Clim. Past Discuss., 10, C1341–C1351, 2014
www.clim-past-discuss.net/10/C1341/2014/

Interactive comment on "Investigating uncertainties in global gridded datasets of climate extremes" *by* R. J. H. Dunn et al.

R. J. H. Dunn et al.

robert.dunn@metoffice.gov.uk

Received and published: 21 August 2014

We thank the reviewer for their detailed comments and address each in turn below.

Reviewer

Apparently a key point in this paper is one that is made on page 2116, lines 26–28, which is that HadEX2 "really only captures the changes in the indices in regions where there are stations". This is essentially a statement that the DLS (decorrelation length scale) is small. But do the authors know that this is really the case? The DLS is not a unique number, even for a single variable, such as surface air temperature. If the primary interest is in changes in index behaviour on decadal or longer scales (as would be the case for detection and attribution studies), then the relevant DLS is that of decadal

(or longer) averages of indices – and these will have longer decorrelation length scales than seen in the monthly or annual index values that are provide in HadEX2 and its relatives. A relevant paper is North et al, 2011 (J Climate, doi: 10.1175/2011JCLI4199.1).

Response

What we were trying to emphasise with this point is that in HadEX2, some grid boxes have been extrapolated from nearby (or not so nearby for large DLS values) stations. Where extrapolation has occurred, then the grid box value depends on stations which could be far away, but no new information has been created in this process. We cannot know what the stations would have measured had they been there. Hence for large regions which have no stations, care should be taken when interpreting HadEX2. However, we agree that by using the ADW gridding and the DLS calculated from the observed data do give confidence that the interpolation will be reasonably accurate in most areas. We have rephrased the section on page 2116 to illustrate this point and also added a reference to North et al. (2011).

Our primary interest in this study is not the absolute behaviour of the indices over time or over different regions, but rather the relative behaviour of different versions of the final index fields resulting from the different methodological choices. Currently in HadEX2, the DLS are calculated from the relevant annual or monthly values. We agree that for other studies a different DLS calculation may be relevant, but not practical to provide multiple versions depending on the timescale of the final application. We have made a comment regarding this in Section 2.

Reviewer

A second comment is that I found the structure of the paper to be a bit challenging. While I understand that the authors chose examples pertaining to different indices to illustrate points, it would perhaps have been more helpful to work through the paper consistently using two indices – one that is representative of temperature indices and another that is representative of precipitation, due to the fundamental differences in station number, distribution, and presumably DLS values, for these two variables. The authors start out by discussing a classification of indices, but they don't really seem to use the classification for the generalization of findings to indices within classifications as they go along.

Response

We appreciate this point and agree that presenting/summarising results for all 27 ETC-CDI indices that are included in HadEX2 can be quite challenging. However, even for temperature and precipitation there is a range of different indices (representing intensity, frequency and duration of events) with different characteristics. As it is our aim to estimate uncertainties for the whole range of indices included in HadEX2, it would seem restrictive to focus on 2 indices only (1 temperature, 1 precipitation). As it is not possible to include material for all the indices in the paper, we provide all figures for each of the indices in supplementary material. However, we have also tried to include more information from the groups of indices which were not previously discussed in our updated manuscript.

Reviewer

A third grumble is that one of the diagnostics of "robustness" used in the paper is the coefficient of variation of trends amongst variants of a dataset. As the authors point out – the coefficient of variation is tough to interpret when the thing in the denominator is near zero. I don't have a good solution, except to suggest that a better diagnostic might be a ratio of variances based on an analysis of variance (ANOVA). One estimate of uncertainty in a trend would indeed be the variation in the trend estimates amongst datasets. But it should also be possible to accompany each individual trend estimate with an estimate of its uncertainty based on variation within the specific dataset that it describes. Those "internally" estimated uncertainties could be pooled, and if the between dataset variance is large compared to the pooled internal estimates, then it seems to me there would be evidence that there is a problem with dataset uncertainty.

