



Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Zhao, P;Wang, QJ;Wu, W;Yang, Q

Title:

Spatial-Mode-Based Calibration (SMoC) of Forecast Precipitation Fields with Spatially Correlated Structures: An Extended Evaluation and Comparison with Gridcell-by-Gridcell Postprocessing

Date:

2023-09-01

Citation:

Zhao, P., Wang, Q. J., Wu, W. & Yang, Q. (2023). Spatial-Mode-Based Calibration (SMoC) of Forecast Precipitation Fields with Spatially Correlated Structures: An Extended Evaluation and Comparison with Gridcell-by-Gridcell Postprocessing. *Journal of Hydrometeorology*, 24 (9), pp.1509-1525. <https://doi.org/10.1175/JHM-D-23-0023.1>.

Persistent Link:

<https://hdl.handle.net/11343/339083>

1 **Spatial Mode-based Calibration (SMoC) of Forecast Precipitation Fields with**  
2 **Spatially Correlated Structures: An Extended Evaluation and Comparison**  
3 **with Grid-cell by Grid-cell Post-processing**

4

5 Pengcheng Zhao,<sup>ab</sup> Quan J. Wang,<sup>a</sup> Wenyan Wu,<sup>a</sup> and Qichun Yang<sup>a,c</sup>

6 *<sup>a</sup> Department of Infrastructure Engineering, The University of Melbourne, Parkville, Australia*

7 *<sup>b</sup> China International Engineering Consulting Corporation, Beijing, China*

8 *<sup>c</sup> Thrust of Earth, Ocean and Atmospheric Sciences, Hong Kong University of Science and Technology (GZ), China*

9

10 *Corresponding author: Pengcheng Zhao, [pengcheng@student.unimelb.edu.au](mailto:pengcheng@student.unimelb.edu.au)*

## ABSTRACT

Post-processing forecast precipitation fields from numerical weather prediction models aims to produce ensemble forecasts that are of high quality at each grid-cell and, importantly, are spatially structured in an appropriate manner. A conventional approach, the grid-cell by grid-cell post-processing, typically consists of two steps: (1) perform statistical calibration separately at individual grid-cells to generate unbiased, skillful, and reliable ensemble forecasts; (2) employ ensemble reordering to link ensemble members of all grid-cells according to certain templates to form spatially structured ensemble forecasts. However, ensemble reordering techniques are generally problematic in practical use. For example, the well-known Schaake shuffle is often criticized for not considering real physical atmospheric conditions. In this context, a fundamentally new approach, namely spatial mode-based calibration (SMoC), has recently been developed for post-processing forecast precipitation fields with inbuilt spatial structures, thereby eliminating the need for ensemble reordering. SMoC was tested on 1-day ahead forecasts of heavy precipitation events and was found to produce ensemble forecasts with appropriate spatial structures. In this paper, we extend SMoC to calibrate forecasts of light and no precipitation events and forecasts at long lead times. We also compare SMoC with the grid-cell by grid-cell post-processing. Results based on multiple evaluation metrics show that SMoC performs well in calibrating both forecasts of light and no precipitation events and forecasts at long lead times. Compared with the grid-cell by grid-cell post-processing, SMoC produces ensemble forecasts with similar forecast skill, improved forecast reliability, and clearly better spatial structures. In addition, SMoC is computationally far more efficient.

## 1. Introduction

Short-term forecasts of precipitation fields play an important role in the meteorological-hydrological forecasting chain and provide critical decision-making information for a wide range of water-related activities, such as flood prediction, irrigation planning, and water resources management (Robertson et al., 2013). In practice, these forecasts are routinely produced from numerical weather prediction (NWP) models, generally over a multitude of grid-cells, in either deterministic or ensemble forms. However, both forms of forecasts often suffer from systematic

39 errors due to various sources of deficiencies in NWP models (Toth and Kalnay, 1993; Gneiting  
40 et al., 2007). Ensemble forecasts also generally suffer from dispersion errors (Buizza et al., 1998;  
41 Roulston and Smith, 2003). These errors should be corrected before the forecasts are used  
42 (Pechlivanidis et al., 2020). In addition, the forecasts may fail to capture the spatial structures of  
43 observed precipitation (Wernli et al., 2008; Shrestha et al., 2015; Radanovics et al., 2018), which  
44 should also be addressed before the forecasts are used, as impacts from precipitation can be  
45 highly sensitive to its spatial distribution (Borga et al., 2007; Li et al., 2017). For these reasons, it  
46 has been a common practice to improve raw forecast precipitation fields through post-  
47 processing. A conventional approach, the grid-cell by grid-cell post-processing (Clark et al.,  
48 2004; Shrestha et al., 2015; Schefzik, 2017; Cattoën et al., 2020; Schepen et al., 2020), is  
49 typically performed following two steps.

50 In the first step, statistical calibration is applied to calibrate raw precipitation forecasts  
51 separately for individual grid-cells, aiming to produce ensemble forecasts that are bias free,  
52 reliable in ensemble spread, and as skillful as possible. Popular calibration models include  
53 ensemble model output statistics (EMOS) (Gneiting et al., 2005; Scheuerer, 2014; Baran and  
54 Lerch, 2015), Bayesian model averaging (BMA) (Raftery et al., 2005; Sloughter et al., 2007;  
55 Wang et al., 2012a; Baran and Möller, 2015), the censored shifted Gamma distribution (CSGD)  
56 (Scheuerer and Hamill, 2015a; Zhang et al., 2017), the Bayesian Joint Probability (BJP) model  
57 (Wang et al., 2009; Robertson et al., 2013; Shrestha et al., 2015; Zhao et al., 2015; Wang et al.,  
58 2019a; Cattoën et al., 2020; Li et al., 2020a), and the seasonally coherent calibration (SCC)  
59 model (Wang et al., 2019b; Zhao et al., 2020; Yang et al., 2021; Du et al., 2022; Zhao et al.,  
60 2022a). As the calibration is carried out separately for each individual grid-cell, members of the  
61 calibrated ensemble forecasts from different grid-cells are disconnected. This problem is  
62 addressed in the second step, where ensemble reordering is applied to establish spatial structures  
63 for the ensemble members produced in the first step, usually by linking ensemble members from  
64 different grid-cells based on certain ordering templates. The two most commonly used ensemble  
65 reordering techniques are the Schaake shuffle (Clark et al., 2004) and the ensemble copula  
66 coupling (Schefzik et al., 2013; Schefzik, 2017).

67 The grid-cell by grid-cell post-processing is widely implemented in a variety of studies and  
68 performs well in improving raw forecast precipitation fields (Shrestha et al., 2015; Cattoën et al.,  
69 2020; Schepen et al., 2020; Shrestha et al., 2020; Li et al., 2022). However, the second step, i.e.,

70 ensemble reordering, is often problematic in practical use. The Schaake shuffle, which reorders  
71 ensemble members using ordering templates derived from historical observations, is not capable  
72 of considering real physical atmospheric conditions of precipitation events. In addition, because  
73 there are often a large number of zero precipitation values in historical records, tied orders will  
74 happen when constructing ordering templates (Bellier et al., 2017). Wu et al. (2018) pointed out  
75 that this could undermine the effectiveness of the Schaake shuffle. The ensemble copula  
76 coupling, which takes raw ensemble members as an ordering template, is only suitable for the  
77 post-processing of ensemble forecasts because there is no ensemble available for deterministic  
78 forecasts to form an ordering template. Furthermore, the use of ensemble copula coupling is  
79 subject to ensemble size and the quality of spatial structures of raw ensembles. Although  
80 ensemble reordering techniques like Schaake shuffle and ensemble copula coupling have been  
81 enhanced by several studies (Hu et al., 2016; Schefzik, 2016; Scheuerer et al., 2017; Bellier et  
82 al., 2018; Scheuerer and Hamill, 2018; Straaten et al., 2018; Wu et al., 2018), mainly in terms of  
83 the construction of ordering templates, the problems mentioned above have not been solved.

84 A fundamentally new approach, namely spatial mode-based calibration (SMoC) (Zhao et al.,  
85 2022b), has recently been developed to deal with the problems of ensemble reordering. SMoC is  
86 capable of calibrating forecast precipitation fields as a whole and producing ensemble forecasts  
87 with inbuilt spatial structures, therefore eliminating the use of ensemble reordering. SMoC is  
88 developed based on spatial modes derived from the empirical orthogonal function (EOF)  
89 analyses of long-term observed precipitation fields (Hannachi et al., 2007; Zhao et al., 2022b).  
90 The calibration is performed through linear regressions of EOF expansion coefficients of  
91 forecasts and corresponding observations that are derived based on the long-term spatial modes.  
92 And by only calibrating expansion coefficients from the first few dominant EOF modes, the  
93 dimension of the post-processing of forecast fields is reduced from hundreds or even thousands  
94 to a much smaller number, which contributes to the synchronous calibration of all forecast grid-  
95 cells. Compared with the grid-cell by grid-cell post-processing that is applied to individual  
96 forecast grid-cells, SMoC is applied to the whole forecast fields, and spatial structures are  
97 therefore inherently constructed in calibrated ensemble members.

98 In the original SMoC study (Zhao et al., 2022b), SMoC was applied to calibrate forecast  
99 precipitation fields of heavy events at 1 day ahead and was found to produce high-quality  
100 ensemble forecasts at both grid-cell and field scales. A more comprehensive evaluation of SMoC

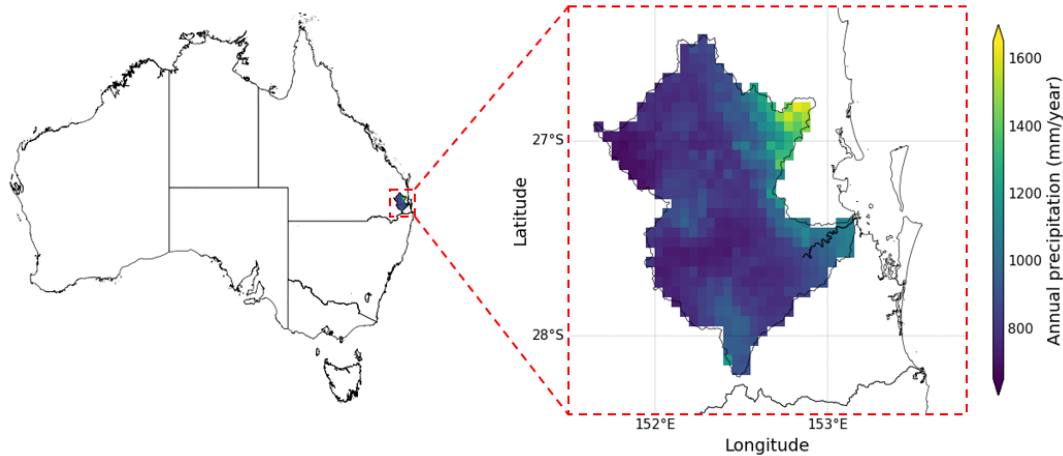
101 is required for forecasts of light and no precipitation events and forecasts at long lead times to  
102 examine its efficacy. In addition, a comparison between SMOc and the grid-cell by grid-cell  
103 post-processing is also required to inform forecast users which approach is better to use. In this  
104 paper, we extend the study of Zhao et al. (2022b) by (1) applying SMOc to calibrate forecasts of  
105 both heavy precipitation events and light and no precipitation events; (2) applying SMOc to  
106 calibrate forecasts at both short and long lead times; and (3) comparing SMOc with the grid-cell  
107 by grid-cell post-processing. Specifically, we apply SMOc to calibrate daily forecast  
108 precipitation fields at lead times of 1, 3, 5, 7, and 9 days ahead for a period of 3 years. For the  
109 grid-cell by grid-cell post-processing, we employ the SCC-SS model, i.e., SCC plus the Schaake  
110 shuffle (SS), which represent the state-of-the-art calibration models and ensemble reordering  
111 techniques, respectively. For SMOc forecasts and SCC-SS forecasts, we apply a comprehensive  
112 set of metrics to evaluate multiple aspects of forecast quality, especially the quality in spatial  
113 characteristics. We also evaluate the computational efficiency of the two approaches, which is  
114 important in practical applications.

115 The remainder of this paper is structured as follows. We introduce precipitation forecast and  
116 observation data, the SCC-SS model, the SMOc model, and forecast evaluation metrics in the  
117 next section. We present forecast evaluation and comparative results in Section 3. After a few  
118 discussions in Section 4, we provide a summary and conclude this paper in Section 5.

## 119 **2. Data and methods**

### 120 *a. Data*

121 In this study, we evaluate and compare the SMOc model and the grid-cell by grid-cell post-  
122 processing by applying them to raw daily forecast precipitation fields of the Brisbane Drainage  
123 Basin (shown in Fig. 1). Located in eastern Australia, the Brisbane Drainage Basin covers a large  
124 spatial area (13,550 km<sup>2</sup>) and has an average annual precipitation of 866 mm.



125

126 Fig. 1. Geographical location (left figure) and annual precipitation map (right figure) of the Brisbane  
 127 Drainage Basin.

