Hydrol. Earth Syst. Sci., 17, 1913–1931, 2013
www.hydrol-earth-syst-sci.net/17/1913/2013/
doi:10.5194/hess-17-1913-2013

Hydrology and Earth System Sciences
Open Access

Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose

D. L. Shrestha[1], D. E. Robertson[1], Q. J. Wang[1], T. C. Pagano[2], and H. A. P. Hapuarachchi[2]

[1]CSIRO Land and Water, Highett, Australia
[2]Bureau of Meteorology, Melbourne, Australia

Correspondence to: D. L. Shrestha (durgalal.shrestha@csiro.au)

Received: 25 October 2012 – Published in Hydrol. Earth Syst. Sci. Discuss.: 5 November 2012
Revised: 18 March 2013 – Accepted: 20 April 2013 – Published: 21 May 2013

Abstract. The quality of precipitation forecasts from four Numerical Weather Prediction (NWP) models is evaluated over the Ovens catchment in Southeast Australia. Precipitation forecasts are compared with observed precipitation at point and catchment scales and at different temporal resolutions. The four models evaluated are the Australian Community Climate Earth-System Simulator (ACCESS) including ACCESS-G with a 80 km resolution, ACCESS-R 37.5 km, ACCESS-A 12 km, and ACCESS-VT 5 km.

The skill of the NWP precipitation forecasts varies considerably between rain gauging stations. In general, high spatial resolution (ACCESS-A and ACCESS-VT) and regional (ACCESS-R) NWP models overestimate precipitation in dry, low elevation areas and underestimate in wet, high elevation areas. The global model (ACCESS-G) consistently underestimates the precipitation at all stations and the bias increases with station elevation. The skill varies with forecast lead time and, in general, it decreases with the increasing lead time. When evaluated at finer spatial and temporal resolution (e.g. 5 km, hourly), the precipitation forecasts appear to have very little skill. There is moderate skill at short lead times when the forecasts are averaged up to daily and/or catchment scale. The precipitation forecasts fail to produce a diurnal cycle shown in observed precipitation. Significant sampling uncertainty in the skill scores suggests that more data are required to get a reliable evaluation of the forecasts. The non-smooth decay of skill with forecast lead time can be attributed to diurnal cycle in the observation and sampling uncertainty.

Future work is planned to assess the benefits of using the NWP rainfall forecasts for short-term streamflow forecasting. Our findings here suggest that it is necessary to remove the systematic biases in rainfall forecasts, particularly those from low resolution models, before the rainfall forecasts can be used for streamflow forecasting.

1 Introduction

Forecasts of streamflow with lead times up to 10 days are important for water resources management and mitigating impacts of floods. Streamflow forecasts are produced by initialising the state variables of a hydrological model to their condition at the forecast time and subsequently forcing the model with future weather conditions for the forecast period. A major source of uncertainty in this process is future precipitation. Numerical Weather Prediction (NWP) models have been used since 1946 to forecast precipitation and other atmospheric variables. However, forecasting precipitation is challenging because it is discontinuous and varies rapidly in space and time. The precipitation process depends not only on the synoptic situation, but also on processes that are not explicitly considered by NWP models, including condensation, vertical convective transport of heat and moisture and phase transitions of water between vapour, clouds and ice (Damrath et al., 2000). Increased computing power and improvement of the NWP models have lead to considerable advancement in the ability to predict precipitation. However, ability of the NWP models to forecast precipitation is still relatively low, especially for very short lead times (e.g. < 12 h), for long lead times (e.g. > 5 days) and for fine scale weather systems such as local-regional convective systems (e.g. thunderstorms) (Cuo et al., 2011).

Some of the earliest experiments linking precipitation forecasts to hydrological applications began three decades ago (see Georgakakos and Hudlow, 1984). Accurate precipitation forecasts can reduce forcing uncertainty in hydrological (e.g. rainfall-runoff) models and can greatly improve the quality of streamflow forecasts. However, NWP precipitation forecasts are inherently uncertain and subject to three types of error (Habets et al., 2004): localisation, timing and intensity of precipitation events, which potentially limit their usefulness for streamflow forecasting. Hydrological models are sensitive to uncertainty in the precipitation forecasts, which are propagated to the model outputs. Thus, it is required to address the quantification of the uncertainty in meteorological observation and forecasting along with their effect on hydrological forecasting (Rossa et al., 2011).

The contribution of precipitation forecasts to the skill of streamflow forecasts is dependent on many factors, including lead time. At lead times that are less than the time of concentration of a catchment, precipitation forecasts will contribute little skill to streamflow forecasts. During this period, catchment and channel storage and the passage of an existing flood wave downstream are the main influences on the streamflow forecasts. NWP model precipitation forecasts are also typically unable to resolve the observed precipitation distribution at very short lead times, and persistence or extrapolation-based methods can provide better forecasts. Hence, NWP model precipitation forecasts are more useful for streamflow forecasting in extending forecast lead time, particularly in the range of a few days to one or two weeks (Cloke and Pappenberger, 2009; Cuo et al., 2011). However, the extent to which precipitation forecasts are beneficial for streamflow forecasts depends considerably on the ability of the NWP models to resolve the scale and processes relevant for hydrological applications and whether the surface hydrology in the catchment is dominated by precipitation (Clark and Hay, 2004; Gebhardt et al., 2008).

Understanding the quality of NWP precipitation forecasts is important step in assessing their potential contribution to the skill of streamflow forecasts. Objective evaluation or verification of precipitation forecasts did not begin until the mid 1990s (e.g. WMO Working Group on Numerical Experimentation, WWRP/WGNE, 2008). The overall purpose of evaluation is to ensure that forecasts are accurate, skilful and reliable from a technical point of view. Evaluation of precipitation forecasts is important to monitor forecast quality over space and time, to compare the quality of different forecast systems and to discover sources of model error to improve the forecast quality (WMO, 2000; WWRP/WGNE, 2008; Casati et al., 2008). However, from a streamflow forecasting perspective, forecast evaluation is also to understand the nature of forecast errors (e.g. bias, error on light precipitation versus heavy precipitation) which can inform the development of methods for post-processing raw forecasts to improve their accuracy and reliability.

Evaluation of NWP model forecasts of precipitation is not a new topic. Numerous forecasters, researchers have verified precipitation forecasts from a meteorological perspective (e.g. Jolliffe and Stephenson, 2012). A much smaller number of them have evaluated precipitation forecasts from a hydrological perspective (see e.g. Pappenberger et al., 2008). Georgakakos and Hudlow (1984) highlighted the relevance of precipitation forecasts products to real-time hydrological forecasting. Golding (2000) identified the critical areas where NWP products fall short, and illustrated techniques being developed to address them. Damrath et al. (2000) verified 7 yr of precipitation forecasts from NWP models of the German Weather Services. Kaufmann et al. (2003) evaluated the quality of 8 yr of precipitation forecasts from the Swiss Model in Switzerland. Hay and Clark (2003) used 40 yr of 8 day ahead precipitation forecasts over the contiguous United States from the National Centres for Environmental Prediction reanalysis project to assess the possibilities for using the medium-range forecast model output. Richard et al. (2003) compared four European and Canadian mesoscale models for precipitation forecasting to reproduce heavy precipitation events. Habets et al. (2004) used precipitation forecasts from two French NWP models as inputs to a hydrologic model. Roy Bhowmik et al. (2007) evaluated precipitation predictive skill of the Indian Meteorological Department operational NWP system over the Indian monsoon region. Roberts (2008) assessed the spatial and temporal variation in the skill of precipitation forecasts from a NWP model. Rotach et al. (2009) tested real-time, end-to-end multi-model hydrometeorological forecasts from heavy precipitation and related flooding in many different catchments in the Alps. Roberts et al. (2009) demonstrated the benefit of using high resolution NWP model precipitation forecasts for flood and short-term streamflow forecasting. Ghile and Schulze (2010) verified the skill and accuracy of the precipitation forecasts by three NWP models over the Mgeni catchment in South Africa. Ament et al. (2011) evaluated the performances of 13 mesoscale NWP models with respect to heavy precipitation alerts by these models in Switzerland during the summer 2007. Ghelli and Ebert (2008) and Jolliffe and Stephenson (2012) presented a comprehensive review and the state of art in forecast verification.

Few studies have verified NWP precipitation forecasts for Australia. McBride and Ebert (2000) verified precipitation forecasts from 7 international (including Australian) NWP models over Australia. They verified 24 h total (daily) precipitation forecasts for the first 24 h of the forecast period over a one-year period using only categorical verification scores. The verification statistics were presented over a standardised 1° latitude-longitude grid over the continent of Australia. Ebert et al. (2003) reported the WGNE assessment of short-term precipitation forecasts from several international NWP global and regional models in different areas of the globe including Australia. Forecasts of 24 h precipitation totals were

verified at lead times of 24 and 48 h over Australia using only two categorical evaluation scores.

This study focuses on comprehensive analysis of the NWP precipitation forecasts in Australia from a hydrological perspective. Unlike many synoptic-scale precipitation verification studies undertaken from a meteorological perspective, this study evaluates the precipitation forecasts on scales relevant to hydrology. The evaluation of the precipitation forecasts and other meteorological variables from a hydrological point of view is challenging because the resolution of the NWP model is often too coarse to resolve the small catchment scale. Irregular catchment boundaries do not necessarily coincide with NWP model grids. This may require an interpolation of the NWP model precipitation forecasts. The verification of precipitation forecasts from a hydrological perspective requires at short temporal resolution (e.g. subdaily). Furthermore, hydrological catchment has a memory of several hours, days, weeks or months based on its size. Response of the catchment depends on the previous events and on the timing of the present events, thus, it requires to evaluate the forecasts on several forecast times. However, most of the meteorological verification is based on evaluating forecasts on several model grid cells at a particular forecast time and does spatial aggregating of forecasts rather than temporal aggregating. For example, in meteorological verification, hit rates are often calculated by counting number of grid cells of correct forecasts at a particular time rather than counting number of events of correct forecasts on several days at a particular location. Spatial aggregating also ignores the location error whereas it is crucial for hydrological application as an error of a few kilometres can lead the precipitation in the wrong catchment (Habets et al., 2004) which does not contribute to the streamflow forecasts to the catchment of interest.

This study is the first part of a research programme to support the production of ensemble streamflow forecasts by the Australian Bureau of Meteorology. The forecasting service seeks to produce ensemble streamflow forecasts out to 10 days using continuous hydrological modelling and NWP rainfall forecasts. The main objectives of this study are to (i) compare the skill of NWP models with different spatial resolutions at station locations and at the catchment scale, (ii) evaluate the effect of lead time, precipitation accumulation period, and precipitation threshold values on forecast skill, and (iii) investigate the effect of diurnal cycle and sampling uncertainty on forecast skill. The contribution and benefit of NWP model rainfall forecasts for use in streamflow forecasting will be presented in a subsequent paper. In comparison with previous studies, the main contributions of this study are to (i) evaluate the quality of the ACCESS model suite which is the latest generation Australian NWP model, (ii) use both continuous and categorical evaluation scores, (iii) analyse the evaluation scores of precipitation forecasts at multiple sub-daily temporal resolutions out to longer forecast lead times, (iv) investigate diurnal cycle and uncertainty analysis of the evaluation scores. The Ovens catchment in Southeast Australia is selected to evaluate the skill of the precipitation forecasts from ACCESS models.

2 Numerical weather prediction models and data

2.1 Description of ACCESS models

The Australian Community Climate Earth-System Simulator (ACCESS) model suite (BoM, 2010) has been the operational NWP system employed by the Australian Bureau of Meteorology (BoM) since August 2010. The ACCESS NWP model system is based on the UK Met Office's Unified Model/Variational Assimilation (UM/VAR) system with multiple resolutions and spatial domains extending from a course resolution global model down to the high resolution city-based models. This study uses the initial rollout of the ACCESS system APS0 (Australian Parallel Suite version 0). The APS0 version of ACCESS uses version 6.4 of the Unified Model from the UK Met Office. Key features of various components and physical parameterisations given below are taken from BoM (2010).

