RESPONSE TO REVIEW D

We would like to thank the reviewer for the time spent reviewing our manuscript. We have found the comments to be very helpful. We do have some points of clarification we would like to make; particularly with respect to statistical significance. Our responses are below.

{1.a: The trend, in addition to being significant (as the authors note) is not statistically different than the Steig et al. trend based on the bounds of uncertainty (0.11 +/- 0.08 C/decade vs 0.20 +/- 0.09 C/decade).}

There are three primary concerns we have with this comment. In summary, they are:

1. *Comparing whether 95% CIs overlap does not yield a 5% significance level for rejection of the two-sample null hypothesis*

2. *Confidence intervals mathematically cannot be added to yield a combined p-value*

3. *The comparison the reviewer makes is only valid under the conditions of independent samples and independent errors*

We would like to take some time to explain each in turn.

1. *Comparing whether 95% CIs overlap does not yield a 5% significance level for rejection of the two-sample null hypothesis*

   Comparing the difference in location (trend) for two samples is not the same as comparing the difference in location for one sample to a fixed point. In the latter case, the fixed point – the null hypothesis – has no associated uncertainty. In the former case, *both* samples have uncertainty.

   Since mutual probabilities are multiplicative (i.e., $p_{event} = p_1 * p_2$, where the event is defined as the simultaneous occurrence of 1 and 2), requiring the difference in location between two samples to exceed the sum of their 95% CIs is equivalent to requiring a two-tailed significance level of 0.25%, not 5%.

2. *Confidence intervals mathematically cannot be added to yield a combined p-value*

   Confidence intervals for linear regressions may be expressed as:

$$CI = c * \frac{s}{\sqrt{n}} = c * SE$$

where *s* is the sample standard deviation, *n* is the number of observations, *SE* is the standard error of the mean, and *c* is a scalar multiplier that scales the standard error to a confidence interval. Since confidence intervals are simply scaled standard deviations, they cannot be added. Instead, one must take the square root of the pooled variance. The corresponding hypothesis test is the two-sample pooled-variance t-test (for samples) or z-test (for populations):

$$t = \frac{\bar{A} - \bar{B}}{\sqrt{\frac{\mathrm{var}(A)}{n_A} + \frac{\mathrm{var}(B)}{n_B}}} = \frac{\bar{A} - \bar{B}}{\sqrt{SE_A^2 + SE_B^2}},$$

where $\bar{A}$ and $\bar{B}$ are the regression coefficients for the series being compared, and $\mathrm{var}(A)/n_A$ and $\mathrm{var}(B)/n_B$ are the error variances (the square of the standard errors). For identical standard deviations and sample sizes, this yields a pooled standard deviation of $\sqrt{2} * SE$, not 2 * *SE*. This means the 5% significance level using this test corresponds to the point at which the 95% CIs overlap by approximately 40%.

3. *The comparison the reviewer makes is only valid under the conditions of independent samples and independent errors*

The null hypothesis for two-sample test discussed above is typically taken to be that the samples were obtained from the same population (with the alternative hypothesis being they were obtained from different populations). The assumptions for this test are that the two samples are comprised of *independent observations* and that the errors are likewise *independent*. The requirement of independent errors is explicit in the formula, which adds the error variances to calculate the pooled standard deviation. Variances only add when the variables are uncorrelated.

Neither assumption holds in the comparison the reviewer makes. The assumption of independent observations is violated since S09 and RO10 use largely the *same data* for conducting the analysis. Even were we to assume that the data used by S09 and RO10 was *different enough* to be considered independent, the errors are clearly not. There is at least one underlying confounding factor that destroys the independence of the errors: time. Only a subset of the population (where the population consists of all possible measurements of near-surface Antarctic temperatures from time zero to the present) is available for observation at any given time, regardless of the source of the observation. Because the possible observations are limited to a *subset* of the population and S09 and RO10 draw the samples out of the same subset, the errors in both are necessarily dependent on the time the observations were made. The errors are not independent, and the pooled variance cannot be accurately calculated by adding the error variances.

If the samples are known not to be independent and/or confounding factors are suspected, the proper test for significance is a one-sample t-test on the residuals (or, equivalently, the paired t-test). When this test is performed, only 4 (RLS) and 3 (E-W) of the 20 regional comparisons (4 regions, once with all seasons and once with each of the 4 seasons) fail to show significance at the 5% level.

