Hydrological Sciences Journal

ISSN: 0262-6667 (Print) 2150-3435 (Online) Journal homepage: http://www.tandfonline.com/loi/thsj20

An evaluation of numerical weather prediction based rainfall forecasts

Mahshid Shahrban, Jeffrey P. Walker, Q. J. Wang, Alan Seed & Peter Steinle

To cite this article: Mahshid Shahrban, Jeffrey P. Walker, Q. J. Wang, Alan Seed & Peter Steinle (2016) An evaluation of numerical weather prediction based rainfall forecasts, Hydrological Sciences Journal, 61:15, 2704-2717, DOI: 10.1080/02626667.2016.1170131

To link to this article: http://dx.doi.org/10.1080/02626667.2016.1170131

Accepted author version posted online: 11 May 2016.
Published online: 05 Aug 2016.

HYDROLOGICAL SCIENCES JOURNAL – JOURNAL DES SCIENCES HYDROLOGIQUES, 2016
VOL. 61, NO. 15, 2704–2717
http://dx.doi.org/10.1080/02626667.2016.1170131

An evaluation of numerical weather prediction based rainfall forecasts

Mahshid Shahrban[a], Jeffrey P. Walker[a], Q. J. Wang[b], Alan Seed[c] and Peter Steinle[c]

[a]Department of Civil Engineering, Monash University, Clayton, VIC, Australia; [b]CSIRO Land and Water, Highett, VIC, Australia; [c]The Centre for Australian Weather and Climate Research, Melbourne, VIC, Australia

ABSTRACT

Assessment of forecast precipitation is required before it can be used as input to hydrological models. Using radar observations in southeastern Australia, forecast rainfall from the Australian Community Climate Earth-System Simulator (ACCESS) was evaluated for 2010 and 2011. Radar rain intensities were first calibrated to gauge rainfall data from four research rainfall stations at hourly time steps. It is shown that the Australian ACCESS model (ACCESS-A) overestimated rainfall in low precipitation areas and underestimated elevated accumulations in high rainfall areas. The forecast errors were found to be dependent on the rainfall magnitude. Since the cumulative rainfall observations varied across the area and through the year, the relative error (RE) in the forecasts varied considerably with space and time, such that there was no consistent bias across the study area. Moreover, further analysis indicated that both location and magnitude errors were the main sources of forecast uncertainties on hourly accumulations, while magnitude was the dominant error on the daily time scale. Consequently, the precipitation output from ACCESS-A may not be useful for direct application in hydrological modelling, and pre-processing approaches such as bias correction or exceedance probability correction will likely be necessary for application of the numerical weather prediction (NWP) outputs.

ARTICLE HISTORY
Received 27 October 2014
Accepted 7 January 2016

EDITOR
M.C. Acreman

ASSOCIATE EDITOR
A. Viglione

KEYWORDS
Hydrological modelling; forecast rainfall; rainfall evaluation; radar rainfall calibration

1 Introduction

Quantitative precipitation forecast (QPF; see Appendix A for a full listing of abbreviations used in this paper) from numerical weather prediction (NWP) models remains the primary source of rainfall data for input into hydrological forecasting models, other than a forecaster's intuition and some research-based products obtained by blending radar-based extrapolation nowcasts and NWP forecasts (Atencia *et al.*, 2010, Bowler *et al.* 2006, Wilson *et al.* 2010). However, the performance of flood forecasts from such hydrological models is highly dependent on the accuracy of the rainfall distribution and intensity.

While a large number of studies have assessed NWP precipitation forecasts, most of the long-term evaluations have been made against observations from raingauges. For example, Damrath *et al.* (2000) evaluated the QPF from the German Weather Service (DWD) using long-time verification statistics against 240 gauge stations over 7 years in Germany and Switzerland, including the frequency bias index (FBI) and the true skill statistics (TSS), and presented examples of application to flood events. They identified a problem in the parameterization of convective precipitation, which was expected to lead to comparatively poor QPF input to hydrological models in the case of summertime flash floods; details of FBI and TSS scores are given in Appendix B. Moreover, Clark and Hay (2004) examined 40 years of 8-day lead time precipitation forecasts from the National Centers for Environmental Prediction (NCEP) against a dense gauge network in the United States and showed that there were systematic precipitation biases exceeding 100% of the mean.

A small number of studies have used gauge observations for evaluation of the forecasts for individual events. Richard *et al.* (2003) assessed precipitation output from four different forecasting models including the Global Model (GM), Europa-Modell (EM), the Deutschland-Modell (DM) and Lokal-Modell (LM, replacing DM) against a high-density gauge station network for several events in Italy and Germany on an hourly and a daily basis. They showed that all models were able to produce the occurrence of the events, but the amount of forecast precipitation was poor, with no specific trend in over- or underestimation. They also indicated that the quality of the forecast is much more case-dependent than model-dependent. More recently, Roberts *et al.* (2009) showed improved forecast performance from the Met Office Unified Model (UM) for an event in 2005 in the northwest of England when using the model outputs with 1, 4 and 12 km grid spacing compared to raingauges. In this work, the 12 km model produced too little rain due to the inadequate representation of the orography. While the 4 km model also predicted too low rainfall for the highest rainfall amounts, it had a more accurate distribution of rainfall. The 1 km model had the most accurate distribution of rainfall but generated too much rain in general.

There are only a few studies that have assessed the forecast precipitation in Australia. McBride and Ebert (2000) verified 24 h precipitation forecasts from seven NWP models including GASP (Global ASsimilation and

CONTACT Mahshid Shahrban ma.shahrban@gmail.com

Prediction) and LAPS (Limited Area Prediction System) against 1° resolution operational daily rainfall analyses for 12 months, focusing on two main subregions in the country: the northern tropical monsoon regime and the southeastern subtropical regime. They used categorical scores including FBI, probability of detection (POD) and false alarm ratio (FAR). Descriptions of FBI, POD and FAR scores are given in Appendix B. They showed that most models significantly overestimated the area of rain throughout the summer months, while in winter all models except one underestimated the frequency or area of rain at thresholds above 2 mm/d. Ebert *et al.* (2003) reported the WGNE (Working Group of Numerical Experimentation) assessment of 24 h precipitation forecasts from several NWP models against gridded raingauge analysis with 1° resolution from 1997 to 2000 in different areas including Australia. They showed that, based on frequency bias, the Australian models consistently overestimated rain frequency in southeastern Australia. Shrestha *et al.* (2013) evaluated the quality of four NWP models from the Australian Community Climate Earth-System Simulator (ACCESS), including ACCESS-VT, ACCESS-A, ACCESS-R and ACCESS-G, against raingauges from 31 March 2010 to 30 March 2011. This evaluation was at point and catchment scales in the Ovens area, located in southeastern Australia. They showed that the skill of the models varied across the gauges and with forecast lead time, using bias score, which is the total difference between observations and forecasts relative to total observations; see Appendix B. The analysis showed that the ACCESS-VT and ACCESS-A models overestimated rainfall by up to 60% in low rainfall areas (low elevation) and underestimated rainfall by up to 30% in high rainfall areas (high elevation); ACCESS-R had a similar pattern but with much greater bias, while ACCESS-G had a systematic bias with underestimation up to 70% across all stations and increasing with altitude.

Although gauge observations have been the most common benchmark used for assessment of model rainfall forecasts, they are based on point measurements which suffer from inaccuracies due to errors in representivity when used in verification of forecasts averaged over a large area (Tustison *et al.* 2001). Furthermore, a dense gauge network is usually required to achieve proper evaluation of forecast rainfall over an area, and there can be large discrepancies between raingauge measurements even when co-located (Wood *et al.* 2000, Ciach 2003). In contrast, weather radar provides an alternative means of determining quantitative precipitation estimates (QPE) with fine spatial and temporal resolution over a large area (Morin and Gabella 2007). Because of its large spatial coverage relative to raingauges, and area-averaged response, radar is a useful source of data for verification of QPF, provided that the errors in radar-based precipitation estimates are corrected (Ebert *et al.* 2007, Rezacova *et al.* 2007). Radar is an active sensor that emits short pulses of microwave energy, and measures the power scattered back by raindrops as a reflectivity factor (Z). This reflectivity is then usually converted to a rain rate (R) through calibration of an empirical Z–R relationship such as:

$$Z = aR^b \quad (1)$$

where Z is radar reflectivity ($mm^6\ m^{-3}$), R is the rainfall rate (mm/h), and a and b are the radar parameters estimated using raingauge observations. The Z–R relationship requires the specification of parameters a and b, which are functions of both radar and rainfall characteristics (Battan 1973, Collier 1989, Rinehart 1991).

While there have been many efforts by researchers to use radar-based rainfall estimates for NWP forecast rainfall verification (Colle and Mass 1996, Johnson and Olsen 1998, Yu *et al.* 1998, Casati *et al.* 2004, Davis *et al.* 2006, Rezacova *et al.* 2007, Roberts 2008, Roberts and Lean 2008), this approach has not yet been conducted for evaluation of forecasts from Australian models. In addition, radar-based verifications have been used mostly for specific events or periods rather than long-time assessment. For example, Johnson and Olsen (Johnson and Olsen 1998) assessed forecast precipitation from the Arkansas–Red River Basin Forecast Center during May–June 1995 against Stage III radar data in the United States, by estimating cumulative frequency of mean bias and some categorical scores. Rezacova *et al.* (2007) applied the area-related root mean square error (RMSE) verification method for two local convective events in the Czech Republic using the Lokal-Modell of the Consortium for Small-Scale Modelling (LM COSMO) and adjusted radar data. This adjustment included combining daily radar precipitation with raingauge data. In addition, Vasić et al. (2007) evaluated precipitation data from the Canadian Global Environmental Multiscale (GEM) model, the United States Eta model and the Geostationary Operational Environmental Satellite (GOES) against radar observations for specific dates in the summer and autumn of 2004 and 2005 using categorical scores and a scale-dependent method. There are some other spatial verification studies using blended gauge-radar products to evaluate QPF data over a large area. For example, Lopez and Bauer (2007) used NCEP stage-IV analyses of precipitation data, which is a combination of raingauge data and high-resolution Doppler weather radars, to evaluate forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) over the mainland United States. For the evaluation method, they used continuous statistics such as bias, RMSE, correlation and categorical scores, over a month in spring 2005. Lopez (2011) used similar observational data and evaluation methods for two different periods (Apr–May and Sep–Oct) in 2010.

