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 #3

Summary: this manuscript presents an evaluation of four NWP models over a 5500 km2 watershed in southeastern Australia, with an intended emphasis in streamflow forecasting. Forecasts are evaluated against individual station observations, without spatial interpolation, and against averaged precipitation over the catchment area. Con-tinuous and categorical skill metrics are employed, and the influence of averaging win-dow (3h up to 24h) as well as lead-time (3h up to 228h) is assessed. The major findings presented in this paper are common to other NWP model evaluation studies elsewhere: i) over (under) estimation of low (high) precipitation amounts, ii) an overall tendency to forecast non-zero precipitation with more frequency than that observed in the field, iii) decrease in forecast skill with increasing lead times. A rather interesting result is the influence of a marked daily precipitation cycle on forecast skill, due to the synchronicity (or lack thereof) between the forecast time, averaging window, lead-time and rainfall hourly stage. It is not easy to evaluate this paper, mainly because in my opinion there is a mismatch between the methods and the intended purpose of the research. While the title indicates that the NWP model evaluation is done in the context of streamflow forecasting, no amount of work is dedicated to linking precipitation forecasts with a hydrological model in a forecasting exercise. The only defensible link would be that of averaging rainfall forecasts at a catchment scale, but this is tricky, because rele-vant catchment scales may go from the tens to the thousands of square kilometers depending on the intended application, local climate, human presence, etc. Therefore, it is not possible to say that this work is relevant for operational streamflow forecasting based on just comparing averaged observations and forecasts over the area of a spe-cific catchment. On the other hand, the methods presented here are appropriate and correctly applied for the evaluation of NWP forecasts per se, against observed precipi-tation. In this sense, as a meteorological evaluation exercise this research falls short in that the comparison area is too small, and one would hope to assess, for example, the skill of the models across different climates (regions), seasons, etc. In the latter case a subset of regions (catchments) throughout Australia could have been sampled to rep-resent the above-mentioned conditions. I would strongly suggest complementing the study in such a way but maintaining the novel aspects indicated in the introduction of this paper. Examples of similar analyses (albeit for different precipitation products) can be found, for instance, in Skomorowski, P., F. Rubel, and B. Rudolf, 2001: Verification of GPCP-1DD global satellite precipitation products using MAP surface observations. Phys. Chem. Earth, 26, 403–409. McPhee, J., S. A. Margulis, 2005: Validation and Error Characterization of the GPCP-1DD Precipitation Product over the Contiguous

C6300 United States. J. Hydrometeor, 6, 441–459. Xie, P. P., and P. A. Arkin, 1996: Analyses of global monthly precipitation using gauge observations, satellite estimates, and nu-merical model predictions. J. Climate, 9, 840–858. Joshi, M. K., Rai, A. and Pandey, A. C. (2012), Validation of TMPA and GPCP 1DD against the ground truth rain-gauge data for Indian region. Int. J. Climatol.. doi: 10.1002/joc.3612. Overall, the paper is well written but there are a few instances of cumbersome use of the English language. Please revise language and grammar thoroughly in an updated manuscript. I do not intend to point out editorial suggestions in the remainder of this review.

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"

As mentioned in the manuscript, the evaluation of the rainfall forecasts in the context of streamflow forecasting is different than that of verification done from meteorological perspective. Besides the reasons mentioned in the original version of the manuscript, we have added the following texts in the third last paragraph of the introduction section.

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 metrological verification is based on evaluating forecasts on several model grid cells at a particular forecast time and does spatial aggregating rather than temporal aggregating. For example, in meteorological verification, hit rates are often calculated by counting the number of grid cells of correct forecasts at a particular time compared to counting the 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 do not contribute to the streamflow forecasts to the catchment of interest.

We have evaluated the NWP rainfall forecasts for a medium size catchment. We agree that from meteorological perspective, the comparison area is small. However as mentioned before our focus is to evaluate in the context of streamflow forecasting. We have also mentioned in the discussion that we are extending this work to other catchments experiencing a range of climatic conditions in Australia.

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. We have added above texts in the 4th paragraph of section 5.

We checked the language and grammar in the revised manuscript (two co-authors are native English speakers).

Specific comments Page Line Comment 12567 18 No hydrological perspective is provided in this work, please delete this statement or modify substantially the manuscript in line with my general comments provided above.

We agree that the hydrological results are not provided in this manuscript. As mentioned before, the main focus of this manuscript is to evaluate the NWP rainfall forecasts on scale relevant to hydrology (i.e. hydrological perspective). Hydrological parts are planning to publish in a subsequent paper.

12571 1 The manuscript mentions a hydrologic model, and shows a figure with subcatchments indicating that their size is roughly similar to the 12-km NWP model resolution. Please provide more information on the hydrological application if a streamflow forecasting focus is to be adopted for this manuscript. Maybe I misunderstood, but later the authors aggregate to the catchment scale arguing that lumped models are used for forecasting. If this is so, what's the point in focusing on subcatchments? In hydrological (flood) applications, usually the relevant scales are neither the point (station) nor the catchment scale, but something in between hillslope and subcatchment scale.

As said before, streamflow forecasting results will be presented in a subsequent paper.

Regarding the subcatchment scale, this comment is similar to one of the comment from the reviewer # 2. We have added the following texts to the first paragraph of section 4.6:

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. Australian Bureau of Meteorology 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. Bureau is planning to use continuous modelling with semi distributed lumped model (connected lumped model) for real time flood forecasting services in Australia.