ACCESS is a non-hydrostatic model with prognostic variables of winds, air density, temperature, mixing ratios of water-vapour, cloud-liquid-water and cloud-frozen-water. The model uses an Arakawa C-grid in the horizontal and a Charney-Phillips grid in the vertical. The model is configured such that each grid point in the horizontal is spaced a constant latitude and longitude increment apart from adjacent grid points. The vertical levels are constructed in a hybrid fashion so they conform to terrain heights near the surface and become constant height surfaces in the upper atmosphere. Two-time-level semi Lagrangian with non-interpolating scheme is used for vertical advection of temperature. Acoustic terms are treated using a semi-implicit approach yielding a Helmholtz equation for the Exner pressure tendency, which is solved using a preconditioned generalised conjugate residual method.

Water clouds are derived from sub-grid scale probability distribution of conserved variables of liquid/frozen water temperature and total water content using an assumed critical relative humidity (Smith, 1990). Ice water content is determined by the prognostic mixed phase microphysics scheme with ice cloud fraction calculated diagnostically from ice water content. Precipitation is computed by single-moment bulk microphysics scheme with explicit calculation of transfers between vapour, liquid and ice phases. The microphysical processes calculated in the scheme are sedimentation of the ice and rain, heterogeneous and homogeneous nucleation of ice particles, deposition and sublimation of ice, riming and melting of ice, collection of cloud droplets by raindrops, etc. The model computes atmospheric radiation using rigorous solution of the two-stream scattering equations including partial cloud cover.

Table 1. Precipitation forecasts available from NWP models for the study.

Domain	NWP system	Resolution (km)	Lead time (days)	Earliest date	Latest date	Forecast dates
Global	ACCESS-G	80	10	27 Aug 2009	18 Apr 2011	597
Regional	ACCESS-R	37.5	3	27 Aug 2009	18 Apr 2011	597
Australia	ACCESS-A	12	2	1 Feb 2010	18 Apr 2011	437
VICTAS	ACCESS-VT	5	1.5	31 Mar 2010	18 Apr 2011	381

Mixing in unstable layers uses a first order non-local scheme that parameterises eddy diffusivity profiles of unstable layers driven either by fluxes at the surface or by cloud-top processes. Cumulus mixing uses the mass-flux convection scheme. Cumulus convection is diagnosed if air at the first model level is unstable to adiabatic ascent above the lifting condensation level. The cloud base mass-flux is calculated based on the reduction of zero convectively available potential energy over a given timescale. The representation of convective momentum transport for deep and shallow convection is based on an eddy viscosity model.

The ACCESS APS0 system comprises a global model (ACCESS-G) with a 80 km resolution and forecast duration of 10 days; regional models (ACCESS-R, and ACCESS-T) with a 37.5 km resolution and forecast duration of 3 days; an Australian model (ACCESS-A) with a 12 km resolution and forecast duration of 2 days, city models (ACCESS-VT, ACCESS-S, ACCESS-P, ACCESS-BR) with a 5 km resolution and forecast duration of 36 h, and a tropical cyclone (ACCESS-TC) with 12 km resolution, a relocatable spatial domain and forecast duration of 3 days. Currently new versions of the ACCESS models (APS1) with improved resolution and model physics are being introduced at the BoM. Figure 1 shows the domains of ACCESS APS0 (ACCESS-G, ACCESS-R, ACCESS-A, and ACCESS-VT) models which are used in this study.

All models except ACCESS-G use boundary conditions that are provided by coarser resolution models, for example, ACCESS-R is nested inside the previous run of ACCESS-G, while ACCESS-A and ACCESS-VT are nested inside the concurrent run of ACCESS-R. ACCESS system uses a four-dimensional variational data assimilation scheme which allows observations made at a range of times and locations to be used to initialise the model in a dynamically consistent way. Data assimilation occurs 4 times daily for nominal assimilation base times of 00:00, 06:00, 12:00 and 18:00 UTC. However, for ACCESS-G and ACCESS-VT, full model forecasts are only run at 00:00 and 12:00 UTC. In contrast, for ACCESS-R and ACCESS-A full model forecasts are run 4 times daily at 00:00, 06:00, 12:00, and 18:00 UTC. For ACCESS-R and ACCESS-A, a second update data assimilation step is run 4 h later than the main run to make use of any additional observational data that were not available at the time of the earlier main assimilation step (BoM, 2010).

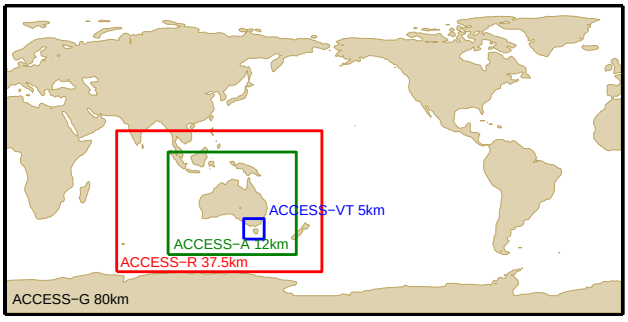

Fig. 1. Domains of initial ACCESS (APS0) NWP models used in this study.

This study uses the archive of precipitation forecasts generated in real time by the ACCESS models. This archive began on late 2009 and has been maintained through to the present. Table 1 shows the archive of precipitation forecasts (issued at 12:00 UTC) available for the study. The BoM expects to run the hydrologic models around 09:00 LT (Fig. 2). The most recent ACCESS model forecasts available at 09:00 LT are those initialised at 12:00 UTC (22:00 LT in Victoria). Therefore, the results presented in this study disregard the first 11 h of the NWP forecasts. NWP forecasts for the first few hours are generally regarded as not reliable because of the so-called "spin-up" time (Kasahara et al., 1992). Thus, our results are considered to be free from model spin-up effects.

Precipitation forecasts from all models are available at hourly intervals for this study, with the exception of ACCESS-G which are available at 3 hourly intervals. In order to compare the skill among the models, only one year period of data from 31 March 2010 to 30 March 2011 (see Table 1) is selected for the analysis.

2.2 Study area

In this study, the Ovens catchment in Southeast Australia is selected to evaluate the skill of the precipitation forecasts from ACCESS models (Fig. 3). The Ovens catchment is the focus of a prototype flood and short-term streamflow forecasting service with lead times up to 10 days run by the BoM. The Ovens catchment provides a significant source of unregulated inflow to the Murray Darling Basin and has

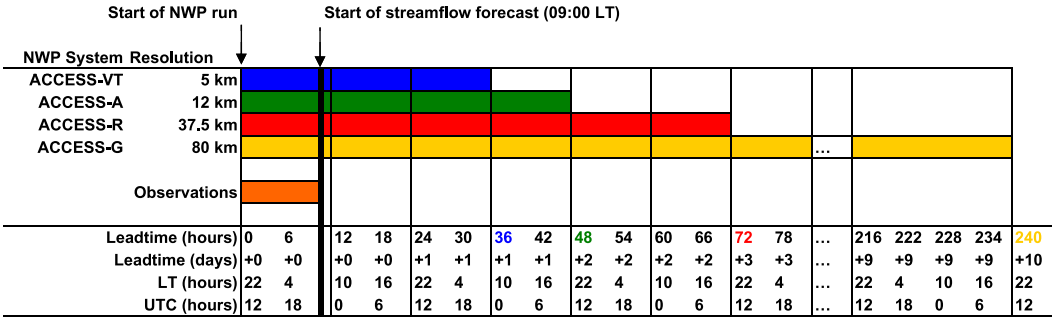

Fig. 2. Schematic of NWP model runs.

several urban centres that have experienced significant economic damage from flooding.

The Ovens river rises in the Victorian Alps and the catchment is bounded by several significant peaks, including Mount Hotham (elevation 1861 m, longitude 147.33°, latitude −37.05°), Mount Feathertop (elevation 1922 m, longitude 147.13°, latitude −36.9°) and Mount Buller (elevation 1805 m, longitude 146.41°, latitude −37.14°). The Wangaratta streamflow gauge (elevation 140 m, longitude 146.30°, latitude −36.42°) drains an area of 5552 km^2. The upper catchment is steep and hilly, and covered by native forest and tree plantations. The lower catchment is relatively flat with a wide floodplain and is mainly used for grazing and cropping. Snow is limited to the higher elevations, but is sufficient to support a seasonal skiing industry when supplemented with artificial snow. The catchment has two small sized (total capacity 37.5 million m^3) reservoirs which support irrigation and hydropower. Average annual potential evapotranspiration is 1164 mm (Raupach et al., 2008), approximately equal to the catchment average annual precipitation. There is a seasonal variability of precipitation and a strong gradient in the average annual precipitation, typically 550 mm in the lowlands near the catchment outlet and 1950 mm in the highlands near catchment headwaters.

Figure 3 also shows the spatial resolution of the ACCESS models with respect to the resolution of the hydrological model (93 sub-catchment areas) currently used in operational streamflow forecasting. The figure shows that the hydrological model resolution is roughly comparable to the 12 km ACCESS-A model grid. Furthermore, the coarser NWP models (viz. ACCESS-R and ACCESS-G) are unlikely to capture gradients of precipitation across the catchment. The 80 km resolution ACCESS-G model has only 4 grid cells across the catchment and more than three-quarter of the catchment is covered by a single grid cell.

Observed precipitation data were collected from 33 measurement stations that are used for operational forecasting in the Ovens catchment (Table 2). The measurement stations are reasonably distributed across the catchment and surroundings as shown in Fig. 3. Some stations at high elevation have heated rain gauge to measure snow fall. Careful

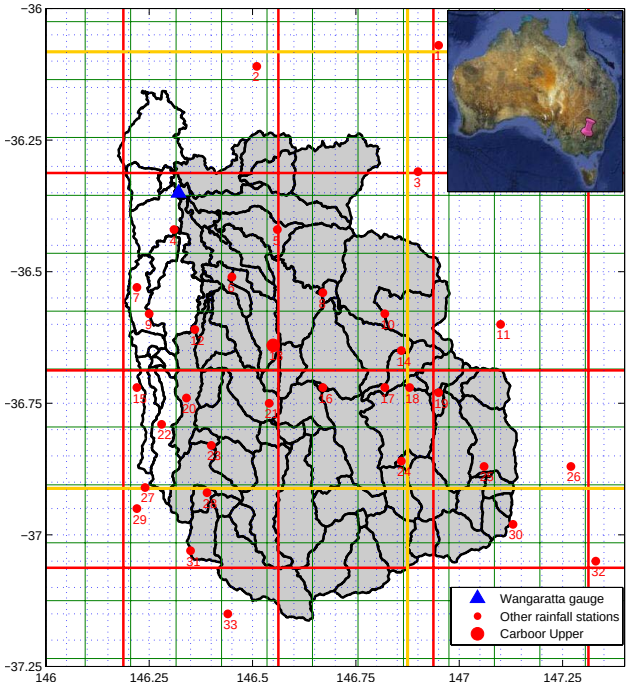

Fig. 3. Map of the Ovens catchment with sub-catchment delineation, precipitation and streamflow gauges. The shown are the overlaying with the ACCESS model grids – dashed, ACCESS-VT (5 km); thin line, ACCESS-A (12 km); thick line, ACCESS-R (37.5 km); and the thickest line; ACCESS-G (80 km). The shaded region is the catchment area draining to Wangaratta gauging station. In inset, location of the Ovens catchment is shown.

preparation of the precipitation observations was necessary and included removal of outliers and infilling of missing values. Data which are significantly different from the neighbouring stations and gridded daily precipitation data or previous time step are marked as missing. A visual inspection of the precipitation hyetograph and corresponding observed streamflow hydrograph was also used to identify outliers in the precipitation. We believe such data are resulted from human error, and there are a few such data records in the evaluation period. The infilling process related daily precipitation

Table 2. Precipitation stations in and near the Ovens catchment. The period of record for the annual average precipitation is September 1991 to February 2011.