While traditional methods of statistics are useful for indicating the overall performance of model predictions in each grid box, especially over a long period, new methods have recently been used for spatial verification. These new methods are especially useful in representing the skill of mesoscale forecasts or event-based evaluations. Ebert and Mcbride (2000) and Marzban and Sandgathe (2006) developed object-based techniques that attribute the error to the displacement, intensity, and structure of precipitation forecasts. The object-based methods indicate an approach for verifying to what extent the forecast matches the observed location, shape and magnitude. However, these methods require sufficiently skilful forecasts to match rain objects from forecasts and

observations. Ebert and Mcbride (2000) applied this object-based method using 24 h forecast from the LAPS NWP model over a 4-year period in Australia, but against operational daily rain analyses. Davis *et al.* (2006) described and applied an object-based method to verify forecasts from the WRF (Weather Research and Forecasting) model against the NCEP stage-IV product for a short period from July to August 2001. They defined a matching approach based on the separation distance of the centroids of observation and forecast objects and calculated skill scores such as critical success index (CSI) and area bias; see Appendix B for more details.

There is also a wide range of neighbourhood verifications (fuzzy methods) that look for approximate agreement between the model and observations within different time and/or space windows (Casati *et al.* 2004, Ebert 2008, Roberts and Lean 2008). For example, Casati *et al.* (2004) used an intensity-scale approach to evaluate the forecast skill of the UK Met Office nowcasting system NIMROD. They compared the model against radar data analyses for six events in the UK as a function of precipitation intensity and spatial scale of the error. These methods provide the temporal or spatial scale at which forecasts reach a specific accuracy. Moreover, Roberts and Lean (2008) introduced and applied the fractions skill score (FSS; see Appendix B) method, comparing the forecast rainfall from the UM of the UK Met Office and radar rain fractional occurrences exceeding a given threshold for 10 convective events in summer 2003 and 2004. Roberts (2008) used spatial and temporal verification with the FSS to compare operational forecasts from the UM of the UK Met Office (grid spacing of 12 km) with radar observations for the whole of 2003. They found that the smallest useful scale for very localized rain with up to 1 h lead time is around 140 km and for 2–24 h lead time is 230 km, while for widespread rain the smallest useful scale is around 40 km for 0–1 h increasing to 85 km for 2–24 h lead times. Methods such as upscaling (Weygandt *et al.* 2004, Yates *et al.* 2006), multi-event contingency table (Atger 2001) and fuzzy logic (Damrath 2004) are other approaches defined for fuzzy verifications. Since the neighbourhood verification methods are based on the agreement within a spatial neighbourhood of the point of interest, these methods may not indicate perfect performance when applied to a perfect forecast, due to the influence of nearby grid boxes with medium or low forecast skill.

Against this background, the main objective of this paper is to evaluate the accuracy of forecast rainfall from the ACCESS-A NWP model over an area in the Murrumbidgee catchment located in southeastern Australia, using radar observations over a long period. Radar-based hourly rain intensities were calibrated to gauge observations for each rainfall event using data from an independent gauge monitoring network. The adopted radar rain intensities were used for evaluating the quality of rainfall forecasts for an area of coincident radar coverage. Different temporal resolutions were assessed by integrating the radar and NWP data into hourly and daily time scales.

2 Study area and data sets

The study area is located in southeastern Australia, including part of the 84 000 km^2 Murrumbidgee catchment (see Fig. 1). Based on streamgauge observations from the New South Wales Office of Water (http://www.water.nsw.gov.au/real time-data/default.aspx), the averages of median and maximum flows at the Wagga Wagga station, during the period of 2007–2012, are, respectively, 5 and 70 m^3/s over the dry years and 25 and 1450 m^3/s over the wet years. These flow values are generated in the tributaries from Burrinjuck and Blowering dams down to the Wagga Wagga station, with an

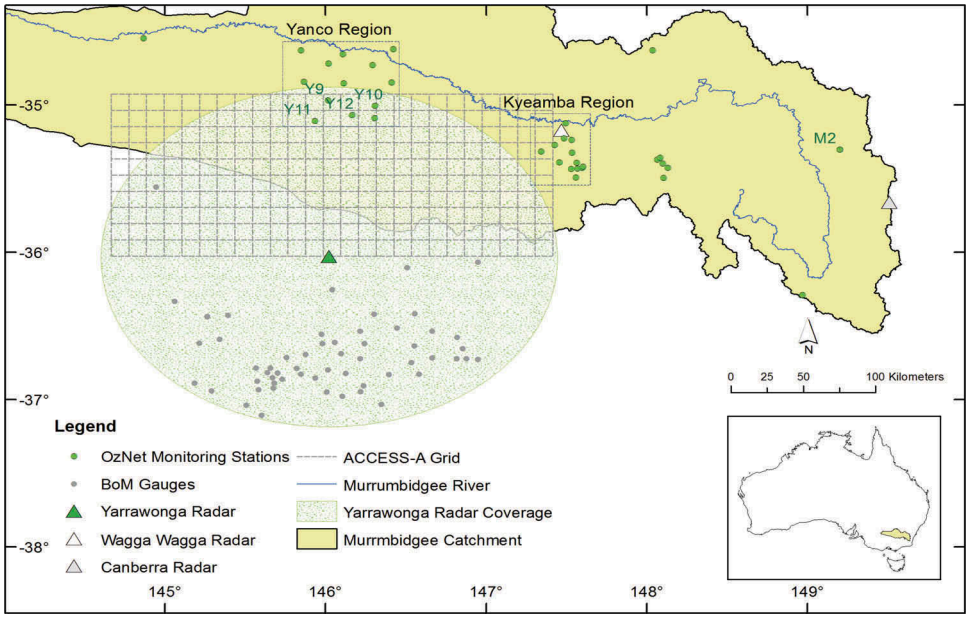

Figure 1. Location of Yarrawonga radar, radar coverage, OzNet raingauges and ACCESS-A grids coinciding with radar coverage in the study area. The horizontal and vertical axis labels are degrees of longitude and latitude, respectively.

area of about 10 886 km^2 (see Fig. 1). This area has been chosen for the study due to the availability of hydrological monitoring sites covered by the Yarrawonga weather radar. The observed gauge data used in this work are from the OzNet monitoring sites (Smith *et al.* 2012) located in the Murrumbidgee catchment (see also www.oznet.org.au).

Based on the Yarrawonga radar observations, stratiform rainfall with long duration and low to medium intensities (below 15 mm/h) is the dominant rain system in the study area in 2010 and 2011, with convective systems occurring only during the warm seasons. According to a review by Green *et al.* (2011), the average annual rainfall from 1898 to 2010 on the focus area in this work ranges from 350 mm on the western plains to 1100 mm in the higher elevations on the eastern part. Elevation in the Murrumbidgee catchment varies from over 2200 m in the eastern parts to less than 50 m on the western plains (Green *et al.* 2011), while the elevation in the study area ranges from about 80 m in the northwestern part to about 330 m in the southeastern part of the northern part of the radar coverage area. Typical hourly radar rain maps over the entire radar coverage are shown in Figure 2. Based on the radar observations, the annual rainfall across the study area in the northern part of the radar coverage area ranges from 350 to 800 mm in 2010 and from 410 to 940 mm in 2011.

Four OzNet raingauges from monitoring sites in the Yanco region (Y9, Y10, Y11 and Y12) are located within the radar coverage. These gauges provide rainfall data in 6-minute intervals and are used for calibrating the radar observations from January 2010 to December 2011 due to availability of gauge and forecast data during this time period. Y13 is not included in this work because of a large gap (March 2010 to May 2011) in the data due to an instrument breakdown during the study period. Calibrated data from the Yarrawonga radar of the Australian weather radar network are used for verification of the NWP forecast rainfall. This C-band Doppler radar, operated by the Bureau of Meteorology (BoM), scans rainfall every 10 minutes with 1 km resolution and a range of 128 km. It has partial coverage of the OzNet sites in the Yanco region, as shown in Figure 1. The radar scans over 14 elevations (0.5°, 0.9°, 1.3°, 1.8°, 2.4°, 3.1°, 4.2°, 5.6°, 7.4°, 10°, 13.3° 17.9°, 23.9° and 32°) with the same range (Rennie 2012), and operated properly during the entire study period of this work. There are two other radars in the Murrumbidgee catchment: the Wagga Wagga radar (C-band) located in the Kyeamba region and the Canberra radar (S-band). However, there is only one independent monitoring site (M2) in the Canberra radar coverage, and the Wagga Wagga radar is an old radar that was installed for qualitative radar observations and is not suitable for use in quantitative precipitation estimation.

