September 27, 2024

Paper #GMD-2024-87 | Model experiment description paper**: 'Design, evaluation and future projections of the NARCliM2.0 CORDEX-CMIP6 Australasia regional climate ensemble'**

Author Comments (ACs) – Referee 2

Table 2. Anonymous Referee 2 (RC2) Comments

#	Issue Description	Discussion	Revision (in re-submitted manuscript)
	Referee #2: General Comments		
1	The authors perform extensive testing of WRF physics schemes for future regional climate projections over SE Australia. Impressively, the model is run at 4km convective permitting resolution. After choosing operational configurations, the authors document the historical biases and future projections. While the analysis is rather simple, it is very helpful that comparisons are made against previous generations of NARCLIM. I think this will form a very important foundational paper. I suggest major revisions based on my comments below, which mostly relate to clarifying important points and improving the presentation and interpretation of results.	We thank the reviewer for reviewing our manuscript and for their constructive comments on our work, including their view that this will form a very important foundational paper. Please see our responses to the reviewer's comments in this table.	Please see our point-by-point responses in this table.
	Referee #2: Specific comments		
2	The authors highlight that NarCLIM2 has large improvements in tasmax biases, with small absolute biases of ~0.5K over many regions. Are these biases also evident when downscaling all individual GCMs, or simply in the ensemble mean? This relates to the order of operations of	The reviewer is asking whether the ensemble mean is made from some models with positive bias and some models with negative bias so in the ensemble mean these biases somewhat cancel out. The answer is yes this is what happens with a reasonably good ensemble	We had stated that NARCliM2.0 shows significant improvement in tasmax biases, primarily based on the ensemble mean. However, this improvement is also evident in several of the individual simulations, though there are exceptions. In NARCliM1.0 and 1.5, most

	where the bias is computed (i.e. before or after the multi-model mean is computed). My concern is that there may be cancelling of biases (e.g. if one downscaled model has a warm bias and the other a cold bias). Can the authors confirm that this is not simply cancelling of biases? Related to this, showing biases for each downscaled model (perhaps in Supplementary material) would help to confirm this.	and indicates that the observations fall within the spread of the ensemble. Results for individual models are provided in the actual revised manuscript. The overall magnitude of the individual biases within the ensemble were smaller in N2.0 compared to N1.x -- though there were some exceptions to that for some N2.0 individual models -- please see revised text in column right now included in the revised manuscript.	simulations exhibited strong systematic cold biases. In contrast, for several ensemble members, NARCliM2.0 reduces these cold biases or replaces them with small warm biases. Overall, individual simulations in NARCliM2.0 generally show a reduction in bias compared to those in NARCliM1.0 and 1.5. This is shown for the individual simulations in the Supporting Information Figures S4-S6 for tasmax. Equivalent plots for tasmin (for which NARCliM 2.0 does not show improved performance versus NARCliM 1.x) are shown in Figures S8-S10, and for precipitation in Figures S12-14. To make this clearer in the revised manuscript, we state the range of per-RCM biases for each variable in the revised main text, and we also highlight RCMs that are in some way exceptions e.g.: “Overall, NARCliM 2.0 RCMs simulate maximum temperature more accurately than NARCliM1.x, with widespread, statistically significant reductions in cold biases in the ensemble mean (Figure 9), as well as for many individual RCMs (Supporting Information Figure S4-S6). These reductions in bias apply for all timescales but are largest for the annual mean, i.e. the area-averaged mean absolute bias is 0.75K (range: 0.61 to 2.03 K) for the NARCliM 2.0 ensemble, 1.73 K (range: 1.1 to 2.37 K) for NARCliM 1.5, and 1.89 K (range: 0.55 to 4.12 K) for NARCliM 1.0 (Figure 9d,g,j). Notably, the ensemble mean annual mean maximum temperature bias magnitudes are very small, i.e. around <0.5 K,

			over south-west WA, southern coastal regions, and several eastern regions. This may be important from a climate change adaptation and mitigation perspective as these regions are heavily populated and economically significant. NARCliM 2.0 retains warm biases of similar magnitude to NARCliM 1.5 along the north-west coast of Australia (Figure 9d,g). Moreover, these warm biases cover additional areas for NARCliM 2.0, especially during DJF (Figure 9e,h). Notably, a wide range of bias signs are evident for the individual NARCliM 2.0 ensemble members (Figures S4-S6) and a minority of NARCliM 2.0 RCMs retain strong cold biases, i.e at an annual timescale NARCliM 2.0-NorESM2-MM R3 (mean absolute bias = 2.03 K) and UKESM-1-0-LL (1.77 K)."
3	Some discussion of observational uncertainty seems warranted, especially if model biases are truly approaching 0.5K.	We agree, it is a good point to raise. The discussion in the panel right is now included on observational uncertainty, which is added at the end of section 4.1 in the revised manuscript.	Revised manuscript now states the text shown below (added to the end of Sect. 4.1). "We have evaluated NARCliM RCM skill via comparison with AGCD observations. Whilst AGCD are a high quality gridded observational data set, like any set of observations, they contain errors and uncertainties. Consequently, the outcomes of our evaluations depend on both the models being evaluated and the AGCD observational dataset. This is clearly a broader issue that applies to any model evaluation versus observations. Uncertainties in AGCD for temperature and precipitation arise from sparse station coverage in some locations, especially in remote areas, and interpolation errors in generating gridded data. More specifically,

