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

Pengcheng Zhao,[a,b] Quan J. Wang,[a] Wenyan Wu,[a] and Qichun Yang[a, c]

[a] *Department of Infrastructure Engineering, The University of Melbourne, Parkville, Australia*

[b] *China International Engineering Consulting Corporation, Beijing, China*

[c] *Thrust of Earth, Ocean and Atmospheric Sciences, Hong Kong University of Science and Technology (GZ), China*

Corresponding author: Pengcheng Zhao, 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

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

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

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

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

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

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

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

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

2. Data and methods

a. Data

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

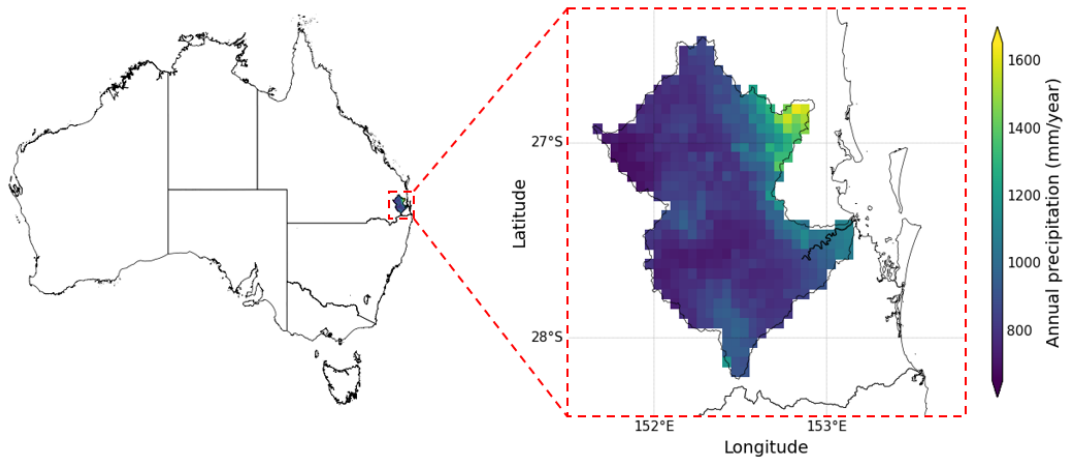

Fig. 1. Geographical location (left figure) and annual precipitation map (right figure) of the Brisbane Drainage Basin.

