Met Office
Hadley Centre

HadISD version 3: monthly updates

January 24, 2019

R.J.H.Dunn

www.metoffice.gov.uk

Abstract

The HadISD is a sub-daily, station-based, quality-controlled dataset with an integrated set of variables. This dataset is currently updated on an annual basis. We outline the changes made to the HadISD codebase to allow for monthly updates to this dataset. For some of the quality control tests, these changes ensure that appending successive months over the course of a calendar year do not affect the properties of the distributions used to calculate threshold values. A more major update early in each calendar year will allow the inclusion of data changes in the deep past as well as reselecting the stations that make up this dataset. We have taken this opportunity to handle better the precipitation data across the range of accumulation periods available in the ISD, and also include station level pressure. Both these quantities are now subject to simple quality control tests.

1 Introduction and Background

The Hadley Integrated Surface Dataset (HadISD Dunn et al., 2012, 2016) is a sub-daily, station-based, quality-controlled dataset with an integrated set of variables. The initial design was as such to facilitate the study of observed extremes, in particular heat health, and so the focus was initially on temperature and humidity variables.

This dataset is built from the data available in the Integrated Surface Dataset (Lott, 2004; Smith et al., 2011), a large collection of sub-daily holdings archived at the National Oceanographic and Atmospheric Administation's (NOAA) National Centers for Environmental Information (NCEI). The ISD itself has been built from a number of sources including collections from the United States Air Force (USAF) and other archives from within NOAA. At present ISD contains over 35,000 stations, starting on January 1, 1901 at 00:00 and is updated daily. A sub-set of stations are selected from this parent dataset that have sufficiently long, truly sub-daily records to create the HadISD.

The initial release (version 1.0.0.2011f, Dunn et al., 2012) contained 6103 stations spread around the globe, with denser coverage in the northern hemisphere, and unfortunately sparse coverage in e.g. South America and Africa. The HadISD data then started in 1973 because of the large increase in the number of stations reporting in the ISD during that year (see e.g. Fig. 2 in Dunn et al., 2016). The station list was static, but the dataset was updated twice a year. The first update occurred each year in January to release a preliminary (p) version allowing users to promptly investigate the weather and climate events of the previous calendar year. Then, to take advantage of the late arrival of some data into the NOAA archives, a further update was carried out each year, usually in April, releasing the final (f) version. During both of these updates, changes in observations throughout all the station records were included (the "deep past"), not just an append of the most recent calendar year. Each version is numbered using an X.Y.Z.YEARi scheme (for example, v2.0.2.2017f). As noted in Dunn et al. (2012), the "X" indicates a major version change, usually accompanied by a peer reviewed paper or technical note. A change in "Y" would point to a more moderate change, for example if the ISD were to update some of their processing or a change to one of the quality control tests. At the moment, the "Z" changes with each annual update, to indicate changes in the input data, both from the append and also in the

"deep past". The year indicates the final complete year of the dataset, so, in the example above, data are available until 31st December 2017 at 2300hrs. The last digit has so far been used to indicate the preliminary or final version as outlined above.

Although quality controlling the data is a major aspect of the dataset creation process, these tests and procedures do not consistently identify systematic changes in station observations resulting from non-climatic changes in the station instrumentation or metadata. In some cases these metadata are available, but as this is a global collection of data mostly these are not. Therefore we investigated whether the dataset could be homogenised, a technique which would correct for these changes. In Dunn et al. (2014) we used the Pairwise Homogenisation Algorithm (PHA) from Menne and Williams Jr (2009) to identify stations and times where these non-climatic change points are likely to have occurred. As currently it is not possible to adjust sub-daily data in a reliable and robust way, we only release the timestamps and the magnitude of the changes to users, so that they can select stations appropriate for their use. The files containing thsis information are available on the HadISD website as a separate download.

The station selection for HadISD.1.0.x was static, so although annual updates appended data to the station timeseries, no new stations were added even if the additional data meant that they would have passed the original selection criteria. The creation of HadISD2 (Dunn et al., 2016) addressed a number of sub-optimal aspects of the HadISD.1.0.x creation scripts. The quality control code was re-written in Python 2.7 from IDL and the start date extended back to 1931 (see their Fig. 2). The station selection procedure was automated (in HadISD.1.0.x this had a large manual component, which was one of the reasons why it was not revisited) with the intention to re-select the stations on each annual update. Hence, although HadISD.2.0.0.2015f had 7677 stations, the most recent version (2.0.2.2017f) has 8103. Furthermore, more detailed quality control tests were added for wind speed and direction, as well as minor changes to other tests and the overall data flow. Finally, a new addition to the HadISD2 was to automatically produce files with humidity and heat stress measures as part of the build process, and these have already been used in climate monitoring.

This document describes the changes made to the HadISD build process to allow monthly updates. This update to HadISD has also allowed the improved handling of precipitation information available in the ISD as well as the inclusion of station level pressure, an important quantity for the reanalysis community.

2 Monthly updates and versioning

An annually updating dataset (as opposed to a static dataset) is already a very useful tool for sectors and users wanting to monitor changes in the weather and climate, at particular locations. However, for the study of short-lived extreme events (e.g. heat waves or mid-latitude storms), having to wait up to 12 months for an update to make the observations available is not ideal. Therefore, to make HadISD more useful and usable the processing flow and tests have been updated to allow for monthly appending of

updated data.

