Response to reviews on the manuscript hess-2018-607 "An evaluation of daily precipitation from atmospheric reanalyses over Australia" by Suwash Chandra Acharya et al.

We would like to thank anonymous referee (R1) and Korbinian Breinl (R2) for their constructive comments and suggestions on our paper. These comments have greatly helped us identify where we need to improve our description of the overall context and framing of the research. There are also several specific points which we could address to improve the way we have provided and discussed the results. We comment first on a general point regarding the purpose of the work raised by both reviewers, and this is followed by responses to more specific issues raised by the individual reviewers.

Comments made by both reviewers

Both reviewers query why we examine the daily performance of BARRA when it is an hourly product, and both reviewers query whether it would be useful to include comparisons with other reanalysis data sets.

We choose to focus our evaluation on daily rainfalls as this time step affords the best means of comparison with gridded data (AWAP) that is well grounded in gauged observations. AWAP is a high-resolution gridded dataset representative of areal rainfalls which is widely accepted as being the best synthesis of gauged daily observations at the 5-25 km resolution (there is no equivalent available dataset for sub-daily rainfalls). We see this as a necessary first step towards examining sub-daily behaviour because any reliance on sub-daily estimates necessarily depends on its ability to correctly represent daily rainfalls. This will inform our future work of conducting sub-daily evaluation, which requires more sophisticated methods and yields more complicated results.

We do include comparison with daily gauged (point) rainfalls and with gridded ERA-interim. We consider daily gauged point rainfalls as these represent the base data on which the AWAP estimates are derived, though as discussed in the paper we would expect there to be differences between point and gridded estimates as the latter data set accounts for some spatial averaging. Our rationale for including ERA-Interim is that it is has been found to be the best performing data set compared to other reanalysis rainfall products in the Australian region (Peña-Arancibia et al., 2013), and it shares similarity with BARRA in that rainfall observations are not assimilated. Importantly, ERA-Interim also provides the boundary condition for BARRA-R simulations and this allows the incremental value of BARRA-R to be assessed. Our focus with these comparisons is solely on the efficacy of the BARRA-R rainfall product, we do not set out to instruct users on the selection of wide-ranging precipitation products that are available across Australia.

Lastly, one additional reason for considering the performance of BARRA-R at a daily time step is that it has the potential to provide more accurate – or at least a credible alternative – estimate of daily rainfalls in regions which are sparsely gauged. There are large areas of Australia with very sparse gauging (see left-hand panel in the figure below), and it may be useful to supplement estimates of rainfalls in these regions with reanalysis products that do not rely on rainfall observations for assimilation. Without observations to inform surface fitting, it is evident that AWAP can yield erroneous results (see right hand panel in the figure below). One potential advantage of BARRA thus lies in its ability to complement the existing datasets in regions where gauged observations are sparse (especially in the semi-arid regions). In this regard, we note that only AWAP grid points that contain rainfall gauges are selected for comparisons in this study.

In the following sections we provide a detailed response to all the remarks raised by the referees. We had posted a short comment earlier to clarify the general concern raised by R1, and we now proceed to address all the comments by listing the reviewers' comments (RC, in blue), our corresponding reply (AR, in black), and proposed modifications in italics.
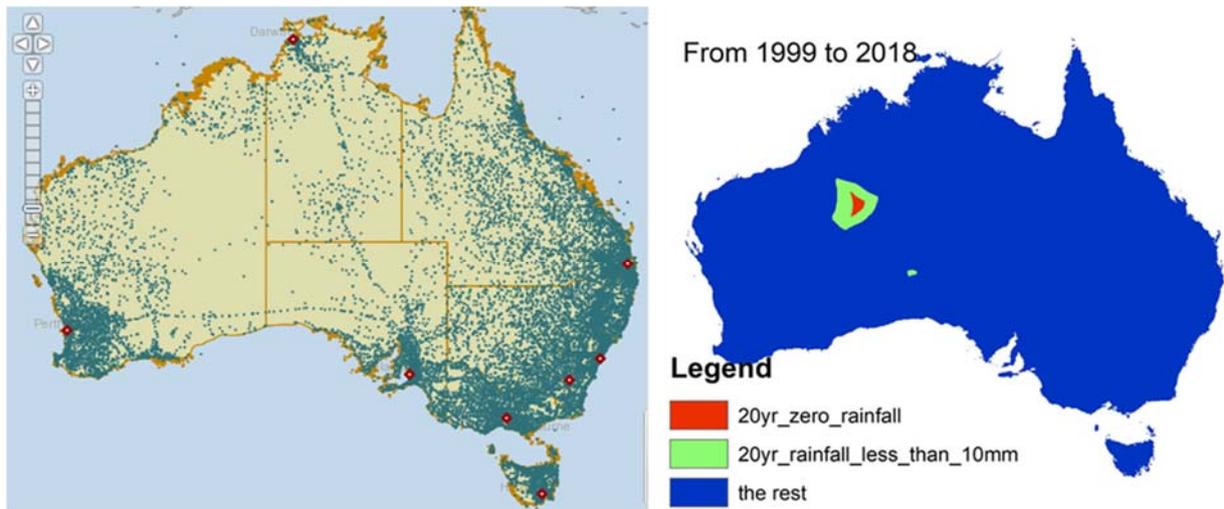
Figure showing location of all daily rainfall stations in Australia (left panel) and AWAP estimates that erroneously indicate zero and less than 10mm total rainfalls over a twenty year period (right panel). The left panel is sourced from the Australian Bureau of Meteorology (http://www.bom.gov.au/climate/data).

Response to Referee #1

Response to general comment

[RC] The manuscript "An evaluation of daily precipitation from atmospheric reanalyses over Australia" aims at comparing the new reanalysis precipitation dataset BARRA with ERA-Interim over Australia using in-situ rainfall data (point-to-grid analysis) and AWAP dataset (grid-to-grid analysis) as benchmarks. I do believe that the paper reads very well, it is properly structured and addresses a relevant topic of uttermost importance. The authors showed that the new dataset BARRA tends to outperform in most of the case ERA-Interim, while provides lower performances when compared to the AWAP dataset. In my opinion, I found the comparison of BARRA with only 1 reanalysis dataset not enough to justify a possible publication in HESS.

[AR] We thank Referee #1 for acknowledging the relevance and importance of the topic discussed.

Responses to specific comments

1.1 The purpose of the current study is to document the performance of the BARRA dataset at a daily scale and to provide a comparative analysis of its strengths and limitations relative to other available datasets. However, only ERA-Interim is used as a comparison. Why did the authors decide to use only 1 dataset for comparison? Why the choice of using another reanalysis dataset (ERA-Interim) and not other based purely on satellite product (e.g. PERSIANN) or corrected satellite (e.g. PERSIANNCDR)? I believe the manuscript (and the comparison) will benefit with the inclusion of additional recent and well-known datasets (e.g. CHIRPS, MSWEPv2.1, SM2RAIN ASCAT, CMORPH-CRT), or other reanalysis datasets (e.g. JRA-55, NCEP-CFSR, PFD, or WFEDEI GPCC) for comparison. Obviously, I am not suggesting to include several datasets in this analysis, but the comparison with 3 or 4 more datasets will definitely strengthen the impact of this research and manuscript.

[AR] As mentioned above, our aim is to evaluate the performance of BARRA dataset rather than provide instructive comments on use of wide-ranging precipitation products. We will provide additional clarification on these points in a revised version of the Introduction. Further, we propose to change the title of the paper to "*An evaluation of daily precipitation from a regional atmospheric reanalysis over Australia*" to better reflect the focus of our paper.

1.2 The authors first mentioned that "The accuracy at a daily scale provides us with an important benchmark as it is applicable to many hydrological applications and also forms the basis for further examination at finer timescales". However, the author then contradicted themselves concluding that "The core attraction of the BARRA dataset is the availability of sub-daily precipitation estimates. Such information is not available in the AWAP data set, and the spatial resolution of the estimates is higher than the currently available global 20 reanalysis and satellite datasets". In fact, in hydrological application at large scale (which is the case for the BARRA dataset due to a spatial resolution of 36km) daily time scale is most used temporal resolution. For this reason, as end-user, I would select the AWAP dataset as input for a large scale model as the resolution is higher and more appropriate to represent complex topographies. Beside the scientific interest in comparing different precipitation datasets, why someone should use BARRA if AWAP is already providing excellent performances at higher spatial resolution?

[AR] The choice of daily scale and its utility is addressed in the general response above. We have revised the introduction section of the manuscript (Page 4 Lines21-25), as follows:

Despite the availability of BARRA precipitation as hourly values, we select a daily timescale to match the temporal resolution of the best available gridded reference dataset (AWAP). In addition, the accuracy at a daily scale serves as a desirable first step towards further examination at finer timescales because any reliance on sub-daily estimates necessarily depends on its ability to correctly represent daily precipitation. The additional value of evaluation at a daily scale is that it assesses the potential of BARRA to provide estimates of daily rainfall in the sparsely gauged regions across Australia.

1.3 Results and discussions of grid-to-grid analysis are very brief and conclusions are somehow similar to the point-to-grid analysis in which BARRA gives better results than ERA: What is the additional value of including such analysis? It would be better to include more dataset for comparison (see the first point) and run only point-to-grid analysis.

[AR] We do agree that the results of grid-to-grid analysis are brief. However, the performance based on frequency metrics (wet day frequency, dry-wet transition probability, and wet-wet transition probability) are similar for both grid-to-grid and point-to-grid analysis and leads to similar conclusions. It is thus appropriate to present the similar results in a brief manner. However, KGE' and its components at grid-to-grid evaluation shows that BARRA exhibits superior performance at grid scale (5 km) which is not obvious from point-to-grid analysis. Given the spatial averaging that is implicit in grid-based products it is not expected that similar results would be obtained with point-to-grid comparisons, or with grid-to-grid comparisons at different resolutions (AWAP is 5 km, BARRA-R is 12km, and ERA-Interim is 80km). This information is reflected in the revised introduction (Page 4 Lines11-16) and methods section (Page 6 Lines 13-21):

The evaluation is performed against gauged measurements from two data sources: one is based on gauged point observations, and the other is a high-resolution gridded AWAP dataset derived from interpolating gauge measurements, which is widely accepted as being the best synthesis of gauged observations. There are inherent differences in gridded and point rainfall estimates due to the spatial averaging of point observations across each grid cell area. Since BARRA provides direct estimates of area-average rainfall, it is useful to compare these estimates with the best available point and areal precipitation estimates.

…

As the gridded datasets represent an areal average, it may be expected that there are differences between point and gridded estimates as the latter account for some spatial averaging. While the gauged data and AWAP rainfall estimates represent the best available reference datasets based on measured data, both are imperfect representations of areal rainfalls. The AWAP estimates contain inaccuracies due to the interpolation method, and the point estimates provide only a coarse estimate of rainfall over a grid cell area. The ability of these point and gridded reference data sets to represent actual areal rainfalls is heavily dependent on the gauging density and local orography, and these factors influence

3

the accuracy of the reference data sets to different degrees across Australia. Accordingly, we compare the BARRA and ERA-Interim estimates to both point gauged and AWAP areal data, where the evaluation of both offer different insights to the quality of model estimates.

1.4 How the different spatial resolution of the reanalysis dataset and interpolation method of the in-situ gauges can affect the results of this analysis?

[AR] The effect of spatial resolution is clearly observed in the comparison of quantiles: coarser datasets are not well suited to representing large rainfall due to averaging over larger area. In addition, the choice of interpolation could also affect the analysis of large rainfalls. In this study, nearest neighbour interpolation is used which preserves the magnitude of precipitation (averaged) over the grid, however, bilinear interpolation is likely to smooth the precipitation field resulting in higher bias for larger rainfall (Accadia et al., 2003).

We have added the reasoning for our choice of interpolation methods in section 3.1 (Page6 Lines 22-26):

Gridded rainfalls are compared to point rainfalls using a nearest neighbour approach. The choice of interpolation scheme is especially important when comparing datasets of different spatial resolutions; bilinear interpolation is likely to smooth the precipitation field resulting in higher bias for larger rainfall (Accadia et al., 2003), whereas nearest neighbour method preserves the magnitude of the precipitation over the grid.

1.5 The methodology is quite straightforward and based on existing approaching for comparing distributed precipitation dataset. Besides the comparison of different datasets over Australia using different performance measures which one is the main research innovation of this paper?

[AR] We agree that the methodology in this study is straightforward and based on existing approaches for evaluation of precipitation dataset. However, the precipitation dataset used in this study is new and is the only regional reanalysis product developed for the Australasian region. The results of this study are useful for the potential users of the dataset for hydrometeorological studies in the Australasian region.

1.6 From Figure 2.d it is difficult to assess where BARRA is performing (on average) better than ERA-Interim. From my point of view, ERA-Interim shows overall higher KGE values than BARRA (blue points). I suggest the authors to estimate the average (and standard deviation) of the values in figure 2.d to see which dataset provides higher KGE.

[AR] The summary of difference in KGE has now been added in the revised plot (Figure 2d)

1.7 Lines 18-20 page 7 "Despite having a slightly lower correlation compared to ERAInterim, the variability of the rainfall is better captured by the BARRA dataset" I do not agree with the authors. From figure 3 it looks that ERA tends to outperform BARRA in almost all the considered performance measures. Also, how the authors can say that rainfall is better captured by BARRA dataset if an aggregated index (KGE) is used?

[AR] We checked the plot and the corresponding statement. The best value for variability ratio is 1 and the variability ratio in BARRA is closer to 1 compared to ERA-Interim in overall as well as seasonal analyses. We have added the lines representing best values in KGE' as well as its components to add interpretation of the plot (Figure 3).

[AR] Based on figure 4, there is slight difference in magnitude of the frequencies across datasets. We agree that BARRA and ERA-Interim are very close to each other. But, in terms of spatial pattern, all three datasets behave similarly as observed in spatial plot shown below.
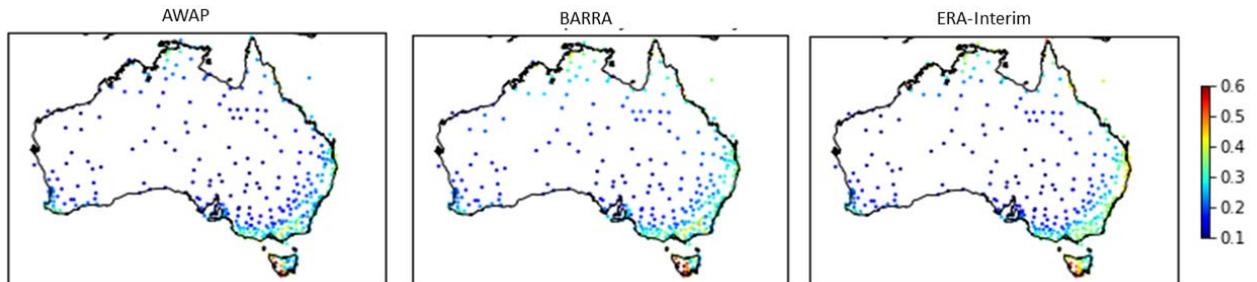

Figure showing frequency of wet days for AWAP, BARRA, and ERA-Interim at gauge locations used in this study.

Response to Referee #2 Korbinian Breinl

Response to general comment

[RC] The present article deals with the evaluation of a new reanalysis dataset called BARRA with other gridded datasets (ERA-Interim and AWAP) and rain gauge observations, for Australia. This is a well written paper, which easy to read. At this stage I think however that some improvements are needed for the publication in HESS.

[AR] Thank you for your comments.

Responses to specific comments

2.1 I wonder if the results of the mean precipitation (Figure 1 and related text in the results section) could be presented in a different way – besides the four maps that are useful for sure – to better capture the changes in spatial variability. Maybe the authors have a good idea for a plot.

[AR] The four maps in Figure 1 visually demonstrate the spatial variability in precipitation across datasets. In addition, we have now computed a correlation measure between BARRA-R and AWAP, and similarly between ERA-Interim and AWAP. We have added the plot in the supplement (Figure S1).

Also, I wonder if cutting off the over-sea precipitation for BARRA and ERA-Interim would help to better read the map, meaning taking the over-land precipitation as the lowest common denominator.

[AR] Our original logic for retaining this was two-fold: to visualize spatial richness in the datasets, and to show that BARRA extends across ocean surrounding Australia which is not the case for gauged measurements and AWAP. However, we are happy to mask the information shown on these plots so they cover the same physical domain. The subplots in the revised plot (Figure 1) now represents precipitation over land for all the datasets.

2.2 I think it would be good to look into spatial correlations. As far as I can see you have not looked into them, although they are relevant. Is there a particular reason for not taking them into account?

[AR] The correlation coefficient was computed for each station and their spatial distribution was visualised. It was not shown in the manuscript as the spatial distribution of correlation coefficient was very close to spatial distribution of KGE'. We have not looked further into spatial correlations, though as responded in 2.1, the added correlation plot (Figure S1) will demonstrate the correlation structure across space.

2.3 I do not fully understand the purpose of the BARRA dataset, at least not in the context of the article, which is focusing on daily values. Considering that the AWAP dataset is superior to BARRA (superior at daily timescale), why would I use BARRA (when sub-daily is not the topic)?

[AR] We address the choice of daily scale and its utility in the general response above. In addition, we have made modifications to the introduction section to clarify the choice of daily timescale to our readers (Page 4 Lines 21-25):

Despite the availability of BARRA precipitation as hourly values, we select a daily timescale to match the temporal resolution of the best available gridded reference dataset (AWAP). In addition, the accuracy at a daily scale serves as a desirable first step towards further examination at finer timescales because any reliance on sub-daily estimates necessarily depends on its ability to correctly represent daily precipitation. The additional value of evaluation at a daily scale is that it assesses the potential of BARRA to provide estimates of daily rainfall in the sparsely gauged regions across Australia.

[AR] As noted earlier, BARRA represent area-average rainfall: while gauged data represents the best estimate of rainfall at a point location, AWAP serves as the best available estimate of areal rainfall. We think it is useful to undertake both grid-to-point and grid-to-grid comparisons as both types of estimates are likely to contain biases due to differences in spatial averaging, and the levels of measurement support also differs between the different products. The two analyses thus offer different insights to the quality of the model estimates. It is worth noting that in terms of performance, KGE' and its components at grid-to-grid evaluation shows that BARRA has superior performance at grid scale (5 km) which is not obvious from point-to-grid analysis.

This is further clarified in the revised introduction (Page 4 Lines11-16) and methods section (Page 6 Lines 13-21):

The evaluation is performed against gauged measurements from two data sources: one is based on gauged point observations, and the other is a high-resolution gridded AWAP dataset derived from interpolating gauge measurements, which is widely accepted as being the best synthesis of gauged observations. There are inherent differences in gridded and point rainfall estimates due to the spatial averaging of point observations across each grid cell area. Since BARRA provides direct estimates of area-average rainfall, it is useful to compare these estimates with the best available point and areal precipitation estimates.

...

