PROJECT DELIVERABLE

Work package 3:	Earth system observations
Deliverable D3.20:	HadISD update
Type:	Report
Author:	Robert Dunn (METO)
Reviewers:	Dick Dee (ECMWF), Hans Hersbach (ECMWF)
Delivered:	11 December 2014
Notes:	

Grant agreement no. 607029

Expanding HadISD: quality-controlled, sub-daily data from 1931

Deliverable D3.20 (December 2014)}

Robert Dunn

Met Office Hadley Centre

December 2014

Abstract

We outline the pre-release version of HadISD (2.0.0.2014) for use in ERA-CLIM2. The updated station selection and quality control codes are introduced. This version will be updated in early 2015 to include data up to the end of 2014, along with further improvements to the quality control tests.

1. Introduction

In this deliverable we outline the update to HadISD in which we extend the temporal coverage back to 1931 and also improve the station selection process. The overall procedure is very similar to the creation of HadISDv1.0.0 as outlined in Dunn et al (2012). This new dataset, a preliminary version of HadISDv2.0.0, is still a quality-controlled subset of the ~28k stations held in the Integrated Surface Database at the National Oceanic and Atmospheric Administration's National Climatic Data Center (NOAA/NCDC). This summary accompanies the pre-release version, v2.0.0.2014, for use by ERA-CLIM2 partners in testing only.

In Section 2 we outline the updated selection and merging procedure, which will also be run on each future annual update. Changes to the quality control tests are outlined in Section 3. The data provision is discussed in Section 4, with a summary in Section 5.

2. Station Selection and Merging

For HadISDv1.0 the stations included in the dataset were fixed at the first release, and no updates were made to this station list. Therefore subsequent annual updates to HadISDv1.0 could not benefit from developments in the ISD made at NOAA/NCDC. In HadISDv2.0.0 the station selection process becomes part of the general update. This means that each year the stations selected from the ISD may be different from the previous version, as different stations satisfy the selection criteria. The methodology of this updated station selection procedure is shown in Fig. 1.

