Section 3.3. After obtaining the RC parameters, we can re-calibrate SCC3 calibrated ensemble members and then transform the re-calibrated ensemble members back to the original scale using the inverse of the log-sinh transformation.

The RC step is essentially an ensemble adjustment of forecast members. It has some similarity to the method of member-by-member post-processing (MBMP) (Johnson and Bowler, 2009; Schefzik, 2017). This point will be further discussed later in the Discussion section.

3.3 Forecast evaluation

In this study, the performances of the SCC1, SCC2, SCC3, and SCC3-RC models are evaluated using a leave-one-month-out cross-validation. Evaluation diagnositics include bias, CRPS (Hersbach, 2000), threshold weighted CRPS (twCRPS) (Gneiting and Ranjan, 2011), and reliability diagram (Wilks, 2011) to assess different aspects of forecast quality.

Bias is the difference between the mean of precipitation forecasts and the mean of corresponding observations:

$$Bias = \frac{1}{T}\sum_{t=1}^{T} x(t) - \frac{1}{T}\sum_{t=1}^{T} y(t) \quad (18)$$

where $x(t)$ and $y(t)$ are forecasts and observed values at time t, respectively; and T is the length of forecast data records. It is important to have minimal bias in precipitation forecasts as bias can be amplified in streamflow forecasting and in other water resource management.

CRPS quantifies the difference between ensemble forecast cumulative distribution and corresponding observations (Hersbach, 2000). The average CRPS for days t=1, 2, …, T is formulated as:

$$CRPS = \frac{1}{T}\sum_{t=1}^{T} \int \left\{F(t,x) - H\big(x - y(t)\big)\right\}^2 dx \quad (19)$$

where $F(t,x)$ is the forecast cumulative density function (CDF), and $y(t)$ is the observation at time t; H is the Heaviside step function that equals 1 if $x - y(t) \geq 0$ and equals 0 otherwise; and T is the length of data records.

We also calculate the CRPS of reference climatology forecasts ($CRPS_{ref}$) using the leave-one-month-out cross-validated climatology ensemble forecasts that are generated from the SCC models. A CRPS skill score can then be calculated as:

$$CRPS\ skill\ score = \frac{CRPS_{ref} - CRPS}{CRPS_{ref}} \times 100(\%) \quad (20)$$

The CRPS skill score is positively oriented and represents the relative improvement of the calibrated forecasts compared to the referenced climatology forecasts. A maximum skill score of 100% indicates that forecasts perfectly match the corresponding observations and a skill score of 0% indicates that forecasts have comparable errors to the reference forecasts. Forecasts poorer than reference forecasts have negative skill scores. Note that the reference climatology forecasts are common to all the SCC models, and therefore CRPS skill scores achieved by the different models can be directly compared with each other.

We further use twCRPS (Gneiting and Ranjan, 2011) to evaluate the model performance on predicting heavy precipitation events (i.e. precipitation amount above a certain threshold). The average twCRPS for days t=1, 2, …, T is formulated as:

$$twCRPS = \frac{1}{T}\sum_{t=1}^{T}\int\left\{F(t,x) - H\left(x - y(t)\right)\right\}^2 \omega(x)dx \quad (21)$$

where $\omega(x)$ is a weight function that equals 1 if $x \geq q$ and equals 0 otherwise; and q is a given threshold, with precipitation above q marked as heavy precipitation events. In this study, q is set to the 95% quantile of observed values. Similarly, we use twCRPS skill score to demonstrate the improvement of calibrated forecasts relative to reference forecasts.

In addition, we apply the reliability diagram as a graphical tool to evaluate the reliability of ensemble forecast uncertainty (ensemble spread not too narrow or too wide). Reliability refers to the statistical consistency between ensemble forecasts and corresponding observations. A reliability diagram illustrates forecast reliability by plotting observed frequencies against predicted probabilities based on a threshold exceedance for ensemble forecasts. In this study, we use 0.2 mm/day and 95% quantile of observed values as two thresholds for constructing the diagrams.

Precipitation values exceeding 0.2 mm/day and the 95% quantile of observations represent the occurrence of light precipitation and the occurrence of heavy precipitation, respectively. Resulting plots in reliability diagrams close to the 1:1 line indicate good reliability, while deviation from the line indicates poor reliability. Perfectly reliable forecasts will show a plot overlapping with the 1:1 line.

4. Results

4.1 Overall forecast evaluation of bias and CRPS skill score

Results of bias and CRPS skill score for raw ensemble forecasts and calibrated forecasts from the SCC1, SCC2, SCC3, and SCC3-RC models are shown in Figure 3. It can be seen from the figure that raw ensemble forecasts overall perform worst in terms of bias and CRPS skill score among these forecasts. The SCC models greatly improve the raw ensemble forecasts, with most of the bias being closer to zero and much greater CRPS skill scores being achieved, especially at short lead times. While there are many negative CRPS skill scores in raw ensemble forecasts, only a few negative skill scores are shown in the calibrated forecasts. Bias has no clear trend over the lead time, while CRPS skill score tends to decrease gradually as the lead time increases, indicating lower skills of forecasts at longer lead times.

[Figure 3]

Besides, SCC2, SCC3, and SCC3-RC have overall better performance than SCC1 on bias. However, it is quite hard to distinguish these four models according to their CRPS skill scores at the scale shown in Figure 3. To further learn about the contributions of the ensemble mean calculations after data transformations, the data augmentation, and the re-calibration for

incorporating raw ensemble spread information, we look more closely at the CRPS skill score of each of the models for each site and each lead time in the following two subsections.

4.2 CRPS skill scores of calibrated forecasts from the three SCC variants

Results of CRPS skill score for calibrated forecasts from the SCC1 model is shown in Figure 4(a). Similar to the results discussed in section 4.1, CRPS skill score of the SCC1 calibrated forecasts is negative just at a few sites and only at some lead times, indicating the evident benefits of the SCC post-processing. Results on CRPS skill score improvements of SCC2 and SCC3 over SCC1 are shown in Figure 4(b) and 4(c) and further summarised in Table 1. As quantitative forecasts of precipitation for a particular location and lead time is highly challenging, even a 1% skill score improvement is often considered a meaningful gain for newly implemented NWP models or post-processing methods (Messner *et al.*, 2014b; Scheuerer and Hamill, 2015). For either the SCC2 or SCC3 model, the CRPS skill score difference from the SCC1 model varies with sites and lead times. The overall difference tends to be positive, indicating improvements over SCC1.

[Figure 4]

[Table 1]

Calculating ensemble mean after transformations (i.e. the SCC2 model) improves the CRPS skill score compared to calculating ensemble mean before transformations (i.e. the SCC1 model), especially for sites 2, 13, 17, and 20, although there are sites with reduced skill scores, such as sites 3, 4, and 12. In addition, the data augmentation for censored ensemble members in the SCC3 model further improves the CRPS skill score of SCC2 calibrated forecasts. The most greatly improved site is site 14. As shown in Table 1, on the basis of the skill score improvement of SCC2

compared to SCC1, the SCC3 model further increases the number of improved cases, and the total accumulated average improved skill score of SCC3 compared to SCC1 reaches 1.13%.

4.3 CRPS skill scores of calibrated forecasts from the two-step calibration

Results of CRPS skill score for SCC3-RC calibrated forecasts are shown in Figure 5 and Table 1. SCC3-RC calibrated forecasts have positive CRPS skill scores at all sites and lead times (Figure 5(a)). This assures that the calibrated forecasts are more skillful than the referenced climatology forecasts in all cases. Compared to SCC1, the accumulated average CRPS skill score improvement of SCC3-RC reaches 1.68%, which is quite prominent considering all the sites and lead times.

[Figure 5]

We also compare the forecast performance of SCC3-RC and SCC3, to learn how the incorporated raw ensemble spread information in the re-calibration influences the model performance. As shown in Table 1, the re-calibration method improves the skill score of 129 cases, with improved skill scores mainly at short lead times and mostly within 4% (Figure 5(c)). The average skill score improvement due to raw ensemble spread utilization is 0.55%, which is comparable to previous studies on distributional regression-based post-processing models (Messner *et al.*, 2014b; Scheuerer and Hamill, 2015). This indicates that the re-calibration method can make SCC calibrated ensemble forecasts more skillful by using raw ensemble spread information. However, it should be noted that there are also reductions in the skill score at some sites, such as sites 2, 17, and 19 (Figure 5(c)).