As the gridded datasets represent an areal average, it may be expected that there are differences between point and gridded estimates as the latter account for some spatial averaging. While the gauged data and AWAP rainfall estimates represent the best available reference datasets based on measured data, both are imperfect representations of areal rainfalls. The AWAP estimates contain inaccuracies due to the interpolation method, and the point estimates provide only a coarse estimate of rainfall over a grid cell area. The ability of these point and gridded reference data sets to represent actual areal rainfalls is heavily dependent on the gauging density and local orography, and these factors influence the accuracy of the reference data sets to different degrees across Australia. Accordingly, we compare the BARRA and ERA-Interim estimates to both point gauged and AWAP areal data, where the evaluation of both offer different insights to the quality of model estimates.

2.5 In that context (point 4), it would be good to see what has already been done in terms of evaluations of gridded rainfall vs. point rainfall (and also grid rainfall to other grid rainfall), as conducted in this study, means I would like to see a more comprehensive literature review. A brief review of such evaluations helps the reader to better understand the implications of the present study. For example, what are the pros and cons of ERA-Interim according to other studies, and what has been concluded in this article? Again, I miss a bit the ability to generalize from the results from this article.

[AR] We have added further details of relevant literature that contain grid-to-grid and point-to-grid evaluations. We have included reference to the studies that document the performance of ERA-Interim (in global and Australian region) and the Unified Model on which the BARRA-R estimates are derived.

The revised text in the introduction section (Page 2 Lines 16-31, Page 3 Lines 1-9) now reads:

Global reanalysis datasets have been evaluated for various applications at global and regional scales. Of the global dataset exclusively based on reanalysis, ERA-Interim is generally considered to provide better performance compared to other reanalysis datasets (Beck et al., 2017, 2019). ERA-Interim has been found to reproduce the climatology of global monsoon precipitation (Lin et al., 2014) and in general it demonstrates high temporal and spatial correlations with interpolated observations (Donat et al., 2014). In a recent global evaluation of gridded precipitation datasets using gauge observations by Beck et al (2017), the ERA-Interim and JRA-55 reanalysis datasets were found to reproduce long-

term trends and temporal correlation more reliably than achieved by satellite datasets. In the Australian continent, ERA-Interim reproduced the observed spatial patterns of long-term rainfall along with other climatic variables and showed an overall better performance compared to NCEP-NCAR (National Centers for Environmental Prediction/National Center for Atmospheric Research) reanalysis (Fu et al., 2016). An evaluation by Peña-Arancibia et al. (2013) of reanalysis datasets, satellite products, and an ensemble of these datasets in Australian and Asian regions showed that the ERA-Interim performed better than other individual datasets across a range of metrics for the Australian region.

The available global reanalysis datasets such as NCEP-CFSR, ERA-Interim, and JRA-55 cover the Australian region, but their horizontal resolutions are relatively coarse (≥ 80 km) and unsuitable for fine-scale application in hydro-meteorological analysis. The resolution of the global reanalysis can be enhanced by downscaling approaches such as dynamic downscaling (Soares et al., 2012) or statistical analysis using high-resolution surface observations (Vidal et al., 2010). Alternatively, the application of a regional model-based data assimilation can provide a better representation of local climate features and extreme events (Bollmeyer et al., 2015). The regional reanalysis, unlike global reanalysis, allows the integration of abundant local surface observations at a finer scale (Bollmeyer et al., 2015; Isotta et al., 2015; Jakob et al., 2017). The studies have also been conducted to evaluate the additional benefit of high-resolution datasets obtained from regional reanalysis. Jermey and Renshaw (2016) found that the overall performance of 12km regional reanalysis was significantly better than 80km ERA-Interim especially for high thresholds of rainfall. Similarly, the performance of a regional analysis against ERA-Interim at grid scale showed an improvement in representing high-threshold events (Isotta et al., 2015; Jermey and Renshaw, 2016; Roberts and Lean, 2008) and convective events (Roberts and Lean, 2008). In general, regional reanalyses based on boundary conditions from ERA-Interim is shown to improve the performance over ERA-Interim albeit these studies are largely focused in Europe.

2.6 I would appreciate a final overview that summarizes the results, ideally in the format of a table. The table could contain for example (i) general information for each dataset (time period, spatial resolution, URL to obtain data), (ii) metrics for evaluation, (iii) performance of each metric etc. The table would make it much easier to get a good overview of the results. And, even better, another column could add some information on each dataset or similar datasets (similar in terms of how they were processed) and some results from other studies if applicable (see comment above).

[AR] We have added a table with information on datasets and another one with summary of performance in the revised manuscript. The added tables are:

Table 1 Overview of gridded precipitation datasets used in this study

Name	Details	Data Source	Spatial coverage and resolution	Temporal coverage and resolution	Reference
BARRA-R	*Bureau of Meteorology Atmospheric high-resolution Regional Reanalysis for Australia* *(http://www.bom.gov.au/research/projects/reanalysis/)*	*Regional reanalysis*	*Australasian region (65 to 196.9° E, -65 to 19.4° N), ~12km*	*1990-February 2019, Hourly*	*(Su et al., 2019)*
ERA-Interim	*European Centre for Medium-range Weather Forecasts ReAnalysis Interim*	*Global reanalysis*		*1979-present,*	*(Dee et al., 2011)*

			Global, ~80km	3- hourly	
AWAP	Australian Water Availability Project (http://www.csiro.au/awap/)	Gauge interpolated	Australia land area, 5km	1900 – present, Daily	(Jones et al., 2009)

Table 2 Summary of performance of BARRA precipitation

Metric	Point-to-grid	Grid-to-grid
KGE' and its components	Performance varies spatially, showing better scores in temperate regions than in the tropical and arid region Variability is well captured	Exhibits less bias in total rainfall than ERA-Interim Overestimates variability in the tropics only
Quantiles	High quantiles (90, 95 and 99%) closely represent high rainfalls and are unbiased.	Similar pattern as for point-to-grid Better representation of higher quantiles than ERA-Interim
Wet day frequency and transition probabilities	Frequency of wet days and wet-wet transition probabilities are over-estimated Dry-wet transition probabilities are well reproduced	Similar pattern as for point-to-grid with improved correlation and reduced bias.
Categorical metrics	Both POD and FAR increase with threshold Frequency bias is on average higher for light/moderate than low and heavy rainfalls.	Improvement along all metrics with respect to ERA-Interim Light rainfalls are over-estimated Large rainfalls are poorly captured yet are better than ERA-Interim

2.7 Reviewer #1 asked for innovation. This is not my main concern as long as I can draw general conclusions from this paper as a reader for my own studies (also outside Australia), but at this stage I miss it a bit. The article tries to explain the results to a certain degree, but I think the discussion should be more detailed, and pros and cons of each dataset should be better explained considering how the datasets were generated, considering seasons, resolution etc. I find statements such as "AWAP estimates of point rainfall are higher than the gauged observations" or "ERA-Interim generally performs better in the central arid region, whereas BARRA exhibits better scores in the temperate region", but I miss good explanations why. The table overview I addressed may help to make this generalization easier. Also, I must admit, looking into more reanalysis data (ideally popular datasets) as suggested by Reviewer #1 may make the paper stronger – at least one more prominent dataset maybe?

[AR] Thank you for acknowledging the relevance of the study. We agree that significance of this paper is enhanced by its ability to provide general conclusions useful for other studies, different data sets, or study areas. In that regard, we have revisited the datasets and discussion sections and explained the physical basis of the results further in order to aid generalization of the research beyond current study area As mentioned in the general response, our focus is on evaluating the new BARRA-R reanalysis dataset rather than providing instruction to users on selection of available precipitation datasets.

Adoption of a changed title (as discussed in response to 1.1 above) should better reflect the intent of the paper.

We have added a discussion of BARRA (Page 5 Line 24 – Page 6 Line 2):

The model includes a comprehensive set of parametrisations, including a modified boundary layer scheme, mixed phase cloud microphysics, a mass flux convection scheme, and a radiation scheme. The model parametrisation in BARRA mainly is inherited from the UKMO Global Atmosphere (GA) 6.0 configurations as described in Walters et al. (2017). Surface and satellite rainfall observations are not assimilated, and the precipitation fields are determined using the assimilated large-scale variables and the physical parameterisation of the model. At 12 km horizontal resolution, BARRA requires the convection scheme to model sub-grid scale convection using an ensemble of cumulus clouds as a single entraining-detraining plume (Clark et al., 2016). The scheme prevents unstable growth of cloud structures on the grid and explicit vertical circulations and can only predict an area-average rainfall instead of a spectrum of rainfall rates.

We have also added the following in the Results section:

(Page 8 Lines 21-26) ERA-Interim generally performs better in the central arid region, whereas BARRA exhibits better scores in the temperate region. The central arid region experiences very few rainy days which is likely to be missed spatially by a sparse network of gauges. This could result in poor correlation metrics as it is sensitive to outliers, and subsequently yields a poor KGE' score. BARRA precipitation is more erroneous in the tropics than in temperate regions which may reflect the limitations of the convection parameterisation adopted in BARRA. This spatially varying performance of BARRA is further discussed in section 5.5).

(Page 9 Lines 2-5) The AWAP estimates of point rainfall are higher than the gauged observations, and a similar degree of overestimation is exhibited by the BARRA dataset. The Barnes interpolation scheme used in AWAP slightly inflates the spatial coverage of light intensity rainfall (Jones et al., 2009; King et al., 2013). In addition, the weight function assigned to gauged point rainfalls may result in wetter bias in the densely gauged (coastal) regions compared to other regions.

(Page 9 Lines 8-10) The performance during winter is better than in summer for both reanalysis datasets. This is likely due to the ability of NWP models to accurately simulate synoptic systems which represent the majority of wintertime rainfall.

(Page 10 Lines 2-4) The model estimates of transition probability p_{11} are greater than p_{01}. Due to the tendency of NWP models to yield more frequent persistent light rainfalls, a high wet-wet frequency is observed.

We have added the following in the Discussion section.

(Page 13 Lines 28-30) The spread in transition probabilities p_{01} is estimated well by all gridded datasets, whereas the p_{11} estimates are overestimated and more varied. As mentioned above, this difference is due to the tendency of NWP models to over-estimate the persistence of light rainfalls (Kendon et al., 2012).

(Page 14 Line 25 – Page 15 Line 1) The possible explanation for this may be the difference in the climatic systems driving the precipitation in those regions and the scheme used to generate precipitation (Su et al., 2019). Convective precipitation is dominant in tropical regions and the parameterisation scheme for sub-grid convection adopted in BARRA is limited in terms of resolving such precipitation. Therefore, a higher accuracy of BARRA can be expected at high latitudes where synoptic rainfall dominates than in low latitudes where convective rainfall is dominant (Ebert et al., 2007; Su et al., 2019). The limitation of NWP models to represent convective precipitation accurately has also been reported by Ebert et al. (2007) and de Leeuw et al. (2015).

10

2.8 Figure 6. I would add x-axis labels to the upper plots

[AR] Agreed, we have added the x-axis labels in the revised plot.

2.9 Optional: The code (R, MATLAB etc.) to analyse the rainfall data could be published with the paper. I always encourage to do so, but this is up to the authors of course.

[AR] We will identify the codes that we consider most relevant for general use. We will upload the codes as a public GitHub repository and add the corresponding link in the manuscript under the sub-section "Data Availability".

Other Changes not mentioned above

1. The Table 1 in the first manuscript which summarizes the performance metrics used in the study is removed from the main text and added as a supplement.

References

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

Isotta, F. A., Vogel, R. and Frei, C.: Evaluation of European regional reanalyses and downscalings for precipitation in the Alpine region, Meteorol. Zeitschrift, 24(1), 15–37, doi:10.1127/metz/2014/0584, 2014.

Peña-Arancibia, J. L., van Dijk, A. I. J. M., Renzullo, L. J. and Mulligan, M.: Evaluation of Precipitation Estimation Accuracy in Reanalyses, Satellite Products, and an Ensemble Method for Regions in Australia and South and East Asia, J. Hydrometeorol., 14(4), 1323–1333, doi:10.1175/JHM-D-12-0132.1, 2013

An evaluation of daily precipitation from <u>a regional</u> atmospheric ~~reanalyses~~<u>reanalysis</u> over Australia

Suwash Chandra Acharya[1], Rory Nathan[1], Quan J Wang[1], Chun-Hsu Su[2], Nathan Eizenberg[2]

[1] Department of Infrastructure Engineering, The University of Melbourne, Melbourne, Australia
5   [2] Bureau of Meteorology, Melbourne, Australia

Correspondence to: S.C. Acharya (suwasha@student.unimelb.edu.au)

Abstract.

An accurate representation of spatio-temporal characteristics of precipitation fields is fundamental for many hydro-
10   meteorological analyses but is often limited by the paucity of gauges. Reanalysis models provide systematic methods of representing atmospheric processes to produce datasets of spatio-temporal precipitation estimates. The precipitation from the reanalysis datasets should, however, be evaluated thoroughly before use because it is inferred from physical parameterization. In this paper, we evaluated the precipitation dataset from the Bureau of Meteorology Atmospheric high-resolution Regional Reanalysis for Australia (BARRA) and compared it against (a) gauged point observations, (b) an interpolated gridded dataset
15   based on gauged point observations (AWAP), and (c) a global reanalysis dataset (ERA-Interim). We utilized a range of evaluation metrics such as continuous metrics (correlation, bias, variability, modified Kling-Gupta efficiency), categorical metrics, and other statistics (wet day frequency, transition probabilities and quantiles) to ascertain the quality of the dataset. BARRA, in comparison with ERA-Interim, shows a better representation of rainfall of larger magnitude at both point and grid scale of 5 km. BARRA also ~~consistently~~<u>more closely</u> reproduces the distribution of wet days and transition probabilities. The
20   performance of BARRA varies spatially, with better performance in the temperate zone than in the arid and tropical zones. A point-to-grid evaluation based on correlation, bias and modified Kling-Gupta efficiency (KGE') indicates that ERA-Interim performs on par or better than BARRA. However, on a spatial scale, BARRA outperforms ~~AWAP~~<u>ERA-Interim</u> in terms of KGE' score and the components of the KGE' score. Our evaluation illustrates that BARRA, with richer spatial variations in climatology of daily precipitation, provides an improved representation of precipitation compared with the coarser ERA-
25   Interim. It is a useful complement to existing precipitation datasets for Australia, especially in sparsely gauged regions.

1 Introduction

Availability of accurate precipitation datasets is an essential requirement for the modelling of natural processes, hydro-meteorological analyses and forecasting, monitoring climatic variations and changes (Kirschbaum et al., 2017; Kucera et al.,

2013; Robertson et al., 2013). A comprehensive knowledge of occurrence and distribution of precipitation is however hindered by the sparseness of the gauging network. Variations in the density and coverage of the gauging network make it difficult to capture information on the spatial and temporal variability of rainfall. This is particularly the case in areas covered by deserts, mountains, and oceans and in large areas with low population densities (Salio et al., 2015; Thiemig et al., 2012). This presents a challenge for the Australian continent where the gauges are mostly located along the densely populated coastal regions. The station network is less dense in the central region which represents the more arid part of the continent (Johnson et al., 2016).

The difficulties inherent in existing observation networks have prompted the development of various gridded datasets with consistent spatial and temporal scale. One such precipitation dataset for Australia is the interpolated precipitation product from the Australian Water Availability Project (AWAP). Because of the uneven gauge network distribution, it is not consistent in terms of accuracy (Jones et al., 2009). Global and regional reanalysis datasets provide another source of precipitation data at a consistent spatial and temporal resolution. Such reanalysis datasets are generated using a numerical weather prediction (NWP) model and a data assimilation scheme to incorporate the available observations thereby providing a consistent method of representation of the atmosphere at a regular interval over larger spatial and temporal domain (Parker, 2016). The availableA range of global reanalysis datasets such as NCEP-CFSR (Saha et al., 2010), ERA-Interim (Dee et al., 2011), JRA-55 (Kobayashi et al., 2015) are readily available.

Global reanalysis datasets have been evaluated for various applications at global and regional scales. Of the global dataset exclusively based on reanalysis, ERA-Interim is generally considered to provide better performance compared to other reanalysis datasets (Beck et al., 2017, 2019). ERA-Interim has been found to reproduce the climatology of global monsoon precipitation (Lin et al., 2014) and in general it demonstrates high temporal and spatial correlations with interpolated observations (Donat et al., 2014). In a recent global evaluation of gridded precipitation datasets using gauge observations by Beck et al (2017), the ERA-Interim and JRA-55 reanalysis datasets were found to reproduce long-term trends and temporal correlation more reliably than achieved by satellite datasets. In the Australian continent, ERA-Interim reproduced the observed spatial patterns of long-term rainfall along with other climatic variables and showed an overall better performance compared to NCEP-NCAR (National Centers for Environmental Prediction/National Center for Atmospheric Research) reanalysis (Fu et al., 2016). An evaluation by Peña-Arancibia et al. (2013) of reanalysis datasets, satellite products, and an ensemble of these datasets in Australian and Asian regions showed that the ERA-Interim performed better than other individual datasets across a range of metrics for the Australian region.

The available global reanalysis datasets such as NCEP-CFSR, ERA-Interim, and JRA-55 cover the Australian region, but their horizontal resolutions are relatively coarse ($\geq$ 80 km) and unsuitable for fine-scale application in hydro-meteorological analysis. The resolution of the global reanalysis can be enhanced by downscaling approaches such as dynamic downscaling (Soares et al., 2012) or statistical analysis using high-resolution surface observations (Vidal et al., 2010). Alternatively, the

2

application of a regional model-based data assimilation can provide a better representation of local climate features and extreme events (Bollmeyer et al., 2015). The regional reanalysis, unlike global reanalysis, allows the integration of abundant local surface observations at a finer scale ~~(Bollmeyer et al., 2015; Isotta et al., 2014; Jakob et al., 2017).~~ (Bollmeyer et al., 2015; Isotta et al., 2015; Jakob et al., 2017). The studies have also been conducted to evaluate the additional benefit of high-resolution

5   datasets obtained from regional reanalysis. Jermey and Renshaw (2016) found that the overall performance of 12km regional reanalysis was significantly better than 80km ERA-Interim especially for high thresholds of rainfall. Similarly, the performance of a regional analysis against ERA-Interim at grid scale showed an improvement in representing high-threshold events (Isotta et al., 2015; Jermey and Renshaw, 2016; Roberts and Lean, 2008) and convective events (Roberts and Lean, 2008). In general, regional reanalyses based on boundary conditions from ERA-Interim is shown to improve the performance

10   over ERA-Interim albeit these studies are largely focused in Europe.