Station number	Station name	Longitude (°)	Latitude (°)	Altitude (m)	Average precipitation (mm yr^{-1})	Missing data (%)
1	Albury AWS	146.95	−36.07	164	660	11.9
2	Rutherglen AWS	146.51	−36.11	175	556	12.5
3	Osbornes Flat	146.9	−36.31	225	855	0.6
4	Wangaratta AWS	146.31	−36.42	153	593	0.6
5	Bloomfield	146.56	−36.42	251	688	1.9
6	Bobinawarrah	146.45	−36.51	202	737	0.6
7	Greta West	146.22	−36.53	174	724	0.6
8	Rocky Point	146.67	−36.54	195	812	0.6
9	Greta South	146.25	−36.58	192	814	0.6
10	Rosewhite	146.82	−36.58	253	949	0.6
11	Mongans Bridge	147.1	−36.6	263	1069	0.6
12	Angleside	146.36	−36.61	189	799	0.6
13	Carboor Upper	146.55	−36.64	298	947	0.6
14	Eurobin	146.86	−36.65	266	1121	0.6
15	Loombah Reservoir	146.22	−36.72	356	967	0.6
16	Lake Buffalo	146.67	−36.72	238	1176	0.6
17	Mount Buffalo	146.82	−36.72	1350	1930	0.6
18	Harris Lane	146.88	−36.72	277	1240	0.6
19	Bright	146.95	−36.73	319	1147	0.6
20	Myrhee	146.34	−36.74	352	1134	0.6
21	Black Range Trout Farm	146.54	−36.75	425	1139	0.6
22	Handcocks	146.28	−36.79	425	1318	0.6
23	Cheshunt	146.4	−36.83	293	1186	0.6
24	Upper Buckland	146.86	−36.86	500	1306	0.6
25	Harrietville	147.06	−36.87	396	1349	13.3
26	Falls Creek AWS	147.27	−36.87	1510	1409	0.3
27	Archerton	146.24	−36.91	907	1243	0.6
28	L Will Hov	146.39	−36.92	440	1235	0.6
29	Mt Tabletop	146.22	−36.95	915	1045	0.6
30	Mt Hotham AWS	147.13	−36.98	1750	1647	11.6
31	Bald Hill	146.35	−37.03	1202	848	0.6
32	Mt Hotham Airport AWS	147.33	−37.05	1750	899	11.9
33	Mt Buller AWS	146.44	−37.15	1707	1494	0.4

totals at the measurement stations to gridded daily precipitation data from the Australian Water Availability Project (Jones et al., 2009) and disaggregated the daily total using the concurrent temporal pattern from the nearest available station. The percentage of missing data is given in Table 2.

Catchment average precipitation was estimated as the area-weighted average of sub-catchment precipitation. Sub-catchment precipitation data were derived by inverse distance weighting of precipitation from the nearby stations. The station precipitation time series were serially complete before inverse distance weighting to sub-catchment centroids. The sub-catchment precipitation was used to drive hydrological model, whereas catchment average precipitation was used to evaluate precipitation forecasts from NWP models at the catchment scale. The spatial resolution of the global model is too coarse to carry out the evaluation at sub-catchment scale.

3 Evaluation methods

The skill of NWP precipitation forecasts is known to vary in space and time. Therefore, an evaluation of the NWP precipitation forecasts should be aimed to reflect this characteristic. WWRP/WGNE (2008) recommended that evaluation be done both against gridded (model-oriented evaluation) observations and station observations (user-oriented evaluation). Model-oriented evaluation includes processing of observation data to match the spatial and temporal scales of the model. User-oriented evaluation uses station observations to evaluate model output from the overlying model grid cell.

In this study the evaluation of the quality of NWP precipitation forecasts is done both at stations and gridded observations. Station-based evaluation is done by directly comparing the station and NWP precipitation amounts at the model

grid cell in which the station exists. While this method is simplistic, any alternative would involve a spatial interpolation of precipitation data from irregularly spaced measurement stations which may introduce further bias (Richard et al., 2003). Although this verification approach has deficiencies (Roberts, 2008), direct comparison facilitates the understanding of skill from a user's perspective (i.e. without any interpolation or reanalysis). Furthermore, hydrological models are commonly calibrated with station observations and, therefore, an evaluation of quality and skill of NWP model has to be performed using observations (Pappenberger et al., 2008). The evaluation scores (described below) are computed for all 33 measurement stations over the Ovens catchment individually by aggregating forecasts over a period of one year (time averaging).

Evaluation using gridded observations is done at catchment scale where the grid is defined by an irregular catchment boundary rather than the NWP model grid. Evaluation is done by comparing interpolated catchment average precipitation and corresponding NWP precipitation forecast. Catchment average precipitation forecast F_c is computed by weighting each precipitation forecast F_i at grid cell i by the fraction of the catchment area within the grid cell i and given by

$$F_c = \frac{\sum_{i=1}^{N_g} A_i F_i}{\sum_{i=1}^{N_g} A_i} \quad (1)$$

where A_i is the area of catchment within the grid cell i, N_g is the number of the grid cells covered partly or fully by the catchment.

As no single evaluation score is adequate to judge the quality of NWP model precipitation forecasts, a large variety of scores are used operationally to verify them (see e.g. Stanski et al., 1989; Wilks, 2006; Wilson, 2001; WWRP/WGNE, 2008). A detailed assessment of the strengths and weaknesses of a set of forecasts usually requires more than one or two summary scores (Jolliffe and Stephenson, 2012). In this study, forecasts of precipitation amount are evaluated using three commonly used continuous verification scores: root-mean-square error (RMSE), bias and correlation coefficient. These scores assess different aspects of forecast quality. RMSE is one of the most basic and widely used methods of verification, and assesses the average magnitude of forecast errors (Stanski et al., 1989). Bias assesses the difference between the mean of forecasts and mean of the corresponding observations. The correlation coefficient reflects linear association between the forecasts and observations. The Pearson product moment correlation coefficient is not sensitive to biases that may be present in the forecasts, it is, however, sensitive to outliers (Wilks, 2006). Thus, Spearman rank correlation coefficient is more appropriate than Person correlation when data are not normally distributed. Note that above three evaluation scores are related according to the following equation (Murphy, 1988)

$$\mathrm{RMSE}^2 = \mathrm{Bias}^2 + S_f^2 + S_o^2 - 2S_f S_o \mathrm{Corr} \quad (2)$$

where S_f^2 and S_o^2 are the sample variances of the forecasts and observations, respectively, Corr is the Pearson correlation between the forecasts and observations.

From user point of view, it is also important to know whether precipitation occurs or not. Continuous precipitation values can be viewed categorically (or binary for "yes" or "no" events) according to whether or not the precipitation exceeds a given threshold value. The "event" here means just an instance of precipitation (not) exceeding a given threshold value at a particular time. Categorical verification scores are then used to evaluate the occurrence of precipitation. These scores are less sensitive to large errors than continuous verifications scores (especially those involving squared errors) which is particularly relevant for highly skewed data such as precipitation amounts. Thus, categorical verification scores may give more meaningful information for precipitation verification (WWRP/WGNE, 2008).

A number of the categorical verification scores are computed by building contingency table (Table 3) which shows the joint distribution of observed and forecast events and non-events. In the Table 3, "Hits" represents the number of events for which both forecasts and observations exceed a given threshold, "Misses" represents the number of events for which only observations exceed the threshold, "False Alarms" represents the number of events for which only forecasts exceed the threshold and "Correct Negatives" represents the number of events for which neither forecasts nor observations exceed the threshold.

In this study, probability of detection (POD), false alarm ratio (FAR), frequency bias (FBI) and critical success index (CSI) have been calculated from the contingency table. Table 4 shows the formulae of categorical verification scores with their perfect and possible ranges values. POD measures the fraction of observed events that were correctly forecast and is insensitive to false alarms. FAR gives the fraction of forecast events that were observed to be non-events and ignores the misses. FBI gives the ratio of frequency of forecast rain to the observed rain and does not take into account accuracy. CSI gives the fraction of all forecast and observed events that were correctly diagnosed and does consider both misses and false alarms.

The value of any evaluation score is limited if uncertainty associated with the score is not quantified (Jolliffe, 2007). Any evaluation score must be regarded as a sample estimate of the "true" value for an infinitely large verification dataset. There is, therefore, some uncertainty associated with the score's value, especially when the sample size is small or the data are not independent, or both (WWRP/WGNE, 2008). In this study, the uncertainty associated with the evaluation scores is estimated using re-sampling technique. Although it is also possible to compute uncertainty of some

Table 3. Contingency table of binary events for categorical verification scores.

Forecast exceeding a given threshold	Observation exceeding a given threshold	
	Yes	No
Yes	Hits	False Alarms
No	Misses	Correct Negatives

scores theoretically assuming some distribution (e.g. Gaussian distribution for correlation coefficient), the distribution of other scores cannot be modelled exactly or approximated by theoretical distributions. Thus, we have used re-sampling techniques in order to generate an empirical distribution for the values of the evaluation scores to compute sampling uncertainty.

A bootstrap procedure (Efron and Tibshirani, 1993) is used to analyse the sampling uncertainty which addresses the question of what range of scores would be obtained given different sets of forecasts from the same forecast system. We sample forecast-observation pairs randomly with replacement, keeping the forecast and the corresponding observation together. The new sample has the same size as the original. Since it is sampled with replacement, it is likely to include some forecast-observation pairs more than once, and some pairings will not be drawn at all. The verification score is computed from the generated sample. This procedure is repeated many times (typically a few thousand) and the various statistics (e.g. mean, percentiles) are computed from the distribution of the verification scores. The bootstrap procedure is given below.

Pseudo-code for bootstrap procedure
Let $\{x_1, y_1\}, \{x_2, y_2\} \ldots, \{x_n, y_n\}$ be forecast-observation paris and n_B be the number of bootstrap sample
for $i = 1$ to n_B
  Sample n pairs of forecast-observation from the original pair $\{x, y\}$ (with replacement)
  Compute verification scores from n pairs of observation and forecasts
end
Compute various statistics (mean, percentiles) from n_B values of the verifications scores.

In this study, we present absolute evaluation scores rather than scores relative to some reference (e.g. climatology, persistence, etc.). This allows for direct comparison of the precipitation forecasts from NWP models of different spatial resolutions between many stations and at different temporal resolutions. Thus, the term "skill" means absolute evaluation score in this study.

4 Results

The first step of the evaluation is to compare the NWP forecasts to the observations at the point scale (rain gauge stations). Although this eliminates any possible errors due to the spatial interpolation of the station data, errors due to subgrid scale variability and representativeness may remain. For example, the frequency of zero precipitation at a grid cell will necessarily be less than at a randomly selected point within that (because if it rains anywhere, the grid cell precipitation will be non-zero). In Sect. 4.6, we evaluate the skill of NWP model forecasts at the catchment scale.