128 We obtain observed daily precipitation data of the Brisbane Drainage Basin from the  
 129 Australian Water Availability Project's (AWAP) climate datasets (Jones et al., 2009). AWAP  
 130 precipitation datasets are well-known as reference precipitation data for Australia and are widely  
 131 used in precipitation-related research (Lerat et al., 2020). Produced from the interpolation of  
 132 available rain gauge observations in Australia, AWAP datasets have a horizontal grid spacing of  
 133  $0.05^\circ \times 0.05^\circ$  across Australia on a daily basis. In this study, the gridded daily AWAP data for a  
 134 period of 20 years from 3 November 1998 to 2 November 2018 are used as long-term  
 135 observations for deriving parameters of the SMoC model, the SCC model, and the Schaake  
 136 shuffle. AWAP data for a period of 3 years from 3 November 2018 to 2 November 2021 are used  
 137 as reference data for forecast calibration and evaluation in the forecast period.

138 We obtain raw forecast precipitation fields of the Brisbane Drainage Basin from the  
 139 Australian Community Climate and Earth-System Simulator Global 3 (ACCESS-G3) model,  
 140 which is operated by the Australian Bureau of Meteorology. ACCESS-G3 is a deterministic  
 141 NWP model and produces forecasts with a horizontal grid spacing of  $0.18^\circ$  (longitude) by  $0.12^\circ$   
 142 (latitude) on an hourly basis. With forecast lead times up to 10 days, ACCESS-G3 precipitation  
 143 forecasts of 1, 3, 5, 7, and 9 days ahead are selected for evaluating forecast post-processing  
 144 performance at different lead times. ACCESS-G3 precipitation forecast data from 3 November  
 145 2018 to 2 November 2021 are used for this study.

146 To match ACCESS-G3 forecasts with AWAP observations, we modify the spatial and  
 147 temporal resolutions of the forecast data. Specifically, we apply bilinear interpolation to produce

148 precipitation forecasts with the horizontal grid spacing of  $0.05^\circ \times 0.05^\circ$ . Under this spatial  
149 coverage, the Brisbane Drainage Basin contains 493 grid-cells for both observations and the  
150 interpolated forecasts. And we accumulate the hourly ACCESS-G3 forecasts to daily  
151 precipitation amounts based on the AWAP data records. Consequently, for each of the 493 grid-  
152 cells in the Brisbane Drainage Basin, we obtain 20 years of long-term daily precipitation  
153 observations and another 3 years of daily precipitation forecasts and corresponding observations.

#### 154 *b. The SCC-SS model*

155 The grid-cell by grid-cell post-processing employed in this study includes two steps, i.e.,  
156 statistical calibration for individual forecast grid-cells using the seasonally coherent calibration  
157 (SCC) model (Wang et al., 2019b) and ensemble reordering of SCC calibrated ensemble  
158 forecasts using the Schaake shuffle (Clark et al., 2004). Both of these two steps are implemented  
159 separately to different lead times.

#### 160 1) THE SCC MODEL

161 The SCC model is used to calibrate raw precipitation forecasts for each grid-cell and produce  
162 calibrated ensemble forecasts that are bias free, reliable in ensemble spread, as skillful as  
163 possible, coherent in seasonal climatology, and consistent with long-term observations, even  
164 when the archive of available precipitation forecasts has limited records. In establishing the SCC  
165 model, we first apply the regionally optimized power transformation in Du et al. (2022) to  
166 normalize both forecast and observation data from each of the 493 grid-cells, to make the highly  
167 skewed precipitation data suitable for a normal distribution modelling framework (Wang et al.,  
168 2012b; Peng et al., 2014; Li et al., 2019). We then establish a joint probability model by fitting a  
169 bivariate normal distribution of raw forecasts and corresponding observations.

170 For deriving SCC model parameters, we use 20 years of observation data to obtain  
171 parameters relevant to the long-term climatology of observations, and we use 3 years of raw  
172 forecasts and corresponding observations to obtain remaining parameters. The optimization of  
173 parameters is achieved by using the maximum likelihood method and the Nelder-Mead searching  
174 algorithm (Nelder and Mead, 1965). After obtaining all of the parameters, we apply SCC to  
175 calibrate new forecasts for each grid-cell. Given a raw forecast, a probability distribution  
176 conditional on the raw forecast can be derived through the SCC model. In this study, we

177 randomly sample an ensemble of 1,000 values from the distribution to represent the whole  
178 probability distribution. And we transform these values back to the original space through an  
179 inverse of power transformation to produce a calibrated ensemble forecast with 1,000 members.

180 For producing SCC calibrated ensemble forecasts over the 3-year forecast period, we use a  
181 leave-one-month-out cross-validation setup. For each month, we first leave the forecast and  
182 observation data out of the 3 years and employ the rest data to derive model parameters. We then  
183 apply the established SCC model to raw forecasts from the left-out month and produce calibrated  
184 ensemble forecasts.

## 185 2) THE SCHAAKE SHUFFLE

186 The Schaake shuffle is used to instill appropriate spatial structures into SCC calibrated  
187 ensemble forecasts. Ensemble members from individual grid-cells are reordered according to  
188 ordering templates that are constructed from historical observations. This in effect connects  
189 ensemble members from different grid-cells by reproducing the spatial patterns of historical  
190 precipitation events.

191 For a given forecast event, we select observed historical events the dates of which (month-  
192 day date) lie within 2 days before and after the forecast date. Therefore, 5 days of observed  
193 events can be selected from a historical year and 100 historical events can be obtained from the  
194 20 years of AWAP observations. These 100 historical field events form an ordering template for  
195 the given forecast event over the whole field. And this ordering template can be further  
196 decomposed into 493 ordering templates that correspond to the given forecast event at each of  
197 the 493 grid-cells. These 493 ordering templates have the same historical date patterns and are  
198 therefore spatially correlated across different grid-cells.

199 We use the 493 ordering templates to reorder the SCC calibrated ensemble members for each  
200 grid-cell. The ranks of ensemble members are reordered based on ranks of selected historical  
201 observations in the ordering template (Clark et al., 2004). For a given grid-cell, as there are 1,000  
202 ensemble members, we randomly divide them to 10 groups, each with 100 members. For each  
203 group, we reorder the 100 ensemble members according to the ordering template of the grid-cell.  
204 After reordering ensemble members for all groups and all grid-cells, ensemble members with the

205 same group numbers and the same event orders are connected to form spatially correlated  
206 forecast fields.

207 In this way, we can obtain an ensemble of forecast precipitation fields for the given forecast  
208 event with an ensemble size of 1,000, and the resulting connected ensemble members have  
209 appropriate spatial structures across different grid-cells. SCC calibrated ensemble forecasts for  
210 all forecast events are reordered using the Schaake shuffle (SS) to produce SCC-SS forecasts for  
211 the 3-year forecast period.

### 212 *c. The SMoC model*

213 The SMoC model is implemented to calibrate forecast precipitation fields as a whole but  
214 separately for different lead times. In establishing the SMoC model, we first apply the regionally  
215 optimized power transformation to normalize the precipitation data. The transformation here is  
216 the same as the transformation of precipitation data for the SCC model. We then employ the  
217 EOF analysis to decompose the 20 years of AWAP observations to spatial modes and expansion  
218 coefficients. The 20 years of spatial modes can be used as representative long-term spatial modes  
219 to derive expansion coefficients of forecasts and corresponding observations in the 3-year  
220 forecast period. After the removal of seasonality which aims to pool data from different months  
221 together, these two derived expansion coefficients from each of the first few dominant EOF  
222 modes are finally related using ordinary least-square linear regressions. Normal distributions of  
223 regression residuals are used as uncertainty information to produce ensemble expansion  
224 coefficients. In this study, we use expansion coefficients from the first 10 EOF modes, following  
225 the setup from the original SMoC study (Zhao et al., 2022b).

226 After obtaining all of the related model parameters, we apply SMoC to calibrate new forecast  
227 fields. Given a raw forecast precipitation field, forecast values can be converted to forecast  
228 expansion coefficients using the long-term spatial modes and then lead to 10 normal distributions  
229 conditional on the raw forecast expansion coefficients from the first 10 EOF modes. We sample  
230 10 expansion coefficient values separately from the 10 distributions and convert them back to the  
231 original spatial space using the long-term spatial modes and an inverse of power transformation  
232 to produce a calibrated forecast precipitation field. This procedure is repeated for 1,000 times to  
233 produce an ensemble of 1,000 calibrated forecast precipitation fields. The flowcharts of the  
234 modelling process of SMoC and the calibration of new forecast fields using SMoC can be

235 referred to Fig. S7 and Fig. S8 in Supplementary Material S1. Similarly, we use a leave-one-  
236 month-out cross-validation for the generation of SMOc forecasts for the 3-year forecast period.

237 In this study, besides the SMOc model established for forecasts of heavy precipitation events  
238 in the original SMOc study, we establish a second SMOc model specifically for forecasts of light  
239 and no precipitation events. The modelling processes of these two SMOc models are the same,  
240 except that they are established based on expansion coefficients of heavy precipitation events,  
241 and light and no precipitation events, respectively. These two expansion coefficients are derived  
242 using the same long-term spatial modes. Daily precipitation events are defined as heavy (light  
243 and no) if the basin average of raw forecast precipitation fields is beyond (below) a threshold,  
244 which is set as 90% quantile of raw forecast basin averages during the 3-year forecast period and  
245 equals 5.47 mm per day. It should be noted that this definition is solely dependent on raw  
246 forecasts as the selection of heavy events based on observations will result in bias effects for  
247 post-processing models (Diks et al., 2011; Gneiting and Ranjan, 2011). Consequently, in the  
248 1,096 days from the 3-year period, the amounts of heavy precipitation events, and light and no  
249 precipitation events are 110 and 986, respectively.

#### 250 *d. Forecast evaluation*

251 For a comprehensive comparison, the quality of SCC-SS forecasts and SMOc forecasts is  
252 evaluated considering several aspects, including forecast skill using continuous ranked  
253 probability score (CRPS) (Hersbach, 2000), forecast reliability using probability integral  
254 transform (PIT) (Renard et al., 2010) and spread-error correlation (Whitaker and Loughe, 1998;  
255 Van Schaeybroeck and Vannitsem, 2016), forecast ruggedness using terrain ruggedness index  
256 (TRI) (Riley et al., 1999), and forecast correlation using variogram score (VS) (Scheuerer and  
257 Hamill, 2015b). The computational time of SCC-SS and SMOc is also calculated to evaluate  
258 their efficiency for practical applications. It should be noted that evaluation of raw forecasts is  
259 not considered in this study, as both SCC-SS and SMOc have shown great capability to improve  
260 the quality of raw forecasts in previous studies (Clark et al., 2004; Shrestha et al., 2015; Wang et  
261 al., 2019b; Zhao et al., 2022a).

262 For demonstration purposes, forecast evaluation is implemented at both grid-cell and basin  
263 scales. Details of evaluation metrics at these two scales are respectively introduced. The  
264 evaluations are applied to forecasts of light and no precipitation events, heavy precipitation

265 events, and all events in the 3-year forecast period to evaluate forecast performance for different  
266 magnitudes of precipitation events.

## 267 1) FORECAST EVALUATION AT GRID-CELL SCALE

### 268 (i) Forecast skill

269 We use continuous ranked probability score (CRPS) to evaluate forecast skill of ensemble  
270 forecasts at each of the 493 grid-cells (Wang et al., 2019b; Cattoën et al., 2020; Du et al., 2022;  
271 Yang et al., 2022a; Zhao et al., 2022a). CRPS assesses the difference between ensemble forecast  
272 cumulative distributions and corresponding observations. For a set of ensemble forecasts at days  
273  $t = 1, 2, \dots, T$ , an average CRPS ( $\overline{CRPS}$ ) is calculated as:

$$274 \quad \overline{CRPS} = \frac{1}{T} \sum_{t=1}^T \int \{F(t, x) - H[x - y(t)]\}^2 dx \quad (1)$$

275 where  $F(t, x)$  is the ensemble forecast cumulative density function (CDF), and  $y(t)$  is the  
276 corresponding observation at day  $t$ ;  $H$  is the Heaviside step function that equals 1 if  $x \geq y(t)$   
277 and equals 0 otherwise; and  $T$  is the number of days in the evaluation period.

278 We calculate CRPS skill score to assess the forecast skill improvement of SCC-SS forecasts  
279 and SMOc forecasts relative to reference forecasts. In this study, we use climatology ensemble  
280 forecasts as the reference forecasts. For each precipitation event, we produce a climatology  
281 ensemble forecast following two steps: (a) we fit a distribution for 20 years of AWAP  
282 observations of the month to which the event day belongs; (b) we draw a random sample from  
283 the fitted distribution to generate an ensemble of 1,000 climatology values for the event. After  
284 climatology ensemble forecasts are generated for all precipitation events, we calculate an  
285 average CRPS of reference forecasts ( $\overline{CRPS}_{ref}$ ) and produce an CRPS skill score:

$$286 \quad CRPS \text{ skill score} = \frac{\overline{CRPS}_{ref} - \overline{CRPS}}{\overline{CRPS}_{ref}} \times 100(\%) \quad (2)$$

287 The maximum value of CRPS skill score is 100%, indicating that forecasts perfectly match  
288 with corresponding observations. A skill score of 0% indicates that forecast errors are equal to  
289 errors of reference forecasts. A positive (negative) skill score indicates that forecasts are more  
290 (less) skillful than reference forecasts.

