

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

4
5
6 Pengcheng Zhao*, Quan J. Wang, Wenyan Wu, Qichun Yang

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

9
10 *: Corresponding author

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

12 ABSTRACT

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

31

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

34

35 **1. Introduction**

36 Short-term precipitation forecasting is routinely performed using numerical weather prediction
37 (NWP) models in meteorological centers around the world (Harper *et al.*, 2007). These models are
38 designed to simulate dynamic physics of the atmosphere forward in time to predict future weather
39 conditions. Traditional deterministic NWP models produce single-value forecasts, hence cannot
40 capture the chaotic nature of the atmosphere (Lorenz, 1963; Epstein, 1969). To account for the
41 uncertainty of weather forecasts, ensemble prediction systems have been developed to generate
42 multiple ensemble members from varying initial conditions or parametrizations of NWP models
43 (Ehrendorfer, 1997; Buizza *et al.*, 1999; Palmer *et al.*, 2007). The superiority of ensemble forecasts
44 compared to deterministic forecasts has been widely demonstrated in the literature (Richardson,
45 2000; Atger, 2001; Rodwell, 2006; Vokoun and Hanel, 2018; Zhao *et al.*, 2020).

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

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

58 Two post-processing schemes are commonly used to produce calibrated probability distributions.
59 The first scheme is based on distributional regressions, such as ensemble model output statistics
60 (EMOS) (Gneiting *et al.*, 2005; Scheuerer, 2014; Baran and Lerch, 2015), ensemble dressing
61 (Fortin *et al.*, 2006), Bayesian model averaging (BMA) (Raftery *et al.*, 2005; Sloughter *et al.*, 2007;
62 Schmeits and Kok, 2010), and machine learning (ML) based regression methods (Dogulu *et al.*,
63 2015). This scheme typically makes use of statistics of raw ensemble forecasts, usually ensemble
64 mean and ensemble spread, to establish probability distributions.

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

73 A couple of advantages can be stated about joint probability models. First, they have been shown
74 to be superior to distributional regression models when only ensemble mean is used as a predictor
75 (Li *et al.*, 2019a). One possible reason is that regression relationship in distributional regressions
76 becomes weak when raw precipitation forecasts approach zero (Gebetsberger *et al.*, 2017). Second,
77 because joint probability models take only the ensemble mean as the model predictor, they allow
78 more flexibility in mathematical treatment to incorporate other complexities. For example, the
79 seasonally coherent calibration (SCC) model (Wang *et al.*, 2019b; Zhao *et al.*, 2020; Yang *et al.*,
80 2021a) based on the joint probability scheme has been developed to produce calibrated forecasts

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

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

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

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

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

106

107 **2. Data**

108 Observed daily precipitation data are obtained from the Australian Water Availability Project's
109 climate datasets (AWAP) (Jones *et al.*, 2009). With a high spatial resolution ($0.05^\circ \times 0.05^\circ$),
110 gridded AWAP data are obtained by interpolating rain gauge observations, which are collected
111 from 0900 h of the previous day to 0900 h of the current day according to Australian local time,
112 including daylight saving. Observations for a period of 30 years from 1 August 1989 to 31 July
113 2019 are used as reference data for SCC model establishments and for forecast evaluations.

114 We select ensemble precipitation forecasts from the Australian Community Climate and Earth-
115 System Simulator Global Ensemble 2 (ACCESS-GE2) NWP model for forecast calibration.
116 ACCESS-GE2 is based on the Met Office Global and Regional Ensemble Prediction System
117 (MOGREPS) (Naughton, 2016) and has a horizontal grid spacing of approximately 60 km. Each
118 ACCESS-GE2 ensemble consists of 24 exchangeable members that result from multiple stochastic
119 disturbances to both model physics and initial conditions of the ACCESS-GE2 model. Ensemble
120 forecasts are issued at 1200 UTC on a daily basis for lead times up to 10 days and are available at
121 a 3-hourly temporal resolution. ACCESS-GE2 forecasts of a 3-year period from 1 August 2016 to
122 31 July 2019 are selected for this study.

123 To match the forecasts with observed data, we modify the spatial and temporal resolutions of
124 ACCESS-GE2 data according to the AWAP data. We apply bilinear interpolation to re-grid the
125 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]

146 3.1 SCC model formulation

147 3.1.1 Log-sinh transformations and a joint probability model

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

$$157 \begin{bmatrix} f(t) \\ o(t) \end{bmatrix} \sim N \left(\begin{bmatrix} u_f[m(t)] \\ u_o[m(t)] \end{bmatrix}, \begin{bmatrix} \sigma_f[m(t)]^2 & \rho[m(t)]\sigma_f[m(t)]\sigma_o[m(t)] \\ \rho[m(t)]\sigma_f[m(t)]\sigma_o[m(t)] & \sigma_o[m(t)]^2 \end{bmatrix} \right) \quad (1)$$

158 where $m(t)$ denotes the month for day t , namely $m(t) \in \{1, 2, \dots, 12\}$; $u_f[m(t)]$ and $\sigma_f[m(t)]$
159 are the mean and standard deviation of the marginal distribution of $f(t)$, and $u_o[m(t)]$ and
160 $\sigma_o[m(t)]$ for $o(t)$; and $\rho[m(t)]$ is the correlation between $f(t)$ and $o(t)$. Because the bivariate
161 normal distribution applies to continuous variables, a censoring technique is used to treat (near)
162 zero precipitation as “censored” data. The term “censored” means these data are treated as below
163 certain thresholds but are not precisely specified (Wang and Robertson, 2011; Messner *et al.*,
164 2014a; Scheuerer and Hamill, 2015; Zhao *et al.*, 2015; Stauffer *et al.*, 2017). Two thresholds x_c
165 and y_c are applied here for $x(t)$ and $y(t)$ with corresponding thresholds f_c and o_c for better fitting
166 $f(t)$ and $o(t)$ in their corresponding marginal distributions (Wang and Robertson, 2011). Here x_c

