<u>Authors responses (in red) to referees comments (in black)</u>

We thank the two reviewers for their constructive comments and suggestions and believe our revisions in response to their comments, outlined below, have significantly improved the manuscript.

<u>Reviewer #1</u>

The article by Peel et al. mainly assesses the 'within-GCM' uncertainty and its impact on modelled runoff for climate change impact studies. Overall, the article is clearly structured and well written. However, I have major concerns about the very premise of the paper and the methods used by the authors. Therefore I seriously doubt it could be a valuable contribution to the journal Hydrology and Earth System Sciences.

Noted.

Comments related to individual chapters:

TITLE:

The title is ambiguous and does not provide a proper summary of the article.

Agree. We have revised the title of the paper to be less ambiguous and to make it independent of the complementary paper in the initial submission.

INTRODUCTION:

The introduction describes concisely the aim of the work and provides a clear overview of the study setup.

Noted with thanks.

However, previous studies on investigating and quantifying uncertainty from various sources other than GCMs are not referenced.

Agree. We assume this comment refers to adding references for the non-GCM uncertainties discussed in the introduction (page 4582). We have revised the Introduction to include appropriate references.

METHODOLOGY AND RELATED LITERATURE:

Section 2.1

The paper claims that the primary aim of this study is to investigate within-GCM uncertainty, but actually what it does is to approximate stochastic replicates of GCM runs based on a single run. The underlying assumption here is that various runs from one GCM have same long-term trend and low frequency signals, which is not necessarily true.

We agree that our methodology assumes the same long-term trend in the GCM runs that we replicate. We state in the paper that our methodology is a likely under-estimate of true within-GCM uncertainty because of this assumption. This stochastic methodology was developed to estimate uncertainty when we have limited GCM runs. This is similar to the approach in water resources system simulation where there is a single historical time-series of data. When there are many runs available for a given GCM and scenario, our methodology need not be used as there will be sufficient runs to estimate within-GCM uncertainty directly from the GCM runs.

We have revised the Introduction and this section of the manuscript to make this issue clearer to the reader.

The authors admit that this approximation represents an under-estimate of the true within-GCM uncertainty. This could be true. In fact the uncertainty within GCM is most likely to be GCM specific since GCMs have different sensitivity to initial conditions. Further, the initial condition is not the only difference between GCM runs. Some of the runs used different forcing (e.g. UKMO-HadGEM1 20C3M runs), some of the runs were simply run on different platforms (e.g. ECHO-G 20C3M runs). Therefore the approximation here can not represent the true within-GCM uncertainty.

We agree that our approximation does not represent true-within-GCM uncertainty as noted by the reviewer. We also agree that each GCM is likely to have its own within-GCM uncertainty. Our approximation was developed to gain some insight in the likely scale of impact of within-GCM uncertainty on hydrologic outputs.

The revisions in response to the previous comment address this issue.

What really assessed in this study is the uncertainty within the stochastically generated data.

In this study we have assessed uncertainty associated with GCM runs using generated sequences of stochastic data. Our approach follows the standard practice in surface hydrology (Hipel & McLeod, 1994) where stochastic replication of the observed record is used to estimate uncertainty. In our case the GCM runs are treated as the record for stochastic replication. We clearly state that this stochastic method is an approximation of within-GCM uncertainty since we are not seeking to replicate the overall trend – only a GCM should be used to produce the long-term trend in response to an emissions scenario.

The revisions in response to the previous comment address this issue.

There are CMIP3 GCMs that provide as many as 8 runs and more for CMIP5 GCMs. It would be much creditable if they use 3 to 10 real GCM runs to quantify within-GCM uncertainty rather than using 100 stochastically constructed replicates. Or at the very least, using those real GCM runs to validate the results and conclusion reached by using this method.

