Review – "Changes in statistical distributions of sub-daily surface temperatures and wind speed" by Dunn et al.

General Comments:

This paper provides a detailed analysis of changes and trends in HadISD temperature and wind data. The analysis uses station data that have been quality controlled and, in some cases, homogenized. Apart from the use of quantile regression to describe changes in the upper and lower tails of the distribution, the paper uses simple statistical methods. Findings are, generally, not surprising, although some trends are found that are somewhat surprising. Comparisons are made with similar papers that have used other datasets.

While this kind of work is fundamental and absolutely necessary, I found the author's apparent decision to limit themselves only to the description of the data unsatisfying. Unfortunately, the paper does not offer physical insight concerning the changes that are observed, and only speculates about the causes of differences between findings reported here and those reported in previous papers based on other datasets. It raises the issue of measurement resolution in the context of quantile regression, but does not concern itself with the impact of measurement resolution on the higher order moments, or how changes in measurement resolution might produce changes in those moments. Also, while the methods are simple, describing them precisely in prose can be very difficult. The authors do write very clearly, but nevertheless, there is sufficient ambiguity in the description of the methods that it would likely not be possible for another scientist to easily reproduce the analyses that are described in the paper. A technical description of the methods using precise mathematical notation could, and should, be included in the supplement.

Detailed comments:

Supplement: The contents of the supplement belong to this paper, but the title does not!

40 – There is a bit of confusion here concerning how to refer to extreme value theory. The correct term is "extreme value theory", not "Generalize Extreme Value theory". The GEV (Generalized Extreme Value) distribution emerges from extreme value theory based on one approach to the analysis of extremes – the so-called block maximum approach. Typical practice is to use one-year blocks and to model resulting collection of annual maxima with the GEV distribution. Another approach that is also often used is called the "peaks over threshold" approach, in which exceedances above a high threshold are modelled. This approach leads to the use of the Generalized Pareto (GP) distribution. Both distributions are asymptotic – in the case of the GEV, it is the limiting distribution that is obtained as blocks grow in length without bound (provided there is a limiting distribution). Similarly, the Generalized Pareto distribution is obtained as the level of the threshold increases without bound (again, provided there is a limiting distribution). Thus, it needs to be understood that for a given block length or threshold, the GEV or GP distributions

respectively, can only approximate the distribution of the extremes that are identified. The Brown et al paper you cite uses the peaks-over-threshold approach.

44 – Delete “easily”. I think it is debatable whether a full distribution approach is best if the objective is to make inferences about extremes. If this were easy, it is likely that we would not have as much effort as is expended on developing and applying extreme value theory.

56 – “significant” in the statistical sense? Please clarify. To avoid confusion and ambiguity, my suggestion would be to avoid the word “significant” in scientific papers except when discussing statistical significance.

83-85 – To what extent are the moments (particularly the higher moments) sensitive to the degree of inhomogeneity that is permitted? Presumably, consideration of that sensitivity should drive the choice of permitted jump. Otherwise the choice of a number like 1°C, while tidy, would be entirely arbitrary. Note that jumps of this size do not have the same impact on higher-order statistics in all climates; the impact might not be easily discernable in a high variability mid- or high-latitude climate, but it might be very large, in relative terms, in a tropical climate with low temperature variability. This is also where the impact of poor measurement resolution would be the largest.

A second point here is that while temperature is mentioned, nothing is said about winds. Is a similar approach used?

Also, I assume that both dry air temperature and dewpoint temperature are treated in the same way. This raises a small question about whether the same limit for jumps should be used for both, given that the impact on higher order moments may be relatively larger in one case versus the other.

114 – Have you thought of using L-moments rather than ordinary moments? L-moments (linear moments) are generally considered to be somewhat more robust (in the statistical sense of the word), and might be less affected by inhomogeneities.

115 – It is unclear, distribution of what? Is it reasonable, for example, to consider all anomalies (relative to the annual cycle) for the entire year to be part of the same distribution? As an example, the processes that produce variability in summer in the northern mid-latitudes are substantially different from those that do so in winter, so it is not clear, really, what the distribution represents. One might also ask the same question about anomalies that are pooled across latitude bands to produce distributions (should one lump anomalies from different climate types together, and still call it a “distribution”?).

118 – A small quibble here; in statistics, a hat (circumflex) is often added to Greek letters denoting distribution parameters if those parameters have been estimated, as is the case here.