4.1 Forecasts of 1–24 h lead time

Figure 4 shows a map of 24 h mean precipitation accumulation for the measurement stations and for the ACCESS model grid cells over the Ovens catchment. The 24 h (daily) precipitation on a given date and time (09:00 LT) is the accumulated forecast precipitation of lead times from 1 to 24 h on that date and time. The precipitation is averaged over a period for 1 April 2010 to 8 February 2011. Here, dark blue colour indicates higher precipitation and white is relatively drier. The ACCESS-VT model has a precipitation maximum adjacent to the eastern extremity of the catchment just to the west of Mount Bogong (elevation 1988 m, longitude 147.2°, and latitude −36.8°, between stations 11 and 26). The average daily precipitation forecast at this location is about 10.5 mm. Regrettably the area of the highest forecast precipitation is without a measurement station. The closest measurement station (26) is about 10 km southeast of the highest precipitation forecast location. This station has observed precipitation of 5.89 mm, while the corresponding grid cell forecast by the ACCESS-VT model is 7.95 mm. The measurement stations with the highest precipitation observations are 17 (8.81 mm), 33 (7.03 mm) and 30 (6.88 mm). The forecast precipitation for the corresponding model grid cells for these stations are 7.26 mm, 4.97 mm and 5.98 mm, respectively. The ACCESS-VT model has a tendency to overforecast in lowland areas (north of the catchment) and underforecast in highland areas (south of the catchment).

The ACCESS-A model places the highest precipitation over Mount Feathertop (southeast, near station 25) and east of Wabonga in the southwest interior of the catchment (near stations 16, 20, 21, and 23). The observed precipitation at station 25 is 6.3 mm and the corresponding ACCESS-A forecast is 6.44 mm. Purely based on elevation, one would not necessarily expect a maximum in long-term average precipitation in this ungauged area, although it may be due to the precipitation patterns in this particular year. Like the ACCESS-VT, the ACCESS-A model has a tendency to overforecast in lowland areas and underforecast in highland areas.

The ACCESS-R model has a precipitation minimum in the headwaters of the catchment (east of station 33 and near stations 25, 26 30, and 32). These are some of the wettest areas

Table 4. Categorical verification scores used in the study.

Score	Formula	Range	Perfect
Probability of detection (POD)	Hits/(Hits + Misses)	[0, 1]	1
False alarm ratio (FAR)	False Alarms/(Hits + False Alarms)	[0, 1]	1
Frequency bias (FBI)	(Hits + False Alarms)/(Hits + Misses)	[0, ∞]	1
Critical success index (CSI)	Hits/(Hits + Misses + False Alarms)	[0, 1]	1

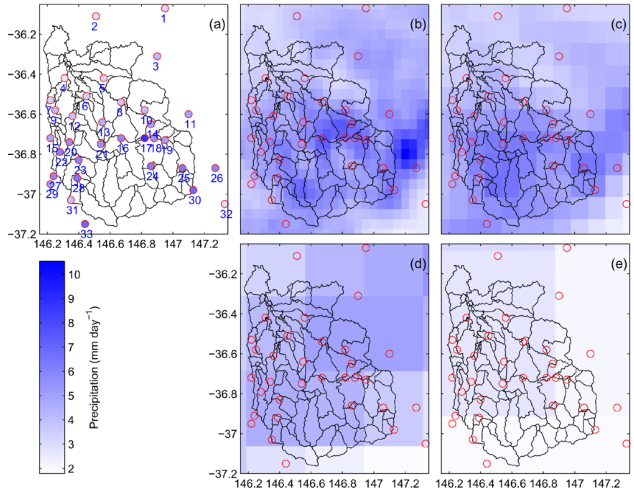

Fig. 4. A comparison of daily average (1–24 h accumulated) observed precipitation at stations and forecasted precipitation by the ACCESS models in Ovens catchment for 1 April 2010 to 8 February 2011: **(a)** Observed station precipitation, **(b)** ACCESS-VT, **(c)** ACCESS-A, **(d)** ACCESS-R, and **(e)** ACCESS-G.

for the high resolution models. Also, the cluster of stations in the southwest part of the catchment has a range of averages that is wide enough to suggest that there is significant within-grid cell variability at this scale. The precipitation maximum is in the northeast corner of the catchment which is a dry region in the higher resolution models. The ACCESS-R model also has a tendency to overforecast in lowland areas and underforecast in highland areas. The coarse resolution model ACCESS-G places the highest precipitation over the northwest of the catchment. Like the ACCESS-R, the ACCESS-G model has a precipitation minimum in the headwaters of the catchment. Note that the ACCESS-G model has only 4 grid cells to cover the entire catchment. Unlike other models, the ACCESS-G underestimates precipitation over the entire catchment. The ACCESS-VT and ACCESS-A model forecasts appear to capture the gradient of precipitation across the catchment although they appear to have less variability than the observations. The ACCESS-R and ACCESS-G model resolutions do not meaningfully represent the fine scale patterns of variability across the catchment. Clearly, downscaling and bias adjustment are operationally recommended for the ACCESS-R and ACCESS-G models.

Figure 5 shows the evaluation scores of the ACCESS models for forecasts of 24 h precipitation accumulations at measurement stations. The RMSE score is shown in Fig. 5a. ACCESS-VT model has a minimum RMSE score of about 5.41 mm day^{-1} for Wangaratta station (4) and a maximum value of 12.06 mm day^{-1} for Falls Creek station (26). ACCESS-A model has the highest RMSE score of 14.64 mm day^{-1} at Cheshunt station (23) and the lowest value of 6.4 mm day^{-1} at Rocky Point station (8). ACCESS-R model has a minimum RMSE value of 6.5 mm day^{-1} at Loombah Reservoir station (15) and a maximum RMSE value of 13.24 mm day^{-1} at Falls Creek station (26). ACCESS-G model has a minimum RMSE value of 4.67 mm day^{-1} at Bloomfield station (5) and a maximum RMSE value of 13.24 mm day^{-1} at Mount Buffalo station (17). Average RMSE values of ACCESS-VT, ACCESS-R and ACCESS-G models for all stations are comparable. In general, the RMSE score does not exhibit any strong spatial pattern with respect to the altitude of the stations.

Figure 5b depicts the bias of the ACCESS model forecasts as a percentage of the observed values. The ACCESS-VT and ACCESS-A models overestimate dry (low elevation) areas by up to 60 % and underestimate wet (high elevation) areas by up to 30 %. This finding supports the hypothesis that orographically enhanced precipitation is underestimated by NWP models. The ACCESS-R model also shows a similar pattern, but the bias is much greater than that of high resolution models. The coarse resolution model ACCESS-G has a systematic positive bias (underforecasting) for all stations and bias generally increases with altitude. The bias of the coarse resolution NWP model is up to 70 %. Other studies have reported the NWP biases on the order of 100 % (see e.g. Clark and Hay, 2004). As the model resolution becomes progressively coarser (i.e. regional and global models), large systematic biases emerge. Unlike RMSE score, the bias shows some spatial pattern.

Spearman rank correlation coefficients between the station precipitation and the corresponding ACCESS model forecasts are shown in Fig. 5c. The correlation coefficients of high resolution models ACCESS-VT, ACCESS-A, and the regional model ACCESS-R are comparable and vary between about 0.7 and 0.8. In some stations like Rosenwhite (10), all these models give consistently lower correlation values (about 0.7) and stations like Cheshnut (23), all models give consistently higher correlation values (about 0.8). The

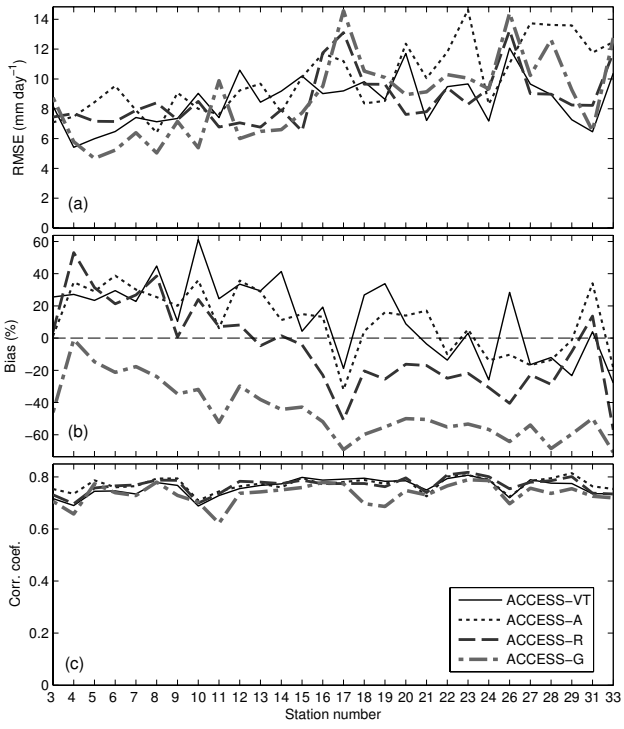

Fig. 5. Evaluation scores of the ACCESS models for daily precipitation forecast at stations: **(a)** RMSE, **(b)** bias, and **(c)** Spearman rank correlation. Stations are ordered according to latitude then longitude.

correlations between station precipitation and ACCESS-G forecasts are generally lower than those of the higher resolution models, and vary across the stations. This may be due to mainly two reasons: (i) the ACCESS-G model resolution is coarse and the spatial variability of precipitation across the stations within a model grid cell is high; and (ii) the spatial variability of the forecast precipitation across the model grid cells is small. Thus, the spatial variability of correlation coefficients (Fig. 5c) comes mainly from the variability of the observed precipitation across the stations and, but not necessarily from the ACCESS-G forecasts. For example, Wangaratta AWS (4) and Mount Buffalo (17) stations share the same value of precipitation forecasts as they lie in the same grid cell of the ACCESS-G model, but have quite different observed precipitation (mean daily values of 2.77 mm vs. 8.81 mm). The variability of mean daily precipitation across the stations (standard deviation of 1.41) is much higher than that of the ACCESS-G model (standard deviation of 0.32). The very low value of correlation at Mongans Bridge (11) may be due to forecast and/or observation outlier in a month of March 2011 (forecast of 150 mm against observation of 5 mm precipitation).

Further analysis has been done to understand the contribution of bias and variance to RMSE (see Eq. 2). The variances of the forecasts and observations are of same order of magnitude. However, the biases of the precipitation forecasts from ACCESS models are much smaller than the standard deviations of the forecasts and observations and, therefore, reducing the biases of the forecasts may not necessarily reduce the RMSE significantly.

4.2 Variation of evaluation scores with forecast lead times

NWP model skill varies with time for three main reasons: the quality of the initial analysis, baroclinic and/or barotropic instability of the large scale flow, and model systematic errors (Stanski et al., 1989). When model forecasts are accurate at the start of a model run it does not necessarily mean it will stay that way or vice versa. Even during the times when the models had more skill overall there can still be some hours where the forecasts are significantly less skilful (Roux and Seed, 2011). In this section, we examine the skill of the NWP model forecasts at different lead times. We present analysis of forecasts from the ACCESS-G model because it has the longest lead time. We focus on a single precipitation station, Carboor Upper (13), which is close to the centre of the Ovens catchment and ACCESS-G model grid cell, and analyse forecast skill of 3 h precipitation accumulations. Analysis of other models and locations produces similar results. The score for 3 h precipitation accumulations at 3 h lead time means the score of total precipitation for the period 09:00–12:00 LT.

Figure 6 shows the forecast skill of 3 h precipitation accumulations for the ACCESS-G model at Carboor Upper station. The RMSE score standardised by standard deviation of observations (sRMSE) is shown in Fig. 6a. The sRMSE value will be greater than 1 when the mean-square error (MSE) exceeds the variance of the observation. This is analogue to a negative value of Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970) when the MSE exceeds the variance of the observation. For the rest of analysis, we present sRMSE instead of RMSE as the former is independent to the magnitude of the data. The sRMSE score displays considerable variation with lead time. The sRMSE score is below 1 for lead times up to 39 h and subsequently fluctuates around 1. Figure 6b shows that the forecast bias varies significantly at different lead times and shows some diurnal cycle. Further investigation into the diurnal cycle is presented in Sect. 4.5. The forecasts have a bias of up to 75 % and consistently underestimate 3 h precipitation accumulations for most lead times. Figure 6c shows the Spearman rank correlation coefficient between forecast and observed of 3 h precipitation accumulations. One can see that the skill with respect to correlation coefficient decreases with lead time which is not obvious in sRMSE and bias mentioned before. The correlation coefficient starts with a value of about 0.6 at the shortest lead time and decreases to a value of about 0.1 at the longest lead time.