			temperature uncertainties include urban heat island effects, inhomogeneities in observation records, and elevation differences. Precipitation uncertainties involve underestimation of extremes, rain gauge measurement errors, and challenges in representing complex terrain. For our purposes, the question of how much of a bias of ~0.5 K is due to the model errors versus the observational uncertainty cannot be currently quantified, because the models are evaluated against this single observational dataset. This leaves the observational uncertainty as implicitly included in our results. In the future observational uncertainty could be explicitly considered using a method like the Observation Range Adjusted (ORA) statistics (Evans and Imran, 2024)."
4	The text and figures swap between K and Celsius units, best to choose one.	Thanks for pointing this out. We have made changes in the text and to the figures to keep the unit consistent as K throughout.	Temperature units are now K throughout the revised manuscript.
5	Obviously a large effort has gone into producing the convection-permitting resolution model output. However, the improvements are mostly seen in temperature and not in precipitation. Perhaps this is because the focus here is on evaluating mean precipitation and not extremes? Can the authors comment further on this? Referring and discussing other international literature here would be useful also.	In this study, the scope was to focus on an initial 'first-order' evaluation of mean precipitation rather than extremes of precipitation. However, clearly much valuable research can now be undertaken into evaluating the skill of NARCliM2.0 in simulating extreme precipitation, subdaily precipitation, etc, using NARCliM 2.0 20 km and 4 km data, especially since these data are now publicly available. A great avenue for further research is to assess the potential value-add in simulating extreme and subdaily precipitation at convection permitting scale versus the convection-parameterised 20 km	Text added to the revised manuscript as per column left / shown below: "In this study, the scope was to focus on an initial "first-order" evaluation of mean precipitation rather than extremes of precipitation. However, clearly much valuable research can now be undertaken into evaluating the skill of NARCliM2.0 in simulating extreme precipitation, subdaily precipitation, etc, using NARCliM 2.0 20 km and 4 km data, noting these data are now publicly available. A good avenue for further research is to assess the potential added value in simulating extreme and subdaily precipitation at

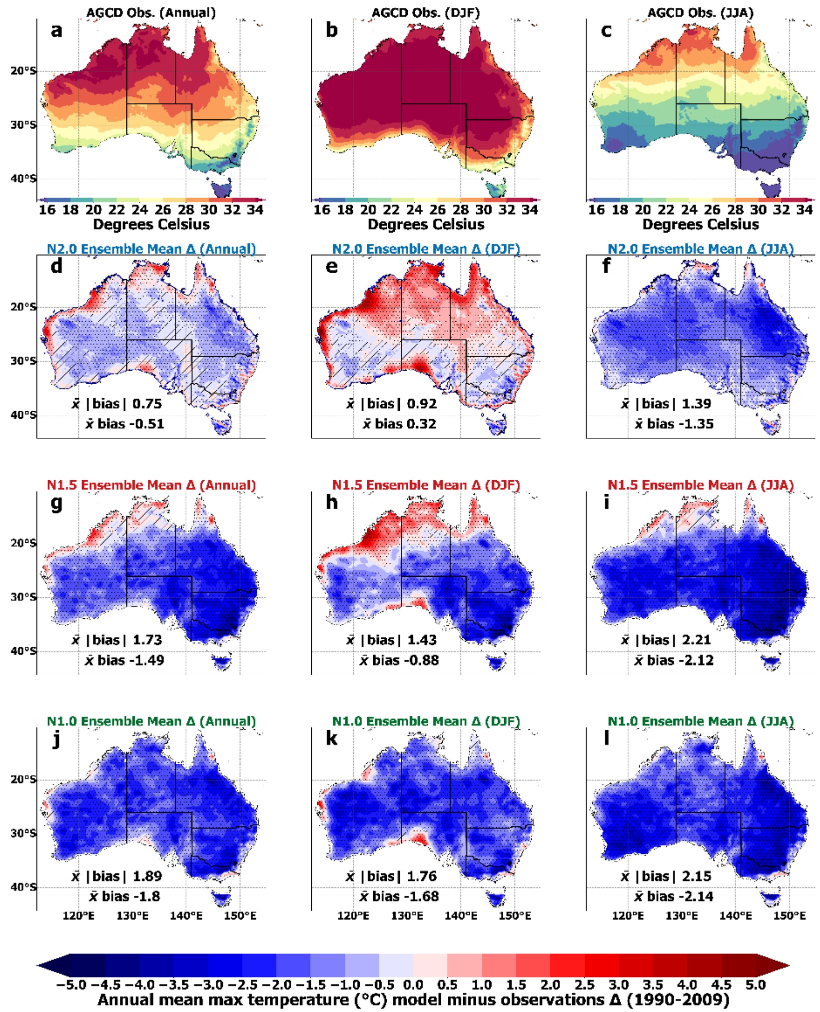
		data. This is now stated in the revised manuscript. In term of previous works: multiple studies have confirmed that convection-permitting resolution model can improve simulating daily and sub-daily rainfall extremes (Xie et al., 2024; Cannon and Innocenti, 2019; Kendon et al., 2017). In future work, we will also assess added value of convection-permitting resolution model in simulating precipitation related extremes. Xie, K., Li, L., Chen, H., Mayer, S., Dobler, A., Xu, C.-Y., and Gokturk, O. M.: Enhanced Evaluation of Sub-daily and Daily Extreme Precipitation in Norway from Convection-Permitting Models at Regional and Local Scales, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2024-68, in review, 2024. Cannon, A. J. and Innocenti, S.: Projected intensification of sub-daily and daily rainfall extremes in convection-permitting climate model simulations over North America: implications for future intensity–duration–frequency curves, Nat. Hazards Earth Syst. Sci., 19, 421–440, https://doi.org/10.5194/nhess-19-421-2019, 2019. Kendon, E. J., and Coauthors, 2017: Do Convection-Permitting Regional Climate Models Improve Projections of Future	convection permitting scale versus the convection-parameterised 20 km data. Several previous studies have confirmed that convection-permitting resolution model can improve simulating daily and sub-daily rainfall extremes (Xie et al., 2024; Cannon and Innocenti, 2019; Kendon et al., 2017).”