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

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

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

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

b. The SCC-SS model

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

1) THE SCC MODEL

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

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

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

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

2) THE SCHAAKE SHUFFLE

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

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

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

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

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

c. The SMoC model

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

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

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

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

d. Forecast evaluation

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

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

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

1) FORECAST EVALUATION AT GRID-CELL SCALE

(i) Forecast skill

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

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

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

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

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

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

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

(ii) Forecast reliability

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

$$\pi(t) = F[t, x = y(t)] \tag{3}$$

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

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

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

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

2) FORECAST EVALUATION AT BASIN SCALE

(i) Forecast skill

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

(ii) Forecast reliability

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

(iii) Forecast ruggedness

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

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

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

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

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

(iv) Forecast correlation

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

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

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

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

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

3) COMPUTATIONAL TIME

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

3. Results

a. Forecast evaluation at grid-cell scale

1) FORECAST SKILL

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

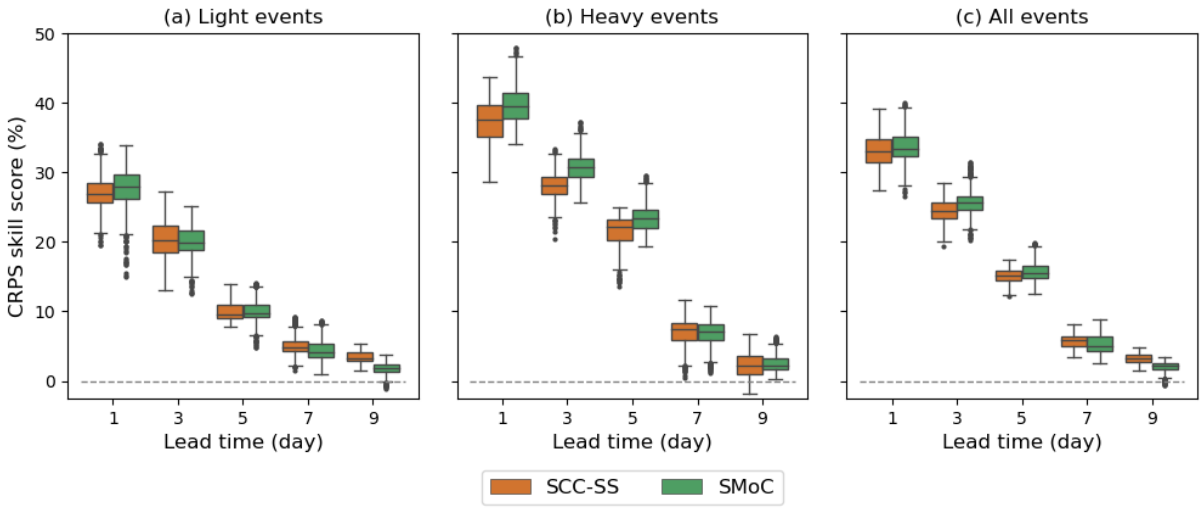

Fig. 2. CRPS skill score values of 493 individual grid-cells for SCC-SS forecasts and SMoC forecasts in the 3-year forecast period for (a) light events including no precipitation events, (b) heavy events, and (c) all events at a set of lead times. For each boxplot, lines of the box portion from bottom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

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

2) FORECAST RELIABILITY

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

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

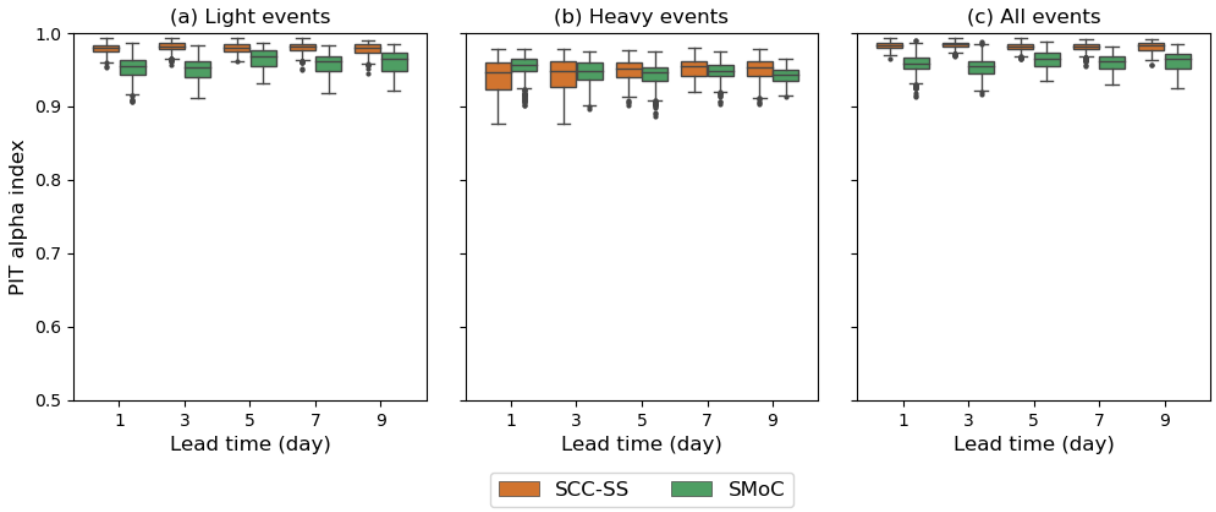

Fig. 3. PIT alpha index values of 493 individual grid-cells for SCC-SS forecasts and SMoC forecasts in the 3-year forecast period for (a) light events including no precipitation events, (b) heavy events, and (c) all events at a set of lead times. For each boxplot, lines of the box portion from bottom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

