Response to Second Review A

We would like to thank the reviewer for the time spent examining our revision. As the main portion of the review consists of three primary points, rather than quote from the review, we will address the main points in order:

1. Choice of $k_{gnd} = 7$
2. Seasonal patterns of change are not statistically different
3. Discussion of GCMs

The review also contains three additional points:

4. Pre-satellite verification
5. Comparison between timeframes showing cooling
6. RLS reconstructions without infilling

{1. The choice of $k_{gnd} = 7$ is suspect due to an invalid cross-validation technique and the manuscript should show other parameter choices.}

A. Toward the end of the review, the reviewer suggests that the editor should require us to display the "most likely" reconstructions in the main text, which the reviewer correctly assumes would be the ridge regression results. We agree that this is the most appropriate choice, and the manuscript has been revised to show the ridge regression results in the main text. The TTLS/TSVD results have been relegated to the Supplemental Information.

   Additional changes to the manuscript to accommodate using the ridge regression results as the primary reconstructions have been made throughout Sections 6, 7, and 8.

   Also note that the ridge results mentioned in the previous response were *multiple* (not individual) ridge regression results and were not optimized for the number of retained satellite PCs or regularization parameter. As this was originally intended as a second-check, we had chosen the faster multiple ridge regression method and used the same parameters for the ridge reconstructions as the TTLS/TSVD reconstructions. We have since performed reconstructions using individual ridge regression and have optimized the number of retained satellite PCs and regularization parameter for both individual and multiple ridge regression.

   For the optimized individual ridge regression reconstruction, the resulting best estimate for West Antarctica is $0.10^{o}C$ $decade^{-1}$. Because the individual ridge regression results display equivalent or better verification statistics and least sensitivity to removal of individual station data (including the manned Byrd station) of all of the methods (TTLS, TSVD, multiple ridge regression), this is what appears in the main text.

B. Due to the choice above to relegate the TTLS/TSVD results to the Supplemental Information, this next portion of the response is moot insofar as the manuscript is concerned. However, the criticism that the cross-validation method is incorrect is not valid, and we would like to take this opportunity to clarify this misunderstanding. We apologize in advance if this sounds repetitive, as some of the information below appears in paragraphs 4.K – 4.M of our original response.

The cross-validation testing performed at the ground station stage consisted of early/late withholding experiments using 35 stations. 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, with a first order approximation of the effect on eigenvalue determination given in North et al. (1982) of $\{n_{orig} / (n_{orig} - n_{withheld})\}^{1/2}$. For a particular station, this effect shows up in the spatial eigenvectors (which distribute the eigenvalues among each variable), where n_{orig} corresponds to the original number of data points for that station. This also has an effect on the eigenvalue/eigenvector determination for the entire data set, in which case n_{orig} corresponds to all non-missing values.

As sampling error increases, modes become increasingly mixed and additional noise is pushed into the lower-order modes. This effect is dependent on the number of data points withheld. The net result is that the optimal truncation parameter for the set in which data was withheld may not be the same as the optimal truncation parameter for the original set. If the number of withheld points is large, all one can say is that the optimal truncation parameter for the complete set is "in the neighborhood" of the optimal parameter for the partially withheld set. The size of the "neighborhood" is, of course, inversely related to the number of points withheld.

The optimal truncation parameter for the partially withheld set will approach that of the complete set as the number of withheld points approaches zero. This is the whole point of *k*-fold cross validation testing, wherein the number of withheld points per run is reduced and the number of runs increased. As the number of points withheld per run approaches zero, the optimal parameter for the partially withheld set approaches that of the complete set.

Given that our screening test involved early/late withholding – so more than 30% of the total available data points were withheld for each test and each station being tested had 50% of the available data withheld – one would fully expect that the optimal truncation parameter for that set would be different (and likely smaller than) the optimal truncation parameter for the complete set. We find that this is, indeed, the case. Similar problems are documented in the literature (e.g., Beckers and Rixen, 2003).

In terms of the present work, there are three potential means of addressing the difference between the optimal parameters between the partially withheld set and the complete set: 1) minimize an analytical function of cross-validation error; 2) increase the number of runs; or, 3) set aside a certain number of predictors that are never included in the regression as minimization targets for the reconstruction. Option (1) is not available for TTLS and TSVD, as no known analytical solution exists. As explained in the previous response, (2) is computationally prohibitive. We therefore chose option (3), in which we set aside 24 stations solely for cross-validation purposes following the reconstruction. This test yields an optimal parameter of $k_{gnd} = 7$.