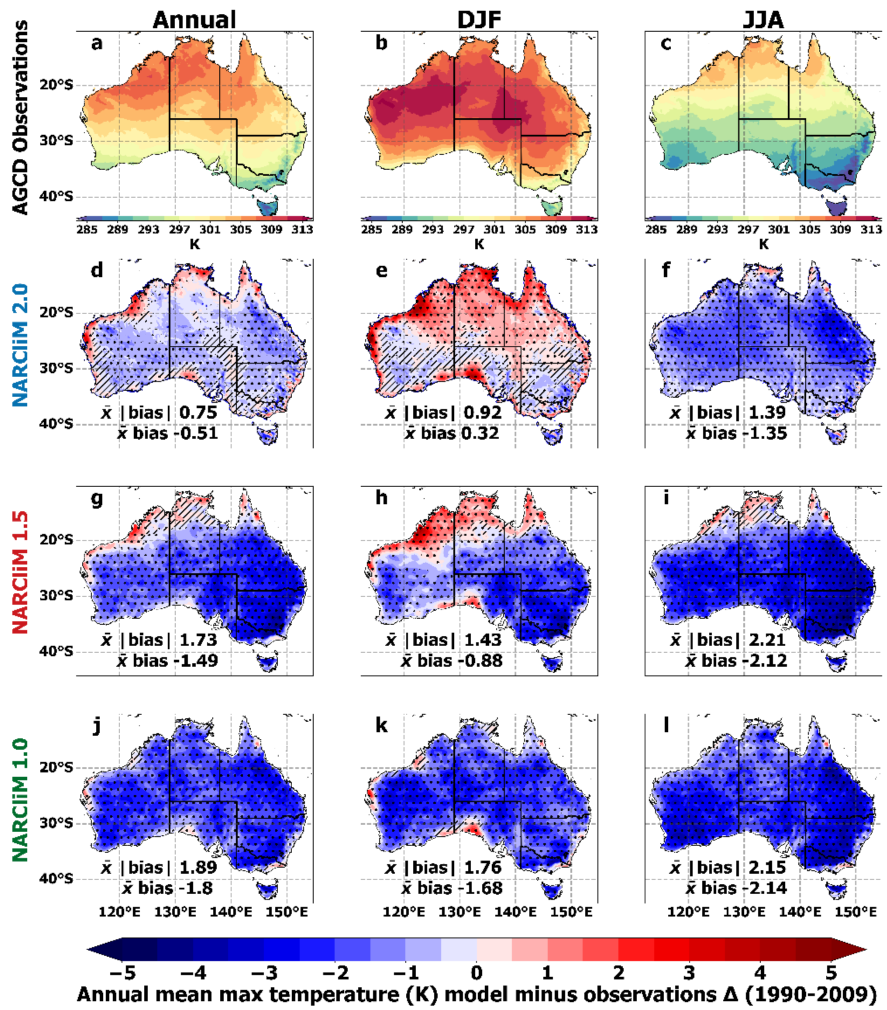
		Precipitation Change?. Bull. Amer. Meteor. Soc., 98, 79–93, https://doi.org/10.1175/BAMS-D-15-0004.1.	
6	On statistical significance. My personal view is that statistical significance is generally misunderstood and misinterpreted in climate science. However, I do think using significance in terms of model agreement is much more defensible (as you have done on top of this). If statistical significance is used, the authors also need to account for multiple testing (e.g. via the false discovery rate), which does not appear to be done: https://journals.ametsoc.org/view/journals/bams/97/12/bams-d-15-00267.1.xml?tab_body=abstract-display	Thank you for your suggestion and the reference you have posted is interesting and something we have applied in the revised version of this manuscript and will continue to apply going forwards. The ensemble mean based plots (Figures 9-14 and panels a, l and s in Figure 15) are the only plots where we combine multiple collections of null hypotheses. For these Figures 9-14 (and panels a, l, and s in figure 15) we have included revised plots with a corrected criterion using Walker's test using Eq.2 from the reference you provided. We applied Walker's test as this is stricter than FDR and easier to implement at this stage. Using this revised method, dependent on the NARCliM ensemble in question, alpha values change from 0.05 to alpha = 0.0051162 (for example). We found no major visible changes to the significance results / significance stippling of our plots for temperature biases and future projections, as can be observed in the comparison of **original versus revised figure versions shown below this table**. Here, the results are similar between the original version and the revised version implementing your suggestion, e.g. temperature climate change signals show widespread significant future changes.	Reviewer's suggestion implemented and Figures 9-15 revised in the revised manuscript as described in column left, and section 4. Evaluation methods in the revised main text now states the additional text below; results/figures in question revised throughout as indicated in column left (please also see example figures below): "Significance thresholds were adjusted to account for multiple testing using Walker's test (Eq.2 in Wilks, 2016)."

