Second referee report: Influence of Urban Heating on the Global Temperature Land Average Using Rural Sites Identified from MODIS Classifications
by Wickham et al.

I now realize that the aim of this paper is much more narrow than I had originally thought it to be. The Rhode et al. paper at http://berkeleyearth.org/pdf/berkeley-earth-averaging-process.pdf is the "flagship" in which the BE data construction and methodological details are presented, and this paper is only focused on the urban heating issue. Consequently I can see that some of the technical details I asked for are written up elsewhere, and in their response, the authors rely heavily on the existence of the Rhode et al. paper to justify leaving so much out of their own. That being the case, however, all the credit attached to the new data set construction and methodology belong to the Rhode et al. paper, so the only grounds for deciding on the publishability of this particular paper is whether it is a good analysis of the topic of urban contamination of the surface record. A weak analysis on an old data set would certainly not be publishable; a good analysis on a new one probably would. This paper presents a weak analysis on a new data set, and the novelty of the data set cannot be weighed in its favour.

I had given some suggestions about how to fix the problems in the methodology in my earlier review, including one idea that would have been relatively straightforward to implement using easily-available data. Unfortunately the authors have made no methodological improvements, and the arguments they offered for keeping their technique unchanged are, as I will explain, unpersuasive. So it will come as no surprise that my view of this draft remains unchanged from before.

On page 7, the sentences on lines 114 to 121 represent an improvement in the discussion of the range of findings in the published literature. But having drawn attention to the contradictory results in previous published analyses, the authors offer a weak explanation as to why some teams find an effect while others do not. They first suggest the issue comes down to a lack of adjustments in CRUTEM products. This is inconsistent with what CRU says about its own data. The CRU web page (http://www.cru.uea.ac.uk/cru/data/hrg/) presents two products: TS and CRUTEM. The TS series are not subject to adjustments for non-climatic influences, and for that reason users are cautioned not to use them for climate analysis, and instead users are directed to the CRUTEM data based on its supposed additional processing:

> **Question One**
> **Q1**. Is it legitimate to use CRU TS 2.0 to 'detect anthropogenic climate change' (IPCC language)?
>
> **A1**. No. CRU TS 2.0 is specifically *not* designed for climate change detection or attribution in the classic IPCC sense. The classic IPCC detection issue deals with the distinctly anthropogenic climate changes we are already experiencing. Therefore it is necessary, for IPCC detection to work, to remove all influences of urban development or land use change on the station data….If you want to examine the detection of anthropogenic climate change, we recommend that you use the Jones temperature data-set. This is on a coarser (5 degree) grid, but it is optimised for the reliable detection of anthropogenic trends. (http://www.cru.uea.ac.uk/cru/data/hrg/timm/grid/ts-advice.html)

Brohan et al. (2006, p. 6) don't claim that their data are unadjusted, they say that the raw data may have been adjusted but they do not have original records so they can't say what was done. Jones and Moberg (2003) say of the CRUTEM2 data set (emph added):

> "All 2000+ station time series used have been assessed for homogeneity by subjective interstation comparisons performed on a local basis. **Many stations were adjusted and some omitted because of anomalous warming trends and/or numerous nonclimatic jumps** (complete details are given by Jones et al. [1985, 1986c])."

So the CRUTEM products are not as raw as the authors imply, even if it is difficult for users to understand what the particular adjustments were. Even if CRUTEM3 is unadjusted, McKitrick and Nierenberg (2010) used both versions 2 and 3 in their analysis, with clear similarity in results between them, so the issue is moot.

The authors then try (lines 117-121) to draw a distinction between analysis of local trends and the global average. I don't follow the logic here, since it is a global sample of local trends. Widespread problems in the local records will carry over to the global average. Had this been properly noted the sentence in question would read (emph added): "McKitrick and Michaels (2004, 2007) and McKitrick and Nierenberg (2010) also focus on finding the heating signal in <u>a global sample of</u> local trends rather than evaluating the effect on a global average." Stated in this way, it would be clear that the authors are saying that the discovery of a global pattern of problems in local trends does not imply a problem exists in the global trend, which is a pretty weak position to take.

The more obvious, and plausible, explanation for the difference in results across the different studies is the difference in testing methodologies. I demonstrated this in my previous report, showing that one set of results can be shown to emerge as restricted estimates from a model whose general form indicates the opposite conclusions, and the restrictions can be rejected.

