Response to Third Review A

Please see our responses below.

{1. 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.}

> We do not intend to change the manuscript. We introduce the seasonal patterns of change discussion by explicitly saying the differences between S09 and our reconstructions in spring and summer are "minor to negligible", and provide a table and three separate figures showing as much. Later, we additionally state that S09 show the largest warming in winter and spring and we show the largest warming in spring and summer for all areas except the Peninsula – again highlighting similarities. For each of the three paragraphs on seasonal change, we use the seasonal *similarities* – not differences – to introduce each paragraph. Additionally, we repeat the spring and summer similarities in the conclusion. We believe this is sufficient emphasis.

{2. The cross validation procedure used is invalid because:

a. RO10 keep the satellite data intact, which forces the estimate to be closer to the last 25 years.
b. Comparisons are made between a reconstruction done during the satellite era and weather station data [AWS stations] that appear primarily only during the satellite era.
c. Cross-validation requires care when data are serially or spatially correlated.
d. The iRidge results are in better agreement with lower values of k.
e. RO10 combine overfitting and underfitting into a single data set.}

> We do not intend to change the manuscript based on the above.
>
> First, the primary results in the paper are now the iRidge results. These results have no dependence whatsoever on the reviewer's concerns (which is even acknowledged by the reviewer). These results also show superior verification statistics as compared to the TTLS / TSVD results, which have been relegated to the Supporting Information. The only purpose the TTLS / TSVD results now serve is to show that – if one uses cross-validation rather than a heuristic to determine the truncation parameter – very similar results to iRidge are obtained.
>
> We emphasize that *even were the reviewer's concerns valid* (which they are not), none of the results or conclusions in the main text are affected. The reviewer's concerns are related solely to the TTLS / TSVD reconstructions. We find it odd that the reviewer would suggest non-publication based on supplementary work, upon which the results and conclusions do not depend.

Second, the reviewer’s arguments concerning the cross-validation performed on the reconstructions in the Supporting Information are incorrect. We also note that the reviewer’s concerns with respect to cross-validation have been somewhat of a moving target. Initially the concern was that the cross-validation was performed to infilled data, which we showed to be a misreading of the text, Supporting Information, and figures. Then the concern became that there was little pre-1982 validation to support the truncation parameters (which is even more applicable to the S09 reconstruction, where S09 computed cross-validation statistics to the *satellite data*, used only the AWS stations as verification targets for the AWS reconstruction, and did not withhold the manned Byrd station for the 15-station restricted reconstruction). The reviewer recommended using the ridge regression results instead.

Not only did we comply with the request to replace the TTLS / TSVD results with ridge regression (making this whole discussion moot insofar as our results and conclusions are concerned), we demonstrated that performing all of the requested additional cross-validation that was *not* present in S09 (including withholding the manned Byrd station) yielded the same results as before. By this time, the amount of cross-validation performed in support of our results greatly exceeded that performed by S09 (as well as most published reconstructions of similar scope), had significantly expanded temporal coverage as compared to S09, demonstrated that there was little difference in the pre-1982 and post-1982 timeframes, and confirmed that the reviewer’s concerns with the original validation method were misplaced.

Now the concern has shifted again, with the reviewer injecting an irrelevant and incorrectly framed spatial autocorrelation argument which, if accepted under the reviewer’s conditions, would effectively render cross-validation impossible for regression analysis. In order to make this argument, the reviewer needed to present his own, nonstandard definition of leave-one-out cross-validation, add words to a quote that are inconsistent with the meaning of the original text, and mischaracterize the use of the satellite data in our reconstructions – mischaracterizations which had been previously addressed in both our initial review response and revisions to the manuscript.

As none of the results in the paper are affected in any way by the reviewer’s concerns, we feel this should be a sufficient response. We have, however, included a section at the end of this response that shows specifically how each of the reviewer’s latest concerns are baseless.

{3. 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.}

This calculation is only applicable for pooled-variance tests (which are joint-probability tests) or when comparing the regression results back to the original data. In our previous reply, we showed these tests were invalid when either a confounding factor (such as time) exists or the observations are not independent. If either disqualifying condition is met, then the appropriate test is the paired t-test or, equivalently, a t-test on the residuals. We chose the latter. This test does not, as the reviewer claims "account for the difference in the 95% confidence intervals". It accounts for the mean and variance of the difference between the test objects.