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

169 3.1.2 Reparameterization of the joint probability model

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

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

$$182 \quad u_f[m(t)] = a + bu_o[m(t)] \quad (2)$$

$$183 \quad \sigma_f[m(t)] = c + d\sigma_o[m(t)] \quad (3)$$

$$184 \quad \rho[m(t)] = r \quad (4)$$

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

189 observed climatology. Equation (4) assumes that the underlying forecast skill of NWP models is
 190 constant for all 12 months. These assumptions are supported by a case study from Wang *et al.*
 191 (2019b) and may help reduce sampling effects of forecast evaluation results. The
 192 reparameterization of the joint probability model replaces 36 parameters ($u_f[m(t)]$, $\sigma_f[m(t)]$, and
 193 $\rho[m(t)]$, $m(t) \in \{1, 2, \dots, 12\}$) with five parameters (a , b , c , d , and r), therefore making it
 194 feasible to estimate model parameters from a short period of archived NWP forecasts. All of the
 195 parameters from the above two procedures can be derived using the method of maximum
 196 likelihood and the Nelder-Mead searching approach (Nelder and Mead, 1965). Related likelihood
 197 functions are given in Wang *et al.* (2019b).

198 3.1.3 Calibration of new forecasts

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

$$201 \quad [o(t)|f(t)] \sim \mathbf{N}\{o(t)|\tilde{u}_o(t), \tilde{\sigma}_o^2(t)\} \quad (5)$$

202 where

$$203 \quad \tilde{u}_o(t) = u_o[m(t)] + \rho[m(t)] \frac{\sigma_o[m(t)]}{\sigma_f[m(t)]} \{f(t) - u_f[m(t)]\} \quad (6)$$

$$204 \quad \tilde{\sigma}_o^2(t) = \{1 - \rho^2[m(t)]\} \sigma_o^2[m(t)] \quad (7)$$

205 We sample an ensemble of values from the conditional distribution to represent the forecast
 206 probability distribution. When $f(t) > f_c$, an ensemble $o(t, n)$ ($n = 1, 2, \dots, N$) of any size N (100
 207 for this study) can be sampled from the conditional distribution. When $f(t) \leq f_c$, a random value
 208 $f'(t)$ from the marginal distribution $\mathbf{N}\{f(t)|u_f[m(t)], \sigma_f^2[m(t)]\}$ is first sampled in the range of
 209 $[-\infty, f_c]$; then a sample can be drawn from the conditional distribution $[o(t)|f'(t)]$ to give an
 210 ensemble member. These two steps are repeated N times to form the ensemble $o(t, n)$. Finally, the

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

214 **3.2 The proposed two-step calibration approach**

215 3.2.1 Step One: Calibration of ensemble mean using SCC

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

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

$$226 \quad x_m(t) = \frac{1}{K} \sum_{k=1}^K x(t, k) \quad (8)$$

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

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

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

$$234 \quad f_m(t) = \frac{1}{K} \sum_{k=1}^K f(t, k) \quad (9)$$

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

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

$$244 \quad f(t, k) \sim \mathbf{N}(f_u(t), \sigma^2) \quad (10)$$

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

$$248 \quad f_u(t) \sim \mathbf{N}(u_u, \sigma_u^2) \quad (11)$$

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

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

$$255 \quad L = \prod_{t=1}^T \prod_{k=1}^K \mathbf{N}(f(t, k)|f_u(t), \sigma^2) \times \mathbf{N}(f_u(t)|u_u, \sigma_u^2) \quad (12)$$

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

$$261 \quad f'_m(t) = \frac{1}{K} \sum_{k=1}^K f'(t, k) \quad (13)$$

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

264 3.2.2 Step Two: Re-calibration of SCC calibrated forecasts using ensemble spread

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

270 Generally, ensemble spread is represented using the standard deviation among ensemble members.
 271 Raw ensemble standard deviation $\sigma_{raw}(t)$ and SCC3 calibrated ensemble standard deviation
 272 $\sigma_{cali}(t)$ in transformed space can be calculated as:

$$273 \quad \sigma_{raw}(t) = \sqrt{\frac{1}{K-1} \sum_{k=1}^K [f'(t, k) - f'_m(t)]^2} \quad (14)$$

274
$$\sigma_{cali}(t) = \sqrt{\frac{1}{N-1} \sum_{n=1}^N [o(t, n) - o_m(t)]^2}$$
 (15)

275 where

276
$$o_m(t) = \frac{1}{N} \sum_{n=1}^N o(t, n)$$
 (16)

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

279
$$o_{RC}(t, n) = \lambda_1 \times o_m(t) + (o(t, n) - o_m(t)) \times (\lambda_2 + \lambda_3 \frac{\sigma_{raw}(t)}{\sigma_{cali}(t)})$$
 (17)

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

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

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

295 **3.3 Forecast evaluation**

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

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

$$302 \quad Bias = \frac{1}{T} \sum_{t=1}^T x(t) - \frac{1}{T} \sum_{t=1}^T y(t) \quad (18)$$

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

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

$$309 \quad CRPS = \frac{1}{T} \sum_{t=1}^T \int \{F(t, x) - H(x - y(t))\}^2 dx \quad (19)$$

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

313 We also calculate the CRPS of reference climatology forecasts ($CRPS_{ref}$) using the leave-one-
314 month-out cross-validated climatology ensemble forecasts that are generated from the SCC models.