The accuracy of radar-based rainfall estimates depends on (i) the reflectivity measurements from the radar and (ii) the parameters used for conversion of the reflectivity (Z) to rain rate (R). The estimation of rainfall from radar has been very challenging due to factors such as radar calibration (Joss and Lee 1995), measurement error and sampling uncertainty (Jordan *et al.* 2000, 2003, Piccolo and Chirico 2005),

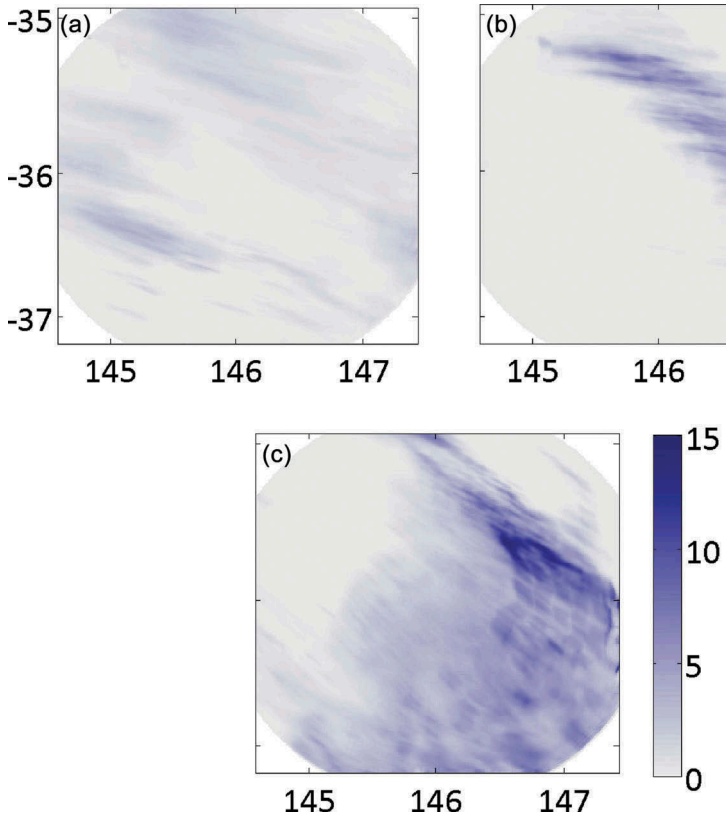

Figure 2. Typical hourly radar rain maps seen over the entire radar coverage area. The rain maps are from events on 24 April 2010 (a), 4 July 2011 (b) and 28 September 2011 (c). The horizontal and vertical axis labels are degrees of longitude and latitude, respectively.

attenuation (Hildebrand 1978), range effects (Chumchean *et al.* 2006, Gabella *et al.* 2006), and variability of raindrop size distributions on the *Z–R* relationship (Lee *et al.* 2009, Alfieri *et al.* 2010). The procedure used by the Australian BoM for estimating real-time radar rainfall consists of three main steps: (i) measurement of reflectivity and removal of measurement errors due to ground clutter, beam blocking, bright band, hail and range-dependent bias; (ii) conversion of the reflectivity to a rainfall rate; and (iii) mean field bias adjustment using the available real-time raingauge network. In the second step, radar rainfall of each pixel is estimated based on the *Z–R* relationship developed separately for stratiform or convective rainfall types. In the last step, based on a Kalman filtering approach, a spatially uniform bias adjustment factor is used to correct the initial radar rainfall estimates on hourly time steps (Chumchean *et al.* 2006, 2008). The raingauges within the radar coverage used operationally by the BoM for radar rainfall estimation from the Yarrawonga radar are shown in Figure 1. The BoM raingauges are mostly located in the southern part of the radar coverage due to flood warning priorities. Therefore, even though the errors in the *Z–R* conversion and mean field bias have been mainly reduced in the three steps of the rainfall estimation procedure, there is still likely to be a bias in the radar data due to the lack of sufficient raingauges. It should be mentioned that the focused study area of this work is approximately flat, so the effect of topography in the radar data used in this study is not significant.

ACCESS-A (Bom 2010, Puri *et al.* 2013) forecast rainfall is used as the forecast data over the years 2010 and 2011, due to the effective resolution (12 km) and coverage of the study area. The new operational ACCESS NWP systems from the Australian BoM replace the GASP, LAPS, TXLAPS and MESOLAPS NWP systems in Australia. ACCESS became operational for NWP application in 2010 and includes several models with different domains, resolutions and forecast lead times. These models include ACCESS-G (global, 80 km), ACCESS-R (regional, 37.5 km), ACCESS-T (tropical, 37.5 m), ACCESS-A (Australia, 12 km), ACCESS-C (cities, 5 km) and ACCESS-TC (tropical cyclone, 12 km). The ACCESS system uses a four-dimensional variational data assimilation (4D-Var) scheme which takes into account various observations with different times or locations for initializing the model in a dynamically consistent way. All models except ACCESS-G use boundary conditions that are provided by a coarser resolution ACCESS mode. For example, ACCESS-R and ACCESS-T are nested inside the previous run of ACCESS-G, while ACCESS-A and ACCESS-C are nested inside the concurrent run of ACCESS-R. ACCESS-A has four runs per day with base times of 00:00, 06:00, 12:00 and 18:00 UTC and forecast duration of 48 h. Based on the study by Shahrban *et al.* (2011), the average RMSE and mean error of ACCESS-A on an hourly time step are lowest for lead times of 13–24 h among other possible lead times (1–12, 25–36 and 37–48 h). Therefore, the forecast data for lead times of 13 to 24 h and from base times of 00:00 and 12:00 are used to produce the continuous forecast time series in this work. The lead time of 13–24 h avoids both the spin-up problem in the shorter lead times and the forecast uncertainties from the longer lead times. The OzNet raingauges, the radar coverage, and ACCESS-A grid in the study area are presented in Figure 1.

3 Methodology

Understanding of radar rainfall uncertainties and rainfall processes is dependent on the availability of a dense raingauge network for the accurate estimation of the parameters for the *Z–R* relationship (Krajewski *et al.* 2010, Peleg *et al.* 2013). The *Z–R* relationship is influenced by the raindrop size distribution, which can vary greatly within a given event, and from one rainfall event to another (Doelling *et al.* 1998, Atlas *et al.* 1999, Steiner and Smith 2000). Therefore, any correction of this relationship required for accurate radar rainfall estimates should be done for individual events rather than over long periods (Alfieri *et al.* 2010). Since the raingauges used by the BoM for estimation of radar rainfall intensities are mainly located in the southern part of the radar coverage, where orographic enhancement is important (see Fig. 1), radar rain rate adjustment in the northern part of the radar domain was needed to decrease the errors brought by calibration to the BoM raingauges alone. Thus, before using radar observations for evaluation of the ACCESS-A forecast rainfall, the radar rainfall intensities were adjusted using a new power-law relationship for each event over the entire northern part of the radar coverage. The adjusted radar rain intensities were then used for evaluating the rainfall forecasts for a coincident area in the northern radar coverage.

For adjusting radar rainfall rates, the radar 10 min rainfall data were accumulated to hourly time steps, by adding six consecutive 10 min accumulations. Then, new power-law relationships between radar rain intensities and independent gauge rainfall rates from four available raingauges were calibrated for each event by estimating the parameters α and β according to:

$$G = \alpha R^{\beta} \tag{2}$$

where G is the gauge rainfall rate intensity (mm/h) and R is the radar rainfall rate (mm/h) in the corresponding radar pixel. This new relationship is based on the power-law relationship typically used in the initial conversion of radar reflectivity measurements to rainfall intensity (Battan 1973, Collier 1989, Rinehart 1991) according to Equation (1). In the adjustment process, the radar rain rates were brought as close as possible to the gauge rates at hourly time steps by minimizing the error between radar and raingauge estimates. Each parameter set (α and β), which was estimated for each event, was used to calculate the new radar rain rates for the individual event over the entire northern part of the radar coverage using the power-law relationship in Equation (2). This event-dependent calibration method accounts for the dependency of the *Z–R* relationship on rainfall characteristics such as rainfall drop size distribution, which varies in both space and time (Atlas *et al.* 1999, Mapiam *et al.* 2009). The methodology for adjusting radar rainfall rates was based on the algorithm proposed by Fields *et al.* (2004). Similarly, Mapiam and Sriwongsitanon (2008) used this method for adjusting the *Z–R* relationship in Equation (1) using a linear regression

between the radar rainfall and the observed gauge rainfall in the Ping River basin in northern Thailand, but the exponent b in Equation (1) was fixed, assuming that b is less sensitive than the parameter a.

For verification of NWP forecast rainfall against radar data, the average of adjusted rainfall rates over the nearest radar pixels which were within the ACCESS grid spacing was calculated. Based on expert judgment, a minimum value of 5 mm/d was used as a threshold for both observation and forecast over the entire study area for separating rain storms from drizzle. This means that all daily rain maps containing at least one pixel with 5 mm/d in the radar observations and/or forecasts were used in the evaluation. RMSE, RE (relative error, or in other words bias) and ME (mean error) were used as traditional verification metrics to identify the pixel-by-pixel differences between the model and the average of radar adjusted rates in all 12 × 12 km ACCESS-A grids over the northern range of the radar coverage where the radar adjustment was implemented. RMSE is one of the most common methods of verification and represents the average magnitude of forecast errors. RE is the total difference between forecasts and radar observations over the time interval divided by total radar observation, and ME is the average of differences between radar observations and forecasts over the same time interval. The ME can be used to identify the arithmetic average of the forecast errors, while the RE is useful to assess the performance of the forecasts compared with the total radar observations. RMSE was calculated for hourly and daily time scales to test the improvement in daily accumulations due to expected decreases in possible timing and location errors in longer term accumulations. As such, the useful time scale of NWP rainfall forecasts could be assessed. Moreover, the contingency table of Ebert and Mcbride (2000), as described in Table 1, was calculated at hourly and daily time scales to better relate the rainfall forecast errors to factors such as wrong timing, wrong location, and error in rain amount. This contingency table is different from the traditional contingency table with standard verifications using categorical statistics such as bias score, probability of detection and false alarm ratio (Doswell *et al.* 1990, Wilks 1995). The method in Table 1 compares the observed and forecast location and magnitude over the entire study domain, and calculates the categorical scores based on the overall rain map in the study area. To identify whether the location of the forecasts was adequately predicted, the distance D separating the centroids of observed and forecast rain objects should be such that $D < R_{\text{eff}}$ and R_{eff} is the effective radius of an observed rain object. This approach is based on the method used by Ebert and Mcbride (2000) and Davis *et al.* (2006) for diagnosing forecast location errors. All rain maps from hourly and daily accumulations with D smaller than R_{eff} are accepted as a good forecast location and categorized as "close" in Table 1.

