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

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

We sincerely thank the reviewer for his thorough reviews and constructive suggestions. We have carefully revised the manuscript to address his comments and suggestions. Our responses to the comments are in blue font, updated text in manuscript in red font.

Anonymous Referee #1

Thanks to the authors for presenting the interesting results of rainfall forecasting from the UM-based NWPs. One fact I like best is that the manuscript describes the research utilising a series of models at different resolutions and its focus on the hydrological use. However, while the efforts are highly appreciated, I have a number of observations that I think need to be addressed in terms of the quality and the science of the paper.

General observations:

G1. The organisation of the paper. It seems to me that the paper is too long or the message has not yet effectively delivered. I understand that the the paper tries to cover several models with a number of experiments. It is still hard to come up with a general conclusion after reading the paper. I would suggest to re-organise the paper to highlight the main points that need to be delivered.

We agree that the paper is a little bit long. We have tried to shorten the paper by skipping some details such as description of ACCESS model, sampling uncertainty in the original version of the paper. The objectives of this study are to i) compare the skills of different spatial resolution NWP models at station locations and catchment scale ii) evaluate the skill with forecast lead times, precipitation accumulation periods, and precipitation threshold values, and iii) investigate the effect of diurnal cycle and sampling uncertainty in the skill. We now explicitly mention the objectives in the introduction section. We believe that the current experiment designs follow logical order to deliver the objectives of the study. We have re-written the abstract and revised the conclusion to deliver the main findings from the study. Second paragraph of the abstract now reads

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.

G2. The hydrological aspect needs to be further strengthened, especially regarding the stream simulation. While the paper uses the catchment boundaries and the areal rainfall to evaluate the rainfall forecast (so that it can differentiate itself from other similar studies), it lacks the details when referring hydrological consequences, e.g., contribution to runoff generation.

This study is the first part of a research program 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. This study mainly focuses on evaluation of NWP model precipitation forecasts for short-term streamflow forecasting purpose. The results from this study are going to be used for streamflow forecasting study. Future work is planned to assess the benefits of using the NWP rainfall forecasts for short term streamflow forecasting. This was explicitly mentioned in the abstract and conclusions of the original version of the paper. However, it seems that the title is somehow misleading as the streamflow forecasting results are not presented. We explicitly mention this also in introduction section of the revised paper and change the title to "Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose"

G3. Some important technical details are missing, e.g., details of the ACCESS models, how the uncertainty sampling techniques are used (see specific observations below).

In the original version of the manuscript, some details are deliberately removed as the paper became a lit bit longer (more than 11K words). Considering the comments from other reviewer as well, we have added the following description of the ACCESS models in section 2.1:

ACCESS is non-hydrostatic model with the prognostic variables 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 are 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 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.

Mixing in unstable layers uses the 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.

We have added the following description of sampling techniques in section 3:

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.

Specific observations:

S1. In section 2.1, I failed to get the reference to the ACCESS models but I would imagine they are linked with each other - e.g., by supplying LBC/IC from coarse model to higher-res models. It would be good to include such description so that readers would know whether these models are running independently or not.

We have added the link http://www.bom.gov.au/australia/charts/bulletins/apob83.pdf to the reference to the ACCESS model in the revised paper. As mentioned in the earlier response, a detailed description of the ACCESS models is also given.

The following text is added in second last paragraph of section 2.1:

All models except ACCESS-G use boundary conditions that are provided by a coarser resolution models, e.g. 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.

S2. Dealing with seasonality (Line 2, page 12576). The period seems to include nearly one year. what is your consideration of the seasons and their impact. Further, during the winter period, how the snowfall is observed and how the NWP models predict the precipitation (overall or separate).

It is believed that the skill of NWP model also depends on the season. Australian 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.

During the winter period, heated rain gauge is used to measure the snow fall. The NWP model predicts the overall precipitation.

The above texts are added to the revised manuscript (last paragraph of section 5, 4th paragraph of section 2.2).

S3. Terminology. Line 14, page 12577. I would suggest to use a different name rather than RMSE to refer to your version of the standardised RMSE. Also, what is the value of the non-standardised RMSE which may make sense to see how large the error is.

Standardised RMSE is named to sRMSE. We have computed non-standardised RMSE for fig 5. For the rest of figures we stick to the sRMSE which is independent to the magnitude of the data (i.e. able to compare 3 h accumulated rainfall to 24 h one).

S4. The use of ACCESS-G model only. Section 4.2, Line 16, page 12579. You stated that the reason is that the G model has the longest lead time. I suspect that this is due to the configuration and other models should also be able to run for the same period as long as you supply them with proper LBC data. The problem is, while you already found that the G model is least useful (in terms for hydrological use), a long section is used to describe its skill.

We agree with the reviewer that it's possible to run high resolution models for the longer periods. However, it is impracticable due to heavy computation requirements. With a current Bureau of Meteorology system, it takes about 3 hour to run the high resolution model ACCESS-VT which has only 36 hour of lead time. We can imagine how much time would be required to run high resolution model for a period of 10 days.

We knew the skill of the ACCESS-G model only after evaluation of the model (this study). As mentioned in the introduction section, the purpose of evolution is also to understand the nature of forecast errors (e.g. bias, error on light versus heavy rain) which can inform the development of methods for post-processing raw forecasts to improve accuracy and reliability. In practice, raw NWP models are rarely directly used to forecast streamflow. The ACCESS-G model has a systematic bias which can be reduced using post processing methods. Currently authors are working on post processing methods to reduce systematic biases from the ACCESS NWP models. Since this model has a longer lead time (10 days), it is useful in extending streamflow forecast lead time after removing forecast bias.

S5. Line 16, page 12588, you stated "Any kind of NWP...". Could you explain why.

We have changed the text to

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.

S6. Line 20, page 12588, "The NWP models ... at their native resolutions (i.e. hourly for individual cells)...". I think it is a misunderstanding of the "native resolution", at least for temporal one. In many models, the hourly resolution is a result of writing out the state variables every hour during the integration which means that you can actually change this to 0.5 hour, or 1.5 hours.

We agree with reviewer and have changed the text to

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).

S7. Sampling uncertainty. A bootstrap procedure is mentioned (Line 7, page 12575) but I am not clear how it is implemented in this study and therefore cannot judge it is proper or not. Could you add a bit more details about the procedure.

As mentioned in the response to G3, we have added the bootstrap procedure in section 3 of the revised manuscript.

S8. The last two paragraphs of page 12589. One fact missing here is that the sampling/interpolation method used in the study may have a more direct impact than the reasons mentioned in the last paragraph. Not only do the neighbouring stations in two grids cause problem, but two stations in the same grid. Also I would expect a quite different result if the IDW method is replaced with any other interpolation.

Evaluations of the rainfall forecasts are done at both rain gauge locations (point) and subareas location (spatial). With regard to rain gauge locations, observed rainfall used in this study is the directly measured rainfall at the gauging location, so there is no interpolation. Regarding to the subareas rainfall, we agree with reviewer that interpolation method may have impact on results as the subareas rainfall is interpolated from rain gauge locations. Note that this paragraph describes only for rain gauge location.

The following texts are added in the second last paragraph of section 5 of the revised manuscript:

“Since the catchment average precipitation is interpolated from nearby stations, 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”.

References

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

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