Response to reviewers' comments for: Interactive comment on "The implications of climate change scenario selection for future streamflow projection in the Upper Colorado River Basin" by B. L. Harding, A. W. Wood and J. R. Prairie.

Response by B. L. Harding and A. W. Wood. April 26, 2012

Reviewer comments are in italics and the authors' response follows

Anonymous Referee #1

Received and published: 24 February 2012

This is an important paper that represents the most complete attempt yet to understand the properties of future climate scenarios in the Upper Colorado River Basin (UCRB). The UCRB is the source of water for millions of people in the western United States and an important driver of economic activity in the region. Accordingly, the results presented here could have a significant impact on stakeholder discussions and real life applications for a large number of people. It is therefore critical to examine the work carefully, and make sure it is treating the problem correctly and the conclusions drawn are supportable and unbiased by the methods and procedures used in the analysis. It is obvious that the manuscript describes an epic undertaking. Working with so many projections, downscaling them, and producing hydrological simulations from all the climate information was surely a monumental task. So I would like to express my appreciation of the authors for the very large effort involved in the work described here.

RESPONSE: The authors thank the reviewer for these comments.

However, I would be doing a discourtesy to my colleagues if I did not convey that I think the work as it currently stands has major procedural flaws that prevent it from being able to answer the issues addressed in an unbiased manner. There are three overriding issues:

1) The downscaling procedure biases the projections to be wet, and we know that this affects the results, yet no attempt is made to take account of this in the results or conclusions.

2) There is unequal weight given to the climate models (some weighted by more than a factor of 5 relative to others), yet the weighting is not done by any metric of model quality, but rather by the nearly arbitrary metric of the number of ensemble members that happened to be available to the investigators.

3) The treatment of the range of uncertainty in the results repeatedly focuses on the difference between the minimum and maximum values found in the distribution, yet this focus is unmotivated by either standard statistical practice or any described application of the audience (stakeholders) that the work is intended to address.

Because of these flaws it is my opinion that the manuscript would be misleading to the intended audience if it were published in its current form. I am suggesting that it be returned to the authors for major revisions. In the comments I have also proffered suggestions as to how some of these issues could be addressed in hope that the authors find them useful.

Major comments:

1) On Page 857, lines 5-8, it is described that the downscaled runs show precipitation changes of "up to 5 percent" wetter than the changes exhibited by the underlying global climate model run. Vogel et al. (1999) J. Irrigation and Drainage Engineering v. 125 p. 148 estimates that the precipitation elasticity of the Upper Colorado River Basin is about 2.5. Sankarasubramanian et al. (2001) WRR v. 37 p. 1771 shows a somewhat lower number in his contour map, around 2.3. If we split the difference and estimate that the precipitation elasticity for the UCRB is 2.4, then the up to 5% wetter precipitation results in an up to 12% greater streamflow in the Colorado River. If the average result (as opposed to the "up to" result) is only half this, it would be about 6% greater streamflow in the Colorado River. Since you find that the mean change in Colorado River flow is -7%, this 6% is an O(1) effect and has the potential to be a significant modification of your results. It can't simply be mentioned once on page 10 of a 37 page report and not referred to again.

I am not disagreeing with your statement that more analysis on the issue is beyond the scope of this paper. But I am very much objecting to the fact that this issue is not clearly pointed out in the conclusions, and a quantitative estimation made as to its effects. The Vogel and Sankarasubramanian elasticity numbers should be quoted so that the reader can understand the size that this effect could have on the presented results. If the size of the correction were trivially small this would not matter, but it is to first order in the mean change and so cannot be simply ignored in the conclusions of the work.

RESPONSE: The reviewer is correct that if the bias correction process does introduce a widespread and consistent positive bias in precipitation, the resulting bias in streamflow would be of the same order as the projected mean change by the end of the century. We agree that adaptation planning should recognize whether the bias correction is itself introducing a new bias or if the wettening represents an improvement over the GCM outputs. Though it is outside our scope to provide that insight, here is a more detailed summary of what is contained in the Reclamation report and a suggestion for revisions.

