Review of version 3 of JCLI-3656 by Ryan ODonnell, Nicholas Lewis, Steve McIntyre, Jeff Condon

This manuscript is again much improved. There remain a number of editorial-level issues, most notably that the remarkable rapid warming in spring over most of Antarctica, as shown by Steig and others is barely mentioned, yet is apparently fully confirmed, or even strengthened by the new analysis. However, I will refrain from commenting further on editorial matters for now, because the manuscript remains flawed in a very basic way and will need to be re-reviewd in any case.

The main issue is that the cross-validation procedures used by O'Donnell et al. in their supplementary infromation are not valid. Although I mentioned this problem in my last review, O'Donnell et al. have not adequateky addressed it.

Here is the problem, once again:

In their optimization procedure for the 'ttls' and 'tsvd' caculations, O'Donnell are doing the following:

A) Iteratively determine an initinail best values of kgnd, with only the ground-station data.

B) Iteratively determine the best value of k-sat and determine a new ('better') value for k-gnd by removing individual weather stations, but keeping all the satellite data intact, and then comparing predicted and original weather station data.

(A) is fine, but (B) is not. This is because for the last 25 years of the 50 year data set, the satellite data provide a very very good estimate of the (removed) weather station data. For the earlier 25 years, no such data are available, only the ground stations. The resulting values of k-gnd and k-sat are necessarily some mix of the two, and overfit the data during the satellite era, and underfit it during the pre-satellite era. This is why they get a higher value for k-gnd (k=7), in spite of the evidence from (A) that this value is too high. I complained about this in my last review:

"..comparison is being made in virtually all cases between a reconstruction done during the satellite era and weather station data during that same time period. The problem with this is that the satellite data themselves provide a very strong constraint on the reconstruction during the satellite era (obviously) but no constraint at all during the pre-satellite era. The optimization of parameters is thus based almost entirely on comparison with station data during the satellite era. This may have very little bearing on the best parameters to use in reconstructing the pre-satellite era, which is of course the primary time period that is being reconstructed."

O'Donnell countered that:

"The reviewer assumes that these sets wherein half of the data is withheld for 35 stations will demonstrate the same ideal truncation parameter as would the complete set. Hence, if the partially withheld sets show an idea parameter of 5, then the ideal parameter to use for the complete set is also 5. This is not true. Withholding data from the regression necessarily increases sampling error. During decomposition, withholding data increases sampling error on that station..."

I agree that withholding large chunks of data leads to sampling error, but this is not an excuse for ignoring all the reasons for doing it! What O'Donnell et al. are doing instead is the equivalent of the much-absued 'leave one out' cross validation procedure, which is entirely inappropriate for serially (in this case, spatially) correlated data (note the 'correlation' in this case is that between the satellite data (kept in) and the weather station data (left out, one at a time)).

The problems with this approach are addressed extensively in the literature, and discussed at length in statistics textbooks, such as Wilks (Statistical Methods in the Atmospheric Sciences, pages 215-216), who writes that:

"Cross-validation requires some special care when the data are serially [or spatially] correlated. In particular, data records adjacent to or near the omitted observations will tend to be more similar to them than randomly selected ones, so the omitted observations will be more easily predicted than the uncorrelated future [in this case, past] observations they are meant to simulate. A solution to this problem is to leave out blocks [of observations...."

Wilks further notes that *in general* use of full data sets without adequate withholding will underpredict the error, resulting in overfitting.

Evidence of the problem is actually quite clear in O'Donnells own results. In their text, they write (p 22) that:

"Because the regularization parameter is fixed for the entire data set, the parameter that is ideal for the data set as a whole causes overfits during period with few predictors and underfits during periods with many predictors, yielding lower overall prediction effectiveness."

Indeed, O'Donnell et al. will find that if they restrict the cross validation *only* to the satellite era, their optimal values of k will increase, whereas if they restrict it to the pre-satellite era, they will decrease.

There may actually be an argument for doing things this way, since there is presumably some value not only in hindcasting to the pre-satellite era, but also in optimizing the infilling of weather station data during the satellite era. These are two different things, though, and they should not be confused with one another, least of all combined into one data set (overfit in the early part, underfit in the latter part). As expected from this, the new method that O'Donnell et al. now emphasize -- using their implementation of the 'regem' algorithm with individual ridge regressions for each data point, and hence individual values of kgnd and ksat being used as well, gives results in much better agreement with lower values of k (e.g. compare Figure S3 with Figure 3).

Simply put, the reasoning used by O'Donnell et al. for optimizing values of kgnd is not defensible, and not in agreement with their own independent finding using the iridge procedure. Before the manuscript can be published, these problems need to be corrected.

A few other notes:

*Both I and the other reviewers stated that O'Donnell's results agree rather well with those of Steig and others, insofar as for most areas and in most seasons the error bars overlap. O'Donnell et al. are correct in point out that it is the joint probability that is of interest, so that overlapping 95% confidence intervals do not show that two populations are indistinguishable. However, if they wish to do this calculation correctly, then they also need to take into account the errors in the regressions (that is, 1-r^2 for the 'unexplained variance'). At the moment, they are only accounting for differences in the 95% confidence levels on trends, using their mean estimates, and ignoring the errors. This probably is not a big deal -- and won't change the results appreciably -- but would be a more honest appraisal of the differences in the results.

*Figure 5 and Figure 6 are nearly identical Figure 6 would be much more useful if it showed the difference between Steig et al. and O'Donnell et al., rather than merely graying out those areas where the difference is small. Among other things, this would allow readers to see easily where O'Donnell et al. find more warming than Steig et al., and where they find less.

* The use of the 'iridge' procedure makes sense to me, and I suspect it really does give the best results. But O'Donnell et al. do not address the issue with this procedure raised by Mann et al., 2008, which Steig et al. cite as being the reason for using ttls in the regem algorithm. The reason given in Mann et al., is not computational efficiency -- as O'Donnell et al state -- but rather a bias that results when extrapolating ('reconstruction') rather than infilling is done. Mann et al. are very clear that better results are obtained when the data set is first reduced by

taking the first M eigenvalues. O'Donnell et al. simply ignore this earlier work. At least a couple of sentences justifying that would seem appropriate.

* An unfortunate aspect to this new manuscript is that, being much shorter, it now provides less information on the details of the various tests that O'Donnell et al. have done. This is not the authors fault, but rather is a response to reviewers' requests for a shorter supplementary section. The main thing is that the 'iridge' procedure is a bit of a black box, and yet this is now what is emphasized in the manuscript. That's too bad because it is probably less useful as a 'teaching' manuscript than earlier versions. I would love to see O'Donnell et al. discuss in a bit more details (perhaps just a few sentences) how the iridget caclculations actually work, since this is not very well described in the original work of Schneider. This is just a suggestion to the authors, and I do not feel strongly that they should be held to it.

In summary, this manuscript needs to be revised again, and sent again to review, before it can be considered acceptable for publication in the Journal of Climate. I emphasize again that I think that it should be published eventually, because it definitey has the potential to be a solid and oft-cited contribution. I thus I hope that the authors are not too put off by the several rounds of review. I do not think the manuscript will require more than minor re-writing to address the above criticisms (though perhaps substantial re-calculating will be needed), and I look forward to seeing a revised version in the near future.