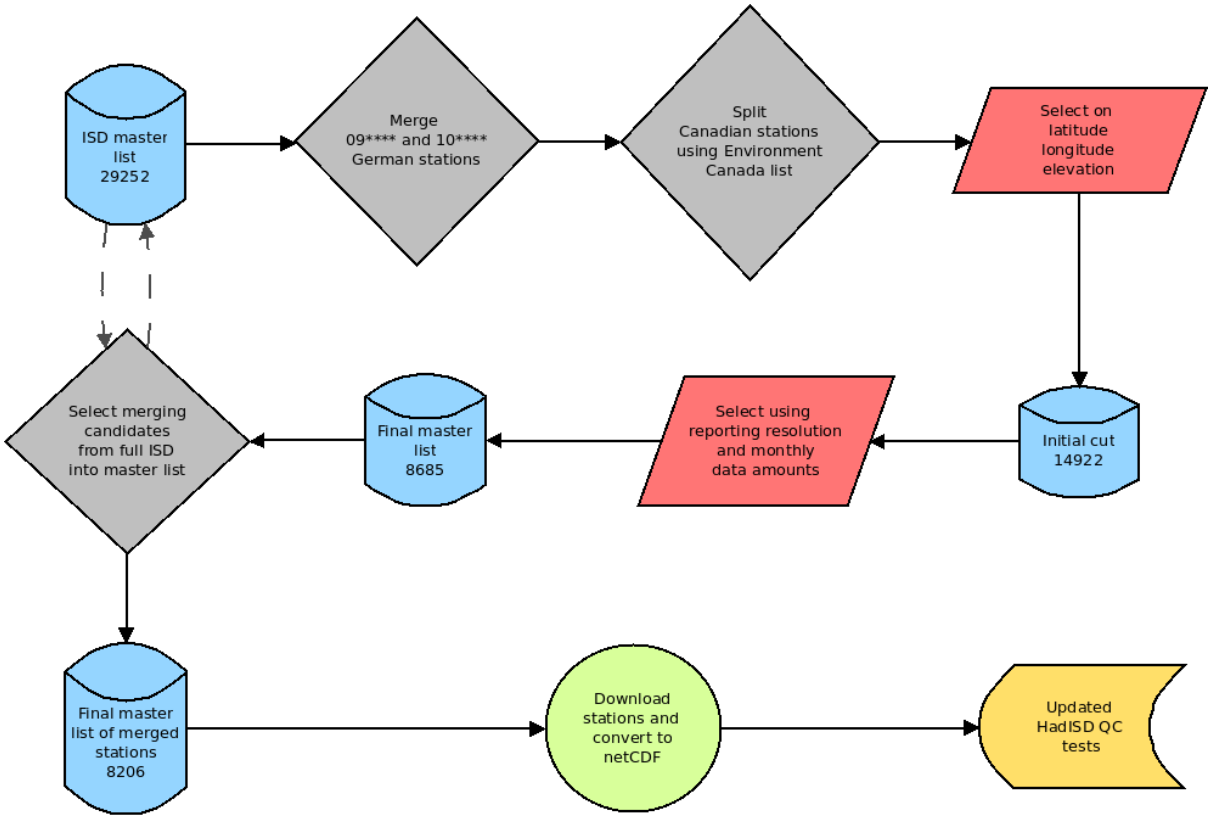

Figure 1: the process used for station selection and merging for HadISD.2.0.0

Using the inventory files on the ISD FTP server, stations are selected on the basis of a number of requirements. Firstly, a station has to have a known latitude, longitude and elevation, and cover a time span of at least 15 years between the first and last observation. This initial cut of 14922 stations are investigated further using the detailed inventory file to ensure that there is an equivalent amount of data present of 15 years of observations every six hours. This results in 8685 stations being taken forward for further processing.

2.1 Merging stations

There are likely to be stations within these 8685 which are non-unique, and so could be merged together. Also, there will be stations in the full ISD catalogue which could supplement the data within these 8685 candidates and so improve the temporal coverage.

To avoid merging stations which are not suitable, we need a simple, yet robust method of selecting stations to merge. We follow a method which is similar to the International Surface Temperature Initiative (ISTI, Rennie et al, 2014). The ISTI methodology maps separations (distance and height) into decaying exponential probability curves. These probabilities are combined and a threshold set below which stations are not merged.

Our selection of merging candidates is based only on the latitude, longitude, elevation and station name. Using the latitude and longitude the Euclidean distance between the two stations is calculated, and a probability returned using an exponential decay with an *e*-folding distance of 25km. For the elevation separation, the probability is obtained using an *e*-folding distance of 100m. The station names are compared using the Jaccard Index (Jaccard 1901) as in the ISTI merging algorithm. If the product of these three probabilities is greater than 0.5, then the stations are deemed similar enough to merge. Merging stations

within the list of candidate stations results in a final list of 8206 stations, of which 2101 contain data from other station IDs. The increase in the data coverage by including stations from the full ISD holdings can be seen in Fig. 2.

