Response to Review C

We would like to thank the reviewer for the time spent examining our paper. We greatly appreciate the helpful suggestions. Our responses, and descriptions of changes to the text, are below. For clarity, statements extracted from the review will be *italicized* and enclosed in brackets {}.

{*1. Items to think about:*}

> {*1.1 The differences between the authors' proposed alternative reconstructions are interesting and significant (according to Figures S15 and S16), yet they are not discussed in the concluding remarks. Given that these two reconstructions are similar, but different, and the authors comment that they are also similar, but different, from the work of Monaghan et al (2008) and Walsh and Chapman (2007), I would think there was an interesting discussion to be had about what this tells us about the uncertainty arising from analysis differences between these four. The authors may believe this is outside the scope of the paper, but I think it would provide the reader with useful information.*}
>
> We agree that such a discussion would be worthwhile, but also feel that much of it would be outside the scope of the present work. A discussion concerning the differences between Monaghan et al. (2008), Chapman & Walsh (2007) and ours would require the same level of deconstruction as the S09 reconstruction for each, as it is not superficially apparent why the differences exist. There is unfortunately insufficient space to do this. In terms of the differences between our own reconstructions, some additional discussion has been added to both the main text and SI based on early results from a work-in-progress. We anticipate being able to publish more comprehensive results in the future. In the meantime, we hope the additional text provides a useful – if preliminary – assessment.
>
> {*1.2 I would be very interested to know how faithfully the new reconstruction reproduces the considerable trend experienced by the Peninsula over the full period. Reconstructions often contain lower trends than the observations they are based upon. A co-located average comparison, i.e. averaging Peninsula station temperature anomalies and reconstructed temperature anomalies only at the locations of those stations and then comparing the temporal evolution of these two averages, might show something interesting and be considered for inclusion somewhere if it does.*}
>
> This is an excellent suggestion. We have added a new figure to the text that compares the 7 most complete Peninsula stations to the RLS, E-W, and S09 reconstructions (Figure 6 in the revised text).
>
> {*1.3 Having concluded my review I am now wondering about the title of the paper and its organisation. The paper is much more than an analysis of someone else's work and I'd like the authors to rethink the title. I'm also now wondering if it should be a paper in two parts, which would allow the authors to make more of the very good material currently in*

the Supplementary Material. If it were split into two parts, the first part would be quite short and be an analysis of the Steig et al work, but the second part could be longer and provide a much clearer documentation of the proposed new analysis (I assume here that the authors would recommend the RLS over the E-W reconstruction). Could the authors please consider whether or not they think that reorganization would be appropriate? It shouldn't be a big job, as the Supplementary Material is well-written. If that is undertaken, I'd be happy to take another look at it.}

We appreciate this comment, and agree that the title could be more appropriate. The amended text uses the title: "Improving methods for PCA-based reconstructions: case study using the Steig et al. (2009) Antarctic temperature reconstruction".

For the latter suggestion, we have amended the text to include the most important information from the SI into the main text (such as Figures S15 and S16 and significant additional information about the RLS and E-W methods). With these changes, understanding the main text does not require the reader to reference the SI. We believe a single paper to be the most appropriate, as analyzing S09 without demonstrating what happens when the concerns are addressed leaves the reader without solid conclusions.

{*2. Items to provide greater clarity:*}

{*2.1 The authors discuss the erroneous temporal information in the AVHRR data set. I presumed this was mainly a high frequency problem after reading the main paper, but I see from the Supplementary Information that a residual trend is found. Perhaps this could be clarified in the main text?*}

The amended text now includes the residual trend between the AVHRR data and the ground data. The trend from NOAA – 7 to NOAA – 14 is a statistically significant 0.19 +/- 0.16 deg C/decade.

{*2.2 I did not understand what the term "on-grid" meant when reading the main paper and I'm still not sure I understand it now. Could the authors please clarify this term on first use?*}

The intent was to indicate those stations located within one of the AVHRR grid cells. However, this is more efficiently expressed by simply stating that stations within 150 km of a grid cell were used. The text has been amended to remove the confusing "on grid" phrase.

{*2.3 Introduction, second paragraph: I suggest maintaining the same style in a) - e), i.e. using "augmentation of" and "estimation of".*}

This has been corrected.

{*2.4 Whilst I understand what the authors mean by "reconstituting the extracted PCs with their corresponding spatial eigenvectors", many readers will not. I like to think of*

this as building up a temperature field from a weighted sum of eigenvectors - I think painting that kind of picture might clarify the concept.}

This is a good suggestion, and we have incorporated this into the text.