Again, as mentioned in the first review response, there is no mathematical or logical reason we could not have skipped the interim, ground-station-only cross-validation testing, and instead relied on testing the full reconstruction against withheld stations. The interim testing was introduced only as a means of limiting the number of full reconstructions that needed to be performed. It was not meant to determine the optimal ground station parameter because it simply is not capable of doing so. There is no reason to expect that the parameter would accurately reflect the ideal parameter for the complete set, and many reasons to expect that the parameter would be *less* than the ideal parameter for the complete set – which is what we found to be the case.

With respect to reviewer's concerns about Byrd station, the ridge regression results push the strong Peninsula warming a bit further into West Antarctica and display a reduced area of cooling on the Ross Ice Shelf. This results in reconstructed trends at Byrd station being much more similar to Byrd AWS.

	1980 - 2003	**1980 - 2002**	**1980 – 2001**	**1981 - 2003**	**1981 - 2002**	**1981 - 2001**
Byrd AWS	0.23 +/- 0.51	0.17 +/- 0.56	-0.02 +/- 0.59	0.70 +/- 0.48	0.63 +/- 0.52	0.46 +/- 0.56
S09	0.56 +/- 0.25	0.55 +/- 0.29	0.56 +/- 0.31	0.72 +/- 0.27	0.71 +/- 0.29	0.73 +/- 0.32
RLS Ridge	0.62 +/- 0.26	0.63 +/- 0.28	0.63 +/- 0.31	0.68 +/- 0.27	0.70 +/- 0.30	0.69 +/- 0.34
E-W Ridge	0.24 +/- 0.16	0.22 +/- 0.18	0.22 +/- 0.20	0.31 +/- 0.18	0.28 +/- 0.19	0.28 +/- 0.21

Mean AWS: 0.36
Mean S09: 0.64
Mean RLS: 0.66
Mean E-W: 0.26

Additionally, we would like to point out that the RLS 1957 – 2006 trend for all of West Antarctica is 0.10, with a 1957 – 2006 trend at Byrd station of 0.27. For E-W, those values are 0.06 and 0.23, respectively. The S09 trends are 0.20 for West Antarctica – double that of RLS and three times that of E-W – but the trend at Byrd is only 0.18 . . . which is ***less*** than either RLS or E-W.

This reinforces our statements in the original response about the difficulties of using the Byrd trend as a proxy for all of West Antarctica. The reason is that a shift of just a few pixels of the border between the high warming from the Peninsula and the lesser warming/cooling on Ross changes the estimate at Byrd by a substantial amount even if the overall regional trend does not change appreciably. Because of Byrd's location, large variance, and large gaps in coverage, attempting to use Byrd AWS as a proxy for all of West Antarctica is destined to give inconsistent results.

{2. The discussion concerning seasonal patterns is misleading because the seasonal patterns are not statistically different in most cases.}

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 for 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 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.

{3. The discussion concerning GCMs is misleading.}

While we feel that the reviewer's arguments apply a different standard to S09 than our text (i.e., it is acceptable for S09 to use a qualitative, visual comparison; yet the reviewer insists that our comparison be statistically quantitative), we do agree that the GCM discussion adds little to the manuscript. All discussion concerning GCMs has been removed.

{4. RO10 have little pre-satellite verification.}

We agree that the amount of pre-satellite verification is minimal. To correct this, we have re-run verification statistics by withholding one station at a time from the ground station infilling, performing the reconstruction, and calculating verification statistics to the withheld ground station (we additionally took this opportunity to verify the optimal parameter for k_{gnd}). This allows us to calculate verification statistics to *every* measured ground station temperature value over the entire reconstruction period. These statistics are summarized in the main text and fully tabulated in the SI. None of our results, optimal value for k_{gnd}, or conclusions are altered as a result of this additional testing.

{5. Comparison of timeframes showing cooling is misleading.}

This text has been removed.

{6. The description of RLS reconstructions without infilling is unclear.}

For the RLS reconstructions without infilling, the baseline is first determined using the long-record stations. The remainder of the stations are then offset to have the same mean as the nearest long-record station for the time during which the observations overlap. However, with the inclusion of the ridge regression results as the primary reconstructions, the value of this test is reduced. To prevent confusion, reference to the RLS reconstructions without infilling have been removed from the text.