Because the test objects are not independent (they utilize the same observations and use similar regression techniques), then changes to those observations are likely to modify the regression coefficients in a similar manner – not in an independent fashion as the reviewer's proposed correction requires. If both regression coefficients move in the same direction, then the contribution of regression coefficient error to the variance of the residuals is reduced. There is no simplistic correction factor or theoretical distribution known to the present authors to account for coupled behavior of the regression coefficients when an iterative technique like expectation-maximization is employed to determine them. To get more precise than what we have already done would require Monte Carlo analysis, in which a change to the observations would be made, both analyses repeated, and a new comparison performed.

Based on relatively large changes in the observations (i.e., withholding more than half the data) resulting in very similar reconstructions and the fact that the posterior probability of a real difference in results is 1.0, such an analysis would serve only to more firmly reject the null hypothesis that the results are the same as the number of trials is increased. In this light, we agree with the reviewer's comment that *some* modification would be a more honest appraisal of the results; however, the appropriate modification would show more significant differences – not fewer.

{4. 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.}

We agree with this comment in principle. However, because the boundaries of statistical significance are so complex, we found that difference plots with the boundaries for statistical significance outlined (or shaded instead of overplotted) were almost uninterpretable. We feel it would be easier for readers to reference Figs. 3 and 4 for trend differences rather than attempt to put all of that information into Fig. 6.

{5. 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.}

We have two topics to discuss here. First, reducing the data set (in this case, the AVHRR data) to the first M eigenvalues is irrelevant insofar as the choice of infilling algorithm is concerned. One could just as easily infill the missing portion of the selected PCs using ridge regression as TTLS, though some modifications would need to be made to extract modeled estimates for ridge. Since S09 did not use modeled estimates anyway, this is certainly not a distinguishing characteristic.

The proper reference for this is Mann et al. (2007), not (2008). This may seem trivial, but it is important to note that the procedure in the 2008 paper specifically mentions that dimensionality reduction was ***not*** performed for the predictors, and states that dimensionality reduction was performed in past studies to guard against collinearity, not – as the reviewer states – out of any claim of improved performance in the absence of collinear predictors. Of the two algorithms – TTLS and ridge – only ridge regression incorporates an automatic check to ensure against collinearity of predictors. TTLS relies on the operator to select an appropriate truncation parameter. Therefore, this would suggest a reason to prefer ridge over TTLS, not the other way around, contrary to the implications of both the reviewer and Mann et al. (2008).

The second topic concerns the bias. The bias issue (which is also mentioned in the Mann et al. 2007 JGR paper, not the 2008 PNAS paper) is attributed to a personal communication from Dr. Lee (2006) and is not elaborated beyond mentioning that it relates to the standardization method of Mann et al. (2005). Smerdon and Kaplan (2007) showed that the standardization bias between Rutherford et al. (2005) and Mann et al. (2005) results from sensitivity due to use of precalibration data during standardization. This is only a concern for pseudoproxy studies or test data studies, as precalibration data is not available in practice (and is certainly unavailable with respect to our reconstruction and S09).

In practice, the standardization sensitivity cannot be a reason for choosing ridge over TTLS unless one has access to the very data one is trying to reconstruct. This is a separate issue from whether TTLS is *more accurate* than ridge, which is what the reviewer seems to be implying by the term “bias” – perhaps meaning that the ridge estimator is not a variance-unbiased estimator. While true, the TTLS estimator is not variance-unbiased either, so this interpretation does not provide a reason for selecting TTLS over ridge. It should be clear that Mann et al. (2007) was referring to the *standardization bias* – which, as we have pointed out, depends on precalibration data being available, and is not an indicator of which method is more *accurate*.

More to [what we believe to be] the reviewer’s point, though Mann et al. (2005) did show in the Supporting Information where TTLS demonstrated improved performance compared to ridge, this was by example only, and cannot therefore be considered a general result. By contrast, Christiansen et al. (2009) demonstrated worse performance for TTLS in pseudoproxy studies when stochasticity is considered – confirming that the Mann et al. (2005) result is unlikely to be a general one. Indeed, our own study shows ridge to outperform TTLS (and to significantly outperform the S09 implementation of TTLS), providing additional confirmation that any general claims of increased TTLS accuracy over ridge is rather suspect.

We therefore chose to mention the only consideration that actually applies in this case, which is computational efficiency. While the other considerations mentioned in Mann et al. (2007) are certainly interesting, discussing them is extratopical and would require much more space than a single article would allow – certainly more than a few sentences.