The Bureau of Meteorology Atmospheric high-resolution Regional Reanalysis for Australia (BARRA) includes a 12km modelling and assimilation system (BARRA-R; in this paper, we refer to BARRA-R as BARRA for convenience) which is the first 12 km regional reanalysis conducted over Australia, New Zealand and southeast Asia ~~(Su et al., 2018). It provides a greater understanding of the past weather, particularly extreme events. It will, therefore, support better planning and~~

15   ~~management to reduce risks for the future.~~(Su et al., 2019). BARRA is expected to provide an improved understanding of the past weather than previously possible, particularly for extreme events, and this should support better planning and management of climate-related risks in the future. BARRA makes use of local surface observations and locally derived wind vectors which are not available to global reanalysis models. However, precipitation observations are not assimilated in BARRA. Precipitation is modelled using a microphysics parameterization and a mass flux convection scheme. The modelled precipitation can be

20   erroneous (Bukovsky and Karoly, 2007; Parker, 2016). Therefore, it is necessary to understand the nature of the uncertainty and inaccuracies involved and to use this understanding to correct for any systematic errors.

The criteria for ~~the~~ evaluation of the reanalysis precipitation need to be rigorous and relevant to the intended purpose of use (Parker, 2016). The type of reference datasets (point or grid), the spatial and temporal scale, and the metrics used vary widely among the studies. In general, gridded precipitation products including reanalysis datasets, satellite precipitation estimates,

25   and/or interpolated datasets are evaluated against benchmark datasets ~~(Baez Villanueva et al., 2018; Beck et al., 2017; Gebremichael, 2010; Isotta et al., 2014; de Leeuw et al., 2015; Peña Arancibia et al., 2013; Zambrano Bigiarini et al., 2017). The suitable benchmark datasets are comprised of~~(Baez-Villanueva et al., 2018; Beck et al., 2017; Gebremichael, 2010; Isotta et al., 2015; de Leeuw et al., 2015; Peña-Arancibia et al., 2013; Zambrano-Bigiarini et al., 2017). Suitable benchmark datasets comprise point measurements (Baez-Villanueva et al., 2018; Chiaravalloti et al., 2018; Salio et al., 2015; Thiemig et al., 2012;

30   Zambrano-Bigiarini et al., 2017) and/or high-quality gridded datasets ~~(Chiaravalloti et al., 2018; Isotta et al., 2014; Peña Arancibia et al., 2013)~~(Chiaravalloti et al., 2018; Isotta et al., 2015; Peña-Arancibia et al., 2013). Catchment-scale

3

representativeness of precipitation datasets is assessed by evaluating catchment average precipitation (e.g. Thiemig et al., 2012). Evaluation studies for hydrological applications focus on a daily timescale analysis (Baez-Villanueva et al., 2018; Thiemig et al., 2012; Zambrano-Bigiarini et al., 2017), though this has been extended to a sub-daily time interval for a purpose of hydrologic risk assessment (Chiaravalloti et al., 2018).

5   Given the complex statistical behaviour of precipitation, a range of evaluation metrics are available which have their own assumptions and limitations. Typical metrics ~~are~~include unconditional scores (e.g. correlation, bias, root mean square error, mean absolute error), categorical metrics (e.g. probability of detection, false alarm ratio, skill scores) and distributional statistics (Gebremichael, 2010). In addition, statistical properties like quantiles, wet day frequency, and transition probabilities help to ascertain if the dataset under evaluation preserves the statistical properties related to sequencing. Any single metric

10 cannot adequately represent the nature of all errors in the precipitation products. Therefore, it is essential to evaluate multiple metrics which ~~describes~~describe different aspects of precipitation to identify the possible sources of mismatch between datasets (Baez-Villanueva et al., 2018).

This study is one of the first comprehensive explorations of the BARRA precipitation estimates. ~~It aims to identify its strengths and limitations relative to other available precipitation datasets (specifically, to AWAP and ERA-Interim, as described below).~~

15 ~~The precipitation estimates in the BARRA reanalysis are available at hourly timescale. While we are interested in the accuracy of the sub-daily estimates as well, this study focuses only on a daily scale. The accuracy at a daily scale provides us with an important benchmark as it is applicable to many hydrological applications and also forms the basis for further examination at finer timescales. We explore a range of metrics to evaluate the depth and statistical distribution of precipitation at a daily level. In addition, we examine the wet day frequency and transition probabilities to assess the suitability of the datasets for climate~~

20 ~~studies. The evaluation is carried out on both point and areal precipitation estimates. This evaluation sheds light into the potential hydro-meteorological applications of the dataset such as climate change studies, water resource management, floods, and droughts.~~It aims to identify strengths and limitations of the BARRA precipitation to ascertain its efficacy for further hydrometeorological applications in the Australian region. The evaluation is performed against gauged measurements from two data sources: one is based on gauged point observations, and the other is a high-resolution gridded AWAP dataset derived

25 from interpolating gauge measurements, which is widely accepted as being the best synthesis of gauged observations. There are inherent differences in gridded and point rainfall estimates due to the spatial averaging of point observations across each grid cell area. Since BARRA provides direct estimates of area-average rainfall, it is useful to compare these estimates with the best available point and areal precipitation estimates. In addition to observed and interpolated datasets, we also compare BARRA with corresponding areal-rainfalls provided by the ERA-Interim global reanalysis. ERA-Interim was selected as it is

30 used to drive BARRA (Su et al., 2019) and thus this comparison reveals the extent to which the high-resolution regional

reanalysis provides additional value over ERA-Interim in capturing finer-scale meteorology. ERA-Interim was also selected as it has been found to be one of the best performing reanalysis datasets (Peña-Arancibia et al., 2013).

Despite the availability of BARRA precipitation as hourly values, we select a daily timescale to match the temporal resolution of the best available gridded reference dataset (AWAP). In addition, the accuracy at a daily scale serves as a desirable first
5 step towards further examination at finer timescales because any reliance on sub-daily estimates necessarily depends on its ability to correctly represent daily precipitation. The additional value of evaluation at a daily scale is that it assesses the potential of BARRA to provide estimates of daily rainfall in the sparsely gauged regions across Australia. We focus our evaluation on a range of metrics related to the suitability of rainfalls for hydrological applications and climate studies. We consider the depth, statistical distribution, wet day frequency and transition probabilities at a daily level for evaluation. The
10 purpose of this evaluation is thus to shed light on the suitability of BARRA for input to hydro-meteorological applications related to climate change studies, water resource management, and the analyses of floods and droughts.

2 Datasets

The sub-hourly time series of recorded rainfall from continuous rainfall stations are obtained from the Bureau of Meteorology. Daily rainfall is generated by aggregating the sub-daily rainfall observations. The period of analysis was determined by the
15 duration at which all the datasets used in the study are available i.e. January 2010 to December 2015. The gauge stations used in the study are chosen based on the availability of information for the entire period over which BARRA data estimates are available. A total of 441 stations are selected, and their spatial distribution is shown in Figure 1(a).

An overview of gridded datasets used in this study is presented in the Table 1.The Australian Water Availability Project (AWAP) provides a daily high-quality 0.05 ° × 0.05 ° (around 5 km) gridded rainfall product, dating back to 1900, based on
20 an extensive network of rain gauges, as described by (Jones et al., 2009). The AWAP product makes use of daily (9 am to 9 am) data from all available gauges across the whole of Australia. The gridded estimates are obtained using a weighting scheme that incorporates an optimized Barnes successive-correction algorithm and orographic corrections (Jones et al., 2009). AWAP uses land-based observations only and therefore does not provide information over the ocean.

The global reanalysis ERA-Interim of the European Centre for Medium-Range Weather Forecasts (ECMWF, www.ecmwf.int)
25 covers the period from 1 January 1979 to present (Dee et al., 2011). The core component of the ERA-Interim data assimilation system is the 12h 4D-variational (4DVar) analysis scheme of the upper-air atmospheric state, which is on a spectral grid with triangular truncation of 255 waves (corresponding to approximately 80 km) and a hybrid vertical co-ordinate system with 60 vertical levels. The precipitation is estimated by the numerical model based on temperature and humidity information derived from assimilated observations originating from Passive Microwave and in-situ measurements. In this study, the sub-daily

precipitation from ERA-Interim is converted to daily by accumulation to 24 hours to be consistent with the gauged and AWAP data.

5 ~~BARRA also uses 4DVar but at the 36 km resolution, and uses the Unified Model as its forecast model (Su et al., 2018). BARRA extends spatially over 65.0° to 196.9° east, -65.0° to 19.4° north at a spatial resolution of 0.11° (approximately 12 km) and with 70 levels up to 80 km into the atmosphere. The project aims to produce the dataset dating back to 1990, however, at present, only a six-year period (2010-2015) is available. The analysis for BARRA is conducted 4 times a day with a 6-hour analysis window centred at time t_0 = 0, 6, 12 and 18 UTC. The forecast cycle is of 12 hours which ranges from t_0-3h to t_0+9h (Su et al., 2018). The BARRA precipitation forecasts from t_0+4 hour to t_0+9 hour are used to generate a daily time series for the study.~~

10 ~~**3 Methodology**~~

~~**3.1 Methodology for data comparison**~~

BARRA also uses 4DVar but at the 36 km resolution, and uses the Unified Model as its forecast model (Su et al., 2019). BARRA extends spatially over 65.0° to 196.9° east, -65.0° to 19.4° north at a spatial resolution of 0.11° (approximately 12 km) and with 70 levels up to 80 km into the atmosphere. The project aims to produce the dataset dating back to 1990, however,

15 at present, only a six-year period (2010-2015) is available. The model includes a comprehensive set of parametrisations, including a modified boundary layer scheme, mixed phase cloud microphysics, a mass flux convection scheme, and a radiation scheme. The model parametrisation in BARRA mainly is inherited from the UKMO Global Atmosphere (GA) 6.0 configurations as described in Walters et al. (2017). Surface and satellite rainfall observations are not assimilated, and the precipitation fields are determined using the assimilated large-scale variables and the physical parameterisation of the model.

20 At 12 km horizontal resolution, BARRA requires the convection scheme to model sub-grid scale convection using an ensemble of cumulus clouds as a single entraining-detraining plume (Clark et al., 2016). The scheme prevents unstable growth of cloud structures on the grid and explicit vertical circulations and can only predict an area-average rainfall instead of a spectrum of rainfall rates. The analysis for BARRA is conducted 4 times a day with a 6-hour analysis window centred at time t_0 = 0, 6, 12 and 18 UTC. The forecast cycle is of 12 hours which ranges from t_0-3h to t_0+9h (Su et al., 2019). The BARRA precipitation

25 forecasts from t_0+4 hour to t_0+9 hour are used to generate a daily time series for the study.

3 Methodology

3.1 Reference Datasets

The study is conducted over a period of 6 years, from January 2010 to December 2015~~. We have performed~~, which represents the full period of available data at the commencement of the study. We undertake a point-to-grid and grid-to-grid evaluation

5   of the reanalysis datasets against suitable ~~benchmark~~reference datasets. The gauged data represents the best estimate of rainfall at a point, and AWAP data provides the best estimate of rainfall over a grid cell. Both estimates are, however, limited by the available gauging density. AWAP ~~is~~estimates are based on all available daily and continuous gauges~~. The~~, and the point data considered here are from selected sub-daily rainfall stations with maximum availability of data over the study period. The selected gauges are distributed ~~all over~~unevenly across Australia and are spatially representative of the available gauging

10   network. As ~~gauges measure point rainfall and~~the gridded datasets represent an areal average, it may be expected that ~~interpolated~~ there are differences between point and gridded estimates as the latter account for some spatial averaging. While the gauged data and AWAP rainfall estimates represent the best available reference datasets based on measured data, both are imperfect representations of areal rainfalls~~differ from point estimates.~~. The AWAP estimates ~~will~~ contain inaccuracies due to the interpolation method, ~~but~~and the point estimates ~~are also an imperfect~~provide only a coarse estimate of rainfall over a grid

15   cell area. ~~Consequently~~The ability of these point and gridded reference data sets to represent actual areal rainfalls is heavily dependent on the gauging density and local orography, and these factors influence the accuracy of the reference data sets to different degrees across Australia. Accordingly, we ~~compared~~compare the BARRA and ERA-Interim estimates to both point gauged and AWAP areal data~~.~~, where the evaluation of both offer different insights to the quality of model estimates.

~~In order to compare gridded rainfalls with point rainfalls, a nearest neighbour method is employed.~~Gridded rainfalls are

20   compared to point rainfalls using a nearest neighbour approach. The choice of interpolation scheme is especially important when comparing datasets of different spatial resolutions; bilinear interpolation is likely to smoothen the precipitation field resulting in higher bias for larger rainfall (Accadia et al., 2003), whereas nearest neighbour method preserves the magnitude of the precipitation over the grid. At each location of point rainfall, the nearest neighbour method assigns the precipitation amount from the nearest grid point. In grid-to-grid evaluation, both reanalysis datasets are interpolated to a common AWAP

25   spatial scale using nearest neighbour method. That means, for each AWAP grid, the precipitation variable is obtained from the nearest grid of reanalysis datasets. Given the spatial inconsistency in the ~~quality~~accuracy of AWAP dataset, the grid-to-grid evaluation is limited to grid points that are nearest to the gauge stations used in this study.

The performance of daily precipitation from the reanalysis data is assessed considering all days of the year and by seasons (Summer: DJF, Autumn: MAM, Winter: JJA, and Spring: SON). In addition, evaluation is stratified across three broad climatic

30   zones (arid, tropical and temperate) as defined by the Köppen-Geiger classification (Peel et al., 2007). Most comparisons are undertaken using estimates from every day in the 6-year period, and thus the number of wet days varies with the different

datasets and locations considered. However, it is recognized that reanalysis products tend to produce a high number of days with light drizzle, and therefore they over-estimate the frequency of wet days. Thus, following Baez-Villanueva et al. (2018), Ebert et al. (2007) and Zambrano-Bigiarini et al. (2017), a threshold of 1 mm/day is used to classify a day as 'wet' or 'dry'.

3.2 Performance indices

The evaluation of precipitation data generally involves an assessment of detection capabilities and biases in the form of continuous and categorical metrics. The continuous metrics used in this study are Modified Kling-Gupta efficiency (KGE') (Gupta et al., 2009; Kling et al., 2012) along with its three individual components: correlation (r), bias ratio (β), and variability ratio (γ). Baez-Villanueva et al. (2018) and Zambrano-Bigiarini et al. (2017) also suggested the use of KGE' and its components for evaluating precipitation datasets as it provides an overall assessment along with an error in the representation of magnitude and variability of the reference precipitation.

The four categorical indices adopted are the probability of detection (POD) or hit rate, false alarm ratio (FAR), critical success (or threat) index (CSI), and frequency bias (fBias). These are evaluated over five different rainfall intensity classes, namely: no rain (<1 mm), light rain (≥1 mm and <5 mm), moderate rain (≥5 mm and <20 mm), heavy rain (≥20 mm and <40 mm) and violent rain (≥40 mm) (Baez-Villanueva et al., 2018; Zambrano-Bigiarini et al., 2017). ~~The~~These standard continuous and categorical metrics ~~used in the study~~ are described in ~~Table 1.~~ Table S2 (in the supplement).

3.3 Precipitation statistics

In this study, we evaluate the capacity of the reanalysis datasets to reproduce a range of precipitation statistics related to the frequency of wet days, transition probabilities between wet and dry days, and the distribution of rainfall amounts based on the 90%, 95%, and 99% daily exceedance values.

The transition probabilities considered here are p_{01} and p_{11}, which denote the probabilities of a dry day followed by a wet day, and a wet day followed by a wet day, respectively. Consider X_0 and X_1 are two consecutive days, and i, j are two states (0: dry, 1: wet), then the transition probability is expressed as

$$p_{ij} \ = \ Pr(X_1 = j | X_0 = i)$$

For dry-wet day: $p_{01} \ = \ Pr(X_1 = 1 | X_0 = 0)$

For wet-wet day: $p_{11} \ = \ Pr(X_1 = 1 | X_0 = 1)$

4 Results

4.1 Mean precipitation

~~Figure 1~~Figure 1 shows the average annual precipitation over Australia for the period of six years (2010-2015) as estimated using the different data sets (gauged point rainfalls, AWAP, BARRA, and ERA-Interim). It is seen that there is a high spatial variability in the average precipitation over Australia. Regions of high rainfall (northern and eastern coasts along with western Tasmania) are similarly represented by all four datasets. ERA-Interim provides a coarser representation of the precipitation field, which expectedly fails to capture the higher spatial variability in the coastal regions and orographic precipitations in the Great Dividing ranges and western Tasmania. By comparison, BARRA precipitation captures this variability in the AWAP precipitation. It should be noted, however, that the AWAP data provides a poor estimate of precipitation over central Australia where there is a paucity of gauging information. BARRA, on the other hand, provides high-resolution precipitation pattern over the ocean as well as the central Australian region where gauges are sparse.

4.2 Point-to-grid assessment

KGE' and its components

The variation of KGE' at the gauge locations is shown in ~~Figure 2.~~Figure 2. As expected, the higher performance of KGE' is observed in the AWAP dataset as this is derived by interpolating gauged data, which includes the selected high-quality point rainfall gauges. The pattern of KGE' score looks similar among both reanalysis datasets. The performance, however, varies spatially with a better score for the gauges located in the southern region in comparison to the central and northern region. A difference in KGE' score between BARRA and ERA-Interim (Figure 2d) shows the difference between the performance of reanalysis datasets in various regions. In most of the locations, the difference is minimal indicating the similar performance by both reanalysis datasets. ERA-Interim generally performs better in the central arid region, whereas BARRA exhibits better scores in the temperate region. The central arid region experiences very few rainy days which is likely to be missed spatially by a sparse network of gauges. This could result in poor correlation metrics as it is sensitive to outliers, and subsequently yields a poor KGE' score. BARRA precipitation is more erroneous in the tropics than in temperate regions which may reflect the limitations of the convection parameterisation adopted in BARRA. This spatially varying performance of BARRA is further discussed in section 5.5).

~~Figure 3~~Figure 3 presents the summary of KGE' and its components at a daily scale calculated with reference to gauge data for overall, summer, and winter seasons. In using all the data at a daily scale, the AWAP precipitation proves to be the best estimate at gauge location with the highest value of KGE'. In decomposing the overall performance of KGE' into its components (linear correlation (r), bias ratio (β), and variability ratio (γ)), AWAP achieves the best score for the correlation and variability terms. The AWAP estimates of point rainfall are higher than the gauged observations, and a similar degree of

9

overestimation is exhibited by the BARRA dataset. The Barnes interpolation scheme used in AWAP slightly inflates the spatial coverage of light intensity rainfall (Jones et al., 2009; King et al., 2013). In addition, the weight function assigned to gauged point rainfalls may result in wetter bias in the densely gauged (coastal) regions compared to other regions. Despite having a slightly lower correlation compared to ERA-Interim, the variability of the rainfall is better captured by the BARRA dataset. The overall KGE' score is slightly better for ERA-Interim compared to BARRA dataset.