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

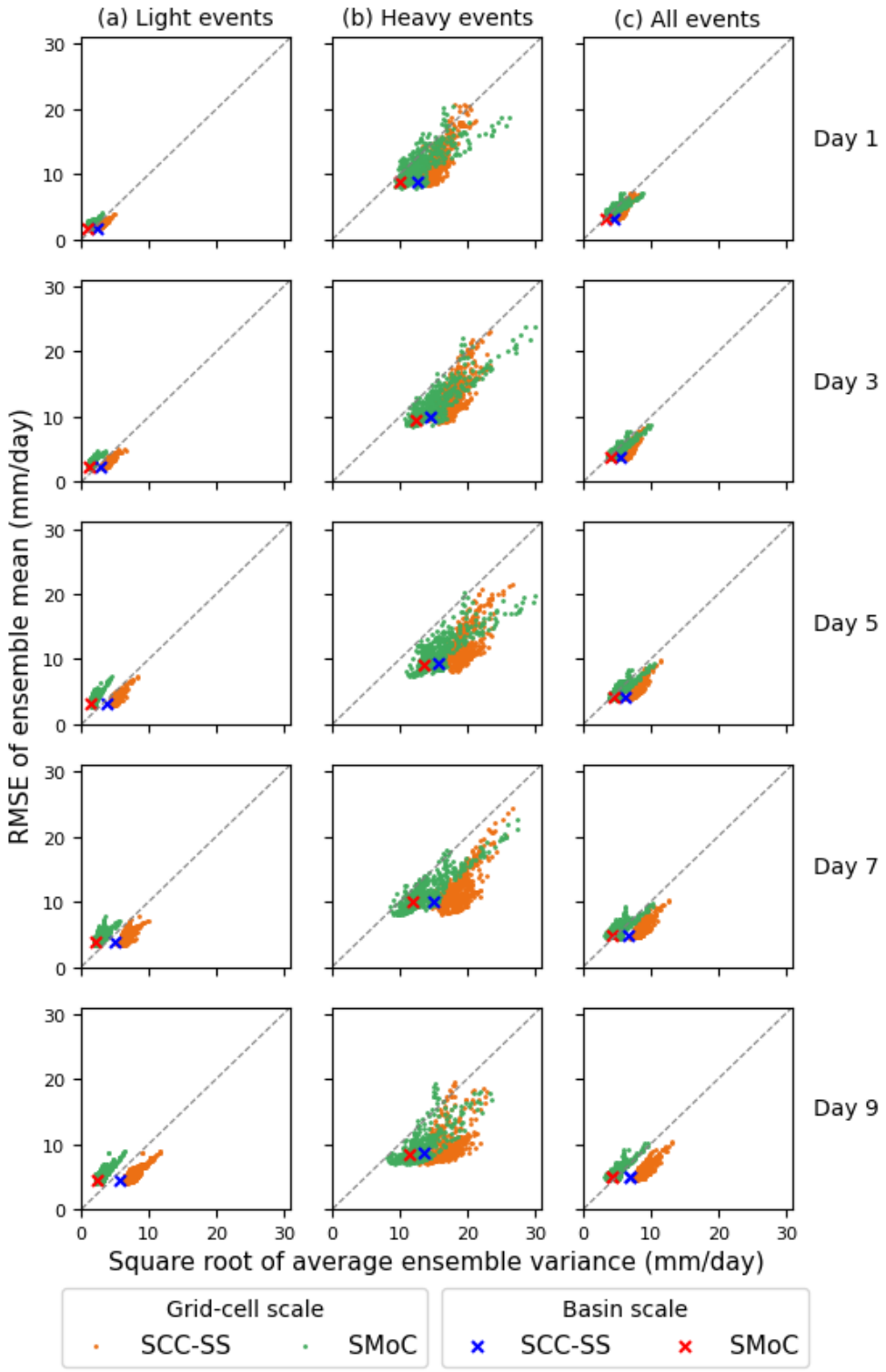

Fig. 4. RMSE of ensemble mean forecasts versus square root values of average ensemble variance for (a) light events including no precipitation events, (b) heavy events, and (c) all events at a set of lead times, both for grid-cell (ensemble forecasts for each of the 493 grid-cells) and basin (basin average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