Figure 6 also shows the 95 % confidence intervals of sampling uncertainty for the evaluation scores using 10 000 number of samples. Although this number seems somewhat

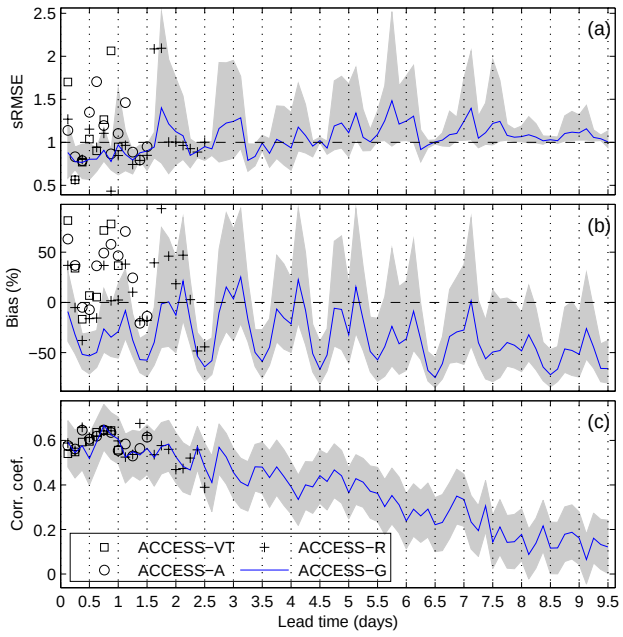

Fig. 6. Evaluation scores of the ACCESS-G model for 3 h accumulated precipitation forecasts at Carboor Upper station at different lead times: **(a)** sRMSE, **(b)** bias, and **(c)** Spearman rank correlation. The shaded area corresponds to the 95 % confidence intervals of sampling uncertainty.

arbitrary, an analysis of the convergence of the mean of evaluation scores (results not shown) suggests that this number is sufficient. The top panel shows that the sRMSE score has a considerable sampling uncertainty (light shaded area) which varies at different lead times. Particularly at 39 and 138 h, the uncertainty is very large, indicating that some extreme events strongly influence the sRMSE score. Further analysis of forecasts at these lead times shows on the one hand the model is not able to forecast some extreme events, but on the other hand the model is producing very large forecasts for some low events.

Figure 6b illustrates sampling uncertainty in the bias score. Like the sRMSE, this score also reveals that there is a considerable sampling uncertainty and particularly at 42, 75, 123 h, and some other forecast hours, uncertainty of the bias score is very large. The 95 % confidence intervals of sampling uncertainty associated with the Spearman rank correlation coefficient are presented in Fig. 6c, which seem to be more symmetrical than for other scores. They are consistent with the correlation coefficients between precipitation forecasts and the corresponding observations and do not fluctuate like other scores as Spearman correlation is less sensitive to the extreme values. The skill of ACCESS-VT, ACCESS-A and ACCESS-R models is also shown in Fig. 6 to compare with ACCEE-G model. sRMSE value of ACCESS-G model is as good as that of other models (up to 2.5 days lead time). The bias of other models is positive (over forecasting) for most of the lead time, whereas it is negative for ACCESS-G. Correlation coefficients of all models are comparable. These results are consistent with the results obtained from Fig. 5.

Figure 7 shows the categorical evaluation scores and their 95 % confidence intervals as a function of forecast lead time. In this study, threshold value of 0.1 mm $(3\,\mathrm{h})^{-1}$ is considered to define the precipitation event "yes" or "no". A non-zero threshold is imposed because there is a minimum measurable precipitation amount for the operational tipping bucket rain gauges. Figure 7a shows the POD score of the model forecasts. As expected the score decreases with increasing lead time. For example, at the shortest lead time more than 70 % of the observed events are correctly detected, while at the longest lead times, the POD score reduces to 30 %. Like continuous scores, the sampling uncertainty is quite large. Figure 7b shows that the FAR score increases with lead time which is consistent with the POD score. The FAR score increases from a value of about 0.5 at the shortest lead time to 0.75 at the longest lead time. As far as uncertainty results are concerned, the FAR score behaves similar to the POD.

The equivalent diagram for the FBI as a function of forecast lead time is shown in Fig. 7c. Unlike the POD and FAR scores, the FBI score does not increase with lead time, rather it fluctuates around a value of 1.3 and shows evidence of a diurnal cycle. Comparing the bias (Fig. 6c) and frequency bias (Fig. 7c) of the forecasts produces an interesting result; forecasts of the precipitation amount tend to be too low, but the occurrence of precipitation is overestimated for most forecast lead times. This indicates that the model forecasts small amounts of precipitation too frequently. This is the well known behaviour of many NWP models and has been reported elsewhere. One can notice a considerable sampling uncertainty in the FBI score as well.

The CSI score reported in Fig. 7d displays results similar to the POD and the FAR scores. The CSI score is similar to the POD except it also considers false alarms. If there are no false alarms, then both scores are equal. Thus, the CSI score is smaller than the POD. For the ACCESS-G model forecasts, the score varies from about 0.45 at the shortest lead time to about 0.15 at the longest lead time. Likewise the uncertainty results are similar to that in the POD score; however, the variation across the lead times is smaller.

The categorical skill of the reference forecasts is also shown in Fig. 7. The reference forecasts are generated using a permutation procedure (see e.g. Mason, 2008; Deque, 2012). The permutation procedure generates a new set of forecasts-observation pairs in which observation are unrelated to the forecasts except by chance. This procedure addresses the question of what is the chance that the given value of evaluation score could have been obtained by accident. The mean scores of 10 000 such reference forecasts are shown in dashed lines. Note that FBI of the reference forecasts is same as that of the ACCESS forecasts; hence, it is not shown in the figure. The results show that ACCESS-G model might not necessarily have significant skill beyond 7 days given sampling uncertainty.

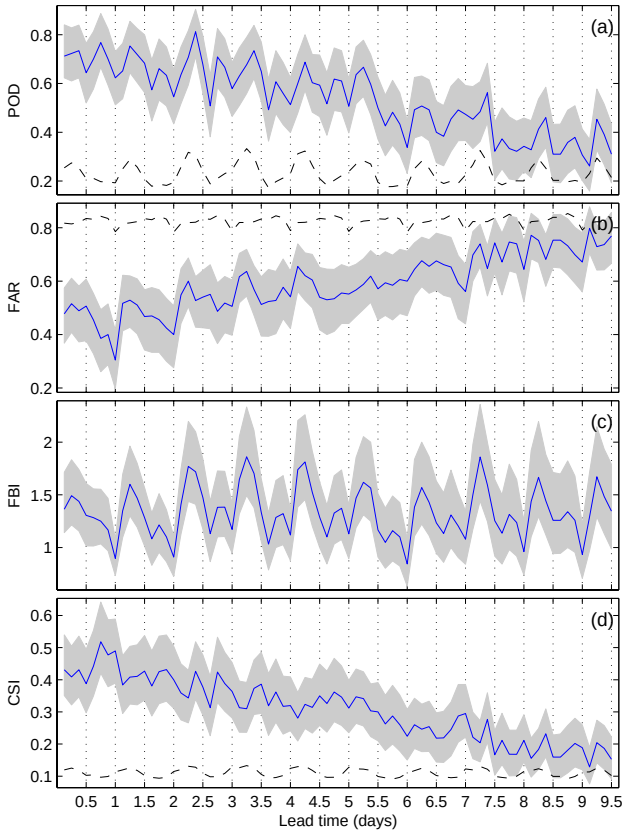

Fig. 7. As in Fig. 6, but for categorical evaluation scores: **(a)** POD, **(b)** FAR, **(c)** FBI, and **(d)** CSI. The mean scores (except FBI) of the reference forecasts are also shown in dashed lines.

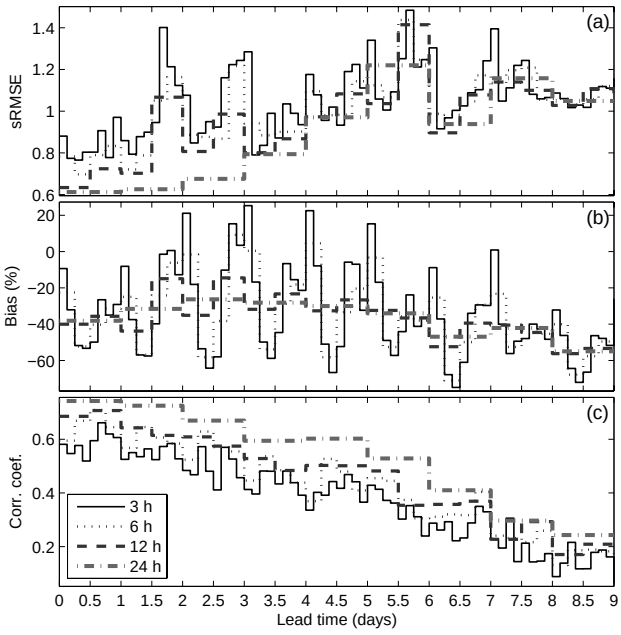

Fig. 8. A comparison of the evaluation scores of the ACCESS-G model for different temporal precipitation accumulation periods at Carboor Upper station: **(a)** sRMSE, **(b)** bias, and **(c)** Spearman correlation coefficient.

4.3 Variation of evaluation scores with precipitation accumulation periods

An analysis of evaluation scores of the ACCESS-models (except ACCESS-G) indicates that the skill of the hourly precipitation forecasts is very low and varies significantly from hour to hour (results not shown). However, there is some skill for forecasts of 3 h precipitation accumulations. Increases in forecast skill due to temporal accumulation arise because errors in the timing of precipitation decrease. In this section, we have further analysed the scores of forecasts from the ACCESS-G model for different accumulation periods (Fig. 8). The sRMSE score is the highest (about 1.48) at 136 h lead time for 3 h precipitation accumulations (Fig. 8a). This drops to 1.44, 1.41 and 1.22 for 6, 12 and 24 h precipitation accumulations, respectively. At the lead times between 36 and 72 h, the sRMSE skill increases (or sRMSE score decreases) significantly from shorter accumulation periods to longer ones. Further analysis of sampling uncertainty (not shown) supports the finding that skill at 24 h accumulation period is significantly better than skill at 3 h accumulation period at shorter lead times. For the longer lead times, the skill at all accumulation periods is not significantly different.

Figure 8b shows that the maximum bias of −75 % is reduced to −46 % when accumulation period increases from 3 to 24 h at the lead times between 144 and 168 h. The bias of forecasts of 24 h precipitation accumulations decreases from −38 % at 1 day to −26 % at 3 days lead time and then increases to −54 % at the longest lead time. The model is overestimating 3 h precipitation accumulations for some lead times (e.g. 51, 75, 99 and 123 h). For the corresponding periods, the biases of the 24 h precipitation accumulations are negative (underestimating) because the biases of other 3 h precipitation accumulations within these periods are negative and the net effect is negative.

Figure 8c shows the Spearman correlation coefficients between forecast and observed precipitation as a function of lead time and accumulation period. The Spearman correlation coefficient displays less variation than the other two scores, because it is less sensitive to outliers and extreme events. The correlation increases from 0.52 at the shortest accumulation period (3 h) to 0.74 for the longest accumulation period (24 h) at the shortest lead time. It is observed that a plot of correlation coefficients for 24 h precipitation accumulations now exhibits smooth monotonic decay which now seems to have less affected by sampling fluctuations.

