Review of revised version of JLI-3656, *Improved methods for PCA-based reconstructions: case study using the Steig et al. (2009) Antarctic temperature reconstruction*.

O'Donnell et al. have substantially improved their manuscript and clarified a series of items that led to some confusion on my part (for example, my impression that they had detrended the satellite data). I appreciate the great amount of work that has gone into this manuscript, and the thorough documentation of the results. I also am convinced that the methods discussed are a substantive contribution to the literature and represent real improvements to the methods used in earlier work. I also think that main findings of the manuscript – that Steig et al.'s overestimate mean Antarctic temperature trends, particularly in winter in the Ross Sea region – are likely to be correct. This is important because it has implications for the causes of recent Antarctic temperature changes, for which the distribution of surface temperature variability and trends is a key test.

Unfortunately, the revised manuscript retains several important flaws in the original version, and I cannot support its publication in the *Journal of Climate* until these are addressed. The main criticism of the manuscript from my first review has not been adequately addressed, and other persistent problems lie in the way that general circulation modeling results and seasonal trends are discussed. All of these aspects of the manuscript need to be revised prior to publication, and another round of reviews conducted.

The major problem pertains to my complaint of the use of a setting of 7 for the parameter (k_{gnd}) for the infilling of the instrumental weather station data. I argued that the use of k_{gnd} =7 is inappropriate since it results in unacceptably low (in fact, negative) CE verification statistics for the key weather stations in West Antarctica. O'Donnell et al. argue in their response that results shown for different k_{gnd} settings are only a sensitivity test, and do not include full optimization of other parameters. But this has no bearing on the problem, which is that it is claimed that k_{gnd} = 7 is the best value to use because the *a posteriori* verification scores for their reconstruction are maximized. I find this argument baffling. It is simply not justified to use predictors (that is, the infilled weather station data) that are demonstrably in error, regardless of whether the reconstruction obtained is *a posteriori* superior.

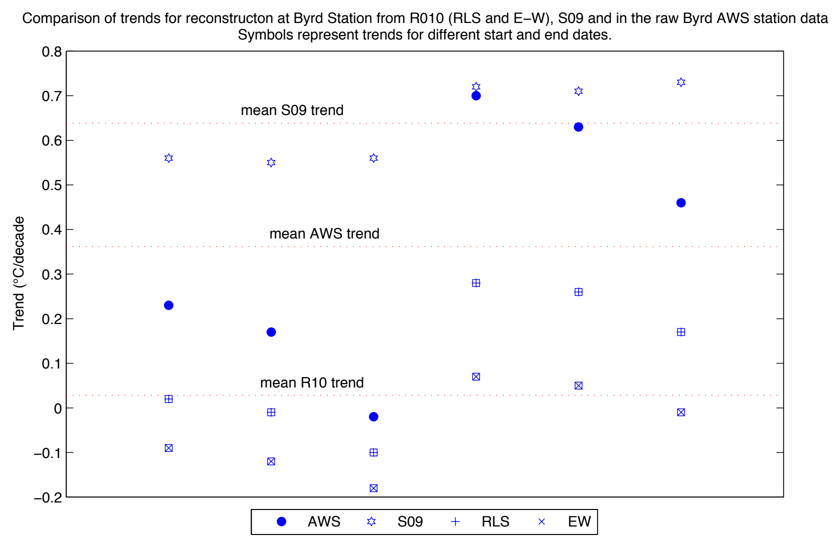
Even if this argument were valid, it is not as if the *a posteriori* verification results actually represent a *marked* improvement. As the authors themselves point out, the improvement is quite small, yet the impact on the results is quite large ("only *k* gnd = 7 yields an insignificant trend in West Antarctica"). Not only that, but the authors note "the West Antarctic regional average is likely to be ~0.10 C/decade, with a low estimate of 0.05 and a high estimate of 0.12." They nevertheless persist in showing figures in the text that are at the low end of this range.

O'Donnell et al. make two arguments in favor of showing the results as they do. First, that "the overall Peninsula average in the optimal solution – which matches ground trends – is *outside* the 95% CIs for the average in the k_{gnd} = 5 solution. These are clear indications that the k_{gnd} = 5 solution produces excessive trends."