Agree. In the revised manuscript we have included a new Section 3 (“Testing the stochastic within-GCM uncertainty approximation”) in which a comparison between multiple runs of the CCSM GCM and our stochastic approximation of within-GCM uncertainty is made. The key result in this section is Figure A (new Figure 4), which compares within-GCM uncertainty based on seven runs from the CCSM GCM with the stochastic approximation of within-GCM uncertainty for (a) annual precipitation and (b) annual temperature for the Herbert River at Gleneagle. In each plot the maximum, median and minimum annual value for a given year are shown for the seven CCSM runs and are compared with the maximum, median and minimum of the 700 (7 x 100) stochastic replicates of the CCSM runs of the within-GCM uncertainty in annual precipitation and temperature. For both precipitation and temperature the median of the 700 stochastic replicates overlies the median of the 7 CCSM runs and difference between the maximum and minimum lines around the median for the two datasets are totally consistent given only seven CCSM runs and 700 stochastic replicates. A comparison of the standard deviation of all annual values calculated for the seven CCSM runs (precipitation = 110 mm, temperature = 1.21 ^{o}C) and the 700 stochastic replicates (precipitation = 111 mm, temperature = 1.20 ^{o}C) confirms that stochastic replicates are replicating the CCSM GCM runs well in terms of overall trend and variability around the trend. These results confirm the credibility of the stochastic methodology for approximating the within-GCM uncertainty when limited GCM runs are available.

Section 2.4

Another serious problem with the methodology is how the GCM climate is related to catchment hydrology. Only 5 of the 17 catchments used in this study are larger than a grid cell of the finest-resolution GCM (MPI) out of the 5 GCMs investigated. Many catchments are smaller than one tenth of a grid cell.

From a hydrological and water management perspective the ability to conduct climate change impact assessments on a large number of small to medium sized catchments for a wide range of GCMs is critical. Since we are dealing with catchments from around the world our resources did not allow us to carry out a comprehensive dynamic downscaling of each of the GCM outputs. Rather a simple statistical approach (quantile-quantile bias correction) was adopted in this proof-of-concept paper. This simple statistical approach is reasonable in this case as the data being bias corrected are monthly, not daily, and the data are not spatially distributed (area weighted average of monthly GCM data). The results shown in Table A (see later comment) confirm that the quality of results from our analysis is independent of catchment area.

We have revised this section to indicate the proof-of-concept nature of this paper and made clear that the bias correction is being applied to monthly catchment average (not spatially distributed) data.

However, regardless of their size, an area weighted average of the GCM data based on the proportion of catchment area associated with each GCM grid cell are calculated for each catchment and used as input to hydrological model PERM after bias correction. As stated in the paper, the GCMs tend to over-estimate low MAP and under-estimate high MAP. Further averaging could only accentuate this. In worst case, a climate series averaged for an area hundreds of times larger than the catchment is forced to represent the catchment climate.

As our time step of analysis is monthly the spatial precipitation and temperature fields are relatively smooth in the GCM and observed data, so area-averaging will not significantly accentuate this problem.

No change to the manuscript.

This is beyond what a quantile-quantile bias correction can fix. The use of bias correction itself is problematic as it impairs the advantages of GCMs by altering spatiotemporal field consistency, relations among variables and by violating conservation principles (Ehret et al. 2012).

Ehret et al. (2012) presents a detailed review of the inadequacies of bias correction and how to deal with the problem in the short, mid and long term. For the short term they offer no alternative methodology but recommend that the uncertainties associated with using the procedure be openly communicated and that the impact of both bias corrected and non-corrected input be provided. We note, however, that Rojas et al. (2011) found that bias correcting monthly temperature using a transfer function preserved the annual statistics. We have revised this section to indicate the proof-of-concept nature of this paper and made clear that the bias correction is being applied to monthly catchment average (not spatially distributed daily) data.

Our choice of quantile-quantile bias correction is supported by Teutschbein and Seibert (2013) who found the quantile-quantile method performed best out of six alternate bias corrections in differential split sample tests for non-stationary conditions. In this study the quantile-quantile bias correction has worked well as confirmed in the runoff results shown in Table A (we will not include Table A in a revised manuscript). Satisfactory runoff estimation requires the bias correction of temperature and precipitation data to be successful. In Table A three sets of mean annual runoffs (MAR) are presented: (1) observed mean annual runoff; (2) MAR estimated from bias-corrected GCM inputs to PERM expressed as a percentage of observed MAR; and (3) MAR estimated from bias-corrected stochastic replicates of GCM precipitation and temperature input to PERM expressed as a percentage of observed MAR. The last two MAR estimates are for the period 1965 – 1994 (20C3M), while the observed MAR is based on all available observed data (not necessarily from the period 1965-1994). Although there are several large biases in MAR, the overall bias is 3.6%, with 73% of MARs being within ±10% of the observed MAR. Overall the modelled results exhibit very small biases, which confirm that the quantile-quantile bias correction of monthly precipitation and temperature inputs to PERM has worked well in this study.