291 By using climatology ensemble forecasts as reference forecasts, we can obtain the results of  
 292 how SCC-SS and SMOc compare with reference forecasts and how SCC-SS and SMOc compare  
 293 with each other. For studies that only aim to assess the comparison between two post-processing  
 294 models (e.g., SCC-SS and SMOc), post-processed forecasts from one model (e.g., SCC-SS) can  
 295 be used as reference forecasts to evaluate the relative performance of the other model (e.g.,  
 296 SMOc) (Li et al., 2020a).

297 *(ii) Forecast reliability*

298 We use probability integral transform (PIT) to evaluate forecast reliability of ensemble  
 299 forecasts at each of the 493 grid-cells (Wang et al., 2020; Zhao et al., 2020; Yang et al., 2022b).  
 300 Forecast reliability shows if ensemble spread is reliable (not too wide or too narrow) and if  
 301 ensemble forecast probability distributions and observed frequency of observations are consistent  
 302 in statistics. The PIT of a forecast-observation pair at day  $t$  is calculated as:

$$303 \quad \pi(t) = F[t, x = y(t)] \quad (3)$$

304 where  $F(t, x)$  is the CDF of the ensemble forecast, and  $y(t)$  is the corresponding observation.  
 305 For reliable ensemble forecasts at days  $t = 1, 2, \dots, T$ ,  $\pi(t)$  follows a uniform distribution. The  
 306 uniformity can be evaluated using the PIT alpha index:

$$307 \quad PIT \text{ alpha index} = 1 - \frac{2}{T} \sum_{t=1}^T \left| \pi^*(t) - \frac{t}{T+1} \right| \quad (4)$$

308 where  $\pi^*(t)$  is the sorted  $\pi(t)$ ,  $t = 1, 2, \dots, T$ , in an increasing order; and  $T$  is the number of days  
 309 in the evaluation period. The value of PIT alpha index ranges from 0 to 1. A value of 1 shows  
 310 perfect reliability and a value of 0 shows poorest reliability.

311 The PIT approach is employed in many studies (Li et al., 2020a; Li et al., 2020b; Schepen et  
 312 al., 2020) and shows good ability to examine ensemble forecast reliability. However, PIT is not  
 313 sensitive to extreme ensemble member values and is therefore not capable of taking into account  
 314 such an effect. For this reason, we also use the spread-error correlation approach to evaluate the  
 315 forecast reliability by comparing ensemble spread with the error of ensemble mean. Here we  
 316 calculate square root values of average ensemble variance and compare them with root mean  
 317 square error (RMSE) values of ensemble mean forecasts (Fortin et al., 2014) at each grid-cell.  
 318 For reliable ensemble forecasts, ensemble spread and the error of ensemble mean are supposed to  
 319 be equivalent if sufficient forecast samples are available.

320 2) FORECAST EVALUATION AT BASIN SCALE

321 (i) Forecast skill

322 For an overall evaluation at basin scale, we calculate the basin average of each member of  
323 ensemble forecasts and evaluate basin average ensemble forecasts, as implemented in the  
324 original SMOc paper (Zhao et al., 2022b). We use CRPS skill score to assess forecast skill  
325 improvement of basin average ensemble forecasts relative to reference forecasts. Similarly, for  
326 each basin average forecast event, we produce a reference ensemble climatology forecast by  
327 sampling from the distribution fitted for 20-year basin average observations of the month to  
328 which the event day belongs.

329 (ii) Forecast reliability

330 We use PIT and spread-error correlation to evaluate forecast reliability of basin average  
331 forecasts. If ensemble forecasts are reliable at grid-cell scale, forecast reliability at basin scale  
332 can be used to indicate whether forecasts are spatially structured in an appropriate manner (Zhao  
333 et al., 2022b). When the spatial structure is not appropriate, ensemble members from different  
334 grid-cells tend to be randomly connected and basin average forecasts will be narrow in ensemble  
335 spread. By contrast, when the spatial structure is appropriate, ensemble members from different  
336 grid-cells are connected in a way that members with large values tend to appear together and  
337 small values together, and basin average forecasts will be appropriate in ensemble spread.  
338 Therefore, when ensemble forecasts are reliable at both grid-cell and basin scales, appropriate  
339 spatial structures are suggested.

340 (iii) Forecast ruggedness

341 Terrain ruggedness index (TRI) was introduced by Riley et al. (1999) to measure the  
342 elevation differences between adjacent grid-cells. Here we use this concept to estimate the  
343 differences of precipitation amounts between adjacent grid-cells. Taking a square of nine grid-  
344 cells as an example, the precipitation amount at the center grid-cell is  $x_0$ , and precipitation  
345 amounts of the surrounding  $K$  ( $K = 8$ ) grid-cells from left top to right bottom are  $x_1, x_2, \dots, x_k$ ,  
346  $k = 1, 2, \dots, K$ . The TRI for the center grid-cell is calculated as:

347 
$$TRI = \sqrt{\sum_{k=1}^K (x_k - x_0)^2} \quad (5)$$

348 where  $x_k$  is the precipitation amount of the  $k$ th grid-cell around the center grid-cell. For an  
 349 ensemble forecast member of a forecast event, TRI values calculated for all grid-cells can be  
 350 averaged across the whole field. And TRI values for all ensemble members and all of the  
 351 evaluated forecasts can be further averaged to give an overall TRI. To evaluate the TRI value of  
 352 forecasts, TRI is also calculated for corresponding observations. Here we calculate a ruggedness  
 353 dissimilarity (RD) using forecast TRI and observation TRI:

$$354 \quad RD = \frac{TRI_f - TRI_o}{TRI_f + TRI_o} \quad (6)$$

355 where  $TRI_f$  and  $TRI_o$  are TRI values of forecasts and observations, respectively. RD is a  
 356 dimensionless measure and ranges from -1 to 1. A RD value close to 0 is preferred, showing  
 357 similar ruggedness of forecasts to the observations. A RD value larger (smaller) than 0 shows  
 358 that forecasts are more (less) rugged than corresponding observations.

359 *(iv) Forecast correlation*

360 Variogram score (VS) was first used for evaluating the correlation structure of wind speed  
 361 fields (Scheuerer and Hamill, 2015b), and was then extended for some other fields, including  
 362 temperature (Schefzik, 2017), streamflow (Hemri et al., 2015), and precipitation (Scheuerer et  
 363 al., 2017; Schepen et al., 2020). Here we use VS to evaluate the spatial correlation of ensemble  
 364 forecast precipitation fields. Specifically, VS measures how ensemble members (with orders)  
 365 from two grid-cells match with each other based on the match of their respective observations.  
 366 Taking any two grid-cells of  $i$  and  $j$  as an example, each grid-cell contains  $M$  ( $M = 1,000$ )  
 367 ensemble members for one forecast event, with orders from 1 to  $M$ . A VS can be calculated as:

$$368 \quad VS_{ij} = w_{ij} \left( |y_i - y_j|^p - \frac{1}{M} \sum_{m=1}^M |x_{m,i} - x_{m,j}|^p \right)^2 \quad (7)$$

369 where  $y_i$  and  $y_j$  are observations of the two grid-cells  $i$  and  $j$ , respectively;  $x_{m,i}$  and  $x_{m,j}$  are the  
 370  $m$ th ordered ensemble members of the two grid-cells  $i$  and  $j$ , respectively;  $w_{ij}$  is a weight and is  
 371 commonly set as 1;  $p$  is an order and is often set as 0.5. For all pairs of grid-cells inside a field  
 372 ( $N$  grid-cells in total and  $N = 493$  for our case study), VS can be calculated as:

$$373 \quad VS = \sum_{i=1}^N \sum_{j=1}^N VS_{ij} \quad (8)$$

374 VS can be calculated as the average value across all of the evaluated forecasts and the 493  
375 grid-cells. A smaller VS is preferable, indicating the spatial correlation of ensemble members  
376 from different grid-cells is closer to corresponding observations.

### 377 3) COMPUTATIONAL TIME

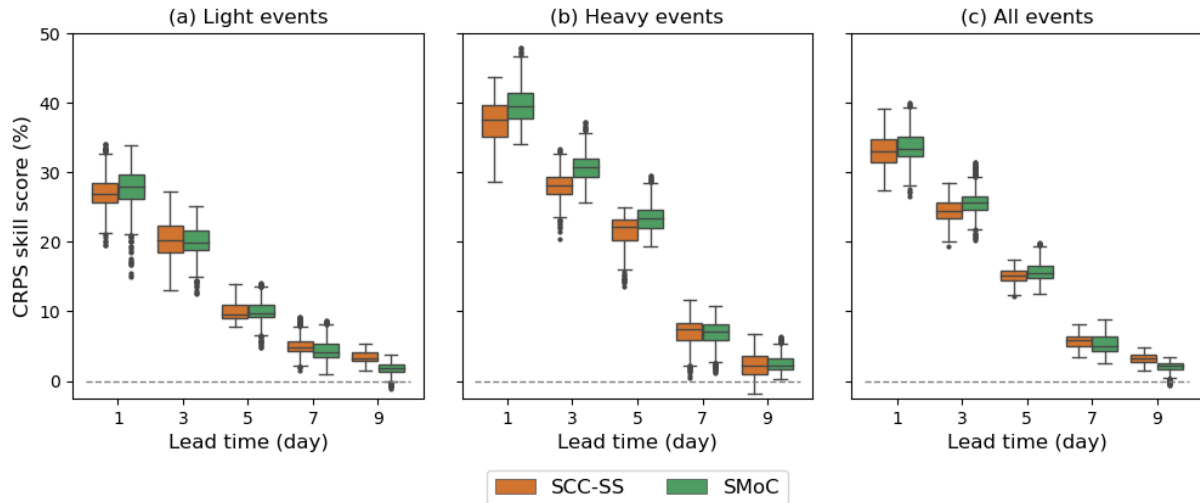
378 We calculate the computational time of SCC-SS and SMoC needed to calibrate raw forecast  
379 precipitation fields and produce ensemble forecasts with spatially correlated structures. In this  
380 study, we employ central processing units (CPUs) from the National Computational  
381 Infrastructure and Python programming to implement these two approaches. The calculated time  
382 is based on the post-processing of forecast precipitation fields of one lead time for all  
383 precipitation events (1,096 daily events in the 3-year period).

## 384 **3. Results**

### 385 *a. Forecast evaluation at grid-cell scale*

#### 386 1) FORECAST SKILL

387 Results of CRPS skill score for SCC-SS forecasts and SMoC forecasts at individual grid-  
388 cells are shown in Fig. 2. CRPS skill scores of these two forecasts tend to decrease gradually  
389 with increasing lead times, which is due to lower forecast skill of raw forecasts at longer lead  
390 times. Almost all of the CRPS skill scores are positive, indicating that these two forecasts are  
391 more skillful than reference climatology forecasts. For light and no precipitation events, SMoC  
392 forecasts have comparable forecast skill to SCC-SS forecasts at a set of different lead times. By  
393 contrast, for heavy events, SMoC forecasts have clearly higher forecast skill than SCC-SS  
394 forecasts at each of the investigated lead times, especially at short lead times. Considering all of  
395 the forecasted precipitation events, SMoC forecasts have marginally higher forecast skill at short  
396 lead times and slightly lower forecast skill at long lead times compared to SCC-SS forecasts. In  
397 particular, for forecasts which forecast users care about the most such as forecasts of heavy  
398 events or forecasts at short lead times, SMoC overall performs better than SCC-SS.



399

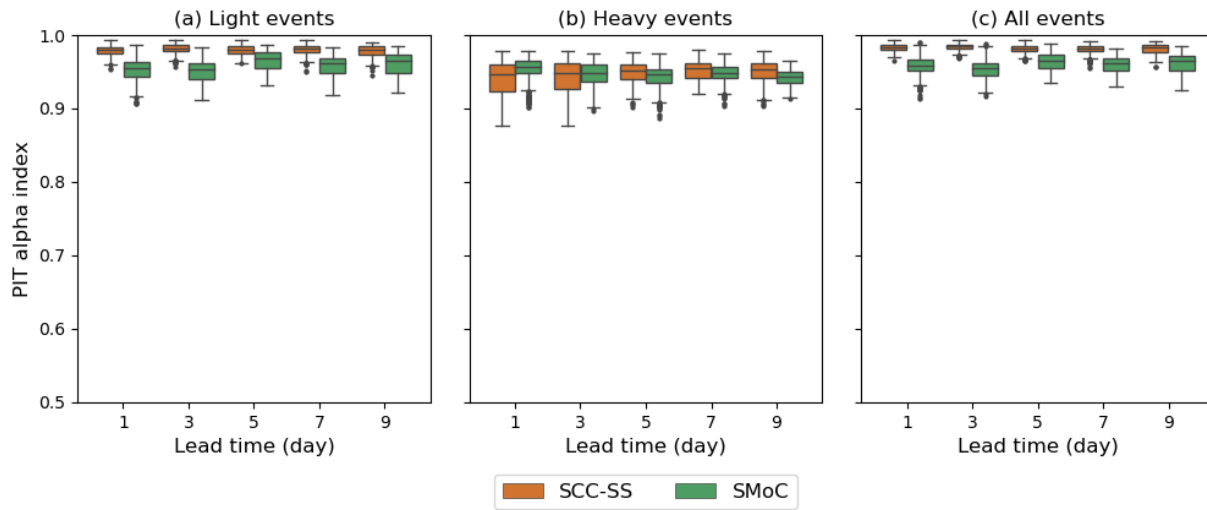
400 Fig. 2. CRPS skill score values of 493 individual grid-cells for SCC-SS forecasts and SMOc forecasts in  
 401 the 3-year forecast period for (a) light events including no precipitation events, (b) heavy events, and (c) all  
 402 events at a set of lead times. For each boxplot, lines of the box portion from bottom to top represent the first  
 403 quartile (Q1, 25th percentile), median (Q2, 50th percentile), and third quartile (Q3, 75th percentile) of the data,  
 404 respectively; lines of the whisker portion from bottom to top represent “minimum” ( $Q1 - 1.5 * (Q3 - Q1)$ ) and  
 405 “maximum” ( $Q3 + 1.5 * (Q3 - Q1)$ ) of the data, respectively; black dots outside the whisker are shown as  
 406 outliers. A positive (negative) CRPS skill score indicates that post-processed forecasts are better (poorer) than  
 407 the referenced climatology ensemble forecasts.