This isn't convincing. These results do not suggest that k_{gnd} = 5 produces excessive trends *in general*. More likely is that while k_{gnd} = 5 may produce excessive trends on the Peninsula, it probably *underestimates* trends in West Antarctica. The authors acknowledge as much in a footnote: "Testing indicates the dependence on k_{gnd} ... may be the result of the fixed truncation parameter providing insufficient filtering when the number of predictors is low, and is the subject of ongoing work by the authors."

Indeed! This would apply precisely to the situation in West Antarctica, and is at least suggestive the lower truncation values used in Steig et al. are actually more appropriate there. While I appreciate that some work may be involved here, it would seem appropriate for O'Donnell et al. to address this main criticism of their work within the current work, rather than leaving it to the future! O'Donnell et al. are effectively arguing that they may have to comprise the results for West Antarctica, in order to better capture the trends on the Antarctica Peninsula. But the chief point of contention here – the primary results in Steig et al. – is West Antarctica, not the Peninsula. This is hence not a compelling argument.

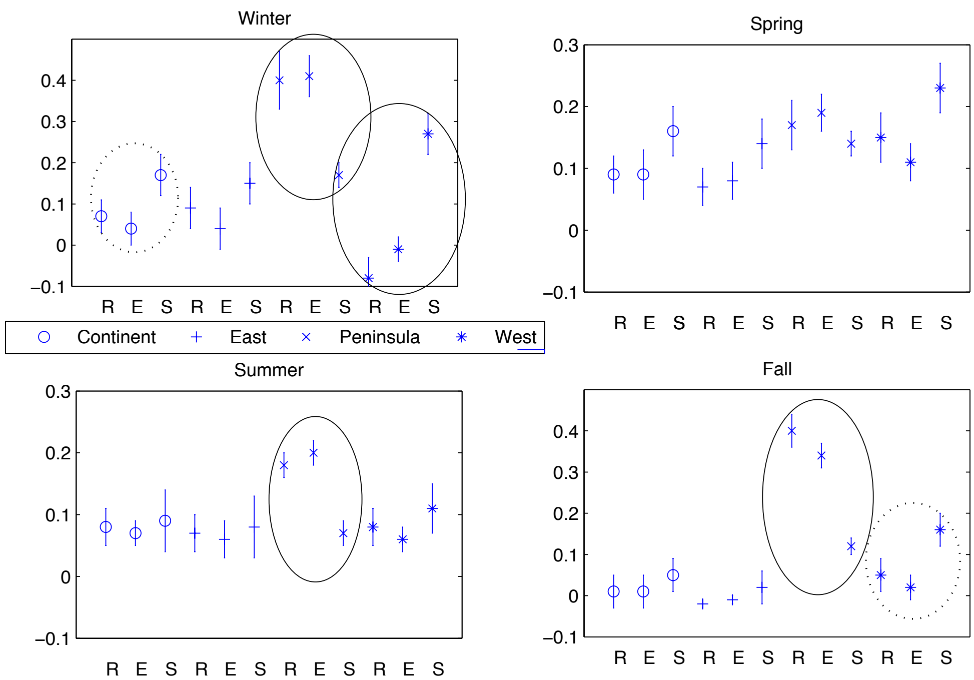
As a second piece of supporting evidence that their choice of reconstruction to show in the main text, O'Donnell et al. offer that that the difference between reconstructed and raw at the Byrd automatic weather station – central West Antarctica – are "within the 95% confidence levels" of each other. But this completely ignores the fact that their reconstructed trends are actually *systematically* lower than at Byrd AWS. And indeed, while these calculations suggest that S09 overestimates the trend at Byrd, on average by about 70%, R10 *underestimates* it on average by more than 90% (see the figure below).



First figure. Comparison of trends for various time periods at Byrd, taken from the Response to Reviewer A.