To decide whether the magnitude of the forecast is correctly predicted or not, several categories have been defined for both hourly and daily intensities. The categories are: 0.5–1, 1–2, 2–5, 5–10, 10–20, and >20 mm/h for hourly rates, and 5–10, 10–20, 20–50, 50–100, and >100 mm/d for daily rates. The categories for daily rates are quite similar to those used by Ebert and Mcbride (2000). However, a minimum value of 5 mm/d was used as a threshold for at least one pixel over the study area, to include a rain map on hourly or daily calculations. For the hourly time scale the categories were approximately derived from the daily categories by converting the ranges to hourly rates with some changes in the values. In a time step with a "close" forecast location, if the forecast maximum intensity was within the same category as the maximum observed value, then the magnitude of the rain in the whole domain was assumed to be well predicted for that time step, making it a "hit". Otherwise, if the maximum forecast rate was more than one category greater than the maximum observed rate the forecast was defined as an "overestimate", while if it was more than one category less than the observation the forecast was defined as an "underestimate". If D was equal to or larger than R_{eff}, the location was not correctly predicted and the forecast was defined to be in the "far" category. If the predicted maximum intensity was approximately similar to the maximum observed value it was defined as a "missed location", but if the maximum intensity was categorized in a group smaller than the maximum observed value it was defined as a "missed event". Otherwise, if it was greater than the observed category it was defined as a "false alarm". This methodology for comparing rain magnitudes is based on the approach used in Ebert and Mcbride (2000) for daily events.

Table 1. Schematic explanation of contingency table for rain events.

	Too little	Approx. correct	Too much
Close	Underestimate	Hit	Overestimate
Far	Missed event	Missed location	False alarm

4 Results

4.1 Calibration of radar rainfall

For calibrating radar rain rates during the study period of 2010–2011, hourly accumulations were calculated for rainfall rates from four Yanco gauges as well as coincident radar pixels in the northern part of the radar coverage. New spatially uniform parameters α and β were estimated for 87 separate events over the 2-year study period, using the average hourly rainfall for the four radar grid cells (1 × 1 km) and the four corresponding gauges. The parameters were selected to yield the best power-law fit between the radar and gauge rainfall rates across the four gauges. The nonlinear least squares method was used for fitting the rates to the new relationships. Before fitting, the outliers for each event were excluded from the fitting. The outliers for each event were identified as the pairs with gauge-based rain rate less than the 10th percentile of gauge rain rates, whilst the radar-based rate was greater than the 90th percentile of radar rates in an event, or *vice versa*.

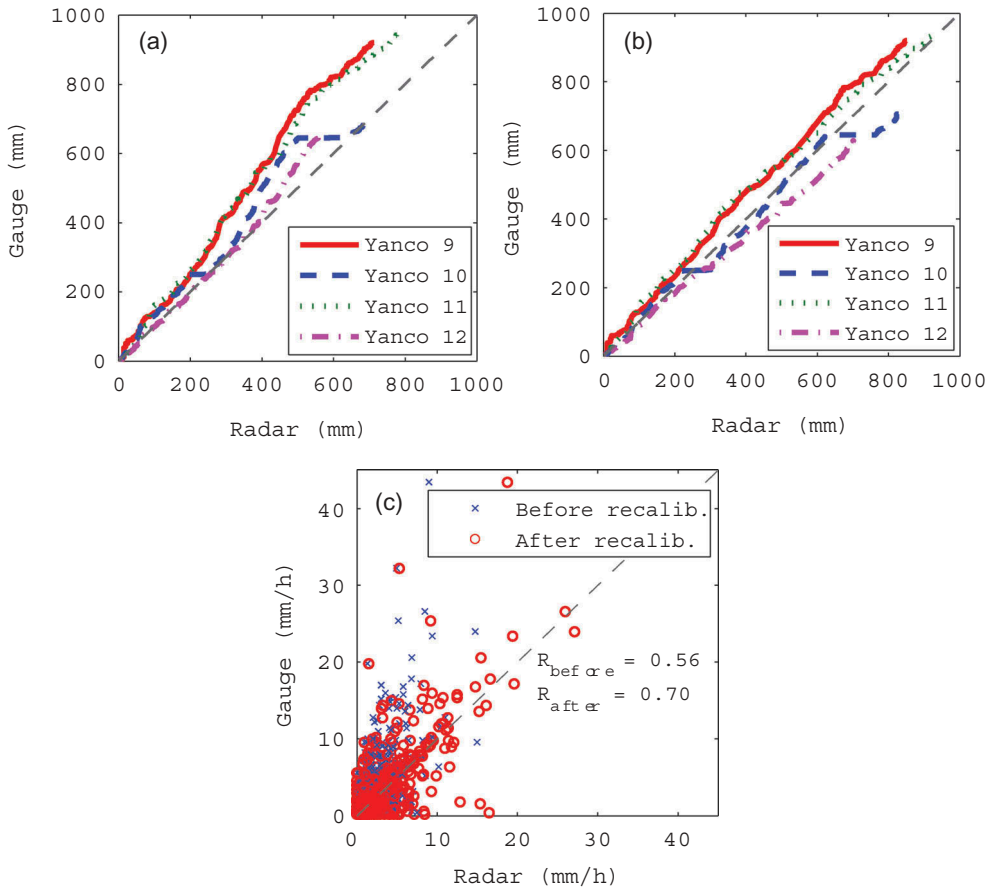

Figure 3. Cumulative rainfall plots for radar before calibration (a) and after calibration (b); scatter plot for radar rainfall rates compared with gauge observations before and after radar recalibration (c). Data from January 2010 to December 2011.

The new parameters were applied to the entire northern domain of the radar coverage for each event, in order to derive calibrated radar hourly rates. The parameters α and β had a temporal range of 0.05–5.18 and 0.05–3.27, respectively, for the events, with no specific seasonal trend seen in the values of the parameters. In the study by Mapiam and Sriwongsitanon (2008), the parameter α was estimated to be equal to 1.868, while the parameter β was fixed (equal to 1). Figure 3 shows the cumulative rainfall for the four raingauge locations and a scatter plot of radar rainfall compared with gauge observations before and after calibration. In the cumulative rainfall plot, created by adding up the hourly rainfall rates from January 2010 to December 2011; the time steps with missing gauge and/or radar values were removed from the calculations. The figure shows data for the full 2-year period. The scatter plot is shown for the four Yanco gauges used for radar calibration. It can be seen from the cumulative rainfall plots that the bias in the radar rainfall estimates was reduced by the calibration. In the scatter plot, apart from some points of overestimation, most of the radar rates (mainly underestimations) were improved, showing that the bias in the overall radar estimates was removed through the calibration process. The bias and RMSE between gauge and radar rainfall rates were decreased from −14% and 2 mm/h to 3% and 1.7 mm/h, respectively, after radar calibration in four grid boxes containing the Yanco sites for the whole study period.

4.2 Evaluation of ACCESS-A using continuous metrics

After calibrating the radar rainfall rates, the radar hourly rainfall accumulations with 1 km grid spacing were aggregated to the ACCESS-A 12 km grid spacing (as explained in Section 3) for verification of the forecast rainfall in the northern half of the radar domain. In order to compare the forecasts with gauge or radar observations, cumulative rainfall from ACCESS-A is shown against cumulative gauge point measurements and cumulative adjusted radar rainfall in Figure 4(a) and (b), and a scatter plot of ACCESS-A is presented against the radar adjusted rates in Figure 4(c) for the entire 2-year period. Here, radar and ACCESS-A are both rainfall over the 12 × 12 km pixels containing the raingauges. From the cumulative plots, ACCESS-A mostly overestimated rainfall compared to the gauges and radar with the ME and RMSE of 12% and 1.3 mm/h, respectively, when compared to the radar data.

The total annual radar-adjusted observations across all 12 × 12 km ACCESS-A pixels over the northern part of the radar coverage is shown for 2010 and 2011 in Figure 5(a), varying from 350 to 800 mm in 2010 and from 410 to 940 mm in 2011 across the area. From the figures, it can be seen that there is a similar pattern in the annual rainfall observations over the area in 2010 and 2011. In these figures, there could be possible underestimation of the rainfall near the edge of the radar range, associated to residual bias due to