315 A CRPS skill score can then be calculated as:

316
$$CRPS \text{ skill score} = \frac{CRPS_{ref} - CRPS}{CRPS_{ref}} \times 100(\%) \quad (20)$$

317 The CRPS skill score is positively oriented and represents the relative improvement of the
 318 calibrated forecasts compared to the referenced climatology forecasts. A maximum skill score of
 319 100% indicates that forecasts perfectly match the corresponding observations and a skill score of
 320 0% indicates that forecasts have comparable errors to the reference forecasts. Forecasts poorer
 321 than reference forecasts have negative skill scores. Note that the reference climatology forecasts
 322 are common to all the SCC models, and therefore CRPS skill scores achieved by the different
 323 models can be directly compared with each other.

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

327
$$twCRPS = \frac{1}{T} \sum_{t=1}^T \int \{F(t, x) - H(x - y(t))\}^2 \omega(x) dx \quad (21)$$

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

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

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

343

344 **4. Results**

345 **4.1 Overall forecast evaluation of bias and CRPS skill score**

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

355 [Figure 3]

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

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

362 **4.2 CRPS skill scores of calibrated forecasts from the three SCC variants**

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

373 [Figure 4]

374 [Table 1]

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

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

383 **4.3 CRPS skill scores of calibrated forecasts from the two-step calibration**

384 Results of CRPS skill score for SCC3-RC calibrated forecasts are shown in Figure 5 and Table 1.
385 SCC3-RC calibrated forecasts have positive CRPS skill scores at all sites and lead times (Figure
386 5(a)). This assures that the calibrated forecasts are more skillful than the referenced climatology
387 forecasts in all cases. Compared to SCC1, the accumulated average CRPS skill score improvement
388 of SCC3-RC reaches 1.68%, which is quite prominent considering all the sites and lead times.

389 [Figure 5]

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

400 **4.4 twCRPS skill scores of calibrated forecasts from the three SCC variants**

401 We further evaluate the ability of the calibration models in predicting heavy precipitation events,
402 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)).

424 Together with results presented in sections 4.2 – 4.4, it can be concluded that on the average of the
425 20 sites and 9 lead times, in the order of SCC1, SCC2, SCC3, and SCC3-RC, both CRPS skill
426 score and twCRPS skill score tend to increase successively, indicating gradual improvements.
427 Besides, the improvements on twCRPS skill score have higher magnitudes than those on CRPS
428 skill score (Tables 1 and 2), meaning that the latter three models (especially SCC3-RC) are far
429 more capable of predicting heavy precipitation events than the SCC1 model, on average.

430 [Figure 7]

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

442 **4.6 Reliability diagrams of calibrated forecasts from different models**

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

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

451 [Figure 8]

452

453 **5. Discussion**

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

469 The re-calibration method employed in our study is similar to member-by-member post-processing
470 (MBMP) methods (Johnson and Bowler, 2009; Van Schaeybroeck and Vannitsem, 2015; Schefzik,
471 2017), which post-processes raw ensemble forecasts by adjusting ensemble members individually.
472 These MBMP methods are often implemented by minimizing objective scoring functions (e.g.
473 CRPS) of the adjusted ensemble forecasts. They can make use of raw ensemble spread information
474 but are generally not sophisticated enough to reflect statistical forecast uncertainty derived from
475 historical weather events. In this context, the re-calibration is developed based on MBMP methods
476 to further post-process the SCC calibrated forecasts using raw ensemble spread information.
477 Forecasts calibrated using the two-step calibration can reflect both the dynamically flow-
478 dependent ensemble spread from raw ensemble forecasts and the SCC calibrated ensemble spread
479 that contains statistically generated uncertainty information.

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

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

505 We acknowledge that our study is only conducted based on ensemble forecasts on a daily basis
506 and with a grid spacing of about 60 km. As temporally and spatially high-resolution weather
507 forecasts have attracted great attention recently, whether the developed models in this study are
508 applicable to, for example, hourly NWP forecasts with a finer grid spacing, needs to be further
509 investigated. Besides, establishing calibration models for precipitation is widely known to be more
510 challenging than other weather variables. Indeed, as a highly sophisticated model compared to
511 traditional joint probability models, SCC has also been shown to be applicable to other variables
512 (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

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

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

541

542 **Declaration of Competing Interest**

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

545

546 **Data Availability Statement**

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

550

551 **Acknowledgements**

552 This work is supported by an Australian Research Council Linkage Project (Grant No.
553 LP170100922) and a collaborative project (Grant No. TP707466) between the University of
554 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 access 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 $X(x_1, x_2, \dots, x_n)$ and $Y(y_1, y_2, \dots, y_n)$,
 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
$$x'_i = x_i - X_m + XY_m, \quad (i = 1, 2, \dots, n) \quad (\text{A3})$$

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

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^* :

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

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:

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

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

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

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

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

$$596 \quad \alpha' = \alpha / M \quad (A8)$$

597 where M is the number of significance tests and is 4 in this study. α' is therefore calculated as
598 0.0125.