12572 17 I think this statement is somewhat confusing/misleading. It is true that precipitation observations are used in construction, calibration and validation of hydrologic models, but it is also true that observations are routinely interpolated in some fashion in order to achieve the spatial coverage required by the hydrologic model. In turn, this is also a function of the spatial discretization used in the hydrologic model. A direct com-parison between NWP forecasts and station data may be appropriate, but probably this has little to do with the fact that the NWP is intended for hydrologic applications.

The first step of the evaluation is to compare with the direct measurement of the observations if available. This will limit the influence of artefacts resulting from missing data that are introduced by the interpolation techniques currently in operational use. We believe that this is also relevant as these are the stations which are directly used to interpolate the sub-catchment rainfall used in this study.

12573 24 I'm curious about the concept of "event" used by the authors for evaluating the NWP skill. Is an event just an instance of non-zero precipitation at any given time window, or are the authors using the word in a more meteorological sense? In other words, an "event" could consist of a series of precipitation intervals, related meteorologically among each other.

The "event" in this paper relates to a single event. In categorical score, "yes event" is the single event when the rainfall exceeds the given threshold value at particular time. The first few sentences of this paragraph now read

From user point of view it is also important to know whether rain 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.

12575 5 A more thorough description of the resampling technique used for estimating uncertainty bounds would be welcome from a reader's perspective.

Agree. This was also pointed out by the reviewer # 1 and added in the second last paragraph of section 3 of the revised paper.

12575 24 This section is confusing in that it is not clear what is being compared here. On one hand there are rainfall observations from a past period, and on the other we have modeled estimates. But the authors refer to "forecasts". . . which forecasts are they talking about? The models can issue forecasts of hourly precip that here are being accumulated at 24h intervals, but the models issue forecasts for many lead times. . . so it there may be a few different forecasts by the same model for the same 24h interval.

24 h (daily) rainfall forecast issued on a particular day and time is the accumulated forecast rainfall of lead time from 1 to 24h on that day and time. Since we have used forecasts which are issued one time a day, there will be only one set of forecasts of 1-24 hour lead time on a given particular date and time.

Mathematically, accumulated 24h forecast issued on date d at 09:00 is
Fd_09 = fd_09_10 + fd_10_11 +…. + fd'_08_09.
Where, fd_09_10 = forecast issued on date d at 09:00 for a period 09:00 to 10:00 h (lead time 1h)
fd_10_11 = forecast issued on date d at 09:00 for a period 10:00 to 11:00 h (lead time 2h)
fd'_08_09 = forecast issued on date d at 09:00 for a period 08:00 to 09:00 h next day (lead time 24h).

The paragraph is revised to clarify this (note that the title of this paragraph explicitly mentioned that it is 1-24 hour lead time). The first few sentences of the first paragraph of section 4.1 now read

Fig. 4 shows a map of 24 hours mean precipitation accumulation for the measurement stations and for the ACCESS model grid cells over the Ovens catchment. 24 hours (daily) precipitation on a given date and time (09:00 LT) is the accumulated forecast precipitation of lead times from 1 to 24 hours on that date and time. The precipitation is averaged over a period for 1 April 2010 to 8 February 2011.

12577 25 The authors state that RMSE does not show a spatial pattern by observing the plot of RMSE values against latitude then longitude. However, elevation may be a factor affecting model skill (an orographic effect does exist), and this variable may have no direct relation with lat-lon. This effect is observed later for the bias statistic. Please comment.

We agree that forecasts are not directly related with latitude and longitude, nor is their skill. Since the southern parts of the catchment are high elevation areas, forecasts and their skill are apparently related to latitude as well. In order to generalise, we now only mention "altitude". This study shows that the altitude influences the bias score more than the other scores.

We have changed the text to

In general, the RMSE score does not exhibit any strong spatial pattern with respect to the altitude of the stations.

12580 3 It is not clear what this sentence means. Correlation coefficient seems to decrease with lead time, whereas RMSE seems to increase and bias shows a varying behavior based on the time of the day.

This sentence was formulated in a context of the model skill. We agree with the reviewer that RMSE value increases with lead time. We can also say that RMSE skill score decreases with lead time.

We have changed text to

One can see that the model skill with respect to correlation coefficient decreases with lead time which is not obvious in RMSE and bias skills mentioned before.

12580 15 Please revise the use of the word "unnecessarily".

Done.

12585 29 This sentence seems gratuitous, given that it mentions for the first time "syn-thetic data" without providing any details about the way it is obtained. Please consider deleting or enhancing this discussion.

The following texts are added to the last paragraph of section 4.5 of the revised paper:

Synthetic data were generated to understand the diurnal cycle in the forecast skill. About 20 years of daily rainfall data from one of the station were disaggregated to hourly using sine curve of one cycle period. The hourly forecast values are generated disaggregating uniformly from daily rainfall values by adding some random noise. RMSE, bias and correlation coefficients scores are 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 RMSE score or correlation coefficient.

12586 7 Do you mean this in general or for the intended application by BoM? Hydrolog-icval models with all kinds of spatial discretizations are used in forecasting application.

We are referring to the lumped model currently employed by BoM for real time streamflow forecasting, where catchment average rainfall is used for the forecast period. Please see also response to earlier comment on page 3.

12586 9 This sentence needs some support, either by evidence shown in the manuscript or by a proper literature review.

Since we are not presenting streamflow forecasting results, we removed this sentence.

12589 17 I'm troubled by this paragraph. Is this result an absolute coincidence? Or is it to be expected? Not only this is only one catchment and only one year of data is analyzed, but no experiments regarding different hydrological model discretizations where performed, so no conclusion should be drawn at all from this result.

We removed this paragraph.

References

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, 10.1016/j.jhydrol.2004.02.004, 2004.
Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275-289, 2003.