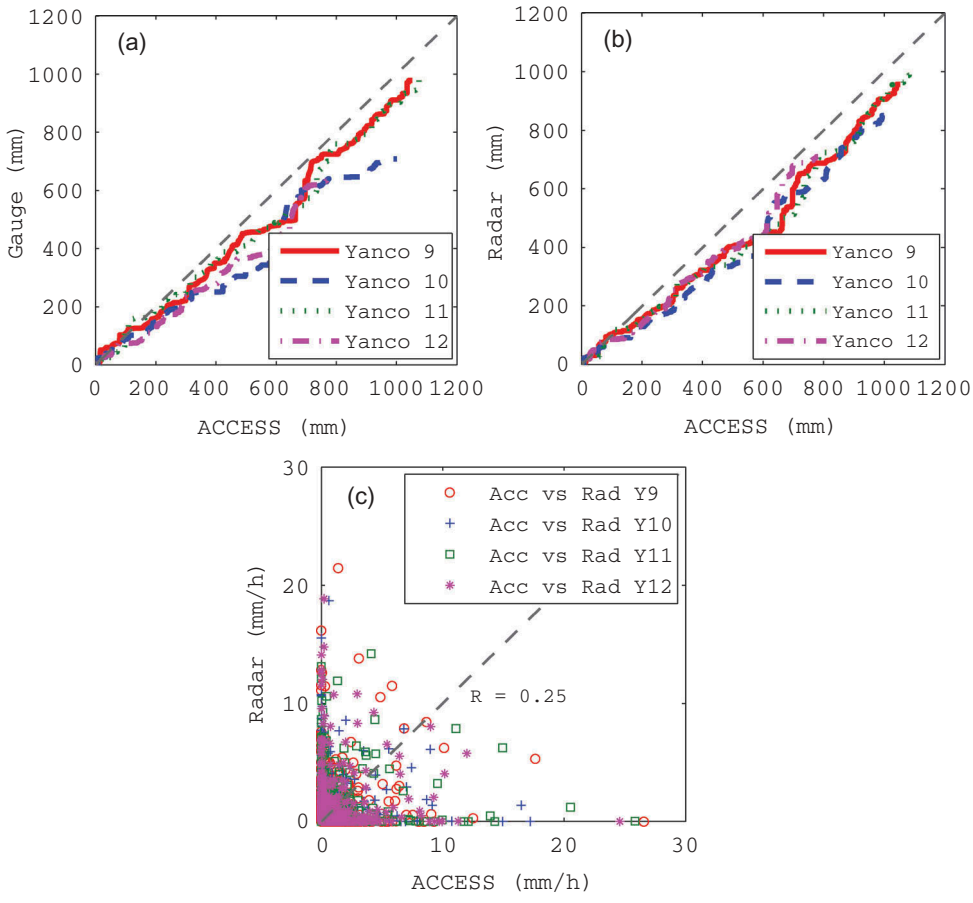

Figure 4. Cumulative rainfall plots for ACCESS-A compared with gauge (a) and calibrated radar (b); scatter plot for ACCESS-A rainfall rates compared with calibrated radar observations (c). Data from January 2010 to December 2011.

the vertical profile of reflectivity, while radar clutter might be a reason for the decrease of the rainfall near the radar location in the central-lower part of the image. Figures 5(b) and (c) depict the spatial variation of annual RE (%) and RMSE (mm/h) for each year across all ACCESS-A pixels in the study area. The RE varied between −22% and +59% in 2010 and −38% and 14% in 2011 across the pixels in the study area, as shown in Figure 5 (b). It can also be seen from this figure that ACCESS-A performance changed across the pixels in the study area, and had a very different response in 2010 to that in 2011. It mainly overestimated rainfall in 2010 (errors are shown in blue), with very small relative errors in the middle parts (grey) in this year. However, it underestimated rainfall in most of the central parts (yellow to red) in 2011. Comparing Figure 5(a) with Figure 5 (b), ACCESS-A can be seen to overestimate the areas with low rainfall observations in 2010 and underestimate the areas with high rainfall observations in 2011, while the error was nearly zero in the areas with moderate rainfall. In Figure 5(c), RMSE was not very different between 2010 and 2011, at 1.4–3.7 mm/h in 2010 and 1.2– 2.9 mm/h in 2011 across the pixels in the study area.

To account for the differences in the errors through the months, the hourly RE, ME and RMSE were investigated separately for 3-month periods, as shown in Figures 6–8, with the total radar observation across the pixels presented for each period in Figure 9. Figures 6 and 7 show that there is no consistent error in ACCESS-A forecasts across the study area through the 3-month periods. The variations in the errors seen between the 3-month periods in Figure 6 are mainly related to the dependency of the model skill on actual rainfall observations, which varied considerably across the study area. This means that the model underestimated rainfall during the periods with heavy rain rates and overestimated light rainfall events. Figure 6 also reveals that ACCESS-A showed strong underestimation in January–March 2011 as the model did not successfully predict the heavy rainfall from convective storms during the summer. In addition, from Figures 6 and 7, it can be seen that that the ME varied between −1 and 1 mm for each 3-month period, while the range of RE was very high for each period across all pixels. For example, in April–June 2010 the ME ranged from −0.22 mm to 0.87 mm across the pixels while RE varied from −32% to 270% of total observed rainfall across the pixels. Indeed, in some periods of the year RE was as high as 60% across the study area with ACCESS-A underestimating rainfall, or it was as much as 270% with ACCESS-A overestimating rainfall. The RE was more than 100% in January–March and July–September 2010, and was more than 200% in April–June 2010 and 2011. Since the model overestimated low rainfall values, the RE was very high where it was positive.

By comparing the RMSE in Figure 8 with total observed rainfall in Figure 9, it is clear that RMSE was high in periods

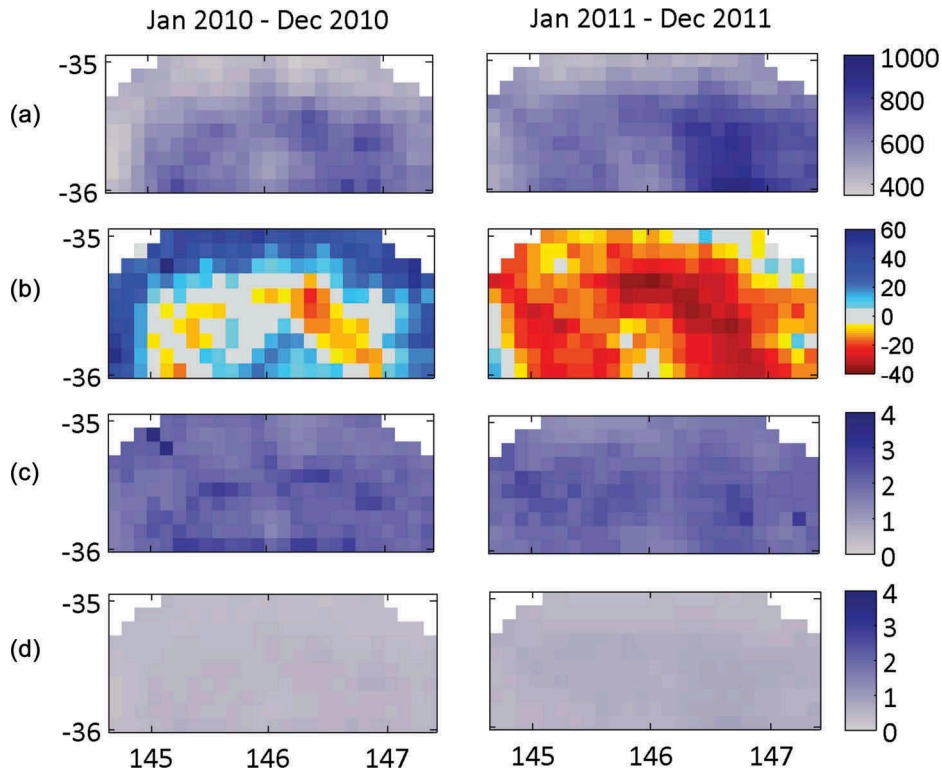

Figure 5. Total calibrated radar observations (mm) (a); relative error (%) between hourly ACCESS-A and calibrated radar (b); RMSE (mm/h) between hourly (c) and daily (d) ACCESS-A and calibrated radar. Data are for 2010 (left) and 2011 (right). Note that the white pixels on the top corners of the images are NA radar data. A consistent colour scale has been used to permit easy cross-comparison. The horizontal and vertical axis labels are degrees of longitude and latitude, respectively.

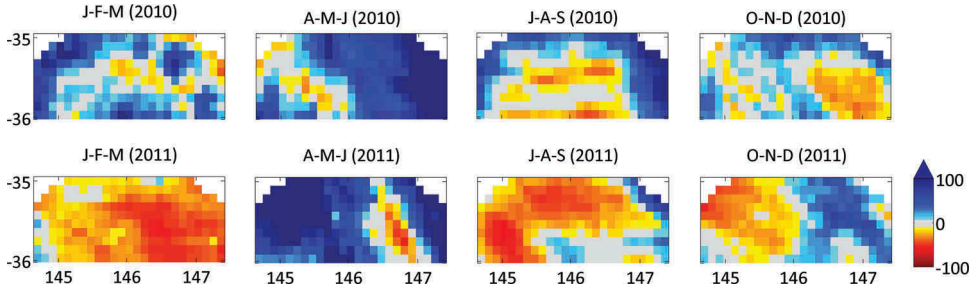

Figure 6. Relative error (%) between hourly ACCESS-A and calibrated radar over 3-month periods for 2010 and 2011. Note that the white pixels on the top corners of the images are NA radar data. A consistent colour scale has been used to permit easy cross-comparison; however, the maximum errors for J-F-M, A-M-J and J-A-S 2010 and A-M-J 2011 are off the scale in the figure, as indicated by the arrow. The horizontal and vertical axis labels are degrees of longitude and latitude, respectively.