Evaluation of daily rainfall in different seasons reveals a mixed performance between the different reanalysis products. The performance during winter is better than in summer for both reanalysis datasets. This is likely due to the ability of NWP models to accurately simulate synoptic systems which represent the majority of wintertime rainfall. However, this difference is larger for BARRA than ERA-Interim. There is less discrepancy in the correlation between BARRA and ERA-Interim. However, the variation of correlation across stations is higher for the BARRA dataset. BARRA tends to overestimate the depth of rainfall and the ERA-Interim under-estimates variability component across all the seasons. The mean and variability for BARRA during winter (JJA) are closer to gauge estimates than in summer (DJF) season. Based on KGE' score, the performance of BARRA is lower than ERA-Interim during summer whereas similar during the winter.

The summary of the KGE' scores based on the climatic zone is presented in Table 2.Table 2. As expected, the KGE' scores are better for AWAP than for BARRA and ERA-Interim for all three climatic zones. The reanalyses match AWAP most closely for the Temperate zone, followed by the Arid zone and performance is worst for the Tropical zone. The KGE' metric shows that BARRA performs better at the temperate zone while ERA-Interim at the tropical zone.

Wet day frequency and transition probabilities

Figure 4Figure 4 shows the comparison of the wet day frequency and transition probabilities (p_{01}, dry day followed by a wet day and p_{11}, wet day followed by a wet day) at different gauge locations. Overall, BARRA shows a better correlation to the gauge estimates than ERA-Interim and is less biased except for dry-wet transition probability. As expected, the frequency of wet days and transition probabilities recorded at gauge locations are lower compared to all the gridded datasets The wet. Wet day frequency in all gridded dataset exhibit positive bias with coarser dataset showing higher bias. The difference is higher for the gauge stations in the northern region. From the colour of scatter, it can be observed that the frequency of wet days is lower over central Australia and higher over the south-east coast and Tasmania. The level of agreement between ERA-Interim and gauge estimates of wet day frequencies varies with the proportion of wet days. At the gauges where the frequency of the wet days is very low (<0.2), the estimates from the ERA-Interim are closer to the gauge. However, for the remaining gauges with higher wet day frequencies, the estimation is positively biased and highly scattered. The wet day frequency varies spatially with the location of the station. The spatial pattern, however, is similar for all datasets. Transition probabilities are also dependent on the location of stations with a similar overall trend for all datasets. The model estimates of transition probability

10

p_{11} are greater than p_{01}. <u>Due to the tendency of NWP models to yield more frequent persistent light rainfalls, a high wet-wet frequency is observed.</u> The variation of p_{01} and p_{11} in the northern region is greater than the southern region. BARRA and AWAP both exhibit higher p_{01} values than gauges for all locations. The p_{01} for the northern region is estimated correctly by ERA-Interim, however, it exhibits more bias and spread for the southern region. In comparison to p_{01}, p_{11} shows more scatter and bias. The general trend of overestimation of p_{11} is similar across all datasets and the estimates are higher for the gauges located in northern region compared to southern. The estimates of p_{11} are highly correlated for AWAP. The spatial pattern and correlation are similar for both reanalysis datasets while BARRA exhibits less bias than ERA-Interim.

Quantiles

~~Figure 5~~Figure 5 shows the comparison of quantiles obtained from the gridded datasets to the corresponding quantiles derived for the observed point rainfalls. Since all days were considered in computing the quantile (i.e. both wet and dry days were considered), the 99% value represents "large" rainfalls that are exceeded on average only three or four times per year. On the other hand, the rainfall represented by the 90% quantile (exceeded on average 36 times per year) corresponds to the more frequent wet day rainfalls which depend upon the climatology of the station considered. The AWAP rainfall corresponding to these quantiles are all higher than the point rainfall estimates. The quantile estimates in BARRA and ERA-Interim shows very different patterns to those of the AWAP data set. BARRA estimates for all quantiles (90%, 95%, and 99%) exhibit no consistent bias with location. In contrast, ERA-Interim greatly underestimates the larger precipitation, where the degree of underestimation increases with increasing quantile. The nature of the differences varies with location and quantile: for example, the ERA-Interim 90% quantile estimates are biased low compared to the point rainfalls in the northern region, yet there is a tendency for all higher quantiles to be biased high. The ERA-Interim reanalysis appears unable to represent higher precipitation magnitudes, and the largest value plateaus at about 40mm for the rarer quantiles.

Categorical Metrics

~~Figure 6~~Figure 6 shows the boxplot of four categorical performance indices computed between three gridded datasets with reference to gauged point rainfalls at a daily scale. The probability of detection (POD) is highest for the AWAP dataset, and the POD values for the two reanalysis data sets are consistently lower and somewhat similar. For rainfall intensities greater than 20 mm and lower than 5mm, BARRA shows higher detection of rainfalls, whereas, for the rainfall intensity 5-20 mm, ERA-Interim shows slightly better detection skill. In contrast, the false alarm ratio (FAR) is higher for BARRA for the heavy rainfall classes (> 20 mm). For the low rainfall classes (<20 mm), FAR is lower for BARRA. The Critical Success Index (CSI) shows that the reanalysis data are not able to adequately represent the distribution of rainfalls greater than 1 mm/day. AWAP, as expected, has higher CSI for each of the rainfall classes. BARRA has higher CSI for all rainfall classes than ERA-Interim except the moderate rainfall (5-20 mm/day) where both reanalysis datasets have similar CSI. The frequency bias shows the

positive bias for AWAP and BARRA for all the rain days. ERA-Interim shows a negative bias for higher rainfall intensities (>20 mm).

4.3 Grid-to-grid analysis

KGE' and its components

5   The boxplot of KGE' and its components shown in ~~Figure 7~~Figure 7 is computed between reanalysis datasets with reference to AWAP dataset at a daily scale. Among the components, the correlation is similar for both reanalysis datasets, though BARRA exhibits a slightly wider range of correlations than ERA-Interim. However, the bias ratio (β) and variability ratio (γ) obtained from BARRA is closer to 1, which is appreciably better than the corresponding statistics for ERA-Interim. The difference in the overall KGE' measure for the two data sets is not as pronounced compared to its components.

10   In ~~Table 2~~Table 2, a summary of the KGE' score is also presented with reference to the AWAP dataset. The reanalysis datasets show mixed performance depending on the climate zone and the metrics used. BARRA shows improved performance over ERA-Interim in the temperate zone with better correlation, bias ratio, variability ratio, and KGE' statistics. The difference in correlation coefficients between the reanalysis datasets is higher for the tropical zone but similar for the arid and temperate zones. An assessment of correlation at grid scale is presented in Figure S1 (in the supplement). It also shows a very high
15   correlation in the southern region and a low correlation at the northern region, although a similar spatial pattern is observed for both reanalysis datasets. Further, the BARRA data set exhibits smaller bias than ERA-Interim across all climatic zones. The variability ratio, however, shows contrasting patterns for the two reanalysis datasets. ERA-Interim shows a marked underestimation in variability across all zones. On the other hand, BARRA closely represents the variability in arid and temperate zones with overestimation in the tropics.

20   *Wet day frequency and transition probabilities*

The frequency of wet day and transition probabilities obtained from reanalysis datasets at AWAP grid locations are shown in ~~Figure 8.~~Figure 8. Overall, the representation of wet day frequency and transition probabilities follow a similar pattern to the point-to-grid assessment (~~Figure 4).~~Figure 4). In comparing the gridded datasets with AWAP data as the benchmark, the estimates of wet day frequency and transition probabilities from BARRA are closer than those of ERA-Interim. However, both
25   reanalysis datasets show an improved correlation and reduced bias in grid-to-grid over point-to-grid evaluation. This is expected as this bias is attributed to the spatial resolution of the data.

Quantiles

The comparison of quantiles in gridded datasets in ~~Figure 9~~Figure 9 shows a similar pattern as observed in ~~Figure 5~~Figure 5 for the gauged point rainfalls. Both reanalysis datasets represent the 90% quantile reasonably well. However, the difference between the reanalysis products increases at higher quantiles. The degree of bias in ERA-Interim is considerably more pronounced compared to BARRA.

Categorical Metrics

~~Figure 10~~Figure 10 shows the boxplot of four categorical performance indices computed between reanalysis datasets with reference to AWAP dataset at a daily scale. In general, the POD for both datasets is similar to ~~the~~point-to-grid comparison ~~against gauge~~. For the larger precipitation intensities, the detection capacity of BARRA is higher whereas, for the smaller rainfall intensities, the detection from both datasets is close to each other. The FAR scores for BARRA estimates of higher rainfalls are slightly greater, indicating the larger occurrence of false alarms while detecting higher rainfall. For the low rainfall classes, BARRA exhibits slightly better FAR than ERA-Interim. The CSI score is higher for BARRA across all rainfall classes and indicates that BARRA is slightly more skilful than ERA-Interim. At larger rainfall class, the CSI is notably higher for BARRA indicating its better performance in representing larger rainfall. The frequency bias shows that the BARRA precipitation tends to produce more events of light rainfall while missing out on some larger rainfalls. The bias in BARRA is, however, smaller in comparison with ERA-Interim which shows a marked underestimation of larger rainfall events (>20 mm).

5 Discussion

The evaluation of daily precipitation from BARRA and its comparison against existing datasets reveals a mixed performance under varying benchmark dataset and the metrics. The ~~key insights~~ general observations on the performance of BARRA ~~reanalysis precipitation obtained~~ are summarised in Table 3, and the key insights from ~~the~~these results are discussed below.

5.1 Representation of spatial precipitation structure

The BARRA dataset exhibits similar spatial patterns of mean annual rainfall as the AWAP dataset in regions where gauge density is highest (~~Figure 1).~~Figure 1). However, the reliability in the evaluation of BARRA in the central regions of Australia is confounded by the lack of gauges. The coarser resolution of ERA-Interim misses the small-scale variability~~. In~~; in contrast, ~~a~~the more realistic representation of complex topography in BARRA helps to capture fine-scale features such as orographic precipitation ~~(Su et al., 2018).~~(Su et al., 2019). A good level of agreement in spatial patterns of rainfalls with AWAP and similarity with ERA-Interim estimates (over both land and ocean) at a coarser scale suggests that the BARRA dataset might provide useful information on the distribution of rainfall in regions where direct measurement of precipitation is not available.

13

5.2 Representation of average precipitation and variability

In general, a slight overestimation is observed in mean rainfall from high-resolution gridded data when compared to gauged point rainfall estimates (bias in mm/day for AWAP = 0.48, BARRA = 0.43). In AWAP, the wet bias observed at low rainfall is due to the Barnes interpolation analysis (Jones et al., 2009). In addition, the weight functions tend to spread rainfalls across the grid ~~making~~which tends to induce a wet bias in coastal stations ~~wetter~~ since the density of stations near the coast is greater. When average AWAP precipitation is considered a benchmark for evaluation, BARRA (R=0.93, bias = -0.05mm/day) shows better agreement than ERA-Interim (R=0.84, bias = -0.49 mm/day). In the locations with higher average precipitation, the ERA-Interim underestimates gauged point rainfalls. This can be attributed to ~~smoothing out of~~ the ~~large rainfall due to~~ coarser resolution of ERA-Interim~~.~~; an 80km scale areal averages are expected to be less than observed for rainfall systems that are spatially localized and/or heterogeneous.

In addition, the temporal variability is represented well by BARRA while this is underestimated by ERA-Interim (~~Figure 3 and Figure 7).~~Figure 3 and Figure 7). As with the assessment of spatial variability, the reason for this underestimation is the coarser resolution of ERA-Interim. The variability ratio varies across climatic zones with largest difference in the tropical zone. BARRA greatly overestimates variability in the tropical region, whereas the opposite is true for ERA-Interim (~~Table 2).~~Table 2). The discrepancy in the performance across climatic zones is further discussed in section 5.5.

5.3 Representation of wet day frequency and transition probabilities

All the gridded datasets exhibit higher frequencies of wet day occurrence compared to gauge point rainfalls. This difference is consistent with the physical reasoning that the likelihood of rainfall occurring over an area is always higher than that over a point location. Moreover, an increase in the area produces a higher difference in the wet days and this is consistent with the increasing levels of bias with grid cell size between AWAP, BARRA, and ERA-Interim (~~Figure 4).~~Figure 4). In comparison with AWAP as a benchmark, BARRA closely represents the wet day frequency and exhibits less uncertainty at higher estimates compared to ERA-Interim (~~Figure 8).~~Figure 8).

The variation of transition probabilities p_{01} and p_{11} at a location can be attributed to the seasonal distribution of precipitation at that location. Most rainy days in the northern region occur during October-March with very low or no rainfall for the rest of the season. Due to distinct dry (and wet) season, the likelihood of a dry day following a dry day (and a wet day following a wet day) is higher. This results in larger p_{11} and smaller p_{01} estimates. In contrast, there is not a distinct wet and dry season in the south, and p_{01} is close to p_{11} for the southern region, mostly due to the light rainfall distributed over the longer period. The ability of the reanalysis datasets to represent transition probabilities also varies with location due to the climatology and varying precipitation mechanisms across locations. ~~The transition probabilities p_{01} spread is estimated well by all gridded datasets, whereas the p_{11} estimates are more varied.~~The spread in transition probabilities p_{01} is estimated well by all gridded datasets,

whereas the p_{11} estimates are overestimated and more varied. As mentioned above, this difference is due to the tendency of NWP models to over-estimate the persistence of light rainfalls (Kendon et al., 2012). Due to the likelihood of higher frequency of wet days in gridded datasets, the transition probability p_{11} is higher in gridded datasets and exhibits higher variability for the tropical zone.

5   **5.4 Over/Under-estimation of large rainfall events**

The comparison of quantiles (Figures 5 and 9) demonstrates that the BARRA dataset is able to represent the higher quantiles more accurately than ERA-Interim. The BARRA estimates of quantiles show less bias, but variance increases with the magnitude of precipitation amount. This uncertainty could be due to displacement error as the precipitation becomes more localized with an increase in magnitude. ERA-Interim, on the other hand, greatly underestimates the higher quantiles. This

10   could be attributed to its coarser resolution resulting in the averaging of rainfall over a larger area. These results are consistent with published findings (Isotta et al., 2014; Jermey and Renshaw, 2016)(Isotta et al., 2015; Jermey and Renshaw, 2016) that high-resolution regional reanalysis improves over ERA-Interim in the representation of large rainfall.

The categorical evaluation (Figure 6Figure 6 and Figure 10) also illustrates the improved performance of the BARRA dataset over ERA-Interim during larger rainfall events. Because of the higher spatial resolution, BARRA provides a more accurate

15   representation of larger rainfalls. Therefore, the probability of detection is higher for the larger rainfall classes. Despite exhibiting a higher hit rate, BARRA also shows higher false alarm ratios. The increased false alarm in BARRA is due to the higher number of large rainfall events reported by the BARRA dataset. With larger rainfall events, the false alarm is also likely to increase. Such a trend is usually observed in the assessment of reanalysis and satellite rainfall estimates (Zambrano-Bigiarini et al., 2017). Critical Success Index, the metric which penalizes both misses and false alarms, represents the overall

20   skill of the data. The performance of BARRA lies between the skilful AWAP and less skilful ERA-Interim. The greater CSI for BARRA compared to ERA-Interim suggests that its improved hit rate outweighs the frequency of false alarms. The difference in CSI between BARRA and ERA-Interim is greater for larger rainfall classes, with the former yielding a better score.

5.5 Superior performance over temperate than over tropical and arid

25   Most metrics suggest that the performance of BARRA is superior in the southern (temperate) region compared to the northern (tropical) region. The summary of KGE' and its component across climatic zones (Table 1) clearly shows the variation in performance over different climatic zones. The possible explanation for this may be the difference in the climatic system driving the precipitation in those regions and the scheme used to generate precipitation (Su et al., 2018). Ebert et al. (2007) and de Leeuw et al. (2015) also observed less accurate performance of NWP models at the regime of convective precipitation.

30   The rainfall in the Northern region of Australia is largely convective in nature during summer which is not well captured by

~~the reanalysis model. In addition, the network of surface observations used to develop BARRA is denser in the southern region compared to the northern. This difference in the information content used to derive the estimates may have caused discrepancies in performance in those~~The summary of KGE' and its component across climatic zones (Table 2) and correlation statistics (Figure S1) clearly shows the variation in performance over different climatic zones. The possible explanation for

5   this may be the difference in the climatic systems driving the precipitation in those regions and the scheme used to generate precipitation (Su et al., 2019). Convective precipitation is dominant in tropical regions and the parameterisation scheme for sub-grid convection adopted in BARRA is limited in terms of resolving such precipitation. Therefore, a higher accuracy of BARRA can be expected at high latitudes where synoptic rainfall dominates than in low latitudes where convective rainfall is dominant (Ebert et al., 2007; Su et al., 2019). The limitation of NWP models to represent convective precipitation accurately

10   has also been reported by Ebert et al. (2007) and de Leeuw et al. (2015). In addition, the network of surface observations used to develop BARRA is denser in the southern region compared to the northern, and this difference in information content also contributes to discrepancies in performance between these regions.

5.6 Varying performance in point-to-grid and grid-to-grid evaluation:

~~AWAP is the best estimate of gridded rainfalls over Australian and not surprisingly it shows the best agreement with point~~

15   ~~rainfall. Despite its limitation in some areas due to poor gauging density, the gridded nature of the estimates provides the best basis for comparison with reanalysis datasets at the area with dense gauging network.~~

This study shows that the general pattern of performance between the reanalysis datasets is similar in the point-to-grid evaluation. However, the evaluation against the gridded AWAP estimates showed a markedly better performance by BARRA in terms of overall bias, variability, wet day frequencies, transition probabilities, quantiles and categorical metrics. As expected,

20   BARRA better represents areal rather than point rainfall as it represents the average precipitation field over a grid cell.

6 Conclusions

The purpose of the current study is to document the performance of the BARRA dataset at a daily scale and to provide a comparative analysis of its strengths and limitations relative to other available datasets. The analysis includes point-to-grid and grid-to-grid evaluations at the gauge locations. A range of metrics representing correlation, daily precipitation statistics,

25   and categorical performance are explored and compared on an annual as well as a seasonal basis.

The high-resolution nature of the BARRA dataset provides more detailed and accurate estimates of rainfall across the Australian region than the coarser ERA-Interim. BARRA precipitation exhibits good agreement with the average annual rainfall from AWAP as well as to the gauge dataset. The correlation statistics of the BARRA estimates are slightly lower than for the global reanalysis (ERA-Interim). However, the depth and variability of daily precipitation from AWAP are better

reproduced by BARRA than by ERA-Interim. We can conclude that BARRA precipitation is representative at the spatial scale of AWAP considering that AWAP data provides the best basis for comparison with the reanalysis datasets.

BARRA provides largely unbiased estimates of larger rainfall quantiles whereas ERA-Interim are clearly underestimated. Categorical evaluation also shows a better correspondence of the larger events in BARRA compared to ERA-Interim. This has

5   important implications for hydrological modelling as simulations of runoff processes are heavily dependent on how realistically the precipitation field is represented both spatially and temporally. Information on large rainfalls also important for many engineering investigations where the design loading conditions of interest are dependent on the characteristics of daily rainfall. The superior performance of BARRA in representing large rainfall strengthen a case for its use over ERA-Interim where information about extremes is required.

