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

Title:
Which precipitation forecasts to use? Deterministic versus coarser-resolution ensemble NWP models

Date:
2021-01-01

Citation:
Zhao, P., Wang, Q. J., Wu, W. & Yang, Q. (2021). Which precipitation forecasts to use? Deterministic versus coarser-resolution ensemble NWP models. Quarterly Journal of the Royal Meteorological Society, 147 (735), pp.900-913. https://doi.org/10.1002/qj.3952.

Persistent Link:
https://hdl.handle.net/11343/276783

Zhao Pengcheng (Orcid ID: 0000-0002-8830-533X)
Wang Quan (Orcid ID: 0000-0002-8787-2738)

Which precipitation forecasts to use? Deterministic versus coarser-resolution ensemble NWP models

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

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

*: Corresponding author

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

Telephone number: +61 0452199537

This work is partially funded by a collaborative project (TP707466) between the University of Melbourne and Australian Bureau of Meteorology and by an Australian Research Council Linkage Project (LP170100922).

November 2020

This revised manuscript will be submitted to Quarterly Journal of the Royal Meteorological Society

ABSTRACT

Deterministic numerical weather prediction (NWP) models and ensemble NWP models are routinely run worldwide to assist weather forecasting. Deterministic forecasts are capable of capturing more detailed spatial features, while ensemble forecasts, often with a coarser resolution, have the ability to predict uncertainty in future conditions. A comparative understanding of the performance of these two types of forecasts is valuable for both users of NWP products and model developers. Past published comparisons tended to be limited in scope, for example, for only specific locations and weather events, and involving only raw forecasts. In this study, we conduct a comprehensive comparison of the performance of a deterministic model and an ensemble model of the Australian Bureau of Meteorology in forecasting daily precipitation across Australia over a period of 3 years. The deterministic model has a horizontal grid spacing of approximately 25 km, and the ensemble model 60 km. Despite the coarser resolution, the ensemble forecasts are found to be superior by a number of measures, including correlation, accuracy, and reliability. This finding holds true for both raw forecasts from the NWP models and forecasts post-processed using the recently developed seasonally coherent calibration (SCC) model. Post-processing is shown to greatly improve the forecasts from both models; however, the improvement is greater for the deterministic model, narrowing the performance gap between the two models. This study adds strong evidence to the general notion that coarser-resolution ensemble NWP forecasts perform better than deterministic forecasts.

Key words: numerical weather prediction, deterministic forecasts, coarser-resolution ensemble forecasts, precipitation, post-processing, forecast verification

1. Introduction

There are increasing needs for weather forecasts that are accurate and reliable, especially in countries with a diverse climate (Bauer *et al.*, 2015). Such weather forecasts can provide critical information for people to utilize weather resources and prepare for weather hazards. Traditional forecasts produced by deterministic numerical weather prediction (NWP) models are single-valued, and viewed as the "best" estimate of future weather. However, the deterministic forecasts cannot fully represent the uncertainty in synoptic conditions (Epstein, 1969; Molteni *et al.*, 1993; Gneiting and Raftery, 2005). Therefore, ensemble NWP models that generate several forecast members based on varying initial conditions or model physics have been developed to capture the chaotic nature of the atmosphere (Toth and Kalnay, 1993; Ehrendorfer, 1997; Hagelin *et al.*, 2017) and to provide an estimation of forecast uncertainty (Hamill and Colucci, 1998; Palmer *et al.*, 2007; Cuo *et al.*, 2011).

Ensemble NWP forecasts can be further improved when either the spatial resolution or the ensemble size is increased (Buizza *et al.*, 1998). However, due to the high computational cost needed to generate ensemble forecasts, operational ensemble NWP models usually have coarser spatial resolutions compared to deterministic models. As both deterministic and coarser-resolution ensemble forecasts are widely deployed by weather forecasting centers (Roebber *et al.*, 2004; Bauer *et al.*, 2015), it is essential to have a good knowledge of the performance of these two types of forecasts. Forecasting service providers would want to know how much resource to invest in each type of the forecasts; and forecast users would want to know which forecast products are better to use (Richardson, 2000). For these reasons, a number of studies were undertaken to compare deterministic forecasts with coarser-resolution ensemble forecasts.

Atger (2001) investigated the intense precipitation forecasts in French network of rain gauges from winter 1998 to winter 1999 and found that coarser-resolution ensemble forecasts performed overall better than deterministic forecasts from both the European Centre for Medium Range Weather Forecasts (ECMWF) and the US National Centers for Environmental Prediction (NCEP). Grimit and Mass (2002) found that forecasts of wind direction from a 12-km ensemble NWP model were overall comparable in forecast errors to forecasts from a 4-km deterministic model, over the Pacific Northwest of the US from January to June 2000. Rodwell (2006) focused on European SYNOP stations whose monthly-mean climatological precipitation exceeded 4 mm and found that coarser-resolution ensemble forecasts had a better ability in predicting the probability of precipitation occurrence than deterministic forecasts. Vokoun and Hanel (2018) evaluated the most significant summer precipitation events during 2011-2015 in Czech Republic and concluded that coarser-resolution ensemble forecasts were more skillful than deterministic forecasts in most verification metrics. In addition, when considering economic values for decision making, Richardson (2000) found that coarser-resolution ensemble forecasts could provide more economic benefits than deterministic forecasts for precipitation from winter 1996 to winter 1997 over Europe; Mylne (2002) came to a similar conclusion for 10 m wind speed at 41 sites in the UK during two winter seasons, particularly at long lead times.

The studies cited above provide valuable insights into the merit of coarser-resolution ensemble forecasts relative to deterministic forecasts. One of the limitations of these studies is that the comparisons were based on limited locations and weather events. It will be valuable to conduct a comprehensive comparison across a large spatial extent with a wide range of weather events and climate conditions. Another limitation of these studies is that they only compared raw forecasts

from the NWP models, while post-processed forecasts that are often used by end users have not been considered.

It is generally recognised that raw forecasts produced by NWP models are not suitable for direct use (Scherrer *et al.*, 2004; Wu *et al.*, 2019; Möller and Groß, 2020). Raw deterministic and ensemble forecasts are often biased and can even be less skillful than naïve climatology forecasts (Li *et al.*, 2017), especially at long lead times. Raw ensemble forecasts are generally not reliable in ensemble spread (Buizza *et al.*, 2005). To overcome these problems, statistical post-processing methods have been developed, such as Model Output Statistics (Wilks, 2011; Schick *et al.*, 2019), Quantile Regression (Koenker and Bassett, 1978), and Bayesian Joint Probability model (Pokhrel *et al.*, 2013; Peng *et al.*, 2014; Zhao *et al.*, 2015; Cattoën C *et al.*, 2020). These methods can significantly improve the quality of raw forecasts, and make the post-processed forecasts much more suited for applications (Gneiting and Katzfuss, 2014; Scheuerer, 2014; Schuhen *et al.*, 2020). In this context, it will be highly valuable to compare forecasts from deterministic and coarser-resolution ensemble NWP models after proper post-processing as well as before post-processing.

The objective of our study is twofold. First, we take into account a wide range of locations and weather events when comparing deterministic forecasts and coarser-resolution ensemble forecasts. Second, we extend the comparison to post-processed forecasts, to further assess the relative performance of these two types of forecasts. Specifically, we compare daily precipitation forecasts across Australia for a period of 3 years from the ACCESS-G2 (deterministic) model and the ACCESS-GE2 (coarser-resolution ensemble) model of the Australian Bureau of Meteorology (BoM).

The remainder of this paper is structured as follows. We introduce data, post-processing models and evaluation metrics in the next section. We present forecast verification and comparison results

in Section 3. After some discussions in Section 4, we provide a summary and draw conclusions in Section 5.

2. Methods

In this study, we first post-process raw precipitation forecasts from a deterministic NWP model and a coarser-resolution ensemble model using the seasonally coherent calibration (SCC) model (Wang *et al.*, 2019b). We then evaluate raw forecasts and post-processed forecasts against reference data using multiple evaluation metrics. Finally, we compare the evaluation results of deterministic forecasts with coarser-resolution ensemble forecasts, both before and after post-processing.

2.1 Data

2.1.1 Reference data

Gridded daily precipitation data are obtained from the Australian Water Availability Project's climate datasets (AWAP) (Jones *et al.*, 2009). The AWAP precipitation dataset has a high spatial resolution (0.05° x 0.05°) across Australia based on the interpolation of rain gauge observations. An AWAP day starts from 0900 h of the previous day to 0900 h of the current day according to Australian local time including daylight saving. The gridded data for a period of 30 years from August 1, 1989 to July 31, 2019 are used as reference data in this study for establishing SCC models and for evaluating forecasts.

The AWAP datasets are commonly used in Australia and represent the best available reference climate data for Australia. For this study, we treat AWAP precipitation data as truth. However, it is prudent to acknowledge that AWAP data are subject to analysis errors, mainly due to sparsely

distributed gauges in the large geographical span of Australia, especially in areas of low population density (Jones *et al.,* 2009). For example, we found a region in northwest Australia where there was, unrealistically, no precipitation for over 30 years in AWAP. The errors in the AWAP data could potentially have an impact on our forecast calibration and evaluation. Therefore, we need to exercise some caution when interpreting the results.