		Before implementing the reviewer's suggestion, the original results for precipitation climate change signals tended to be non-significant over most regions for most models. Having implemented the reviewer's suggestion, there are fewer locations showing statistically significant future changes for mean precipitation (see comparison figures below).	
7	In Figure 15, is there an understanding of why the projections for ACCESS-ESM1-5 projections are so dry? Presumably this is in the GCM also? Do we know why that is from the physical perspective?	ACCESS-ESM1-5 driven RCM simulations project very dry futures for Australia, which is mostly inherited from the GCM. There are 40 realisations for ACCESS-ESM1-5, but only realisation 6 provides sub-daily outputs that can be used in dynamical downscaling using WRF. This realisation simulates a particularly dry projection over Australia, especially for eastern Australia, making it a useful "stress test" case. It also shows that internal variability within the GCM is a factor in producing this dry projection. Please see more details in: https://research.csiro.au/access/model-ensembles-to-understand-climate-variability-and-change/ In terms of GCM skill versus observations, globally, this GCM is dry biased over a few regions owing to a location bias with the Inter-tropical Convergence Zone (ITCZ), e.g. see Ziehn et al. (2020): CSIRO PUBLISHING | Journal of Southern Hemisphere Earth Systems Science	Text shown in column left added to the revised manuscript.