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

605 [Table A1]

606

607

608

609

610

611

612 **Reference**

- 613 Atger, F., 2001. Verification of intense precipitation forecasts from single models and ensemble
614 prediction systems. *Nonlinear Processes in Geophysics*. 8(6), 401-417.
615 <https://doi.org/10.5194/npg-8-401-2001>.
- 616 Baran, S. and Lerch, S., 2015. Log-normal distribution based Ensemble Model Output Statistics
617 models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological*
618 *Society*. 141(691), 2289-2299. <https://doi.org/10.1002/qj.2521>.
- 619 Bennett, J. C., Wang, Q. J., Pokhrel, P. and Robertson, D. E., 2014. The challenge of forecasting
620 high streamflows 1-3 months in advance with lagged climate indices in southeast Australia.
621 *Natural Hazards and Earth System Sciences*. 14(2), 219-233. [https://doi.org/10.5194/nhess-14-](https://doi.org/10.5194/nhess-14-219-2014)
622 [219-2014](https://doi.org/10.5194/nhess-14-219-2014).
- 623 Brocker, J. and Smith, L. A., 2008. From ensemble forecasts to predictive distribution functions.
624 *Tellus Series a-Dynamic Meteorology and Oceanography*. 60(4), 663-678.
625 <https://doi.org/10.1111/j.1600-0870.2008.00333.x>.
- 626 Buizza, R., 2018. Chapter 2 - Ensemble Forecasting and the Need for Calibration, in: Vannitsem,
627 S., Wilks, D. S. and Messner, J. W. (Eds.), *Statistical Postprocessing of Ensemble Forecasts*.
628 Elsevier, pp. 15-48. <https://doi.org/10.1016/B978-0-12-812372-0.00002-9>.
- 629 Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y. and Wei, M., 2005. A Comparison
630 of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. *Monthly Weather Review*.
631 133(5), 1076-1097. <https://doi.org/10.1175/MWR2905.1>.

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

635 Cattoën, C., Robertson, D. E., Bennett, J. C., Wang, Q. J. and Carey-Smith, T. K., 2020.
636 Calibrating Hourly Precipitation Forecasts with Daily Observations. Journal of Hydrometeorology.
637 21(7), 1655-1673. <https://doi.org/10.1175/jhm-d-19-0246.1>.

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

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

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

646 Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H. and Shrestha, D. L., 2015.
647 Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods
648 and their comparison on contrasting catchments. Hydrology and Earth System Sciences. 19(7),
649 3181-3201. <https://doi.org/10.5194/hess-19-3181-2015>.

650 Ebert, E. E., 2001. Ability of a Poor Man's Ensemble to Predict the Probability and Distribution of
651 Precipitation. Monthly Weather Review. 129(10), 2461-2480. [https://doi.org/10.1175/1520-](https://doi.org/10.1175/1520-0493(2001)129<2461: Aoapms>2.0.Co;2)
652 [0493\(2001\)129<2461: Aoapms>2.0.Co;2](https://doi.org/10.1175/1520-0493(2001)129<2461: Aoapms>2.0.Co;2).

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

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

657 Epstein, E. S., 1969. Stochastic dynamic prediction. Tellus B: Chemical and Physical Meteorology.
658 21(6), 739-759. <https://doi.org/10.3402/tellusa.v21i6.10143>.

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

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

667 Gelfand, A. E., 2000. Gibbs Sampling. Journal of the American Statistical Association. 95(452),
668 1300-1304. <https://doi.org/10.1080/01621459.2000.10474335>.

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

672 Gneiting, T., Raftery, A. E., III, A. H. W. and Goldman, T., 2005. Calibrated Probabilistic
673 Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. Monthly
674 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 Gritmit, 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/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(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>.

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

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

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

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

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

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

715 Li, Y., Wang, Q. J., He, H., Wu, Z. and Lu, G., 2020b. A method to extend temporal coverage of
716 high quality precipitation datasets by calibrating reanalysis estimates. Journal of Hydrology. 581,
717 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](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](https://doi.org/10.5194/hessd-4-655-2007).

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](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/10.1175/1520-0434\(2001\)016<0463:Tuoeti>2.0.Co;2](https://doi.org/10.1175/1520-0434(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>.

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

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

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

812 Wang, Q. J., Shao, Y., Song, Y., Schepen, A., Robertson, D. E., Ryu, D. and Pappenberger, F.,
813 2019a. An evaluation of ECMWF SEAS5 seasonal climate forecasts for Australia using a new
814 forecast calibration algorithm. *Environmental Modelling & Software*. 122, 104550.
815 <https://doi.org/10.1016/j.envsoft.2019.104550>.

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

819 Wang, Q. J., Zhao, T., Yang, Q. and Robertson, D., 2019b. A Seasonally Coherent Calibration
820 (SCC) Model for Postprocessing Numerical Weather Predictions. *Monthly Weather Review*.
821 147(10), 3633-3647. <https://doi.org/10.1175/mwr-d-19-0108.1>.

822 Wilks, D. S., 2011. *Statistical methods in the atmospheric sciences*, third ed. Academic Press,
823 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>.

847 **Tables**

848

849 Table 1. Differences in CRPS skill score between SCC2 and SCC1, SCC3 and SCC1, SCC3-RC and SCC1, and
 850 SCC3-RC and SCC3. The CRPS skill score values for all 180 cases (20 sites and 9 lead times) in the 3-year period
 851 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%

852

853 Table 2. Differences in twCRPS skill score between SCC2 and SCC1, SCC3 and SCC1, SCC3-RC and SCC1, and
 854 SCC3-RC and SCC3. The twCRPS skill score values for all 180 cases (20 sites and 9 lead times) in the 3-year period
 855 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%