{*2.5 Section 2: "We restrict our replication of ... S09 ... to steps that follow ..." according to the earlier summary of the S09 process, these are all their steps. Without having re-read S09 at this point, I didn't understand the point of this sentence.*}

We neglected to mention that the first 2 steps of the S09 procedure are the cloud masking and regridding of the AVHRR data. This has been corrected.

{*2.6 Section 3a: "These factors all highlight a need to calibrate the AVHRR ... (or vice versa)" Not vice versa, I think, given what the aforementioned "factors" are.*}

We agree, and this has been amended.

{*2.7 Section 3b: I believe the authors use "x" and "X" in Equations 1 to 6 to describe rather different types of things. I suggest that it is confusing to the reader to use the same letter (albeit in a different case), because they naturally assume that they might be similar quantities. Could the authors think of an alternative to one of them?*}

This was an oversight on our part. Later on in the text (and also in the code), we use "Y" to denote the ground station matrix. We have revised the earlier equations to use the same notation – which has the added benefit of being consistent with the notation used by S09. We thank the reviewer for noticing this.

{*2.8 Section 3b: "X=(A|B)" is used at the start of this section, whereas a variant of "X=(A B)" is used in Equation 6. I found the latter more straightforward so, if it is correct to use this, please do in both cases.*}

The latter notation is now present throughout the text.

{*2.9 Section 3b: The equations in the Supplementary Information are generally better explained than the equations in the main text and so are easier to follow, e.g.:*

- Please define what n and p are when introducing the "nxp matrix of observations"

- Please define U, Lambda and V.

- What is xk in Equation 4?}

The suggested amendments (along with other unclear variable assignments) have been made.

{*2.10 Even after re-reading the paper and reading the Supplementary information, I don't understand why the AVHRR PCs would be temporally incomplete. I don't see any evidence of AVHRR data gaps in Figure S3. I can see that there would be missing values*

in some locations due to cloud, but I wouldn't expect that to preclude calculation of a PC for the field as a whole, unless the data gaps were very numerous. Could the authors please explain that?}

Our choice of words here was poor. The intent was to indicate that the AVHRR data exists only for the latter half of the reconstruction (1982 – 2006) and is entirely absent in the earlier half. However, the sentences seem to imply that gaps exist in the data (which, as the reviewer observes, is not true). We have rewritten this to avoid confusion.

{*2.11 Section 3c: When "suggests the possibility of mutual reinforcement" first appeared in the text, I found it rather opaque. I understand the concept now, but perhaps its meaning could be clarified earlier in the paper?*

2.12 Section 3c: "These observations present a major difficulty in ascribing a calibration function to RegEM". Because of the use of words that might be more commonly intended to mean something different, this sentence is hard to understand. I suggest something along the lines of "These features present a major difficulty in using RegEM for calibration." I also found the rest of this paragraph difficult to follow, so could the authors please consider rewriting it?

2.13 Section 3c, last paragraph: suggest change "components" to "PCs" to be consistent.

2.14 Section 3c, last paragraph: "truncation parameter k will be less effective" because if errors are random, they are found in the lower order modes and so removed. Without stating that, it is assumed the reader understands this.}

We appreciate the reviewer noting these issues. The entire calibration section of our paper was somewhat obtuse. We have rewritten the section to more clearly explain and delineate the calibration issues.

{*2.15 Section 4: this section was hard to understand, particularly the last sentence in the first paragraph. I recommend that it be rewritten.*

2.16 Section 4: Does "i" indicate a station location?

2.17 Section 4: I wondered if because the authors had neglected the contribution of a series to itself, this explained any of the discrepancy found.

2.18 Section 4: "less than half are weighted similarly" this is subjective - can the authors be more specific?

2.19 Section 4: I understand the last sentence after having read the Supplementary Information, but didn't without.}

We entirely rewritten this section and incorporated the relevant material from the SI. For the question concerning the contribution of a series to itself, this is always unity for the

series being investigated. Neglecting it will not change the relative magnitudes of the remaining coefficients, and therefore cannot account for any discrepancy.

{*2.20 Section 5: I suggest replacing "boundary conditions" by "shape of the boundary"*

2.21 Section 5: Suggest replacing "the statistical authority cited by S09 as their source for determining k" by "North et al (1982)", as I feel the former comes across as a bit argumentative.