The authors dismissed this demonstration by saying something that I confess I can't make much sense of:

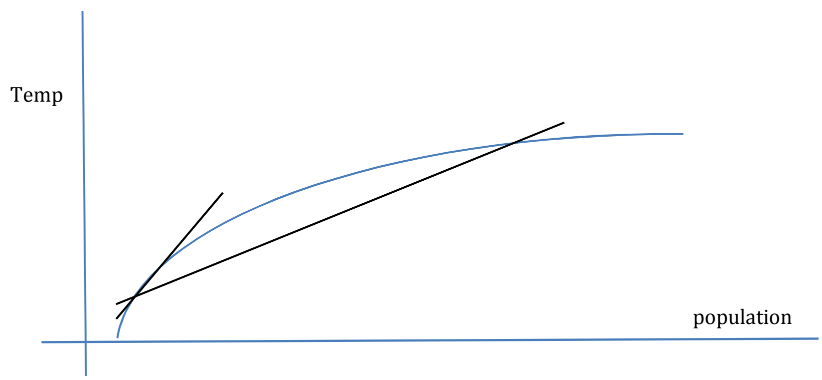
> The empirical demonstration is interesting, but we view it as a way to do the "trend analysis" part of our paper "correctly". That isn't our goal. Our conclusions are based on the Berkeley Average on the very-rural stations compared to all stations.

Are they really saying it is *not* their goal to do the trend analysis "correctly"? I don't think I have ever encountered a situation where authors have said of their own work that it was not their goal to do it correctly. I am sure they did not mean this, but I draw a blank at trying to figure out what they did mean. Later they say:

> We are not asserting that surface temperature data are unaffected by urbanization, but that a global average based on data that includes stations that may have warmed due to urbanization is not significantly different to one based only on stations that are assumed not to contain urban effects.

I suspect that any reasonable reader, upon completing the paper, would be startled to learn that the authors did not intend to assert that surface temperature data are unaffected by urbanization. I think the above sentence was meant to say something like: "We are not claiming there are no contaminating influences in individual locations, only that they are too small and isolated to affect the global average." Unfortunately the whole issue is whether their methodology reliably supports this conclusion, and in this draft they have done nothing to deal with the evidence that it does not, instead they simply assumed the problems away.

The authors dismissed the conceptual example with an argument that is both incorrect and beside the point. Ignoring their observation that the convex function could be a square root (which would look pretty much the same), they say that the argument relies on each tangent line being defined over an infinitesimal domain of the same length, and that the population change in the urban region would likely be larger in magnitude than in the rural region. If the diagram were redrawn to reflect this case, the underlying point would emerge even more strongly, since for any convex function, an arc connecting two points has a flatter slope than does a tangent at the first point, and the farther apart the points, the flatter is the arc line. Hence a steeper slope in the rural sample is what we would expect if urbanization were a large effect in the data and the urban population increased more than did the rural population.



The *point* of this argument, to which the authors did not respond, was that their method is, in principle, unable to support the conclusions they draw, since their findings are consistent both with the absence or the presence of a significant urban warming bias. Nothing in their response or their revised paper addresses this problem. Instead they seem to rule out one interpretation by assumption and then claim to have proven the other interpretation.

Moreover their empirical results are becoming harder and harder to reconcile with their own preferred interpretation. Between the last draft and this one, the negative rural/all trend divergences got even larger. Over the full sample the trend difference was -0.10C/100yr before, and is now -0.14 C/100yr. On the subset of records ≥30 yrs, the trend difference was -0.12 C/100yr before, and is now -0.15 C/100yr. The authors downplay the negative divergence in their conclusions, and try to portray it as essentially a zero difference, but the number reported in the Conclusion, -0.10 ±0.24 C/100yr, seems to have been derived in a very different way than by differencing the trends in Table 1, not least since the standard error is on a far larger scale. (Unfortunately the reader is not informed how this was computed. Was it a time series regression on the post-1950 in Figure 5B?)

The larger the size and sign of such divergences, the less consistent their data get with their preferred story, namely that there is no difference between samples; but they more consistent they get with the existence of a global-scale urbanization contamination problem as conjectured in the above figure.

Or maybe there are other explanations. For instance, the very rural data set is heavily dominated by stations in North America and northern Europe (Fig 2). If recent regional warming in northern midlatitudes is stronger than in the SH, the very rural sample is more heavily drawn from faster-warming regions. Then the Kriging method has to do more work to compensate for this. So one interpretation of the stronger relative warming in the very rural sample is that the Kriging method is not providing an adequate offset for the sample change through the spatial weighting system. In other words, we have to assume the validity of their method to accept the interpretation of their results, since otherwise the results could just as well be interpreted as evidence against the validity of the methods. The authors do not present any evidence to suggest they considered how to rule this possibility out.

I had hoped that in response to my previous review the authors would have made some attempt to strengthen their methodology and rule out rival interpretations, and I even suggested a relatively straightforward test they could have done using data readily available online. The authors chose not to do any of these things. As before, I would be willing to re-read a major revision that deals with the methodological problems, but at this point the authors appear determined to leave their methodology unchanged, so not surprisingly my original recommendation against publication is also unchanged.

Signed review: Ross McKitrick