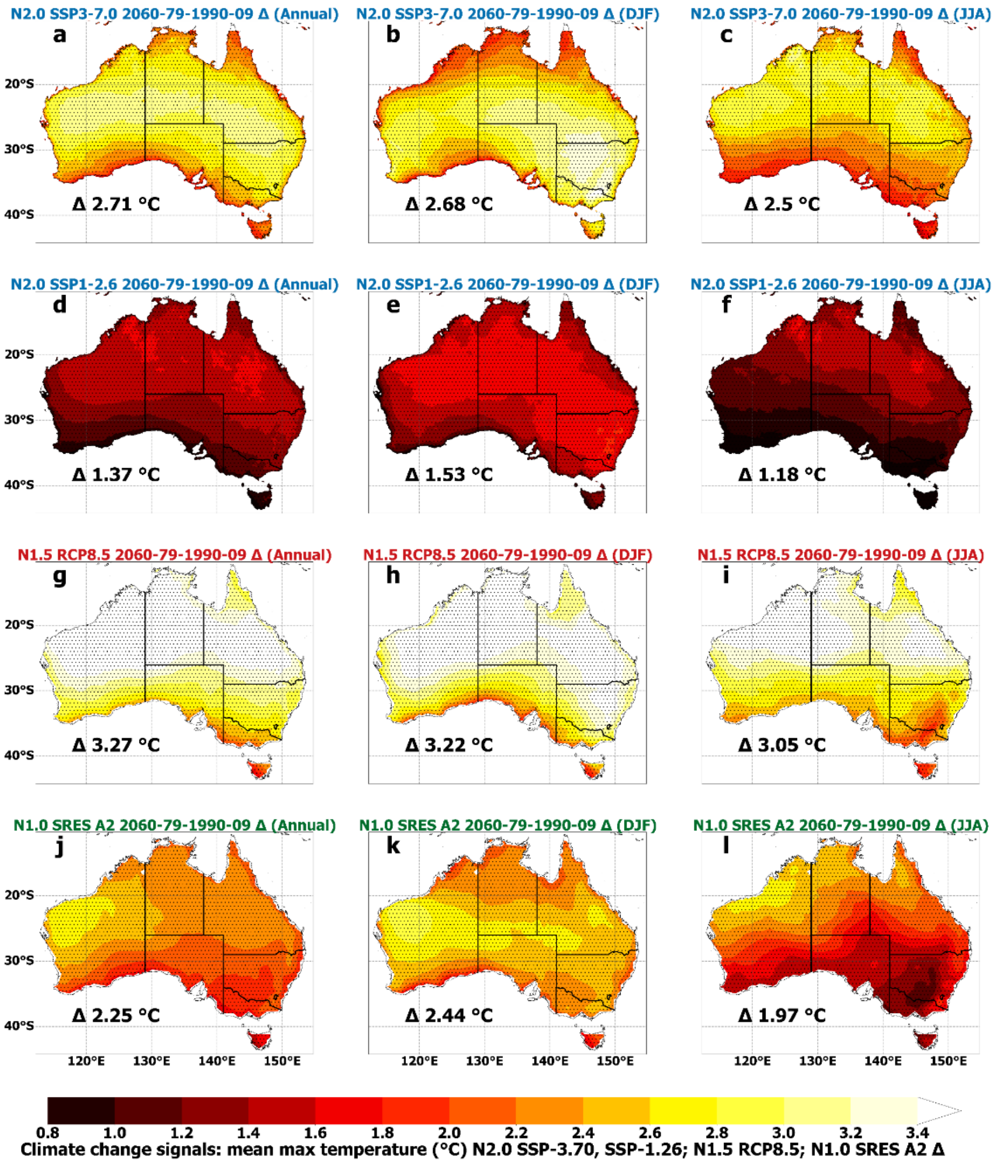
		and: Rashid et al. (2022): https://www.publish.csiro.au/es/fulltext/es21028	
8	Table 1 is very helpful. Can an extra row on computational resources (core hours) be added? This would help emphasise how much more of an effort going to 4km resolution is.	For NARCliM 2.0, during production phase of running both the 20 km and convection-permitting 4 km simulations, we used approximately **1060M core hours**. Note that these domains were run simultaneously, we do not have separate usage for the 4km resolution domain only. For NARCliM1.5, figures used are from when we were performing cost estimates for NARCliM 2.0 estimates (i.e. not actual logs): we consumed in total **30M core hours**. Unfortunately, NCI (the HPC facility we used) discarded historical SU usage when they replace their main HPC, so we can not confirm the original billing logs. Records for core hour usage for the original NARCliM 1.0 are unfortunately no longer available, but core hour usage per ensemble member year should be broadly similar to NARCliM 1.5.	Table 1 is updated accordingly in the revised manuscript.
9	Figure 4, for precip, are the units mm/day?	Thanks for asking this question, yes, the units are mm/day – figure caption revised accordingly.	Figure caption revised for units.
10	Figure 9 (and others), I found it difficult to see the stippling/hatching. The resolution of the file	We agree that Figure 15 is difficult to read, e.g. the original version was 300 DPI; we have	Figure 15 revised (please see below this table).