2.1.2 NWP forecasts

We select ACCESS-G2 (Australian Community Climate and Earth-System Simulator Global 2) and ACCESS-GE2 (Australian Community Climate and Earth-System Simulator Global Ensemble 2) from the BoM to represent the deterministic and coarser-resolution ensemble models, respectively. As the new BoM models ACCESS-G3 and ACCESS-GE3 have only a very short period of archived forecasts, we choose ACCESS-G2 and ACCESS-GE2 for this study to minimize possible impacts of sample size on model performance evaluation.

The ACCESS-G2 model has a horizontal resolution of approximately 25 km. With a forecasting horizon of 10 days, hourly ACCESS-G2 forecasts are produced at 0000 UTC, 0600 UTC, 1200 UTC and 1800 UTC on a daily basis.

The ACCESS-GE2 model has a coarser horizontal resolution of approximately 60 km. Each ACCESS-GE2 ensemble forecast includes 24 members resulting from both perturbed initial conditions and stochastic model physics. ACCESS-GE2 forecasts also have a forecasting horizon of 10 days. 3-hourly ACCESS-GE2 forecasts are produced at 0000 UTC and 1200 UTC every day.

We modify the spatial and temporal resolutions of ACCESS-G2 and ACCESS-GE2 forecasts to match the AWAP data. Specifically, we use bilinear interpolation to regrid the forecasts to the spatial resolution of the AWAP. In view of Australia's UTC offsets, which range from UTC+8 to

UTC+11, we select the ACCESS-G2 and ACCESS-GE2 forecasts produced at 1200 UTC to ensure that the most recent forecasts are used. The NWP hours used in the aggregations are chosen to exactly match the AWAP day. We then aggregate the hourly ACCESS-G2 forecasts and the 3-hourly ACCESS-GE2 forecasts to daily values by accumulating forecasts according to the AWAP day. Hence, the aggregated forecasts in each grid cell represent 24-hour accumulations of precipitation. Due to the adjustments of the time difference between the AWAP day and ACCESS forecasts, and available time of the forecasts, we obtain ACCESS-G2 and ACCESS-GE2 daily precipitation forecasts for 9 days ahead. Forecasts for days 1, 5 and 9 ahead are selected for this study to evaluate forecast performance at different lead times. We evaluate forecasts across 278129 grid cells (as in AWAP) covering Australia during a 3-year period from August 1, 2016 to July 31, 2019.

We should point out that Naughton (2016) made a comparison of raw forecasts from the two models for a 1-month period from February 1, 2014 to February 28, 2014. In that study, ACCESS-GE2 mean precipitation forecasts were found to be more skillfull than ACCESS-G2 forecasts at long lead times. We are extending the evaluation to a period of 3 years to better understand performances of the two models.

2.2 Post-processing methods

The SCC modelling method was developed to post-process deterministic forecasts and produce calibrated ensemble forecasts that were unbiased, reliable in ensemble spread, as skillful as possible, and coherent in seasonal climatology consistent with long-term observations (Wang *et al.*, 2019b). Given a raw forecast, in the form of a deterministic value or ensemble mean, a forecast distribution conditional on the raw forecast can be derived through the SCC model. We sample an ensemble of values from the forecast distribution to represent the forecast probability distribution.

In this study, we use an ensemble size of 100. In establishing an SCC model, we take two steps: (i) we use the AWAP data from August 1, 1989 to July 31, 2019 to obtain model parameters related to long-term climatology of observations; and (ii) we use a 3-year period from August 1, 2016 to July 31, 2019 of raw NWP forecasts and corresponding observations to obtain remaining model parameters. In both steps, we leave one month of data out of the three years. We then apply the established SCC model to the left-out month to calibrate the raw forecasts to produce ensemble forecasts for that month. This procedure is repeated to enable a complete leave-one-month-out cross-validation for the 3-year period. Because the SCC model currently only applies to single-value forecasts, it is used in this study to post-process the ACCESS-G2 deterministic forecasts and the ACCESS-GE2 ensemble mean forecasts. This model is applied separately to different grid cells and lead times, on the basis of its high computational efficiency for practical use.

The SCC modelling method is highly sophisticated. It is worthwhile to compare forecasts post-processed using SCC with post-processed forecasts using a simple bias correction as well as with raw forecasts. In this study, we apply a simple multiplicative bias correction to post-process all ensemble members of ACCESS-GE2 forecasts for the 3-year period. We follow the method of Wang *et al.* (2019a) as described by

$$x'(t) = \frac{\sum_{i=1}^{k} y(i)}{\sum_{i=1}^{k} x_m(i)} * x(t) \quad (1)$$

where $y(i)$ denotes observation and $x_m(i)$ denotes ensemble mean at time i; $x(t)$ is an ensemble member of a new raw forecast and $x'(t)$ is the bias-corrected ensemble member at time t; and k is the number of days in the training period. The simple bias correction is applied to each of the ensemble members. We also adopt the leave-one-month-out cross-validation procedure when applying this simple bias-correction.

2.3 Forecast evaluation

We evaluate both raw and post-processed forecasts at each of the 278129 grid cells. To facilitate the evaluation, we group the forecasts into single-value and ensemble forecasts, as shown in Table 1. Details of the corresponding evaluation metrics for these two groups of forecasts are described in sections 2.3.1 and 2.3.2.

[Table 1]

2.3.1 Evaluation of single-value forecasts

For ensemble forecasts, ensemble mean is usually extracted to provide a single-value forecast, and ensemble spread is used to represent forecast uncertainty (Whitaker and Loughe, 1998; Fortin *et al.*, 2014). Compared to ensemble forecasts, ensemble mean can be more intuitive for comparison with deterministic forecasts (Rodwell, 2006). Furthermore, single-value forecasts are still preferred by many users for ease of interpretation (Ramos *et al.*, 2010). For these reasons, we include in our study a comparison of coarser-resolution ensemble mean forecasts with deterministic forecasts, as in previous studies (Grimit and Mass, 2002; Rodwell, 2006; Vokoun and Hanel, 2018).

We evaluate the correlation between single-value forecasts and corresponding observations to measure the degree of their correspondence (Murphy, 1993). The correlation concept is widely used in many statistical post-processing methods (Li *et al.*, 2020), including the SCC model. A higher correlation shows higher correspondence. However, seasonality embedded in forecasts and observations could lead to misleadingly high correlations, which would not effectively

demonstrate true correspondence. To resolve this problem, we evaluate the forecasts by using the anomaly correlation coefficient as follows.

We first apply a spectral method by Narapusetty *et al.* (2009) to the AWAP data from August 1, 1989 to July 31, 2019 to estimate the climatological mean for each day during the study period (from August 1, 2016 to July 31, 2019). We then calculate the anomaly correlation coefficient between the forecast anomalies and observation anomalies:

$$ACC = \frac{\sum_{t=1}^{n}\{(x(t)-c(t))-(\bar{x}-\bar{c})\}\{(y(t)-c(t))-(\bar{y}-\bar{c})\}}{\sqrt{\sum_{t=1}^{n}\{(x(t)-c(t))-(\bar{x}-\bar{c})\}^2}\sqrt{\sum_{t=1}^{n}\{(y(t)-c(t))-(\bar{y}-\bar{c})\}^2}} \quad (2)$$

where $x(t)$, $y(t)$ and $c(t)$ are the single-value forecast, observation and climatological mean at time t, respectively; $\bar{x}$, $\bar{y}$ and $\bar{c}$ are the average of single-value forecasts, observations, and climatological mean values during the 3-year period, respectively; and n is the total number of days in the 3-year period. The ACC ranges from -1 to 1 and is positively oriented , with a value of 1 showing a perfect linear relationship between the forecasts and observations.

We also apply the continuous ranked probability score (CRPS) to evaluate forecast accuracy, which measures the difference between forecasts and their corresponding observations (Hersbach, 2000). CRPS calculation for single-value forecasts is a special case of calculation for ensemble forecasts. For this reason, we will introduce CRPS in the next sub-section.

2.3.2 Evaluation of ensemble forecasts

We apply the CRPS to evaluate the accuracy of ensemble forecasts:

$$CRPS(t) = \int\{F(t,x) - H(x - y(t))\}^2 dx \quad (3)$$

$$\overline{CRPS} = \frac{1}{n}\sum_{t=1}^{n} CRPS(t) \quad (4)$$

where $F(t,x)$ is the cumulative density function of an ensemble forecast, and $y(t)$ is the observation at time t; H is the Heaviside step function ($H = 1$ if $x - y(t) \geq 0$ and $H = 0$ otherwise); the overbar represents averaging across the n days; and n is the total number of days in the 3-year period. We also evaluate CRPS for ensemble forecasts generated from the SCC fitted climatology of observations. This second CRPS is used as a reference. Then the percentage of reduction from the reference CRPS can be calculated to give a CRPS skill score:

$$CRPS\ skill\ score = \frac{\overline{CRPS}_{ref} - \overline{CRPS}}{\overline{CRPS}_{ref}} \times 100(\%) \quad (5)$$

The CRPS skill score is positively oriented. A positive (negative) skill score indicates that forecasts are better (poorer) than the reference forecasts. Forecasts that perfectly match the corresponding observations will have a skill score of 100%.