Reclamation provides an example of the precipitation difference in the form of an empirical CDF of projected precipitation change over the 112 CMIP3 projections for the period 2040-2069 relative to the period 1970-1999 for one location in the western Sierra Nevada mountains. The CDF exhibits a wet difference above about the 30th percentile. The difference at the 50th percentile is about plus 2% and reaches a maximum of about plus 3% at about the 75th percentile. Reclamation also provides small maps of the differences for the 25th, 50th and 75th percentiles for a 2° grid covering roughly the continental U.S. Interpreting the color gradation of these maps indicates that at the 75th percentile up to a plus 5% difference exists in the western U.S. However, Reclamation notes, that the positive differences are greater in wetter areas, and this appears to be the case in the water-producing areas of the Colorado River Basin. A careful examination of the map indicates that the difference in those areas may be plus 2% or less at the 75th percentile.

Given the low resolution of the data that have been published, and the difficulty, given limited published information of integrating the effect over the entire distribution of projections, we believe it would be a disservice to provide an estimate based on the upper bound of possible effects. This is particularly true because the bias correction process may actually be improving the GCM outputs. However, we agree with the reviewer that this is an issue of central importance for adaptation planning—this uncertainty increases the overall uncertainty of projected impacts by approximately 50%. We suggest that we revise our paper 1) to summarize efficiently some of the complexity of the issue as described above, including providing an estimate of impacts of a 5%

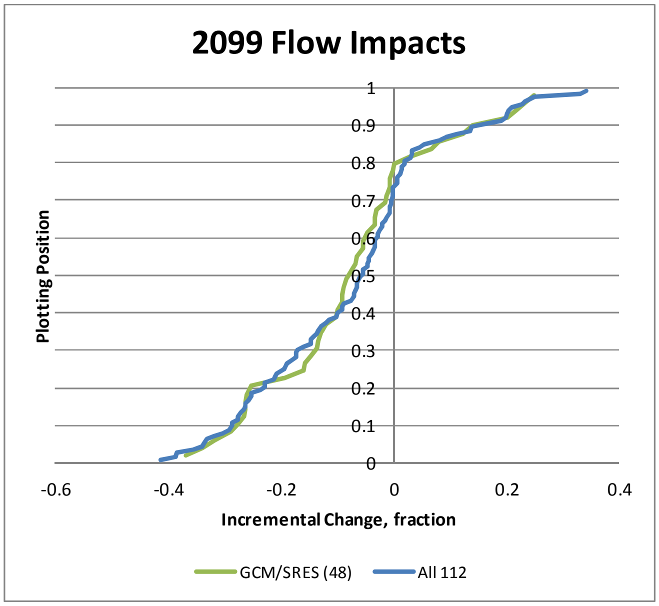
wettening for some of the key results, 2) to provide a clearer cautionary note about interpreting impacts based on bias-corrected projections, and 3) to provide a sharp recommendation for additional analysis to understand this effect.

2) The methodology is fundamentally flawed in that it does not take proper account of the uneven number of ensemble members included across models. This is not a trivial omission considering that three models have _5 ensemble members per scenario, while 8 other models have only 1 ensemble member per scenario. There is no logic nor justification for having scientific results that weight some models by a factor of 5 over others for reasons unrelated to model skill. The weighting used is essentially that of the donating institution's computing budget, which is the main thing that determined the number of ensemble members available. Yet computing budget has not been shown to be linked to model skill, and so should not have an effect on the results shown here. The methods used in this work need to be altered to weight results from all models equally, rather than preferentially weighting models from institutions with big computing budgets.

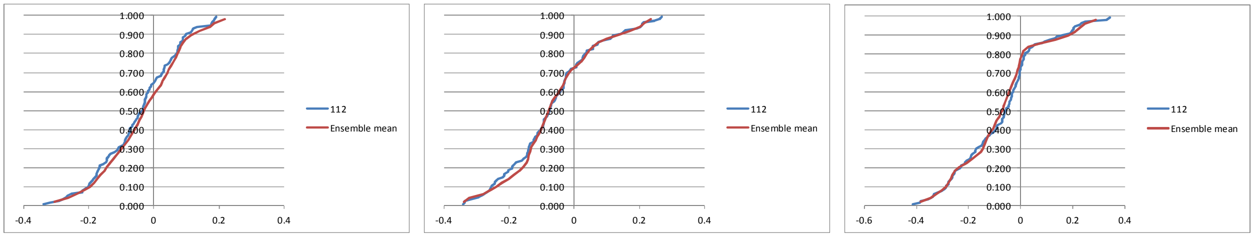
This is a particularly acute problem for the results shown here because Table 1 shows that the model with the most ensemble members, and therefore weighting the results 5 times more than other well-regarded models such as CSIRO, CNRM, or GFDL, is NCAR PCM1. While Dr. Washington and his group have nothing but my admiration for producing a coupled climate model almost 14 years ago, one has only to look at Peter Gleckler's muti-model comparative analysis of the CMIP-3 models to see that PCM is notably worse than all the other models included in CMIP-3. PCM was an outmoded model even by the time CMIP-3 came around. The way it's weighted in the results shown here relative to other models yields misleading results and conclusions. But ultimately this isn't a PCM issue – the results presented here need to weight all models equally, not according to the arbitrary computing budgets of the institutions that donated the runs.