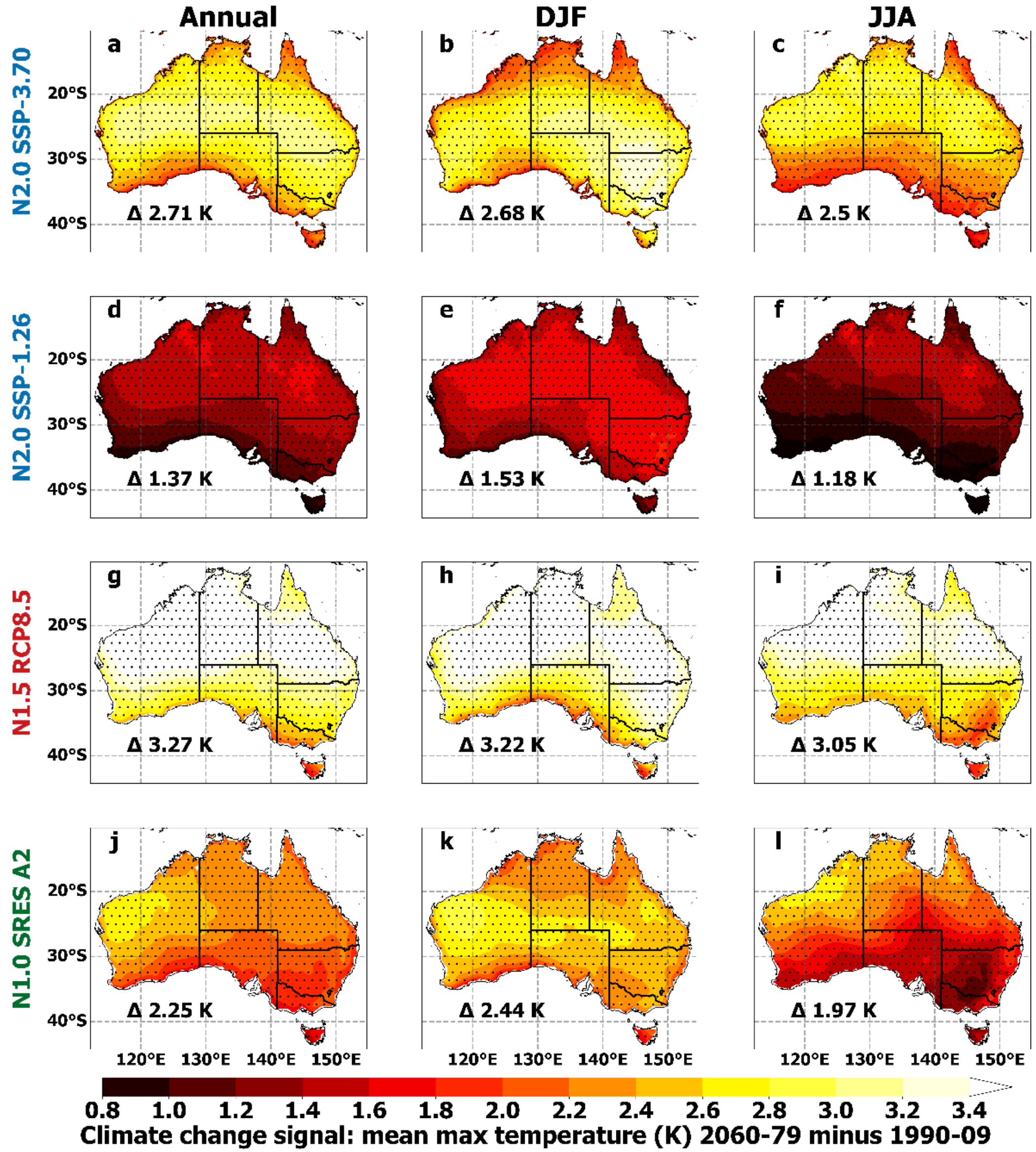
	was low (not sure if this was an issue with the pdf preprint?) but please ensure that high resolution figures are used and that the journal isn't compressing these in the final version. The resolution is particularly low for Figure 15 and very difficult to read.	now increased the DPI to 600, among other modifications. We have revised this plot, please see the example new Figure 15 below this table.	
11	I think in some figures there is a lot more repetitive text than there needs to be. Rethinking the layout headers of certain figures would help. For example, Figure 9 and 12, (Annual, DJF, JJA) could simply be headers at the top of each page, and the different versions of Narclim could be along the LHS of page. The text is often also too small to read. E.g. the colorbar caption in Figure 15 is excessively long and this information could simply be in the caption.	Thanks for these suggestions. We have revised these figures as you have suggested – please see examples below this table.	Figures revised as suggested (please see below this table).