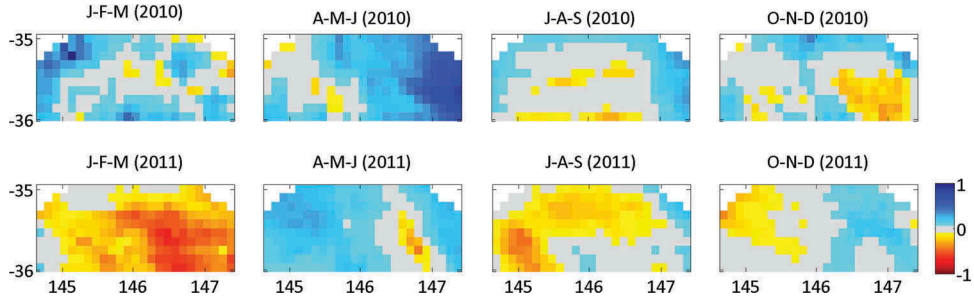

Figure 7. Mean error (mm/h) between hourly ACCESS-A and calibrated radar over 3-month periods for 2010 and 2011. Note that the white pixels on the top corners of the images are NA radar data. A consistent colour scale has been used to permit easy cross-comparison. The horizontal and vertical axis labels are degrees of longitude and latitude, respectively.

with medium to high rainfall observations over the periods. For example, maximum RMSE was in the periods October–December 2010 and January–March 2011, with total observed rainfall over 3-month period varying from 1.4 to 3.4 mm/h and 1.6 to 4.6 mm/h, respectively. In order to investigate the extent to which the error decreases with accumulation period,

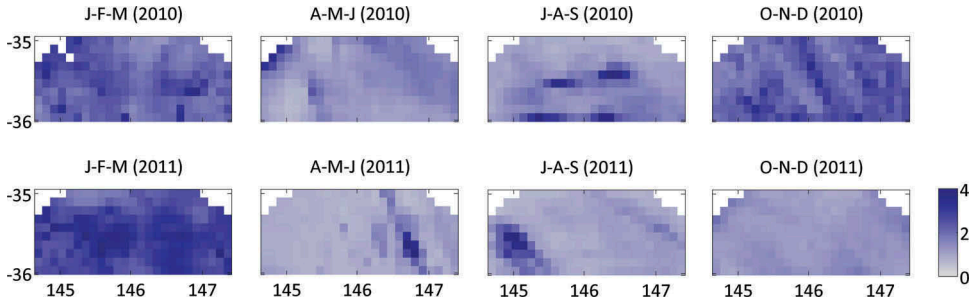

Figure 8. RMSE (mm/h) between hourly ACCESS-A and calibrated radar over 3-month periods for 2010 and 2011. Note that the white pixels on the top corners of the images are NA radar data. A consistent colour scale has been used to permit easy cross-comparison. The horizontal and vertical axis labels are degrees of longitude and latitude, respectively.

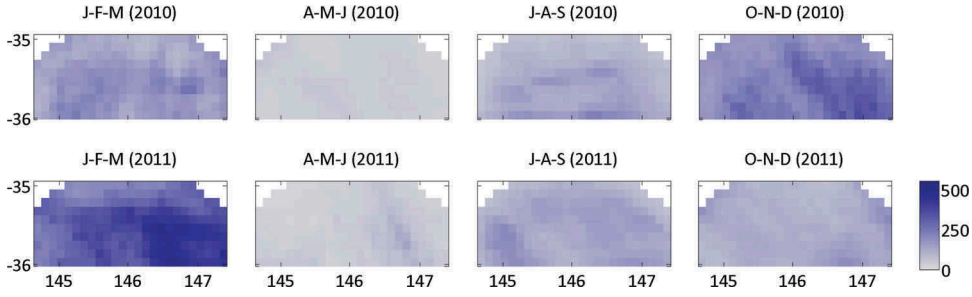

Figure 9. Total calibrated radar rainfall (mm) over 3-month periods for 2010 and 2011. Note that the white pixels on the top corners of the images are NA radar data. A consistent colour scale has been used to permit easy cross-comparison. The horizontal and vertical axis labels are degrees of longitude and latitude, respectively.

the RMSE was also calculated on daily accumulations, with the results shown in Figure 5(d) for 2010 and 2011 respectively. The daily RMSE ranged from 0.3 to 0.7 mm/h (7.2–16.8 mm/d) in 2010 and from 0.4 to 0.9 mm/h (9.6–21.6 mm/d) in 2011 across the pixels. From these results, the areal averages of RME were decreased by 78% and 68% for 2010 and 2011, respectively, in daily time steps. However, the range of RMSE on the daily time scale was still high (7.2–16.8 mm/d for 2010 and 9.6–21.6 mm/d for 2011).

4.3 Evaluation of ACCESS-A using contingency table

To evaluate the importance of timing as a source of error in the forecasts, relative to errors in the rainfall volume and location over the entire study domain, the contingency table in Table 1 was calculated for hourly and daily accumulations. To produce this table, rain events that did not contain at least one observed and/or forecast pixel with more than 5 mm/d rainfall were removed, and thresholds of 0.1 mm/h and 1.0 mm/d were used to distinguish between rain and no-rain pixels for hourly and daily analysis, respectively. All hourly and daily rainfall amounts below these thresholds were considered zero. To distinguish whether the forecast location was sufficiently good or not, the effective radius of the observed rain object and the distance between centroids of observed and forecast rain objects were calculated for each time step (see Section 3). The effective radius was estimated as the radius of a circular region having the same area as the observed rain area, and the centroid of the observed (or forecast) rain object was calculated as the arithmetic mean location of all observed (or forecast) rain pixels in a rain map. For comparing forecast magnitude and radar rain rates, the categories for hourly and daily time scales defined in Section 3 were used. The results for the contingency table are presented in Table 2 as a percentage, being the number of hours/days for each event type as defined in Table 1, divided by the total hours/days (excluding no rain observations and forecasts). This table indicates that 53% of the hours had wrong locations including 14%, 24% and 15% for missed location, missed events and false alarms, while 47% of the forecasts were within the correct location. For daily accumulations, the percentages of wrong locations decreased to 6% for missed location, missed events and false alarms, and consequently the total forecasts with correct locations increased to 82%, due to reducing the timing errors by using longer accumulation time. However, only 21% of these days were well forecast (hits), showing that a large proportion of daily rain images (79%) had forecasts with wrong magnitude and/or location.

Table 2. Contingency table for hourly and daily rainfall over the entire area from January 2010 to December 2011.

	Hits (%)	Underestimates (%)	Overestimates (%)	Missed locations (%)	Missed events (%)	False alarms (%)
Hourly	13	25	9	14	24	15
Daily	21	28	33	6	6	6

5 Discussion

The goal of this work was to assess the errors in operational ACCESS-A NWP rainfall forecasts over a 2-year period in Australia. This assessment is important for understanding the impacts on flood forecasting when using NWP rainfall forecasts as input. The Australian-domain model, ACCESS-A, was evaluated against adjusted rainfall observations from the Yarrawonga radar from January 2010 to December 2011. The evaluation of NWP data was based on RE (relative error), ME (mean error), RMSE and a contingency table. For this purpose, radar rainfall intensities were adjusted to independent raingauges by estimating a new relationship between radar and gauge rates. Radar-rainfall estimates can provide the broad-scale observations required for verifying model precipitation forecasts, provided the errors in radar-based rainfall are corrected. This work was based on the assumption that, after adjusting the radar, the error in radar data has been sufficiently minimized to be useful in evaluation of forecast rainfall data.

Based on the results from annual accumulations of RE and ME, the forecast skill was found to be different in 2010 and 2011, being highly dependent on the rainfall observations over the study area. Overall, the ACCESS forecasts overestimated rainfall in areas with low total rainfall and underestimated rainfall in high rainfall areas across the study area. The variation of these errors through the 3-month periods also showed that the skill of the model varied across the study area.

The range of RE across the study area was from −22% to +59% in 2010 and from −38% to +14% in 2011 (see Fig. 5(b)). The range obtained here is similar to the errors estimated by Shrestha *et al.* (2013) from March 2010 to March 2011 in the Ovens catchment in southeastern Australia using gauge data alone. They showed that ACCESS-A overestimated precipitation in dry, low elevation areas by up to 60% and underestimated it in wet, high elevation areas by up to 30%. However, the study area here is nearly flat with an average slope between 0% and 1.8%. Therefore, this study shows that the error is more likely to be dependent on observed rain magnitude through time than on elevation, as was proposed by Shrestha *et al.* (2013).

The range of RMSE found in this work is quite consistent between 2010 and 2011 across the study area, being 1.4–3.7 mm/h and 1.2–2.9 mm/h for the hourly time scale for 2010 and 2011, respectively (see Fig. 5(c)).While there was a large decrease in the annual RMSE on the daily time scale across the area, as shown in Figure 5(d), compared to the hourly time scale, the errors were still high across the area for daily accumulations (7.2–16.8 mm/d in 2010 and 9.6–21.6 mm/d in 2011). The range of RMSE across the area on daily accumulations here agrees with the RMSE found in the study by Shrestha *et al.* (2013), with values from 6.4 to 14.6 mm/d for the ACCESS-A model.

The continuous verification used in this study for a long period of data was able to give a proper view of timing errors when comparing hourly and daily scales, but could not differentiate between other sources of error such as location and rain volume. A contingency table was used for investigation of these different sources of error. In order to distinguish between forecasts with correct location and displacement, the distance between the centroids of rain objects was compared to the effective radius of the observed rain object. This approach allowed for an approximate evaluation of forecast location assuming that the rain forecast object initially matches the observed object. From the contingency table (Table 2), it is seen that a large fraction of the ACCESS-A forecasts on hourly time steps (53%) were found to suffer from a spatial displacement, from which 39% had the wrong magnitude (missed events or false alarms). However, 34% of the forecasts had the correct location but wrong magnitude and only 13% of the forecasts were identified as a "hit".

The results from the contingency table showed that the deficiency in ACCESS-A forecast on the hourly time scale is related to both imperfect location and wrong magnitude. The large effect of displacement error in forecast uncertainties obtained here is consistent with the results from Ebert *et al.* (2004), which indicated that 1 h forecasts from nowcast algorithms may have position errors of up to 80 km, with a mean error of about 15–30 km. The use of daily accumulations in the contingency table (Table 2) shows that the frequency of events with wrong location decreased substantially regardless of the magnitude, due to removal of the timing errors. However, only 21% of the days had perfect forecast location and magnitude and a large fraction of the forecasts (61%) had wrong magnitude with correct location. Velasco-Forero *et al.* (2009) have shown previously that the spatial correlation of radar rainfall fields could be as small as 0.3 over distances as short as 20 km. Therefore, for the study area here (100×250 km^2), it is expected that forecasts with wrong location would have similarly low correlations with observations. The moderate improvement from hourly to daily accumulation indicated that the location deficiency on the hourly scale, which was mainly related to the timing errors, was reduced on daily accumulations, while wrong magnitude was still the main source of errors on the daily time scale. Moreover, Kobold and Sušelj (2005) showed that 15% deviation in rainfall input into rainfall–runoff models led to 20% error in peak discharge predictions. Consequently the errors obtained in this study indicate that the raw ACCESS-A forecasts may not be sufficiently accurate to be used in hydrological forecasting, since a large fraction of the study area had relative errors more than 15% (Fig. 7). Post-processing methods, such as the probability modelling approach or exceedance probability correction, can be used to remove biases and reliably quantify forecast uncertainties. For example, using exceedance probability of observational data, the forecasts could possibly be corrected so that their probabilities match those observed.