{6. 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.}

We hold a rather different opinion of which algorithm is a “black box”. Tikhonov regularization (which is called ridge regression primarily in the statistical literature, but Tikhonov regularization elsewhere) has a substantial body of published literature dating back to the 1960s. Much more has been written concerning ridge regression than any other shrinkage estimator of which the present authors are aware. It is a far more common tool in applied mathematics, statistics, and signal / image processing than TTLS.

Schneider’s 2001 paper spends but two paragraphs (page 866) on TTLS in a 12,000+ word article. The remainder of the article is dedicated to EM and ridge regression. We disagree rather strongly that the ridge regression procedure in Schneider (2001) is not well described – it is quite thoroughly described. On the other hand, TTLS is hardly mentioned, and most of the important calculations that appear in the algorithm are not even shown, much less discussed.

The reason much of the supporting information is gone is because the algorithm that actually requires additional explanation is TTLS, and the TTLS reconstructions are no longer the source for the results and conclusions of the paper. We feel the ridge regression algorithm is well-documented – both in Schneider (2001) and elsewhere – and adding additional explanation would be redundant.

Specific responses to the reviewer's cross validation concerns:

{RO10 keep the satellite data intact, which forces the estimate to be closer to the last 25 years. [paraphrased]}

The reviewer claims (without any evidence or supporting calculations) that "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 concept seems to be taken from our newly-added paragraph describing the problems with TTLS, but the reviewer has taken it out of context or misunderstood the intent. The fit problems we described are *the ground station data to itself in the complete <u>absence</u> of satellite data.* In this case, if there were a fit problem, it would be *overfitting* during the pre-satellite era, not underfitting, as there are fewer available station values with which to produce a fit. This increases the influence of individual station error on the solution and results in a greater possibility of overfit. Satellite data has nothing to do with this. Additionally:

1. There is no satellite data at all during the pre-satellite period, so one wonders how something that does not exist could be "underfit" or "overfit". If the reviewer means the relationships determined in the satellite era might be different in the pre-satellite era, then this particular difficulty is shared not only by every single reconstruction that has been performed, but also by every single regression analysis that has been performed where the entire population of the predictors and predictands is not known. A key diagnostic to determine this effect is to compare validation statistics (Table 3) using the two different periods. It is clear that this is a minimal concern for our iRidge reconstructions, and, since the TTLS / TSVD agree most closely at k_{gnd} = 7, it is therefore a minimal concern for them as well.

2. As mentioned in the previous response, we additionally re-performed the cross-validation for the TTLS and TSVD reconstructions and split it into pre-1982 and post-1982 portions. Just like iRidge, the statistics for the two periods are very comparable and *exceed the skill of the raw AVHRR data* even in the pre-1982 period. We noted that including the pre-1982 verification resulted in the same optimal value for k_{gnd} of 7. The reviewer appears to have missed this information from our second response.

3. The fit problem the reviewer proposes – were it to exist – would affect only the E-W reconstructions, as the RLS reconstructions utilize only the spatial component to weight the station data and do not rely on <u>any</u> temporal fitting between satellite and ground data. We provided clarification on this in both previous review responses and in the main text. Since the E-W and RLS yield the same basic results and the RLS version cannot have the reviewer's hypothesized fit problem, it also follows that the "overfitting" of the satellite data is not a concern for the E-W reconstructions, either.

4. With respect to the reviewer's claim that we "keep the satellite data intact", the reviewer has again misread the text and our previous explanations concerning the

RLS method. During the prediction, *only the grid points corresponding to the stations actually used* appear in equation (10). The corresponding satellite data from the withheld station(s) *is not included*. This is both a logical and mathematical requirement. Were the unused grid cells included, the column span of $\mathbf{L}^{\mathbf{T}}$ would not match the row span of **Y,** and vector **a** could not be computed. Unlike the E-W reconstructions – which regress the PCs – absolutely no satellite information that corresponds to the withheld stations is used for the RLS regression. Moreover, absolutely no regression of ground station data against satellite temporal data is performed for RLS, rendering irrelevant the reviewer's argument that the satellite provides a good fit to the ground data.

5. We would also like to point out that our pre-1982 verification statistics *exceed the fit of the raw AVHRR data to the ground station data during the satellite period*. This was also true of the TTLS / TSVD reconstructions, and was noted (and tabulated) in previous revisions of the text. If our reconstructions were the result of over/underfitting the satellite data, then one might wonder how the reconstructions fit even the pre-1982 data *better* than the raw or rank-reduced satellite data fits the post-1982 data. "Overfitting" and "underfitting" are, by definition, a worsening of the fit, not an improvement. This misconception is perhaps due to the reviewer conflating our reconstruction methods (which take great care to minimize use of AVHRR temporal data) with S09's method, which directly utilizes AVHRR data.