RESULTS AND DISCUSSION

The boxplots in Figures 4, 8 and 11 show that the within-GCM ranges are usually larger than the between-GCM (raw) ranges. This is to say that the initial condition used by a GCM has greater impact than the model structure and parametrisation. I suspect this is strongly related to the method used in this study.

We suspect the reviewer has been misled by the misleading caption to Figures 4, 8 & 11. The 'Raw' values have actually been bias-corrected and are not raw (original) GCM values. The captions have been modified to correct this (the body of text reflects the correct interpretation).

CONCLUSIONS AND IMPLICATIONS

The authors conclude that the with-in GCM should not be neglected and has significant implications for interpreting climate change impact assessments and warned the decision makers the risk of sense of certainty that is unjustified. In reality, the large uncertainty in climate change impact assessments is well known, and it is also well established that the largest uncertainty is usually associated with GCM simulations. There are extensive discussions around how to improve this situation, including clearer communication, using multi-model ensembles, and eventually, improving models themselves. In my opinion, this paper adds limited value to the research community.

We agree that the large uncertainty in climate change impact assessments is well known, but suggest within-GCM uncertainty has not been well quantified due to the limited number of GCM runs available for each GCM and scenario combination. Our methodology is a contribution toward quantifying within-GCM uncertainty and provides an objective approach for communicating the uncertainty in climate change impact assessments in a quantitative manner. In this paper we applied our procedure to estimate the impact of within-GCM uncertainty on annual runoff and reservoir yield. Such information is crucial to water resources engineers and management decision makers in the short to medium term planning horizon. For the research community the stochastic data generation methodology provides a way to assess within-GCM uncertainty on a temporary basis until the number of GCM runs for a given GCM and scenario combination becomes adequate to estimate within-GCM uncertainty from GCM runs directly.

The research reported in this paper was conducted as part of a collaboration with the water industry where the industry partners were extremely interested to see the level of uncertainty in estimates of runoff and reservoir yield from our approximation of within-GCM uncertainty.

We have revised the Conclusions and Implications section to make the contribution of this manuscript more evident.

References

Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., Liebert, J., 2012. HESS Opinions "Should we apply bias correction to global and regional climate model data?". Hydrol. Earth Syst. Sci., 16(9), 3391-3404, doi:10.5194/hess-16-3391-2012.

Noted.

Reviewer #2

Attempting to comprehensively address all sources of uncertainty is challenging and it is reasonable for the authors to focus on some particular aspects of the problem. Here it is clear that the main focus lies on addressing the representation of internal variability by generating synthetic time series conditioned on information about lowhigh frequency variability in the GCM time series. This looks interesting and is well worthy of publication though with some tweaks could probably address some concerns that otherwise could be directed towards this study (note that I'm not commenting on the Hydrological aspects of this paper, e.g. the hydrological model and the reservoir calculations, these areas are outside my expertise).

Noted.

1. Can the stochastic method really re-place dynamically simulated ensemble members? If this is a proof of concept paper, it would be good if the authors had selected a GCM with several runs so that we could see the spread of the dynamically simulated ensemble members in relation to the stochastically simulated ensemble members. Further, why 100 simulations - does the spread stabilise around 100 samples?

We agree that this paper should be read as a proof of concept paper and we have modified the text accordingly. In the Introduction we stated that 'A complete assessment of the magnitude of within-GCM uncertainty requires numerous, for example 100, runs of each scenario from each GCM, which are currently unavailable across all GCMs' and that our method is an approximation. In the methodology we also stated that 'An ideal assessment of within-GCM uncertainty would involve analysis of at least 100 runs of a single GCM for a given scenario with each run having slightly different, but equally plausible, initial conditions.' We agree our method should not replace dynamically simulated ensemble members, as the reviewer notes, but is a temporary approximation that can be used until substantially more GCM ensemble members become available for each GCM and scenario combination. We have modified the text to reflect this.