The CRPS is applicable to both single-value and ensemble forecasts. Forsingle-value forecasts, CRPS reduces to the mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|x(t) - y(t)| \quad (6)$$

where $x(t)$ and $y(t)$ are the single-value forecast and observation at time t, respectively; and n is the total number of days in the 3-year period. Similarly, we use the CRPS skill score to demonstrate the accuracy of single-value forecasts relative to reference forecasts. To have a fair evaluation, we use the same reference CRPS for single-value and ensemble forecasts.

We calculate the probability integral transform (PIT) for assessing the reliability of ensemble forecast uncertainty (ensemble spread not too wide or too narrow). Statistically, reliability represents the consistency between ensemble forecasts and the corresponding observations (Gneiting *et al.*, 2007). The PIT for a forecast-observation pair at time t is defined as:

$$\pi(t) = F\big(t, x = y(t)\big) \tag{7}$$

where $F(t,x)$ is the cumulative density function of the ensemble forecast, and $y(t)$ is the observation. For reliable forecasts, $\pi(t)$ follows a uniform distribution. The uniformity can be visually examined by either histograms or uniform probability plots of the PIT values (Schepen *et al.*, 2018). In this study, we summarize the reliability as α-index (Renard *et al.*, 2010):

$$\alpha = 1 - \frac{2}{n}\sum_{t=1}^{n}\left|\pi^*(t) - \frac{t}{n+1}\right| \tag{8}$$

where $\pi^*(t)$ is the sorted $\pi(t)$, $t = 1,2,\dots,n$, in an increasing order; and n is the total number of days in the 3-year period. The α-index ranges from 0 to 1, with a value of 1 showing perfect reliablility, and a value of 0 showing worst reliability. The calculation of PIT becomes problematic when an observation is known to be below or equal to a certain value y_c. Here y_c is 0.2 mm per day, which reflects the precision of available observed precipitation data. In this case, $\pi(t)$ cannot be precisely found. To overcome this problem, we randomly generate a pseudo-PIT value from a uniform distribution with a range $[0, F(t, x = y_c)]$ and subsequently use it to calculate the PIT (Wang and Robertson, 2011).

3. Results

3.1 Correlation between single-value forecasts and observations

Results of ACC for ACCESS-G2 forecasts and ACCESS-GE2 mean forecasts are shown in Fig. 1. Here the left triangle symbol on the color bar of the maps means that ACC can be negative. According to the cumulative density function (CDF) plots, there are only few grid cells with negative ACC values at day 9 ahead. The high ACC area of both forecasts are mainly distributed

in eastern and southwestern parts of Australia, especially at short lead times. ACC decreases as lead time increases, indicating lower correspondence between forecast and observed anomalies at longer lead times. For instance, most grid cells possess ACC values over 0.6 at day 1 ahead, but hardly any grid cell has an ACC over 0.6 at day 9 ahead.

[Fig. 1]

At each lead time, both the ACC maps and CDF plots show similar ACC ranges for these two forecasts. Nevertheless, ACCESS-GE2 mean forecasts have overall higher ACC values than ACCESS-G2 forecasts (more obvious from the CDF plots). Their difference is more pronounced at longer lead times. This suggests that there are more advantages of an ensemble approach over a deterministic approach when NWP models are used to generate forecasts for longer lead times.Generally, the ensemble mean is superior to each ensemble member in terms of the correlation as it tends to average out the less predictable detail (Richardson, 2000). By extension, the ensemble mean can also outperform a higher-resolution deterministic model run.

3.2 Accuracy of single-value forecasts

Results of CRPS skill score for ACCESS-G2 forecasts and ACCESS-GE2 mean forecasts are shown in Fig. 2. The left (right) triangle symbol on the color bar means that skill score can be below (above) the displayed minimum (maximum) value. The spatial distributions of the skill scores for these two forecasts are similar, with low values in midwest and high values in eastern and southwestern parts of Australia. The skill score tends to decrease rapidly with increasing lead times, as similarly presented by Golding (1998). And both forecasts generally have negative skill scores in most grid cells, particularly at long lead times. At day 1 ahead, nearly 80% of the grid cells have negative skill scores. However, at days 5 and 9 ahead, there is almost no positive value on the skill score map. This shows that both ACCESS-G2 forecsts and ACCESS-GE2 mean

forecasts are less accurate than the ensemble climatology forecasts in most parts of Australia and across different lead times. Although having similar skill score ranges at each lead time, ACCESS-GE2 mean forecasts have overall higher skill scores than ACCESS-G2 forecasts (more obvious from the CDF plots). And their difference is also more pronounced at longer lead times.

[Fig. 2]

3.3 Accuracy of ensemble forecasts

Results of CRPS skill score for ACCESS-GE2, simple bias-corrected ACCESS-GE2, SCC-calibrated ACCESS-G2, and SCC-calibrated ACCESS-GE2 mean forecasts are shown in Fig. 3. Raw ACCESS-GE2 ensemble forecasts have positive skill scores in most grid cells, especially at short lead times. For example, at day 1 ahead, there are positive values at about 80% of the grid cells. This indicates the benefit of the ensemble spread of raw ensemble forecasts, since ACCESS-GE2 ensemble forecasts have much higher skill scores than ACCESS-GE2 mean forecasts (latter shown in Fig. 2). Besides, ACCESS-GE2 ensemble forecasts are considerably improved by the simple bias correction, particularly at grid cells with negative skill scores shown in Fig. 3.

[Fig. 3]

The SCC model greatly improves the skill scores of ACCESS-G2 forecasts and ACCESS-GE2 mean forecasts (latter shown in Fig. 2). Almost all negative values are removed, and positive values are further enhanced. The spatial pattern of the skill scores for both forecasts is also retained by the SCC model. The skill scores of both SCC-calibrated ACCESS-G2 and SCC-calibrated ACCESS-GE2 mean forecasts are mostly positive, with higher values at short lead times and approaching zero at long lead times. This is because the SCC model is designed to make the calibrated forecasts no worse than climatology forecasts when forecast skill becomes low at long

lead times (Wang *et al.*, 2019b). Likewise, the CDF plots become sharper as lead time increases, especially for grid cells where CRPS skill score is positive.

Similarly, SCC-calibrated ACCESS-GE2 mean forecasts generally have higher skill scores than SCC-calibrated ACCESS-G2 forecasts at all lead times (more obvious from the CDF plots). And the skill difference between ACCESS-G2 forecasts and ACCESS-GE2 mean forecasts (Fig. 2) is narrowed by the SCC model (Fig. 3). Furthermore, the simple bias-corrected ACCESS-GE2 forecasts generally have lower skill scores than SCC-calibrated ACCESS-GE2 mean forecasts, despite the fact that the simple bias correction makes use of the ACCESS-GE2 ensemble spread information while the SCC model only makes use of the ACCESS-GE2 mean information. This confirms the advantage of the sophisticated SCC model relative to simple calibration models.

To further investigate the difference between SCC-calibrated ACCESS-GE2 mean and SCC-calibrated ACCESS-G2 forecasts under different climate conditions, we compare their accuracies by seasons and across regions with different annual precipitation. Results of the skill score difference between these two forecasts are shown in Fig. 4 and Fig. 5, with positive difference indicating higher skill scores in SCC-calibrated ACCESS-GE2 mean forecasts over SCC-calibrated ACCESS-G2 forecasts, and vice versa.

[Fig. 4]

[Fig. 5]

At all lead times, the skill score difference is positive for the majority of grid cells for all seasons and regions. Therefore, SCC-calibrated ACCESS-GE2 mean forecasts can provide overall improved accuracy compared to SCC-calibrated ACCESS-G2 forecasts under different types of climate conditions. Their difference changes slightly with seasons and regions, and generally

decreases as lead time increases. Besides, when the annual precipitation is smaller than 200 mm, there is a large percentage of grid cells where SCC-calibrated ACCESS-GE2 mean forecasts are less skillful than the SCC-calibrated ACCESS-G2 forecasts.

3.4 Reliability of ensemble forecasts

Results of α-index for ACCESS-GE2, simple bias-corrected ACCESS-GE2, SCC-calibrated ACCESS-G2, and SCC-calibrated ACCESS-GE2 mean forecasts are shown in Fig. 6. Clearly, the ensemble spread of ACCESS-GE2 and simple bias-corrected ACCESS-GE2 forecasts is not reliable at any lead time, although the simple bias correction slightly increases α-index of ACCESS-GE2 forecasts. In contrast, SCC-calibrated ACCESS-G2 and SCC-calibrated ACCESS-GE2 mean forecasts have α-index values close to 1 at almost all grid cells, meaning that both forecasts can reliably quantify forecast uncertainties. It should be noted that the CDF lines of these two SCC-calibrated forecasts are almost identical and their difference in reliability is unnoticeable.