In short O'Donnell's arguments for using a low-end estimate of West Antarctic temperature trends as the basis for their figures and discussion are simply not convincing. The improvement in average CE verification statistics is quite small, while the impacts on the reconstruction are large, especially in the critical area of West Antarctica, and results in a systematic underestimate of the trends there. This is not objectively the best reconstruction, and it provides a very misleading picture both of differences between their results and those of Steig et al., and – what is more important – a misleading picture of what is actually likely happening in West Antarctica.

The second main problem is that the discussion of seasonal differences remains quite misleading. This is exacerbated by the focus on the low-end estimates of West Antarctic trends, but even if one were to accept those results, the discussion would still be misleading. The authors report "substantially different seasonal patterns", and contrast the results of Steig et al. showing "maximum warming in winter and spring for all areas" with their new results showing "maximum warming in spring and summer." The misleading implication here is that the differences in the results pertain to all areas in at least two seasons. But of course, the only reason that summer now shows "maximum warming" is that something has to take the place of winter! And in fact, average spring and summer trends are indistinguishable between O'Donnell et al. and Steig et al. in all regions. Although the magnitude of trends is underestimated in Steig et al. relative to O'Donnell et al. (though not by a significant amount in Spring) the seasonal distribution of Peninsula trends is no different. In fact, the *only* statistically seasonal differences for averages in each region occur in West Antarctica, and this almost certainly only in winter. These points are illustrated in the figure below.



Second figure. Mean trends for Antarctica, with values uncertainties taken directly from Table 4 in O'Donnell et al. R and E refer to the "RLS" and "E-W" reconstructions of O'Donnell et al., and S to that of Steig et al. Those groups where the O'Donnell et al. and Steig et al. trends do not overlap are circled. (Dotted circles are for those trends that would likely be overlapping if a more accurate reconstruction for West Antarctica were used, rather than that emphasized in the main text. Note that this is just a guess, because O'Donnell et al. do not show the impact of using other choices on the seasonal reconstruction, but it seems likely that the use of will almost certainly result in only winter being left as distinct.)

The third problem with the current manuscript is in the discussion of general circulation modeling results. I stated in my first review that the claim that there was a discrepancy between the GISS ModelE runs Antarctic temperatures was invalid. The response is that "We simply point out (accurately) that the model result S09 *cited* deviates further from both our reconstruction and the ground information than it does from the S09 reconstruction." In fact, what they say in the manuscript is that they obtain a "significantly poorer match" with GISS ModelE. But Steig et al. do not claim a "significant" match in the first place. They merely state that "the model reproduces many of the basic features of our reconstruction, with warming over most of the continent and persistent in West Antarctica". The confusion here perhaps arises because Steig et al. also note that difference in trends between West Antarctica and the Peninsula is also captured by GISS ModelE, a result not supported by O'Donnell et al. But to claim that a 'poorer match' has been found, one would need to quantify what constitutes a 'good' match, and show that by this measure, the match is reduced. Indeed, it is entirely possible GISS ModelE could matches better in some respects with O'Donnell et al's. results than with Steig et al., but this exploring this seems rather beyond the scope of either works. All that has been done here in either case is a simple visual comparison, which does not provide sufficient grounds for making any formal claims about the goodness of fit between model and observations, and even less so on whether such fit is improved or degraded.

In summary, the argument for showing the results in the main figures of the main text, in which the West Antarctic trends are only ¼ those found in Steig et al., is at best very weak, and presents a very misleading comparison. Additionally, the discussion of the seasonal trends and the comparison with climate models remains misleading, and needs to be corrected.

One might argue that this is all rather academic, on the grounds that there is insufficient information to resolve the question of West Antarctic temperature trends at all, and that the main point in O'Donnell et al. is simply to illustrate that the results are quite sensitive to parameter choices. This might be a valid point, were it not for the *primae facie* evidence that the trends in this region shown in O'Donnell et al. are too small, and had O'Donnell et al. themselves not concluded that "the West Antarctic regional average is likely to be 0.10/C decade."