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

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

b. Forecast evaluation at basin scale

1) FORECAST SKILL

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

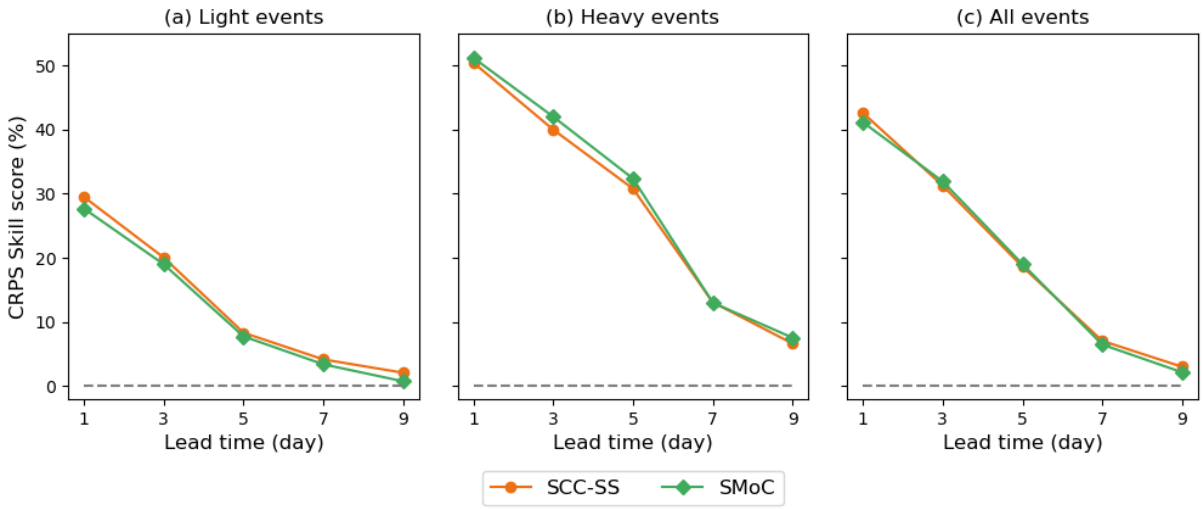

Fig. 5. CRPS skill score values of basin average forecasts for SCC-SS and SMoC in the 3-year forecast period for (a) light events including no precipitation events, (b) heavy events, and (c) all events at a set of lead times. A positive (negative) CRPS skill score indicates that post-processed forecasts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

2) FORECAST RELIABILITY

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

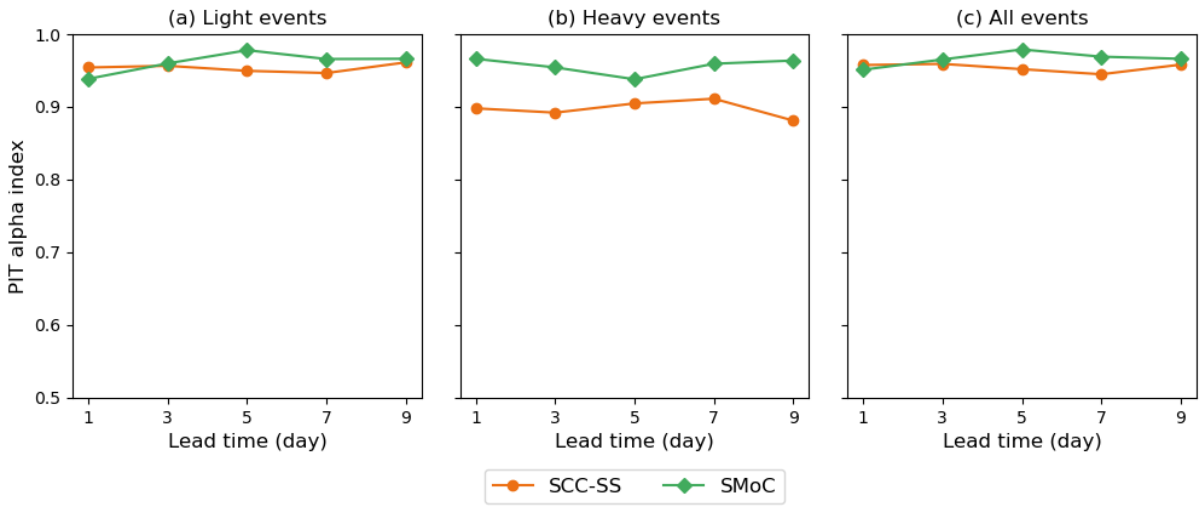

Fig. 6. PIT alpha index values of basin average forecasts for SCC-SS and SMoC in the 3-year forecast period for (a) light events including no precipitation events, (b) heavy events, and (c) all events at a set of lead times.