Response

This is what we have attempted to do in Table 2. We show the statistical range in global "average" trends and compare these to the range in trends from each of the choices. Also, if the denominator is near zero, then if this does result in large values for the coefficient of variation, this is important; users may try to read too much into a small but not-quite-zero trend when this value is very uncertain.

Reviewer

A fourth general comment is that comparison, particularly of structural differences, is probably made difficult through the (unavoidable) use of a common base period. That has the effect of bring differently constructed datasets together during the base period (through the constraint that anomalies necessarily average to zero over this period). That constraint does not exist outside the base period, so all else being equal, spread will still be greater outside the base period.

Response

We agree that centring the time series plots to a common base period does give the impression of greater variability outside of this span, and have added a caveat to this effect in all relevant figure captions. As you say, it is unavoidable, and in some cases is built into the indices during their construction. We expect that users will use HadEX2 to look at changes over time rather than the actual amounts, and so the common base period will have less of an effect, though for presentational purposes it is important.

Reviewer

I hadn't previously encountered the first differencing method, but it occurs to me that a concern with the method (which I'm sure must have been written about somewhere) is that errors accumulate as differences are cumulated (error variance would necessarily increase in time). This perhaps contributes to its steeper trend, eg, as seen in Figure 14. Since there is no restriction on the direction in which one cumulates differences, why not calculate and accumulate differences for time running into the past rather than towards the future? The denser networks of modern times would, presumably, contribute smaller errors, and thus error variance growth would occur relatively more slowly from the start of the accumulation point if time ran backwards, than if it runs forwards. Differences between forward and reverse accumulation could give a further indication of robustness, or the lack thereof.

Response

This is a very good point, and we have run another version of the FDM in reverse. This is now included in the time series plots as well as all subsequent summary diagrams.

Reviewer

I hadn't previously encountered the term "jackknifing" in quite the sense that it is used in this paper. In statistics, this refers to a particular subsampling approach that is used for variance estimation and bias correction. To avoid any confusion – since the authors do not specifically use the jackknife in this way – it would be better simply to refer to subsampling (which also more directly describes what was done without using jargon).

Response

We have updated the name of the section and how it is referred to in the text.

Reviewer

Finally, I thought the conclusion on line 8, page 2132, was a bit bleak. The methods could, at minimum, be applied to high resolution climate models or reanalyses sampled to mimic observations and then compared with "truth" as represented by the full–field model or reanalysis output that was sampled. This would provide information about the adequacy, or lack thereof, of existing observing networks for estimating changes in different types of indices. For example, see Wan et al (2012, JGR, doi: 10.1002/jgrd.50118).

Response

This sentence has been removed in light of comments by Reviewer 2. We agree that the methods have uses elsewhere, but we were trying to emphasise the point that HadEX2 is limited by the (underlying) data availability (especially for gridding methods that interpolate).

A few specific comments (page and line number):

Reviewer

2109, 6–7: Assuming perfect correlation at zero distance is a bit strong. Often one sees a "nugget effect" relating to instrumental and related (e.g., siting, etc) uncertainty. That effect is presumably small for temperature, but you could imagine it being larger for precipitation.

Response

We place a datapoint at (0,1) as this is what was done in HadEX2, however we do not force the line to pass through this point (see the comments by Reviewer 2). During initial investigations, this point was omitted, but to ensure consistency with HadEX2, it was reinstated. However the effects were small. We have added an extra sentence to clarify this section.

Reviewer

2110, 11: Do you mean "normalized" (which involves subtracting the base period climatology and dividing by an estimate of the standard deviation), or rather, do you mean "centered"?

Response

We mean centred, and have update the text accordingly.

Reviewer

2110, 19: Insert “at” before “around”.

Response

Done.

Reviewer

2111, 12: Correlation coefficients describing correlations between what and what?

Response

Between HadEX2 and the versions with different completeness requirements – the text has been clarified.

Reviewer

2111, 19: The use of “masks” is a bit unclear – is it being used as a noun or a verb?

Response

In this case the mask is a noun. The sentence has been clarified.

Reviewer

2111, 25: Caption should be plural. There are quite a lot of minor editoral issues of this type, particularly in this part of the paper, so I suggest proofreading it carefully again.

