1. Response to general comments from Reviewer R1

This paper is an evaluation of BARRA v1.0 reanalysis. The authors compare BARRAC, a 1.5 km downscaled re-analysis against various observations, and investigate added value to BARRA-R, which is the 12 km continental version. This is a highly valuable paper and generally very well written. I recommend acceptance in GMD subject to the following revisions.

We would like to thank the reviewer for their thorough and constructive review. We respond to the comments, in turn below, which we believe has led to an improved paper.

2. Response to specific questions from Reviewer R1

1. *The paper would benefit by having an "Observational data-sets" section or similar, where all observations used are described in a bit of detail. Currently details of observations are only provided when results are shown. This would fit well in section 3 of the paper. More details are needed about the point observations and gridded analyses. Could you also provide some background on MERRA2? Why do you choose to use MERRA2 specifically? Some context is needed here. Details about AWAP also needed.*

We agree that more information can be provided for the various reference data sets. We have considered adding another section or subsection within or before Sec. 3, but find this disrupts the flow of the paper. We will add a section S1 in the Supplement to include (1) Table that summarizes the characteristics of the reference data sets (including spatiotemporal resolutions, references and parameters we used), (2) details and issues of AWAP and Rainfields2, and (3) distinctions between the global reanalyses.

In Sec. 3., we will refer readers to Sec. S1 of the Supplement, and add a comment "To increase the diversity of models used in our inter-comparison, we also include the Modern-Era Retrospective analysis for Research and Applications-2 (MERRA2, Gelaro et al., 2017) hindcasts.". We find that we need three distinctive systems to distinguish model biases in view of limitations to observational data sets; here we have BARRA-R with UM and 4DVar, ERA reanalyses with IFS and 4DVar, and MERRA2 with GEOS and 3DVar.

2. *Figure 3 – comparisons with AWAP. Interpolation errors within AWAP are available as "RMSE Analysis" from BOM's web-page. It would be useful to show or talk about this to put the biases in context of errors in AWAP?*

We agree. A section S1 will be added to the Supplement to discuss issues of AWAP. In particular for AWAP, S1 notes
"*AWAP provides gridded daily 0.05 × 0.05° analysis of station observed maximum and minimum 2 m temperature data, and raingauge-based daily accumulation of precipitation. The grids for temperature are generated using an optimized Barnes successive-correction method that applies weighted averaging to the station data. Topographical information is included by using anomalies from long-term (monthly) averages in the analysis process. By contrast, the ratio of observed rainfall to monthly average is used in the analysis process of precipitation. Readers are referred to Jones et al. (2009) for details. The root-mean-squared error (RMSD) of the daily maximum temperature analysis is around 1-1.5 K over the BARRA-C domains. The error is higher around Nullarbor Plain (northwest of the BARRA-AD domain) due to a relatively sparse network, and the Southeastern highlands (BARRA-SY domain) due to strong temperature gradients between the coast and mountains. The analysis errors are larger for daily minimum temperature than for daily maximum values, with RMSD around 1.5-2 K and larger errors in the regions. For daily precipitation analysis, the RMSD over the BARRA-C domains is relatively uniform at around 2.5 mm. Higher RMSD of 5 mm*

is noted over northwestern coastal regions of BARRA-SY domain. Further, Chubb et al. (2016) has shown that the AWAP error for wintertime precipitation over the Snowy Mountains (BARRA-SY domain) can be as high as 4.5 mm, due to the lack of gauges and steep topography exposed to prevailing winds. At high elevations where frozen precipitation is challenging to measure, AWAP analysis has underestimated the total precipitation amount by more than 10%. Therefore, the comparisons with AWAP for the Southeastern Highlands and the Nullarbor Plain need to be interpreted in view of these limitations. King et al. (2012) has also found AWAP tends to report lower extreme rainfall estimates (e.g., climatological 95th percentile rainfall) than those observed at stations, which is characteristic of an interpolated product."
This may partly explain the warm bias observed over the Nullarbor Plain for daily maximum temperature in Figure 3. Apart from this, it is difficult to definitively attribute biases in various reanalyses to error patterns of AWAP.

3. *Section 3.2, Figure 3 – you do not discuss the very large bias in Daily max temp in South Australia, which is up to 7.6 degrees C, in the north-west part of the domain. This is a very large bias in both BARRA-C and BARRA-R, and this needs some more attention. MERRA2 has a similar bias in south Australia to BARRA-C and BARRA-R.*

We agree. BARRA, MERRA2 and ERA5 all show warm bias, with respect to AWAP, over the Nullarbor Plain. This region has very few observing stations. According to http://www.bom.gov.au/climate/data/, there are only around 5 stations reporting daily maximum temperature over this 2x2 degree region. Differences in the land cover classification between BARRA and ERA reanalyses may account for some of their differences; BARRA is based on IGBP while ERA is likely based on CCI land cover. The differences in the land cover classification between BARRA and ERA5 appear to contribute to the differences in temperature bias seen over the salt lakes in the AD domain. To investigate this further, work is in progress to map CCI raw data to the local land cover types used by UM; the current transform table has led to large differences between IGBP and CCI in terms of shrub and bare soil covers in this region. These comments will be added to Sec. 3.2, and with reference to the Supplement where the quality of AWAP is noted.