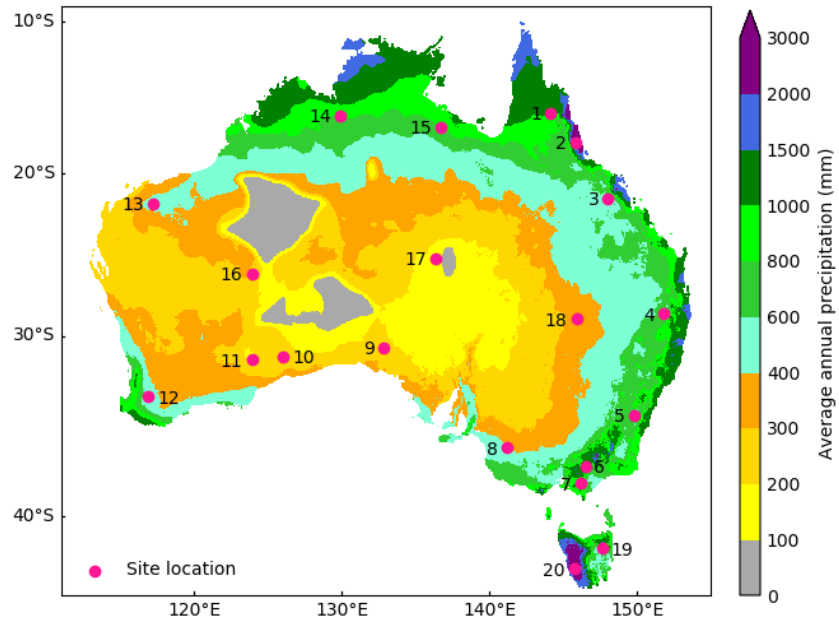
856

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

	<i>p</i> -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

860 **Figures**

861

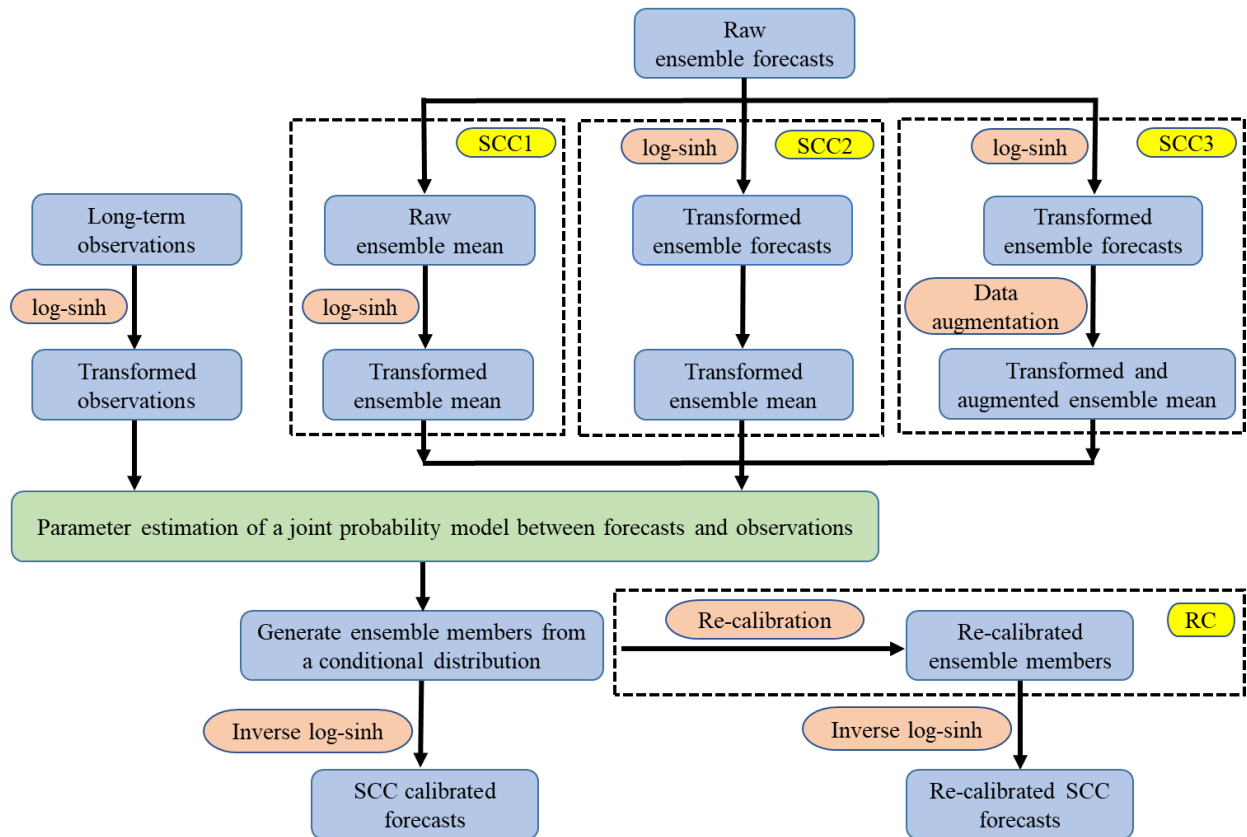


862