4.4 twCRPS skill scores of calibrated forecasts from the three SCC variants

We further evaluate the ability of the calibration models in predicting heavy precipitation events, which are crucial for some common hydrological applications such as irrigation and flood

403 forecasting. The twCRPS skill score for calibrated forecasts from the SCC1 model is shown in
404 Figure 6(a). Clearly, SCC1 calibrated forecasts have more negative values of the twCRPS skill
405 score compared to the CRPS skill score (Figure 4(a)). This highlights again the challenge of
406 forecast post-processing for predicting heavy precipitation events as previously identified in the
407 literature (Taillardat *et al.*, 2019; Li *et al.*, 2020a). Results on the improvements in the twCRPS
408 skill score of SCC2 and SCC3 over SCC1 are shown in Figure 6(b) and 6(c) and further
409 summarised into Table 2. Both SCC2 and SCC3 have overall higher twCRPS skill score than
410 SCC1, and the improvements are larger than those in the CRPS skill score (Table 1). This shows
411 that SCC2 and SCC3 overall perform better than SCC1 for predicting heavy precipitation events.
412 Also, it is worth noting that there are still reduced twCRPS skill scores at several sites (Figure 6(b)
413 and 6(c)).

414 [Figure 6]

415 [Table 2]

416 **4.5 twCRPS skill scores of calibrated forecasts from the two-step calibration**

417 The twCRPS skill scores of SCC3-RC calibrated forecasts are shown in Figure 7 and Table 2.
418 SCC3-RC greatly improves SCC1 in terms of the twCRPS skill score, with the average
419 improvement exceeding 3%, which is almost twice as much as the average CRPS skill score
420 improvement (Table 1). Besides, SCC3-RC also has higher twCRPS skill score than SCC3, with
421 the average improvement close to 1%. The two-step calibration is shown to gain considerable
422 twCRPS skill score improvements, although there are slightly degraded skill scores at some sites
423 (Figure 7(b) and 7(c)).

Together with results presented in sections 4.2 – 4.4, it can be concluded that on the average of the 20 sites and 9 lead times, in the order of SCC1, SCC2, SCC3, and SCC3-RC, both CRPS skill score and twCRPS skill score tend to increase successively, indicating gradual improvements. Besides, the improvements on twCRPS skill score have higher magnitudes than those on CRPS skill score (Tables 1 and 2), meaning that the latter three models (especially SCC3-RC) are far more capable of predicting heavy precipitation events than the SCC1 model, on average.

[Figure 7]

To further demonstrate the significance level of the advantages of SCC2, SCC3, and SCC3-RC over SCC1, we carry out a strict statistical significance testing with bootstrapping on the CRPS (twCRPS) skill score results of forecasts calibrated using these models. According to the testing results (shown in Table A1), the CRPS skill score difference between any two models is not statistically significant. However, the twCRPS skill score difference between SCC3-RC and SCC1 is found to be significant. This further confirms the superiority of the SCC3-RC model on producing more skillful calibrated forecasts for heavy precipitation events than SCC1. Besides, it should also be noted that the differences in the CRPS (twCRPS) skill score obtained from SCC3-RC and SCC3 are not statistically significant, although the improvement of SCC3-RC compared to SCC3 is shown to be overall considerable. Details of the statistical testing and interpretation can be found in Appendix A.

4.6 Reliability diagrams of calibrated forecasts from different models

The reliability diagrams of calibrated forecasts from SCC1, SCC2, SCC3, and SCC3-RC as well as raw ensemble forecasts are shown in Figure 8. As can be seen from the figure that all of the calibrated forecasts clearly show better reliability than raw ensemble forecasts for the two thresholds selected, i.e. 0.2 mm/day and the 95% quantile of observed values. For the occurrence

of light precipitation events, SCC2 and SCC3 have poorer reliability performance than SCC1, while SCC3-RC has slightly better reliability than SCC1 (Figure 8(a)). For the occurrence of heavy precipitation events, SCC2, SCC3, and SCC3-RC have similar reliability performance and all of them perform better than SCC1 (Figure 8(b)).

[Figure 8]

5. Discussion

The calculation of ensemble mean and ensemble spread with censored ensemble members is problematic as there are no exact values for these members in transformed space. One feasible solution is to augment exact values for censored data. Many mathematical methods have been developed for the augmentation of datasets with censored records (Cohen, 1961; Hornung and Reed, 1990; Perkins *et al.*, 1990). These methods are generally developed based on the maximum likelihood estimation to augment values for one single dataset. However, when applied to multiple datasets which are a series of ensembles in our study, augmenting values for each ensemble individually cannot take into consideration the relationship among different ensembles, given that ensemble members from all ensembles are transformed using the same set of parameters in the data normalization process. Besides, these methods might become problematic when most or even all of the ensemble members are censored. By contrast, the data augmentation algorithm we propose in this study can connect all ensembles by establishing a hierarchical model. And ensembles even with a large proportion of censored members can still be augmented in a sensible way. These advantages make our algorithm more appropriate for the data augmentation of censored ensemble weather forecasts.

The re-calibration method employed in our study is similar to member-by-member post-processing (MBMP) methods (Johnson and Bowler, 2009; Van Schaeybroeck and Vannitsem, 2015; Schefzik, 2017), which post-processes raw ensemble forecasts by adjusting ensemble members individually. These MBMP methods are often implemented by minimizing objective scoring functions (e.g. CRPS) of the adjusted ensemble forecasts. They can make use of raw ensemble spread information but are generally not sophisticated enough to reflect statistical forecast uncertainty derived from historical weather events. In this context, the re-calibration is developed based on MBMP methods to further post-process the SCC calibrated forecasts using raw ensemble spread information. Forecasts calibrated using the two-step calibration can reflect both the dynamically flow-dependent ensemble spread from raw ensemble forecasts and the SCC calibrated ensemble spread that contains statistically generated uncertainty information.

Accurate and reliable ensemble forecasts for heavy precipitation events are crucial for hydrological forecasting and have attracted much attention from the post-processing perspective. It is therefore valuable to evaluate the performance of post-processing models over heavy precipitation events. However, the restriction of routine forecast evaluations to only observed heavy events often has bias effects and may degrade even the most skillful forecasts available (Diks *et al.*, 2011; Gneiting and Ranjan, 2011). This is also referred to as the forecaster's dilemma in Lerch *et al.* (2017). In this study, we apply the twCRPS to evaluate calibrated ensemble forecasts with emphasis on heavy precipitation events. As a proper weighted scoring rule, the twCRPS metric takes into consideration all precipitation events while evaluating the forecast distribution tails. It avoids possible bias due to the stratified sampling conditioned on observations and therefore provides a remedy for the dilemma.

In this study, the three SCC variants differ in how we calculate the ensemble mean for the SCC post-processing. Indeed, a number of studies take ensemble mean as a summarized statistic in operational ensemble forecasting as well as post-processing, as ensemble mean generally has smaller errors than any individual ensemble member when averaged over many cases (Scheuerer and Hamill, 2015; Wang *et al.*, 2019a). However, it has also been recognized that in complex spatial fields, ensemble mean can sometimes "smear out" some important features by decreasing high amplitudes and increasing the spatial coverage of low values, especially for ensemble forecasts with extreme events (Ebert, 2001; Surcel *et al.*, 2014). Some approaches, such as the probability matching (PM) method (Ebert, 2001; Clark, 2017), have been developed to solve this issue by modifying the ensemble mean in some statistical ways. Therefore, a simple ensemble mean (as implemented in SCC1) may not be the most appropriate representation of ensemble forecasts for the SCC post-processing. Further experiments are needed in the future to investigate if the other two ensemble mean calculation approaches (as implemented in SCC2 and SCC3) can help alleviate the smearing effect.

We acknowledge that our study is only conducted based on ensemble forecasts on a daily basis and with a grid spacing of about 60 km. As temporally and spatially high-resolution weather forecasts have attracted great attention recently, whether the developed models in this study are applicable to, for example, hourly NWP forecasts with a finer grid spacing, needs to be further investigated. Besides, establishing calibration models for precipitation is widely known to be more challenging than other weather variables. Indeed, as a highly sophisticated model compared to traditional joint probability models, SCC has also been shown to be applicable to other variables (Yang *et al.*, 2021b). In view of the forecast improvements in our study, we therefore anticipate

513 our models to be robust for application to other joint probability models and to post-processing
514 other weather variables.

515

516 ## 6. Summary and conclusions

517 In this study, we aim to integrate the ensemble spread into joint probability models. A two-step
518 calibration approach is developed for ensemble precipitation forecasts with the seasonally coherent
519 calibration (SCC) model as an example of joint probability models. In the first step, we employ
520 the SCC model to calibrate ensemble mean. In the second step, we re-calibrate the SCC calibrated
521 forecasts to incorporate the ensemble spread information.

522 As SCC for precipitation forecasts involves transformations for data normalization and special
523 treatments of zero values, we investigate three different ways to estimate ensemble mean values
524 when establishing the SCC model in the first step. We find that calculating ensemble mean after
525 transformations overall performs better than that before transformations for SCC. We therefore
526 recommend ensemble mean to be calculated in transformed space when post-processing ensemble
527 precipitation forecasts. In addition, we propose a data augmentation algorithm to estimate
528 ensemble mean (and ensemble spread for use in the second step) to handle zero precipitation values.
529 The resulting ensemble mean values are shown to lead to even better forecast calibration.