*I will note that addressing this issue would be straightforward, if computationally tedious. One could simply construct a "super-ensemble" composed of a larger number of ensemble runs than the 112 you started with, with each model contributing equally to the super-ensemble. As a simplified example, the figures could be redrawn based on analysis of an ensemble of 112*7 = 784 runs, constructed such that each model donates exactly 7 ensemble members to the super-ensemble (7 because that is the number of ensemble members donated by the model with the most ensemble members, PCM1 in scenario B1). PCM1 donates 7 ensemble members by donating each of its 7 individual ensemble members exactly once, while HadCM3 donates its single ensemble member 7 times. That way you analyze a set that each model has contributed to equally – which is key – and yet you can still construct your figures 3-11, which would not be possible if you ensemble averaged of the results from each individual model. I've simplified the description of the process here because some models have an number of ensemble members that does not evenly divide into 7, but I assume the basic idea is clear.*

RESPONSE: This is an important point for adaptation planning and one that we should discuss. We were aware of this issue and have elsewhere advocated for development of larger, equally weighted ensembles of projections. We had made, but did not report, an analysis to evaluate the impact of uneven weighting. Our approach was to fit a linear trend to each model run for the period 2000-2099 and then average these trends across each of the 48 model/SRES scenario combinations. The ECDF of this ensemble at the end of the century is virtually the same as the ensemble of 112 individual projections as shown in the following figure.



Following the reviewer's suggestion we implemented a Monte Carlo resampling of the 112 projections, using one randomly-selected run from each of the 48 GCM/SRES combination in each ensemble to develop a super-ensemble of 100 members, each a 48-member ensemble. The results are shown in the following figure for 2039 (left), 2069 (middle) and 2099 (right). Each panel in the figure compares the ECDF for the equally-weighted, 48-member ensemble with the ECDF for the arbitrarily-weighted, 112-member ensemble. A subjective comparison indicates only slight differences.



We applied the Kolmogorov-Smirnov test to compare each of the members of the super-ensemble to the full 112-member, arbitrarily weighted CMIP3 ensemble at each time frame. With the exception of one member of the 2039 ensemble, the null hypothesis of equality of the two distributions could not be rejected at a p value of 0.2. Most p-values were considerable higher, with an average of about 0.9. We conclude that the arbitrary weighting of models arising from the unequal ensemble members for each did not lead to findings which were statistically different from those that would have been achieved had each model/scenario only had a single member, thus an equal weighting.

We suggest that we revise our paper to describe briefly these findings.

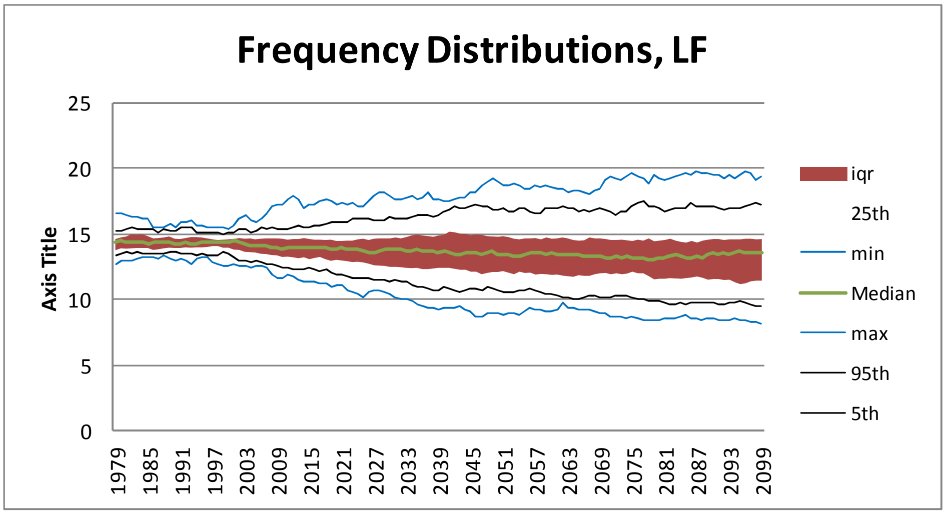
3) In several important places in the paper, including the entire motivation for the work (page 853 line 23) and the discussion around the beginning of page 863, the uncertainty in the model results is characterized by the range between the minimum and maximum obtained value. Since this range of uncertainty is self-described as the motivation for the paper (page 853 line 23), this statement deserves some thought. Yet as presented, this seems like an imprecise statement that is not meaningful.

