Clim. Past Discuss., 8, C1659–C1665, 2012
www.clim-past-discuss.net/8/C1659/2012/

Climate of the Past Discussions

Interactive comment on "HadISD: a quality controlled global synoptic report database for selected variables at long-term stations from 1973–2010" *by* R. J. H. Dunn et al.

R. J. H. Dunn et al.

robert.dunn@metoffice.gov.uk

Received and published: 24 September 2012

We thank Neal Lott for his positive review, and address each point he raised in turn below.

Referee:

In the introduction, reference is made to the ISD database being "non-trivial for the non-expert" to access. It should be noted that NCDC provides access to ISD via a GIS interface and with the ability for users to select desired parameters into a space or comma-delimited file. Providing this information would be useful for readers to know.

Response:

We have added a sentence indicating the availability of an alternative access to the ISD database.

Referee:

The paper references ISD-Lite. Was this input source used in the effort? If so, there may be some questions regarding the ISD vs ISD-Lite that should be addressed. If not, then it's a non-issue.

Response:

The ISD-Lite was only used in the compositing section to check whether two station records are likely to come from the same physical station. Only the 00UTC temperature anomalies, the data count per month and the daily distribution of observing times were used. This choice was one of expediency. We state explicitly the merge is 'research quality' with plenty of room for improvement in future versions. The ISD-Lite was not used in any other stage of the analysis or as input into the QC suite.

Referee:

In section 3, para 1, the paper states that a large number of stations report only rarely. This should be quantified in some way. Readers may infer that well over half of the stations have very little data.

Response:

We have done a quick analysis of the current state of the ISD database (on 31st July 2012). There are 29368 unique station IDs in the database. Of these almost half report in fewer than 10 years (14062), 18726 for less than 20, and 21825 for less than 30. The largest number of years in which a single station reports is 80. The total size of the data file for a station is a rough indicator of the number of years with data. Most stations had total file sizes between 10^5 and 10^7 bytes with a long tail down to 100 bytes, and

CPD

8, C1659–C1665, 2012

a sharp drop off at higher amounts.

We plot the number of years a station reports in against the amount of data in the final file in the attached figure. A large fraction of stations report for few years (<20) and also have small total record sizes (in bytes).

Unfortunately we do not have the state of the ISD database when the download occurred to create the HadISD dataset, however this brief analysis gives a flavour of the problem we were trying to highlight. We shall amend the text in the following way at the beginning of Section 3:

"The ISD consists of a large number of stations, some of which have reported only rarely. Of the ∼30,000 stations, about 2/3 have observations in 30 years or fewer and several thousand have small total file sizes, corresponding to few observations. However, almost 2000 stations have long records extending 60 or more years between 1901 and end-2011. Most of these have large total file sizes indicating quasi-continuous records, rather than only a few observations per year. To simplify selection..... "

Referee:

In section 3, para 2, the paper refers to the first 250 lines of each annual file. It might be useful to refer to the first 250 observations or the first 250 data records, so it's obvious to the reader what this means.

Response:

We have changed the wording to read "..the first 250 observations of each.."

Referee:

Regarding section 4, were the authors aware of NCDC's online documentation regarding known problems in ISD? It would be interesting to know if the QC checks performed by the authors "caught" some of these problems.

Response:

CPD

8, C1659–C1665, 2012

We were aware of the online documentation regarding the problems already found in the ISD. However, we thank Neal for reminding us to go back and check through this list. Of the 27 known problems listed in the known problem file (from 08-Jun-2011) many relate to variables, stations or time periods which are not in the HadISD selection. However there are six problems which are relevant (numbers 6, 7, 8, 22, 24 and 25). Our QC suite is not able to test for the problem outlined in number 24 (station 725020-14734) because, although it looks for changes in reporting accuracy, these are not used to flag any periods. The compositing of stations outlined in number 22 was done successfully, the final station reports under 725765-24061 containing both 725765-99999 and 726720-99999. The data issues described by the remaining four tests have been successfully identified by the QC suite (stations 718790, 722053, 722051 and 722010 – we have merged 722053 and 722051). We have added a paragraph in Section 4.3 outlining these checks.

Referee:

Regarding section 4.1.11, providing an example for a station here might be helpful. Also, several of the figures refer to the affect that station compositing may have on the QC results (eg, increasing the flagging rate) – very good to see that this was evaluated. For the QC checks in 4.1.11, could compositing have resulted in some artefacts in the data which would have affected the QC results? (NCDC recently ran a check which showed no occurrences of dew point higher than temperature in the ISD database – ie, for specific stations in each annual file.)