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

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

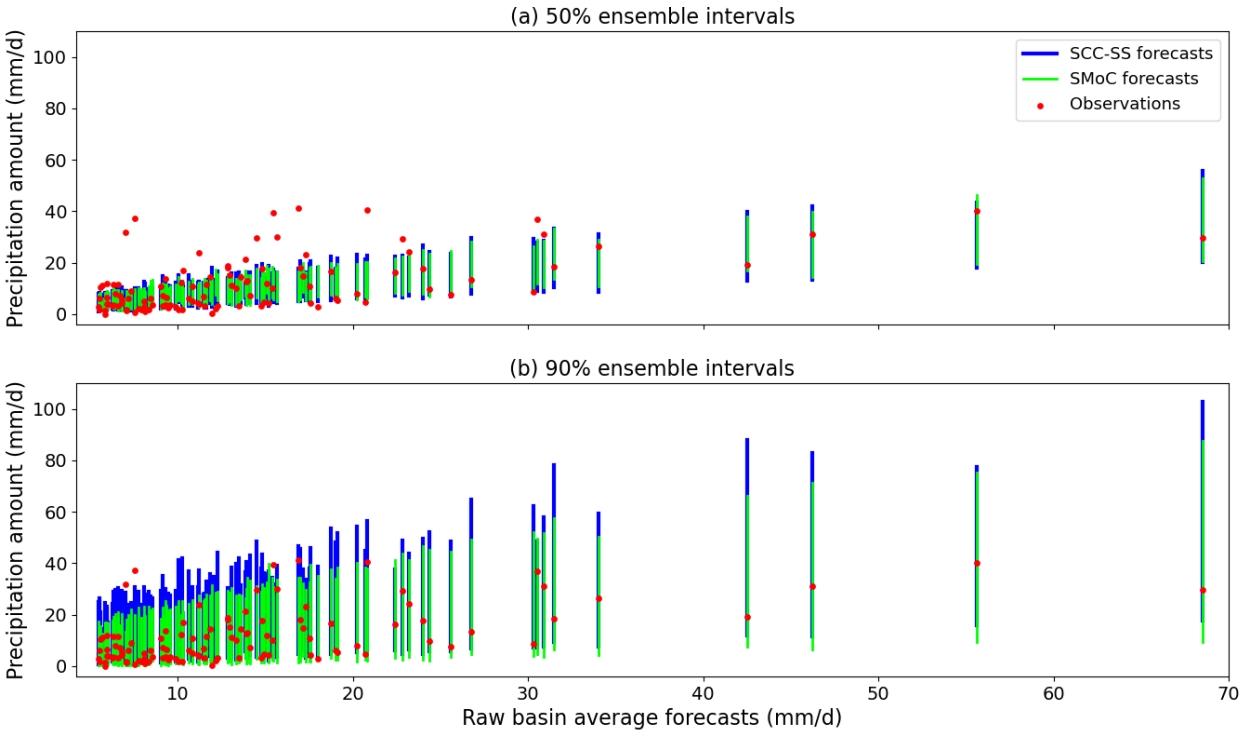

Fig. 7. 50% and 90% ensemble intervals of SCC-SS and SMoC basin average ensemble forecasts against raw basin average forecasts for heavy precipitation events at 1 day ahead during the 3-year forecast period. Each vertical line represents an ensemble interval of a basin average ensemble forecast for one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