*Including more samples is *expected* to give a higher likelihood of finding extremes. Nothing interesting or unexpected about that. For example, consider an ordinary Gaussian distribution. The range of the distribution is unbounded, so a Gaussian has "infinite" uncertainty if for some reason you consider uncertainty to be the range between the largest and smallest value present in the distribution. By this manuscript's terminology, ANY shift in the mean of a Gaussian is tiny compared to the "uncertainty" in the distribution if uncertainty is characterized as the distribution's range. This is nonsensical.*

The statements in the text concerning the range of uncertainty should be recast in terms of a standard measure of the width of the distribution, not the range between extremes. Has the interquartile range expanded with increasing numbers of projections? What about the 90% confidence interval? Either would be a meaningful statement. As sampling continues to increase, the "range" defined as the most negative to most positive value will tend to get larger, but the interquartile range should settle down much more quickly, and is a more physically meaningful result. I know that the authors are perfectly aware of the standard statistical tools for evaluating these kinds of trends, yet they aren't used in the manuscript. First, is the trend in mean streamflow statistically meaningful? That at least should be computed. Second, I am completely understanding of what I infer to be the authors' motivation in this exercise, which is to avoid the case where a statistically meaningful result is found, but one which is not meaningful in a practical sense to the intended audience. I.e., with enough sampling, even a tiny trend can be found to be statistically meaningful, yet a tiny trend embedded in large year to year variability may not be meaningful to a water manager. The problem is that the manuscript has not properly addressed this issue. The approach taken has been to compare the change in mean to the entire range of identified values, which is nonsensical for the reasons quoted above. It is obvious that as more and more samples are taken, the width of sampled values found (range between minimum and maximum) will tend to increase. By this manuscript's logic, that means that the more samples you have, the less meaningful any trend will be compared to this ever-widening range! This is backwards from what properly treated sampling should do.

The trend needs to be compared to some fixed measure of the width of the distribution, not the range between the sampled extrema. Since you are using 30-yr means, the obvious choice of width is the standard deviation of 30-yr mean values. Estimating this from Fig. 7, the standard deviation is about 13 percentage points. The mean change is about -7%. So the change is on the order of one-half of a standard deviation. Conservatively assuming 16 degrees of freedom (one per independent model, conservative because it discounts the extra information of the multiple ensemble members per model) that size shift is statistically significant. Furthermore, I would also think that it would be noticeable and important to water managers. However having useful information on "how big a shift is big enough to matter to our target audience" is a place where the authors could bring value in their analysis.

RESPONSE: As we note in our discussion, there is considerable disagreement about how to use an ensemble of climate projections. Our presentation of the full distribution in our ECDFs was in part motivated by pessimistic reports that hold that the only meaningful information that can be gotten from the ensemble is its range and that range can only be treated as a representation of the minimum value of the "true" range (see, e.g., Stainforth, et al, 2007 and Wilby, 2010). We included the ensemble mean in our time-series charts because other research suggests that it provides useful information (see, e.g., Glecker, et al., 2008, Pierce, et al., 2009). We believe that water managers should consider both the mean and the range of projections. Another motivation for our presentations was to illustrate the contribution of the SRES scenarios to the overall uncertainty in runoff, which contribution is much smaller than is informally asserted in the "conventional wisdom" related to adaptation planning. The reviewer is correct that the influence of the SRES scenarios (and the trends) will be correspondingly larger if, say, the interior 80 percent of runs is used as a basis for comparison, but an examination of Figure 7 shows that such a truncation would not affect our conclusions. Further, we think that readers will be able to make that interpretation from the figures. The figure below shows the evolution of the distribution of projected change with time for the 112-member ensemble of 30-year means. The reviewer is correct that the interquartile range is much more stable than the extremes. However, we believe that Figure 7 illustrates this result and does not presume a particular perspective on interpretation of the ensemble. Nonetheless, we will include results based on the interquartile range in our findings.