4. *Figures 3, 4 and 7 – Some analysis of trend would also be useful? Rather than just means. Also, what about variability? It may be useful to examine the standard deviation from AWAP versus the models?*

We find that the trends are only apparent in few models and for few domains (noted in text), and are generally not systematic across time period, across the models and domains to warrant trend analysis. Further, comments on possible trend of the temperature max/min bias for these cases are made in Sec. 3.2 and 3.4.

We will extend the bias analysis to compare their standard deviation of daily min/max temperature in each month and present the results in the Supplement. Section 3.2 will be extended with the following observations of this additional analysis:
"*This analysis of variability of bias is also repeated for the standard deviation of the modelled temperature and AWAP in Figure S4 and S5 of the Supplement. BARRA-C shows a slightly wider dispersion of daily maximum temperature than AWAP (by 0.4 K) and BARRA-R (by 0.1 K), with the exception for the TA domain. For BARRA-TA, the standard deviation of BARRA is similar to AWAP and is higher than the global reanalyses. For daily minimum temperature, both BARRA are similar and they are generally under-dispersed by 0.3 K compared to AWAP.*

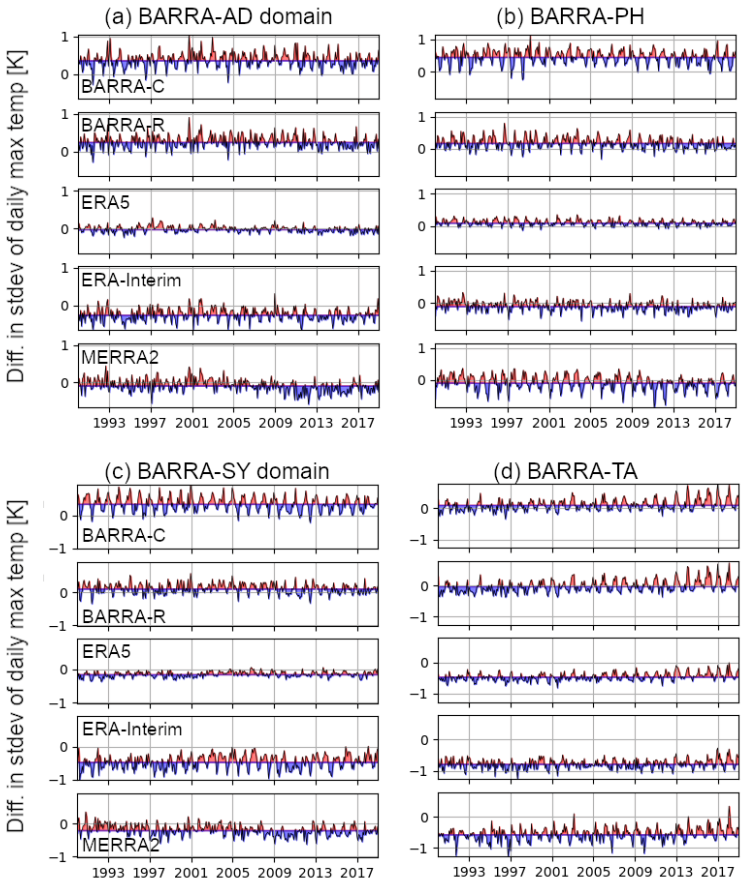

Figure S4: As with Figure 4, but for monthly difference in standard deviation of daily maximum temperature over various BARRA-C domains, with respect to AWAP. The timeseries are shaded around their individual 1990-2018 means.