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

864

865

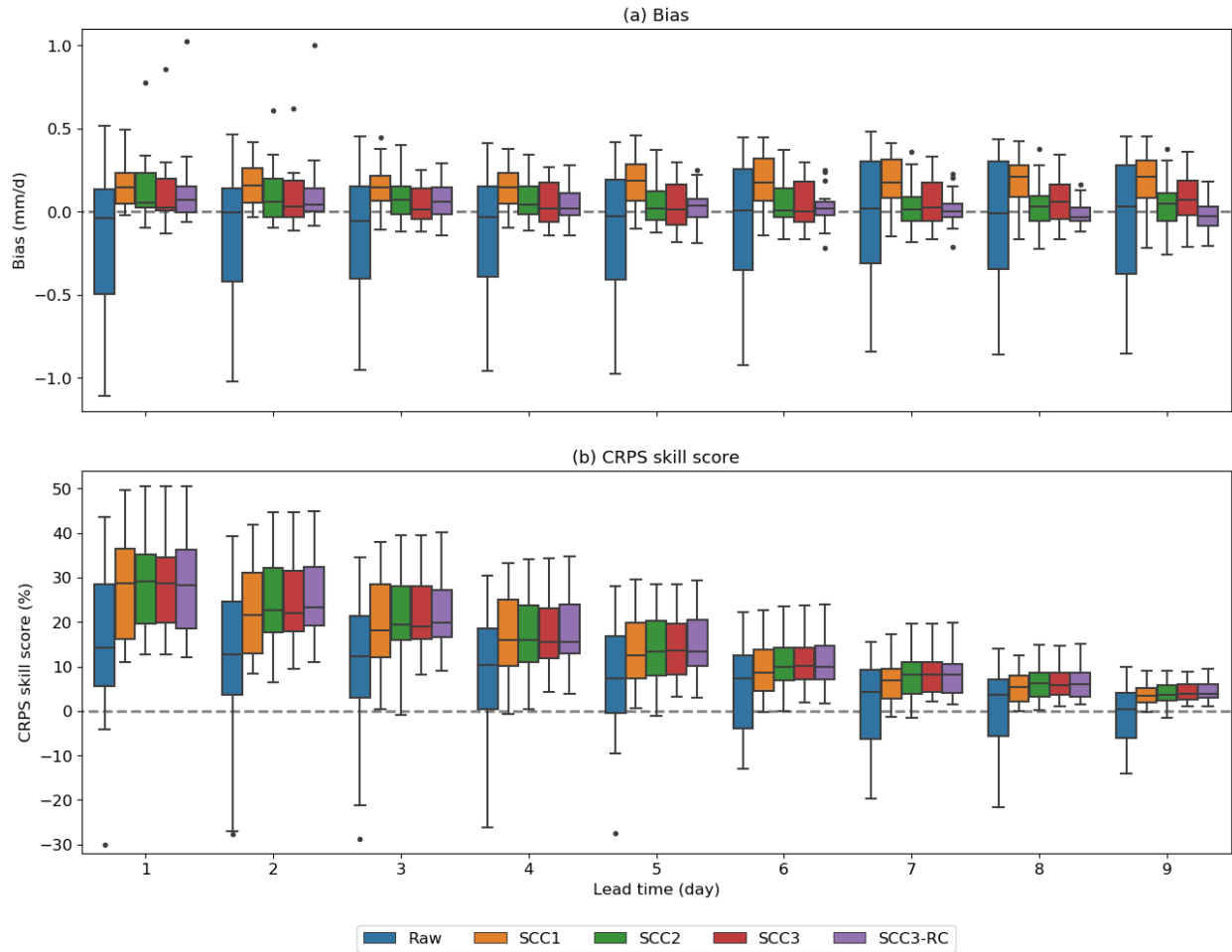


866

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

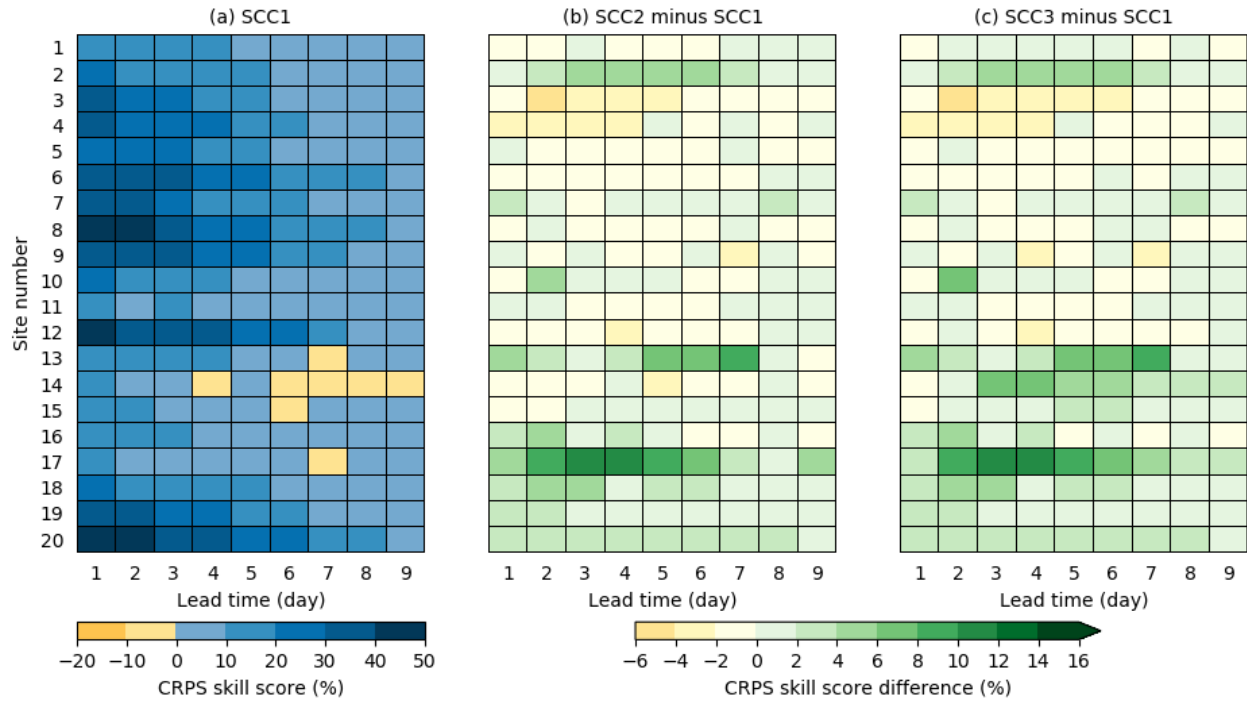


873