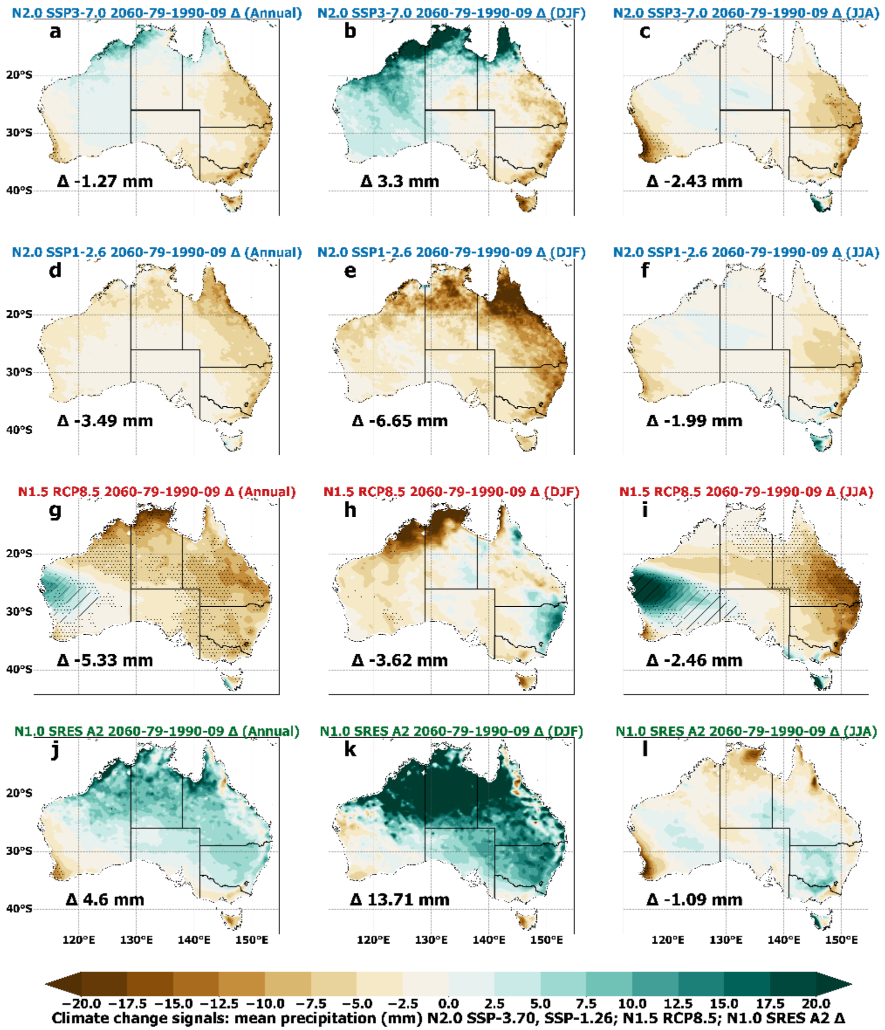


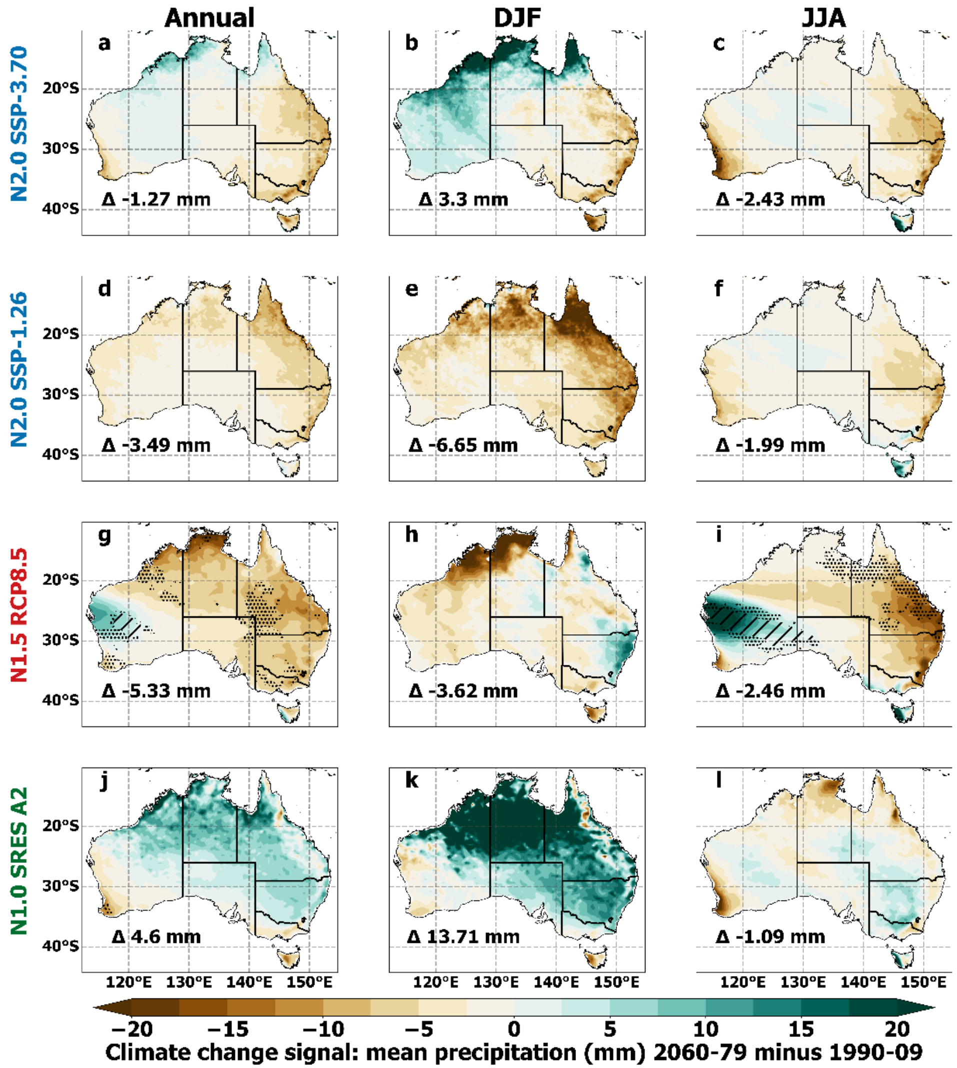
Reviewer 2, Comments #6 and #11. Left: original **Figure 9** from initial submission; Right: revised **Figure 9** using revised statistical significance method (please see #6) and revised plot layout/headers and labelling and increased DPI (please see #11)