We don’t think that the statistical significance of the trends really helps planners determine “how big a shift is big enough.” Any real trend will eventually become important given enough time. However, the reviewer is correct in the sense that tests of statistical significance will help planners ignore results that may not be real. Following the approach suggested by Deser (2012) we analyzed the statistical significance of the trends at 2039, 2069 and 2099. All three were statistically significant to a 95% confidence. The minimum number of ensemble members required in order that the trend be significant was 53, 32 and 36 for 2039, 2069 and 2099, respectively. These estimates implied that the size of the ensembles used by CL07 was insufficient to estimate a trend. We analyzed the projections selected by CL07 and found that for their analysis of the A2 emissions scenario, only the 2099 trend was significant; and for their analysis of the B1 emissions scenario, only the 2069 trend was significant.

The 48 members of the un-weighted ensemble described above is smaller than the minimum number of ensemble members (53) we estimated were necessary to obtain a significant trend for 2039). This observation motivated us to analyze our 100-member super-ensemble (each member being a 48-member ensemble of model runs); we found that only 35% of the ensemble members had a significant trend at 2039, whereas 99% and 100% of the ensemble members had a significant trend at 2069 and 2099, respectively.

Specific comments:

1) page 851, line 14: "...for basins that contain mountainous areas." Please specify geographical region where this conclusion is applicable.

RESPONSE: We will provide some examples.

2) Page 858, line 20: Was the tree line allowed to migrate to higher elevations as temperatures warmed? Please specify either way in the text. Allowing the tree line to migrate to higher elevations as temperatures climb presumably would allow greater evapotranspiration loss at those elevations, so neglecting the movement would arguably understate runoff declines a bit.

RESPONSE: Tree line adjustment was used, as noted in lines 19 and 20 on page 858.

3) Page 859, line 15: "differ from those reported in the earlier studies." Please briefly indicate in what manner they differ.

RESPONSE: The differences were slight; we will identify them.

*4) Figures 4, 5, and others: Spell out and describe what units you are using for precipitation. "BCM" is not a typical precipitation unit. Also, you should not use "billion cubic meters", since the word "billion" means 10^9 for most North Americans but means 10^12 for many Europeans. I note that HESS is a European-based journal, so the use of "billion" would be particularly confusing. Spell out units instead: "10^9 m**3 year**-1 totaled over the Upper Colorado River basin," or whatever. Finally, note that all flows are per unit time. Units of river flow should be, for instance, 10^9 m***3 year**-1. I know that Southwest locals usually drop the "per unit time" bit for historical reasons, but for a published article in an international journal the units should be correct and unambiguous to the entire readership.*

RESPONSE: We used units in a similar manner to Christensen and Lettenmaier, 2007. We used a volumetric measure for total basin precipitation to allow comparison to runoff so that the reader can appreciate the efficiency of the basin. We will address this suggestion with our editor. In general, we feel that if the units are clearly defined, any quantitative audience will not have trouble interpreting the import of the figures.

5) Page 861, line 17: Re Joe Barsugli's result, this is interesting, you should encourage him to publish it.

RESPONSE: We agree and will do so.

6) Figure 6a: Am I right in thinking these are 30-yr running means (since values extend out to 2100)? If so, please specify in caption.

RESPONSE: The caption specifies the chart as presenting "30-yr average streamflows". These are trailing means, that is the plotted value is the mean of the previous 30 years. We will work with our editor to clarify this caption.

7) Page 862 and Table 2: Please add a line or two specifying how these correlations were calculated.

RESPONSE: We will clarify the text accordingly.

8) Table 3 (Average projected percent change in streamflow): It is fine if you want to mix very different sources of uncertainty (emissions scenarios vs. natural internal variability and hydrological models), but not everyone wishes to do so. This table needs to be augmented with the values broken out by emissions scenario.

RESPONSE: We believe that figure 7(b) and the coloration of our time series plots makes the relative contribution of the SRES scenarios apparent.

9) Figure 7: A curious oddity about Figure 7 is the notable break in panel b) between the values above and below 0.8. This is probably a result of weighting PCM so heavily via using so many ensemble members from this model, and so an artifact of the analysis (see my major comment 2).