10   BARRA closely reproduces the frequency of wet days and dry-wet transition probabilities. This evaluation broadly supports BARRA precipitation for its ability to reproduce precipitation statistics at a daily scale. BARRA precipitation could be useful for assessing variation in the spatial and temporal characteristics of precipitation in a consistent manner that is not influenced by differences in gauge density. In addition, it could potentially be used as a source of data in the central arid zone where AWAP estimates are poor or not available. Considering the limitations in the availability of gauged datasets, BARRA could

15   be considered as a valuable reference dataset for hydro-climatic analysis across the whole of Australia, particularly where gauging density is low.

The core attraction of the BARRA dataset is the availability of sub-daily precipitation estimates. Such information is not available in the AWAP data set, and the spatial resolution of the estimates is higher than the currently available global reanalysis and satellite datasets. Accordingly, future work will be directed towards an analysis of BARRA estimates of sub-

20   daily precipitation to assess its ability to represent precipitation at finer time scales.

References

Accadia, C., Mariani, S., Casaioli, M., Lavagnini, A. and Speranza, A.: Sensitivity of Precipitation Forecast Skill Scores to Bilinear Interpolation and a Simple Nearest-Neighbor Average Method on High-Resolution Verification Grids, Weather

Forecast., 18(5), 918–932, doi:10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2, 2003.

Baez-Villanueva, O. M., Zambrano-Bigiarini, M., Ribbe, L., Nauditt, A., Giraldo-Osorio, J. D. and Thinh, N. X.: Temporal and spatial evaluation of satellite rainfall estimates over different regions in Latin-America, Atmos. Res., 213(December 2017), 34–50, doi:10.1016/J.ATMOSRES.2018.05.011, 2018.

5   Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J. and Wood, E. F.: Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling, Hydrol. Earth Syst. Sci., 21(12), 6201–6217, doi:10.5194/hess-21-6201-2017, 2017.

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., van Dijk, A. I. J. M., Huffman, G. J., Adler, R. F. and Wood, E. F.: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS, Hydrol. Earth Syst. Sci., 10   23(1), 207–224, doi:10.5194/hess-23-207-2019, 2019.

Bollmeyer, C., Keller, J. D., Ohlwein, C., Wahl, S., Crewell, S., Friederichs, P., Hense, A., Keune, J., Kneifel, S., Pscheidt, I., Redl, S. and Steinke, S.: Towards a high-resolution regional reanalysis for the european CORDEX domain, Q. J. R. Meteorol. Soc., 141(686), 1–15, doi:10.1002/qj.2486, 2015.

Bukovsky, M. S. and Karoly, D. J.: A Brief Evaluation of Precipitation from the North American Regional Reanalysis, J. 15   Hydrometeorol., 8(4), 837–846, doi:10.1175/JHM595.1, 2007.

Chiaravalloti, F., Brocca, L., Procopio, A., Massari, C. and Gabriele, S.: Assessment of GPM and SM2RAIN-ASCAT rainfall products over complex terrain in southern Italy, Atmos. Res., 206(December 2017), 64–74, doi:10.1016/j.atmosres.2018.02.019, 2018.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., 20   Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N. and Vitart, F.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc., 137(656), 553–597, doi:10.1002/qj.828, 2011.

25   Donat, M. G., Sillmann, J., Wild, S., Alexander, L. V., Lippmann, T. and Zwiers, F. W.: Consistency of Temperature and Precipitation Extremes across Various Global Gridded In Situ and Reanalysis Datasets*, J. Clim., 27(13), 5019–5035, doi:10.1175/JCLI-D-13-00405.1, 2014.

18

Ebert, E. E., Janowiak, J. E. and Kidd, C.: Comparison of near-real-time precipitation estimates from satellite observations and numerical models, Bull. Am. Meteorol. Soc., 88(1), 47–64, doi:10.1175/BAMS-88-1-47, 2007.

Fu, G., Charles, S. P., Timbal, B., Jovanovic, B. and Ouyang, F.: Comparison of NCEP-NCAR and ERA-Interim over Australia, Int. J. Climatol., 36(5), 2345–2367, doi:10.1002/joc.4499, 2016.

5   Gebremichael, M.: Framework for satellite rainfall product evaluation, pp. 265–275., 2010.

Gupta, H. V., Kling, H., Yilmaz, K. K. and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377(1–2), 80–91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

Isotta, F. A., Vogel, R. and Frei, C.: Evaluation of European regional reanalyses and downscalings for precipitation in the
10   Alpine region, Meteorol. Zeitschrift, 24(1), 15–37, doi:10.1127/metz/2014/0584, 20142015.

Jakob, D., Su, C.-H., Eizenberg, N., Kociuba, G., Steinle, P. and Al., E.: An atmospheric high-resolution regional reanalysis for Australia, Bull. Aust. Meteorol. Oceanogr. Soc., 30(3), 16–23, 2017.

Jermey, P. M. and Renshaw, R. J.: Precipitation representation over a two-year period in regional reanalysis, Q. J. R. Meteorol. Soc., 142(696), 1300–1310, doi:10.1002/qj.2733, 2016.

15   Johnson, F., Hutchinson, M. F., The, C., Beesley, C. and Green, J.: Topographic relationships for design rainfalls over Australia, J. Hydrol., 533, 439–451, doi:10.1016/j.jhydrol.2015.12.035, 2016.

Jones, D., Wang, W. and Fawcett, R.: High-quality spatial climate data-sets for Australia, Aust. Meteorol. Oceanogr. J., 58(04), 233–248, doi:10.22499/2.5804.003, 2009.

Kendon, E. J., Roberts, N. M., Senior, C. A. and Roberts, M. J.: Realism of Rainfall in a Very High-Resolution Regional
20   Climate Model, J. Clim., 25(17), 5791–5806, doi:10.1175/JCLI-D-11-00562.1, 2012.

King, A. D., Alexander, L. V. and Donat, M. G.: The efficacy of using gridded data to examine extreme rainfall characteristics: A case study for Australia, Int. J. Climatol., 33(10), 2376–2387, doi:10.1002/joc.3588, 2013.

Kirschbaum, D. B., Huffman, G. J., Adler, R. F., Braun, S., Garrett, K., Jones, E., McNally, A., Skofronick-Jackson, G., Stocker, E., Wu, H. and Zaitchik, B. F.: NASA's Remotely Sensed Precipitation: A Reservoir for Applications Users, Bull.
25   Am. Meteorol. Soc., 98(6), 1169–1184, doi:10.1175/BAMS-D-15-00296.1, 2017.

19

Kling, H., Fuchs, M. and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, J. Hydrol., 424–425, 264–277, doi:10.1016/j.jhydrol.2012.01.011, 2012.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., ONODA, H., ONOGI, K., KAMAHORI, H., KOBAYASHI, C., ENDO, H., MIYAOKA, K. and TAKAHASHI, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, J. Meteorol. Soc. Japan. Ser. II, 93(1), 5–48, doi:10.2151/jmsj.2015-001, 2015.

Kucera, P. A., Ebert, E. E., Turk, F. J., Levizzani, V., Kirschbaum, D., Tapiador, F. J., Loew, A. and Borsche, M.: Precipitation from space: Advancing earth system science, Bull. Am. Meteorol. Soc., 94(3), 365–375, doi:10.1175/BAMS-D-11-00171.1, 2013.

de Leeuw, J., Methven, J. and Blackburn, M.: Evaluation of ERA-Interim reanalysis precipitation products using England and Wales observations, Q. J. R. Meteorol. Soc., 141(688), 798–806, doi:10.1002/qj.2395, 2015.

Lin, R., Zhou, T. and Qian, Y.: Evaluation of Global Monsoon Precipitation Changes based on Five Reanalysis Datasets, J. Clim., 27(3), 1271–1289, doi:10.1175/JCLI-D-13-00215.1, 2014.

Parker, W. S.: Reanalyses and Observations: What's the Difference?, Bull. Am. Meteorol. Soc., 97(9), 1565–1572, doi:10.1175/BAMS-D-14-00226.1, 2016.

Peel, M. C., Finlayson, B. L. and McMahon, T. A.: Updated world map of the Köppen-Geiger climate classification, Hydrol. Earth Syst. Sci., 11(5), 1633–1644, doi:10.5194/hess-11-1633-2007, 2007.

Peña-Arancibia, J. L., van Dijk, A. I. J. M., Renzullo, L. J. and Mulligan, M.: Evaluation of Precipitation Estimation Accuracy in Reanalyses, Satellite Products, and an Ensemble Method for Regions in Australia and South and East Asia, J. Hydrometeorol., 14(4), 1323–1333, doi:10.1175/JHM-D-12-0132.1, 2013.

Roberts, N. M. and Lean, H. W.: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events, Mon. Weather Rev., 136(1), 78–97, doi:10.1175/2007MWR2123.1, 2008.

Robertson, D. E., Shrestha, D. L. and Wang, Q. J.: Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting, Hydrol. Earth Syst. Sci., 17(9), 3587–3603, doi:10.5194/hess-17-3587-2013, 2013.

Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y. T., Chuang, H. Y., Juang, H. M. H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Van Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Van Den Dool, H., Kumar, A.,

Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J. K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C. Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R. W., Rutledge, G. and Goldberg, M.: The NCEP climate forecast system reanalysis, Bull. Am. Meteorol. Soc., 91(8), 1015–1057, doi:10.1175/2010BAMS3001.1, 2010.

Salio, P., Hobouchian, M. P., García Skabar, Y. and Vila, D.: Evaluation of high-resolution satellite precipitation estimates over southern South America using a dense rain gauge network, Atmos. Res., 163, 146–161, doi:10.1016/j.atmosres.2014.11.017, 2015.

Soares, P. M. M., Cardoso, R. M., Miranda, P. M. A., de Medeiros, J., Belo-Pereira, M. and Espirito-Santo, F.: WRF high resolution dynamical downscaling of ERA-Interim for Portugal, Clim. Dyn., 39(9), 2497–2522, doi:10.1007/s00382-012-1315-2, 2012.

Su, C.-H., Eizenberg, N., Steinle, P., Jakob, D., Fox-Hughes, P., White, C. J., Rennie, S., Franklin, C., Dharssi, ~~D~~I. and Zhu, H.: BARRA v1.0: ~~The~~the Bureau of Meteorology Atmospheric high-resolution Regional Reanalysis for Australia, ~~submitted to Geoscientific~~Geosci. Model ~~Development, 2018~~Dev., 12(5), 2049–2068, doi:10.5194/gmd-12-2049-2019, 2019.

Thiemig, V., Rojas, R., Zambrano-Bigiarini, M., Levizzani, V. and De Roo, A.: Validation of Satellite-Based Precipitation Products over Sparsely Gauged African River Basins, J. Hydrometeorol., 13(6), 1760–1783, doi:10.1175/JHM-D-12-032.1, 2012.

Vidal, J. P., Martin, E., Franchistéguy, L., Baillon, M. and Soubeyroux, J. M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, Int. J. Climatol., 30(11), 1627–1644, doi:10.1002/joc.2003, 2010.

Walters, D., Boutle, I., Brooks, M., Melvin, T., Stratton, R., Vosper, S., Wells, H., Williams, K., Wood, N., Allen, T., Bushell, A., Copsey, D., Earnshaw, P., Edwards, J., Gross, M., Hardiman, S., Harris, C., Heming, J., Klingaman, N., Levine, R., Manners, J., Martin, G., Milton, S., Mittermaier, M., Morcrette, C., Riddick, T., Roberts, M., Sanchez, C., Selwood, P., Stirling, A., Smith, C., Suri, D., Tennant, W., Vidale, P. L., Wilkinson, J., Willett, M., Woolnough, S. and Xavier, P.: The Met Office Unified Model Global Atmosphere 6.0/6.1 and JULES Global Land 6.0/6.1 configurations, Geosci. Model Dev., 10(4), 1487–1520, doi:10.5194/gmd-10-1487-2017, 2017.

Zambrano-Bigiarini, M., Nauditt, A., Birkel, C., Verbist, K. and Ribbe, L.: Temporal and spatial evaluation of satellite-based rainfall estimates across the complex topographical and climatic gradients of Chile, Hydrol. Earth Syst. Sci., 21(2), 1295–1320, doi:10.5194/hess-21-1295-2017, 2017.
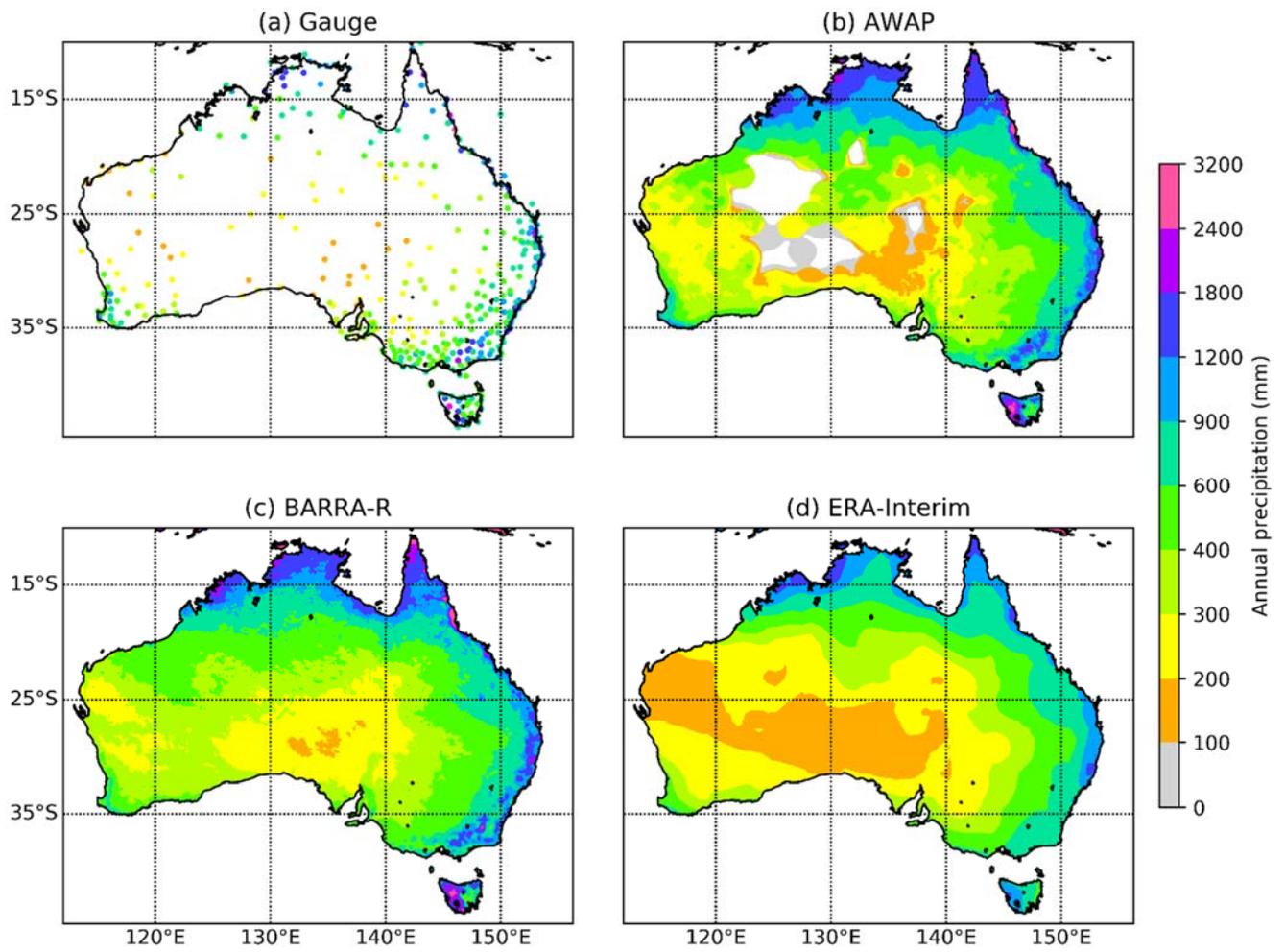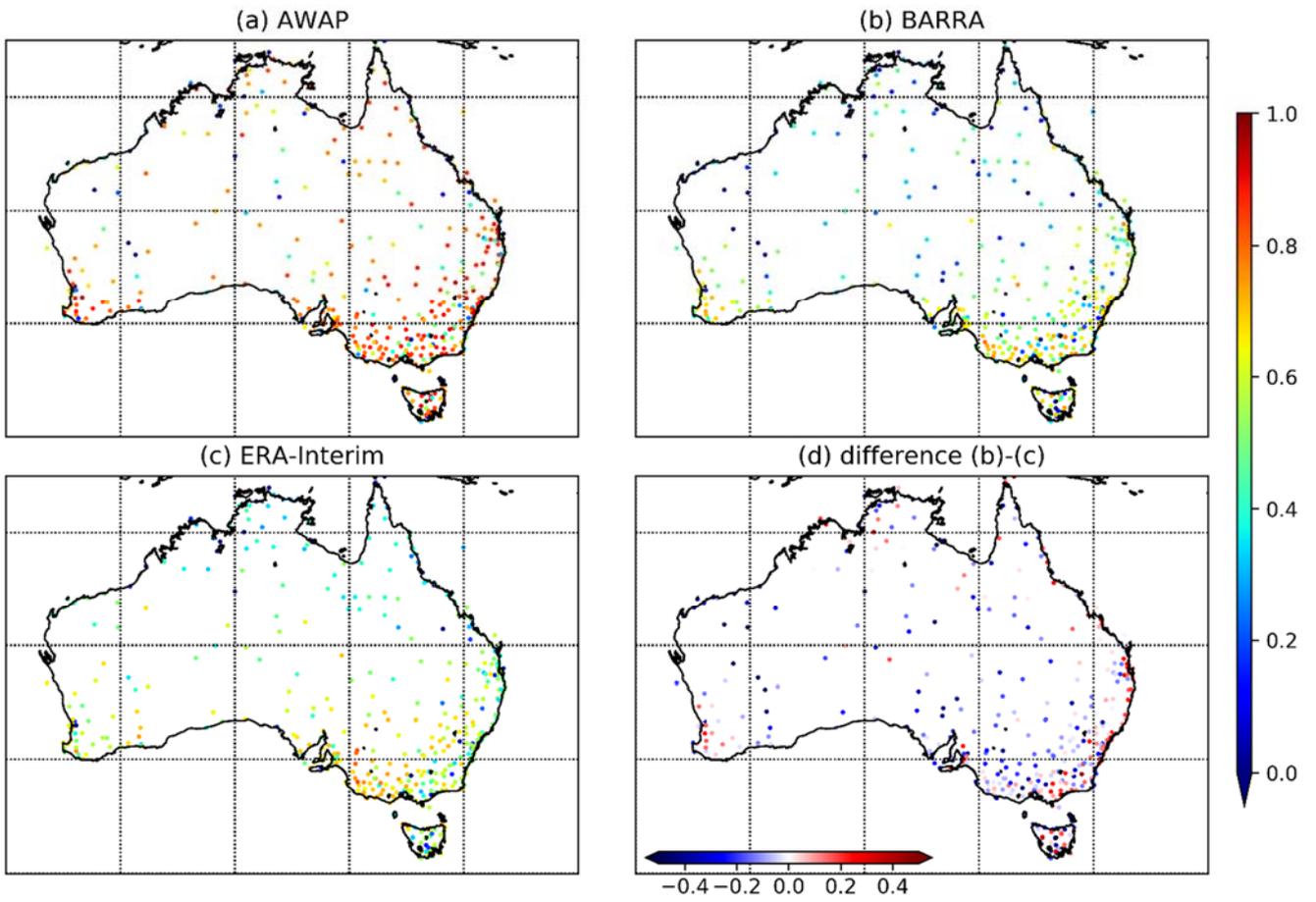
21
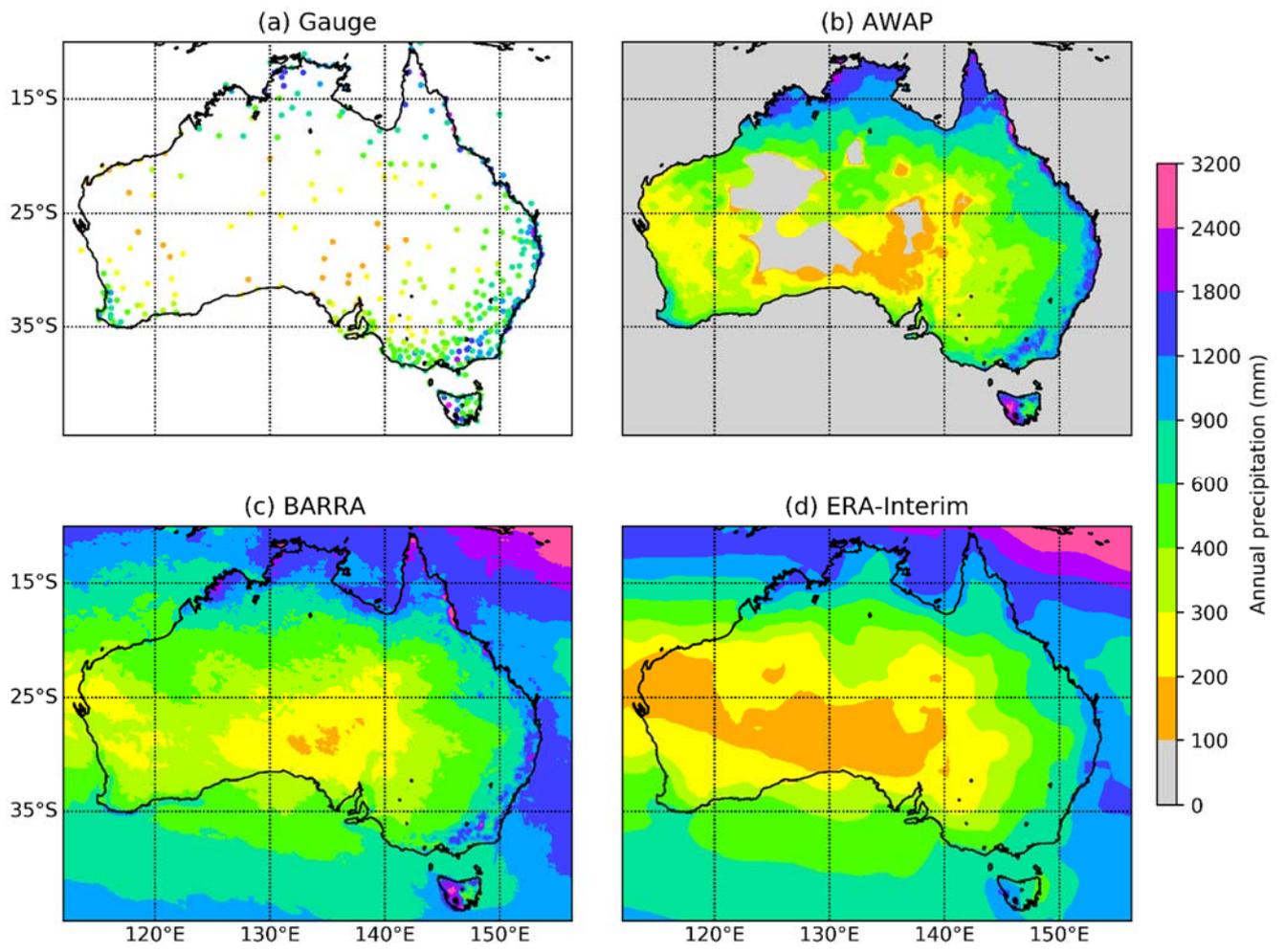
Figure 1 Average annual precipitation over Australian region (a)Gauge data, (b) AWAP, (c) BARRA, and (d) ERA-Interim. Missing values and ocean are masked in ~~AWAP and represented by gray colour~~all datasets.