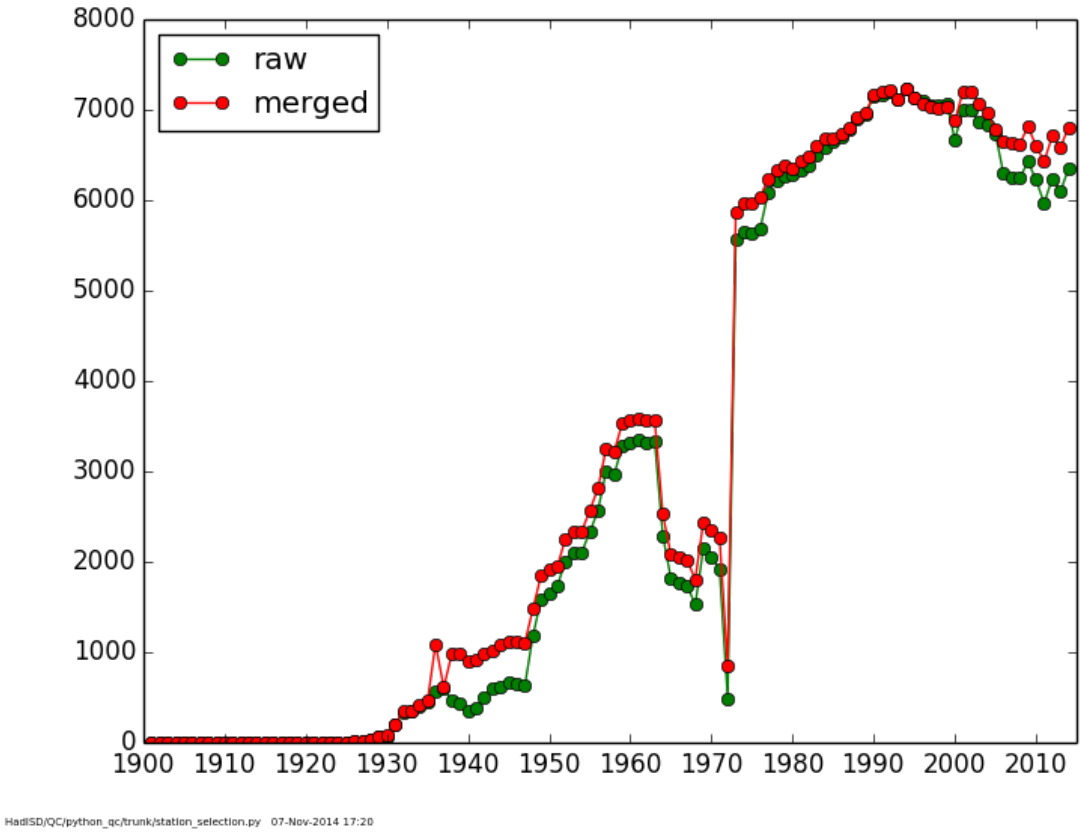

Figure 2: The distribution of stations with time before (green) and after (red) merging.

Fig. 2 also shows why we have not included stations prior to 1931 in this extension of HadISD. However, by checking in the full ISD catalogue for stations to merge with, the coverage has been significantly improved prior to 1950, as well as smaller improvements at other times. The final locations of the stations are shown in Fig. 3, along with the location of the merged stations.

2.2 Extra processing for specific countries

Since the release of HadISDv1.0.0 a number of specific problems with the data held from certain countries have come to light in ISD. In this pre-release version, stations with WMO ID numbers starting 09 and 10 have been specifically targeted for merging. Information from DWD (Andreas Becker, private communication) allowed us to use the last 4 digits of the WMO ID of stations beginning 10 to test against the last 4 digits of those starting 09 instead of using the merging algorithm outlined above. This process is done before the station selection on location and length of record, as shown in Fig. 1.

Issues with Canadian stations, which required splitting because station identifiers were re-used by different stations, are still being resolved, but the resulting files will be included in the final release version of v2.0.0.2014.

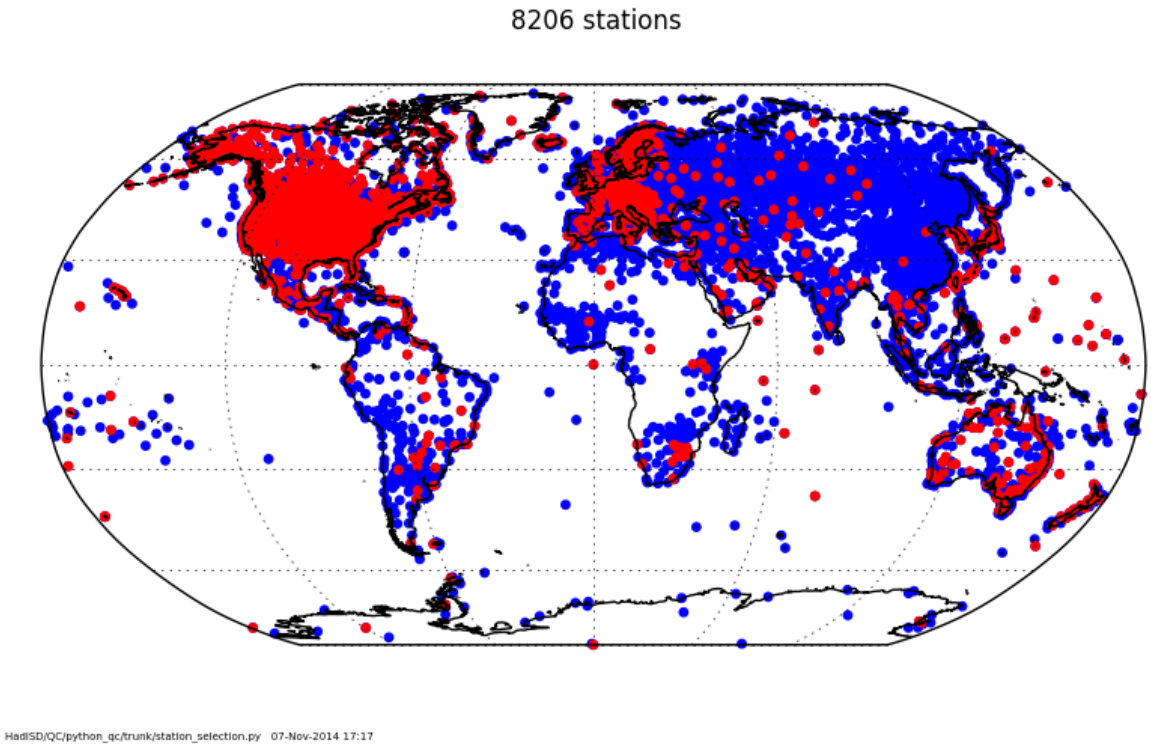

Figure 3: The location of the final set of stations (blue) with the merged stations highlighted in red.