RESPONSE: See response to comment 2. This result is not an artifact of weighting.

10) Page 863, lines 10-12: "Figure 7 also shows that approximately one third of the scenarios suggest a wetter future for Colorado River flow." To my eye this is not an accurate summary of Fig 7. If we look at the period 2070-2099, then about 20% of the runs across all emissions scenarios show a wetter future, 60% show a drier future, and 20% show little to no change.

RESPONSE: We will re-word the text accordingly.

11) In the figures where 2 particular PCM runs are called out (8, 9, and 10) please draw one of the black lines as dashed, so a reader can distinguish the two black lines. Otherwise, it's impossible to tell if the lines cross and switch relative positions, or simply intersect.

RESPONSE: We will revise the chart as suggested.

*12) Page 865, lines 15-20: When comparing this work to previous results, the interpretation is not clear (and the text as worded is confusing) because previous results have looked at *models*, while this result looks at *ensemble members*. Note how line 16 says "there is a broad consensus among climate MODELS" while line line 20 says "about one-third of the available climate PROJECTIONS...." You can't compare the models to the projections since you have very uneven numbers of projections per model, as per my major comment #2. This section needs to be rewritten after doing the analysis with all models treated equally.*

RESPONSE: See response to comment 2. Lines 15 and 16 on Page 865 refer to Seager et al. (2007) and Seager and Vecchi (2010); both framed their discussions in terms of models.

13) Page 867, line 25-26: "GCMs differ substantially ... particularly as to the phasing of low-frequency (decadal) variability." The wording makes it sound as if this is somehow a deficiency of the models, rather than an inevitable outcome of the chaotic nature of atmospheric processes.

RESPONSE: We will note this. We have compared the two PCM runs highlighted in Figure 8 to reconstructed prehistoric flows and the variability exhibited in those projected flows is not unprecedented.

14) Page 869, line 1-5: I think you are underselling valuable aspects of dynamical downscaling based on your own particular application. Other applications may come to the opposite conclusion. For example, dynamical downscaling is presently the only way to obtain many variables for which no suitable statistical downscaling exists. If the big, statistically downscaled archive doesn't have the variable you need, it's not much use. I know the authors know this, but the text on the quoted lines makes awfully sweeping statements from only one particular point of view. Not mentioned in the text, but should be, is the fact that the climate modeling community at large tends to be skeptical of statistical downscaling methods of the future climate, seeing as how they assume stationarity in exactly the situation (anthropogenic climate change) where stationarity is strongly suspected to be lost.

RESPONSE: We will revise our comments to recognize these issues. Also, while we still wish to make the point that small-sample impact analyses cannot be support based on the findings of this paper, including, essentially, all such RCM-based assessments, we will be careful to highlight that in non-impact assessment context, RCM-based assessments are quite valuable.

Technical comments

1) *Figure 1: Labeling and legend so small as to be almost unreadable.*

RESPONSE: We will work with our editor and revise the figure as necessary to insure that it is legible.

2) *Christensen et al. 2004 is sometimes referred to as "CO4" (with the letter "oh") and sometimes as "C04" (with the digit zero). Please be consistent, so people can search the text for use of the reference and find all instances.*

RESPONSE: Thank you; we thought we had caught all of those.

3) *Page 860, line 5: spell out "ECDF" at first instance of using it. It would also be appropriate to add full phrase to legend in Fig. 3.*

RESPONSE: We will revise the text and caption accordingly.

4) *Page 862, line 6: I think you mean "Table 2" here, not "Table 3".*

RESPONSE: We will correct the text. We thought we had fixed this once.

5) *Page 862 line 27: I think you mean "Table 3" here, not "Table 2".*

RESPONSE: We will correct the text.

Anonymous Referee #2

Received and published: 7 March 2012

Summary: The manuscript presents a study that incorporates far more future climate projections, run through a hydrology model, than have been attempted yet for the Colorado basin. Given the importance of this basin, and the controversies surrounding the implications of past work on potential impacts in this stressed basin, the article could be a valuable contribution. Overall it is well written and easy to follow, but I find the interpretation to be missing some important considerations. While there is some solid analysis in the paper, the advantages of using a larger ensemble is not exploited in a way to add quantitative and convincing information to the understanding of uncertainties in projections for the Colorado basin or in how scenarios should be constructed for impacts analysis.