[Fig. 6]

From the results above, ACCESS-GE2 forecasts have a clear advantage over ACCESS-G2 forecasts in terms of ACC and CRPS skill score. To demonstrate the significance level of this advantage, we carry out Kolmogorov-Smirnov (K-S) statistical significance testing on the CDF distributions of the ACC and CRPS skill score values of these two forecasts. On interpreting the K-S testing results, we conclude that the difference between ACCESS-GE2 forecasts and ACCESS-G2 forecasts is statistically significant, both before and after post-processing. And as expected, the significance increases with lead time. Details of the K-S testing and interpretation can be found in the online supplementary material.

4. Discussion

The key motivation for this work is to thoroughly evaluate the performance of deterministic NWP models and coarser-resolution ensemble models by comparing their precipitation forecasts in terms of correlation with reference data, accuracy, and reliability. We conduct a comparative analysis of these two types of forecasts of precipitation over a period of 3 years across Australia. Results show that the coarser-resolution ensemble forecasts overall perform better than deterministic forecasts. Our investigation further confirms findings from previous studies (Atger, 2001; Mylne, 2002; Rodwell, 2006; Vokoun and Hanel, 2018), by performing spatially and temporally explicit comparisons, and extending the comparison to post-processed forecasts.

We acknowledge that our study is based on only ACCESS-G2 and ACCESS-GE2 models operated by the BoM. Indeed, evaluation of different forecasting systems may lead to different conclusions, which may be influenced by the spatial resolution gaps between these two types of models as well as different NWP model mechanisms (Buizza *et al.*, 2005). Nonetheless, we hope that this study can be a valuable reference for various forecasting centers, in terms of the comparison methods and the results. In this study, we select precipitation as our target variable of weather forecasts.Our results are also consistent with findings on other weather variables and weather-related variables, as demonstrated by studies on wind direction (Grimit and Mass, 2002) and hydrological forecasts (Boucher *et al.*, 2011).

The superiority of coarser-resolution ensemble forecasts over deterministic forecasts may increase with ensemble size (Richardson, 2000; Mullen and Buizza, 2002). Therefore, increasing the ensemble size of ensemble forecasts may be a better strategy than increasing the resolution of deterministic forecasts. Additionally, other than ensemble forecasts produced by multiple runs of a single NWP model, coarser-resolution ensemble forecasts generated by collecting forecast

members from different NWP models or by analog ensemble techniques also show better performance than deterministic forecasts (Stensrud *et al.*, 1999; Wandishin *et al.*, 2001; Bowler *et al.*, 2008; Zhang *et al.*, 2015). Therefore, any loss in spatial resolution of ensemble forecasts can generally be compensated by the ensemble approach.

Often, NWP forecasts need to be regridded in accordance with the grid spacing of the reference data. However, using interpolation techniques may affect the forecast skill scores in a statistically significant way (Accadia *et al.*, 2003). The use of bilinear interpolation in this study and its potential statistical effects on the forecast quality deserve to be further investigated. Besides, we interpret the forecast verification results across all grid cells of Australia, including the region with no precipitation for more than 30 years according to the AWAP. To alleviate the possible influence of the AWAP analysis errors on forecast evaluation, it would be more sensible to exclude such regions in future studies.

The post-processing of raw forecasts presented in this study is conducted individually for each lead time and each grid cell. In practice, when multiple lead times and multiple locations are involved, their mutual relationships should be included in post-processed ensemble forecasts, so as to capture the temporal and spatial correlation structures of observed precipitation (Shrestha *et al.*, 2015; Schepen *et al.*, 2020). For this purpose, methods can be applied to connect post-processed ensemble members from different lead times and locations. One of the popularly used methods, the Schaake shuffle, has shown to be effective (Clark *et al.*, 2004; Scheuerer *et al.*, 2017).

We also notice that the ensemble spread of ACCESS-GE2 forecasts contains remarkably beneficial information, which could be utilized to enhance post-processing (Möller *et al.*, 2013; Williams *et al.*, 2014). Many studies have incorporated ensemble spread in forecast post-processing (Gneiting *et al.*, 2005; Veenhuis, 2013; Messner *et al.*, 2014; Scheuerer and Hamill, 2015), but the

improvements are generally not significant. How to appropriately make use of the spread information in post-processing still needs further investigation. Besides, although ACCESS-GE2 forecasts perform better than ACCESS-G2 forecasts in most grid cells, there are some grid cells where ACCESS-GE2 forecasts demonstrate worse performance. A natural idea is to combine these two forecasts appropriately to generate new forecasts that are better than either one across all locations. Many methods have been developed to merge available data information from different sources (Rodwell, 2006; Wang *et al.*, 2012; Barnes *et al.*, 2019; Xu *et al.*, 2019; Leutbecher and Ben Bouallègue, 2020). It will be useful to evaluate how to apply some of these techniques to combine these two types of NWP forecasts to take advantages of their strengths.

5. Summary and conclusions

A comparison of forecasts from deterministic NWP models and coarser-resolution ensemble models is valuable for forecast users and model developers. However, previous comparisons typically focused on evaluating raw forecasts for limited locations and weather events. In this study, we provide a comprehensive comparison of these two types of forecasts, both before and after post-processing, and for a wide range of locations and weather events.

Specifically, we evaluate and compare the ACCESS-G2 and ACCESS-GE2 models of the BoM for daily precipitation forecasts across Australia for a period of 3 years. ACCESS-G2 is a deterministic model with a resolution of 25 km, and ACCESS-GE2 is an ensemble model with a resolution of 60 km. In post-processing the forecasts, we employ the recently developed seasonally coherent calibration (SCC) model.

We find that ACCESS-GE2 forecasts are overall superior than ACCESS-G2 forecasts at all lead times, under all climatic conditions, both before and after post-processing. Specifically, for raw forecasts, ACCESS-GE2 (mean) forecasts are more accurate and have a higher correlation with corresponding observations than ACCESS-G2 forecasts, especially at long lead times. For post-processed forecasts, SCC-calibrated ACCESS-GE2 mean forecasts are more accurate than SCC-calibrated ACCESS-G2 forecasts. The conclusion holds true for different seasons and regions. In general, the performance gap between the two models is narrowed by post-processing with SCC.

With equivalent computation costs, increasing the number of model runs usually brings more benefits to NWP models than increasing the spatial resolution (Mullen and Buizza, 2002). Therefore, coarser-resolution ensemble NWP models tend to be more resource-efficient than deterministic models. This further adds to the advantage of ensemble forecasts over deterministic forecasts.

Both SCC-calibrated ACCESS-G2 and SCC-calibrated ACCESS-GE2 mean forecasts are found to be reliable in ensemble spread. The SCC calibration greatly improves the performance of raw forecasts in terms of both accuracy and reliability. The value of post-processing NWP forecasts is clearly demonstrated.

Acknowledgments

This study is linked to a collaborative project (TP707466) between the University of Melbourne and Australian Bureau of Meteorology and by an ARC Linkage Project (LP170100922). We thank the Australian Bureau of Meteorology for providing access to the ACCESS-G2, ACCESS-GE2, and AWAP data. We also thank the National Computational Infrastructure for providing

computation resources to support our work. We gratefully acknowledge the editor and the two reviewers for their thorough reviews and constructive suggestions.

References

Accadia, C., Mariani, S., Casaioli, M., Lavagnini, A. and Speranza, A. (2003) 'Sensitivity of Precipitation Forecast Skill Scores to Bilinear Interpolation and a Simple Nearest-Neighbor Average Method on High-Resolution Verification Grids', *Weather and Forecasting*, 18(5), pp. 918-932. https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

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

Barnes, C., Brierley, C. M. and Chandler, R. E. (2019) 'New approaches to postprocessing of multi-model ensemble forecasts', *Quarterly Journal of the Royal Meteorological Society*, 145(725), pp. 3479-3498. https://doi.org/10.1002/qj.3632.

Bauer, P., Thorpe, A. and Brunet, G. (2015) 'The quiet revolution of numerical weather prediction', *Nature*, 525, p. 47. https://doi.org/10.1038/nature14956.

BoM (2016) *APS2 Upgrade to the ACCESS-G Numerical Weather Prediction System*.

Boucher, M. A., Anctil, F., Perreault, L. and Tremblay, D. (2011) 'A comparison between ensemble and deterministic hydrological forecasts in an operational context', *Advances in Geosciences*, 29, pp. 85-94. https://doi.org/10.5194/adgeo-29-85-2011.

Bowler, N. E., Arribas, A. and Mylne, K. R. (2008) 'The Benefits of Multianalysis and Poor Man's Ensembles', *Monthly Weather Review*, 136(11), pp. 4113-4129. https://doi.org/10.1175/2008MWR2381.1.