530 For the second step (i.e. the re-calibration), we develop an ensemble adjustment method to adjust
531 the SCC calibrated ensemble members individually using raw ensemble spread information. The
532 additional forecast information extracted by the re-calibration leads to forecast skill increase
533 comparable to that achieved in past studies of using raw ensemble spread information. This means

that the two-step approach is able to utilize the ensemble spread information of raw ensemble precipitation forecasts while preserving the strengths of the SCC model.

The two-step calibration approach, namely SCC3-RC in our study, has been found to fairly improve the performance of the original SCC in terms of the bias, forecast skill and forecast reliability. The improvement is especially notable for heavy precipitation events. It is expected that the two-step calibration approach can be adapted for other joint probability models and for post-processing other weather variables.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement

Data used in this study are produced by the Australian Bureau of Meteorology and accessed via the National Computational Infrastructure system. Please contact the Bureau of Meteorology to request data access.

Acknowledgements

This work is supported by an Australian Research Council Linkage Project (Grant No. LP170100922) and a collaborative project (Grant No. TP707466) between the University of Melbourne and Australian Bureau of Meteorology. The co-author Wenyan Wu acknowledges the

555 support of the Australian Research Council via the Discovery Early Career Researcher Award
556 (DE210100117). We would like to thank the Australian Bureau of Meteorology for supplying the
557 ACCESS-GE2 and AWAP data. We also thank the National Computational Infrastructure for
558 providing assess to computation resources to support our work. We gratefully acknowledge the
559 two reviewers for their thorough reviews and constructive comments.

560

561 Appendix A. Statistical significance testing of two samples with bootstrapping

562 We employ the algorithm in Efron and Tibshirani (1994) to implement the statistical significance
563 testing between two samples. In our study, assuming that the CRPS (twCRPS) skill score values
564 of forecasts calibrated by two different models are samples $X(x_1, x_2, \cdots, x_n)$ and $Y(y_1, y_2, \cdots, y_n)$,
565 the sample size n of both samples will be 180 (20 sites and 9 lead times in total). Steps of the
566 significance testing with bootstrapping are as follows:

567 (1) Calculate the test statistic t for X and Y:

568 $$t = \frac{X_m - Y_m}{\sqrt{\frac{{\sigma_X}^2}{n} + \frac{{\sigma_Y}^2}{n}}} \quad \text{(A1)}$$

569 where X_m and σ_X are the mean and standard deviation of sample X, and Y_m and σ_Y for sample Y.

570 (2) Calculate the mean XY_m of the combined sample of X and Y:

571 $$XY_m = \frac{X_m + Y_m}{2} \quad \text{(A2)}$$

572 (3) Construct two new samples $X'(x'_1, x'_2, \dots, x'_n)$ and $Y'(y'_1, y'_2, \dots, y'_n)$:

573 $$x'_i = x_i - X_m + XY_m, \ (i = 1,2, \dots, n) \quad \text{(A3)}$$

574 $$y'_i = y_i - Y_m + XY_m, \ (i = 1,2, \dots, n) \quad \text{(A4)}$$

575 (4) Apply bootstrapping to draw two random samples X^* and Y^* of size n from X' and Y',
576 respectively.

577 (5) Calculate the test statistic t^* for X^* and Y^*:

578 $$t^* = \frac{X_m^* - Y_m^*}{\sqrt{\frac{{\sigma_{X^*}}^2}{n} + \frac{{\sigma_{Y^*}}^2}{n}}} \quad \text{(A5)}$$

579 where X_m^* and σ_{X^*} are the mean and standard deviation of sample X^*, and Y_m^* and σ_{Y^*} for sample
580 Y^*.

581 (6) Repeat Step (4) and Step (5) K (i.e., a large value) times to obtain K t^* values.

582 (7) Count the number of t^* values that are equal to or greater than t as N.

583 (8) Estimate the p-value:

584 $$p = N/K \quad \text{(A6)}$$

585 The null hypothesis of the significance testing is that X and Y are from a distribution with the same
586 mean, and the alternative hypothesis is that they are not. The null hypothesis will be rejected at
587 significance level α if

588 $$p < \alpha \quad \text{(A7)}$$

589 In this study, we first choose the most often used α value 0.05 as the significance level and choose
590 $K = 100000$ to keep the sampling errors as small as possible. For CRPS (twCRPS) skill score
591 results, we implement 4 significance tests between SCC2 and SCC1, SCC3 and SCC1, SCC3-RC
592 and SCC1, and SCC3-RC and SCC3, respectively. However, conducting multiple statistical tests
593 simultaneously often comes with the problem of multiple comparisons (Miller, 1981). To

counteract this problem, we employ the Bonferroni correction (Haynes, 2013), which is the most conservative multiple testing correction method, to give the strictest significance level:

$$\alpha' = \alpha / M \tag{A8}$$

where M is the number of significance tests and is 4 in this study. α' is therefore calculated as 0.0125.

Results of the significance testing on CRPS (twCRPS) skill scores are shown in Table A1. The twCRPS skill score comparison between SCC3-RC and SCC1 has a p-value smaller than 0.0125. In this case, we can reject the null hypothesis and conclude that the twCRPS skill score difference between these two models is statistically significant, even with the strictest significance level. For the rest cases, we cannot reject the null hypothesis so there is not enough evidence to suggest that the CRPS (twCRPS) skill score difference is significant.

[Table A1]

Reference

Atger, F., 2001. Verification of intense precipitation forecasts from single models and ensemble prediction systems. Nonlinear Processes in Geophysics. 8(6), 401-417. https://doi.org/10.5194/npg-8-401-2001.

Baran, S. and Lerch, S., 2015. Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting. Quarterly Journal of the Royal Meteorological Society. 141(691), 2289-2299. https://doi.org/10.1002/qj.2521.

Bennett, J. C., Wang, Q. J., Pokhrel, P. and Robertson, D. E., 2014. The challenge of forecasting high streamflows 1-3 months in advance with lagged climate indices in southeast Australia. Natural Hazards and Earth System Sciences. 14(2), 219-233. https://doi.org/10.5194/nhess-14-219-2014.

Brocker, J. and Smith, L. A., 2008. From ensemble forecasts to predictive distribution functions. Tellus Series a-Dynamic Meteorology and Oceanography. 60(4), 663-678. https://doi.org/10.1111/j.1600-0870.2008.00333.x.

Buizza, R., 2018. Chapter 2 - Ensemble Forecasting and the Need for Calibration, in: Vannitsem, S., Wilks, D. S. and Messner, J. W. (Eds.), Statistical Postprocessing of Ensemble Forecasts. Elsevier, pp. 15-48. https://doi.org/10.1016/B978-0-12-812372-0.00002-9.

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y. and Wei, M., 2005. A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. Monthly Weather Review. 133(5), 1076-1097. https://doi.org/10.1175/MWR2905.1.

Buizza, R., Milleer, M. and Palmer, T. N., 1999. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. Quarterly Journal of the Royal Meteorological Society. 125(560), 2887-2908. https://doi.org/10.1002/qj.49712556006.

Cattoën, C., Robertson, D. E., Bennett, J. C., Wang, Q. J. and Carey-Smith, T. K., 2020. Calibrating Hourly Precipitation Forecasts with Daily Observations. Journal of Hydrometeorology. 21(7), 1655-1673. https://doi.org/10.1175/jhm-d-19-0246.1.

Clark, A. J., 2017. Generation of Ensemble Mean Precipitation Forecasts from Convection-Allowing Ensembles. Weather and Forecasting. 32(4), 1569-1583. https://doi.org/10.1175/WAF-D-16-0199.1.

Cohen, A. C., 1961. Tables for Maximum Likelihood Estimates: Singly Truncated and Singly Censored Samples. Technometrics. 3(4), 535-541. https://doi.org/10.2307/1266559.

Diks, C., Panchenko, V. and van Dijk, D., 2011. Likelihood-based scoring rules for comparing density forecasts in tails. Journal of Econometrics. 163(2), 215-230. https://doi.org/10.1016/j.jeconom.2011.04.001.

Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H. and Shrestha, D. L., 2015. Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments. Hydrology and Earth System Sciences. 19(7), 3181-3201. https://doi.org/10.5194/hess-19-3181-2015.

Ebert, E. E., 2001. Ability of a Poor Man's Ensemble to Predict the Probability and Distribution of Precipitation. Monthly Weather Review. 129(10), 2461-2480. https://doi.org/10.1175/1520-0493(2001)129<2461:Aoapms>2.0.Co;2.

Efron, B. and Tibshirani, R. J., 1994. An Introduction to the Bootstrap, first ed. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9780429246593.

Ehrendorfer, M., 1997. Predicting the uncertainty of numerical weather forecasts: A review. Meteorologische Zeitschrift. 6, 147-183. https://doi.org/10.1127/metz/6/1997/147.

Epstein, E. S., 1969. Stochastic dynamic prediction. Tellus B: Chemical and Physical Meteorology. 21(6), 739-759. https://doi.org/10.3402/tellusa.v21i6.10143.