3) FORECAST RUGGEDNESS

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

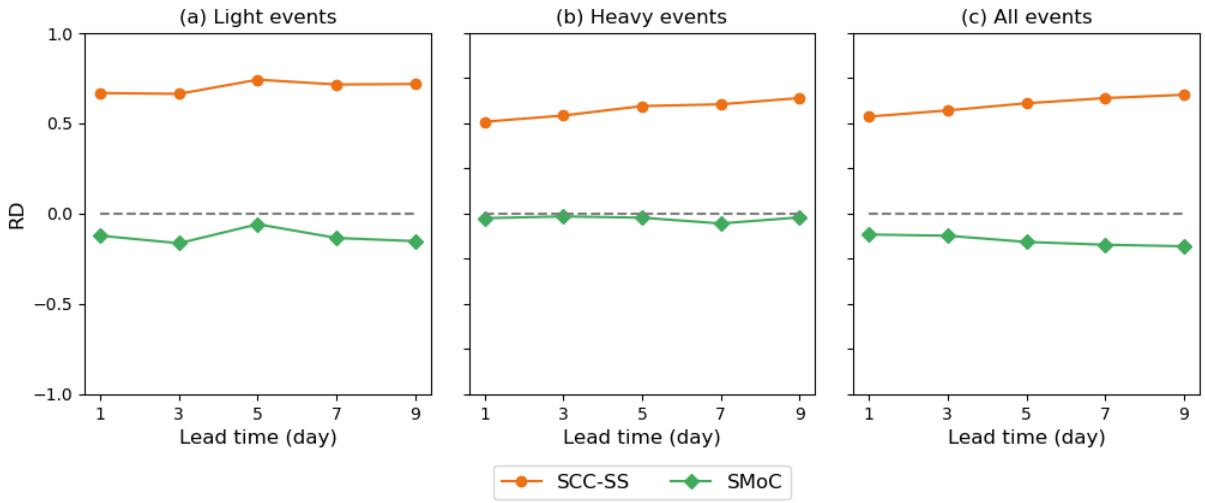

Fig. 8. Ruggedness dissimilarity (RD) values of SCC-SS forecasts and SMoC forecasts in the 3-year forecast period for (a) light events including no precipitation events, (b) heavy events, and (c) all events at a set of lead times. RD ranges from -1 to 1. A value close to 0 is preferred, showing similar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

4) FORECAST CORRELATION

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

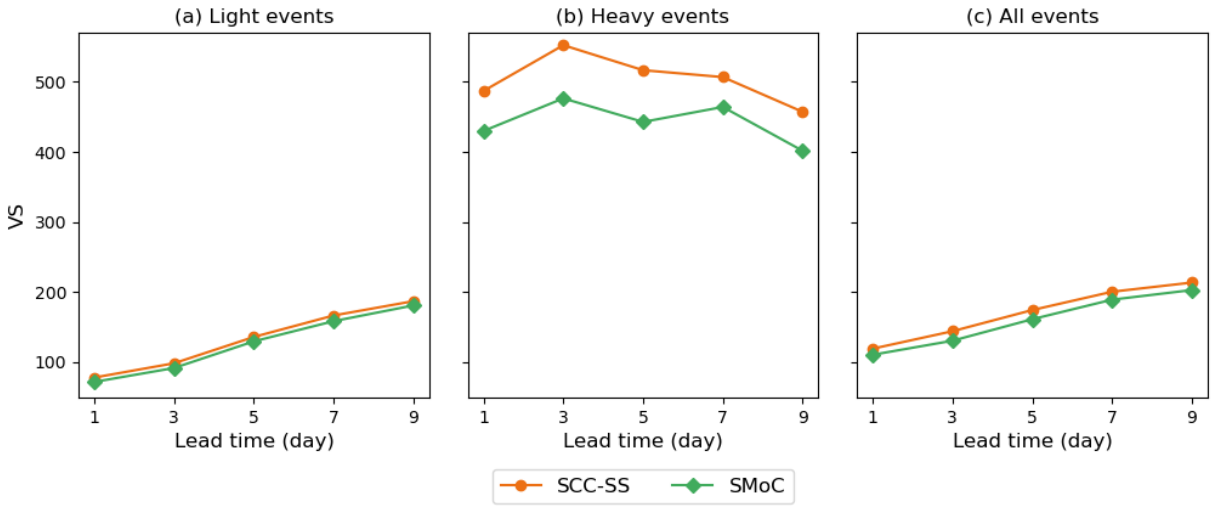

Fig. 9. Variogram score (VS) values of SCC-SS forecasts and SMoC forecasts in the 3-year forecast period for (a) light events including no precipitation events, (b) heavy events, and (c) all events at a set of lead times. A smaller VS is preferable, indicating the spatial correlation of ensemble members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

To facilitate visualization, we plot a few spatial precipitation fields of SCC-SS forecasts and SMoC forecasts for two selected forecast examples at 1 day ahead in Supplementary Material S1, as shown in Fig. S10 and Fig. S11 for a light event and a heavy event, respectively. Comparing with corresponding observations, these two forecasts can be intuitively examined for intensity and spatial structures of precipitation. For each event, we plot spatial fields of 5 ensemble members from SCC-SS forecasts and SMoC forecasts, corresponding to 10%, 30%, 50%, 70%, and 90% quantiles based on basin averages of the 1,000 post-processed ensemble members for the whole basin.

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

c. Computational time

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

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

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

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

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

4. Discussions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5. Summary and conclusions

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

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

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

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

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

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

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

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

Acknowledgments.

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

Data Availability Statement.

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Nelder, J. A., Mead, R., 1965: A Simplex Method for Function Minimization. *The Computer 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.

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

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

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

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

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

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

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

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

Wang, Q. J., Robertson, D. E., 2011: Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resources Research*, **47**, W02546, 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., Loughe, 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-
983 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.

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

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

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

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

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

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

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