2.22 Section 5: I noted here that a picture of the EOFs would be useful to refer to, so perhaps part of S4 could be reproduced in the main paper? I also regretted not having S2 in the main text to aid interpretation of Figures 2 and 3.}

This section needed to be revised due to a misunderstanding on our part concerning the reference used by S09 to determine truncation parameters. In the course of that rewrite, the above concerns were addressed. Additionally, Figure S2 has been moved from the SI into the main text.

{*2.23 Section 6b: "we address this .. by simply infilling a matrix" How?*}

We have clarified that we used RegEM for this purpose.

{*2.24 Section 6c: I don't understand the first sentence.*}

We have added some introductory text to clarify this section, and eliminated some extraneous and confusing text.

{*2.25 I wanted to see whether or not the spatial trends shown in the main paper were significant. I think it helps to understand the importance of the differences. I was pleased to find it in the Supplementary Information, perhaps that information could be added to Figures 4 and 5?*}

We agree, and have moved Figures S15 and S16 into the main text.

{*2.26 Why are the diagonal elements of Table 1 not 1?*}

During regularization, all components greater than *k* are discarded. The values on the diagonal (like the remainder of the values) are the linear sum of the first *k* retained components. Values close to 1 indicate that the retained components explain a great deal of variance in that data series. Values closer to zero indicate that the discarded components are necessary to explain the variance in that data series.

{*2.27 I wondered if the verification results were sensitive to the choice of which 28 stations were in the subset?*}

Yes. Removing too many of the long-record length stations for verification degrades verification statistics markedly. Given that there are only about 10 stations that have reasonably complete records over the entire reconstruction period, the results are naturally sensitive if they are withheld for verification. With that exception, however, the verification statistics are fairly insensitive to which stations one chooses to withhold.

{*Figure 1: could the black and blue time series be plotted on different axes? It looks to me like there is a constant offset in the blue line, but because of the scale, I can't see whether or not there is and what size it might be.*}

Based on other review comments, Figure 1 has been moved to the SI and replaced with a higher-resolution version.

{*2.30 Figure 2: blue and black are indistinguishable here. Please state in the captions that the locations are the locations of the stations.*}

This has been amended, and Figure S2 (which is the geographic key) has been moved into the main text to aid interpretation of these figures.

{*2.31 Figure 3: does this show different locations to Figure 2?*}

Yes, since the E-W reconstructions use a larger station set than S09. However, with the exception of the two Southern Ocean stations used in S09 (Orcadas and Signy), all of the S09 stations are present.

{*3. Comments related to Supplementary Material*}

{*3.1 Section S2d: is the difference in trend of 0.08 C/decade statistically significant too? It is hard to detect this trend in the plot.*}

It is not significant. The trend from NOAA-7 to NOAA-14 is statistically significant (0.19 +/- 0.16 after correction for serial correlation of the residuals), but the full-period trend is not. This has been clarified in the SI.

{*3.2 Section S2d: the authors discuss the "dramatic change in NOAA-11 ICT variability and the mid-2000 jump in NOAA-14 variability" however, I don't see a significant effect of the latter in the bottom panel of Figure S3, so I'm not sure it's worth mentioning.*}

Figure S3 is a based on a running 24-month sample, so rapid events would not show clearly. However, as we do not show a figure that complements the above text, we agree that they contribute little and we have removed that portion.

{*3.3 Section S2d: "NOAA-9 demonstrates" should be "NOAA-7 and NOAA-9 demonstrate"*}

This has been corrected.

{*3.4 Section S2d: for completeness, the authors might recognise here that the station data is also likely to contain some errors (I note this is done later, so a reference forward might be made).*}

This is a good suggestion and we have added a forward reference at the end of S2.d.

{*3.5 Section S3: "Due to the vastly larger number of data points in the Peninsula" I think this is overstating it. The number is much larger, but "vastly" is over-doing it, I think.*}

We have amended the text to drop the "vastly".

{*3.6 Section S3: I understood here why the authors had used the West Antarctic stations alone in one of the verification statistics in the main paper, but I think that needs to be explained better there, as it wasn't clear to me when reading that stand-alone.*}

This is an excellent comment. We have broken out a separate section on West Antarctic trend sensitivity in the Results section and incorporated some of the material from the SI, such that the main text can stand alone.

{*3.7 Section S3: the issue with the Peninsula stations swamping the West Antarctic stations discussed here is not explained clearly in the main paper, perhaps a few more words could be added on this?*}

Based on other review comments, we have decided to remove the AVHRR eigenvector discussion from the main text, as this detail is not essential to the criticism that the Peninsula contamination in S09 is significant. However, because some readers might be interested why this affected West Antarctica more than East Antarctica, we have left this section in the SI.