Figure 2 – A small editorial comment is that yellow curves often almost disappear into the background (many people find them hard to see). Another colour scale that doesn't fade away to light colours as you approach the present would be helpful.

161 – What aspect of wind speed does the paper consider? Are these hourly mean values?

163-164 – I think many would dispute that temperature is "normally" (Gaussian) distributed. Indeed, if temperature did have a normal distribution then the study of skewness and kurtosis would be entirely uninteresting, since the normal distribution is fully determined by its first two moments.

167 – Confidence bounds? I'm fine with simply describing +/- one standard deviation as an uncertainty range without a quantified level of confidence, but as soon as you call this a confidence bound, a question arises about the confidence level.

178 – How does seasonality play into these findings about changes in the shape of the temperature distribution? Is there a physical or sampling interpretation to the change in kurtosis that is noted? Could this be due to undetected data artefacts (such as changes in instrumentation over time).

Figure 3 (and others) – the labelling of the figures could be improved throughout the paper and the supplement. In this case, the key piece of information that tells the reader what is in each panel (mean, stdev, etc), is buried in in the middle of a small-font machine generated label.

Figure 3 (and others) – the decision to show only stations with +/- one standard deviation uncertainty ranges that exclude zero seems arbitrary to me. My general preference is not to censor results in this way, since changes in the underlying field are likely relatively smooth and continuous. Imposing a noisy mask (which corresponds to conducting a test with unknown properties) could be argued to be at least as deleterious to describing changes in the dataset as including all of the effects of internal variability and sampling noise at all locations. The locations that are retained are still affected by these uncertainties.

235 – See previous comment.

237-238 – Why is US station density apparently so low?

242 – Replace “Over 90%” with “Approximately 95%” since 2806/2956=0.949.

257 – “visible in more widely in” → “visible more widely in”

272 – I don’t see how greater station density would increase trends (do stations emit heat? ☺).

286-295 – Here and elsewhere, it would be really useful to have more depth in the analysis of differences between datasets (or analyses). There is merit in pointing out differences, of course, but it would be much more useful to those assessing datasets if the authors could delve into the causes of these differences.

297-303 – What fraction of stations are affected by expanding urban heat islands, which can induce apparent trends when formerly rural stations come to be affected by urban heat islands as nearby cities develop? This is a huge problem in China, for example. China has a very dense observing network, but it was not built for climate purposes, and only a small fraction of Chinese stations can be classified as being rural throughout their observing history (this is particularly the case in Eastern China). It has been estimated that this difficult-to-detect cause of inhomogeneity has resulted in recorded temperature in China warming substantially more than the country has actually warmed (despite intense development, urban areas still represent only a small fraction of the Chinese land area). See Sun et al., 2016, *Nature Climate Change*. Other parts of the developing world are, presumably, similarly affected.

328-330 – It would be really useful if the authors could dig more deeply and provide more than speculation on the causes of the differences that are noted.

337-338 – Is this decrease in relatively humidity confirmed in HadISDH?

392 – See a previous comment pointing out that changes in the tails are not necessarily easily inferred from changes in the full distribution.

395 – Replace “calculates” with “can be used to calculate”. Implicitly, the statement here describes quantile regression as calculating linear trends. The method, is in fact, a lot more flexible than that (other trend models can also be fitted, if appropriate).

408 – As mentioned previously, I think the question of measurement (or perhaps better, data recording) resolution should have been introduced much sooner. Also, the statement that temperatures are reported to the nearest 0.1°C or 1°C does not really convey the full complexity of the problem (e.g., associated with conversion of °F to °C, and the subsequent rounding or truncation to increments of 0.1°C or 1°C). See Rhines et al, 2015.

421 – How many stations in a polygon?

435 – Define N.

451 – This sentence seems to be grammatically challenged.

500-501 – Have there been changes in instrumental design or the processing of instrumental readings (e.g., approaches that might have been applied to compensate for variations in cup anemometer drag with velocity) that might have contributed to the apparent "stilling"?

519 – I'm not sure that spatial smoothing would result in "less extreme fields". The idea that the magnitude of extremes could be reduced is reasonable, but the idea that the result could be "less extreme" seems less well founded. "Extremeness", the relative position of an observation in the tail of its distribution, can only be evaluated within the context of the variability that a given data product tries to represent. Saying that the magnitudes of point rainfall observations can be larger than grid box mean rainfall amounts doesn't help one decide whether a given grid box mean value is more "extreme" (situated deeper in the tail) than a given point observation.