3. Updating the Quality Control Tests

As part of this update we took the opportunity to re-write the quality control software in Python instead of IDL, as Python is becoming more commonly used and is Open Source. All the code used to create HadISDv2.0.0 is written in Python, and will be made available alongside the dataset.

We attempted to match the performance and outputs of the tests between the two languages. In some cases we were able to correct bugs present in the IDL, and some tests could be written to result in bit-wise reproducibility. However for some tests, this was not possible.

Using the 167 stations in the UK and Eire in HadISDv1.0, we compared the number of observations identified by each test between the Python (HadISDv2.0.0) and IDL (HadISDv1.0.0) versions.

Test Name	Changes
Frequent Values	Bug fixed in DJF section of seasonal check
Distributional Gap	Threshold changes resulting from minor differences in Gaussian fitting routine
Known Record	Values updated to account for recent world record changes
Repeated Streaks	Bug in annual string expectance fixed – first year of record now also checked
Climatological Outlier	Threshold changes resulting from minor differences in Gaussian fitting routine
Spike	Changes from the way that missing/flagged observations are handled. Bug resulting from single/double precision comparison fixed
Wet-bulb cut-offs	Improvement in reporting frequency calculation
Variance	Bug fixed as was only working on filtered (cleaned) values
Unflagging (neighbour)	Not run for spike check as retained obvious spikes. Improved for odd cluster

Table 1: Changes to the QC tests on translation from IDL to Python

The change in the unflagging section of the neighbour check results in more odd clusters being retained as they can be compared to observations at neighbouring stations and the changes to the spike check have resulted in fewer spikes being identified.

Also, for this preliminary version of HadISDv2.0.0 we do not re-run the spike and odd-cluster checks a second time around (as was done in HadISDv1.0 – see Dunn et al, 2012, Fig. 3). This may be addressed in the future as for some stations this results in a large reduction in the number of observations identified by the spike check. However, for most stations there is no change as a result of this part of the updated methodology.

4. Data Provision

HadISDv2.0.0 is provided as Network Common Data Format version 4 files (NetCDF4) at www.metoffice.gov.uk/hadobs/hadisd/v200_2014/download.html. We have moved from NetCDF3 files as used in HadISDv1.0 to NetCDF4 as these have internal compression, and so result in smaller file sizes on disc, which will hopefully make them easier to process.

The versioning scheme will be the same as for HadISDv1.0, with annual updates occurring at the beginning of each calendar year. To maximise the inclusion of data from the latest year in the updates, these are carried out in a two stage process. A preliminary dataset will be released early in the year (e.g. v2.0.1.2015p in January 2016) with a final version a few months later to ensure that late-arriving data are included.

5. Summary

We have outlined the pre-release version of HadISD.2.0.0.2014 for use by the ERA-CLIM2 project. The span of the dataset has been extended to January 1st 1931 and currently runs until 31st December 2013. An update in early 2015 will include data from 2014. There are now 8206 stations in HadISD of which 2101 are composites. Improvements have been made to the QC suite, along with other changes, on conversion from IDL to Python. A full assessment of the differences and further improvements will be summarised in a forthcoming paper.

Acknowledgements

This project has received funding from the European Union's Framework Programme under grant agreement number 607029.

References

Dunn, R., Willett, K., Thorne, P., Woolley, E., Durre, I., Dai, A., Parker, D., and Vose, R.: HadISD: a quality-controlled global synoptic report database for selected variables at long-term stations from 1973–2011, Climate of the Past, 8, 1649–1679, 2012.

Jaccard, P.: Etude comparative de la distribution florale dans une portion des Alpes et du Jura, vol. 37, Impr. Corbaz, 1901.

Rennie, J., Lawrimore, J., Gleason, B., Thorne, P., Morice, C., Menne, M., Williams, C., Almeida, W. G., Christy, J., Flannery, M., et al.: The international surface temperature initiative global land surface databank: monthly temperature data release description and methods, Geoscience Data Journal, 2014.