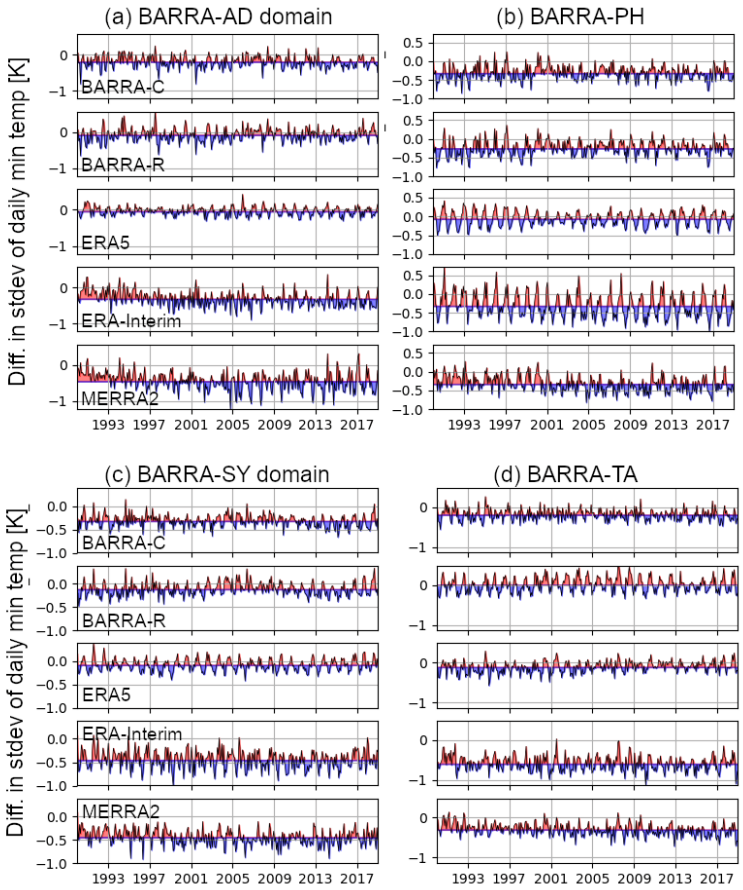

Figure S5: As with Figure S4, but for daily minimum temperature.

”

5. *Figure 4 – this is a very busy figure, and I urge the authors to find better ways to summarize/plot the data. It is very hard to compare with the different y-axis limits. I suggest to have fewer plots, 1 plot per domain, for tmax and tmin separately. Don't plot the difference, but plot AWAP as solid black line, and each model for that domain as a different color/marker. This is currently too much to digest. By plotting on the same plot, with similar axis, one can actually makes sense of all of this. This Figure is much too busy to digest. Reduce the number of plots, it is very common to plot up to 4-5 lines on a single plot.*

Figure 4 aims to demonstrate the seasonal and inter-annual variations in the model biases (w.r.t. AWAP). As the bias is in the order of 1-2K, revising the plots as per the reviewer's suggestion will not capture this. We agree that the figure is currently small and busy but is so in order to capture differences between the domains, different reanalyses, and between daily maximum and minimum temperatures. For the readability of the subfigures, we will split the figure into two sets of subfigures, one for daily maximum temperature (Figure 4), and another for daily minimum temperature (Figure 5).

6. *Figure 5 – could you make the models dotted lines do they are easier to distinguish from AWAP in solid blue. Line 270 – BARRA-R and BARAA-C have too many warm days in Perth, but also, all other models seem to have too few, or close to AWAP. For Adelaide in SA, it seems that BARRA-C does worst than BARRA-R for warm extremes?*

We will improve Figure 5 by changing the plotting line styles. In particular, we expand this with the added value analysis (see reply to comment R1#8) to look at warm extremes across the domains, not limiting to few selected points. We move Figure 5 and this discussion to the Supplement.

7. *Figure 8 and related text – same comment as for Figure 4.*

See our reply to R1#4.

8. *Can you come up with some objective measure of added value of BARRA-C over BARRA-R? There are many metrics used to quantify added value. There is a lot of literature on quantifying added value of downscaling GCMs using RCMs for example. The same concepts could be applied to quantify added value of BARRA-C over BARRA-R. Producing re-analysis at 1.5 km resolution, takes a lot of effort and the data storage is difficult, as I am sure the authors would know very well. Quantifying the added value would be useful I think.*

In response to this comment, we report an added value analysis in a new Sec. 3.6, which reads,
“

3.6 Added value analysis for temperature and rainfall extremes

We apply an approach similar to Di Luca et al. (2015) to quantify the added value (AV) in the representation of climatological extremes from BARRA-C by comparing the skill between the BARRA-C and BARRA-R. The warm extremes of daily maximum temperature, the cold extremes of daily minimum temperature and the wet extremes of daily precipitation are assessed against AWAP, noting that the true AV from BARRA-C at its native resolution is not fully determined here. The statistics for extremes (X) are given by the percentiles of the daily temperature and precipitation values over the 29-year time period. We use AV_d=[d(X_(BARRA-R),X_AWAP)-d(X_(BARRA-C),X_AWAP)]/[d(X_(BARRA-R),X_AWAP)+d(X_(BARRA-C),X_AWAP)] of Di Luca et al. (2016) where d defines a distance metric between the model-derived and AWAP-derived statistics computed across the grid cells. To capture both the total errors and spatial patterns of the statistics, we let d=MSE(A,B)=E[(A-B)^2] defining the mean squared error and d=Corr(A,B)=1-R(A,B), where E(∘) is the expectation operator and R their Pearson’s correlation. Larger positive AV values suggest smaller errors in BARRA-C than in BARRA-R and thus substantial added value by the downscaling of BARRA-R.