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

Buizza, R., Petroliagis, T., Palmer, T., Barkmeijer, J., Hamrud, M., Hollingsworth, A., Simmons, A. and 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(550), pp. 1935-1960. https://doi.org/10.1002/qj.49712455008.

Cattoën C, Robertson DE, Bennett JC, Wang QJ and TK, C.-S. (2020) 'Calibrating hourly precipitation forecasts with daily observations', *Journal of Hydrometeorology, Accepted*.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. and Wilby, R. (2004) 'The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields', *Journal of Hydrometeorology*, 5(1), pp. 243-262. https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

Cuo, L., Pagano, T. C. and Wang, Q. J. (2011) 'A Review of Quantitative Precipitation Forecasts and Their Use in Short- to Medium-Range Streamflow Forecasting', *Journal of Hydrometeorology*, 12(5), pp. 713-728. https://doi.org/10.1175/2011JHM1347.1.

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

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

Fortin, V., Abaza, M., Anctil, F. and Turcotte, R. (2014) 'Why Should Ensemble Spread Match the RMSE of the Ensemble Mean?', *Journal of Hydrometeorology*, 15(4), pp. 1708-1713. https://doi.org/10.1175/JHM-D-14-0008.1.

Gneiting, T., Fadoua, B. and Raftery, A. E. (2007) 'Probabilistic Forecasts, Calibration and Sharpness', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 69(2), pp. 243-268. https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Gneiting, T. and Katzfuss, M. (2014) 'Probabilistic Forecasting', *Annual Review of Statistics and Its Application*, 1(1), pp. 125-151. https://doi.org/10.1146/annurev-statistics-062713-085831.

Gneiting, T. and Raftery, A. E. (2005) 'Weather forecasting with ensemble methods', *Science*, 310, p. 248+. https://doi.org/10.1126/science.1115255.

Gneiting, T., Raftery, A. E., III, A. H. W. and Goldman, T. (2005) 'Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation', *Monthly Weather Review*, 133(5), pp. 1098-1118. https://doi.org/10.1175/MWR2904.1.

Golding, B. W. (1998) 'Nimrod: a system for generating automated very short range forecasts', *Quarterly Journal of the Royal Meteorological Society*, 5(1), pp. 1-16. https://doi.org/10.1017/s1350482798000577.

Grimit, E. P. and Mass, C. F. (2002) 'Initial Results of a Mesoscale Short-Range Ensemble Forecasting System over the Pacific Northwest', *Weather and Forecasting*, 17(2), pp. 192-205. https://doi.org/10.1175/1520-0434(2002)017<0192:IROAMS>2.0.CO;2.

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N. and Tennant, W. (2017) 'The Met Office convective-scale ensemble, MOGREPS-UK', *Quarterly Journal of the Royal Meteorological Society*, 143(708), pp. 2846-2861. https://doi.org/10.1002/qj.3135.

Hamill, T. M. and Colucci, S. J. (1998) 'Evaluation of Eta–RSM Ensemble Probabilistic Precipitation Forecasts', *Monthly Weather Review*, 126(3), pp. 711-724. https://doi.org/10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2.

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

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

Koenker, R. and Bassett, G. (1978) 'Regression Quantiles', *Econometrica*, 46(1), pp. 33-50. https://doi.org/10.2307/1913643.

Leutbecher, M. and Ben Bouallègue, Z. (2020) 'On the probabilistic skill of dual-resolution ensemble forecasts', *Quarterly Journal of the Royal Meteorological Society*, 146(727), pp. 707-723. https://doi.org/10.1002/qj.3704.

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

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

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

Möller, A. and Groß, J. (2020) 'Probabilistic temperature forecasting with a heteroscedastic autoregressive ensemble postprocessing model', *Quarterly Journal of the Royal Meteorological Society*, 146(726), pp. 211-224. https://doi.org/10.1002/qj.3667.

Möller, A., Lenkoski, A. and Thorarinsdottir, T. L. (2013) 'Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas', *Quarterly Journal of the Royal Meteorological Society*, 139(673), pp. 982-991. https://doi.org/10.1002/qj.2009.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. (1993) 'The ECMWF ensemble prediction system: Methodology and validation', *Fourth Workshop on Meteorological Operational Systems, 22-26 November 1993*. Shinfield Park, Reading, 1993. ECMWF. Available at: https://www.ecmwf.int/node/11190.

Mullen, S. L. and Buizza, R. (2002) 'The Impact of Horizontal Resolution and Ensemble Size on Probabilistic Forecasts of Precipitation by the ECMWF Ensemble Prediction System', *Weather and Forecasting*, 17(2), pp. 173-191. https://doi.org/10.1175/1520-0434(2002)017<0173:Tiohra>2.0.Co;2.

Murphy, A. H. (1993) 'What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting', *Weather and forecasting*, (8 (2)), pp. 281–293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Mylne, K. R. (2002) 'Decision-making from probability forecasts based on forecast value', *Meteorological Applications*, 9(3), pp. 307-315. https://doi.org/10.1017/S1350482702003043.

Narapusetty, B., DelSole, T. and Tippett, M. K. (2009) 'Optimal Estimation of the Climatological Mean', *Journal of Climate*, 22(18), pp. 4845-4859. https://doi.org/10.1175/2009JCLI2944.1.

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

Palmer, T. N., Roberto, B., Martin, L., Renate, H., Jung, T., Mark, R., Frédéric, V., Berner, J., Hágel, E., Lawrence, A. R., Florian, P., Park, Y. Y., Bremen, L. v. and Gilmour, I. (2007) 'The Ensemble Prediction System - Recent and Ongoing Developments'. ECMWF. Available at: https://www.ecmwf.int/node/12527.

Peng, Z., Wang, Q. J., Bennett, J. C., Schepen, A., Pappenberger, F., Pokhrel, P. and Wang, Z. (2014) 'Statistical calibration and bridging of ECMWF System4 outputs for forecasting seasonal precipitation over China', *Journal of Geophysical Research: Atmospheres*, 119(12), pp. 7116-7135. https://doi.org/10.1002/2013JD021162.

Pokhrel, P., Robertson, D. E. and Wang, Q. J. (2013) 'A Bayesian joint probability post-processor for reducing errors and quantifying uncertainty in monthly streamflow predictions', *Hydrology and Earth System Sciences*, 17(2), pp. 795-804. https://doi.org/10.5194/hess-17-795-2013.

Ramos, M.-H., Mathevet, T., Thielen, J. and Pappenberger, F. (2010) 'Communicating uncertainty in hydro-meteorological forecasts: mission impossible?', *Meteorological Applications*, 17(2), pp. 223-235. https://doi.org/10.1002/met.202.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M. and Franks, S. W. (2010) 'Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors', *Water Resources Research*, 46(5). https://doi.org/10.1029/2009WR008328.

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

Rodwell, M. J. (2006) 'Comparing and combining deterministic and ensemble forecasts: How to predict rainfall occurrence better', *ECMWF Newsletter*(106), pp. 17-23. https://doi.org/10.21957/cd347812th.

Roebber, P. J., Schultz, D. M., Colle, B. A. and Stensrud, D. J. (2004) 'Toward Improved Prediction: High-Resolution and Ensemble Modeling Systems in Operations', *Weather and Forecasting*, 19(5), pp. 936-949. https://doi.org/10.1175/1520-0434(2004)019<0936:TIPHAE>2.0.CO;2.

Schepen, A., Everingham, Y. and Wang, Q. J. (2020) 'On the Joint Calibration of Multivariate Seasonal Climate Forecasts from GCMs', *Monthly Weather Review*, 148(1), pp. 437-456. https://doi.org/10.1175/MWR-D-19-0046.1.

Schepen, A., Zhao, T., J. Wang, Q. and Robertson, D. (2018) 'A Bayesian modelling method for post-processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments', *Hydrology and Earth System Sciences*, 22, pp. 1615-1628. https://doi.org/10.5194/hess-22-1615-2018.

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

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

Scheuerer, M. and Hamill, T. M. (2015) 'Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions', *Monthly Weather Review*, 143(11), pp. 4578-4596. https://doi.org/10.1175/mwr-d-15-0061.1.

Scheuerer, M., Hamill, T. M., Whitin, B., He, M. and Henkel, A. (2017) 'A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation', *Water Resources Research*, 53(4), pp. 3029-3046. https://doi.org/10.1002/2016WR020133.

Schick, S., Rössler, O. and Weingartner, R. (2019) 'An Evaluation of Model Output Statistics for Subseasonal Streamflow Forecasting in European Catchments', *Journal of Hydrometeorology*, 20(7), pp. 1399-1416. https://doi.org/10.1175/jhm-d-18-0195.1.

Schuhen, N., Thorarinsdottir, T. L. and Lenkoski, A. (2020) 'Rapid adjustment and post-processing of temperature forecast trajectories', *Quarterly Journal of the Royal Meteorological Society*, 146(727), pp. 963-978. https://doi.org/10.1002/qj.3718.