{*3.8 Section S5: I got to the end of the last paragraph and wondered eagerly if the spatial structures were incompatible. Later on I found the answer, but it might be helpful to look forward the outcome here.*}

We have cleaned up the wording in S5 such that it fits all on one page, followed by the relevant tables and figures. This arrangement should make it easier for readers to navigate the information for this section.

{*3.9 Caption for Figure S7: Is the sentence "For the AVHRR data ..." essentially repeating the previous sentence?*}

This has been corrected.

{*3.10 What do italics denote in Table S3?*}

We apologize for the formatting error. The italics have been removed.

{*3.11 Section S6, second sentence following equation S1: should "estimation-maximization" be "expectation-maximization"?*}

This has been corrected.

{*3.12 Figure S10: why is there a separation in results after k=5?*}

We are still investigating why the covariance networks are more susceptible to overfitting than correlation networks. A plausible reason is that, in a covariance setting, sampling error on the high variance stations results in estimates that are further from the mean than a correlation setting, where all stations have a variance of unity. The bottom panel of this figure (now corrected) shows the results of random withholding (which minimizes the increase in sampling error due to withholding data), and seems to support this hypothesis. Additionally, the high-variance stations (which tend to be in the Antarctic interior) may contain a higher percentage of local weather "noise" that is not useful for prediction. Because these stations are preferentially selected in a covariance setting, this results in poorer predictions. Further work is required in this area.

{*3.13 Section S9b: I am worried by the fact that the authors have selected c=0.1 as the optimum parameter, yet it is the lowest value tried and there is no evidence from Figure S14 that the verification is converging at c=0.1. Why not test lower values, given that this is the case? I am not convinced that c=0.1 is optimum here.* }

Because the optimal number of included AVHRR spatial eigenvectors is greater than the number of stations, a non-zero regularization parameter is required. Based on testing with fewer AVHRR spatial eigenvectors, the difference in trend between c=0 and c=0.1 is in the third significant digit. Therefore, determining c with a resolution finer than 0.1 does not result in any material benefit. We did test c = 0.05, but that did not always provide sufficient regularization to prevent computational singularities during inversion.

{*3.14 Section S9c: it is very interesting that the covariance results are more sensitive to kgnd than the correlation results and this is shown nicely by Figures S18 and S20. I would like to understand why. Any ideas?*}

See response to 3.12.

{*3.15 Section S9c, last para: Please rephrase the sentence starting "Most importantly, trends ..." because I don't think the authors intended to imply that the trends for all locations are similar*}

This has been corrected.

{*3.16 Section S9d, last para: I didn't understand why the truncation parameters in the RLS and E-W reconstructions should not provide filtering.*}

This was poorly worded and has been corrected. The intent was to explain that the nearest-station method is unfiltered, but that the RLS and E-W method *are* filtered.

{*3.17 There is no text to go with S15 and S16. I think they are worth discussing, particularly in the context of my first comment.*}

We agree, and have moved them into the main text, along with providing some discussion.

{*3.18 I liked figures S21-S26. I felt they should be more prominent.*}

We felt that having them at the back of the SI – rather than buried in the middle – would make it easier for readers to find them. Additionally, as the SI is laid out in approximately the same order as the main text, moving them to a different location may cause additional confusion.

{*3.19 using "R squared or CE" in the column headings in Table S5 is confusing. It suggests they are interchangeable, yet footnote d in the Table suggests they are not.*}

Mathematically, R^2 and CE are equivalent measures. R^2 compares reconstructed explained variance in the calibration period to the mean of the actual data, while CE compares reconstructed explained variance in the verification period to the mean of the actual data. The form is the same. In our case, stations were either entirely included (i.e., the entire record length of the station is the calibration period) or entirely withheld (i.e., the entire record length of the station is the verification period). Rather than have an R^2 column – which would be entirely blank for the verification stations – and a separate CE column – which would be entirely blank for the calibration stations, we felt it made sense to combine them and indicate whether the measure was to an included or withheld station. We have clarified this in the footnote.

{*4.1 Should there be a subscript j on the A tilde in Equation 6?*}

This has been corrected.

{*4.2 Section 5: "influenced based visual similarity" should be "influenced by visual similarity"*}

This portion of the text has been removed.

{*4.3 Section 7c, last para: "Table 6" should be "Table 7"*

4.3 Section 7c, last para: "Table 6" should be "Table 7"

4.5 Section S4: "TIR" here "IR" should be subscript.}

These errors have been corrected.