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

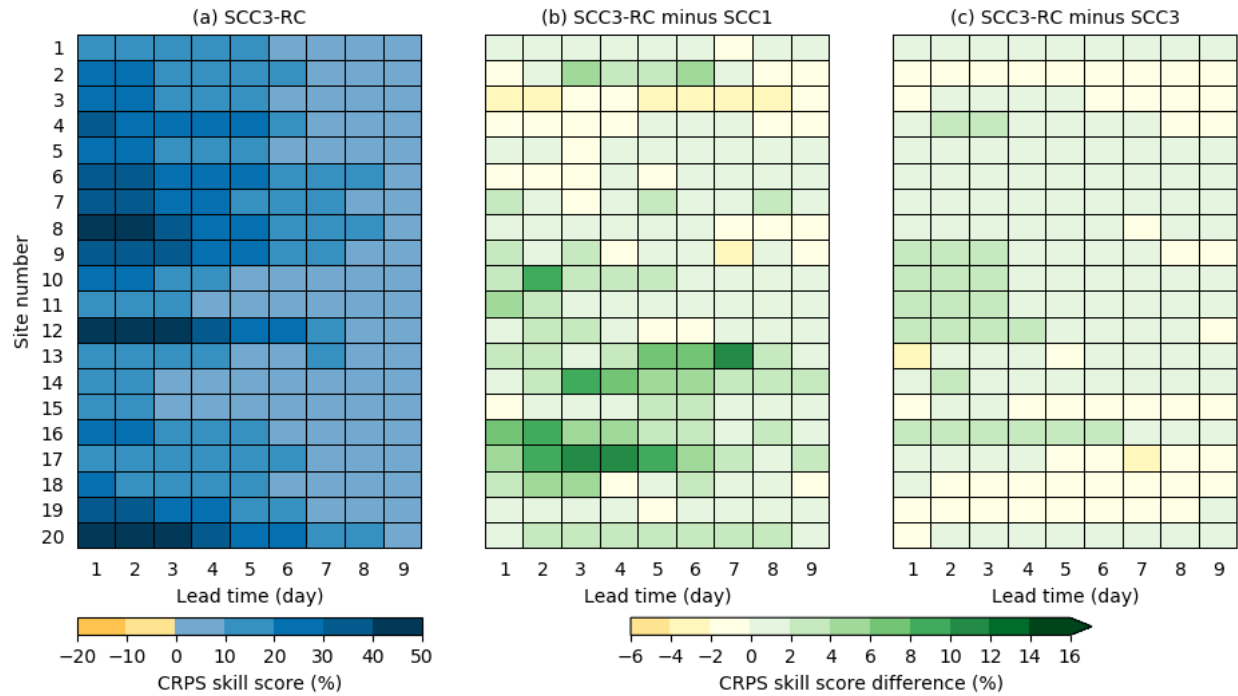
881



882

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

888



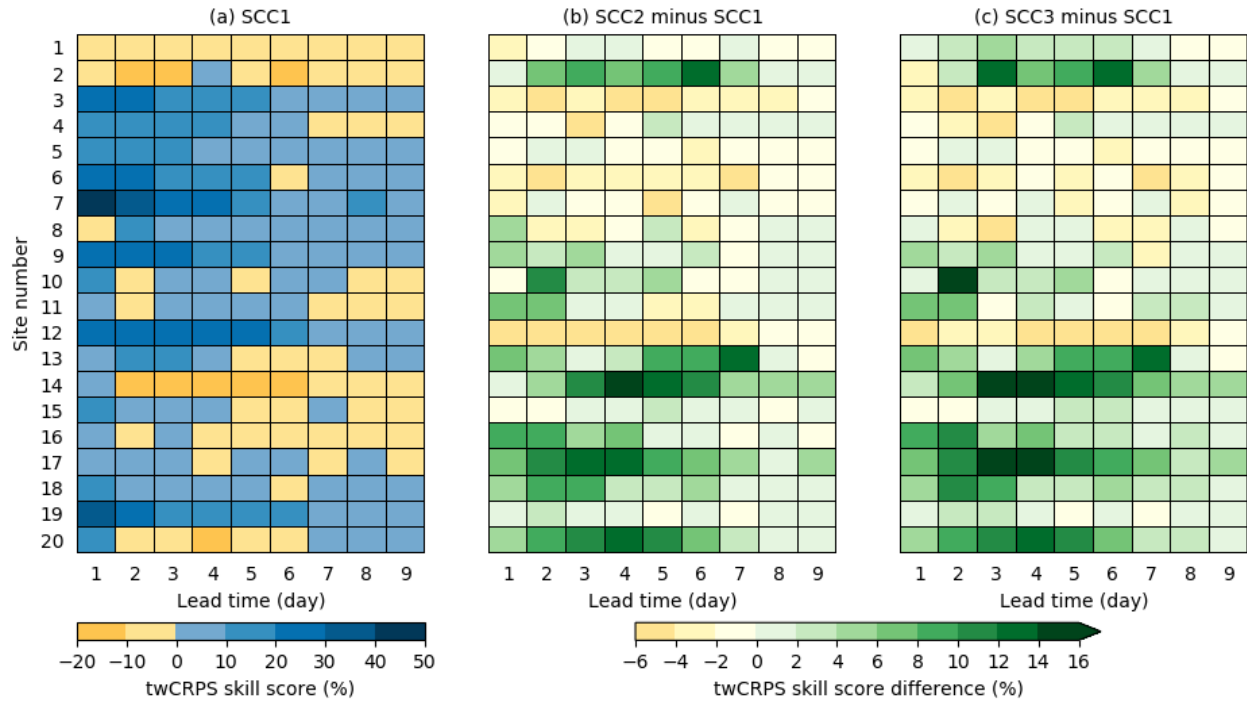
889

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