There is nothing special about 100 stochastic replicates. As a proof of concept paper we selected 100 replicates to ensure we have a reasonable sample size on which to base the summary statistics. In stochastic hydrology, generally 100 replicates are adopted (see McMahon et al.; 2008; Adeloye et al, 2010; Potter et al., 2010).

In response to reviewer #1 we have introduced a comparison of the seven CCSM runs from CMIP3 against our stochastic methodology (see Figure A). See response to reviewer #1.

2. Can within GCM variability really be greater than between GCM variability? Well, I guess it is possible in the near term for variables with large natural variability such as rainfall. The selected ’future’ time period here falls in the ’near to mid-term’ category, so perhaps it isn’t impossible. However, as the authors note - the GCMs have been bias corrected and the sample of GCMs is small so is this conclusion robust? It would be good to relate the spread of the selected sub-sample of GCMs (before and after bias correction) to that of the entire CMIP3 archive in terms of projected precip and temp (regionally and globally).

Within-GCM uncertainty can be higher than between-GCM uncertainty particularly at regional scales as illustrated in Figure 4 of Hawkins & Sutton (2011). They found the magnitude of within-GCM uncertainty is greater than between-GCM uncertainty at smaller spatial and temporal scales, so one would expect it would be even larger at catchment scale.

From our bias corrected results we do observe that within-GCM variability is generally larger than between-GCM variability for our sub-set of GCMs. As noted in the Conclusions and Implications section we believe the observed limited between-GCM variability is due to the bias correction forcing all GCMs to have the same mean and variance as the observed P & T data over the observed period of record at each catchment. The between-GCM variability observed in Figures 4, 8 & 11 is due to differences in GCM trend from the observed period to the two periods of assessment (1965-1994 and 2015-2044). If original data were used (un-bias corrected) the between-GCM variability would be significantly larger.

The reviewer requests that we “relate the spread of the selected sub-sample of GCMs (before and after bias correction) to that of the entire CMIP3 archive in terms of projected precip and temp (regionally and globally)”. In Figure B we compare un-bias corrected mean annual precipitation against mean annual temperature for (a) 1965-1994 and (b) 2015-2034 for the selected five GCMs runs compared with the 23 CMIP3 GCMs including all ensemble members for the global land surface. The figures show the relative positions of the selected five GCM runs compared with the 23 CMIP3 GCMs including all 44 ensemble members for the two periods. The five selected GCM runs sit within the cloud of 44 GCM ensemble members, which indicate that they are reasonably representative of the range of current and future GCM projections. The wide range of mean annual precipitation and temperature values for these two periods shows there is significant between-GCM variability prior to bias-correction.

In Table A we show post bias-corrected results for our 17 regional catchments for our selected sub-set of GCMs. Across the five GCMs we observe very similar average bias in the runoff results following bias-correction. To extend this Table to include results from all 44 GCM runs from CMIP3 for these catchments would be a major undertaking that is not expected to yield a significantly different outcome.

No change to the manuscript.

3. Rather than using bias corrected GCM data as input to the hydrological model, could you not use simple daily scaling whereby you apply the change signal on observed data? There just seems to be a bit of a scale mis-match between the GCM output and required catchment scale. Daily scaling has its obvious limitations, but would better represent the regional variability in precip. Maybe a few words on why you choose to use bias corrected GCM input over this very simple downscaling method.

It would be possible to use daily scaling as opposed to bias correction, if by ‘simple daily scaling’ the reviewer means scaling ‘the historical climate series to reflect a future climate by the relative difference between GCM simulations of future and historical precipitation and other climate variables' (Chiew, 2010, page 218). However, simple scaling would not make full use of the stochastic replicates. Re-scaling the observed series multiple times keeps the observed sequence of precipitation and temperature in all cases. In contrast the stochastic replicates have a different sequence of precipitation and temperature events each time. We have modified the manuscript to reflect this discussion.

References

Adeloye, A. J., Pal, S., and O'Neill, M.: Generalised storage-yield-reliability modelling: Independent validation of the Vogel–Stedinger (V–S) model using a Monte Carlo simulation approach, Journal of Hydrology, 388, 234–240, 2010.