Fortin, V., Favre, A.-c. and Saïd, M., 2006. Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. Quarterly Journal of the Royal Meteorological Society. 132(617), 1349-1369. https://doi.org/10.1256/qj.05.167.

Gebetsberger, M., Messner, J. W., Mayr, G. J. and Zeileis, A., 2017. Fine-Tuning Nonhomogeneous Regression for Probabilistic Precipitation Forecasts: Unanimous Predictions, Heavy Tails, and Link Functions. Monthly Weather Review. 145(11), 4693-4708. https://doi.org/10.1175/mwr-d-16-0388.1.

Gelfand, A. E., 2000. Gibbs Sampling. Journal of the American Statistical Association. 95(452), 1300-1304. https://doi.org/10.1080/01621459.2000.10474335.

Gneiting, T., Fadoua, B. and Raftery, A. E., 2007. Probabilistic Forecasts, Calibration and Sharpness. Journal of the Royal Statistical Society. Series B (Statistical Methodology). 69(2), 243-268. https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Gneiting, T., Raftery, A. E., III, A. H. W. and Goldman, T., 2005. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. Monthly Weather Review. 133(5), 1098-1118. https://doi.org/10.1175/MWR2904.1.

675 Gneiting, T. and Ranjan, R., 2011. Comparing Density Forecasts Using Threshold- and Quantile-
676 Weighted Scoring Rules. Journal of Business & Economic Statistics. 29(3), 411-422.
677 https://doi.org/10.1198/jbes.2010.08110.

678 Grimit, E. P. and Mass, C. F., 2007. Measuring the ensemble spread-error relationship with a
679 probabilistic approach: Stochastic ensemble results. Monthly Weather Review. 135(1), 203-221.
680 https://doi.org/10.1175/mwr3262.1.

681 Harper, K., Uccellini, L. W., Kalnay, E., Carey, K. and Morone, L., 2007. 50th Anniversary of
682 Operational Numerical Weather Prediction. Bulletin of the American Meteorological Society.
683 88(5), 639-650. https://doi.org/10.1175/bams-88-5-639.

684 Haynes, W., 2013. Bonferroni Correction, in: Dubitzky, W., Wolkenhauer, O., Cho, K.-H. and
685 Yokota, H. (Eds.), Encyclopedia of Systems Biology. Springer New York, New York, pp. 154-
686 154. https://doi.org/10.1007/978-1-4419-9863-7_1213.

687 Hersbach, H., 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble
688 Prediction Systems. Weather and Forecasting. 15(5), 559-570. https://doi.org/10.1175/1520-
689 0434(2000)015<0559:DOTCRP>2.0.CO;2.

690 Hopson, T. M., 2014. Assessing the Ensemble Spread–Error Relationship. Monthly Weather
691 Review. 142(3), 1125-1142. https://doi.org/10.1175/mwr-d-12-00111.1.

692 Hornung, R. W. and Reed, L. D., 1990. Estimation of Average Concentration in the Presence of
693 Nondetectable Values. Applied Occupational and Environmental Hygiene. 5(1), 46-51.
694 https://doi.org/10.1080/1047322X.1990.10389587.

695 Johnson, C. and Bowler, N., 2009. On the Reliability and Calibration of Ensemble Forecasts.
696 Monthly Weather Review. 137(5), 1717-1720. https://doi.org/10.1175/2009mwr2715.1.

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

Krzysztofowicz, R. and Evans, W. B., 2008. Probabilistic Forecasts from the National Digital Forecast Database. Weather and Forecasting. 23(2), 270-289. https://doi.org/10.1175/2007waf2007029.1.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. and Gneiting, T., 2017. Forecaster's Dilemma: Extreme Events and Forecast Evaluation. Statistical Science. 32(1), 106-127. https://doi.org/10.1214/16-STS588.

Li, W., Duan, Q., Ye, A. and Miao, C., 2019. An improved meta-Gaussian distribution model for post-processing of precipitation forecasts by censored maximum likelihood estimation. Journal of Hydrology. 574, 801-810. https://doi.org/10.1016/j.jhydrol.2019.04.073.

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

Li, W., Wang, Q. J. and Duan, Q., 2020a. A Variable-Correlation Model to Characterize Asymmetric Dependence for Postprocessing Short-Term Precipitation Forecasts. Monthly Weather Review. 148(1), 241-257. https://doi.org/10.1175/MWR-D-19-0258.1.

Li, Y., Wang, Q. J., He, H., Wu, Z. and Lu, G., 2020b. A method to extend temporal coverage of high quality precipitation datasets by calibrating reanalysis estimates. Journal of Hydrology. 581, 124355. https://doi.org/10.1016/j.jhydrol.2019.124355.

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

721 Lorenz, E. N., 1963. Deterministic Nonperiodic Flow. Journal of Atmospheric Sciences. 20(2),
722 130-141. https://doi.org/10.1175/1520-0469(1963)020<0130:Dnf>2.0.Co;2.

723 Messner, J. W., Mayr, G. J., Wilks, D. S. and Zeileis, A., 2014a. Extending Extended Logistic
724 Regression: Extended versus Separate versus Ordered versus Censored. Monthly Weather Review.
725 142(8), 3003-3014. https://doi.org/10.1175/mwr-d-13-00355.1.

726 Messner, J. W., Mayr, G. J., Zeileis, A. and Wilks, D. S., 2014b. Heteroscedastic Extended Logistic
727 Regression for Postprocessing of Ensemble Guidance. Monthly Weather Review. 142(1), 448-456.
728 https://doi.org/10.1175/mwr-d-13-00271.1.

729 Miller, R. G., 1981. Simultaneous statistical inference, second ed. Springer-Verlag, New York.
730 http://dx.doi.org/10.1007/978-1-4613-8122-8.

731 Naughton, M., 2016. ACCESS Numerical Weather Prediction resources for the national research
732 community, OzEWEX 3rd National Workshop. Canberra, 14-15 December 2016.

733 Nelder, J. A. and Mead, R., 1965. A Simplex Method for Function Minimization. The Computer
734 Journal. 7(4), 308-313. https://doi.org/10.1093/comjnl/7.4.308.

735 Palmer, T. N., Roberto, B., Martin, L., Renate, H., Jung, T., Mark, R., Frédéric, V., Berner, J.,
736 Hágel, E., Lawrence, A. R., Florian, P., Park, Y. Y., Bremen, L. v. and Gilmour, I., 2007. The
737 Ensemble Prediction System - Recent and Ongoing Developments. ECMWF Technical
738 Memorandum 540. ECMWF: Reading, UK. Available at: https://www.ecmwf.int/node/12527.

739 Perkins, J. L., Cutter, G. N. and Cleveland, M. S., 1990. Estimating the Mean, Variance, and
740 Confidence Limits from Censored (<Limit of Detection), Lognormally-Distributed Exposure Data.
741 American Industrial Hygiene Association Journal. 51(8), 416-419.
742 https://doi.org/10.1080/15298669091369871.

743 Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M., 2005. Using Bayesian Model
744 Averaging to Calibrate Forecast Ensembles. Monthly Weather Review. 133(5), 1155-1174.
745 https://doi.org/10.1175/mwr2906.1.

746 Richardson, D. S., 2000. Skill and relative economic value of the ECMWF ensemble prediction
747 system. Quarterly Journal of the Royal Meteorological Society. 126(563), 649-667.
748 https://doi.org/10.1002/qj.49712656313.

749 Robertson, D. E., Shrestha, D. L. and Wang, Q. J., 2013. Post-processing rainfall forecasts from
750 numerical weather prediction models for short-term streamflow forecasting. Hydrology and Earth
751 System Sciences. 17(9), 3587-3603. https://doi.org/10.5194/hess-17-3587-2013.

752 Rodwell, M. J., 2006. Comparing and combining deterministic and ensemble forecasts: How to
753 predict rainfall occurrence better. ECMWF Newsletter. 106, 17-23.
754 https://doi.org/10.21957/cd347812th.

755 Saminathan, S., Medina, H., Mitra, S. and Tian, D., 2021. Improving short to medium range GEFS
756 precipitation forecast in India. Journal of Hydrology. 126431.
757 https://doi.org/10.1016/j.jhydrol.2021.126431.

758 Schaake, J., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X. and
759 Seo, D. J., 2007. Precipitation and temperature ensemble forecasts from single-value forecasts.

760 Hydrology and Earth System Sciences. 2007(4), 655-717. https://doi.org/10.5194/hessd-4-655-
761 2007.

762 Schefzik, R., 2017. Ensemble calibration with preserved correlations: unifying and comparing
763 ensemble copula coupling and member-by-member postprocessing. Quarterly Journal of the Royal
764 Meteorological Society. 143(703), 999-1008. https://doi.org/10.1002/qj.2984.