Figure 11 plots AVs for different BARRA-C domains, showing that AV is not gained consistently across the percentiles, variables and domains. For warm extremes of daily maximum temperature, BARRA-C shows positive AVMSE over BARRA-R in the TA and AD domains. However, there are low or negative AVMSE for AD, PH and SY (inland region) mainly due to the warm bias in BARRA-C, also seen in Figure 3(c) and 6(a,b). With positive AVCorr, BARRA-C captures the spatial patterns of the warm extremes across the domains, particularly over the coastal and high topography regions (Figure S6 of the Supplement). For cold extremes in Figure 11(b), BARRA-C still shows positive AVMSE over all but the SY domain, due to AV over the coastal regions. The negative AVMSE in SY is related to warmer cold extremes, particularly over the Great Dividing Range. Positive AVCorr is seen in TA but not in the other domains, although it should be noted that the BARRAs are generally strongly correlated with AWAP with R mostly between 0.7 to 0.9.

AV from BARRA-C for wet extremes of precipitation relates more to the spatial patterns of the extremes (Figure 11(c)). Given the tendency of BARRA-C to overestimate heavy rainfall, the wet bias relative to AWAP, particularly over the PH domains (Figure 7(b)), is responsible for the low AVMSE. For the SY domain, positive AVCorr for precipitation agrees with the above FSS analysis, which somewhat avoids the issue of bias through percentile-based thresholding. Furthermore, AV¬Corr is positive for extreme rainfall for all but the AD domain indicating that despite the wet bias, rainfall extremes in BARRA-C can be better spatially correlated with AWAP. Assessing AV for wet extremes may also be problematic with AWAP. As an interpolated dataset, AWAP tends to underestimate the intensity of extreme heavy rainfall observed at stations and the issue is more

pronounced at locations with sparse observational sampling or high topography, particularly in SY and TA (Chubb et al. 2016; King et al., 2012).

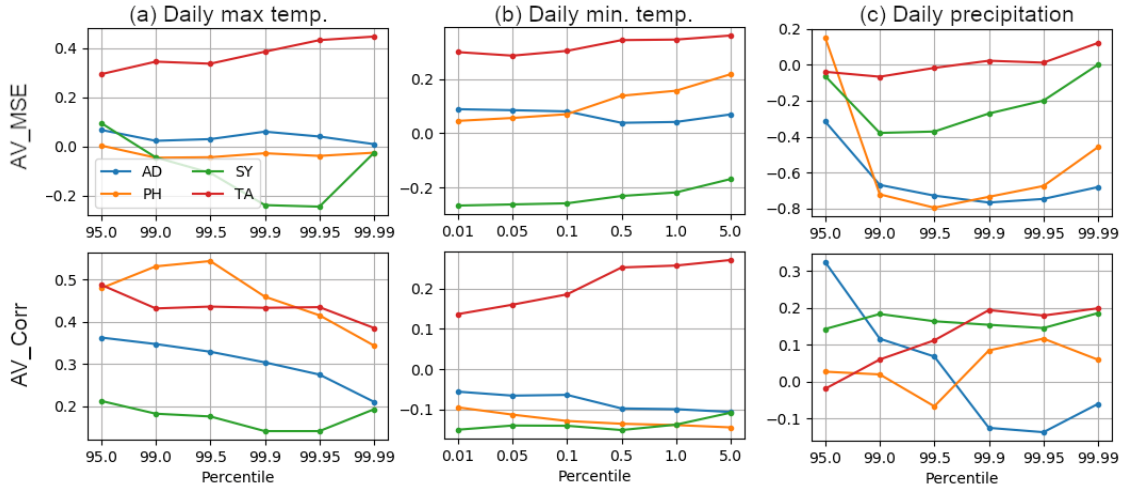

Figure 11: Added value (AV) analysis of the (a) warm extreme of daily maximum temperature, (b) cold extreme of daily minimum temperature, and (c) wet extreme of daily precipitation, performed for different BARRA-C domains.

“

9. *The discussion and conclusion mostly speculates for reasons to explain model biases, that is to be expected. But the paper would be much more interesting if some dynamical analysis were carried out to better understand differences between the models. This could be comparing MSLP patters during very hot extreme events, just as one example. I think this would make the paper much more interesting. I would like the authors to think a bit more about this. The paper would benefit from some more actual analysis of model dynamics. You cannot do everything, I understand, but there is room for the basic dynamics analysis I think.*

This paper is submitted to Geoscientific Model Development journal as a model experiment description paper. It is outside the scope of this work to provide detailed analysis of model dynamics. We have provided references, namely Bush et al. (2020, regional configuration of the Unified Model), Champion and Hodges (2014, model spin up issues), Lean et al. (2008) and Hanley et al. (2016, convection) for some of the model issues also observed in this work.