408 To investigate if there is spatial difference of CRPS skill score between SCC-SS forecasts  
 409 and SMOc forecasts, we also plot spatial distributions of CRPS skill score for these two forecasts  
 410 at grid-cell scale, as shown in Fig. S1, Fig. S2, and Fig. S3 (from Supplementary Material S1),  
 411 respectively for light and no precipitation events, heavy events, and all events. According to  
 412 these three figures, similar skill score values can be found in the same grid-cells for SCC-SS  
 413 forecasts and SMOc forecasts, indicating that there is no evident difference in spatial distribution  
 414 of CRPS skill score for these two forecasts.

## 415 2) FORECAST RELIABILITY

416 Results of PIT alpha index for SCC-SS forecasts and SMOc forecasts at individual grid-cells  
 417 are shown in Fig. 3. PIT alpha index values for both of these two forecasts are close to 1,  
 418 indicating that the ensemble spread of these two forecasts at grid-cell scale is overall reliable for  
 419 light and no precipitation events, heavy events, and all events. Overall, SCC-SS forecasts have  
 420 better reliability than SMOc forecasts in terms of PIT alpha index. Similarly, spatial distributions  
 421 of PIT alpha index at grid-cell scale shown in Fig. S4, Fig. S5, and Fig. S6 (from Supplementary

422 Material S1) also suggest that there is no evident difference in spatial distribution of PIT alpha  
423 index for these two forecasts.

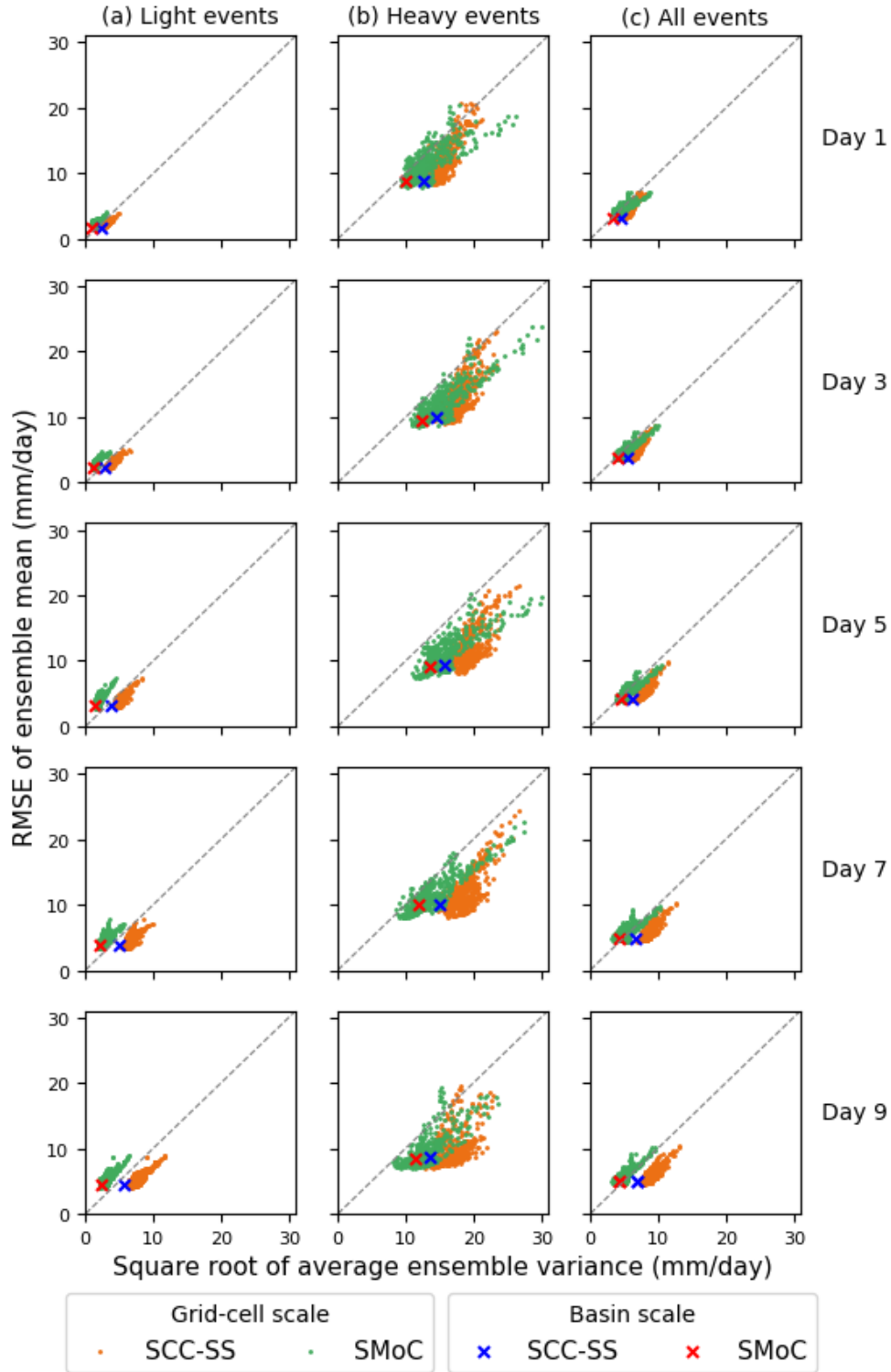


424

425 Fig. 3. PIT alpha index values of 493 individual grid-cells for SCC-SS forecasts and SMOc forecasts in the  
426 3-year forecast period for (a) light events including no precipitation events, (b) heavy events, and (c) all events  
427 at a set of lead times. For each boxplot, lines of the box portion from bottom to top represent the first quartile  
428 (Q1, 25th percentile), median (Q2, 50th percentile), and third quartile (Q3, 75th percentile) of the data,  
429 respectively; lines of the whisker portion from bottom to top represent “minimum” ( $Q1 - 1.5 * (Q3 - Q1)$ ) and  
430 “maximum” ( $Q3 + 1.5 * (Q3 - Q1)$ ) of the data, respectively; black dots outside the whisker are shown as  
431 outliers.

432 Forecast reliability of SCC-SS forecasts and SMOc forecasts at grid-cell scale evaluated  
433 using the spread-error correlation is shown in Fig. 4. The RMSE of ensemble mean and the  
434 square root of average ensemble variance both increase with lead times, indicating larger forecast  
435 uncertainty at longer lead times. For light and no precipitation events, scatter plots of these two  
436 forecasts are found to basically follow the “perfect” diagonal line, indicating good forecast  
437 performance in spread-error correlation. However, it should be noted that SCC-SS forecasts tend  
438 to be over-dispersed (under the diagonal) while SMOc forecasts tend to be under-dispersed (over  
439 the diagonal). For heavy events, scatter plots of these two forecasts cover a wide range of values  
440 of RMSE of ensemble mean and square root of average ensemble variance, indicating large  
441 forecast uncertainty of heavy events. Both of these two forecasts are over-dispersed (under the  
442 diagonal) for heavy events, but the over-dispersion of SMOc forecasts is slighter than SCC-SS  
443 forecasts, especially at long lead times.

444



445

446 Fig. 4. RMSE of ensemble mean forecasts versus square root values of average ensemble variance for (a)  
 447 light events including no precipitation events, (b) heavy events, and (c) all events at a set of lead times, both  
 448 for grid-cell (ensemble forecasts for each of the 493 grid-cells) and basin (basin average ensemble forecasts)  
 449 scales. The spread-error correlations of SCC-SS forecasts and SMOc forecasts are plotted together for an  
 450 intuitive comparison.

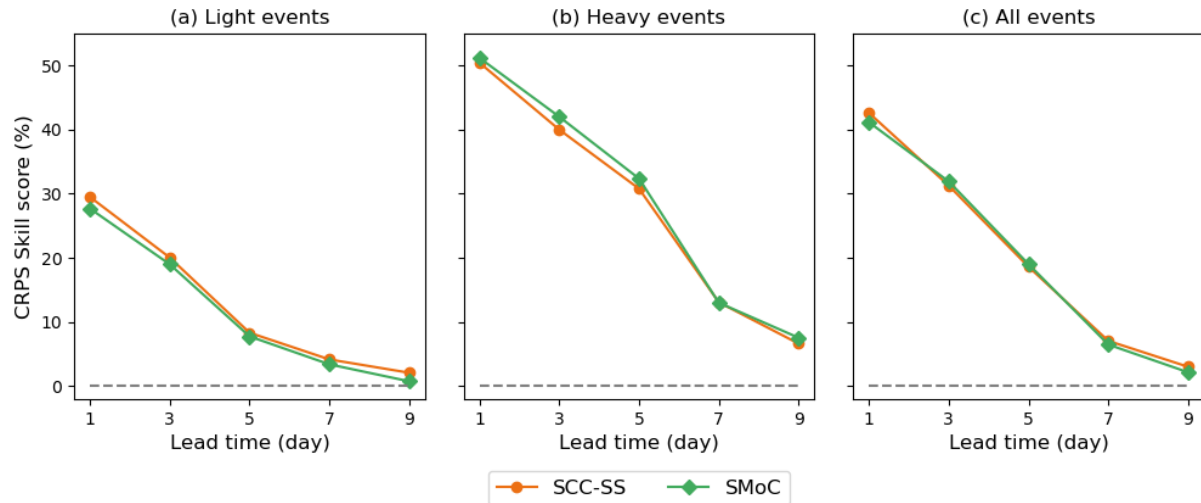
451 The under-dispersion of light and no precipitation events and over-dispersion of heavy events  
452 for SMOc forecasts counter-act to some extent and lead to an almost perfect spread-error  
453 correlation when considering all of the evaluated events. By contrast, the over-dispersion for  
454 both light and no precipitation events and heavy events for SCC-SS forecasts leads to over-  
455 dispersion for all of the evaluated events. Therefore, SMOc forecasts have overall better forecast  
456 reliability in terms of spread-error correlation than SCC-SS forecasts at grid-cell scale.

457 Forecast evaluation of SCC-SS forecasts at grid-cell scale is equivalent to forecast evaluation  
458 of SCC forecasts at individual grid-cells, as the Schaake shuffle does not impact the grid-cell  
459 forecast performance. The SCC model is tuned specifically for each grid-cell and is capable of  
460 producing ensemble forecasts that are of high quality in forecast skill and forecast reliability. The  
461 SMOc model, established for the whole forecast fields, is found to produce grid-cell ensemble  
462 forecasts with similar quality to SCC forecasts. This indicates that the SMOc model, although  
463 calibrating all forecast grid-cells as a whole, does not come at the cost of impaired grid-cell  
464 forecast performance.

#### 465 *b. Forecast evaluation at basin scale*

##### 466 1) FORECAST SKILL

467 Results of CRPS skill score for basin average forecasts of SCC-SS and SMOc are shown in  
468 Fig. 5. Similar to the grid-cell forecasts, CRPS skill score values of these two basin average  
469 forecasts are positive, indicating that they are more skillful than reference climatology forecasts  
470 at basin scale. For light and no precipitation events, SCC-SS forecasts have higher forecast skill  
471 than SMOc forecasts at all lead times, while for heavy events, SMOc forecasts have higher  
472 forecast skill than SCC-SS forecasts. Considering all of the evaluated events, these two basin  
473 average forecasts are overall comparable in forecast skill.