765 Scherrer, S. C., Appenzeller, C., Eckert, P. and Cattani, D., 2004. Analysis of the Spread–Skill
766 Relations Using the ECMWF Ensemble Prediction System over Europe. Weather and Forecasting.
767 19(3), 552-565. https://doi.org/10.1175/1520-0434(2004)019<0552:AOTSRU>2.0.CO;2.

768 Scheuerer, M., 2014. Probabilistic quantitative precipitation forecasting using Ensemble Model
769 Output Statistics. Quarterly Journal of the Royal Meteorological Society. 140(680), 1086-1096.
770 https://doi.org/10.1002/qj.2183.

771 Scheuerer, M. and Hamill, T. M., 2015. Statistical Postprocessing of Ensemble Precipitation
772 Forecasts by Fitting Censored, Shifted Gamma Distributions. Monthly Weather Review. 143(11),
773 4578-4596. https://doi.org/10.1175/mwr-d-15-0061.1.

774 Schmeits, M. J. and Kok, K. J., 2010. A Comparison between Raw Ensemble Output, (Modified)
775 Bayesian Model Averaging, and Extended Logistic Regression Using ECMWF Ensemble
776 Precipitation Reforecasts. Monthly Weather Review. 138(11), 4199-4211.
777 https://doi.org/10.1175/2010mwr3285.1.

778 Shrestha, D. L., Robertson, D. E., Bennett, J. C. and Wang, Q. J., 2015. Improving Precipitation
779 Forecasts by Generating Ensembles through Postprocessing. Monthly Weather Review. 143(9),
780 3642-3663. https://doi.org/10.1175/MWR-D-14-00329.1.

781 Sloughter, J. M. L., Raftery, A. E., Gneiting, T. and Fraley, C., 2007. Probabilistic Quantitative
782 Precipitation Forecasting Using Bayesian Model Averaging. Monthly Weather Review. 135(9),
783 3209-3220. https://doi.org/10.1175/mwr3441.1.

784 Stauffer, R., Umlauf, N., Messner, J. W., Mayr, G. J. and Zeileis, A., 2017. Ensemble
785 Postprocessing of Daily Precipitation Sums over Complex Terrain Using Censored High-
786 Resolution Standardized Anomalies. Monthly Weather Review. 145(3), 955-969.
787 https://doi.org/10.1175/mwr-d-16-0260.1.

788 Surcel, M., Zawadzki, I. and Yau, M. K., 2014. On the Filtering Properties of Ensemble Averaging
789 for Storm-Scale Precipitation Forecasts. Monthly Weather Review. 142(3), 1093-1105.
790 https://doi.org/10.1175/MWR-D-13-00134.1.

791 Taillardat, M., Fougères, A.-L., Naveau, P. and Mestre, O., 2019. Forest-Based and
792 Semiparametric Methods for the Postprocessing of Rainfall Ensemble Forecasting. Weather and
793 Forecasting. 34(3), 617-634. https://doi.org/10.1175/waf-d-18-0149.1.

794 Toth, Z., Zhu, Y. and Marchok, T., 2001. The Use of Ensembles to Identify Forecasts with Small
795 and Large Uncertainty. Weather and Forecasting. 16(4), 463-477. https://doi.org/10.1175/1520-
796 0434(2001)016<0463:Tuoeti>2.0.Co;2.

797 Van Schaeybroeck, B. and Vannitsem, S., 2015. Ensemble post-processing using member-by-
798 member approaches: theoretical aspects. Quarterly Journal of the Royal Meteorological Society.
799 141(688), 807-818. https://doi.org/10.1002/qj.2397.

800 Veenhuis, B. A., 2013. Spread Calibration of Ensemble MOS Forecasts. Monthly Weather Review.
801 141(7), 2467-2482. https://doi.org/10.1175/mwr-d-12-00191.1.

Verkade, J. S., Brown, J. D., Reggiani, P. and Weerts, A. H., 2013. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. Journal of Hydrology. 501, 73-91. https://doi.org/10.1016/j.jhydrol.2013.07.039.

Vokoun, M. and Hanel, M., 2018. Comparing ALADIN-CZ and ALADIN-LAEF Precipitation Forecasts for Hydrological Modelling in the Czech Republic. Advances in Meteorology. 2018, 14. https://doi.org/10.1155/2018/5368438.

Wang, Q. J. and Robertson, D. E., 2011. Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. Water Resources Research. 47(2). https://doi.org/10.1029/2010WR009333.

Wang, Q. J., Shao, Y., Song, Y., Schepen, A., Robertson, D. E., Ryu, D. and Pappenberger, F., 2019a. An evaluation of ECMWF SEAS5 seasonal climate forecasts for Australia using a new forecast calibration algorithm. Environmental Modelling & Software. 122, 104550. https://doi.org/10.1016/j.envsoft.2019.104550.

Wang, Q. J., Shrestha, D. L., Robertson, D. E. and Pokhrel, P., 2012. A log-sinh transformation for data normalization and variance stabilization. Water Resources Research. 48(5). https://doi.org/10.1029/2011WR010973.

Wang, Q. J., Zhao, T., Yang, Q. and Robertson, D., 2019b. A Seasonally Coherent Calibration (SCC) Model for Postprocessing Numerical Weather Predictions. Monthly Weather Review. 147(10), 3633-3647. https://doi.org/10.1175/mwr-d-19-0108.1.

Wilks, D. S., 2011. Statistical methods in the atmospheric sciences, third ed. Academic Press, Oxford.

824 Wu, L., Seo, D.-J., Demargne, J., Brown, J. D., Cong, S. and Schaake, J., 2011. Generation of
825 ensemble precipitation forecast from single-valued quantitative precipitation forecast for
826 hydrologic ensemble prediction. Journal of Hydrology. 399(3), 281-298.
827 https://doi.org/10.1016/j.jhydrol.2011.01.013.

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

831 Yang, Q., Wang, Q. J. and Hakala, K., 2021a. Achieving effective calibration of precipitation
832 forecasts over a continental scale. Journal of Hydrology: Regional Studies. 35, 100818.
833 https://doi.org/10.1016/j.ejrh.2021.100818.

834 Yang, Q., Wang, Q. J., Hakala, K. and Tang, Y., 2021b. Bias-correcting input variables enhances
835 forecasting of reference crop evapotranspiration. Hydrology and Earth System Sciences. 25(9),
836 4773-4788. https://doi.org/10.5194/hess-25-4773-2021.

837 Zhao, P., Wang, Q. J., Wu, W. and Yang, Q., 2020. Which precipitation forecasts to use?
838 Deterministic versus coarser-resolution ensemble NWP models. Quarterly Journal of the Royal
839 Meteorological Society. 147(735), 900-913. https://doi.org/10.1002/qj.3952.

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

843 Zhao, T., Wang, Q. J., Bennett, J. C., Robertson, D. E., Shao, Q. and Zhao, J., 2015. Quantifying
844 predictive uncertainty of streamflow forecasts based on a Bayesian joint probability model. Journal
845 of Hydrology. 528, 329-340. https://doi.org/10.1016/j.jhydrol.2015.06.043.

846

Tables

Table 1. Differences in CRPS skill score between SCC2 and SCC1, SCC3 and SCC1, SCC3-RC and SCC1, and SCC3-RC and SCC3. The CRPS skill score values for all 180 cases (20 sites and 9 lead times) in the 3-year period are pooled together to construct the table.

	SCC2 minus SCC1	SCC3 minus SCC1	SCC3-RC minus SCC1	SCC3-RC minus SCC3
Number of improved cases	106	117	143	129
Average CRPS skill score difference	0.83%	1.13%	1.68%	0.55%

Table 2. Differences in twCRPS skill score between SCC2 and SCC1, SCC3 and SCC1, SCC3-RC and SCC1, and SCC3-RC and SCC3. The twCRPS skill score values for all 180 cases (20 sites and 9 lead times) in the 3-year period are pooled together to construct the table.