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

Stensrud, D., Brooks, H., Du, J., Tracton, M. and Rogers, E. (1999) 'Using Ensembles for Short-Range Forecasting', *Monthly Weather Review*, 127. https://doi.org/10.1175/1520-0493(1999)127<0433:UEFSRF>2.0.CO;2.

Toth, Z. and Kalnay, E. (1993) 'Ensemble Forecasting at NMC: The Generation of Perturbations', *Bulletin of the American Meteorological Society*, 74(12), pp. 2317-2330. https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

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

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

Wandishin, M. S., Mullen, S. L., Stensrud, D. J. and Brooks, H. E. (2001) 'Evaluation of a Short-Range Multimodel Ensemble System', *Monthly Weather Review*, 129(4), pp. 729-747. https://doi.org/10.1175/1520-0493(2001)129<0729:EOASRM>2.0.CO;2.

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

Wang, Q. J., Schepen, A. and Robertson, D. E. (2012) 'Merging Seasonal Rainfall Forecasts from Multiple Statistical Models through Bayesian Model Averaging', *Journal of Climate*, 25(16), pp. 5524-5537. https://doi.org/10.1175/JCLI-D-11-00386.1.

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

Wang, Q. J., Zhao, T., Yang, Q. and Robertson, D. (2019b) 'A Seasonally Coherent Calibration (SCC) Model for Postprocessing Numerical Weather Predictions', *Monthly Weather Review*, 147(10), pp. 3633-3647. https://doi.org/10.1175/mwr-d-19-0108.1.

Whitaker, J. S. and Loughe, A. F. (1998) 'The Relationship between Ensemble Spread and Ensemble Mean Skill', *Monthly Weather Review*, 126(12), pp. 3292-3302. https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2.

Wilks, D. S. (2011) *Statistical methods in the atmospheric sciences*. Academic Press.

Williams, R. M., Ferro, C. A. T. and Kwasniok, F. (2014) 'A comparison of ensemble post-processing methods for extreme events', *Quarterly Journal of the Royal Meteorological Society*, 140(680), pp. 1112-1120. https://doi.org/10.1002/qj.2198.

Wu, Y., Yang, X., Zhang, W. and Kuang, Q. (2019) 'Mixture probabilistic model for precipitation ensemble forecasting', *Quarterly Journal of the Royal Meteorological Society*, 145(725), pp. 3516-3534. https://doi.org/10.1002/qj.3637.

Xu, J., Anctil, F. and Boucher, M.-A. (2019) 'Hydrological post-processing of streamflow forecasts issued from multimodel ensemble prediction systems', *Journal of Hydrology*, 578, p. 124002. https://doi.org/10.1016/j.jhydrol.2019.124002.

Zhang, J., Draxl, C., Hopson, T., Monache, L. D., Vanvyve, E. and Hodge, B.-M. (2015) 'Comparison of numerical weather prediction based deterministic and probabilistic wind resource assessment methods', *Applied Energy*, 156, pp. 528-541. https://doi.org/10.1016/j.apenergy.2015.07.059.

Zhao, T., Wang, Q. J., Bennett, J. C., Robertson, D. E., Shao, Q. and Zhao, J. (2015) 'Quantifying predictive uncertainty of streamflow forecasts based on a Bayesian joint probability model', *Journal of Hydrology*, 528, pp. 329-340. https://doi.org/10.1016/j.jhydrol.2015.06.043.

Tables

Table 1. Groups of single-value forecasts and ensemble forecasts and evaluation metrics

Groups	**Forecasts**	**Evaluation metrics**
Single-value forecasts	ACCESS-G2, ACCESS-GE2 mean	Correlation (ACC), Accuracy (CRPS skill score)
Ensemble forecasts	ACCESS-GE2, Simple bias-corrected ACCESS-GE2, SCC-calibrated ACCESS-G2, SCC-calibrated ACCESS-GE2 mean	Accuracy (CRPS skill score), Reliability (α-index)

Figures

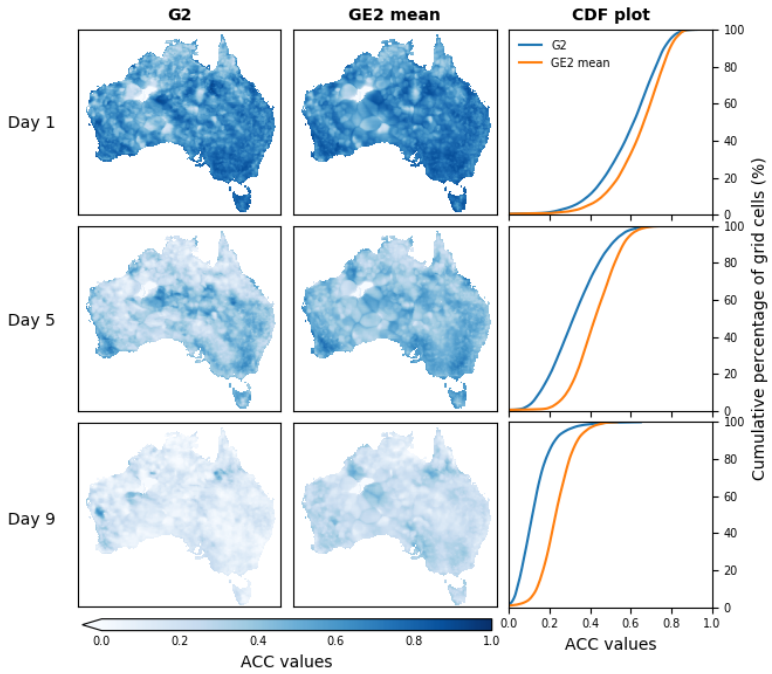


Fig. 1. ACC maps of ACCESS-G2 forecasts and ACCESS-GE2 mean forecasts at three lead times (first two columns), and cumulative density function (CDF) plots of the ACC across all grid cells (third column).

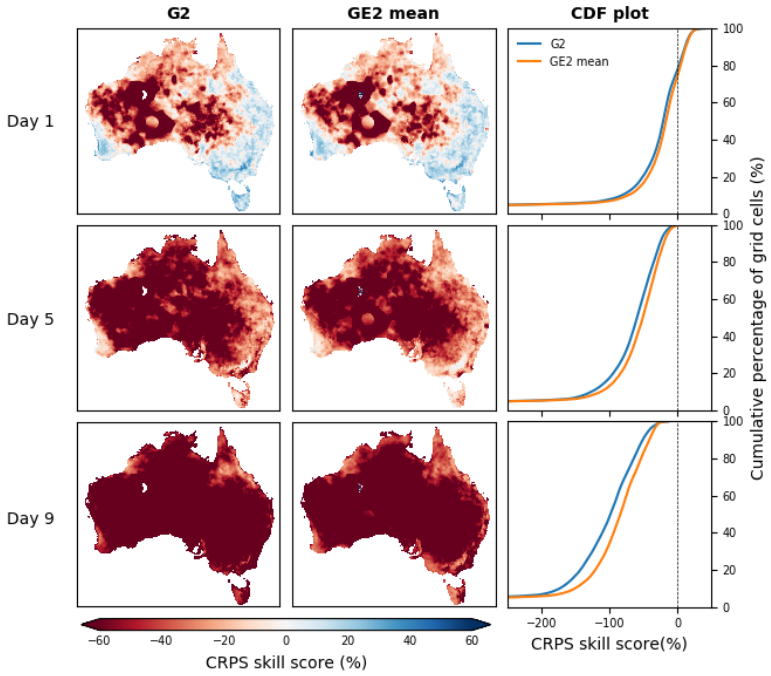


Fig. 2. CRPS skill score maps of ACCESS-G2 forecasts and ACCESS-GE2 mean forecasts at three lead times (first two columns), and CDF plots of the CRPS skill score across all grid cells (third column).