5

24

(a) AWAP
(b) BARRA
(c) ERA-Interim
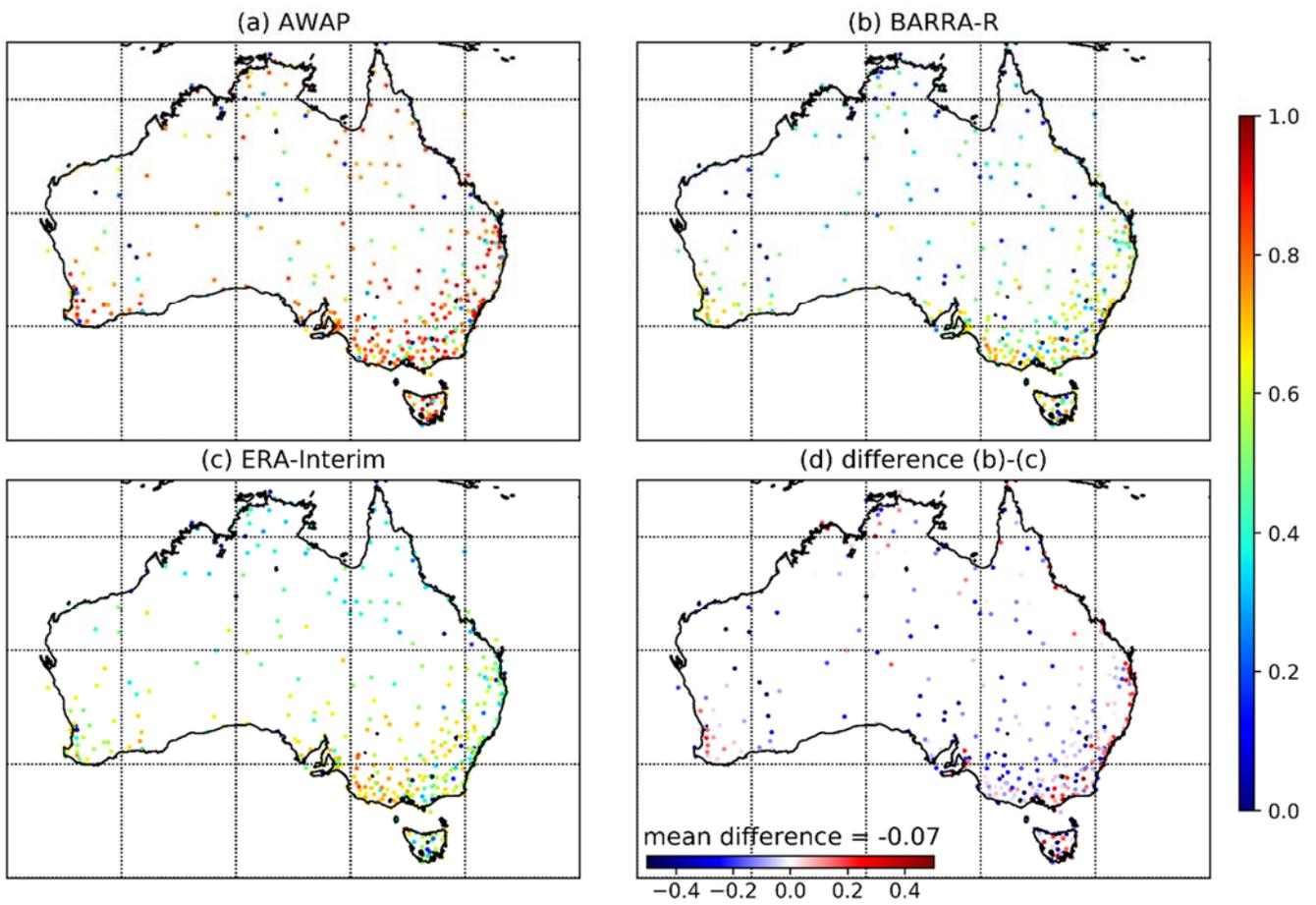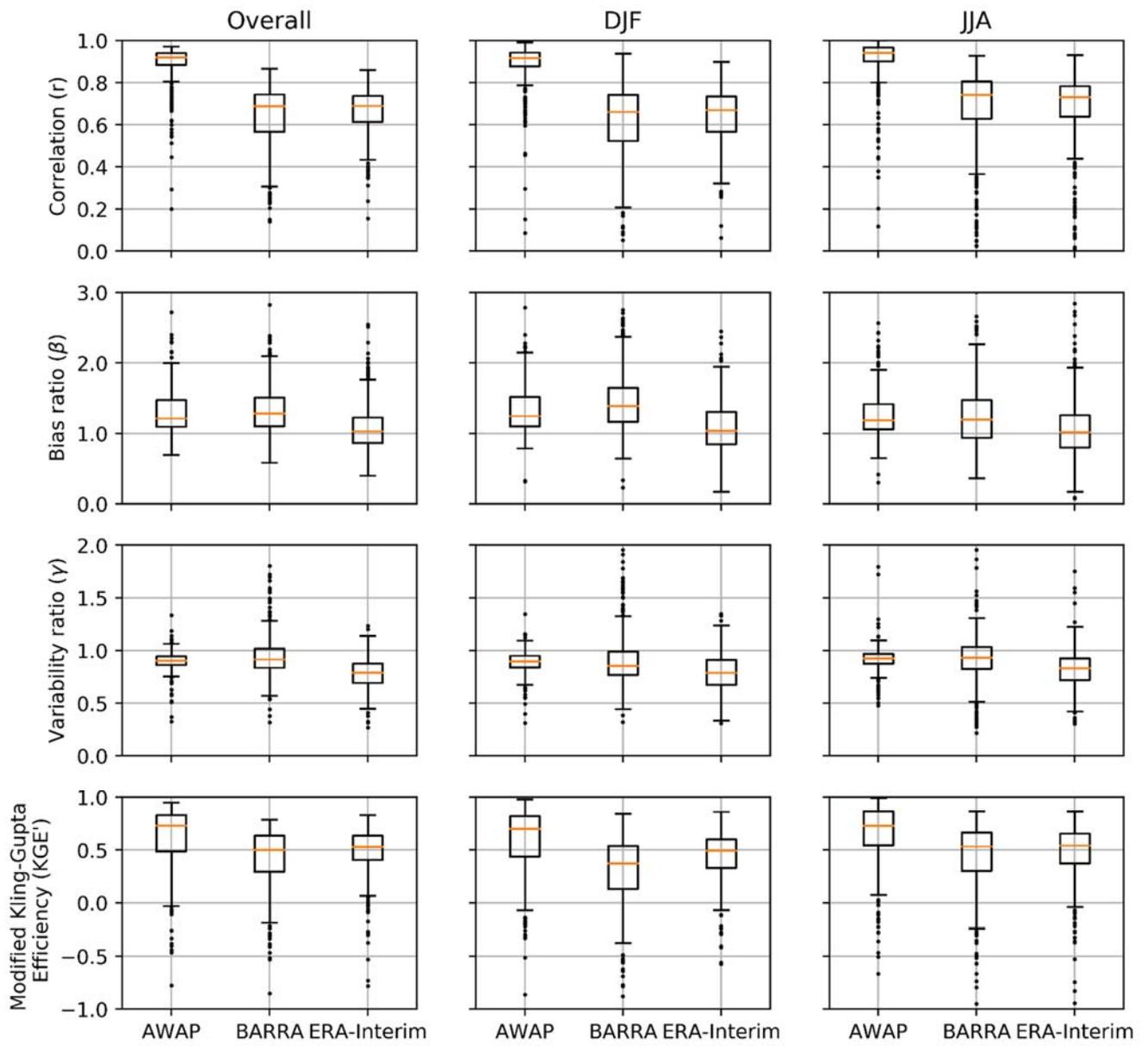(d) difference (b)-(c)