Response

Text updated.

Reviewer

2112, 3: Insert “judged” ahead of “likely” (it’s your assessment of what is likely, not an absolute.

Response

Done.

Reviewer

2112, 13: “weighting function has been” → “weighting parameter has been” (since the discussion is about the value of m).

Response

Done.

Reviewer

2113, 22 and 2144, etc: I found the terms in the caption for Figure 5, used various places in the text, to be confusing. The conjunction of words “mean detrended correlation coefficient” and the shorthand “detrended–r” both allude to detrending of the correlation coefficient. Rather, this is about correlation coefficients that are calculated from detrended time series.

Response

We have corrected the description of this quantity where it appears throughout the text.

Reviewer

2115, 3: I know what is being alluded to by “regional workshops”, but many people will not, so I think this indirect reference to the ETCCDI and the APN workshops needs to be clarified a bit.

Response

A footnote has been added to explain these (see also comment from Reviewer 2).

Reviewer

2116, 24: But the weighting of the stations could be quite different – so the “not surprising” result is nevertheless, not completely expected.

Response

This section has been expanded and clarified.

Reviewer

2118, 14: “knock–on effect” is colloquial jargon, I think, that might not be understood by all CPD readers.

Response

Sentence simplified.

Reviewer

2120, 7: Doesn’t this formulation for the spatial correlation function essentially include a “nugget” at zero?

Response

All of the fitted forms allow for a nugget at zero (see Fig. 11) – it is only in the binned observations that perfect correlation is artificially set at zero distance.

Reviewer

2120, 16: What makes it physically reasonable?

Response

This section has been expanded upon in light of comments from Reviewer 1. The phrasing now is “closest fit to the data”.

Reviewer

2123, 11–12: The words here suggest that all selected stations end up with the climatology of the reference station – so somehow uncertainty in adjustments that are applied across the DLS are reflective of the uncertainty in the climatology of that one reference station. Doesn’t that counter the benefit of grid averaging somehow?

Response

In the reference station method (RSM), the stations are merged adjusting the mean over the common period of the two stations rather than over a fixed climatology period, which allows for more variation. Also distance-weighted averages rather than just simple means are re-calculated at the addition of each station. However, we do agree with the concern that this method requires that the reference station is clean and representative. None of the four gridding methods presented herein are ideal.

Reviewer

2123, 27: This seems to be a rather strong assumption. It might be reasonable for temperature (at least in jurisdictions that have worked on temperature in more than perfunctory manner), but has anyone really been able to tackle precipitation homogenization?

Response

We have expanded on this section (also in light of comments by Reviewer 2) and made more cautious comments as to the data quality and homogeneity.

Reviewer

2124, 26–27: Doesn’t Fig 14b draw “relatively good” into question?

Response

We have discussed the early and late periods of the time series separately to clarify where the agreement is good and where it is not (see also comments by Reviewer 2). The later period does have relatively good agreement, which was our focus when writing the text, but agree that the early period does not. We hope that this is now clearer.

Reviewer

2126, 14–16: I don't understand how a less dense station network should lead to a larger DLS. The DLS is presumably a property of the index rather than a property of the network. Subsampling would, presumably, lead to greater uncertainty in DLS estimates, and perhaps that is what is reflected?

Response

On the whole, the DLS is a property of an index, but in the sub-sampling runs, values for the binned correlations (Fig. 11) will change depending on which stations remain in the network, resulting in different estimates of the underlying DLS value. These could result in slightly higher or lower DLS values than in the full network, showing the uncertainty in the DLS values.

Reviewer

2128, 27–28: Need to explain how the "DLS fitting method" is "a second order Taylor expansion".

Response

We have placed this in a footnote, and clarified that it is the second-order polynomial which is the Taylor expansion of the exponential.

Reviewer

2159: I didn't see a discussion of Figure 20 in the text.

Response

All Taylor diagrams now are presented as a single figure in the main text.

Interactive comment on Clim. Past Discuss., 10, 2105, 2014.