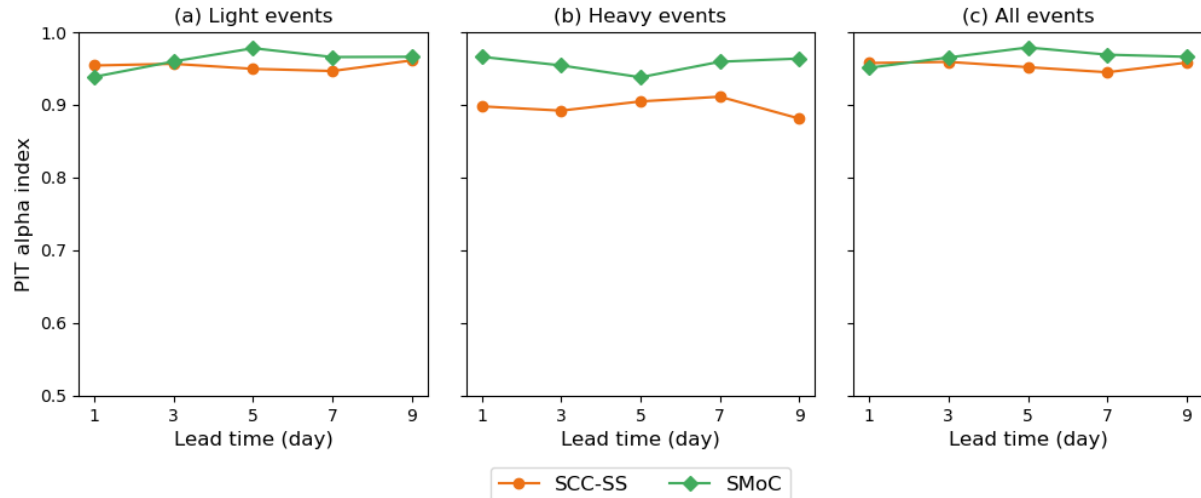


474

475 Fig. 5. CRPS skill score values of basin average forecasts for SCC-SS and SMoC in the 3-year forecast  
 476 period for (a) light events including no precipitation events, (b) heavy events, and (c) all events at a set of lead  
 477 times. A positive (negative) CRPS skill score indicates that post-processed forecasts are better (poorer) than  
 478 the referenced climatology ensemble forecasts.

479 2) FORECAST RELIABILITY

480 Results of PIT alpha index of basin average forecasts for SCC-SS and SMoC are shown in  
 481 Fig. 6. Similar to the grid-cell forecasts, PIT alpha index values of these two basin average  
 482 forecasts are close to 1, indicating that both forecasts are reliable in ensemble spread. Together  
 483 with the PIT alpha index results at grid-cell scale, it can be concluded that these two forecasts are  
 484 reliable at both grid-cell and basin scales, indicating appropriate spatial structures embedded in  
 485 calibrated ensemble members. SMoC forecasts have overall better forecast reliability than SCC-  
 486 SS forecasts in terms of the PIT alpha index. Compared with SCC-SS forecasts, SMoC forecasts  
 487 have smaller PIT alpha index values at grid-cell scale but higher values at basin scale, indicating  
 488 that SMoC forecasts have more appropriate spatial structures.

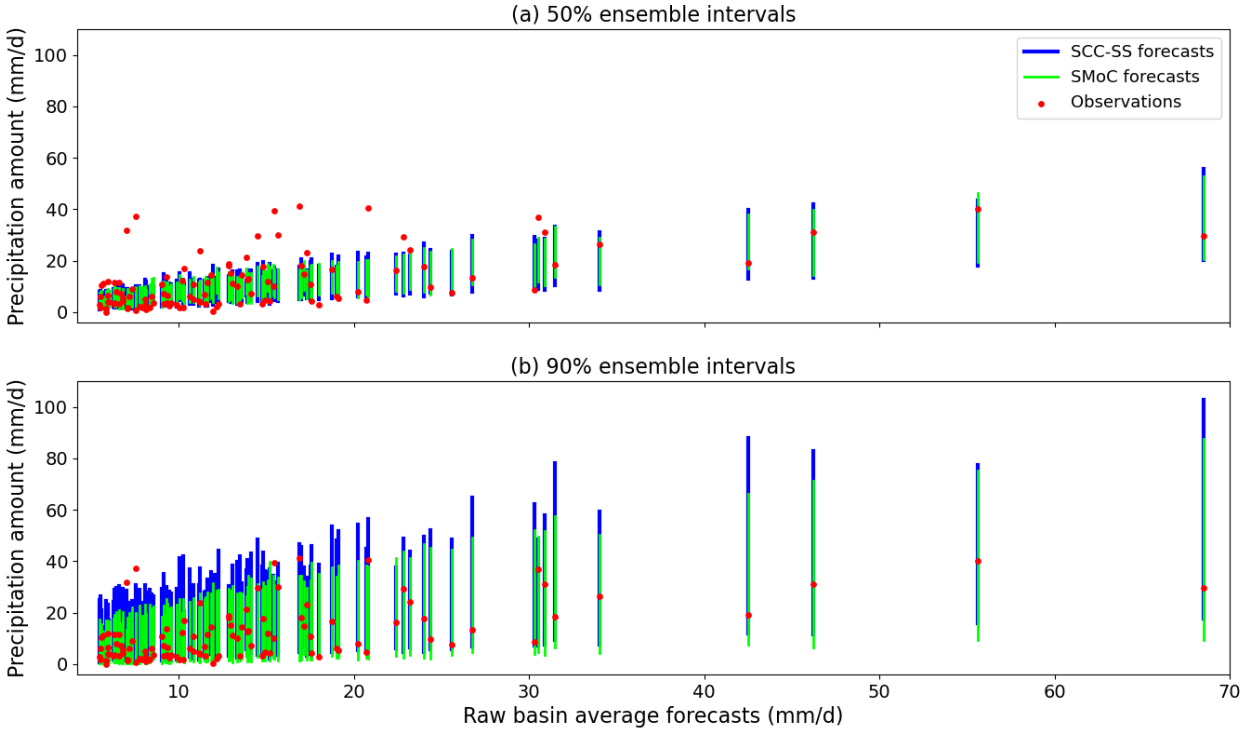


489

490 Fig. 6. PIT alpha index values of basin average forecasts for SCC-SS and SMoC in the 3-year forecast  
 491 period for (a) light events including no precipitation events, (b) heavy events, and (c) all events at a set of lead  
 492 times.

493 In addition, the spread-error correlation results of basin average forecasts for SCC-SS and  
 494 SMoC are shown in Fig. 4. Similar conclusions to the spread-error correlation at grid-cell scale  
 495 can be made. For light and no precipitation events, SCC-SS forecasts are over-dispersed and  
 496 SMoC forecasts are under-dispersed. For heavy events, both forecasts are over-dispersed, but  
 497 SMoC forecasts have a slighter over-dispersion. And overall, SMoC performs better than SCC-  
 498 SS in terms of the spread-error correlation as well as forecast reliability of basin average  
 499 forecasts.

500 To illustrate the ensemble dispersion, we also provide example plots of 50% and 90%  
 501 ensemble intervals for basin average ensemble forecasts of SCC-SS and SMoC at 1 day ahead in  
 502 Fig. 7. We choose to present the dispersion of forecasts of heavy precipitation events as it often  
 503 calls for special attention in the study of statistical post-processing. Fig. 7 clearly shows that  
 504 SMoC forecasts have narrower ensemble dispersion than SCC-SS forecasts, and therefore have  
 505 better performance in terms of the over-dispersion issue. This is in line with the spread-error  
 506 results in Fig. 4 and PIT alpha index results in Fig. 6. Alleviating the over-dispersion issue of  
 507 forecasts of heavy precipitation events is practically meaningful and will boost forecast users'  
 508 confidence in employing ensemble precipitation forecasts.



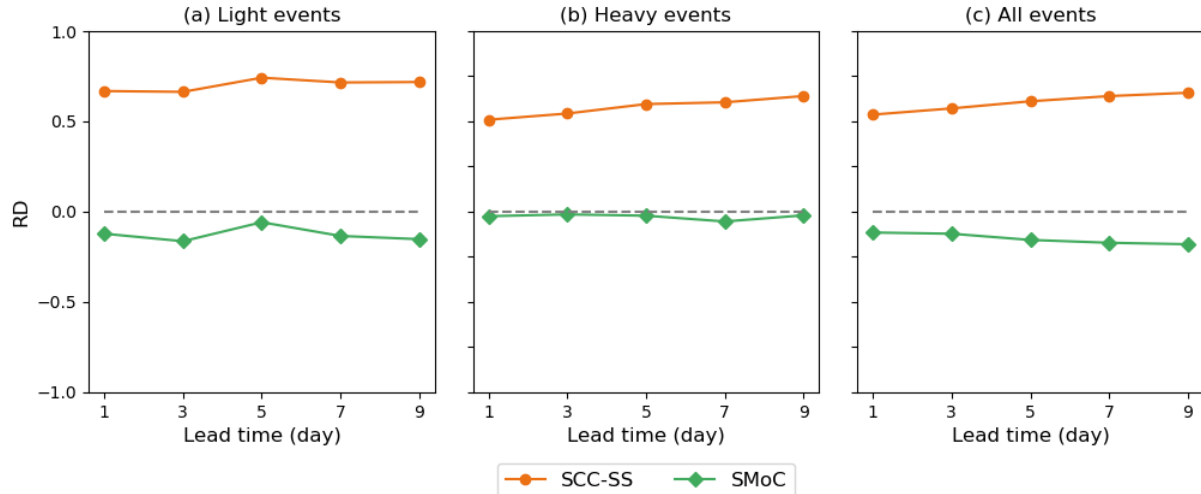
509

510 Fig. 7. 50% and 90% ensemble intervals of SCC-SS and SMOc basin average ensemble forecasts against  
 511 raw basin average forecasts for heavy precipitation events at 1 day ahead during the 3-year forecast period.  
 512 Each vertical line represents an ensemble interval of a basin average ensemble forecast for one heavy  
 513 precipitation event and each red dot represents the corresponding observation.

514 3) FORECAST RUGGEDNESS

515 Results of RD for SCC-SS forecasts and SMOc forecasts are shown in Fig. 8. It is obvious  
 516 that RD values of SMOc forecasts are close to 0. This indicates that the ruggedness of basin  
 517 precipitation values of SMOc forecasts is similar to that of corresponding observations. RD  
 518 values of SCC-SS forecasts, however, are much larger than 0, especially at long lead times. This  
 519 indicates that SCC-SS forecasts have much greater ruggedness of basin precipitation values than  
 520 corresponding observations. Therefore, SMOc forecasts have a clear advantage compared to  
 521 SCC-SS forecasts in terms of the forecast ruggedness.

522



523

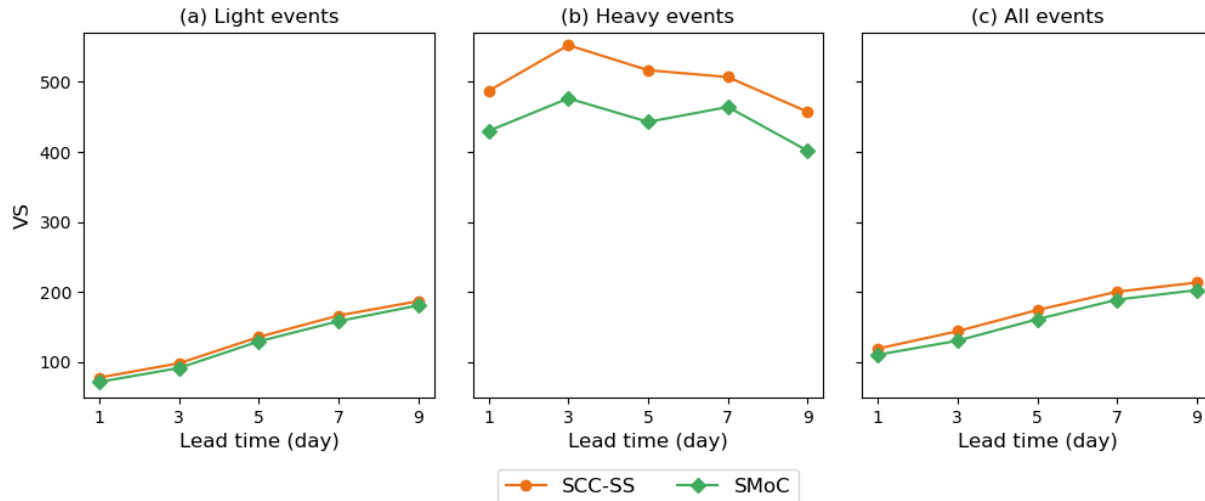
524 Fig. 8. Ruggedness dissimilarity (RD) values of SCC-SS forecasts and SMOc forecasts in the 3-year  
 525 forecast period for (a) light events including no precipitation events, (b) heavy events, and (c) all events at a set  
 526 of lead times. RD ranges from -1 to 1. A value close to 0 is preferred, showing similar ruggedness of forecasts  
 527 to the observations.

528 4) FORECAST CORRELATION

529 Results of VS for SCC-SS forecasts and SMOc forecasts are shown in Fig. 9. For either light  
 530 and no precipitation events or heavy events, SMOc forecasts have smaller VS values than SCC-  
 531 SS forecasts at all investigated lead times. This indicates that, compared with SCC-SS ensemble  
 532 members, the spatial correlation structure of SMOc ensemble members is more similar to that of  
 533 corresponding observations. Therefore, SMOc performs better than SCC-SS in terms of the  
 534 spatial correlation across different grid-cells of calibrated ensemble members.

535

536



537

538 Fig. 9. Variogram score (VS) values of SCC-SS forecasts and SMOc forecasts in the 3-year forecast  
 539 period for (a) light events including no precipitation events, (b) heavy events, and (c) all events at a set of lead  
 540 times. A smaller VS is preferable, indicating the spatial correlation of ensemble members from different grid-  
 541 cells is closer to corresponding observations.