Figure 2 Modified Kling-Gupta efficiency (KGE') against gauge dataset at daily scale for (a) AWAP, (b) BARRA, (c) ERA-Interim, and (d) difference of KGE' between BARRA and ERA-Interim.
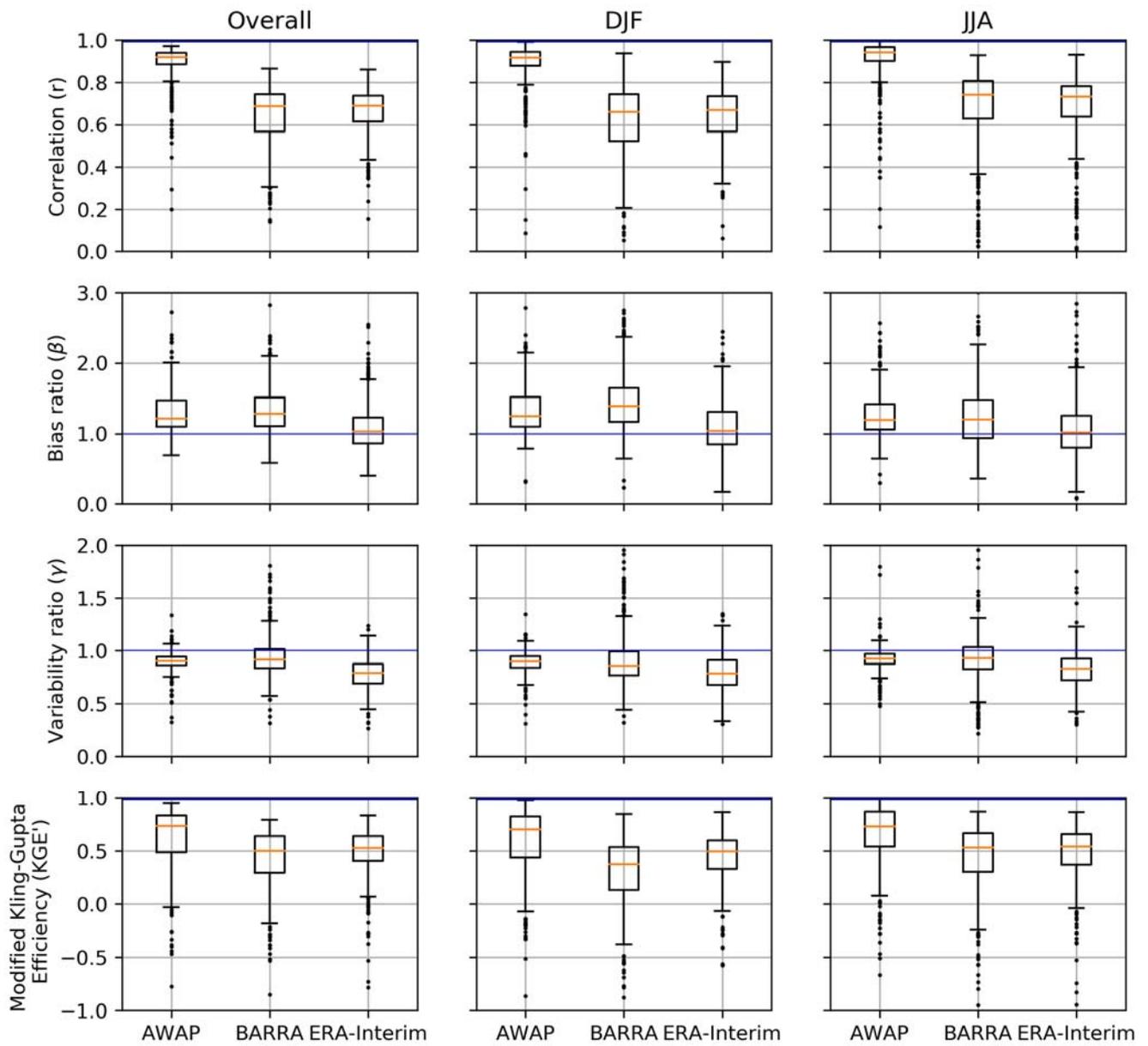
Figure 3 Boxplot of correlation (r), bias ratio (β), variability ratio (γ), and modified Kling-Gupta Efficiency (KGE') between daily precipitation estimates from gridded datasets and observations at gauge locations. Columns represent overall data, summer (DJF), and winter (JJA) respectively. Each box extends from first to third quartile, medians are marked in each box, and whisker extends to furthest point or 1.5 times the interquartile range whichever is closer. The blue horizontal line represents the best value.
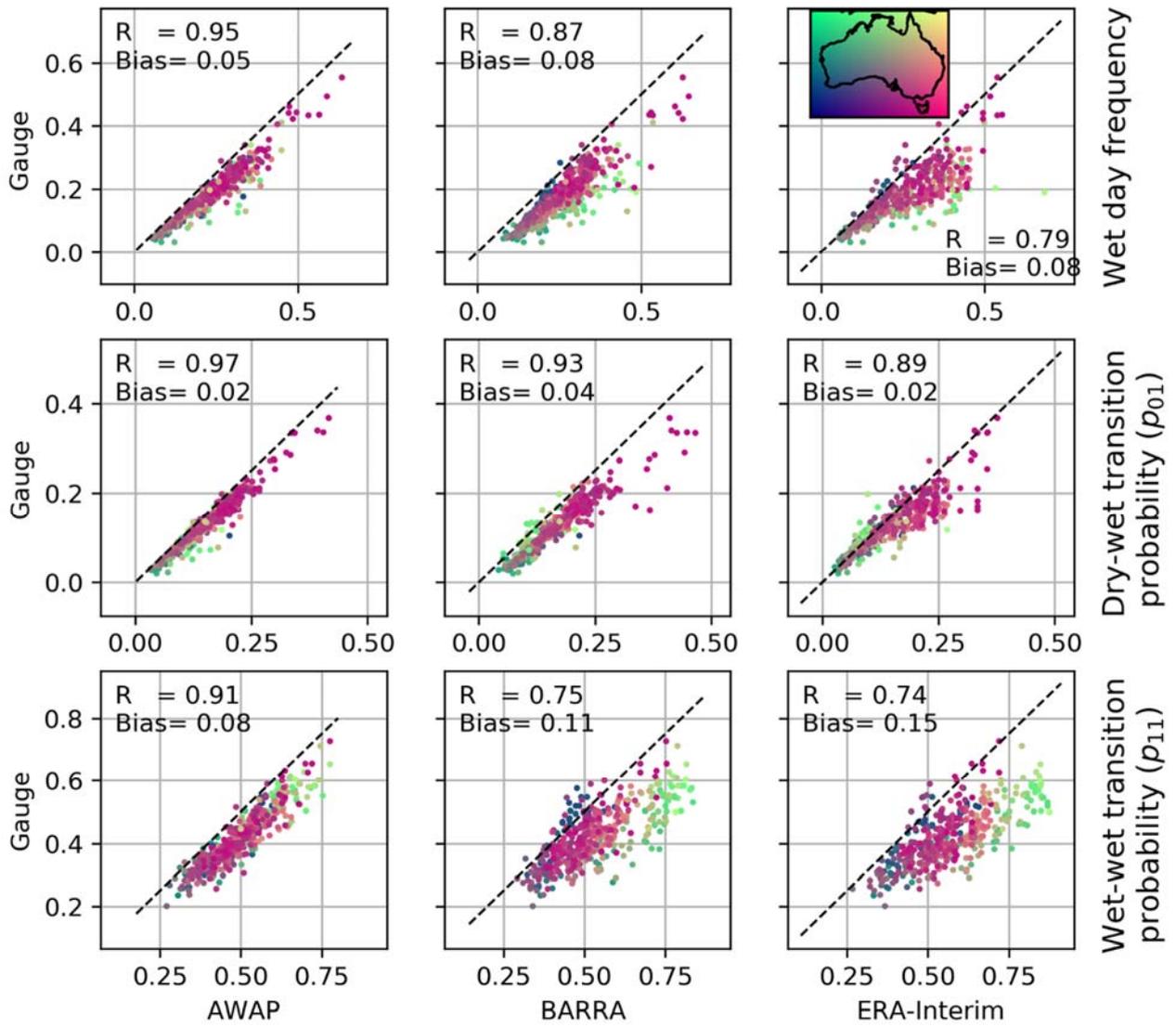
Figure 4 Wet day frequency estimates from gridded datasets (AWAP, BARRA, and ERA-Interim) and observations at gauge locations (first row). Transition probabilities: dry-wet (p_{01}, second row), and wet-wet (p_{11}, third row). Colour of scatter indicates the location of the station.
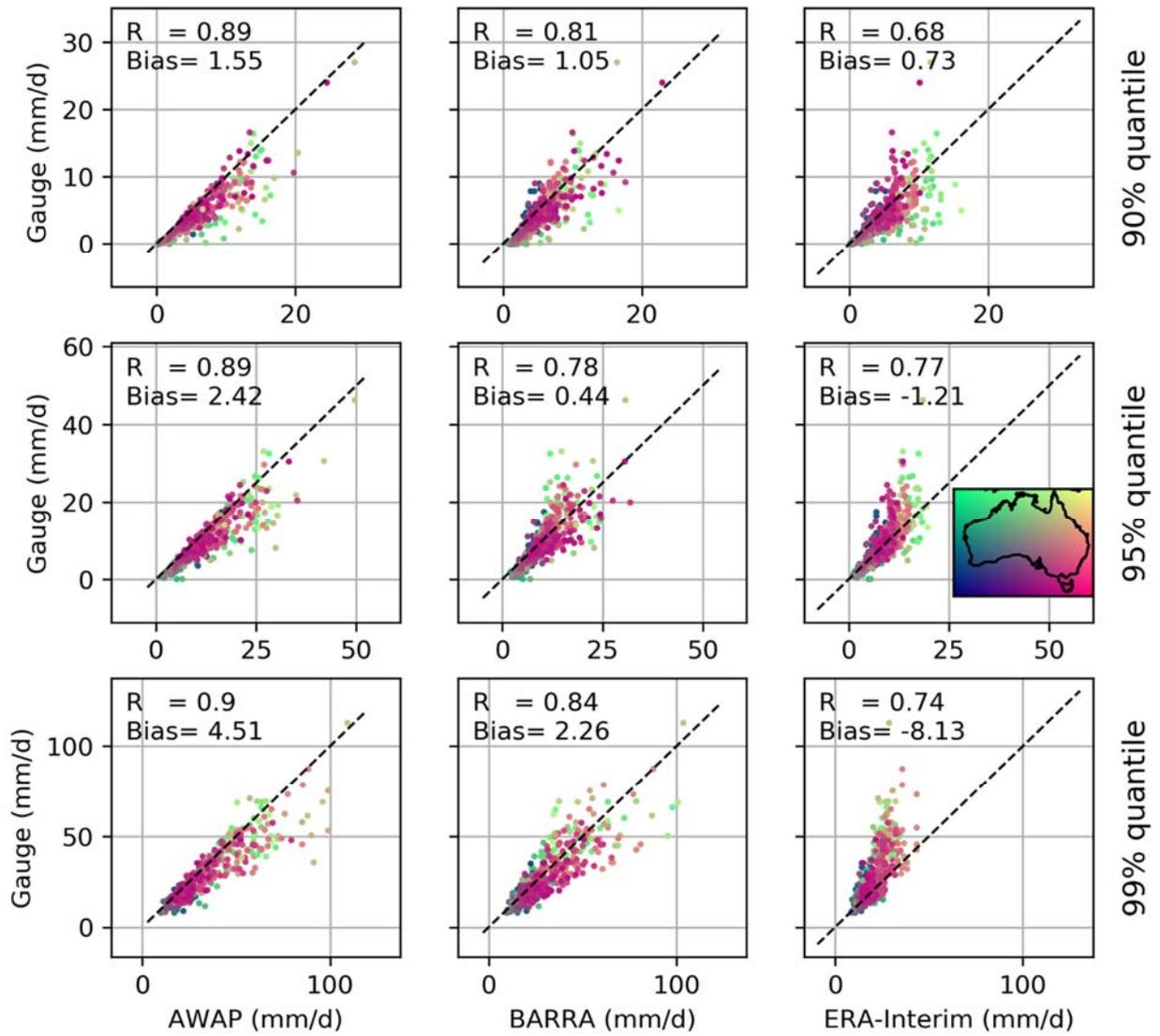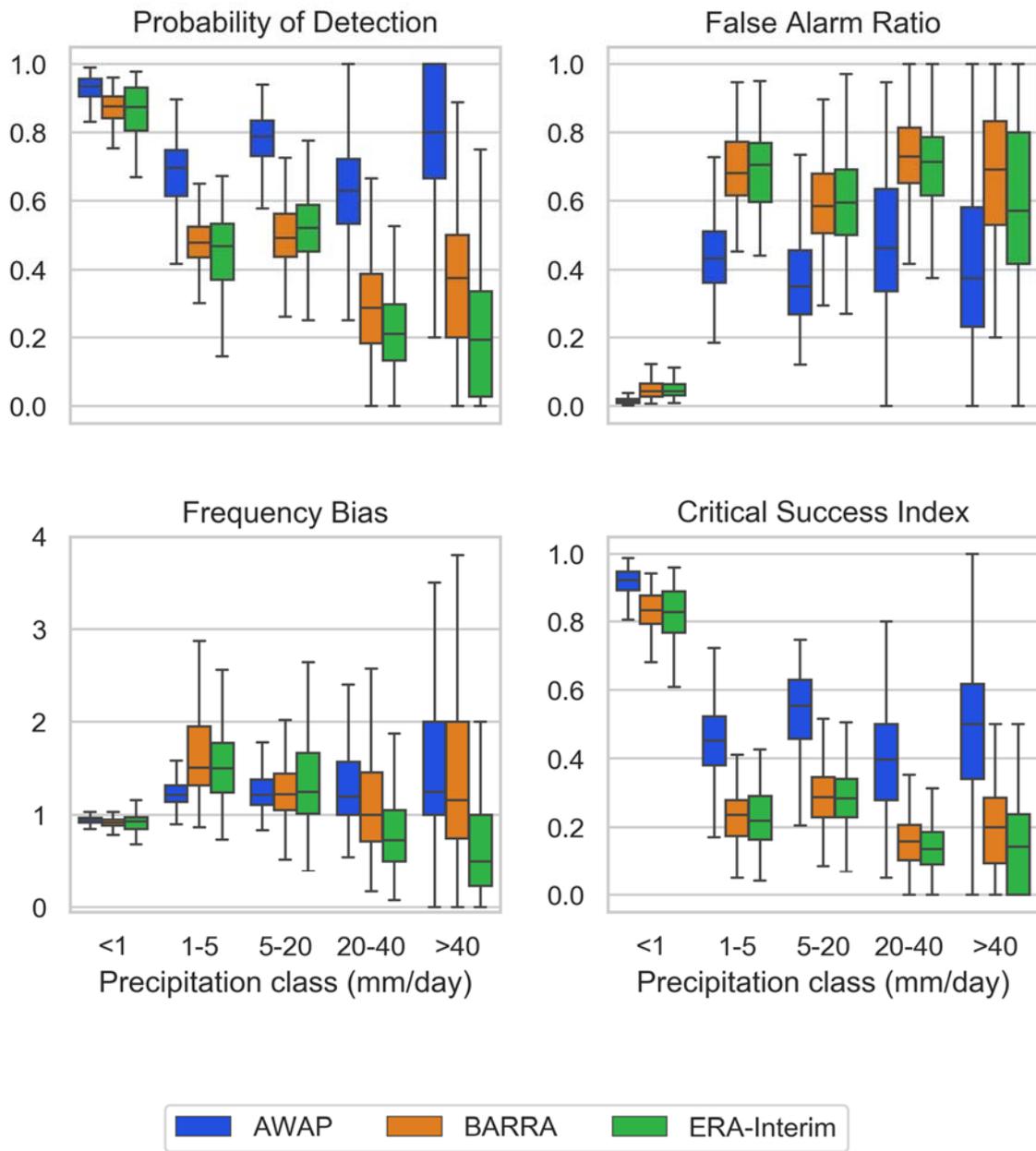
Figure 5 Quantile estimates (mm/day; 90, 95 and 99% in the top, middle and bottom row respectively) from gridded datasets (AWAP, BARRA, and ERA-Interim) and observations at gauge locations. The colour of the scatter indicates the location of the stations.

5
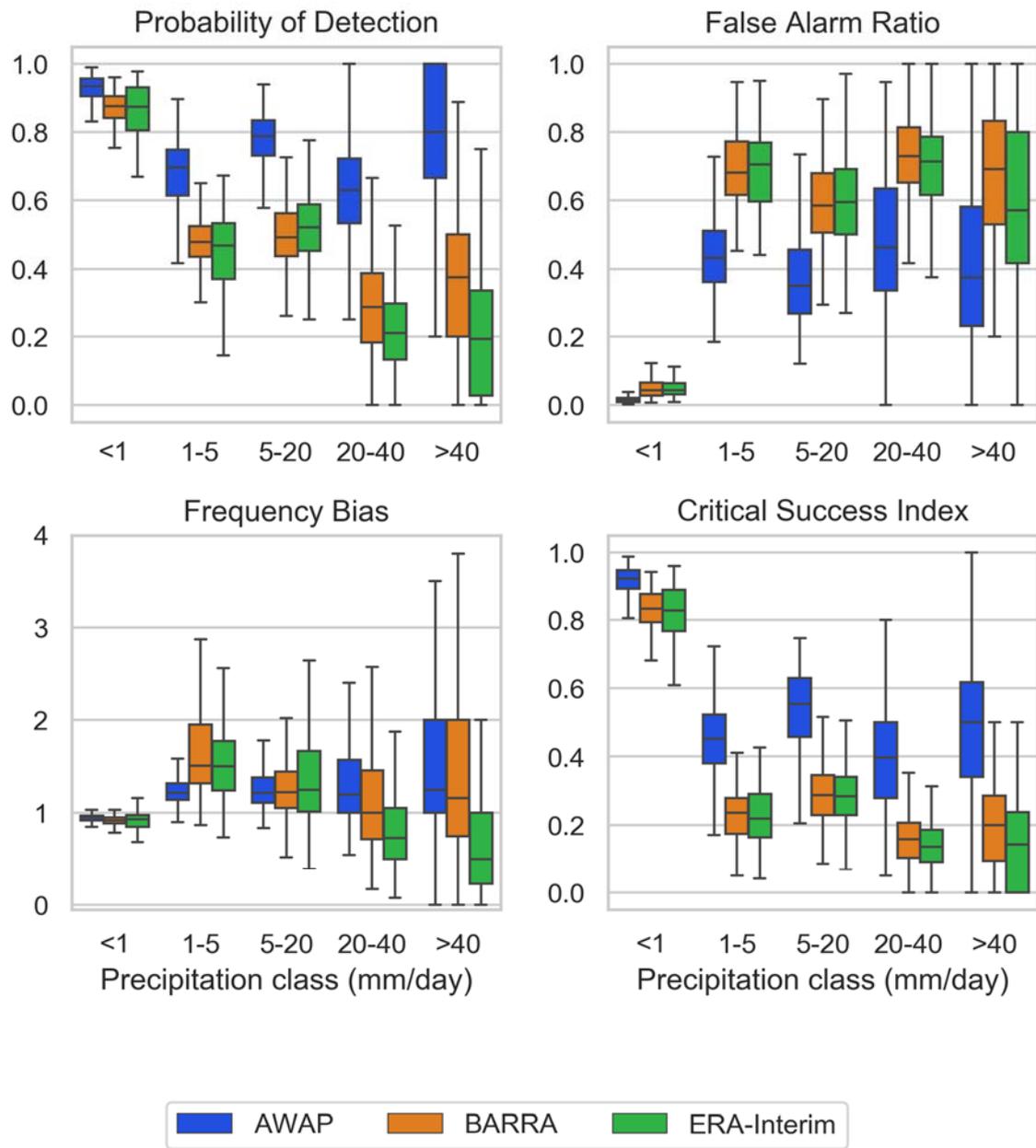
Figure 6 Boxplot of categorical performance indices (Probability of detection, false alarm ratio, frequency bias , and critical success index-) calculated against gauge data for five classes of rainfall intensity. Each box extends from first to third quartile, medians are marked in each box, and whisker extends to furthest point or 1.5 times the interquartile range whichever is closer.
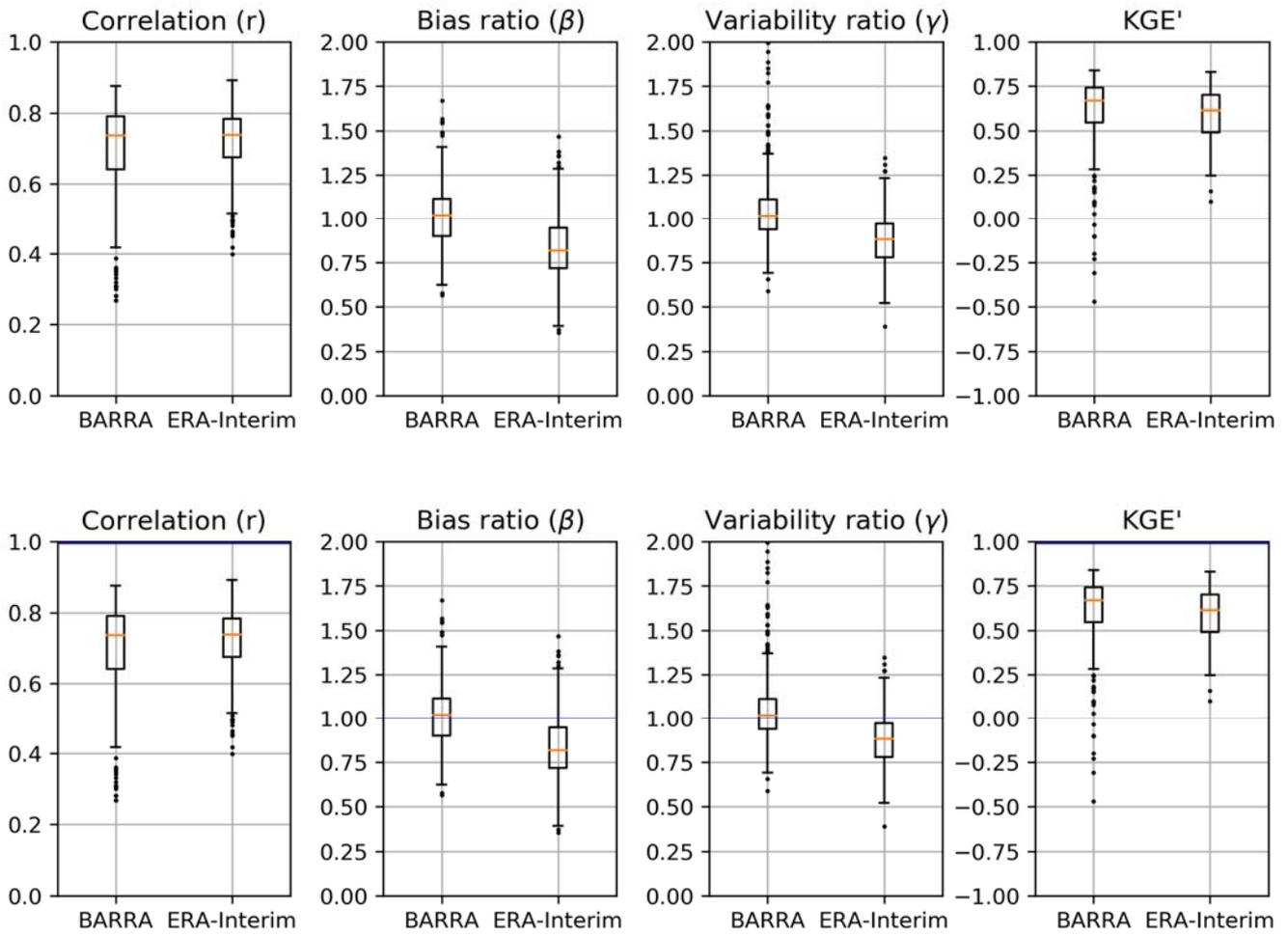
5