Chiew, F. H. S.: Lumped Conceptual Rainfall-Runoff Models and Simple Water Balance Methods: Overview and Applications in Ungauged and Data Limited Regions, Geography Compass, 4/3, 206–225, 2010.

Hawkins, E., and Sutton, R.: The potential to narrow uncertainty in projections of regional precipitation change, Clim. Dynam., 37, 407–418, 2011.

Hipel, K. W. and McLead, A. I.: Time series modelling of water resources and environmental systems, Elsevier, Amsterdam, 1013 pp.

Potter, N. J., Chiew, F. H. S., and Frost, A. J.: An assessment of the severity of recent reductions in rainfall and runoff in the Murray–Darling Basin, Journal of Hydrology, 381, 52–64, 2010.

McMahon, T. A., Kiem, A. S., Peel, M. C., Jordan, P. W., and Pegram, G. G. S.: A New Approach to Stochastically Generating Six-Monthly Rainfall Sequences Based on Empirical Mode Decomposition, Journal of Hydrometeorology, 9, 2008.

Rojas, R., Feyen, L., Dosio, A., and Bavera, D.: Improving pan-European hydrological simulation of extreme events through statistical bias correction of RCM-driven climate simulations, Hydrol. Earth Syst. Sci., 15, 2599–2620, doi:10.5194/hess-15-2599-2011, 2011.

Teutschbein, C., and Seibert, J.: Is bias correction of regional climate model (RCM) simulations possible for non-stationary conditions? Hydrol. Earth Syst. Sc., 17, 5061-5077, 2013.

Table A Mean annual runoffs expressed as percentage of observed mean annual runoff for the five bias-corrected GCM runs and for the average of 100 bias-corrected stochastically generated replicates for each GCM.

Ref. no.	River	Catchment area (km^2)	Obs. MAR (mm $year^{-1}$)	Years of obs. runoff	HadCM3	HadCM3 stoch.	MIROCM (1)	MIROCM (1) stoch.	MIUB (1)	MIUB (1) stoch.	MPI(1)	MPI(1) stoch.	MRI(3)	MRI(3) stoch.
1202	Bafing	15500	568	27	1.06%	9.15%	4.23%	3.70%	3.52%	4.75%	6.51%	3.35%	-2.64%	-0.35%
1325	Oueme	10326	195	33	6.15%	15.9%	30.8%	39.5%	-5.13%	-3.08%	-7.69%	3.59%	4.10%	-17.4%
1333	Sabi	11000	127	30	-0.79%	15.0%	-14.2%	-5.51%	2.36%	2.36%	-8.66%	-9.45%	-3.15%	0.79%
2270	Songhuajiang	391000	92.9	46	-5.38%	-1.83%	-15.7%	-6.35%	-3.34%	6.89%	-5.06%	-2.69%	7.64%	7.64%
2274	Tapi	61575	224	31	25.9%	15.7%	21.5%	20.1%	19.6%	15.6%	6.70%	7.59%	35.7%	25.9%
2288	Wujiang	58300	619	44	0.16%	-3.39%	-10.3%	-6.95%	-1.13%	3.39%	-4.20%	-6.62%	-2.42%	0.97%
3195	Kiamichi	3686	418	47	-6.70%	-2.63%	12.0%	11.5%	11.5%	0.24%	22.5%	15.6%	28.7%	14.8%
3279	Black	3243	257	56	7.39%	-0.78%	6.23%	7.39%	10.9%	5.84%	0.39%	6.23%	2.33%	14.4%
3543	Umpqua	9539	719	20	-1.25%	4.59%	-1.95%	2.09%	3.20%	-5.56%	-0.83%	9.04%	4.17%	11.4%
4014	Magdalena	74410	1043	47	4.31%	4.03%	-1.53%	-3.45%	0.67%	-6.23%	0.38%	0.38%	1.53%	0.10%
4019	Cuyuni	53354	638	31	8.62%	28.2%	-12.8%	9.25%	8.15%	7.52%	-8.15%	1.88%	9.40%	5.49%
4145	Lumaco	1054	569	42	-0.35%	4.57%	-8.61%	-2.81%	7.73%	7.73%	4.22%	2.64%	-4.04%	-0.88%
4179	Rio Jaguaribe	21770	39.4	52	25.6%	30.7%	35.8%	23.6%	13.4%	15.0%	17.0%	64.2%	29.2%	43.2%
5255	Clyde	1704	745	25	1.48%	2.15%	3.22%	3.89%	-3.76%	-3.76%	0.40%	5.23%	-2.82%	-5.50%
6058	Herbert	5236	195	82	5.13%	-5.64%	-3.59%	18.0%	-5.64%	7.69%	1.54%	3.59%	-3.08%	3.08%
6103	Nymboida	1660	483	85	2.69%	2.07%	-8.49%	-1.66%	-9.94%	-0.62%	6.83%	5.59%	-13.7%	0.41%
6279	Ovens	5410	207	98	-7.25%	3.86%	-62.4%	-48.8%	-7.73%	-10.1%	-10.1%	-5.80%	-20.8%	-12.1%
		Average			3.93%	7.15%	-1.53%	3.73%	2.61%	2.80%	1.28%	6.14%	4.13%	5.41%