542 To facilitate visualization, we plot a few spatial precipitation fields of SCC-SS forecasts and  
 543 SMOc forecasts for two selected forecast examples at 1 day ahead in Supplementary Material  
 544 S1, as shown in Fig. S10 and Fig. S11 for a light event and a heavy event, respectively.

545 Comparing with corresponding observations, these two forecasts can be intuitively examined for  
 546 intensity and spatial structures of precipitation. For each event, we plot spatial fields of 5  
 547 ensemble members from SCC-SS forecasts and SMOc forecasts, corresponding to 10%, 30%,  
 548 50%, 70%, and 90% quantiles based on basin averages of the 1,000 post-processed ensemble  
 549 members for the whole basin.

550 In Fig. S10, differences in precipitation amounts among adjacent grid-cells can be clearly  
 551 found in SCC-SS ensemble members while not in SMOc ensemble members. This finding is  
 552 especially evident in Fig. S11, where precipitation amounts in grid-cells are not spatially  
 553 continuous across the basin. Consequently, SMOc forecast fields are smoother, and more  
 554 reasonable in spatial structures, compared with SCC-SS forecast fields. These spatial plots  
 555 further strengthen and explain the earlier results that SMOc forecasts perform better than SCC-  
 556 SS forecasts in terms of the forecast ruggedness and forecast correlation.

557 *c. Computational time*

558 The computational time of SCC-SS and SMOc cost to calibrate raw forecast precipitation  
 559 fields and produce ensemble forecasts with spatially correlated structures are shown in Table 1.

560 Considering practical applications, 60 CPUs are employed for SCC-SS and 1 CPU is employed  
 561 for SMoC in computational implementations. The 60 CPUs run in parallel to implement SCC-SS  
 562 for forecasts from each of the 493 grid-cells. Consequently, 8 running cycles of 47 CPUs and 9  
 563 running cycles of the remaining 13 CPUs are needed to finish the post-processing. The 1 CPU  
 564 for SMoC runs for 1 cycle. Despite this, the 60 CPUs for SCC-SS take 103 minutes while the 1  
 565 CPU for SMoC only takes 11 minutes. This indicates that SMoC is about 560 times more  
 566 computationally efficient than SCC-SS.

Approaches	Number of CPU	Computational time (minutes)	
SCC-SS	60	SCC	72
		SS	31
		103	
SMoC	1	11	

567 Table 1. Computational time cost to calibrate forecast precipitation fields of one lead time for all  
 568 precipitation events through Python programming.

569 The reasons for the huge difference between the two computational times are as follows.  
 570 When applying SCC-SS for post-processing, both SCC and the Schaake shuffle are applied  
 571 separately to individual 493 grid-cells. And the computation time will increase linearly with the  
 572 number of the grid-cells. By contrast, SMoC is applied to the whole fields, and therefore does  
 573 not need much time. It should be noted that the speed of SCC-SS and SMoC could be improved  
 574 with further parallel setups. Because we use a leave-one-month-out cross-validation for forecast  
 575 post-processing, we can divide the SMoC implementation into 36 parallel fractions (36 months  
 576 in the 3-year period) and employ 36 CPUs to speed up the computation. Similarly, the leave-one-  
 577 month-out cross-validation allows us to employ more CPUs for SCC. Furthermore, the ordering  
 578 templates for reordering 10 groups of ensemble members allows us to employ more CPUs for the  
 579 Schaake shuffle. After these speed improvements, SMoC is still much more efficient in terms of  
 580 the computational cost. This is a huge advantage of SMoC, especially when applying the post-  
 581 processing to forecasts of large drainage basins that consist of a large number of grid-cells.

## 582 4. Discussions

583 Prior to this study, there was a lack of a comprehensive spatial evaluation of forecast fields in  
 584 the literature. For the first time, a number of metrics are brought together in this study to evaluate  
 585 spatial characteristics of forecast fields, including forecast reliability of basin average forecasts

586 using probability integral transform (PIT) and spread-error correlation (Zhao et al., 2022b),  
587 forecast ruggedness using ruggedness dissimilarity(RD), and forecast correlation using  
588 variogram score (VS) (Scheuerer et al., 2017; Schepen et al., 2020). Results of these forecast  
589 quality measures in different aspects lead to the same conclusion that SMOc forecasts have more  
590 appropriate spatial structures than SCC-SS forecasts, most notably in terms of the forecast  
591 ruggedness. In addition, the spatial evaluation of forecast fields can be further improved using  
592 more detailed spatial information. For example, dividing the study basin into a few sub-basins  
593 and evaluating all sub-basin average ensemble forecasts will provide more valuable insights into  
594 the evaluation of spatial fields.

595 TRI was originally used to measure elevation differences between adjacent grid-cells in Riley  
596 et al. (1999) and was never used for evaluating precipitation fields. In this study, we borrow the  
597 TRI concept to measure the differences of precipitation amounts between adjacent grid-cells.  
598 The ruggedness dissimilarity (RD) used in this study takes into account a usual case where  
599 precipitation amounts at adjacent grid-cells are similar and an unusual case where there is a  
600 heavy precipitation event at one grid-cell while no precipitation at adjacent grid-cells. This  
601 indicates that a large forecast TRI is acceptable as long as it is close to the observation TRI.  
602 Despite this, it's worth noting that when averaged across time and space, the TRI value is likely  
603 to be dominated by a few days where precipitation is particularly heterogenous. It would be  
604 sensible to evaluate the TRI individually for those extreme precipitation events in future studies.  
605 In the comparison of SMOc and SCC-SS, it is found that SMOc has much better performance in  
606 terms of the RD values, showing the superiority of SMOc in producing ensemble forecasts with  
607 appropriate ruggedness features. In particular, this superiority is visually reflected in Fig. S10  
608 and Fig. S11 from Supplementary Material S1, where spatial precipitation distributions of SMOc  
609 forecast fields are clearly smoother than those of SCC-SS forecast fields. This is because the  
610 Schaake shuffle in SCC-SS only considers the ranks of historical records but ignores actual  
611 precipitation amounts of ensemble members. This results in discontinuous values across  
612 neighbouring grid-cells. By contrast, SMOc post-processes all grid-cells as a whole and produces  
613 ensemble forecasts with inbuilt spatial structures across different grid-cells. In practice, the use  
614 of TRI can also be easily extended for evaluating forecast fields of other forecasting variables.

615 In this study, long-term precipitation observations are employed to solve a number of  
616 intractable problems when constructing the SMOc model, the SCC model, and the Schaake

617 shuffle. In the SMOc model, long-term observations are used to derive representative EOF  
618 spatial modes for producing expansion coefficients of forecasts and corresponding observations  
619 in a way that the two sets of expansion coefficients are comparable and can be modelled using  
620 linear regressions. In the SCC model, long-term observations are used to generate representative  
621 climatology of observations for deriving climatology parameters of short-period forecasts so that  
622 both forecasts and observations of the joint probability model can be represented using  
623 seasonally coherent parameters. And in the Schaake shuffle, long-term observations are used to  
624 produce historical ordering templates for reordering the calibrated ensemble members with  
625 spatial structures that are present in the historical records. Therefore, long-term precipitation  
626 observations can be leveraged to improve forecast post-processing, especially when the archived  
627 record of precipitation forecasts is short due to the frequently updated NWP models. In addition,  
628 it would be sensible to also take into account the impact of temporal (e.g., climate change) and  
629 spatial (e.g., topography) patterns (Hannachi et al., 2007; Duan and Duan, 2020; Shao et al.,  
630 2022) which might affect the effectiveness of long-term observations.

631 We employ PIT and spread-error correlation for a comprehensive evaluation of forecast  
632 reliability. Both of these two metrics are widely used, but often separately, to verify reliability of  
633 ensemble forecasts (Grimm and Mass, 2007; Wang and Robertson, 2011; Shao et al., 2020; Guo  
634 et al., 2022; Zhao et al., 2022b). When applying PIT to an ensemble forecast, we pool ensemble  
635 members together with the corresponding observation and sort them in an increasing order to  
636 obtain the order of the observation. For a number of ensemble forecasts, we count the frequency  
637 of observation orders. If the probability of the observation being in any order among the sorted  
638 ensemble members is roughly the same, good forecast reliability is implied. PIT relies on the  
639 ranking of observations with ensemble members, with an appropriate ranking indicating that the  
640 observation is indistinguishable from ensemble members. However, PIT ignores the exact values  
641 of observations and ensemble members and is therefore hardly affected by the heavy tails of  
642 ensemble member values. For example, the PIT results shown in Fig. 3 and Fig. 6 indicate  
643 slightly better forecast reliability compared with the spread-error correlation results shown in  
644 Fig. 4. That's probably due to the lack of capability of PIT to involve the impact of extreme  
645 ensemble member values. Different from PIT, spread-error correlation compares ensemble  
646 spread with the error of ensemble mean. Ensemble mean and ensemble spread are often used to  
647 represent the whole ensemble forecast. And ensemble spread provides forecast uncertainty to

648 predict errors of ensemble mean. If ensemble spread is roughly equal to the error of ensemble  
649 mean based on a lot of ensemble forecasts, forecasts will be considered reliable in ensemble  
650 spread. Spread-error correlation can take into account heavy tails of ensemble forecast values,  
651 however, cannot provide the ranking information of the observation among ensemble members.  
652 In this context, it is sensible to employ both PIT and spread-error correlation to obtain multifold  
653 evaluation results of the forecast reliability, as implemented in this study.

654 The normalization of precipitation data was implemented using log-sinh transformation with  
655 two parameters in the original SCC model (Wang et al., 2019b). Log-sinh transformation was  
656 originally developed to stabilize the variance of streamflow data (Wang et al., 2012b) and was  
657 then extended to normalize precipitation data, showing good performance in several studies  
658 (Robertson et al., 2013; Shrestha et al., 2015; Cattoën et al., 2020; Zhao et al., 2020). Recently,  
659 Li et al. (2019) found that in precipitation forecast post-processing, power transformation had  
660 similar performance to log-sinh transformation in most regions of the Huai River basin. Du et al.  
661 (2022) further confirmed this finding and pointed out that power transformation had only one  
662 parameter and was more suitable for fixing normalization parameters for a whole region as  
663 needed in the SMOc model. Therefore, we adopt power transformation to normalize precipitation  
664 data for this study. The exponents of power transformation for the region are calculated based on  
665 the maximum likelihood estimation method considering precipitation data from all grid-cells and  
666 are therefore capable of keeping the power transformation consistent across the whole region  
667 (Du et al., 2022; Zhao et al., 2022b).

668 In view of different statistics of forecasts of heavy precipitation events and forecasts of light  
669 and no precipitation events (shown in the original SMOc study as well as in Fig. S9 from  
670 Supplementary Material S1), we establish two SMOc models respectively for these two  
671 forecasts. For forecasting heavy precipitation events, SMOc forecasts are found more skillful  
672 than SCC-SS forecasts. This finding offers a remarkable improvement in forecast post-  
673 processing, as forecasting heavy precipitation events is often challenging and draws the most  
674 public attention (Scheuerer and Hamill, 2015a; Zhao et al., 2022a). However, for forecasting  
675 light and no precipitation events, SMOc forecasts are less skillful than SCC-SS forecasts. As  
676 pointed out by Zhao et al. (2022b), there are a large number of no and little precipitation events  
677 in the Brisbane Drainage Basin (typically in most regions), and consequently there are a large  
678 number of small values of EOF expansion coefficients that are not suitable for the use of linear

679 regressions. In our future work, we will refine the SMOc model to improve the calibration of  
680 forecasts of light and no precipitation events by using extended linear regression models (Huffel  
681 and Vandewalle, 1991; Golub and Loan, 1996; Vannitsem, 2009) or some other advanced  
682 techniques such as machine learning methods (Dogulu et al., 2015).

683 When constructing ordering templates for the Schaake shuffle, we select 100 historical  
684 observation records from the 20-year observations. The number of the selected historical records  
685 increases with the decrease in the overall similarity between forecast event dates and historical  
686 event dates. It would be of interest to investigate this trade-off and see if a higher number of  
687 historical records will lead to a better performance of SCC-SS forecasts. It's also sensible to  
688 construct ordering templates based on longer-term observations (more than 20 years) so that the  
689 number of the selected historical records does not come at the cost of the decreased similarity  
690 between forecast dates and historical dates.

691 In the original SMOc study (Zhao et al., 2022b) as well as this study, SMOc is found to  
692 perform well in the Brisbane Drainage Basin. The first 10 EOF modes used to establish the  
693 SMOc model account for 96% of the total data variance. In other words, the rest 483 EOF modes  
694 only account for 4% of the variance information, which actually has minor impacts on the  
695 intensity and spatial patterns of precipitation. Despite this, it remains unclear whether EOF and  
696 SMOc are still so effective in larger regions, especially when forecast post-processing is  
697 implemented at continental scale. Investigating the performance of SMOc at different spatial  
698 scales and accommodating SMOc to work in complex regions where precipitation is quite  
699 spatially heterogeneous will be an essential study for practical applications. Furthermore, for  
700 applications on larger regions like the Australian continent, we could also divide the country to  
701 major drainage divisions and apply the SMOc model to each of the divisions individually.