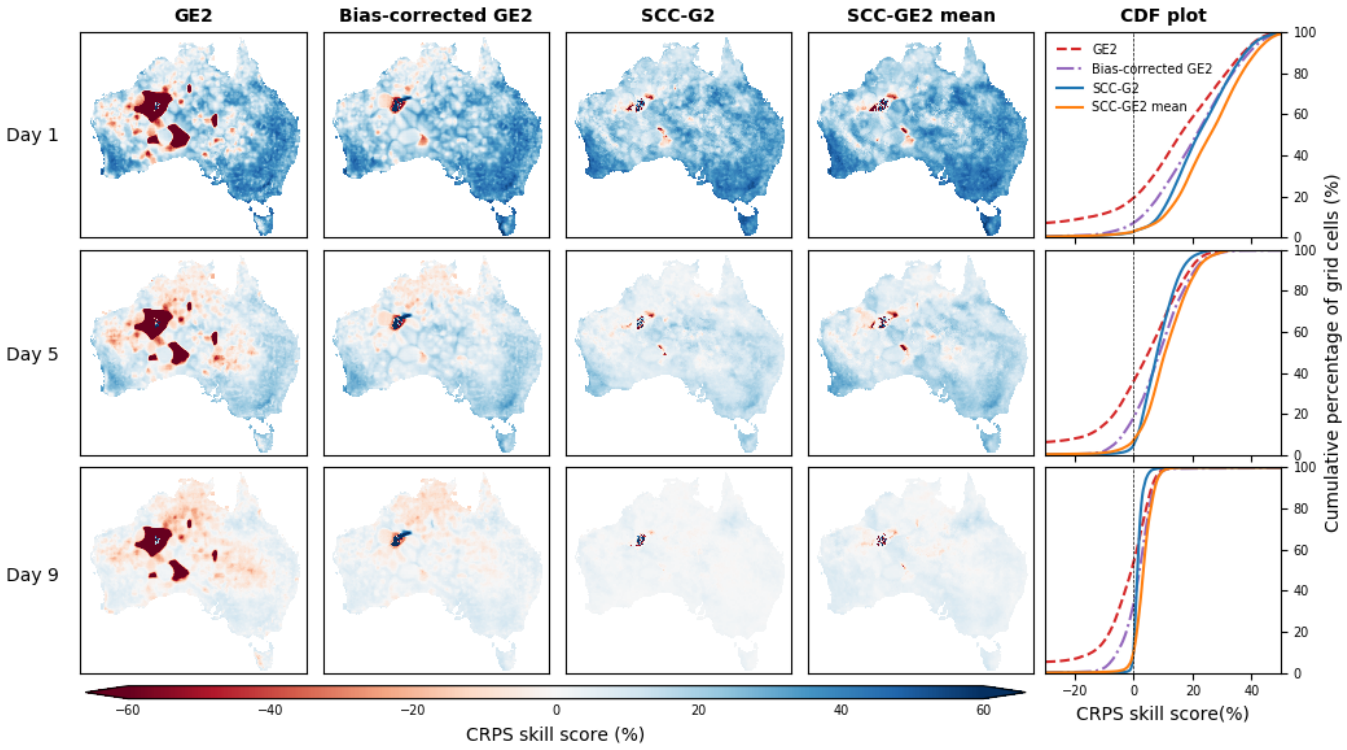


Fig. 3. CRPS skill score maps of ACCESS-GE2, simple bias-corrected ACCESS-GE2, SCC-calibrated ACCESS-G2, and SCC-calibrated ACCESS-GE2 mean forecasts at three lead times (first four columns), and CDF plots of the CRPS skill score across all grid cells (fifth column). It is worth noting that for easy interpretation, the scale of the CRPS skill score axis in the CDF plots is much smaller than that in Fig. 2.

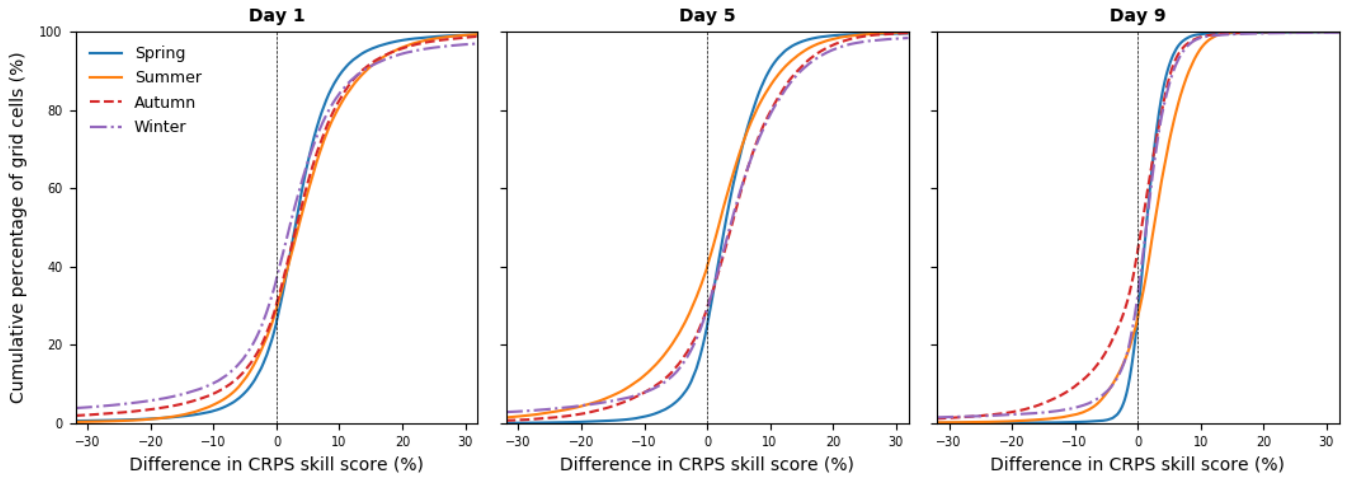


Fig. 4. CDF plots of the difference in CRPS skill score between SCC-calibrated ACCESS-GE2 mean and SCC-calibrated ACCESS-G2 forecasts across all grid cells in different seasons.

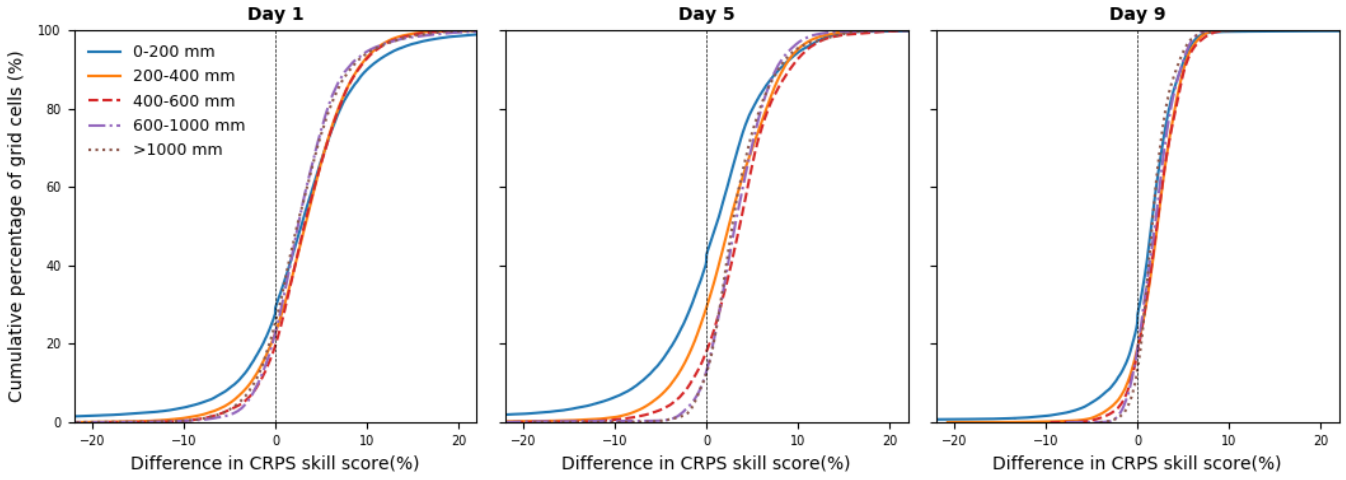


Fig. 5. CDF plots of the difference in CRPS skill score between SCC-calibrated ACCESS-GE2 mean and SCC-calibrated ACCESS-G2 forecasts in regions with different amounts of annual precipitation.