{Comparisons are made between a reconstruction done during the satellite era and weather station data [AWS stations] that appear primarily only during the satellite era. [paraphrased]}

The reviewer has failed to note that we re-performed the verification studies using *all of the pre-satellite ground data* – not just the AWS data (and, in the case of the verification reconstructions, *entirely excluding* the AWS data) – and doing so produces the same result that $k_{gnd} = 7$ is optimal. The implication that our cross-validation statistics are only to AWS stations, and, hence, do not account for the 1957 – 1982 period, is incorrect. This is already clear in both the previous review response and the main text, which now breaks out the cross-validation statistics into pre- and post-1982 periods.

{Cross-validation requires care when data are serially or spatially correlated. [paraphrased]}

The reference to Wilks and subsequent explanation of the serial correlation problem with cross-validation is, unfortunately, contradictory and incoherent. First, the reviewer somehow equates temporal correlations between the satellite and ground data to spatial autocorrelation. Given that spatial autocorrelation is a measure of spatial homogeneity *within* a data set (or a set of *regressors*) – not across two separate data sets – and that the reviewer's measure is temporal rather than spatial, the proposed equality is absurd.

Second, the reviewer states that with "spatially correlated" data, one must use blockwise rather than random withholding. To support this, the the reviewer quotes Wilks, who states that the solution to serial correlation is to leave out blocks of observations. First, we agree with Wilks. Second, this has nothing to do with spatial autocorrelation. Third, we left out *the entire station*. If this is not blockwise withholding, then nothing would suffice.

The reviewer has inappropriately used the quote from Wilks by adding bracketed comments that are neither present nor consistent with the original text to imply that spatial autocorrelation (which, as before, the reviewer erroneously equates with the temporal relationship between the ground data and satellite data) and serial correlation affect cross-validation in an identical manner. Wilks most certainly does not make this argument (nor does any other reference of which we are aware). We are not sure what the reviewer intends with this particular argument regardless, as the proposed conclusion of blockwise withholding is something we have already done. The only implied requirement that we can think of is that the reviewer is attempting to say that spatial autocorrelation mandates *retaining* some unspecified minimum portion of the validation target, which is both illogical and unsupported by the reviewer's own reference.

Contributing to the confusing argument, the reviewer seems to misunderstand the difference between spatial and serial [auto]correlation. In the case of serial correlation, each observation *within a variable* has some degree of dependence on the previous observation, meaning that the effective number of observations is less than the total available observations. Random withholding does not capture this dependence, which motivates the use of blockwise withholding. This is the subject of the Wilks quote.

Spatial autocorrelation, however, has nothing to do with serial dependence of the observations within a variable; rather, spatial autocorrelation is a measure of dependency *between variables*. In the case of regression analysis, it is a measure of the dependency *between regressors*. While it is analogous to serial correlation, the effect is quite different. Rather than reducing the effective number of observations, spatial autocorrelation reduces the *effective number of regressors.* Adjustment for spatial autocorrelation requires dimensionality reduction, spatial weighting, resampling, or some combination thereof, not blockwise withholding. These techniques increase calibration period error in order to decrease the validation period error, and they are why Schneider (2001) suggests that RegEM – which reduces the dimension of the regressors – can provide better results over standard EM even if the problem is not rank-deficient.

We would be stunned if the reviewer could produce any literature that supports the notion that temporal blockwise withholding (which we do anyway, to address the proper concern of serial correlation) is preferred over dimensionality reduction (inherent in RegEM) or weighting/sampling techniques for addressing spatial autocorrelation in regression analysis.

However, perhaps the reviewer's intended meaning is that we should leave out *more than one station* at a time, as he specifically mentioned that "*in general* use of full data sets without adequate withholding will underpredict the error". In this case, we point out that the verification reconstructions – which also yield the same conclusions as our full reconstructions – leave out all of the AWS stations (35 of the 63 stations). Even after leaving all of these stations out, the optimal parameter computed by *withholding the remaining stations* per the procedure in the main text and our second response is still k_{gnd} = 7, and the resulting reconstructions still show more skill *to each of the additional withheld stations* (i.e., as each of the remaining 28 stations are withheld) than either the full S09 reconstruction (with no stations left out) or the raw AVHRR data.