702 In this study, we compare the performance of SMOc and SCC-SS in calibrating forecast  
703 precipitation fields and producing ensemble forecasts with spatially correlated structures. The  
704 calibration and comparison are implemented separately for a set of different lead times. Apart  
705 from the spatial structures across different grid-cells, ensemble forecast fields are ought to have  
706 appropriate temporal structures across different lead times (Clark et al., 2004; Wu et al., 2018).  
707 For example, the commonly observed precipitation events, whether heavy or light, often occur  
708 over a few consecutive days. NWP forecasts at consecutive lead times are also temporally

709 correlated, as they represent different temporal phases of the evolution of atmospheric simulation  
710 in NWP systems. The second step of the SCC-SS method, i.e., the Schaake shuffle, is capable of  
711 reconstructing not only spatial structures of forecasts from different grid-cells, but also temporal  
712 structures of forecasts from different lead times (Shrestha et al., 2015; Schefzik, 2016; Shrestha  
713 et al., 2020). The SMoC model, however, is currently only capable of building spatial structures  
714 separately for individual lead times. How to enable SMoC to calibrate forecast precipitation  
715 fields at multiple lead times as a whole and produce ensemble forecasts with both spatial  
716 structures and temporal structures needs to be investigated in future study.

## 717 **5. Summary and conclusions**

718 In this paper, the spatial mode-based calibration (SMoC) and the grid-cell by grid-cell post-  
719 processing are evaluated and compared for post-processing forecast precipitation fields with  
720 spatially correlated structures. The seasonally coherent calibration (SCC) model and the Schaake  
721 shuffle (SS) are used as two representative examples of forecast calibration and ensemble  
722 reordering respectively to form the grid-cell by grid-cell post-processing (SCC-SS). In evaluating  
723 SMoC, we also extend it to calibrate forecasts of light and no precipitation events and forecasts  
724 at long lead times. To adapt to different statistical characteristics of forecasts of heavy  
725 precipitation events and forecasts of light and no precipitation events, we establish two SMoC  
726 models separately for these two groups of forecasts.

727 SCC-SS includes two steps, i.e., statistical calibration separately for individual grid-cells  
728 using SCC and reordering of SCC calibrated ensemble members at different grid-cells using the  
729 Schaake shuffle. The main problem of SCC-SS is that Schaake shuffle relies on historical  
730 records and ignores real atmospheric conditions. By contrast, SMoC is a one-step model that  
731 calibrates forecast precipitation fields as a whole and produces ensemble forecasts with inbuilt  
732 spatial structures, and thereby avoids the need for ensemble reordering.

733 SCC-SS and SMoC are evaluated by applying them to precipitation forecasts of a set of lead  
734 times over the Brisbane Drainage Basin during a 3-year period. A number of metrics are  
735 employed to evaluate different aspects of both approaches, including continuous ranked  
736 probability score for forecast skill, probability integral transform and spread-error correlation for  
737 forecast reliability, and terrain ruggedness index and variogram score for spatial structure. SMoC  
738 is found to perform well in calibrating forecasts of both light and no precipitation events and

739 heavy precipitation events, as well as in calibrating forecasts at long lead times. The comparison  
740 of SMOc and SCC-SS is conducted at both grid-cell and basin scales and is summarized as  
741 below.

742 At grid-cell scale, compared with SCC-SS forecasts, SMOc forecasts have similar forecast  
743 skill for light and no precipitation events and higher forecast skill for heavy events. For all  
744 precipitation events in the 3-year period, SMOc forecasts have higher forecast skill at short lead  
745 times and lower forecast skill at long lead times. Overall, SMOc has comparable forecast skill to  
746 SCC-SS. Both SMOc forecasts and SCC-SS forecasts are reliable in ensemble spread. SCC-SS  
747 forecasts have higher PIT alpha index values and SMOc forecasts have better spread-error  
748 correlations.

749 At basin scale, compared with SCC-SS basin average forecasts, SMOc basin average  
750 forecasts have slightly lower forecast skill for light and no precipitation events and higher  
751 forecast skill for heavy events. And overall, these two forecasts have comparable forecast skill.  
752 SMOc basin average forecasts perform better than SCC-SS basin average forecasts in terms of  
753 both PIT alpha index and the spread-error correlation. Besides, SMOc forecasts are superior to  
754 SCC-SS forecasts in spatial structures, according to the results of forecast ruggedness and  
755 forecast correlation.

756 Compared with SCC-SS, i.e., the representative grid-cell by grid-cell post-processing, SMOc  
757 can produce ensemble forecasts that have similar forecast skill, improved forecast reliability, and  
758 clearly better spatial structures. In addition, SMOc is much more computationally efficient,  
759 which is a huge advantage in practical implementations. Therefore, SMOc can be considered a  
760 significant advance in post-processing theory and practice.

761 Despite the superiority of SMOc, there is still room for improvement. Our future research  
762 will focus on the refinement of SMOc to improve the calibration of forecasts of light and no  
763 precipitation events, the accommodation of SMOc to work in large and complex regions, and the  
764 construction of temporal structures of SMOc calibrated ensemble forecasts from different lead  
765 times.

766

767

768 *Acknowledgments.*

769 This work is linked to an Australian Research Council Linkage Project (Grant No.  
770 LP170100922) and a collaborative project (Grant No. TP707466) between the University of  
771 Melbourne and Australian Bureau of Meteorology. Wenyan Wu acknowledges support from the  
772 Australian Research Council via the Discovery Early Career Researcher Award (DE210100117).  
773 We thank the Australian Bureau of Meteorology for supplying the ACCESS-G3 and AWAP  
774 precipitation data. We also thank the National Computational Infrastructure for providing  
775 computation resources to support our work.

776

777 *Data Availability Statement.*

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

781

782

## REFERENCES

783 Baran, S., Lerch, S., 2015: Log-normal distribution based Ensemble Model Output Statistics  
784 models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological*  
785 *Society*, **141**, 2289-2299, <https://doi.org/10.1002/qj.2521>.

786 Baran, S., Möller, A., 2015: Joint probabilistic forecasting of wind speed and temperature  
787 using Bayesian model averaging. *Environmetrics*, **26**, 120-132, <https://doi.org/10.1002/env.2316>.

788 Bellier, J., Bontron, G., Zin, I., 2017: Using Meteorological Analogues for Reordering  
789 Postprocessed Precipitation Ensembles in Hydrological Forecasting. *Water Resources Research*,  
790 **53**, 10085-10107, <https://doi.org/10.1002/2017WR021245>.

791 Bellier, J., Zin, I., Bontron, G., 2018: Generating Coherent Ensemble Forecasts After  
792 Hydrological Postprocessing: Adaptations of ECC-Based Methods. *Water Resources Research*,  
793 **54**, 5741-5762, <https://doi.org/10.1029/2018WR022601>.

794 Borga, M., Boscolo, P., Zanon, F., Sangati, M., 2007: Hydrometeorological Analysis of the  
795 29 August 2003 Flash Flood in the Eastern Italian Alps. *Journal of Hydrometeorology*, **8**, 1049-  
796 1067, <https://doi.org/10.1175/JHM593.1>.

797 Buizza, R., Petroliagis, T., Palmer, T., Barkmeijer, J., Hamrud, M., Hollingsworth, A.,  
798 Simmons, A., Wedi, N., 1998: Impact of model resolution and ensemble size on the performance  
799 of an ensemble prediction system. *Quarterly journal of the royal meteorological society*, **124**,  
800 1935-1960, <https://doi.org/10.1002/qj.49712455008>.

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

804 Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., Wilby, R., 2004: The Schaake  
805 Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and  
806 Temperature Fields. *Journal of Hydrometeorology*, **5**, 243-262, [https://doi.org/10.1175/1525-7541\(2004\)005<0243:TSSAMF>2.0.CO;2](https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2).

808 Diks, C., Panchenko, V., van Dijk, D., 2011: Likelihood-based scoring rules for comparing  
809 density forecasts in tails. *Journal of Econometrics*, **163**, 215-230,  
810 <https://doi.org/10.1016/j.jeconom.2011.04.001>.

811 Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H., Shrestha, D. L., 2015:  
812 Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC  
813 methods and their comparison on contrasting catchments. *Hydrology and Earth System Sciences*,  
814 **19**, 3181-3201, <https://doi.org/10.5194/hess-19-3181-2015>.

815 Du, Y., Wang, Q. J., Wu, W., Yang, Q., 2022: Power transformation of variables for post-  
816 processing precipitation forecasts: Regionally versus locally optimized parameter values. *Journal*  
817 *of Hydrology*, **610**, 127912, <https://doi.org/10.1016/j.jhydrol.2022.127912>.

818 Duan, Q., Duan, A., 2020: The energy and water cycles under climate change. *National*  
819 *Science Review*, **7**, 553-557, <https://doi.org/10.1093/nsr/nwaa003>.

820 Fortin, V., Abaza, M., Anctil, F., Turcotte, R., 2014: Why Should Ensemble Spread Match  
821 the RMSE of the Ensemble Mean? *Journal of Hydrometeorology*, **15**, 1708-1713,  
822 <https://doi.org/10.1175/JHM-D-14-0008.1>.

823 Gneiting, T., Fadoua, B., Raftery, A. E., 2007: Probabilistic Forecasts, Calibration and  
824 Sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 243-  
825 268, <https://doi.org/10.1111/j.1467-9868.2007.00587.x>.

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

829 Gneiting, T., Ranjan, R., 2011: Comparing Density Forecasts Using Threshold- and  
830 Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*, **29**, 411-422,  
831 <https://doi.org/10.1198/jbes.2010.08110>.

832 Golub, G. H., Loan, C. F. V., 1996: *Matrix computations*. Johns Hopkins University Press,  
833 694 pp.

834 Gritmit, E. P., Mass, C. F., 2007: Measuring the ensemble spread-error relationship with a  
835 probabilistic approach: Stochastic ensemble results. *Monthly Weather Review*, **135**, 203-221,  
836 <https://doi.org/10.1175/mwr3262.1>.

837 Guo, D., Wang, Q. J., Ryu, D., Yang, Q., Moller, P., Western, A. W., 2022: An analysis  
838 framework to evaluate irrigation decisions using short-term ensemble weather forecasts.  
839 *Irrigation Science*, **41**, 155–171, <https://doi.org/10.1007/s00271-022-00807-w>.

840 Hannachi, A., Jolliffe, I. T., Stephenson, D. B., 2007: Empirical orthogonal functions and  
841 related techniques in atmospheric science: A review. *International Journal of Climatology*, **27**,  
842 1119-1152, <https://doi.org/10.1002/joc.1499>.

843 Hemri, S., Lisniak, D., Klein, B., 2015: Multivariate postprocessing techniques for  
844 probabilistic hydrological forecasting. *Water Resources Research*, **51**, 7436-7451,  
845 <https://doi.org/10.1002/2014WR016473>.

846 Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for  
847 Ensemble Prediction Systems. *Weather and Forecasting*, **15**, 559-570,  
848 [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).

849 Hu, Y., Schmeits, M. J., van Andel, S. J., Verkade, J. S., Xu, M., Solomatine, D. P., Liang, Z.  
850 M., 2016: A Stratified Sampling Approach for Improved Sampling from a Calibrated Ensemble

851 Forecast Distribution. *Journal of Hydrometeorology*, **17**, 2405-2417,  
852 <https://doi.org/10.1175/jhm-d-15-0205.1>.

853 Huffel, S. V., Vandewalle, J., 1991: *The total least squares problem: computational aspects*  
854 *and analysis*. Frontiers in applied mathematics, 300 pp.

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

858 Lerat, J., Thyer, M., McInerney, D., Kavetski, D., Woldemeskel, F., Pickett-Heaps, C., Shin,  
859 D., Feikema, P., 2020: A robust approach for calibrating a daily rainfall-runoff model to monthly  
860 streamflow data. *Journal of Hydrology*, **591**, 125129,  
861 <https://doi.org/10.1016/j.jhydrol.2020.125129>.

862 Li, W., Duan, Q., Wang, Q. J., 2019: Factors Influencing the Performance of Regression-  
863 Based Statistical Postprocessing Models for Short-Term Precipitation Forecasts. *Weather and*  
864 *Forecasting*, **34**, 2067-2084, <https://doi.org/10.1175/WAF-D-19-0121.1>.

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

868 Li, W., Pan, B., Xia, J., Duan, Q., 2022: Convolutional neural network-based statistical post-  
869 processing of ensemble precipitation forecasts. *Journal of Hydrology*, **605**, 127301,  
870 <https://doi.org/10.1016/j.jhydrol.2021.127301>.

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

874 Li, Y., Wang, Q. J., He, H., Wu, Z., Lu, G., 2020b: A method to extend temporal coverage of  
875 high quality precipitation datasets by calibrating reanalysis estimates. *Journal of Hydrology*, **581**,  
876 124355, <https://doi.org/10.1016/j.jhydrol.2019.124355>.

877 Nelder, J. A., Mead, R., 1965: A Simplex Method for Function Minimization. *The Computer*  
878 *Journal*, **7**, 308-313, <https://doi.org/10.1093/comjnl/7.4.308>.

879 Pechlivanidis, I. G., Crochemore, L., Rosberg, J., Bosshard, T., 2020: What Are the Key  
880 Drivers Controlling the Quality of Seasonal Streamflow Forecasts? *Water Resources Research*,  
881 **56**, e2019WR026987, <https://doi.org/10.1029/2019WR026987>.

