Review “Investigating uncertainties in global gridded datasets of climate extremes” by Dunn, Donat and Alexander.

This paper describes an ambitious and very useful study of uncertainty in HadEX2 and related datasets, and before providing comments, I would like to congratulate the authors on their accomplishment.

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).

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.

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.

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.

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.

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).

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).

A few specific comments (page and line number):

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.

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"?

2110, 19: Insert "at" before "around".

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

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

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.

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

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

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.

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.

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

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

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

2120, 16: What makes it physically reasonable?

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?

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?

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

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?

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

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