	SCC2 minus SCC1	SCC3 minus SCC1	SCC3-RC minus SCC1	SCC3-RC minus SCC3
Number of improved cases	110	119	133	113
Average twCRPS skill score difference	1.72%	2.12%	3.04%	0.92%

Table A1. Significance testing results of the CRPS (twCRPS) skill score values of calibrated forecasts from SCC2 and SCC1, SCC3 and SCC1, SCC3-RC and SCC1, and SCC3-RC and SCC3. The CRPS (twCRPS) skill score values from all of 20 sites and 9 lead times in the 3-year period are pooled together to conduct the significance testing.

	p-value of the statistical significance testing			
	SCC2 vs. SCC1	SCC3 vs. SCC1	SCC3-RC vs. SCC1	SCC3-RC vs. SCC3
CRPS skill score	0.2442	0.1711	0.0803	0.3192
twCRPS skill score	0.0486	0.0208	0.0018	0.1661

Figures

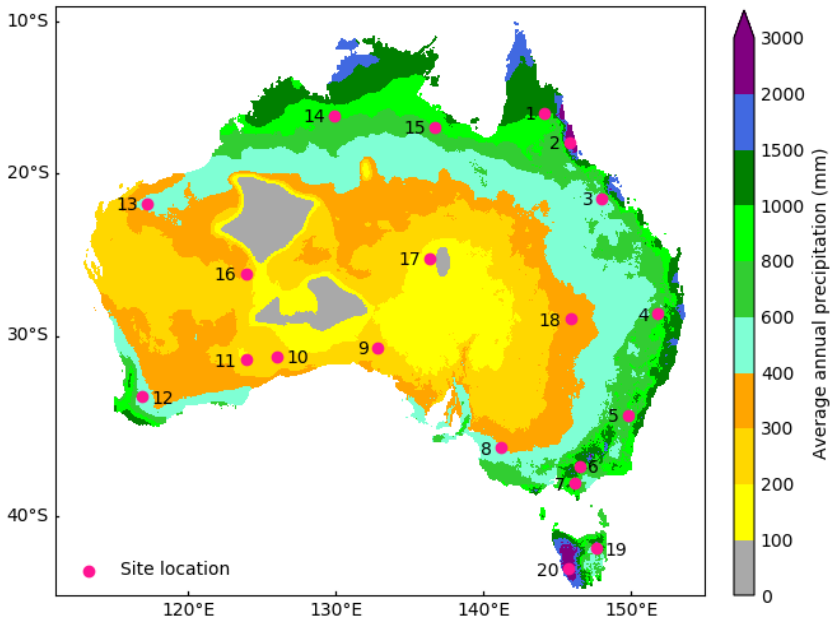
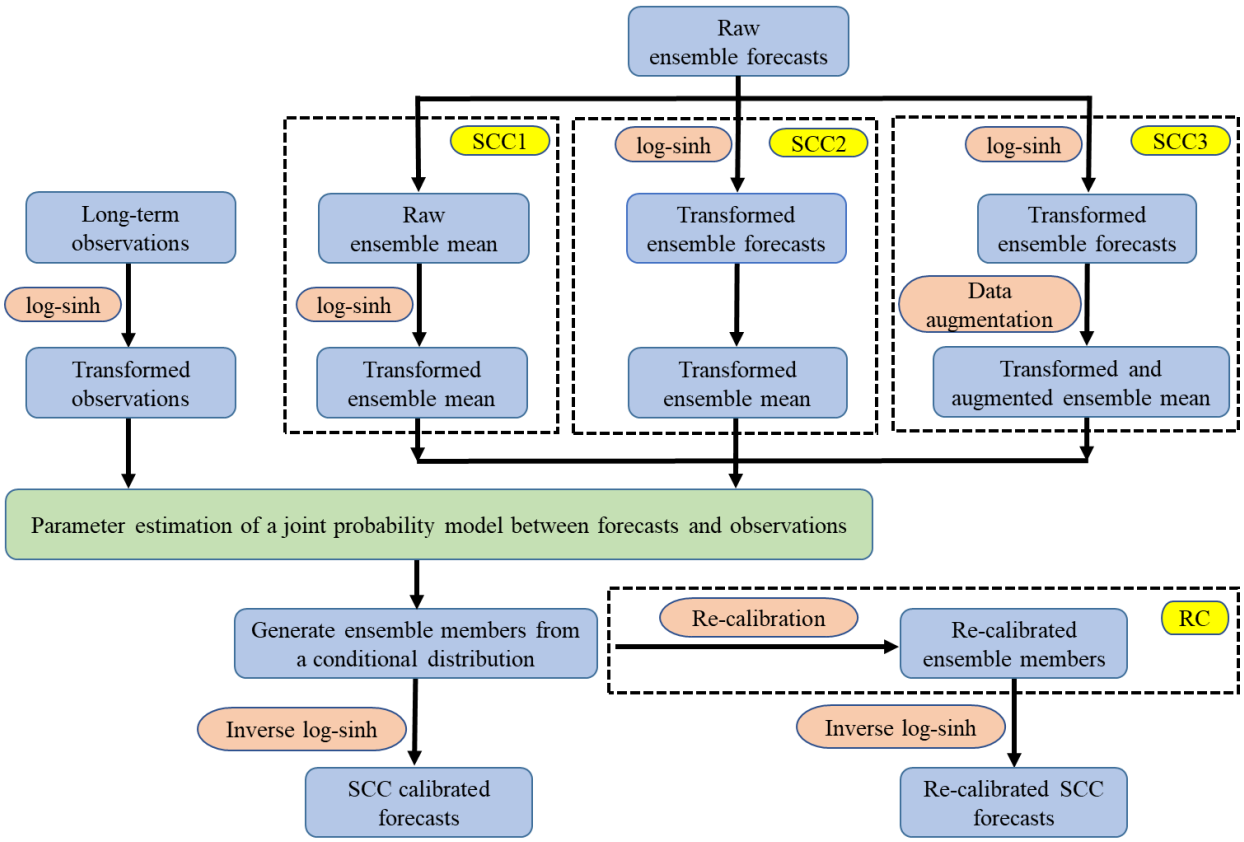


Figure 1. Locations and average annual precipitation map for the selected 20 sites.

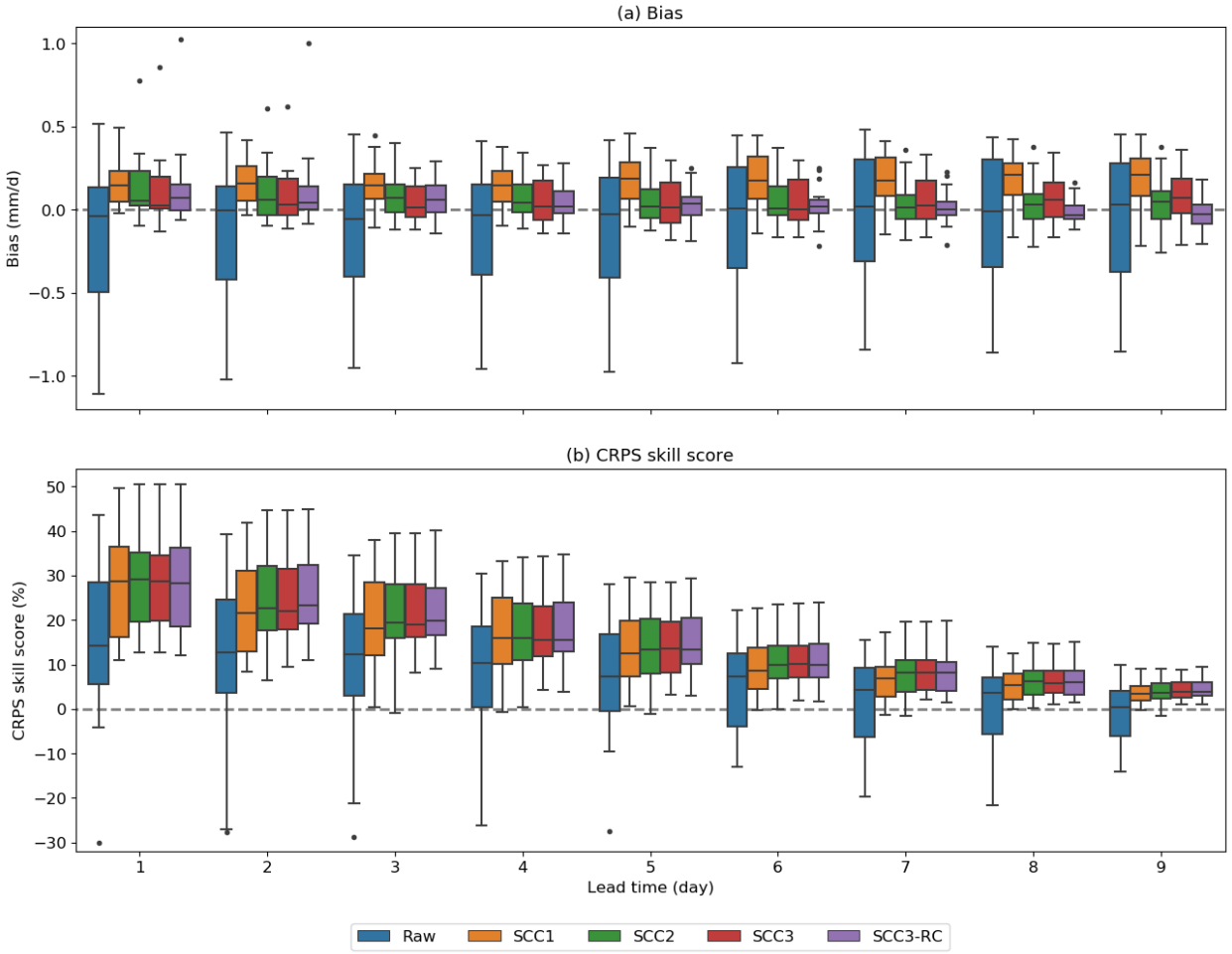


866

867 Figure 2. The modelling process of the two-step calibration approach. Log-sinh is the transformation algorithm we
868 apply for normalizing precipitation data. SCC1, SCC2, and SCC3 represent the SCC model with three different
869 methods for estimating ensemble mean values in the first step, and RC represents the use of the ensemble spread
870 information in the second step.