6 Conclusions

Forecast precipitation data are necessary for forecasting of flood events. Evaluation of precipitation from NWP models has been an important subject during the past decade. This study has evaluated ACCEES-A precipitation forecasts during a 13–24 h period against adjusted weather radar data.

The results revealed that the skill of the NWP precipitation forecasts varied across the study area and through time, being highly dependent on the rainfall observations over the study area. Use of daily accumulations of ACCESS-A resulted in decreased errors compared with the hourly time scale, but the forecast skill was still not appropriate for hydrological modelling applications. In addition, based on a contingency table, both location and magnitude errors were the main sources of forecast uncertainties on hourly accumulations, while wrong magnitude was the dominant source of error on the daily time scale. Consequently, the results from this work suggest that without error correction (i) the raw hourly forecasts are not sufficiently accurate to be used for flood forecasting at the scale of ACCESS-A and (ii) the improvement in daily forecast accumulations is still not enough to allow for hydrological applications at the spatial scale of the NWP forecast model.

Acknowledgements

The authors would like to thank David Robertson from CSIRO Land and Water for his helpful comments on this work. In addition, they would like to acknowledge the cooperation of Monash University and Melbourne University for supporting the OzNet hydrological monitoring data and Sandra Monerris for processing that data. The authors also wish to thank Dr Tomeu Rigo for reviewing the paper and his valuable comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study has been supported by scholarships from Monash University and CSIRO.

References

Alfieri, L., Claps, P., and Laio, F., 2010. Time-dependent *Z-R* relationships for estimating rainfall fields from radar measurements. *Natural Hazards and Earth System Science*, 10, 149–158. doi:10.5194/nhess-10-149-2010

Atencia, A., *et al*., 2010. Improving QPF by blending techniques at the meteorological service of Catalonia. *National Hazards and Earth System Science*, 10, 1443–1455. doi:10.5194/nhess-10-1443-2010

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

Atlas, D., *et al*., 1999. Systematic variation of drop size and radar-rainfall relations. *Journal of Geophysical Research: Atmospheres*, 104, 6155–6169. doi:10.1029/1998JD200098

Battan, L.J., 1973. Radar observation of the atmosphere. L. J. Battan (The University of Chicago Press) 1973. PP X, 324; 125 figures, 21 tables. £7·15. *Quarterly Journal of the Royal Meteorological Society*, 99, 793–793. doi:10.1002/qj.49709942229

Bom, 2010. Operational implementation of the ACCESS numerinumerical weather prediction systems. *NMOC Operations Bulletin*, 83, Melbourne, Australia.

Bowler, N.E., Pierce, C.E., and Seed, A.W., 2006. STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quarterly Journal of the Royal Meteorological Society*, 132, 2127–2155. doi:10.1256/qj.04.100

Casati, B., Ross, G., and Stephenson, D.B., 2004. A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Applications*, 11, 141–154. doi:10.1017/S1350482704001239

Chumchean, S., Seed, A., and Sharma, A., 2008. An operational approach for classifying storms in real-time radar rainfall estimation. *Journal of Hydrology*, 363, 1–17. doi:10.1016/j.jhydrol.2008.09.005

Chumchean, S., Sharma, A., and Seed, A., 2006. An integrated approach to error correction for real-time radar-rainfall estimation. *Journal of Atmospheric and Oceanic Technology*, 23, 67–79. doi:10.1175/JTECH1832.1

Ciach, G.J., 2003. Local random errors in tipping-bucket rain gauge measurements. *Journal of Atmospheric and Oceanic Technology*, 20, 752–759. doi:10.1175/1520-0426(2003)20<752:LREITB>2.0.CO;2

Clark, M.P. and Hay, L.E., 2004. Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. *Journal of Hydrometeorology*, 5, 15–32. doi:10.1175/1525-7541(2004)005<0015:UOMNWP>2.0.CO;2

Colle, B.A. and Mass, C.F., 1996. An observational and modeling study of the interaction of low-level southwesterly flow with the Olympic mountains during COAST IOP 4. *Monthly Weather Review*, 124, 2152–2175. doi:10.1175/1520-0493(1996)124<2152:AOAMSO>2.0.CO;2

Collier, C.G., 1989. *Applications of weather radar systems: a guide to uses of radar data in meteorology and hydrology*. Chichester: Horwood; 1989.

Damrath, U., 2004. Verification against precipitation observations of a high density network – what did we learn? *International Verification Methods Workshop*, Montreal, 15–17 September 2004.

Damrath, U., *et al*., 2000. Operational quantitative precipitation forecasting at the German weather service. *Journal of Hydrology*, 239, 260–285. doi:10.1016/S0022-1694(00)00353-X

Davis, C., Brown, B., and Bullock, R., 2006. Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, 134, 1772–1784. doi:10.1175/MWR3145.1

Doelling, I.G., Joss, J., and Riedl, J., 1998. Systematic variations of Z–R-relationships from drop size distributions measured in northern Germany during seven years. *Atmospheric Research*, 47–48, 635–649. doi:10.1016/S0169-8095(98)00043-X

Doswell, C.A., Davies-Jones, R., and Keller, D.L., 1990. On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting*, 5, 576–585. doi:10.1175/1520-0434(1990)005<0576:OSMOSI>2.0.CO;2

Ebert, E.E., 2008. Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, 15, 51–64. doi:10.1002/(ISSN)1469-8080

Ebert, E.E., *et al*., 2003. The WGNE assessment of short-term quantitative precipitation forecasts. *Bulletin of the American Meteorological Society*, 84, 481–492. doi:10.1175/BAMS-84-4-481

Ebert, E.E., Janowiak, J.E., and Kidd, C., 2007. Comparison of near-real-time precipitation estimates from satellite observations and numerical models. *Bulletin of the American Meteorological Society*, 88, 47–64. doi:10.1175/BAMS-88-1-47

Ebert, E.E. and Mcbride, J.L., 2000. Verification of precipitation in weather systems: determination of systematic errors. *Journal of Hydrology*, 239, 179–202. doi:10.1016/S0022-1694(00)00343-7

Ebert, E.E., *et al*., 2004. Verification of nowcasts from the WWRP Sydney 2000 forecast demonstration project. *Weather and Forecasting*, 19, 73–96. doi:10.1175/1520-0434(2004)019<0073:VONFTW>2.0.CO;2

Fields, G., *et al*., 2004. Calibration of weather radar in South East Queensland. *In*: *Sixth international symposium on hydrological applications of weather radar*, Melbourne, Australia.

Gabella, M., *et al*., 2006. Range adjustment for ground-based radar, derived with the spaceborne TRMM precipitation radar. *IEEE Transactions on Geoscience and Remote Sensing*, 44, 126–133. doi:10.1109/TGRS.2005.858436

Green, D., *et al*., 2011. *Water resources and management overview: Murrumbidgee catchment*. Sydney: NSW Office of Water.

Hildebrand, P.H., 1978. Iterative correction for attenuation of 5 cm radar in rain. *Journal of Applied Meteorology*, 17, 508–514. doi:10.1175/1520-0450(1978)017<0508:ICFAOC>2.0.CO;2

Johnson, L.E. and Olsen, B.G., 1998. Assessment of quantitative precipitation forecasts. *Weather and Forecasting*, 13, 75–83. doi:10.1175/1520-0434(1998)013<0075:AOQPF>2.0.CO;2

Jordan, P., Seed, A., and Austin, G., 2000. Sampling errors in radar estimates of rainfall. *Journal of Geophysical Research: Atmospheres*, 105, 2247–2257. doi:10.1029/1999JD900130

Jordan, P.W., Seed, A.W., and Weinmann, P.E., 2003. A stochastic model of radar measurement errors in rainfall accumulations at catchment scale. *Journal of Hydrometeorology*, 4, 841–855. doi:10.1175/1525-7541(2003)004<0841:ASMORM>2.0.CO;2

Joss, J. and Lee, R., 1995. The application of Radar–Gauge comparisons to operational precipitation profile corrections. *Journal of Applied Meteorology*, 34, 2612–2630. doi:10.1175/1520-0450(1995)034<2612:TAORCT>2.0.CO;2

Kobold, M. and Sušelj, K., 2005. Precipitation forecasts and their uncertainty as input into hydrological models. *Hydrology and Earth System Sciences*, 9, 322–332. doi:10.5194/hess-9-322-2005

Krajewski, W.F., Villarini, G., and Smith, J.A., 2010. RADAR-rainfall uncertainties. *Bulletin of the American Meteorological Society*, 91, 87–94. doi:10.1175/2009BAMS2747.1

Lee, C.K., *et al.*, 2009. A preliminary analysis of spatial variability of raindrop size distributions during stratiform rain events. *Journal of Applied Meteorology and Climatology*, 48, 270–283. doi:10.1175/2008JAMC1877.1

Lopez, P., 2011. Direct 4D-var assimilation of NCEP stage IV Radar and Gauge precipitation data at ECMWF. *Monthly Weather Review*, 139, 2098–2116. doi:10.1175/2010MWR3565.1

Lopez, P. and Bauer, P., 2007. "'1D+4DVAR" assimilation of NCEP stage-IV radar and gauge hourly precipitation data at ECMWF. *Monthly Weather Review*, 135, 2506–2524. doi:10.1175/MWR3409.1

Mapiam, P.P. and Sriwongsitanon, N., 2008. Climatological Z-R relationship for radar rainfall estimation in the upper Ping river basin. *Science Asia*, 34, 215–222. doi:10.2306/scienceasia1513-1874.2008.34.215

Mapiam, P.P., *et al.*, 2009. Effects of rain gauge temporal resolution on the specification of a Z-R relationship. *Journal of Atmospheric and Oceanic Technology*, 26, 1302–1314. doi:10.1175/2009JTECHA1161.1

Marzban, C. and Sandgathe, S., 2006. Cluster analysis for verification of precipitation fields. *Weather and Forecasting*, 21, 824–838. doi:10.1175/WAF948.1

Mcbride, J. L. and Ebert, E. E., 2000. Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. *Weather and Forecasting*, 15, 103–121.