My recommendation is that the editor insist that results showing the 'mostly likely' West Antarctic trends be shown in place of Figure 3. While the written text does acknowledge that the rate of warming in West Antarctica is probably greater than shown, it is the figures that provide the main visual 'take home message' that most readers will come away with. I am not suggesting here that $k_{gnd} = 5$ will necessarily provide the best estimate, as I had thought was implied in the earlier version of the text. Perhaps, as the authors suggest, k_{gnd} should not be used at all, but the results from the 'iridge' infilling should be used instead. The authors state that this "yields similar patterns of change as shown in Fig. 3, with less intense cooling on Ross, comparable verification statistics and a statistically significant average West Antarctic trend of 0.11 +/- 0.08 C/decade." If that is the case, why not show it? I recognize that these results are relatively new – since they evidently result from suggestions made in my previous review – but this is not a compelling reason to leave this 'future work'.

With respect to seasonal trends, the rather than referring to "substantial differences in seasonal trends", O'Donnell et al. should state what they actually find: smaller winter time trends in West Antarctica, and larger trends in most seasons on the Peninsula.

With respect to the GCMs, this section should probably just be eliminated.

Other issues

There are a few other aspects of the current manuscript that seems problematic, and need to be clarified or corrected.

1) Reference is made to a 63-predictor reconstruction throughout the text. But this number of weather stations only applies to the most recent period, when AWS stations

are available. None of these can be used in the reconstruction of pre-satellite era. This makes me suspicious that the verification / validation results (which are generally restricted to the later period) do not actually apply.

2) On page 22 it is stated that "Whilte cooling in S09 is restricted primarily to East Antarctica in the 1969 – 2000 period, the RLS and E-W reconstructions provide evidence of cooling in various areas of the continent for all periods analyzed including in the Ross area of West Antarctica during 1957 – 1981." This is a misleading comparison. Steig et al. only *showed* cooling during the 1969-2000 period, and this was for comparison with other work using those dates. I'm sure that Steig et al. also find cooling for other start/end dates. This comparison needs to be made and reported accurately.

3) On page 26, it is stated that "For the second experiment, we avoid infilling the ground stations altogether. Because not all of the stations are complete for any given period, the appropriate offsets are determined based on periods of mutual overlap." I find this obscure. What is actually being done here? It is not clear either in the main text or in the supplementary information.

Additional review notes from Reviewer A.

The O'Donnell et al. paper uses an iterative method to optimize various parameters to use in their reconstruction. The parameters are chosen on the basic of verification statistics, from comparing weather station data withheld from the reconstruction with estimates of the data at those same locations. For the vast majority of the 63 stations used, only of a handful of them contain data in the pre-satellite era (pre-1982) and not the satellite era (post 1982). Furthermore, in the critical area of West Antarctica, there is only one station (Byrd) that contains any data prior to the satellite era.

This means that 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.

Even more serious, in the 28-station reconstruction, during which some weather stations are withheld, the only weather station of any length in West Antartica (Byrd) is /not /withheld. I was confused on this point because it is stated in the main text that Byrd has been withheld, but the Supplementary information make it clear that this is only Byrd AWS that has been withheld. Byrd AWS only has data during the satellite era, and no other station in West Antarctica has pre-satellite data. Consequently, any optimization of parameters based on verification statistics in West Antarctica is based only on the satellite era.

In short, the parameters chosen by O'Donnell et al., which they claim to be the optimal parameters for West Antarctica (and hence Byrd), have only been shown to be the optimal parameters when they are heavily constrained by satellite estimates. It is possible, of course, that it can be shown that when Byrd Station itself (not Byrd AWS) is used, the same parameters choices will result, but the authors have not shown this, and indeed it seems rather unlikely. Given the strong evidence that the parameters used by O'Donnell et al. result in a strongly non-representative reconstruction for West Antarctica, it is critical to address this. Given the clear underestimation of trends at Byrd, even during the satellite era (as discussed elsewhere in this review), it appears very likely that other parameter choices will turn out to be optimal.

The way that verification tests and parameter choices have been made in this paper is not valid. I believe that this can be corrected, but it is very likely that it will substantially change the final results. It is therefore absolutely critical that this be done. If the authors continue to insist on presenting their existing results 'as is', then I cannot recommend publication of the paper in Journal of Climate, and it will need to be rejected. A much better outcome would be for the authors to revise the paper, properly taking these concerns into account. In this case, it will be very important that another round of reviews be conducted.