871

872



873

874 Figure 3. Overall forecast evaluation over 20 sites of raw ensemble forecasts, and calibrated forecasts from SCC1,
875 SCC2, SCC3, and SCC3-RC models in the 3-year period. For each boxplot, lines on the box portion from bottom to
876 top represent first quartile (Q1, 25th percentile), median (Q2, 50th percentile), and third quartile (Q3, 75th percentile)
877 of the data, respectively; lines on the whisker portion from bottom to top represent "minimum" (Q1 – 1.5 * (Q3 – Q1))
878 and "maximum" (Q3 + 1.5 * (Q3 – Q1)) of the data, respectively; black points outside the whisker are shown as
879 outliers of the data.

880

881

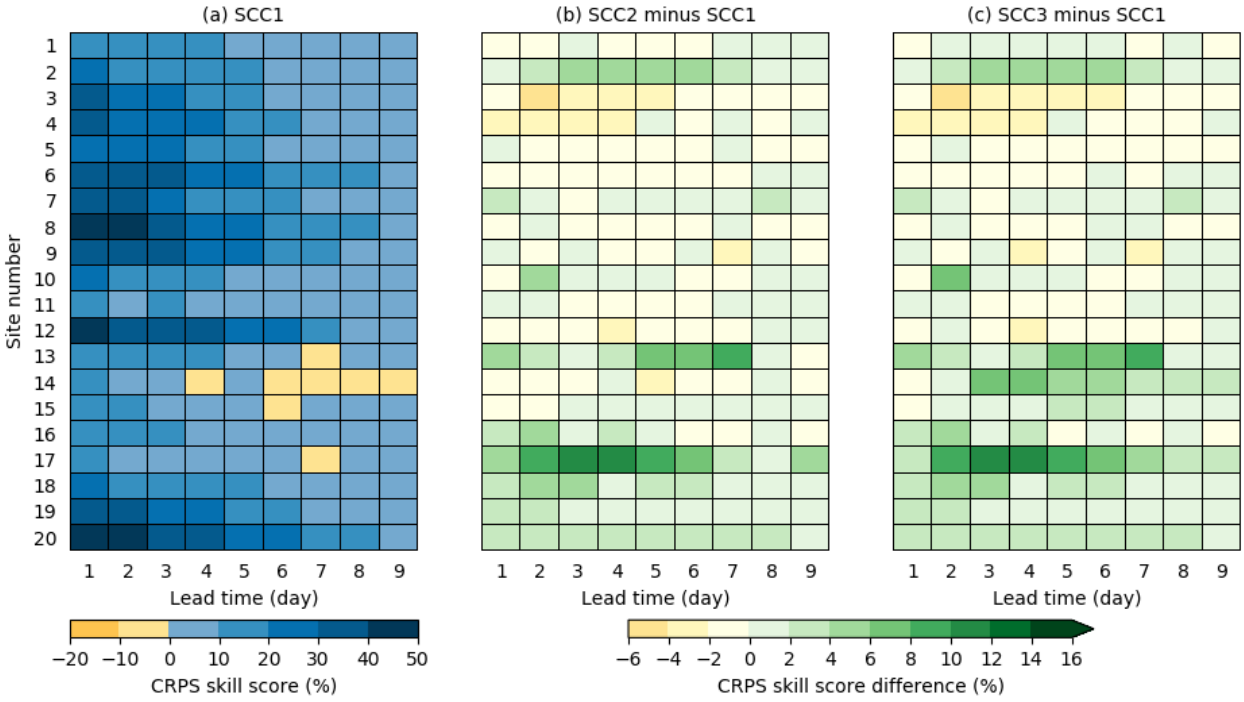


Figure 4. CRPS skill score verification of calibrated forecasts from SCC1, SCC2, and SCC3 models in the 3-year period. (a) CRPS skill score of SCC1; (b) CRPS skill score difference between SCC2 and SCC1; (c) CRPS skill score difference between SCC3 and SCC1. A positive (negative) CRPS skill score indicates that calibrated forecasts are better (poorer) than the referenced climatology forecasts. The arrow of the right color bar indicates that values can be above the displayed maximum CRPS skill score difference.

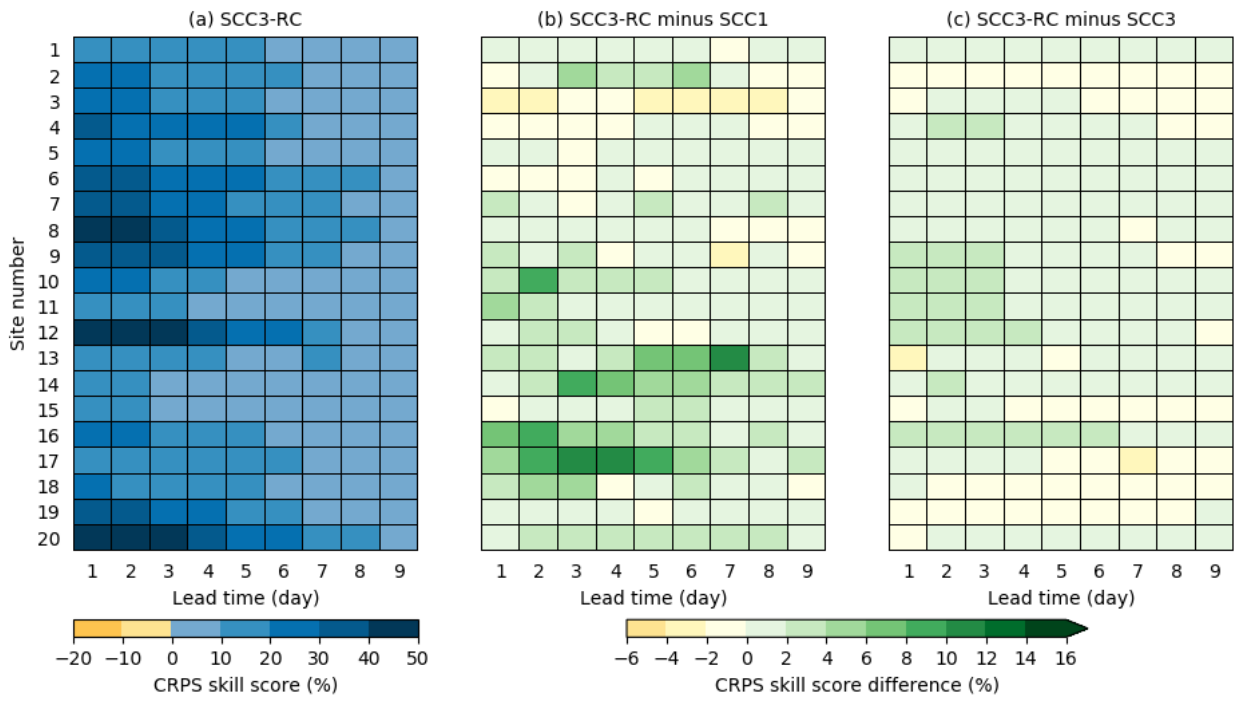


Figure 5. CRPS skill score verification of SCC3-RC calibrated forecasts in the 3-year period. (a) CRPS skill score of SCC3-RC; (b) CRPS skill score difference between SCC3-RC and SCC1; (c) CRPS skill score difference between SCC3-RC and SCC3. A positive (negative) CRPS skill score indicates that calibrated forecasts are better (poorer) than the referenced climatology forecasts. The arrow of the right color bar indicates that values can be above the displayed maximum CRPS skill score difference.