The analysis presented in this section suggests that, in general, the skill of ACCESS-G precipitation forecasts increases with increasing accumulation period. However, the appropriate accumulation period to adopt will depend not only upon the forecast skill, but also upon the intended use of NWP

precipitation forecasts. For example for flood forecasting applications, daily forecasts are likely to be too coarse as the flood peak may remain for only a few hours. For other purposes such as water resources management, hourly precipitation forecasts may not be needed. Further analysis is required to select the optimal temporal resolution for streamflow forecasting purposes.

4.4 Variation of evaluation scores with precipitation threshold values

In Sect. 4.2, we presented the categorical evaluation scores of 3 h total precipitation forecasts from the ACCESS-G model for a threshold value of 0.1 mm $(3\,\mathrm{h})^{-1}$. The skill of the NWP precipitation forecast may also be expected to vary with precipitation intensity. We evaluate the skill of the ACCESS-G model forecasts for threshold values of 0.1, 1, 2, 5, 10, and 20 mm day^{-1} based on recommendations of WWRP/WGNE (2008).

Figure 9 depicts the categorical evaluation scores of the ACCESS-G forecasts as a function of precipitation threshold value. The scores are computed for forecasts of 24 h precipitation accumulations for lead times of 1 to 9 days. The categorical evaluation scores are strongly related to the threshold and in general, decrease with increasing threshold values. For example the POD score (Fig. 9a) decreases from about 0.8 for low threshold value (1 mm day^{-1}) to about 0.35 for precipitation amounts above 20 mm day^{-1} for forecasts of the first 24 h. Furthermore, as expected the POD score decreases with increasing lead times. The scores for the high threshold values must be used with care because only few cases may occur, for example 11.9 % of all cases occur for threshold greater than 10 mm day^{-1}, 7.2 % for threshold greater than 20 mm day^{-1}.

The remaining panels show the FAR (Fig. 9b), the FBI (Fig. 9c), and the CSI (Fig. 9d) scores. Consistent with the POD score, the FBI and the CSI decreases with increasing threshold values whereas the FAR score increases with increasing threshold values. At 1 and 2 days lead time, the FAR score decreases, whereas at 3 to 6 days lead time, it first decreases for low threshold values and then increases for higher values. From Fig. 9c it can be seen that, for a low threshold value (e.g. 0.1 mm day^{-1}) the FBI score is greater than 1 at all lead times whereas for higher threshold values it is less than 1. This indicates that occurrence of precipitation or light precipitation is overestimated while the heavy precipitation events are consistently underestimated. As far as CSI score is concerned, it increases slightly at low threshold values of 1 and 2 mm day^{-1} at shorter lead times (1 to 3 days) and then decreases, which is consistent with FAR score.

One sample t test indicates that the evaluation scores for higher precipitation threshold values are significantly different (at 5 % significant level) than that of lower threshold values for all lead times. All evaluation scores except FAR for lower threshold value are significantly different for longer

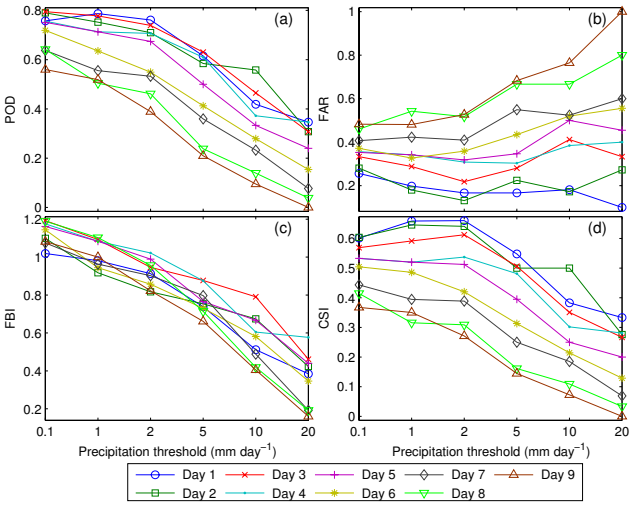

Fig. 9. Categorical evaluation scores of the ACCESS-G model for 24 h accumulated precipitation forecasts as a function of precipitation threshold and lead time: **(a)** POD, **(b)** FAR, **(c)** FBI, and **(d)** CSI.

lead times. Further analysis shows that all evaluation scores except FBI for longer lead times (day 8 and 9) is significantly different for all precipitation threshold values. FAR and CSI scores for shorter lead times (day 1 and 2) are significantly different for all precipitation threshold values. Note that sample sizes for the significant test of evaluation scores at different precipitation thresholds and forecast lead time are 6 and 9, respectively.

4.5 Further results

Results from Fig. 6 indicate that there might be some diurnal cycle in the evaluation scores, particularly for the bias. We investigate the diurnal cycle of the observed precipitation and corresponding ACCESS-R model forecasts at Carboor Upper station. ACCESS-R is chosen for this analysis because ACCESS-G precipitation forecasts are not available at hourly temporal resolution and a more thorough analysis of the diurnal cycle would require a forecast length beyond 24 h. Figure 10 shows the diurnal cycle of observed precipitation at the station and the ACCESS-R forecasts for the corresponding model grid cell. Observed precipitation displays a diurnal cycle, with maximum at 07:00, then 10:00 and 11:00 UTC and minimum at 01:00 UTC. This finding for Carboor Upper station is consistent with results reported by Westra and Sharma (2010) that the hourly maximum and minimum in precipitation occurrence was found between 08:00 and 10:00 UTC and between 23:00 and 24:00 UTC, respectively for more than 80 % of Australian stations. The precipitation forecasts do not seem to have a diurnal cycle except the outlier at 13:00 UTC which is the first hour of the forecast. Poorly representing the timing and magnitude of the diurnal cycles, particularly in precipitation, is a known

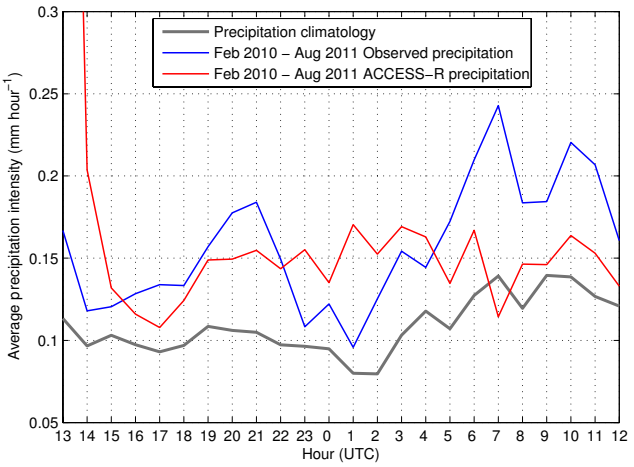

Fig. 10. Diurnal cycle of the observed and the ACCESS-R precipitation forecasts. Precipitation climatology (thick gray line) is based on a period from September 1991 to February 2011.

problem with many NWP models and is commonly related to the representation and parameterisation of convective processes (Kaufmann et al., 2003; Dai and Trenberth, 2004; Evans and Westra, 2012).

Since it is difficult to see the diurnal cycle of the evaluation scores of the ACCESS-R model (because the lead time is only up to 60 h), we have further analysed the diurnal cycle for the bias of the ACCESS-G model. From Fig. 6b, some of the lowest biases are at 27, 51, 75, 99 h lead times, which corresponds to 12:00 LT. This is consistent with the minimum value of observed precipitation which occurs at 11:00 (12:00 during daylight saving) LT. Similarly the maximum bias occurs at 21:00 LT while the maximum values of observed precipitation are around 18:00–21:00 LT (daylight saving time). Thus, there is some consistent between the timing of hourly maximum and minimum of the observations and the bias score. The cyclic nature of the biases in the ACCESS-G model precipitation forecasts is likely the product of the limited ability of the model to describe the diurnal cycle. Furthermore, given that the precipitation forecasts do not seem to have a pronounced diurnal cycle, the bias score being linear (as opposed to e.g. RMSE which is quadratic) exhibits similar cyclic patterns as the observations. Synthetic data were generated to understand the diurnal cycle in the forecast skill. About 20 yr of daily precipitation data from one of the stations were disaggregated to hourly precipitation using sine curve of one cycle period. The hourly forecast values were generated disaggregating uniformly from daily precipitation value by adding some random noise. RMSE, bias and correlation coefficients scores were computed from these synthetic data. The results (not shown) support the finding that the evidence of diurnal cycle in observation is likely to be seen in the bias score compared to sRMSE or correlation coefficient score.

4.6 Evaluation at catchment scale

Previous sections presented the evaluation scores of the ACCESS model precipitation forecasts at point scale (i.e. at rain gauge station). For hydrological applications the localisation of precipitation is important at the catchment scale so that it is useful to evaluate precipitation forecasts on catchment averages (e.g. Oberto et al., 2006; Rossa et al., 2008). We are using lumped model GR4J (Perrin et al., 2003) for each sub-catchment and the flow from each sub-catchment is routed to the outlet of the catchment using Muskingum channel routing algorithms. Thus, average precipitation over sub-catchment is used input to the GR4J model for hydrological forecasting. BoM currently uses the event-based model URBS (Malone, 1999) for real-time flood forecasting in Australia. URBS is a lumped model which uses a single catchment average forecast rainfall as compared to sub-catchment average rainfall for the GR4J model. BoM is planning to use continuous modelling with semi distributed lumped model (connected lumped model) for real-time flood forecasting services in Australia.

Figure 11 gives the performance scores for forecasts for 3 h accumulations of catchment average precipitation. These results smooth over some of the errors related to displacement and are a better indicator of the quality of forecasts of precipitation volume. Compared to station precipitation (Fig. 6a), the sRMSE score (Fig. 11a) of the catchment average precipitation exhibits similar pattern, but the magnitude of the score is lower. As expected, the sRMSE score of the ACCESS-G model, in general, increases with increasing lead times (e.g. 0.8 at 3 h lead time to about 1.0 at the longest lead time). The 95 % sampling uncertainty plot shows that there is a considerable sampling variation in the sRMSE score. For several lead times (e.g. at 42, 180 h), the uncertainty is very large, indicating that some extreme events strongly influence the sRMSE score.

Figure 11b depicts the bias score of catchment average precipitation forecasts from the ACCESS-G model. Systematic biases in the forecasts are evident, where it struggles to produce high enough intensity forecasts. The bias of the ACCESS-G model forecasts is around −48 % at the shortest lead time, then fluctuates around −50 %, and finally reaches to about −67 % at the longest lead time. The ACCESS-G model forecasts tend to be lower than the interpolated catchment average precipitation. Like station precipitation, a diurnal cycle is present in the bias score of catchment average precipitation. The uncertainty analysis shows that there is also a considerable sampling uncertainty in the bias score and this is not surprising given that only one year of data is used which has some extreme precipitation events.

Figure 11c shows the Spearman correlation coefficient between observed and catchment average rainfall forecasts of 3 h total from ACCESS-G model. Unlike other two scores, correlation exhibits a relatively smooth decay as the lead time increases. The correlation coefficient declines from about 0.7

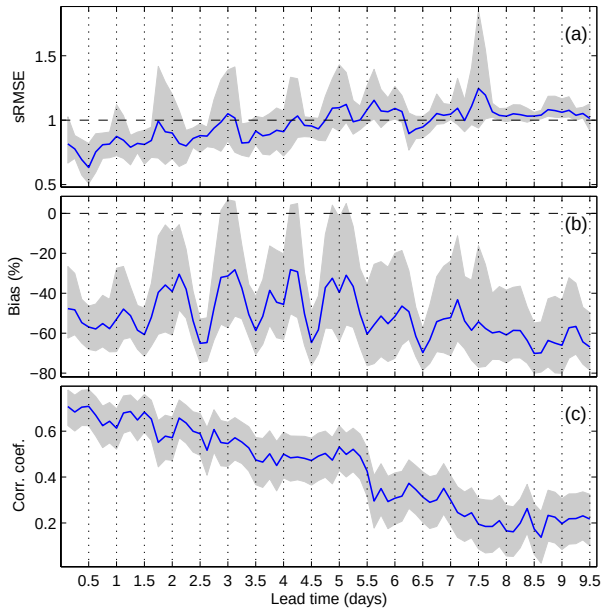

Fig. 11. As in Fig. 6, but for the catchment average precipitation above Wangaratta stream gauging station: **(a)** aRSME, **(b)** bias, and **(c)** Spearman rank correlation.

at the shortest lead time to about 0.22 at the longest lead time. The 95 % sampling uncertainty shows similar behaviour like that of the station precipitation (Fig. 6c).

