Hydrol. Earth Syst. Sci. Discuss., 9, C123–C131, 2012
www.hydrol-earth-syst-sci-discuss.net/9/C123/2012/

Hydrology and Earth System Sciences Discussions

Interactive comment on "The implications of climate change scenario selection for future streamflow projection in the Upper Colorado River Basin" *by* B. L. Harding et al.

Anonymous Referee #1

Received and published: 24 February 2012

Comments on "The implications of climate change scenario selection for future streamflow projection in the Upper Colorado River Basin" by Harding, Wood, and Prairie, submitted to HESS.

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.

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.

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 $\sim$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.

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.

Minor comments:

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

2) Page 858, line 20: Was the tree line allowed to migrate to higher elevations as tem-

peratures 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.

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

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 10ˆ9 for most North Americans but means 10ˆ12 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: "10ˆ9 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, 10ˆ9 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.

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

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.

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

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.

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).

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.

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.

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.

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.

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.

Technical comments

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

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.

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.

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

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

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 9, 847, 2012.