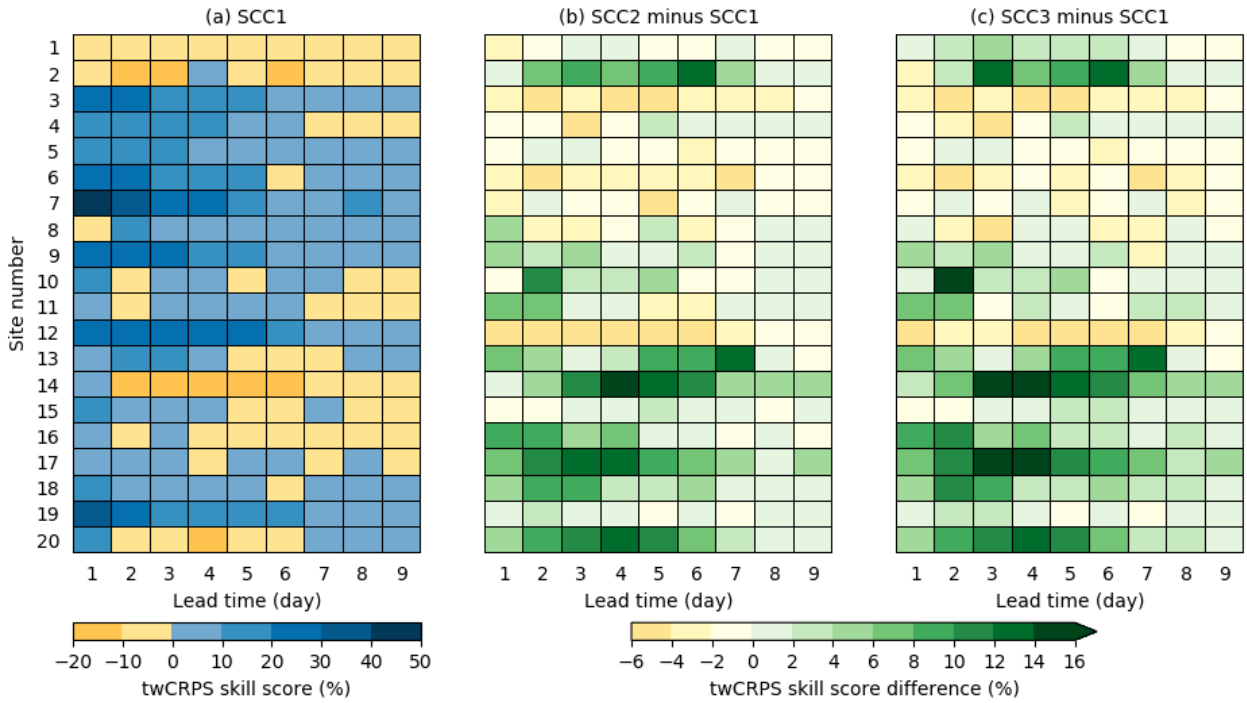
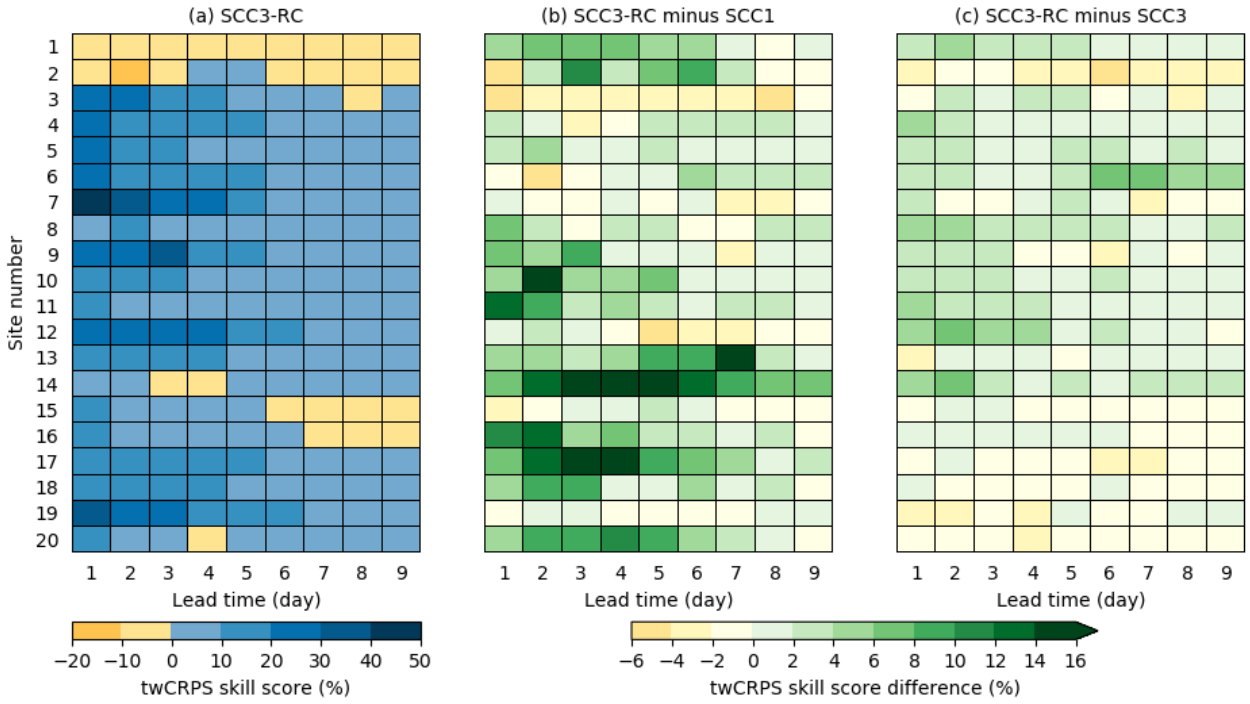


Figure 6. twCRPS skill score verification of calibrated forecasts from SCC1, SCC2, and SCC3 models in the 3-year period. (a) twCRPS skill score of SCC1; (b) twCRPS skill score difference between SCC2 and SCC1; (c) twCRPS skill score difference between SCC3 and SCC1. The threshold for calculating twCRPS is the 95% quantile of observed precipitation values in each case. A positive (negative) twCRPS skill score indicates that calibrated forecasts are better (poorer) than the referenced climatology forecasts. The arrow of the right color bar indicates that values can be above the displayed maximum twCRPS skill score difference.



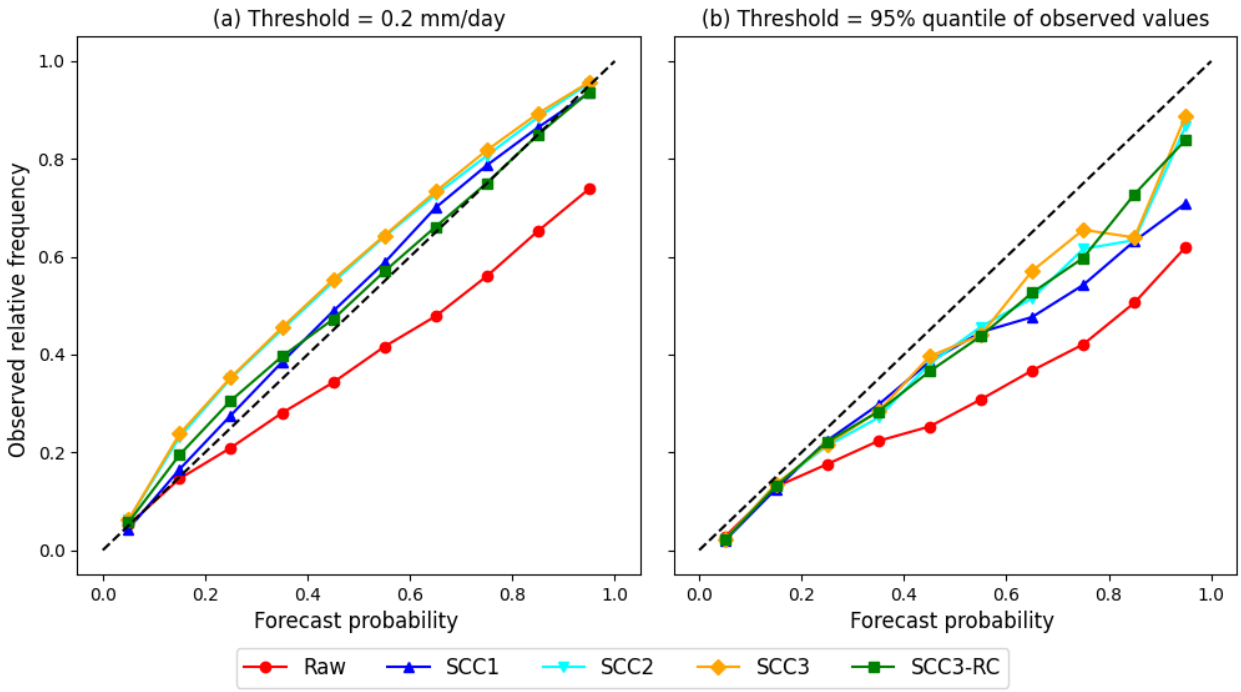
906

907 Figure 7. twCRPS skill score verification of SCC3-RC calibrated forecasts in the 3-year period. (a) twCRPS skill
908 score of SCC3-RC; (b) twCRPS skill score difference between SCC3-RC and SCC1; (c) twCRPS skill score difference
909 between SCC3-RC and SCC3. The threshold for calculating twCRPS is the 95% quantile of observed precipitation
910 values in each case. A positive (negative) twCRPS skill score indicates that calibrated forecasts are better (poorer)
911 than the referenced climatology forecasts. The arrow of the right color bar indicates that values can be above the
912 displayed maximum twCRPS skill score difference.

913

914

915



916

917 Figure 8. Reliability diagrams for raw ensemble forecasts, and calibrated forecasts from SCC1, SCC2, SCC3, and
918 SCC3-RC models with exceedance probabilities considered at the threshold of (a) 0.2 mm/day and (b) 95% quantile
919 of observed values. Results from all 180 cases (20 sites and 9 lead times) in the 3-year period are pooled together to
920 construct the diagrams.