If leaving out more than half of the predictors and *then* additionally leaving out the remaining predictors in sequence, testing each predictor by entirely withholding it, and preferentially leaving out predictors during the satellite-ground station overlap does not sufficiently address the reviewer's cross-validation concerns, then we submit that the reviewer has established an impossible set of criteria. We note that the cross-validation procedure in S09 is far less rigorous than this. S09 measures skill from the modeled PCs (which are not even used in the reconstruction post-1982) *back to the satellite data*. Since the satellite data is the dependent variable in the S09 reconstruction and S09 retain the 1982-2006 satellite data in its entirety, not only does the S09 reconstruction not measure skill to the proper variables, but it *discards the only period of the modeled PCs for which they have validation information.* Moreover, S09 never evaluate how any of these statistics change if the truncation parameter is changed.

In summary, the reviewer attempts to use semantics to redefine "leave-one-out" cross validation – which has a clear mathematical definition of leaving out one *observation* at a time (e.g., Wahba, 1990, Golub et al. 1979) – to categorize leaving out an entire *station* as "leave-one-out" cross-validation. With this categorization, he proceeds to incorrectly equate a degree-of-freedom correction for *serial* correlation to an improperly specified *spatial autocorrelation* problem (that would already be resolved by the dimensionality reduction in RegEM regardless), glosses over the fact that one cannot possibly address a spatial autocorrelation problem by blockwise temporal withholding, ignores that if the real concern is serial correlation, then withholding the entire predictor as we do is the *optimal* blockwise method (since it entirely eliminates the issue of temporally adjacent withheld and retained points), and finally takes our own discussion concerning the problems inherent with the fixed truncation parameter inherent to TTLS out of context to somehow equate it to satellite fitting problem when our discussion concerns fits during which the satellite data is entirely *excluded*.

{The iRidge results are in better agreement with lower values of k.}

By any objective measure, this is false. While the patterns for all variants are similar, the only two values of *k* that yield cooling on Ross shown in the iRidge reconstructions during the both the 1957-1982 period and the full period are 7 and 8. For RLS, the regional trends are only comparable to iRidge for *k* = 7 or 8 (indeed, the Peninsula trends for the two lower values of *k* have lower 95% CIs that exceed the iRidge estimate). For E-W, the results are even more clear, with only *k* = 7 producing trends in all areas that are consistent with the iRidge estimate. Regardless, in both cases, the <u>*k* = 8</u> results are more consistent with iRidge than either *k* = 5 or 6. We find it inconsistent that the reviewer – who absolutely insisted on a quantitative comparison with respect to GCMs – continues to use qualitative visual comparisons to draw conclusions that are unsupported by the quantitative data (which is available on the immediately preceding page) and then attempts to use these comparisons to suggest non-publication when the results and conclusions *have no dependence* on the reviewer's comparison anyway.

{RO10 combine overfitting and underfitting into a single data set. [paraphrased]}

We agree that TTLS overfits in some periods and underfits in others. We specifically point this out as a *disadvantage* of TTLS. Use of the RegEM TTLS algorithm *requires* a fixed truncation parameter that unavoidably causes fit problems if the number of predictors changes. The reviewer seems to be implying that the S09 reconstruction is better; however, we note that the S09 reconstruction simply underfits the ground data for *the entire reconstruction*, leaving the unaltered satellite PCs to exercise undue influence. How is this better? Our argument is that fixed truncation parameters should be chosen based on minimizing cross-validation error – which, unless the ideal truncation parameter is exactly the same for all time steps (and the reviewer admits it is not), necessarily results in some underfits and some overfits. It is not possible to avoid this effect using a fixed truncation parameter.

We are unsure what the reviewer is attempting to argue. He seems to equate the rigidity of the truncation parameter to some measure of fit between the satellite data and ground data. However, as mentioned before, neither the RLS reconstructions nor the iRidge reconstructions depend on any measure of temporal fit between the satellite data and ground data, and all verification statistics are calculated to ground station data anyway – even for the E-W reconstructions.

Lastly, we re-emphasize that the only purpose for the TTLS and TSVD reconstructions is to demonstrate that cross-validation (rather than using a heuristic) to select the truncation parameters leads to reconstructions that are similar to our main results. None of the reviewer's comments – or, indeed, the existence of the TTLS / TSVD reconstructions in the SI – affect any of the conclusions in the text.