Figure 12 shows the categorical evaluation scores for catchment average precipitation forecasts. In general, these scores exhibit similar patterns to station precipitation. However, as expected the skill of the catchment average precipitation forecasts are higher. The POD of the ACCESS-G model forecasts decreases from about 0.68 at the shortest lead time to about 0.38 at the longest lead time (Fig. 12a). Similarly, the FAR of the ACCESS-G model forecasts increases from about 0.2 at the shortest lead time to about 0.53 at the longest lead time (Fig. 12b). Figure 12c shows that there is a significant variation in the FBI score across the lead times and a diurnal cycle similar to that of station precipitation is present (Fig. 7c). However, the FBI of the catchment average precipitation forecasts is less than 1 for most lead times, whereas it is greater than 1 for the station precipitation. This difference is logical because if there is precipitation at any measurement station within the catchment, then the catchment average precipitation is non-zero and the probability of observed rain events is higher. The last panel shows that the CSI score, like other scores, is higher than that of the station precipitation. It decreases from about 0.6 at the shortest lead time to about 0.27 at the longest lead time. Uncertainty analysis of the categorical evaluation scores is also reported in Fig. 12. The results are consistent with the continuous evaluation scores. The mean scores of the reference forecasts are shown in Fig. 12. The results are consistent with the station precipitation that the ACCESS-G model is unlikely to have significant skill beyond 7–8 days.

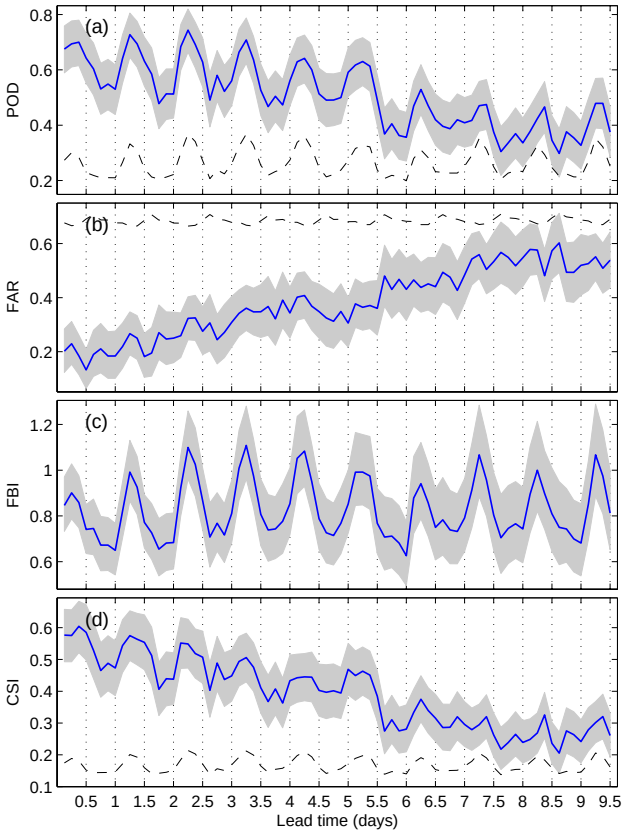

Fig. 12. As in Fig. 7, but for the catchment average precipitation above Wangaratta stream gauging station: **(a)** POD, **(b)** FAR, **(c)** FBI, and **(d)** CSI. The mean scores (except FBI) of the reference forecasts are also shown in dashed lines.

5 Discussion

There is a general perception that the variability of NWP model output does not match the observed variability. Specifically, it is thought that there is a tendency for too frequent small amounts of precipitation in the NWP model output. NWP models are much less successful in their handling of low level stratiform cloud, and generally have a tendency to overestimate light precipitation (Golding, 2000). The results (not shown) reveal that the ACCESS models have a tendency to have too many small precipitation events. For events less than 0.13 mm h^{-1} the model frequency is greater than that of the observed data. For events between 0.13 and 2 mm h^{-1}, the models do not produce enough events. For events greater than 2 mm h^{-1}, the sample size is not big enough to draw reliable conclusions. For all precipitation frequencies the ACCESS-G global model does not produce enough intense events.

The NWP forecasts and observations are highly skewed and the error does not necessarily appear to be linear in log-transformed space. Specifically, both time series contain many zeros and the relative error can be very large for small

precipitation amounts. NWP post-processing method relying on Gaussian error distribution would need to transform the observations and forecasts in a way that the variables or residuals are relatively normally distributed. However, undue weight should not be placed on the small precipitation amounts as these are relatively inconsequential for flood and streamflow forecasting applications.

The NWP models do not appear to be the most skilful at 1 or 3 hourly temporal resolutions and their native spatial resolutions (i.e. individual grid cells). There is greater skill when the NWP model forecasts are averaged over coarser spatial and temporal resolutions (e.g. catchment average, daily average). Further analysis is necessary to determine the optimal resolution for extracting useful information from NWP, as this resolution may depend on the catchment and/or season. However, any techniques for quantifying NWP forecast uncertainty that use only the native resolution data may unnecessarily conclude that the NWP forecasts contain no skill.

Prior to the commencement of this study, it was anticipated that the NWP precipitation forecasts would have significant and systematic biases that would have to be corrected to make them useful for predicting streamflow. Even if the precipitation forecasts were good at predicting the "true" precipitation (i.e. what actually fell on the catchment), the "measured" precipitation may depend on the mix of available station data. Operational datasets used for streamflow forecasting contain a subset of the full station network because of the requirement that data be available in real time and have a long records. As a result the geographic characteristics of stations used for operational streamflow forecasting may not be representative of the catchment as a whole (e.g. clustered in valley bottoms). The data for the Ovens catchment used in this study passed through a thorough quality control and infilling process which produced serially complete hourly data at stations, checked against an independent gridded precipitation dataset. Such processes often cannot be performed in real time and, therefore, the observed data used in this study are closer to the true precipitation than the data currently used in the operational system (which does not check for flat lined sensors and does not infill missing data). The satellite observations (see e.g. Xie and Arkin, 1996; Skomorowski et al., 2001; McPhee and Margulis, 2005; Joshi et al., 2012) can be used to estimate precipitation in area where the density of rain gauge networks are very poor (e.g. in the central part of the Australia). These precipitation estimates are useful for NWP data assimilation. However, the temporal and spatial resolution of satellite observations is too coarse to be used for the short-term streamflow forecasting purpose.

The skill of the precipitation forecasts from the NWP models at two nearby stations can be quite different because (i) they (the stations) are in same model grid cell, but have different precipitation observation (observed variability), or (ii) they are in different model grid cells, thus, have different forecasts (forecast variability), or (iii) they are in different model grid cells, have similar forecasts, but different precipitation observation (model's inability to resolve the scale). In this study, precipitation forecasts from the NWP models are compared with station and catchment average precipitation whose spatial resolution is different than that of the model. A better understanding of the quality of forecasts would be gained if the spatial resolution of the model matches with that of the observation. Since the catchment average precipitation is interpolated from nearby station, the skill of spatial evaluation results would have been influenced by interpolation method used. Cherubini et al. (2002) showed that evaluation scores computed by comparing model grid box values to gridded rainfall data were more favourable than those computed by comparing interpolated model output to the original point observations.

It is believed that the skill of NWP model also depends on the season. Australia has a highly variable climate. Rainfall in this continent is largely influenced by El Niño and La Niña events. Thus, it is very difficult to draw any conclusions about seasonality based on only one year of data. This study has evaluated the precipitation forecasts for conditions where precipitation is principally due to large scale synoptic systems. Large scale synoptic systems tend to be better predicted by NWP models because they tend to evolve relatively slowly and occur on spatial scales that are resolved by the models (Roux and Seed, 2011; Roux et al., 2012). NWP models tend not to predict precipitation from convective systems well because there processes evolve rapidly and commonly occur on spatial scales finer than those resolved by the model. Further work has been planned to extend experiments for catchments experiencing a range of climatic conditions in Australia, particularly in areas where significant precipitation is the result of convective processes.

6 Conclusions

This study evaluates the performance of precipitation forecasts from the latest generation of Australian Numerical Weather Prediction (NWP) models over the Ovens catchment in Southeast Australia. The precipitation forecasts from four NWP models (viz. ACCESS-G, ACCESS-R, ACCESS-A and ACCESS-VT) are compared to observed precipitation at measurement stations and to interpolated catchment average precipitation over one year period. A number of continuous and categorical evaluation scores have been used to assess the skill of the ACCESS models at different lead times and temporal resolutions. The effect of diurnal cycle of the precipitation observations and sampling uncertainty in the model performance is also investigated.

The results show that the skill of the NWP precipitation forecasts varies a lot across the stations with some structure with respect to the altitude of the stations. The high resolution models ACCESS-VT and ACCESS-A overestimate 24 h precipitation accumulations in dry, low elevation areas by up to 60 % and underestimate 24 h precipitation accumulations

in wet, high elevation areas up to 30 %. The low resolution model ACCESS-G underestimates 24 h precipitation accumulations by up to 70 % over all stations and in general, the bias increases with the altitude. The correlation of the high resolution NWP (ACCESS-VT, and ACCESS-A) and the regional (ACCESS-R) models is as good as of the low resolution model (ACCESS-G). Overall, high resolution NWP models capture the variability of the precipitation across the stations and perform better at predicting aggregated precipitation amount that the precise location or timing of the precipitation. There is a tendency for small amounts of precipitation to be forecasted too frequently by the NWP models.

The skill of the NWP model forecasts varies significantly with forecast lead time. In general, forecast skill decreases with the lead time, however, there are many instances where the skill at shorter lead times is lower than at longer lead times. This can be attributed to mainly sampling and diurnal variation. Observed precipitation displays a diurnal cycle, with maximum mean precipitation occurring between 17:00 and 21:00 LT, while the NWP precipitation forecasts fails to capture the cycle. Consequently some evaluation scores such as bias and frequency bias show the evidence of the diurnal cycle which is consistent with that of the observation. Uncertainty analysis reveals that the evaluation scores have a significant sampling variation. The NWP forecasts appear to have little skill when evaluated at a short temporal resolution (e.g. hourly or 3 hourly). The skill of the forecasts increase with increasing precipitation accumulation periods (at least up to 24 h) because timing errors in individual periods will tend to compensate for each other.

The skill of the ACCESS model forecasts is higher at the catchment scale than for measurement stations. Spatial averaging of precipitation over a catchment reduces displacement errors and provides a better indicator of the quality of the forecast of precipitation volume. Systematic biases in the global ACCESS model are also evident in catchment average precipitation forecasts. The model struggles to produce high enough intensity forecasts. The resolution of the global model is too coarse to resolve the small catchment scale.

Future work is planned to assess the benefits of using the NWP precipitation forecasts for short-term streamflow forecasting. Our findings here suggest that it is necessary to remove the systematic biases in precipitation forecasts, particularly those from low resolution models, before the forecasts can be used for streamflow forecasting. Post-processing techniques to remove biases and reliably quantify precipitation forecast uncertainty are being currently developed and tested by the authors.