Note: Observed (obs.) mean annual runoffs (MAR) are based on the available historical data which covered different periods to that used in the GCM and stochastic data analysis (1965 – 1994, 20C3M). This difference would explain some of the discrepancies between the historical data and the GCM runoff. Although there are some large biases, 73% of MARs are within ±10% of the observed MAR. Overall, the modelled results exhibit very small biases, which confirm that the quantile-quantile bias corrections of monthly precipitation and temperature inputs to PERM have worked well in this study.

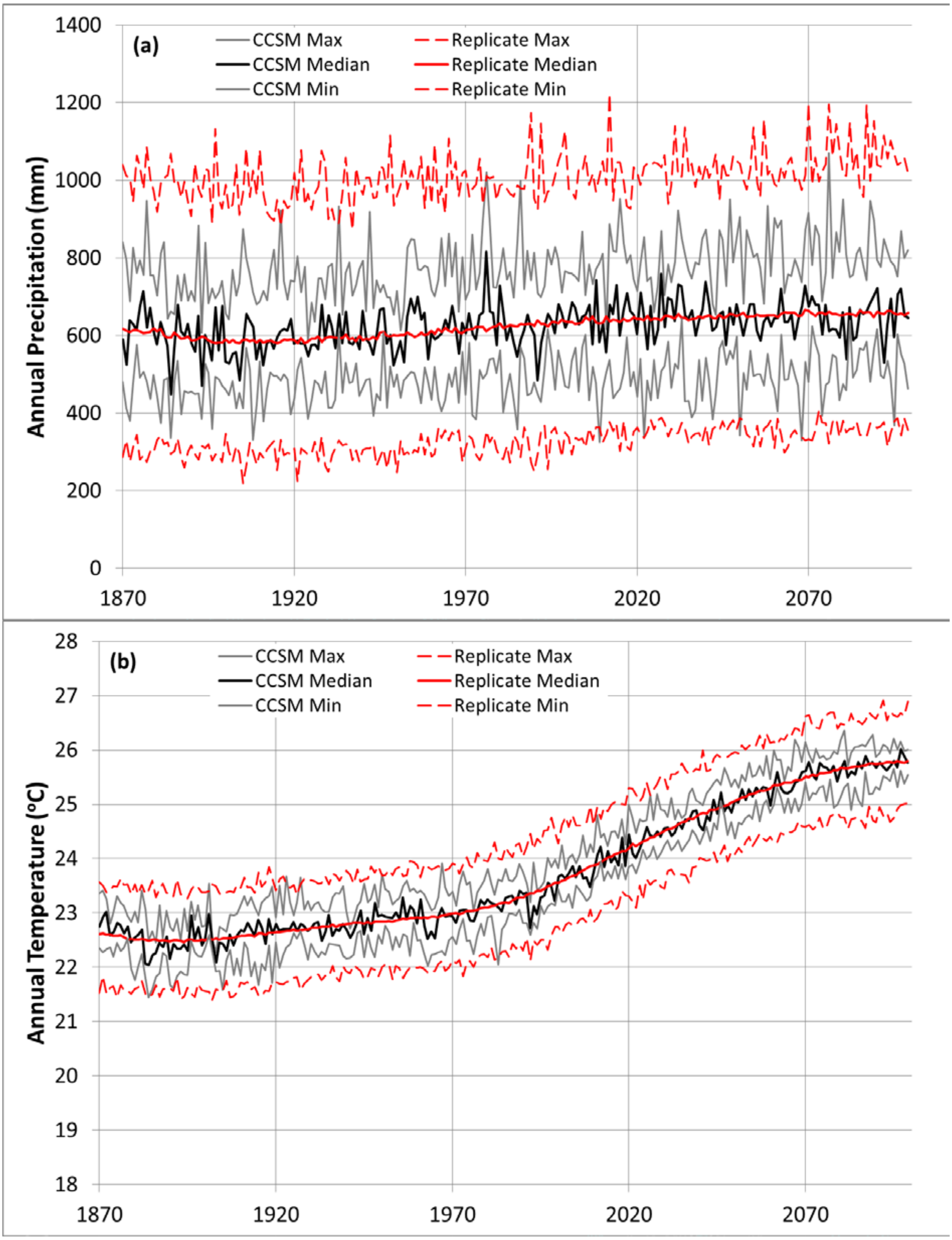
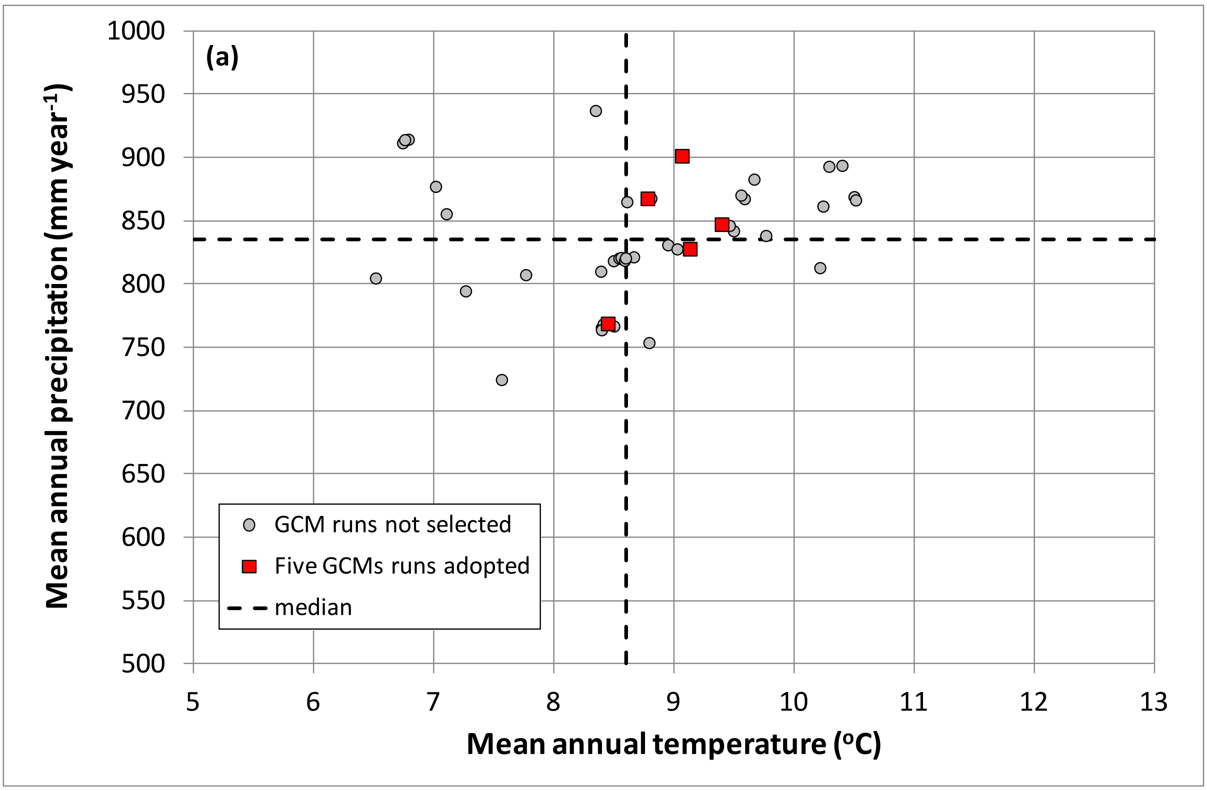