Reviewer 2, Comments #6 and #11. Left: original **Figure 12** from initial submission; Right: revised **Figure 12** using revised statistical significance method (please see #6) and revised plot layout/headers and labelling and increased DPI (please see #11)





Reviewer 2, Comments #6 and #11. Left: original **Figure 14** from initial submission; Right: revised **Figure 14** using revised statistical significance method (please see #6) and revised plot layout/headers and labelling and increased DPI (please see #11)

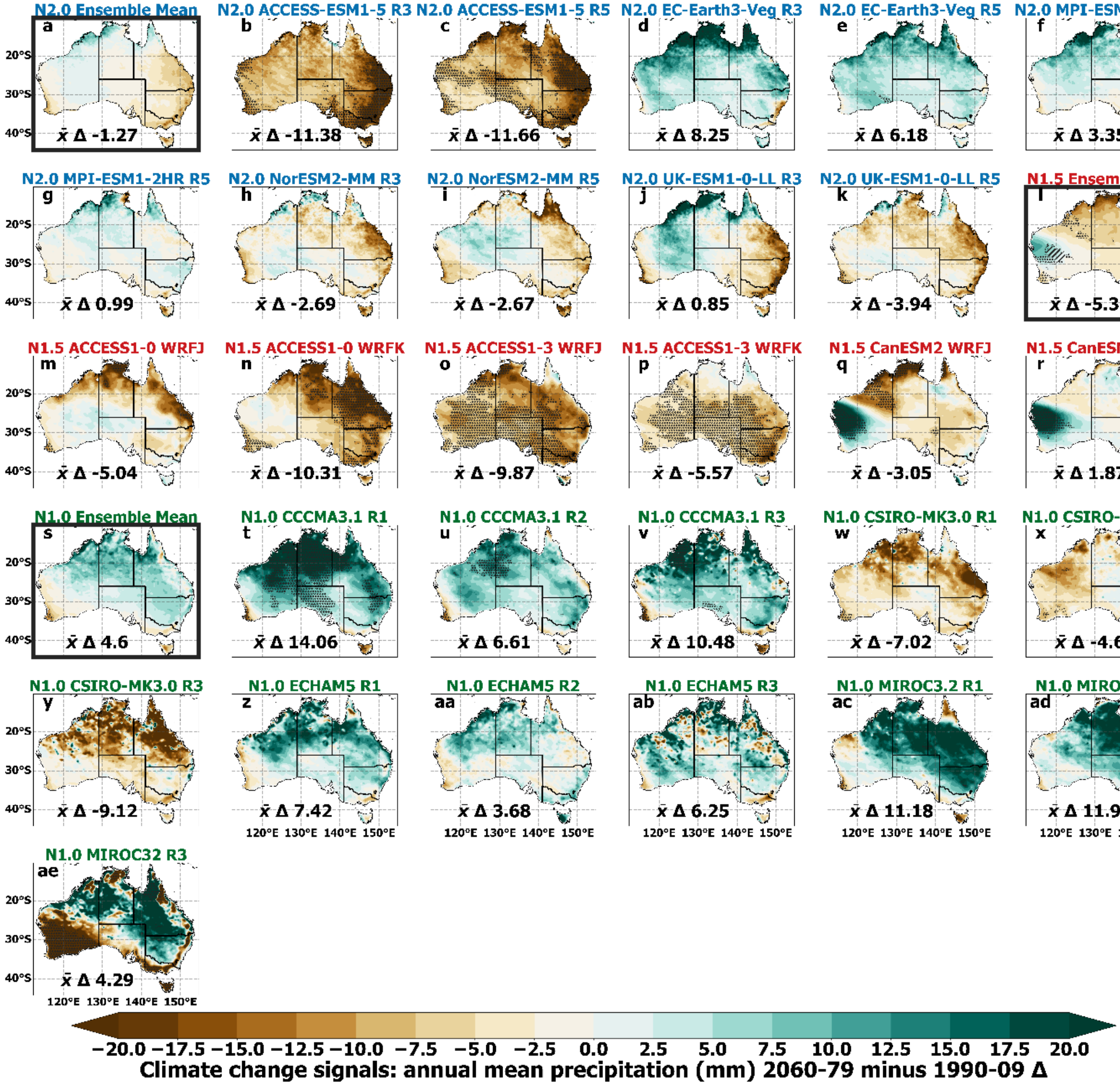


Figure 15: revised version (Reviewer 2, comment #10)