Acknowledgements. This work is a part of the water information research and development alliance between CSIRO's Water for a Healthy Country Flagship and the Bureau of Meteorology. David Enever (Bureau of Meteorology) and Chris Leahy (Bureau of Meteorology) provided some of the data for this work. Alan Seed (Bureau of Meteorology) and Belinda Roux (Bureau of Meteorology) provided precipitation forecasts from the NWP models. Phil Ward (CSIRO) helped extracting and infilling station precipitation data. The authors sincerely thank Massimiliano Zappa and other two anonymous referees for their helpful and constructive comments.

Edited by: E. Morin

References

Ament, F., Weusthoff, T., and Arpagaus, M.: Evaluation of MAP D-PHASE heavy precipitation alerts in Switzerland during summer 2007, Atmos. Res., 100, 178–189, doi:10.1016/j.atmosres.2010.06.007, 2011.

BoM: Operational implementation of the ACCESS numerical weather prediction systems, NMOC, Operations Bulletin, No. 83, available at: http://www.bom.gov.au/australia/charts/bulletins/apob83.pdf (last access: March 2013), Melbourne, Australia, 2010.

Casati, B., Wilson, L., Stephenson, D., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U.,Ebert, E., Brown, B., and Mason, S.: Forecast verification: current status and future directions, Meteorol. Appl., 15, 3–18, 2008.

Cherubini, T., Ghelli, A., and Lalaurette, F.: Verification of precipitation forecasts over the Alpine region using a high-density observing network, Weather Forecast., 17, 238-249, 2002.

Clark, M. P. and Hay, L. E.: Use of medium-range numerical weather prediction model output to produce forecasts of streamflow, J. Hydrometeorol., 5, 15–32, 2004.

Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: a review, J. Hydrol., 375, 613–626, doi:10.1016/j.jhydrol.2009.06.005, 2009.

Cuo, L., Pagano, T. C., and Wang, Q.: A review of quantitative precipitation forecasts and their use in short-to medium-range streamflow forecasting, J. Hydrometeorol., 12, 713–728, doi:10.1175/2011JHM1347.1, 2011.

Dai, A. and Trenberth, K. E.: The diurnal cycle and its depiction in the community climate system model, J. Climate, 17, 930–951, doi:10.1175/1520-0442(2004)017<0930:tdcaid>2.0.CO;2, 2004.

Damrath, U., Doms, G., Fruhwald, D., Heise, E., Richter, B., and Steppeler, J.: Operational quantitative precipitation forecasting at the German Weather Service, J. Hydrol., 239, 260–285, 2000.

Deque, M.: Continous variables, in: Forecast Verification: a Practitioner's Guide in Atmospheric Science, edited by: Jolliffe, I. T. and Stephenson, D. B., Wiley, West Sussex, UK, 97–120, 2012.

Ebert, E. E., Damrath, U., Wergen, W., and Baldwin, M. E.: The WGNE assessment of short-term quantitative precipitation forecasts, B. Am. Meteorol. Soc., 84, 481–492, 2003.

Efron, B. and Tibshirani, R.: An Introduction to the Bootstrap, Chapman & Hall/CRC, Boca Raton, Florida, US, 1993.

Evans, J. and Westra, S.: Investigating the mechanisms of diurnal rainfall variability using a regional climate model, J. Climate, 25, 7232–7247, doi:10.1175/jcli-d-11-00616.1, 2012.

Gebhardt, C., Theis, S., Krahe, P., and Renner, V.: Experimental ensemble forecasts of precipitation based on a convection-resolving model, Atmos. Sci. Lett., 9, 67–72, doi:10.1002/asl.177, 2008.

Georgakakos, K. P. and Hudlow, M. D.: Quantitative precipitation forecast techniques for use in hydrologic forecasting, B. Am.

Meteorol. Soc., 65, 1186–1200, 1984.
Ghelli, A. and Ebert, E.: Special issue on forecast verification, Meteorol. Appl., 15, p. 1, doi:10.1002/met.69, 2008.
Ghile, Y. and Schulze, R.: Evaluation of three numerical weather prediction models for short and medium range agrohydrological applications, Water Resour. Manage., 24, 1005–1028, doi:10.1007/s11269-009-9483-5, 2010.
Golding, B.: Quantitative precipitation forecasting in the UK, J. Hydrol., 239, 286–305, 2000.
Habets, F., LeMoigne, P., and Noilhan, J.: On the utility of operational precipitation forecasts to served as input for streamflow forecasting, J. Hydrol., 293, 270–288, doi:10.1016/j.jhydrol.2004.02.004, 2004.
Hay, L. and Clark, M.: Use of statistically and dynamically downscaled atmospheric model output for hydrologic simulations in three mountainous basins in the Western United States, J. Hydrol., 282, 56–75, 2003.
Jolliffe, I. T.: Uncertainty and inference for verification measures, Weather Forecast., 22, 637–650, doi:10.1175/waf989.1, 2007.
Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: a Practitioner's Guide in Atmospheric Science, Wiley, West Sussex, England, 2012.
Jones, D., Wang, W., and Fawcett, R.: High-quality spatial climate data-sets for Australia, Aust. Meteorol. Oceanogr. J., 58, 233–248, 2009.
Joshi, M. K., Rai, A., and Pandey, A.: Validation of TMPA and GPCP 1DD against the ground truth rain-gauge data for Indian region, Int. J. Climatol., online first, doi:10.1002/joc.3612 2012.
Kasahara, A., Mizzi, A. P., and Donner, L. J.: Impact of cumulus initialization on the spinup of precipitation forecasts in the tropics, Mon. Weather Rev., 120, 1360–1380, 1992.
Kaufmann, P., Schubiger, F., and Binder, P.: Precipitation forecasting by a mesoscale numerical weather prediction (NWP) model: eight years of experience, Hydrol. Earth Syst. Sci., 7, 812–832, doi:10.5194/hess-7-812-2003, 2003.
Malone, T.: Using URBS for real time modelling, 25th Hydrology and Water Resources Symposium, Brisbane, 1999.
Mason, S.: Understanding forecast verification statistics, Meteorol. Appl., 15, 31–40, 2008.
McBride, J. L. and Ebert, E. E.: Verification of quantitative precipitation forecasts from opera-tional numerical weather prediction models over Australia, Weather Forecast., 15, 103–121, doi:10.1175/1520-0434(2000)015<0103:voqpff>2.0.CO;2, 2000.
McPhee, J. and Margulis, S. A.: Validation and error characterization of the GPCP-1DD precipitation product over the contiguous United States, J. Hydrometeorol., 6, 441–459, 2005.
Murphy, A.: Skill scores based on the mean square error and their relationships to the correlation coefficient, Mon. Weather Rev., 116, 2417–2424, 1988.
Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, 1970.
Oberto, E., Turco, M., and Bertolotto, P.: Latest results in the precipitation verification over Northern Italy, COSMO Newsletter, 6, 180–184, available at: http://www.cosmo-model.org/content/model/documentation/newsLetters/newsLetter06/cnl6_oberto.pdf (last access: March 2013), 2006.
Pappenberger, F., Scipal, K., and Buizza, R.: Hydrological aspects of meteorological verification, Atmos. Sci. Lett., 9, 43–52, doi:10.1002/asl.171, 2008.
Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol, 279, 275–289, 2003.
Raupach, M. R., Briggs, P. R., Haverd, V., King, E. A., Paget, M., and Trudinger, C. M.: Australian Water Availability Project, CSIRO Marine and Atmospheric Research Component: Final Report for Phase 3. CAWCR Technical Report No. 013, Canberra, Australia, 2008.
Richard, E., Cosma, S., Benoit, R., Binder, P., Buzzi, A., and Kaufmann, P.: Intercomparison of mesoscale meteorological models for precipitation forecasting, Hydrol. Earth Syst. Sci., 7, 799–811, doi:10.5194/hess-7-799-2003, 2003.
Roberts, N. M.: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model, Meteorol. Appl., 15, 163–169, 2008.
Roberts, N. M., Cole, S. J., Forbes, R. M., Moore, R. J., and Boswell, D.: Use of high-resolution NWP rainfall and river flow forecasts for advance warning of the Carlisle flood, North-West England, Meteorol. Appl., 16, 23–44, 2009.
Rossa, A., Nurmi, P., and Ebert, E.: Overview of methods for the verification of quantitative pre-cipitation forecasts, in: Precipitation: Advances in Measurement, Estimation, and Prediction, edited by: Michaelides, S., Springer-Verlag, Berlin, 419–452, 2008.
Rossa, A., Liechti, K., Zappa, M., Bruen, M., Germann, U., Haase, G., Keil, C., and Krahe, P.: The COST 731 Action: A review on uncertainty propagation in advanced hydrometeorological forecast systems, Atmos. Res., 100, 150–167, doi:10.1016/j.atmosres.2010.11.016, 2011.
Rotach, M. W., Ambrosetti, P., Appenzeller, C., Arpagaus, M., Fontannaz, L., Fundel, F., Germann, U., Hering, A., Liniger, M. A., Stoll, M., Walser, A., Ament, F., Bauer, H.-S., Behrendt, A., Wulfmeyer, V., Bouttier, F., Seity, Y., Buzzi, A., Davolio, S., Corazza, M., Denhard, M., Dorninger, M., Gorgas, T., Frick, J., Hegg, C., Zappa, M., Keil, C., Volkert, H., Marsigli, C., Montaini, A., McTaggart-Cowan, R., Mylne, K., Ranzi, R., Richard, E., Rossa, A., Santos-Muñoz, D., Schär, C., Staudinger, M., Wang, Y., and Werhahn, J.: MAP D-PHASE: Real-time demonstration of weather forecast quality in the alpine region, B. Am. Meteorol. Soc., 90, 1321–1336, doi:10.1175/2009bams2776.1, 2009.
Roux, B. and Seed, A. W.: Assessment of the accuracy of the NWP forecasts for significant rainfall events at the scales needed for hydrological prediction, Bureau of Meteorology, Melbourne, Australia, 2011.
Roux, B., Seed, A. W., and Dahni, R.: An evaluation of the possibility of correcting the bias in NWP rainfall forecasts, Bureau of Meteorology, Melbourne, Australia, 42 pp., 2012.
Roy Bhowmik, S., Joardar, D., and Hatwar, H.: Evaluation of precipitation prediction skill of IMD operational NWP system over Indian monsoon region, Meteorol. Atmos. Phys., 95, 205–221, 2007.
Skomorowski, P., Rubel, F., and Rudolf, B.: Verification of GPCP-1DD global satellite precipitation products using MAP surface observations, Phys. Chem. Earth Part B, 26, 403–409, 2001.

Smith, R. N. B.: A scheme for predicting layer clouds and their water content in a general circulation model, Q. J. Roy. Meteorol. Soc., 116, 435–460, doi:10.1002/qj.49711649210, 1990.

Stanski, H. R., Wilson, L. J., and Burrows, W. R.: Survey of common verification methods in meteorology, Research Rep. 89-5, Environment Canada, 114 pp., 1989.

Westra, S. and Sharma, A.: Australian rainfall and runoff revision project 4: continuous rainfall sequences at a point, UNSW Water Research Centre, No. P4/S1/002, 100, Sydney, Australia, 2010.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 2nd Edn., Academic Press, San Diego, California, US, 2006.

Wilson, C.: Review of current methods and tools for verification of numerical forecasts of precipitation, COST717 Working Group Report on Approaches to Verification, available at: http://www.smhi.se/hfa_coord/cost717/doc/WDF_02_200109_1.pdf (last access: May 2012), Met Office, UK, 2001.

WMO: Guidelines on performance assessment of public weather services, Geneva, Switzerland, No. WMO/TD No. 1023, 67 pp., 2000.

WWRP/WGNE: Recommendations for the verification and intercomparison of QPFs and PQPFs from operational NWP models, 37 pp., Geneva, Switzerland, 2008.

Xie, P. and Arkin, P. A.: Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions, J. Climate, 9, 840–858, 1996.