Along with the three items above, from a Bayesian point of view, the value of this test is rather limited. If the samples are identical, unless the mathematical treatments – and, hence, subsequent results – are *exactly* equivalent (and in this case they are not), the posterior probability of a real difference in results is precisely 1.0. The situation is analogous to using a hypothesis test to answer the question of whether using $n - 1$ or n degrees of freedom to calculate sample variance yields different results. It is an absolute certainty that a real difference exists, regardless of the outcome of the hypothesis test or whether the difference "matters". Since the probability is already known prior to the test being conducted, one might question whether the test adds confusion rather than value.

It is important to remember that the question of "where is A located?" and "what is the difference in location between A and B?" are *different* questions that can sometimes be answered with very different precision. In practice, one is rarely able to use the former to accurately estimate the latter. The former – "where is A located" – uses the *sample* variance to calculate uncertainty. The latter – "what is the difference in location between A and B" – uses the *residual* variance between A and B to calculate the uncertainty. When the samples are the same (or nearly so), or a confounding factor can be identified, the latter question can be answered with much higher precision than the former.

In the event that one wishes to estimate the *magnitude* of the difference and associated uncertainty, knowing only that there *is* a difference is not very informative. In this case, the t-test on the residuals will yield the desired information. We agree that this information can be useful (though potentially subject to misinterpretation), and have provided both regional summaries and spatial maps that indicate whether the estimate of the difference is significant at the 5% level.

We caution that one should evaluate these results in the context that the posterior probability of a real difference in results is 1.0, regardless of the calculated significance level of the hypothesis test. The important information is the residual variance, not the *p*-value itself.

With respect to the original comment – that the West Antarctic trends between S09 and RO10 are not statistically different – when the correct test is used, they are, indeed, statistically different. The residual trend is 0.09 +/- 0.05, which, if one is curious, yields a *p*-value of about $5*10^{-5}$.

{1.b: Even at half the magnitude of Steig et al., O'Donnell et al.'s West Antarctic trend is still equivalent to 0.55 degrees of warming over the past 50 years, a number that is approximately consistent with the global mean rate of warming over the same period.}

We agree that this point should be highlighted, as well as the fact that one of the key results of S09 – significant warming over a large part of West Antarctica – is corroborated by our analysis. The text has been revised accordingly.

{1.c: That the region of warming over West Antarctica, while smaller than Steig et al. found, covers a key region where glacial recession has been most prominent: the Pine Island and Thwaites glacial drainages (e.g., Rignot et al. 2008, Nature Geoscience). While the wastage of these glaciers has been attributed primarily to regional ocean warming "eating away" at the ice where it terminates into the ocean, it is possible that enhanced surface melting may be helping to lubricate the base of these glaciers, as studies over Greenland have already shown.}

Throughout our manuscript, we have intentionally avoided discussion of Antarctic temperature change in terms of any hypothesized or known physical mechanisms and have likewise avoided hypothesizing any physical consequences should our results prove accurate. Instead, we have approached the topic from a mathematical point of view, and demonstrate what happens to the reconstruction when more appropriate assumptions are made and certain clear mistakes (like failure to calibrate the AVHRR data) are corrected. We do not postulate what would physically happen if the actual, unknown temperature history of Antarctica matched our results. While this may seem regrettable, the reason is that we do not have the requisite background to competently do this.

Because our expertise is with the mathematics, we prefer to limit our paper to the mathematics. We feel it is more appropriate to allow the community to decide how important our results are in the understanding of the physical dynamics driving Antarctic climate than for us to give [comparatively] uninformed commentary of our own.

{2. The authors do not relate their findings to physical mechanisms.}

Please see the response to 1.c.

{3. Table 3: I assume these trends span 1957-2006. If so, please state this in the caption.}

This has been corrected.

{4. Table 3: Your trends for West Antarctica for RLS and E-W in this Table (0.05+/- 0.8 and 0.04+/- 0.08) are different from what you discuss in the text (0.10+/- 0.07 on page 22 and 0.11+/- 0.08 on page 26). Is this a typo?}

Table 3 is correct . . . as are the later statements. The trends listed on page 22 and 26 are our *best estimates* based on additional results that did not appear in the main text. The confusion that this adds is the reason one of the other reviewers requested that we replace the TTLS/TSVD solutions (which were present primarily to demonstrate that the S09 infilling algorithm *can* be used to obtain more optimal results) with the ridge regression solutions (which are objectively better estimates).