Response:

In the knowledge that no supersaturation events were found in the ISD, we have gone back to our data to check what could have occurred to result in super saturation flags. In the initial conversion from the ISD ascii files to the netcdf HadISD files there was some merging of observations across each hour period (i.e., an 11am observation could come from individual observations between 10.31am and 11.30 am), favouring

individual elements which were closest to the full hour. This approach had been chosen to maximise the data within HadISD. However, this led to some observations in HadISD having temperature and dewpoint temperature coming from slightly different timestamps. We realise now that this is not desirable, especially if using this dataset to study humidity.

Therefore, we have amended our routines to require that all elements in a single timestep in HadISD come from a single timestep in ISD. However, rather than only favouring the timestep closest to the full hour, we favour the one which has both temperature and dewpoint temperature observations, or if no dewpoint is reported, then temperature alone is favoured. If neither is reported, then the timestep closest to the full hour is favoured. Therefore for a single timestamp in HadISD, all non-missing elements come from the same timestamp in ISD, though this may not be the one closest to the full hour.

This has required that we re-run the conversion to netcdf and also the QC suite (which was planned anyway as a result of the first referee's review). We have taken the opportunity to update the dataset at the same time, so that the end date is now 00h00 UT on 1st January 2012, adding 1.5 years of observations into the initial release of HadISD. The selection of the climate quality stations (Section 6) has also been redone.

We have updated all the relevant sections in the paper, as well as all the figures and tables to reflect these changes. The paper title has also changed as a result of the extra data.

The final dataset now no longer has any instances of super saturation. The validation results are still accurate, with no loss of data or extra removals of extremes as a result of these changes.

Referee:

Was consideration given to process the QC checks both before and after station com-

CPD

8, C1659–C1665, 2012

positing, at least for certain algorithms?

Response:

Unfortunately, the QC suite was not run before and after compositing on those relevant stations. In theory we could reverse engineer to do this and undo merges conditional on the merge decision to see what difference it makes. We would prefer to undertake such analyses in a formalised framework as part of development of a next version as it is significant work and also would extend an already very long paper.

Interactive comment on Clim. Past Discuss., 8, 1763, 2012.

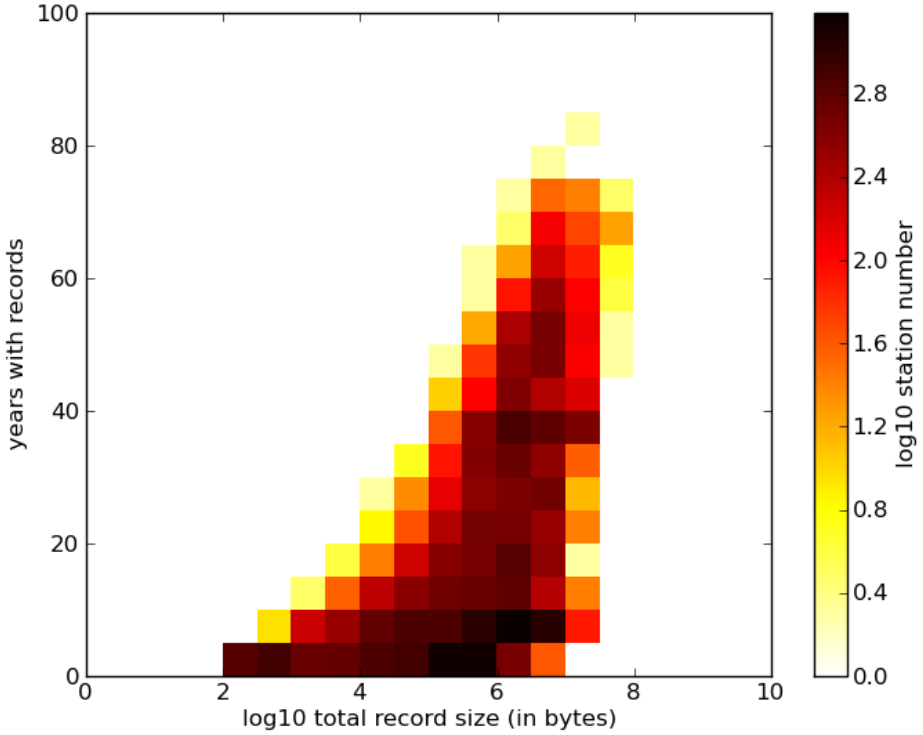

Fig. 1. Years with records against record file size for the full ISD database.

CPD
8, C1659–C1665, 2012

Interactive Comment