Clim. Past Discuss., 11, C2850–C2855, 2016
www.clim-past-discuss.net/11/C2850/2016/

Climate of the Past Discussions

Interactive comment on "Expanding HadISD: quality-controlled, sub-daily station data from 1931" *by* R. J. H. Dunn et al.

R. J. H. Dunn et al.

robert.dunn@metoffice.gov.uk

Received and published: 6 January 2016

We thank the reviewer for their comments which have helped improve the station selection procedure. Our responses to each comment are detailed below.

Reviewer

This paper describes an update of an already published dataset. The modifications do not justify a new paper, even though the science behind is good and solid. Perhaps the authors could try to publish in a journal specialized in climate datasets description.

Response

We have submitted this paper to Climate of the Past as this is where all the HadISD

papers have been published so far (Dunn et al, 2012, Dunn et al, 2014). We therefore thought it would be best to keep the update paper with the same journal. We understand that this is an update to a dataset, but there have been some modifications to the station selection and merging procedures, which are fundamental to the make-up of the dataset. We have also added extra quality control procedures and also adjusted the way the quality control checks have been applied. It is our belief that these changes are sufficient to merit an update in the same journal in which the initial release was described.

The editors have informed us that they believe this paper would better sit in another journal of the EGU family (Geoscientific Instrumentation, Methods and Data Systems). We will resubmit there, including the improvements suggested by both reviewers.

Some comments:

Reviewer

4571-6: use a more friendly name for public version control.

Response

We have used the HadISD.x.y.z version control since the dataset was created, and chose this from the similar versioning system used in HadCRUT. This allows us to clearly identify which dataset is being used, given the annual updates that are applied. In this manuscript we have used an "x" in the final digit to indicate all possible versions when indicating a feature that goes across all minor versions.

Reviewer

4572-1: users might have problems with this as they might rely on a fixed set of stations

Response

This is a good point, and until we have run a number of yearly updates we will not be sure how large the changes are to the input station lists. What we will do in light of this

comment is to release on the updates the list of new stations that are included, and a list of stations that are no longer included - so that any changes to the station list are clear. We decided on a dynamic station list so that over time the station number, and hence coverage of the dataset would improve. We have added a comment in light of this in Section 2.1.

Reviewer

4572-15: The selection proess may introduce stations with l5 years of data spread over the 1931- 2015 period. Would a station with data only in January between 1931 and 1990 make it to the dataset? Perhaps I am missundertanding the way you do the selection. It is not clear to me.

Response

The station selection would allow 15 years worth of months with the equivalent of 6 hourly observations to be spread over the entire 1931-2014 record - there has been no requirement on continuity - for example 5 years at the beginning and 10 years at the end of the record. This was explicitly stated in Section 2. Given the distribution of stations in each year, as shown in Figure 2, a significant number of stations will have large gaps in the data.

Your second point is interesting, and yes, the station selection would have allowed a station with only data in January (at least 120 observations) through. What we have done to address this is to test each month to ensure that the median observing interval is at least 6 hourly (so 120 observations in a month), rather than just the entire record.

Reviewer

4573.10: is there any criteria based on the data itself? For example . . . does the merge extend a station (good) rather than filling little gaps (bad); do you check if correlation over overlapping sections if they exist? Do you check if values of both stations are compatible?

Response

At the moment we do not use any criteria based on the data itself. This was done using the ISD-lite database when selecting stations for HadISD.1.0.0. The merge also does not check whether extra observations are extending a record or filling a gap.

In our current workflow,stations are merged together during the conversion from the ISD ASCII fixed-field format files to netCDF - and so there is no chance to check on correlation or compatible values during the line-by-line read. Also, given the size of the ISD archive, it would be prohibitive to download all the stations to check on the correlation and for compatible values.

We do appreciate that both these suggestions would be a useful additional checks to the process and is something we will look do to in future releases. It would require a two-stage process and a further adaptation of our selection process. For example, having identified the merging candidates, create individual netCDF files without merging, then do some compatibility tests, before merging to create the final netCDF files.

This is something we will bear in mind for a future update of HadISD - we would need to ensure that these work in an objective and automated fashion, and including these would delay an update to HadISD for longer than we think is appropriate.

Reviewer

4577-8 & 18: Gaussian vs gaussian. I would recommend not to capitalize the distribution name. Better write "gaussian distribution" than "gaussian"

Response

We now use "gaussian distribution" throughout.

Reviewer

4577 : distributional gap and streak test are not well explained. Streaks: does the sentence: 'To allow these thresholds to be calulated dynamically, the distribution of

repeated values is analysed.' Imply that a series with lots of repeated values will have a higher threshold? IF so, this is not convenient. If not, please, explain better.

Response

Our intention was that readers would refer to the original data paper for a more complete description of the tests. However we see that it would be useful to have a concise outline of the tests so readers can clearly see the improvements. We have added extra text to clarify how these tests work.

For the streak check this is not what is meant, and we have clarified the text to improve our explanation. By fitting the distribution of streak-lengths, we aim to ensure that naturally occurring repeated values are retained, but erroneous ones are not. And this cut off value is set by the properties of the station (see Figure 5).

Reviewer

4580: Neighbour checks: 'The closest 20 neighbours within the limits of 500m elevation and 300km distance are obtained for each station. For each of these neighbours, the data overlap with the target is calculated. Also, the correlation between the neighbour and target is obtained after removing the annual and diurnal cycles.' : even understanding the need to find neighbours, I wonder if 500 km is not a very large distance, specially at the resolution of this dataset. I am wondering if by removing daily and annual cycles you are not more likely to correlate stations with very different climates.

Response

Our use of 300km distance and 500m elevation matches the neighbour selection criteria for HadISD.1.0.x. In that dataset there were 704/6103 stations for which no neighbour check was possible. For HadISD.2.0.0 there are 383 stations out of 8113 for which fewer than three neighbours were found. Unfortunately this does not guarantee that the neighbours that were found had contemporaneous observations such that each observation could be checked.

Our intention of removing the daily and annual cycles was to ensure that the anomalies correlated. This dataset has a use for studying extreme events, and so keeping anomalously high or low (but valid) values is important. By correlating the anomalies it means that we will allow stations with very different climates - which also increases the number of neighbours that are available. As the test works using the distribution of the difference between the station-neighbour pairs, then our thought was that this was a valid approach.

Interactive comment on Clim. Past Discuss., 11, 4569, 2015.