RESPONSE: The authors thank the reviewer for these comments. Our specific responses follow.

Specific Comments:

1) For regions like this, which have low runoff ratios (from almost zero to 0.3, if basin efficiency, p. 850 line 22, is the same as runoff ratio, which it appears to be), it may be difficult to support the claim that the approximately 5% wettening observed in the BCSD process is negligible (p. 965, line 24). Referring *to the simple runoff formulation of Wigley and Jones (Nature, 1985), discounting the direct CO2 effect, for a runoff ratio of 0.15, a 5% increase in P could result in a runoff increase of 13%, not a negligible number.*

RESPONSE: See discussion of comment 1 from reviewer 1. We will revise the discussion on page 865.

2) p. 853, line 16, the use of a large ensemble of 112 projections is presented as an improvement over the recent C07 use of 22. The newer techniques apparently produce equivalent results for the same set of GCMs (p. 864, line 20). However, Table 1 shows that, by using all 112 projections, essentially the ensemble created for this study rewards prolific GCMs with many runs archived. The cited Pierce et al study provides some perspective on this, clearly saying that adding more runs of a single GCM does much less to improve the ability of an ensemble to capture uncertainty than using runs of different GCMs. Furthermore, Pierce et al show that once somewhere around 10-14 GCMs are included in an ensemble, little is gained by adding more runs. Thus, I would not find support for the claims in this paper that a "more realistic weighting" (p. 866, line 1) has been achieved here, or that there has been a "more comprehensive accounting" (p. 866, line 24) of variability. Perhaps a quantitative assessment of this could be done by doing something like using this larger ensemble by randomly constructing ensembles of 10-14 GCMs (not more than 1 member per GCM) and seeing how the variability in projections varies among these ensembles. As it stands, I do not think this large ensemble convincingly provides any better information than the prior studies.

RESPONSE: See the responses to comment 2 and comment 3 of reviewer 1. The pessimists, cited above, will not agree that any subset is sufficient. However, the reviewer is correct that there is some ensemble size that will represent the ensemble mean with sufficient confidence. However, representing the shape of the distribution will likely require more runs. C07 represented the ensemble mean adequately in two of the three time frames, but the CDF was not well represented; they used the same models for all time frames so instability in the shape of the distribution is due to the internal variability of the models. Following the approach in Desers et al. (2012) indicates that the mean changes estimated by C07 are only significant for two out of the six cases presented (see response to comment 2 of reviewer 1).

The results cited from Pierce, et al. (2009) are based on temperature, are at a larger scale than our analysis (9 mountainous regions in the western U.S.) and are focused on detection and attribution over the observation period, so they are not directly comparable to our analysis of runoff over the projection period for a much smaller region. Our analysis in response to comment 2 from reviewer 1 indicates that determining the mean change in runoff with a confidence of 95% requires larger ensembles than indicated by Pierce, et al. (2009). We will address these differences in the text.

The reference to weighting in Line 1 of Page 866 refers to spatial resolution and is not related to the issue of weighting of GCMs in the ensemble.

3) The analysis of low frequency variability is interesting, though aside from supporting a warming against using a few projections to inform policy (p. 864, line 14; p. 867, line 26) it seems to be too qualitatively discussed to be conclusive. The Hawkins and Sutton (BAMS 2009) article includes a nice discussion of the internal variability contribution to projections; their followup article (climate dynamics 2010) shows that for precipitation at 100 years out internal variability is a minor contribution to uncertainty. If the claim here is that it is a much more significant contribution to total uncertainty, then that should be more quantitatively asserted, relative to other sources of uncertainty.

Our results are conceptually consistent with those of Hawkins and Sutton (2009 and 2011) regarding the apportionment of uncertainty for precipitation. This is not clear from the paper, but the first figure provided here in response to comment 2 from reviewer 1 shows that virtually all of the uncertainty at the end of the century can be attributed to model-to-model disagreement about the linear trend in runoff. However, our work shows that the total uncertainty in *runoff from the Colorado River Basin*, even at the end of the century, is substantially larger than the projected mean change (i.e. the fractional uncertainty is much greater than 1) which disagrees with Hawkins and Sutton results for *global mean precipitation.* We will add a figure to the paper and recognize these differences in the text.

Typos: p. 850, line 18, is -> are p. 856 line 12 a -> an; "dynamics…are" p850, l17. "…from an interpolated…" p856, l12

RESPONSE: We will make these corrections.