Figure 7 Boxplot of correlation (r), bias ratio (β), variability ratio (γ), and modified Kling-Gupta Efficiency (KGE') against AWAP dataset. Each box extends from first to third quartile, medians are marked in each box, and whisker extends to furthest point or 1.5 times the interquartile range whichever is closer.
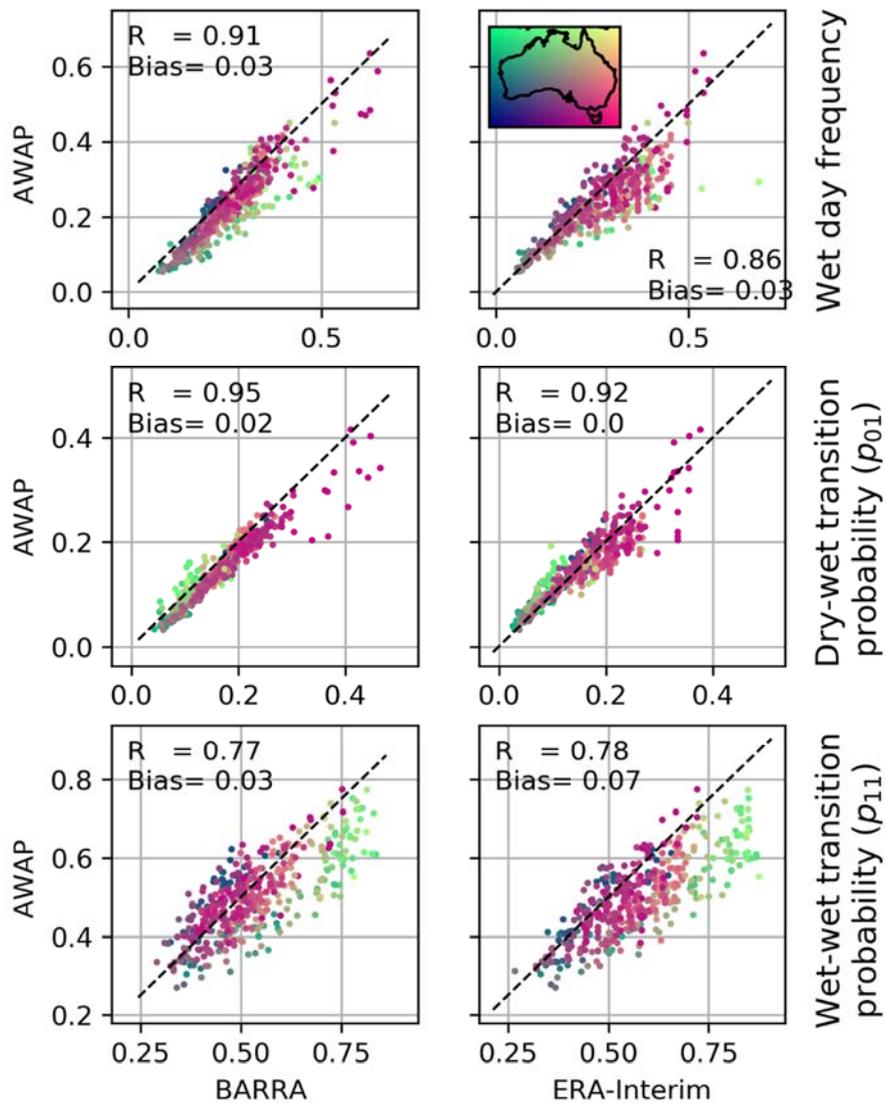
Figure 8 Wet day frequency estimates from reanalysis datasets (BARRA, and ERA-Interim) and AWAP at gauge locations (first row). Transition probabilities: dry-wet (p_{01}, second row), and wet-wet (p_{11}, third row). Colour of scatter indicates the location of the AWAP grid.
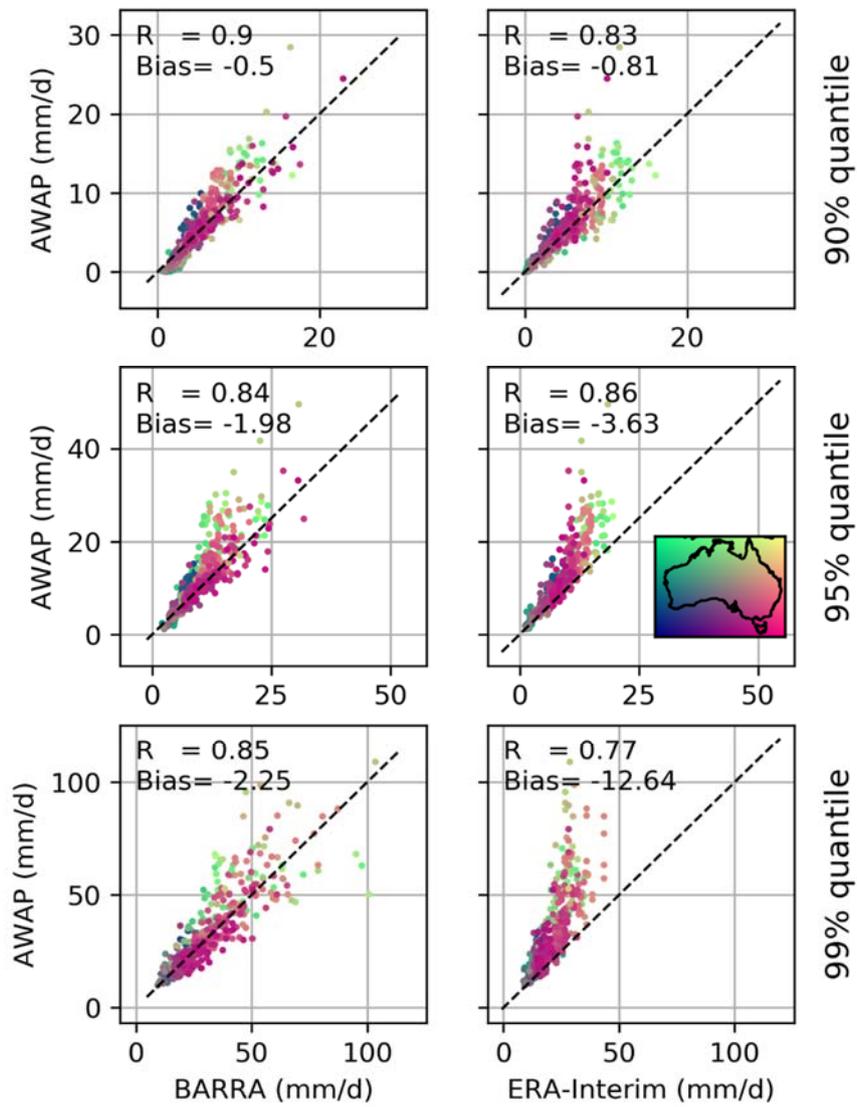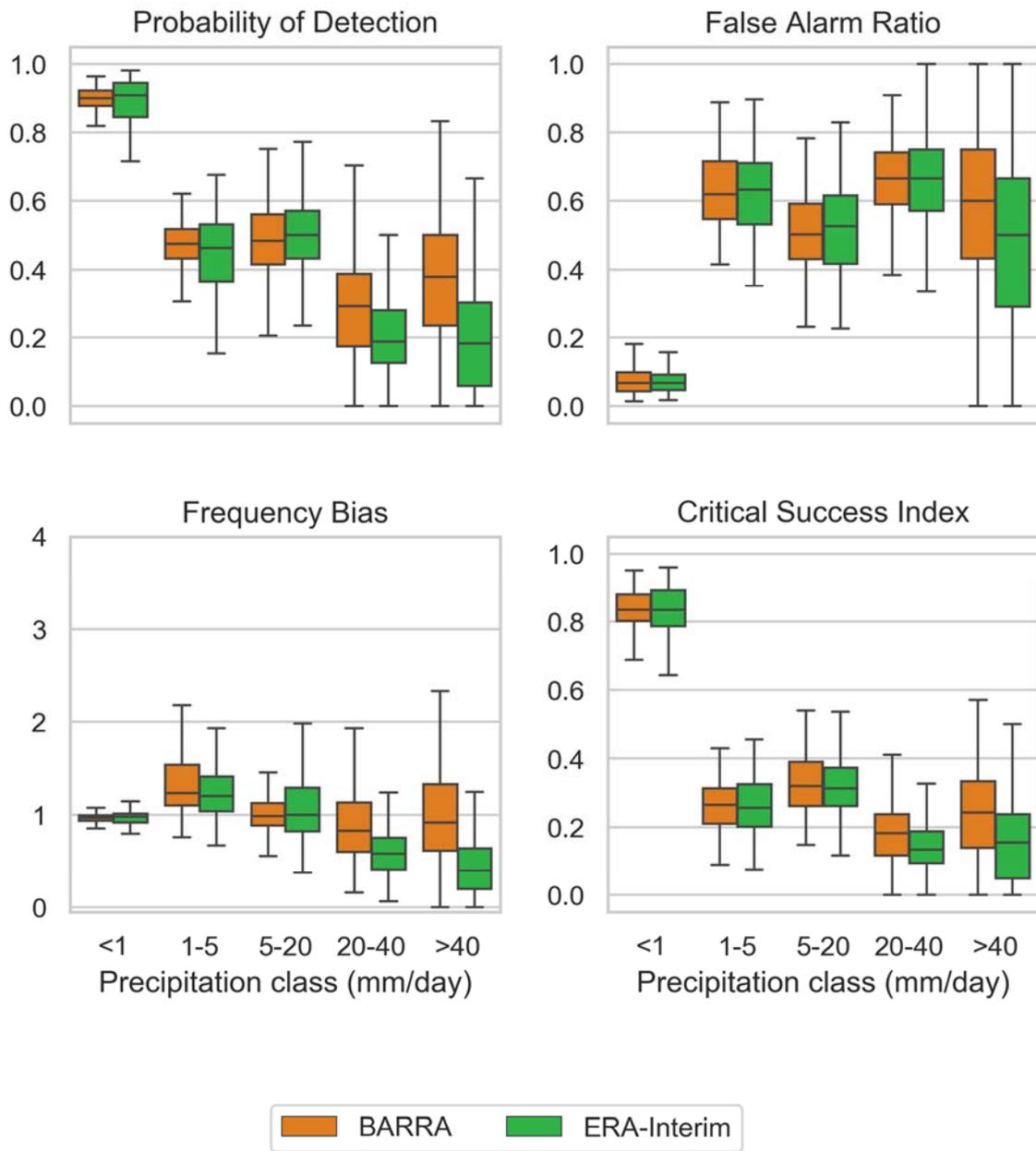
5

Figure 9 Quantile estimates (mm/day; 90, 95 and 99%, in the top, middle and bottom row respectively) from reanalysis datasets (BARRA, and ERA-Interim) and AWAP at gauge locations. The colour of the scatter indicates the location of the AWAP grid.
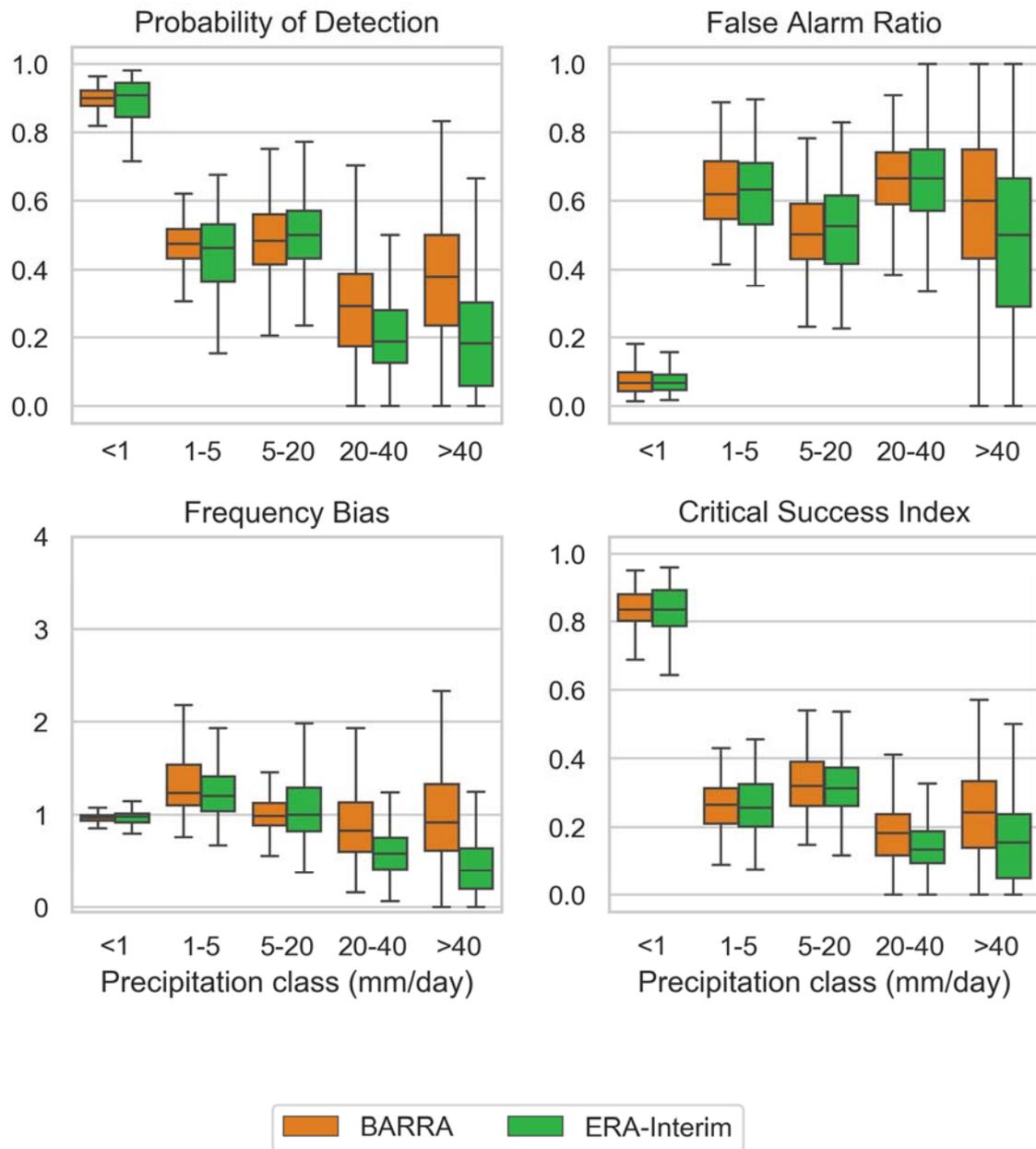
5

Figure 10 Boxplot of categorical performance indices (Probability of detection, false alarm ratio, frequency bias , and critical success index) calculated against AWAP data for five classes of rainfall intensity. Each box extends from first to third quartile, medians are marked in each box, and whisker extends to furthest point or 1.5 times the interquartile range whichever is closer.

5

Table 1 Metrics used in the evaluation of precipitation data.

Measures	Equation	Description
Pearson correlation coefficient (r)	$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$	Pearson product moment correlation measures the linear correlation of the observed and modelled values. Range [-1,1] Best value =1
Bias ratio (β)	$$\frac{\bar{y}}{\bar{x}}$$	Bias ratio measures the agreement between means of modelled and observed values. Evaluates bias in total volume over-estimation (β>1) under-estimation (β<1) Range [0, +∞] Best value =1
Variability ratio (γ)	$$\frac{CV_s}{CV_o} = \frac{s_{\bar{y}}/\bar{y}}{s_{\bar{x}}/\bar{x}}$$	Variability ratio measures the agreement between Coefficient of Variation (CV) of modelled and observed values. The use of CV in the calculation of γ ensures that the bias and variability ratio are not cross-correlated (Kling et al., 2012) Evaluates spread . over-dispersed (γ>1) or under-dispersed (γ<1) Range [0, +∞] Best value =1

Continuous metrics

Categorical metrics	Modified Kling-Gupta Efficiency (KGE')	$1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}$	Summarises correlation, bias ratio and variability ratio Range [-∞,1] Best value =1
	Probability of Detection (POD) or Hit Rate	$\dfrac{H}{H+M}$	POD measures the fraction of number of events that were correctly reported by the model. Sensitive to hits but ignores false alarms. Suitable for rare events Range [0,1] Best value =1
	False Alarm Ratio (FAR)	$\dfrac{FA}{H+FA}$	FAR measures the fraction of number of non-events that were incorrectly reported as events by the model. Sensitive to false alarm but ignores misses Range [0,1] Best value =0
	Threat Score (TS) or Critical Success Index	$\dfrac{H}{H+FA+M}$	Measures the fraction of observed events that were correctly modelled. It penalizes both misses and false alarms Range [0,1] Best value =1
	Frequency Bias (fBias)	$\dfrac{H+FA}{H+M}$	Measures the ratio of number of events modelled and observed. Range [0, +∞] Best value =1

39

Table 1 Overview of gridded precipitation datasets used in this study

Name	Details	Data Source	Spatial coverage and resolution	Temporal coverage and resolution	Reference
BARRA-R	Bureau of Meteorology Atmospheric high-resolution Regional Reanalysis for Australia (http://www.bom.gov.au/research/projects/reanalysis/)	Regional reanalysis	Australasian region (65 to 196.9° E, -65 to 19.4° N), ~12km	1990-February 2019, Hourly	(Su et al., 2019)
ERA-Interim	European Centre for Medium-range Weather Forecasts ReAnalysis Interim (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim)	Global reanalysis	Global, ~80km	1979-present, 3-hourly	(Dee et al., 2011)
AWAP	Australian Water Availability Project (http://www.csiro.au/awap/)	Gauge interpolated	Australia land area, 5km	1900 – present, Daily	(Jones et al., 2009)

5 **Table 2 Median values of KGE' and its components grouped for broad Koppen-Geiger climate categories for Australia. The number in parenthesis indicates the number of stations in the climatic zones. Value in bold represent the best score in each group**

	Tropical (50)			Arid (125)			Temperate (266)		
	AWAP	BARRA	ERA Interim	AWAP	BARRA	ERA Interim	AWAP	BARRA	ERA Interim

40

Against Gauge	Correlation (R)	**0.88**	0.42	0.55	**0.91**	0.60	0.65	**0.93**	0.72	0.71
	Bias ratio (β)	1.24	1.28	**0.97**	1.22	1.35	**1.01**	1.20	1.22	**1.06**
	Variability ratio (γ)	**0.87**	1.15	0.61	**0.90**	0.85	0.86	0.91	**0.93**	0.79
	KGE'	**0.67**	0.27	0.35	**0.72**	0.41	0.53	**0.74**	0.58	0.56
Against AWAP	Correlation (R)		0.49	**0.63**		0.66	**0.67**		**0.77**	0.76
	Bias ratio (β)		**0.99**	0.76		**1.08**	0.83		**0.99**	0.82
	Variability ratio (γ)		1.35	**0.72**		**0.95**	0.94		**1.02**	0.88
	KGE'		0.38	**0.43**		**0.60**	**0.60**		**0.71**	0.64

Table 3 Summary of performance of BARRA precipitation

Metric	Point-to-grid	Grid-to-grid
KGE' and its components	• Performance varies spatially, showing better scores in temperate regions than in the tropical and arid region • Variability is well captured	• Exhibits less bias in total rainfall than ERA-Interim • Overestimates variability in the tropics only
Quantiles	• High quantiles (90, 95 and 99%) closely represent high rainfalls and are unbiased.	• Similar pattern as for point-to-grid • Better representation of higher quantiles than ERA-Interim

Wet day frequency and transition probabilities	• Frequency of wet days and wet-wet transition probabilities are over-estimated • Dry-wet transition probabilities are well reproduced	• Similar pattern as for point-to-grid with improved correlation and reduced bias.
Categorical metrics	• Both POD and FAR increase with threshold • Frequency bias is on average higher for light/moderate than low and heavy rainfalls.	• Improvement along all metrics with respect to ERA-Interim • Light rainfalls are over-estimated • Large rainfalls are poorly captured yet are better than ERA-Interim