882 Peng, Z., Wang, Q. J., Bennett, J. C., Schepen, A., Pappenberger, F., Pokhrel, P., Wang, Z.,  
883 2014: Statistical calibration and bridging of ECMWF System4 outputs for forecasting seasonal  
884 precipitation over China. *Journal of Geophysical Research: Atmospheres*, **119**, 7116-7135,  
885 <https://doi.org/10.1002/2013JD021162>.

886 Radanovics, S., Vidal, J.-P., Sauquet, E., 2018: Spatial Verification of Ensemble  
887 Precipitation: An Ensemble Version of SAL. *Weather and Forecasting*, **33**, 1001-1020,  
888 <https://doi.org/10.1175/waf-d-17-0162.1>.

889 Raftery, A. E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005: Using Bayesian Model  
890 Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, **133**, 1155-1174,  
891 <https://doi.org/10.1175/mwr2906.1>.

892 Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S. W., 2010: Understanding  
893 predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural  
894 errors. *Water Resources Research*, **46**, W05521, <https://doi.org/10.1029/2009WR008328>.

895 Riley, S. J., DeGloria, S. D., Elliot, R., 1999: A terrain ruggedness index that quantifies  
896 topographic heterogeneity. *Intermountain Journal of Sciences*, **5**, 23-27.

897 Robertson, D. E., Shrestha, D. L., Wang, Q. J., 2013: Post-processing rainfall forecasts from  
898 numerical weather prediction models for short-term streamflow forecasting. *Hydrology and  
899 Earth System Sciences*, **17**, 3587-3603, <https://doi.org/10.5194/hess-17-3587-2013>.

900 Roulston, M. S., Smith, L. A., 2003: Combining dynamical and statistical ensembles. *Tellus  
901 A*, **55**, 16-30, <https://doi.org/10.1034/j.1600-0870.2003.201378.x>.

902 Schefzik, R., 2016: A Similarity-Based Implementation of the Schaake Shuffle. *Monthly  
903 Weather Review*, **144**, 1909-1921, <https://doi.org/10.1175/MWR-D-15-0227.1>.

904 Schefzik, R., 2017: Ensemble calibration with preserved correlations: unifying and  
905 comparing ensemble copula coupling and member-by-member postprocessing. *Quarterly  
906 Journal of the Royal Meteorological Society*, **143**, 999-1008, <https://doi.org/10.1002/qj.2984>.

907 Schefzik, R., Thorarinsdottir, T. L., Gneiting, T., 2013: Uncertainty Quantification in  
908 Complex Simulation Models Using Ensemble Copula Coupling. *Statistical Science*, **28**, 616-640,  
909 <https://doi.org/10.1214/13-STS443>.

910 Schepen, A., Everingham, Y., Wang, Q. J., 2020: On the Joint Calibration of Multivariate  
911 Seasonal Climate Forecasts from GCMs. *Monthly Weather Review*, **148**, 437-456,  
912 <https://doi.org/10.1175/MWR-D-19-0046.1>.

913 Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using Ensemble  
914 Model Output Statistics. *Quarterly Journal of the Royal Meteorological Society*, **140**, 1086-  
915 1096, <https://doi.org/10.1002/qj.2183>.

916 Scheuerer, M., Hamill, T. M., 2015a: Statistical Postprocessing of Ensemble Precipitation  
917 Forecasts by Fitting Censored, Shifted Gamma Distributions. *Monthly Weather Review*, **143**,  
918 4578-4596, <https://doi.org/10.1175/mwr-d-15-0061.1>.

919 Scheuerer, M., Hamill, T. M., 2015b: Variogram-Based Proper Scoring Rules for  
920 Probabilistic Forecasts of Multivariate Quantities. *Monthly Weather Review*, **143**, 1321-1334,  
921 <https://doi.org/10.1175/MWR-D-14-00269.1>.

922 Scheuerer, M., Hamill, T. M., 2018: Generating Calibrated Ensembles of Physically  
923 Realistic, High-Resolution Precipitation Forecast Fields Based on GEFS Model Output. *Journal*  
924 *of Hydrometeorology*, **19**, 1651-1670, <https://doi.org/10.1175/JHM-D-18-0067.1>.

925 Scheuerer, M., Hamill, T. M., Whitin, B., He, M., Henkel, A., 2017: A method for  
926 preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal  
927 forecast fields of temperature and precipitation. *Water Resources Research*, **53**, 3029-3046,  
928 <https://doi.org/10.1002/2016WR020133>.

929 Shao, Y., Wang, Q. J., Schepen, A., Ryu, D., 2020: Embedding trend into seasonal  
930 temperature forecasts through statistical calibration of GCM outputs. *International Journal of*  
931 *Climatology*, **41**, E1553-E1565, <https://doi.org/10.1002/joc.6788>.

932 Shao, Y., Wang, Q. J., Schepen, A., Ryu, D., Pappenberger, F., 2022: Improved Trend-  
933 Aware Postprocessing of GCM Seasonal Precipitation Forecasts. *Journal of Hydrometeorology*,  
934 **23**, 25-37, <https://doi.org/10.1175/JHM-D-21-0099.1>.

935 Shrestha, D. L., Robertson, D. E., Bennett, J. C., Wang, Q. J., 2015: Improving Precipitation  
936 Forecasts by Generating Ensembles through Postprocessing. *Monthly Weather Review*, **143**,  
937 3642-3663, <https://doi.org/10.1175/MWR-D-14-00329.1>.

938 Shrestha, D. L., Robertson, D. E., Bennett, J. C., Wang, Q. J., 2020: Using the Schaake  
939 shuffle when calibrating ensemble means can be problematic. *Journal of Hydrology*, **587**,  
940 124991, <https://doi.org/10.1016/j.jhydrol.2020.124991>.

941 Sloughter, J. M. L., Raftery, A. E., Gneiting, T., Fraley, C., 2007: Probabilistic Quantitative  
942 Precipitation Forecasting Using Bayesian Model Averaging. *Monthly Weather Review*, **135**,  
943 3209-3220, <https://doi.org/10.1175/mwr3441.1>.

944 Straaten, C. v., Whan, K., Schmeits, M., 2018: Statistical Postprocessing and Multivariate  
945 Structuring of High-Resolution Ensemble Precipitation Forecasts. *Journal of Hydrometeorology*,  
946 **19**, 1815-1833, <https://doi.org/10.1175/JHM-D-18-0105.1>.

947 Toth, Z., Kalnay, E., 1993: Ensemble Forecasting at NMC: The Generation of Perturbations.  
948 *Bulletin of the American Meteorological Society*, **74**, 2317-2330, [https://doi.org/10.1175/1520-0477\(1993\)074<2317:EFANTG>2.0.CO;2](https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2).

950 Van Schaeybroeck, B., Vannitsem, S., 2016: A Probabilistic Approach to Forecast the  
951 Uncertainty with Ensemble Spread. *Monthly Weather Review*, **144**, 451-468,  
952 <https://doi.org/10.1175/mwr-d-14-00312.1>.

953 Vannitsem, S., 2009: A unified linear Model Output Statistics scheme for both deterministic  
954 and ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **135**, 1801-1815,  
955 <https://doi.org/10.1002/qj.491>.

956 Wang, Q. J., Bennett, J. C., Robertson, D. E., Li, M., 2020: A Data Censoring Approach for  
957 Predictive Error Modeling of Flow in Ephemeral Rivers. *Water Resources Research*, **56**,  
958 e2019WR026128, <https://doi.org/10.1029/2019WR026128>.

959 Wang, Q. J., Robertson, D. E., 2011: Multisite probabilistic forecasting of seasonal flows for  
960 streams with zero value occurrences. *Water Resources Research*, **47**, W02546,  
961 <https://doi.org/10.1029/2010WR009333>.

962 Wang, Q. J., Robertson, D. E., Chiew, F. H. S., 2009: A Bayesian joint probability modeling  
963 approach for seasonal forecasting of streamflows at multiple sites. *Water Resources Research*,  
964 **45**, W05407, <https://doi.org/10.1029/2008WR007355>.

965 Wang, Q. J., Schepen, A., Robertson, D. E., 2012a: Merging Seasonal Rainfall Forecasts  
966 from Multiple Statistical Models through Bayesian Model Averaging. *Journal of Climate*, **25**,  
967 5524-5537, <https://doi.org/10.1175/JCLI-D-11-00386.1>.

968 Wang, Q. J., Shao, Y., Song, Y., Schepen, A., Robertson, D. E., Ryu, D., Pappenberger, F.,  
969 2019a: An evaluation of ECMWF SEAS5 seasonal climate forecasts for Australia using a new  
970 forecast calibration algorithm. *Environmental Modelling & Software*, **122**, 104550,  
971 <https://doi.org/10.1016/j.envsoft.2019.104550>.

972 Wang, Q. J., Shrestha, D. L., Robertson, D. E., Pokhrel, P., 2012b: A log-sinh transformation  
973 for data normalization and variance stabilization. *Water Resources Research*, **48**, W05514,  
974 <https://doi.org/10.1029/2011WR010973>.

975 Wang, Q. J., Zhao, T., Yang, Q., Robertson, D., 2019b: A Seasonally Coherent Calibration  
976 (SCC) Model for Postprocessing Numerical Weather Predictions. *Monthly Weather Review*, **147**,  
977 3633-3647, <https://doi.org/10.1175/mwr-d-19-0108.1>.

978 Wernli, H., Paulat, M., Hagen, M., Frei, C., 2008: SAL—A Novel Quality Measure for the  
979 Verification of Quantitative Precipitation Forecasts. *Monthly Weather Review*, **136**, 4470-4487,  
980 <https://doi.org/10.1175/2008mwr2415.1>.

981 Whitaker, J. S., Lough, A. F., 1998: The Relationship between Ensemble Spread and  
982 Ensemble Mean Skill. *Monthly Weather Review*, **126**, 3292-3302, [https://doi.org/10.1175/1520-0493\(1998\)126<3292:TRBESA>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2).

984 Wu, L., Zhang, Y., Adams, T., Lee, H., Liu, Y., Schaake, J., 2018: Comparative Evaluation  
985 of Three Schaake Shuffle Schemes in Postprocessing GEFS Precipitation Ensemble Forecasts.  
986 *Journal of Hydrometeorology*, **19**, 575-598, <https://doi.org/10.1175/JHM-D-17-0054.1>.

987 Yang, Q., Wang, Q. J., Hakala, K., 2021: Achieving effective calibration of precipitation  
988 forecasts over a continental scale. *Journal of Hydrology: Regional Studies*, **35**, 100818,  
989 <https://doi.org/10.1016/j.ejrh.2021.100818>.

990 Yang, Q., Wang, Q. J., Hakala, K., 2022a: Calibrating anomalies improves forecasting of  
991 daily reference crop evapotranspiration. *Journal of Hydrology*, **610**, 128009,  
992 <https://doi.org/10.1016/j.jhydrol.2022.128009>.

993 Yang, Q., Wang, Q. J., Western, A. W., Wu, W., Shao, Y., Hakala, K., 2022b:  
994 Reconstructing climate trends adds skills to seasonal reference crop evapotranspiration  
995 forecasting. *Hydrology and Earth System Sciences*, **26**, 941-954, [https://doi.org/10.5194/hess-26-](https://doi.org/10.5194/hess-26-941-2022)  
996 [941-2022](https://doi.org/10.5194/hess-26-941-2022).

997 Zhang, Y., Wu, L., Scheuerer, M., Schaake, J., Kongoli, C., 2017: Comparison of  
998 Probabilistic Quantitative Precipitation Forecasts from Two Postprocessing Mechanisms.  
999 *Journal of Hydrometeorology*, **18**, 2873-2891, <https://doi.org/10.1175/jhm-d-16-0293.1>.

1000 Zhao, P., Wang, Q. J., Wu, W., Yang, Q., 2020: Which precipitation forecasts to use?  
1001 Deterministic versus coarser-resolution ensemble NWP models. *Quarterly Journal of the Royal*  
1002 *Meteorological Society*, **147**, 900-913, <https://doi.org/10.1002/qj.3952>.

1003 Zhao, P., Wang, Q. J., Wu, W., Yang, Q., 2022a: Extending a joint probability modelling  
1004 approach for post-processing ensemble precipitation forecasts from numerical weather prediction  
1005 models. *Journal of Hydrology*, **605**, 127285, <https://doi.org/10.1016/j.jhydrol.2021.127285>.

1006 Zhao, P., Wang, Q. J., Wu, W., Yang, Q., 2022b: Spatial mode-based calibration (SMoC) of  
1007 forecast precipitation fields from numerical weather prediction models. *Journal of Hydrology*,  
1008 **613**, 128432, <https://doi.org/10.1016/j.jhydrol.2022.128432>.

1009 Zhao, T., Wang, Q. J., Bennett, J. C., Robertson, D. E., Shao, Q., Zhao, J., 2015: Quantifying  
1010 predictive uncertainty of streamflow forecasts based on a Bayesian joint probability model.  
1011 *Journal of Hydrology*, **528**, 329-340, <https://doi.org/10.1016/j.jhydrol.2015.06.043>.