Morin, E. and Gabella, M., 2007. Radar-based quantitative precipitation estimation over Mediterranean and dry climate regimes. *Journal of Geophysical Research: Atmospheres*, 112, n/a-n/a. doi:10.1029/2006JB004485

Peleg, N., Ben-Asher, M., and Morin, E., 2013. Radar subpixel-scale rainfall variability and uncertainty: lessons learned from observations of a dense rain-gauge network. *Hydrology and Earth System Sciences*, 17, 2195–2208. doi:10.5194/hess-17-2195-2013

Piccolo, F. and Chirico, G.B., 2005. Sampling errors in rainfall measurements by weather radar. *Advances in Geosciences*, 2, 151–155. doi:10.5194/adgeo-2-151-2005

Puri, K., *et al.*, 2013, Implementation of the initial ACCESS numerical weather prediction system. *Australian Meteorological and Oceanographic Journal*, 63, 265–284.

Rennie, S.J., 2012, Doppler weather radar in Australia. *CAWCR Technical Report*, 55, 1–42.

Rezacova, D., Sokol, Z., and Pesice, P., 2007. A radar-based verification of precipitation forecast for local convective storms. *Atmospheric Research*, 83, 211–224. doi:10.1016/j.atmosres.2005.08.011

Richard, E., *et al.*, 2003. Intercomparison of mesoscale meteorological models for precipitation forecasting. *Hydrology and Earth System Sciences*, 7, 799–811. doi:10.5194/hess-7-799-2003

Rinehart, R.E., 1991. *Radar for meteorologists or, you too can be a radar meteorologist*. Grand Forks, ND: R.E. Rinehart.

Roberts, N., 2008. Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorological Applications*, 15, 163–169. doi:10.1002/(ISSN)1469-8080

Roberts, N.M., *et al.*, 2009. Use of high-resolution NWP rainfall and river flow forecasts for advance warning of the Carlisle flood, northwest England. *Meteorological Applications*, 16, 23–34. doi:10.1002/met.v16:1

Roberts, N.M. and Lean, H.W., 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136, 78–97. doi:10.1175/2007MWR2123.1

Shahrban, M., *et al.*, 2011. Comparison of weather radar, numerical weather radar, numerical weather prediction and gauge-based rainfall estimates. *In*: *MODSIM 19th International Congress on Modelling and Simulation*, Modelling and Simulation Society of Australia and New Zealand, Perth, Australia.

Shrestha, D.L., *et al.*, 2013. Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose. *Hydrology and Earth System Sciences*, 17, 1913–1931. doi:10.5194/hess-17-1913-2013

Smith, A.B., *et al.*, 2012. The Murrumbidgee soil moisture monitoring network data set. *Water Resources Research*, 48. doi:10.1029/2012WR011976

Steiner, M. and Smith, J.A., 2000. Reflectivity, rain rate, and kinetic energy flux relationships based on raindrop spectra. *Journal of Applied Meteorology*, 39, 1923–1940. doi:10.1175/1520-0450(2000)039<1923:RRRAKE>2.0.CO;2

Tustison, B., Harris, D., and Foufoula-Georgiou, E., 2001. Scale issues in verification of precipitation forecasts. *Journal of Geophysical Research: Atmospheres*, 106, 11775–11784. doi:10.1029/2001JD900066

Vasić, S., *et al.*, 2007. Evaluation of precipitation from numerical weather prediction models and satellites using values retrieved from radars. *Monthly Weather Review*, 135, 3750–3766. doi:10.1175/2007MWR1955.1

Velasco-Forero, C.A., *et al.*, 2009. A non-parametric automatic blending methodology to estimate rainfall fields from rain gauge and radar data. *Advances in Water Resources*, 32, 986–1002. doi:10.1016/j.advwatres.2008.10.004

Weygandt, S.S., *et al.*, 2004. Scale sensitivities in model precipitation skill scores during IHOP. *In*: *22nd Conference Severe Local Storms, American Meteorological Society*, Hyannis, MA, 4–8 October.

Wilks, D.S., 1995. *Statistical methods in the atmospheric sciences. An introduction*. San Diego: Academic Press, 467pp.

Wilson, J.W., *et al.*, 2010. Nowcasting challenges during the Beijing Olympics: Successes, failures, and implications for future nowcasting systems. *Weather and Forecasting*, 25, 1691–1714. doi:10.1175/2010WAF2222417.1

Wood, S.J., Jones, D.A., and Moore, R.J., 2000. Accuracy of rainfall measurement for scales of hydrological interest. *Hydrology and Earth System Sciences*, 4, 531–543. doi:10.5194/hess-4-531-2000

Yates, E., *et al.*, 2006. Point and areal validation of forecast precipitation fields. *Meteorological Applications*, 13, 1–20. doi:10.1017/S1350482705001921

Yu, W., *et al.* 1998. *High resolution model simulation of precipitation and evaluation with Doppler radar observation*. *In*: Proceedings of the 1997 3rd International Workshop on Rainfall in Urban Areas, December 4, 1997–December 7, 1997, Pontresina, Switz. Elsevier, 179–186.

Appendix A: Abbreviations

4D-Var	Four-dimensional variational data assimilation
ACCESS	Australian Community Climate Earth-System Simulator
ACCESS-A	Australian ACCESS model
ACCESS-G	Global ACCESS model
ACCESS-R	Regional ACCESS model
ACCESS-T	Tropical ACCESS model
ACCESS-VT	Victoria–Tasmania ACCESS model
ACESS-TC	Tropical cyclone ACCESS model

BoM	Bureau of Meteorology
CSI	Critical success index
CSIRO	Commonwealth Scientific and Industrial Research Organisation
DM	Deutschland Modell
ECMWF	European Centre for Medium-Range Weather Forecasts
EM	Europa-Modell
FAR	False alarm ratio
FBI	Frequency bias index
FSS	Fractions skill score
GASP	Global Assimilation and Prediction
GEM	Global Environmental Multiscale
GM	Global model
GOES	Geostationary Operational Environmental Satellite
LAPS	Limited area and prediction system
LM COSMO	Lokal-Modell of the Consortium for Small-Scale Modelling
LM	Lokal-Modell
ME	Mean error
NCEP	National Centres for Environmental Prediction
NWP	Numerical weather prediction
POD	Probability of detection
QPE	Quantitative precipitation estimates
QPF	Quantitative precipitation forecast
RE	Relative error
RMSE	Root mean square error
TSS	True skill statistics
UM	Unified model
WGNE	Working Group of Numerical Experimentation
WRF	Weather Research and Forecasting

Appendix B: Verification metrics

The RMSE and ME measure the average error magnitude, while the RE or bias is the total difference between the observed and forecast values divided by the total observed values:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i)^2} \tag{A1}$$

$$\mathrm{ME} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i) \tag{A2}$$

$$\mathrm{RE} = \frac{\sum_{i=1}^{N}(Y_i - X_i)}{\sum_{i=1}^{N}(X_i)} \tag{A3}$$

where Y_i is the forecast value, X_i is the corresponding observed value, and N is the number of forecast–observation pairs.

The area bias is defined by:

$$\mathrm{Bias}_{\mathrm{area}}^{(h)} = \frac{\sum_{i=1}^{N}(A_{\mathrm{f},i}^{(h)} - A_{\mathrm{o},i}^{(h)})}{\sum_{i=1}^{N}A_{\mathrm{o},i}^{(h)}} \tag{A4}$$

where f and o refer to forecast and observation respectively, h refers to an hour of the day, and the summation is over all matching pairs at a given hour.

The FBI is the ratio of the forecast frequency to the observed frequency, while the POD calculates the fraction of observed events that were correctly forecast, and the FAR indicates the fraction of predicted events that were observed to be non-events:

$$\mathrm{FBI} = \frac{H+F}{H+M} \tag{A5}$$

$$\mathrm{POD} = \frac{H}{H+M} \tag{A6}$$

$$\mathrm{FAR} = \frac{H}{H+F} \tag{A7}$$

where H (hits) is the number of actual rain events predicted by the radar/model, M (misses) is the number of actual rainfall events missed by them, F (false alarm) is the non-observed rain predicted by the radar/model, and R is the number of correct non-forecast cases.

The CSI is the fraction of all forecast and/or observed events that were correctly forecast:

$$\mathrm{CSI} = \frac{H}{H+M+F} \tag{A8}$$

The TSS measures the ability of the model to distinguish between occurrences and non-occurrences of an event:

$$\mathrm{TSS} = \frac{(H\times R)-(M\times F)}{(H+M)(F+R)} \tag{A9}$$

The FSS is a variation on the fractions Brier score (FBS):

$$\mathrm{FSS} = \frac{\mathrm{FBS}}{\mathrm{FBS}_{\mathrm{worst}}} \tag{A10}$$

$$\mathrm{FBS} = \frac{1}{N}\sum_{i=1}^{N}(M_i - O_i)^2 \tag{A11}$$

where M and O are the model and observation fractions respectively with values between 0 and 1. N is the number of pixels in the verification area. $\mathrm{FBS}_{\mathrm{worst}}$ is given by:

$$\mathrm{FBS}_{\mathrm{worst}} = \frac{1}{N}\left[\sum_{i=1}^{N}M_i^2 + \sum_{i=1}^{N}O_i^2\right]. \tag{A12}$$