Figure A Within-GCM uncertainty for the Herbert River at Gleneagle based on seven runs from the CCSM GCM compared with the stochastic approximation of within-GCM uncertainty for un-bias corrected (a) annual precipitation and (b) annual temperature. In each plot the maximum, median and minimum annual value for a given year are shown for the seven CCSM runs compared with the maximum, median and minimum of the 700 (7 x 100) stochastic replicates of the CCSM runs.



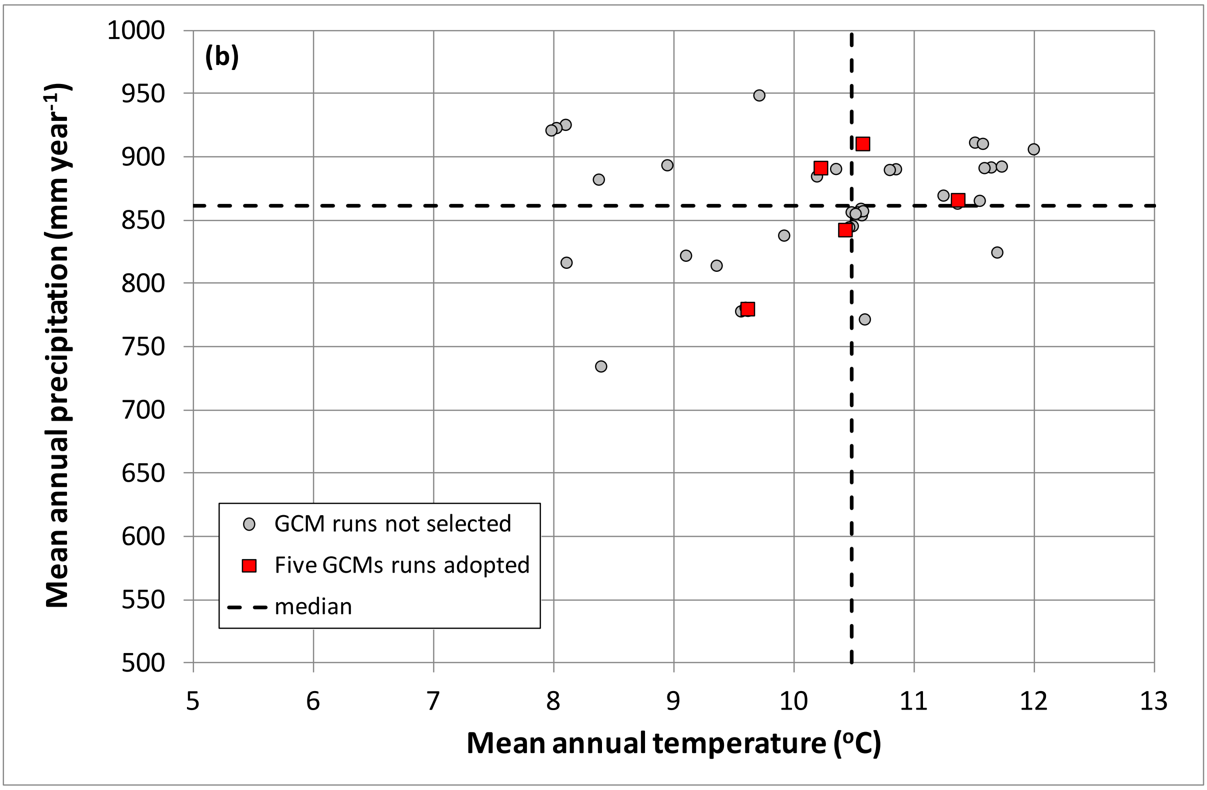


Figure B Mean annual precipitation versus mean annual temperature for (**a**) 1965-1994 and (**b**) 2015-2034 for the selected five GCMs runs compared with the 23 CMIP3 GCMs including all ensemble members for the global land surface