895

896

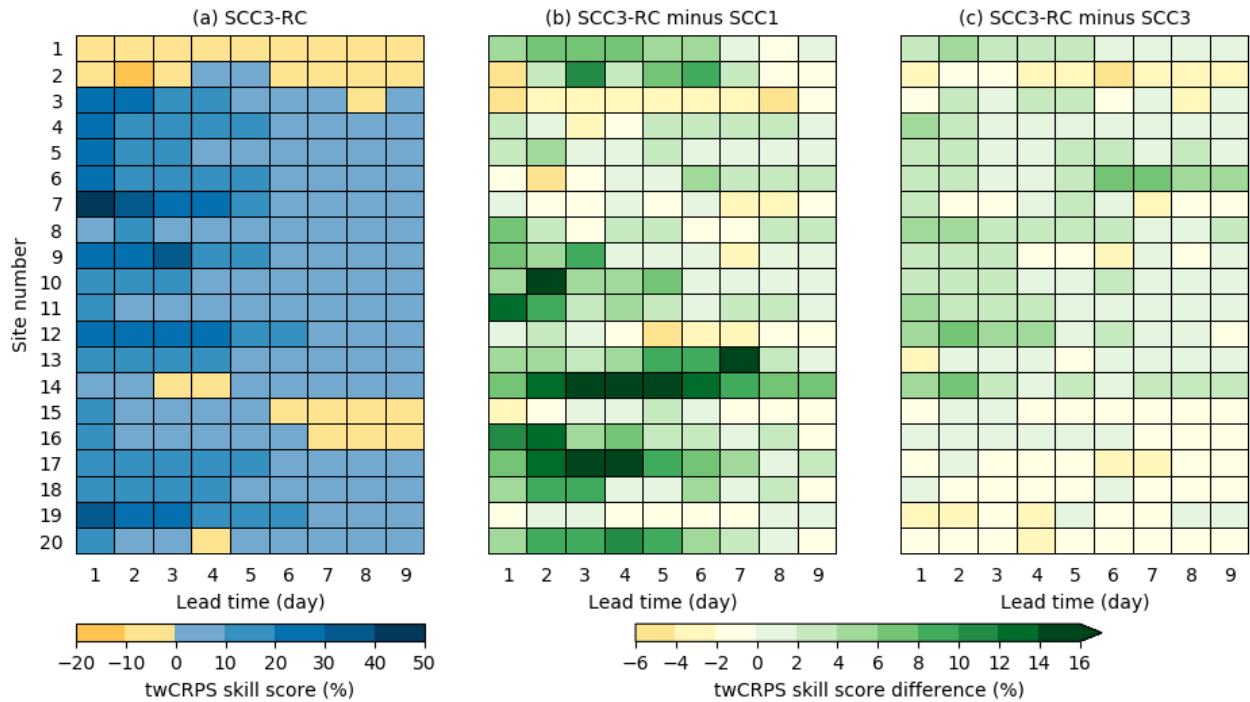
897



898

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

905



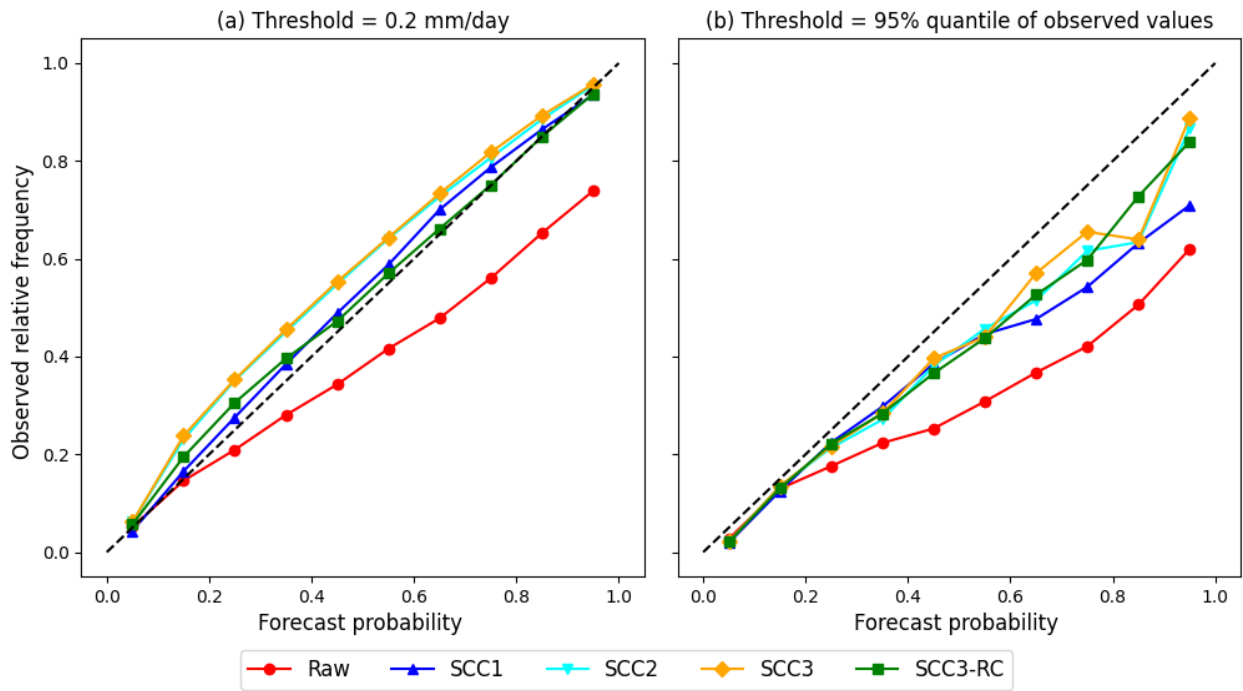
906

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.