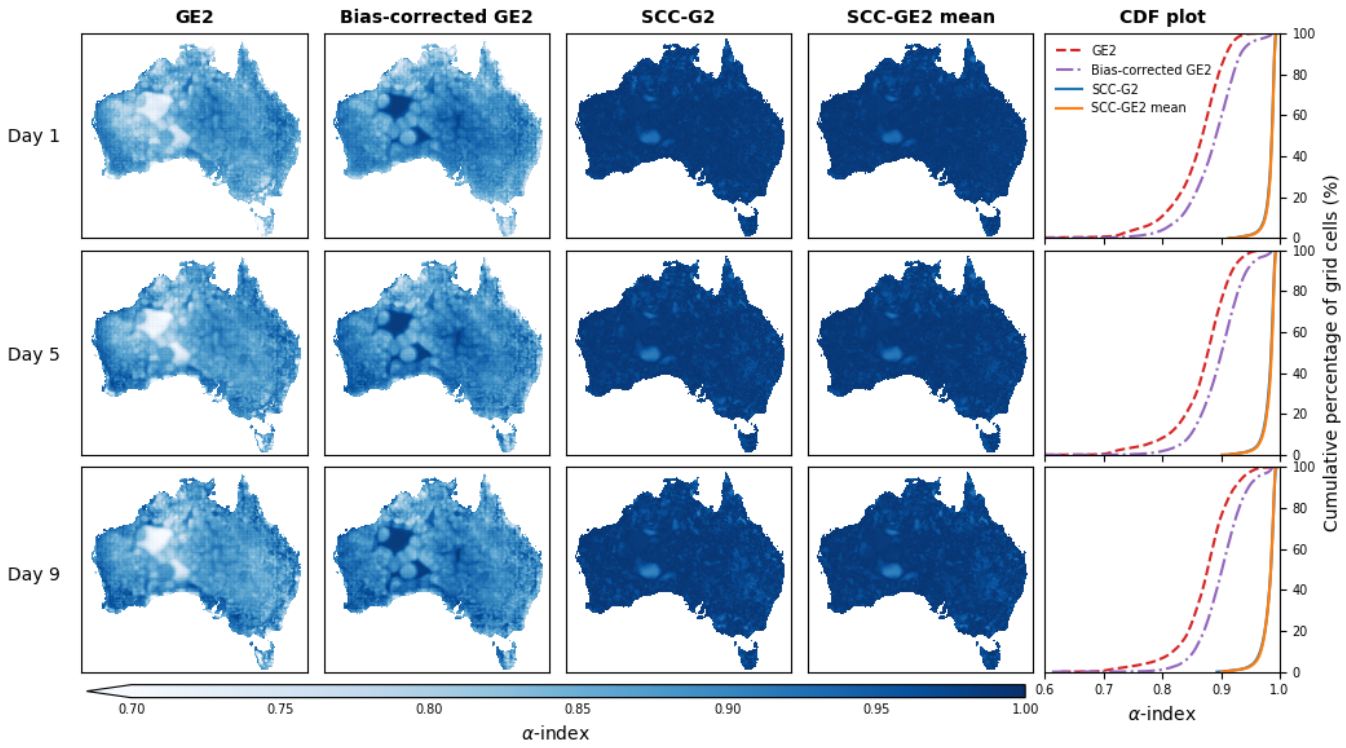
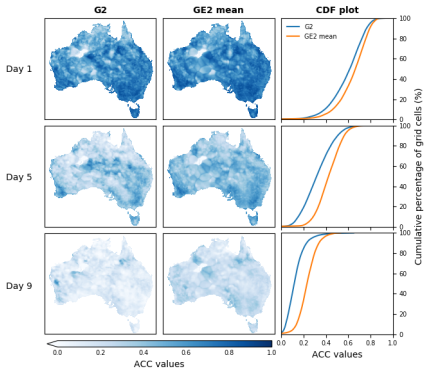
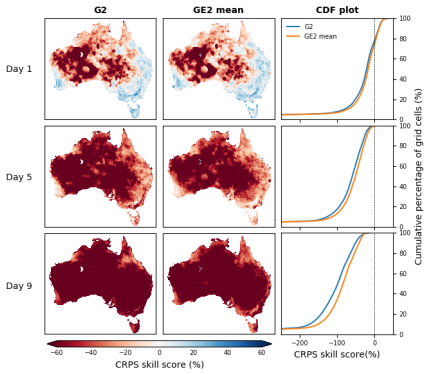


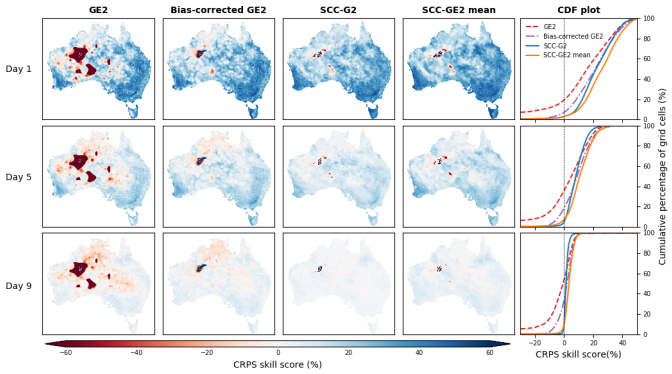
Fig. 6. α-index of ACCESS-GE2, simple bias-corrected ACCESS-GE2, SCC-calibrated ACCESS-G2, and SCC-calibrated ACCESS-GE2 mean forecasts at three lead times (first four columns), and CDF plots of the α-index across all grid cells (fifth column).



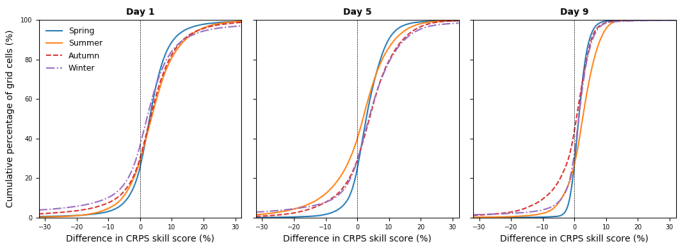
QJ_3952_Fig. 1.tiff



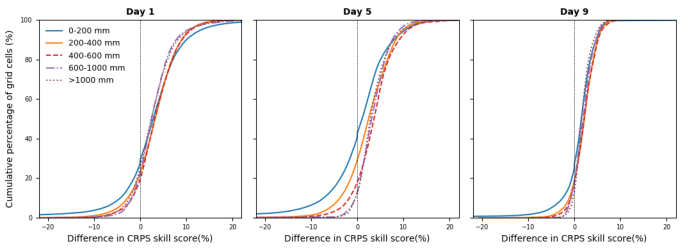
QJ_3952_Fig. 2.tiff



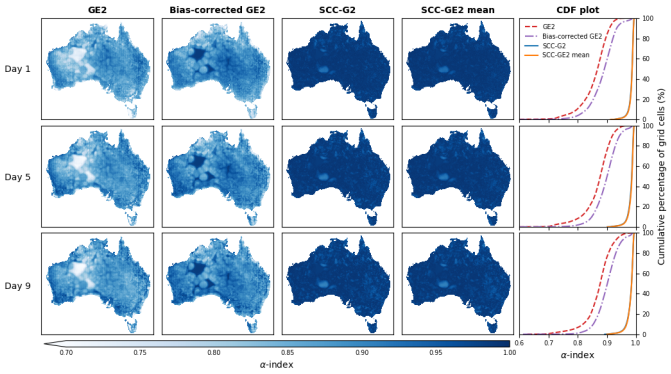
QJ_3952_Fig. 3.tiff



QJ_3952_Fig. 4.tiff



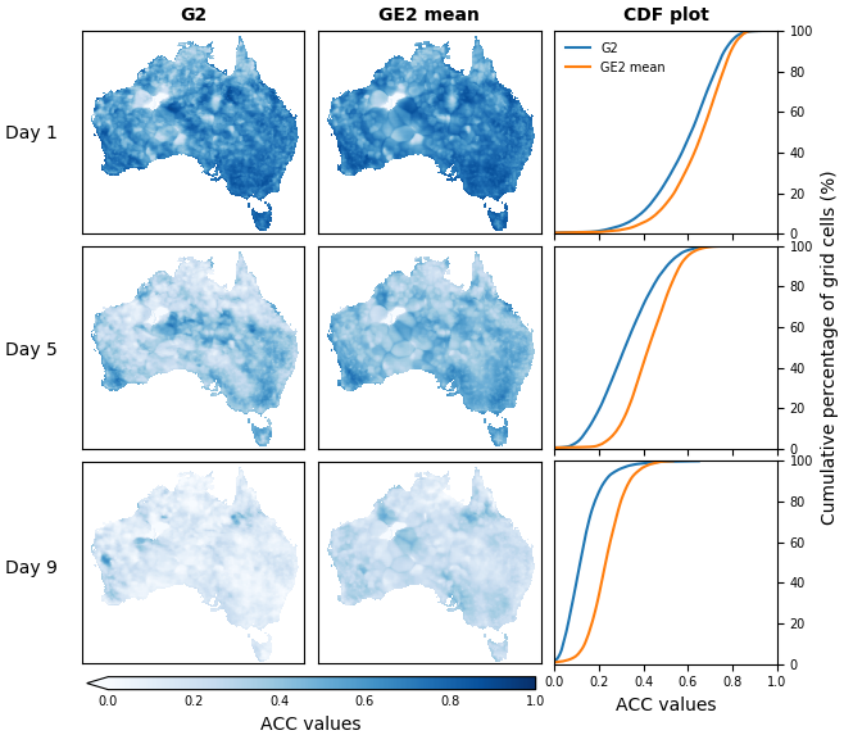
QJ_3952_Fig. 5.tiff



QJ_3952_Fig. 6.tiff

Which precipitation forecasts to use? Deterministic versus coarser-resolution ensemble NWP models

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



This study focuses on the comparative analysis of deterministic numerical weather prediction (NWP) forecasts and coarser-resolution ensemble NWP forecasts. A comprehensive comparison between these two kinds of forecasts is of significant reference value to both forecast users and NWP model developers. Our results suggest that for precipitation, coarser-resolution ensemble forecasts overall outperform deterministic forecasts, both before and after post-processing, under different types of climate conditions.

Table 1. Groups of single-value forecasts and ensemble forecasts and evaluation metrics

Groups	**Forecasts**	**Evaluation metrics**
Single-value forecasts	ACCESS-G2, ACCESS-GE2 mean	Correlation (ACC), Accuracy (CRPS skill score)
Ensemble forecasts	ACCESS-GE2, Simple bias-corrected ACCESS-GE2, SCC-calibrated ACCESS-G2, SCC-calibrated ACCESS-GE2 mean	Accuracy (CRPS skill score), Reliability (α-index)