As mentioned in the general note, the TTLS/TSVD solutions will be retained in the SI to demonstrate the improvement by using cross-validation to set truncation parameters.

{5. Table 4: Similar to comment #3, are these trends for 1957-2006? If so, please state in the caption.}

This has been corrected.

{6. Table 3: Assuming that the trends for West Antarctica in this table are a typo (see comment 4 above), both the RLS and E-W trends for Continental, East Antarctica, and West Antarctica are not statistically different from those of S09. This is an important statistical aspect to point out in the discussion, as it objectively demonstrates that the reconstructions in relative agreement.}

Please see the response to 1.a. above.

{7. Table 4: Similar to comment #6, the continental trends among RLS, E-W, and S09 are not statistically different in Spring, Summer and Fall, nor are the East Antarctic trends for the same 3 seasons, nor are the West Antarctic trends for summer. As for comment #6, this should be discussed in the text.}

Please see the response to 1.a. above.

{8. One of the most important results of this paper – and something that was not pointed out by S09 – is that there is statistically significant warming in summer across the entire continent, and in every region. This is a robust result among all three reconstructions (RLS, E-W, and S09). Considering that summer is currently the only season in which melt occurs over continental Antarctica, this is by far the most important season to by monitoring for warming trends due to the potential impact of enhanced melting on the mass balance of Antarctica (and subsequently sea level rise). This key result should be mentioned not just in the discussion, but also in the abstract.}

While we feel it is perfectly appropriate for us to highlight the fact that the summer warming is robust across the three reconstructions, we feel it is necessary for us to refrain from making any statement on why this may be *physically* important. As mentioned above, we feel competent to discuss the mathematics, but we do not feel we are qualified to discuss the physical theories.

{9. S09 presented results from two AVHRR reconstructions – one that used "trended" AVHRR, and another that used detrended AVHRR data. Even though S09 focused primarily on the "trended" version, they included discussion of the detrended version, including presenting the continent-average 1957-2006 trend from the S09-detrended reconstruction in the main body of the text: 0.08 C/decade (not statistically different from zero). O'Donnell et al mention the S09-detrended reconstruction briefly in a footnote on page 3, but they do not mention the resulting S09-detrended 1957-2006 trend anywhere in the text. Considering that the O'Donnell et al continent-average trends for 1957-2006 (both RLS and E-W) are in relatively close agreement

with the result from S09-detrended (0.06, 0.05, and 0.08 respectively), and that the S09-detrended result was included in the S09 paper, the S09-detrended result merits discussion by O'Donnell et al. An additional reason that the S09-detrended reconstruction deserves mention is because the good agreement between RLS, E-W, and S09-detrended suggests that potential problems with AVHRR data may have had a first-order influence on the S09-trended results, in addition to S09's statistical assumptions that are the main subject of the O'Donnell et al. critique.

We agree that additional clarity should be provided here. The magnitude of the first-order influence on S09's results is already discussed in the text and tabulated in Table 7 (now in an updated Table 4). The first-order effect of retained trends in the AVHRR data is wholly covered by Mod 2, which uses the modeled AVHRR PCs instead of retaining the original AVHRR PCs. This results in a difference of 0.02 (decreasing the S09 result from 0.12 to 0.10). Failure to use the modeled PCs is, indeed, one of the three primary statistical assumptions of S09 that we find to be less than optimal.

While the S09 detrended reconstruction does result in continental trends similar to our reconstructions, fully half of the decrease is due not to problems in the AVHRR data but rather to a combination of excessive detrending (since the ground data do not show a zero trend from 1982 – 2006) and the detrending operation removing spatial covariance information, which affects the regression results during the 1957 – 1982 period (this latter point is also recognized by S09). In this case, the combination of these effects happen to yield a result that is closer to ours, but this is by no means a general property. Depending on the location, behavior, and/or numbers of predictors, loss of covariance information could just as easily had the opposite effect. The importance of the statistical criticisms is that more appropriate assumptions can avoid the potentially unpredictable behavior associated with the type of detrending used by S09.

Additional discussion has been added to ensure that this subtle – but important – point is more clear.