The ISD itself is updated daily, and although most changes occur to the data from the current year, NOAA-NCEI do reprocess complete years of data when needed or make other changes to data in the "deep past". The ISD data files are arranged in one file per station per calendar year, and separated in folders, one for each calendar year. This has made it simple to perform annual updates as well as check which other files need to be updated. Although it is desirable to include changes across the entire station record for the annual updates, this is not always the case for a monthly update to maintain some stability for users.

The approach here is to provide a single, complete update only on an annual basis, with monthly appends in between. The update run in December 2018 would include data up to the end of November 2018 and be released as 3.0.0.201811p. To round off the data for the calendar year 2018, there will be a release in January 2019 which will be the final one appending 2018 data, and includes updates to the end of December (3.0.0.2018f). This update would be the equivalent of 3.0.0.201812p, but as it is the final monthly update of 2018 data, we use the slightly different label. At this point the homogeneity assessment will be run using the PHA and the outputs made available on the HadISD download pages along with the data files.

Then, in early February 2019, as the first update to include 2019 data, the ISD inventory will be reassessed, and a new station listing for HadISD will be created. Also, as well as data from the previous calendar year (2018), updates that have been applied to years in the deep past (1931-2017) will also be incorporated and the updated files downloaded for processing. Running this update in February gives the opportunity for as much data from previous calendar years (1931-2018 inclusive) to have been processed into the ISD. However, this will be the final time that 2018 data will be updated (until February 2020) as the subsequent monthly updates only append the 2019 data. As this is the first monthly update to include 2019 data and also having a new station listing, this will be released as 3.0.1.201901p. Then in March, there will be a normal monthly append, to release 3.0.1.201902p, where only the 2019 data files are updated, and so on.

In January 2020, the version 3.0.1.2019f would be released, with 3.0.2.202001p available a month later in February 2020. So, during each monthly update, all the months for that current (incomplete) calendar year are reprocessed. Therefore, there could be some change in the values of the observations in the NetCDF file from the current calendar year due to data improvements and insertions in prior months, but the data for previous years will stay the same. This means that the data from the in-progress year may not be stable between each monthly release.

2.1 Changes to Quality Control tests

A number of the quality control (QC) tests use the entire station record to determine threshold values from the distribution of observations available for that station. This allows the tests to be adapted to the climatological conditions measured at the station location rather than using fixed values valid either globally or for defined regions. However, by appending data during a monthly update, these thresholds

could change, resulting in observations which are removed by a QC test one month, but retained in the following. As this is undesirable, a number of the QC tests have been adapted as follows to mitigate this effect.

Frequent Values This test uses the distribution of observed values to identify those which occur more frequently than expected. Only data from the period up to and including the last complete year are used to identify values which appear more frequently than expected for each three month season. When checking these pre-identified values on an annual basis, then all data are used.

Diurnal Cycle By calculating the offset from a sine-curve, the phase of the diurnal cycle is estimated, and periods which differ significantly are flagged. The temporal offset of the diurnal cycle is identified using data from the period up to and and including the last complete year. Thereafter, all data are processed to identify where the diurnal cycle is offset from the identified location.

Distributional Gap This test looks for gaps in the distribution of observations, either on monthly averages or for all observations in each calendar month. For the monthly test, the monthly climatologies are calculated from data up to the end of the last complete year. However all months are assessed for asymmetries in the distribution. In the test assessing all observations, again, distributions for each month are calculated from the data up to the end of the last complete year, but all data are assessed for gaps.

Repeating Streaks Strings of repeating observations are identified, and if long enough, flagged. The thresholds are now calculated from data up to the end of the last complete year only. However, all data are assessed with these thresholds.

Climatological Outlier Outliers from calculated climatological values are identified and if sufficiently different, flagged. Now, climatologies are calculated from data up to the end of the last complete year. But anomalies are created and assessed for all months.

Spike Short-duration spikes in the data are identified using the spread in the first-differences. The thresholds are now calculated from data up to the end of the last complete year only. However, all data are assessed with these thresholds.

Excess Variance The climatological variance of the observations is calculated for each calendar month. These climatologies are now calculated from data up to the end of the last complete year. But normalised variances are assessed for all months.

Winds Using the wind speed and direction, a climatological wind rose is created, and one for each calendar year is compared with it. The wind rose is created from observations up to the end of the last complete year rather than the entire record. Each year is then compared with this wind rose using the root-mean square difference. Years are only removed if they fall beyond the threshold as outlined in Dunn et al. (2016) and if they have at least 100 observations. If there is a large annual

cycle in the average wind direction, then in early months of a year the resulting wind rose may not match well with that derived from all previous years, and so more flags may be set.

3 Changes to variables and additional QC tests

As the change to monthly updates caused a relatively significant change to the processing code, the opportunity was taken to add in two new variables in which users had expressed an interest.

3.1 Precipitation

As the ISD itself comprises of a number of different archives, precipitation information can be found in up to four separate entries[1]. Each of these has (accumulation) period, depth, condition and quality values. To assist in the quality control of the dewpoint temperature observations, the first of these four precipitation fields was extracted when building HadISD v1.0.x. For completeness and transparency the accumulation period and amount from this field was included in the final netCDF files as `precip1_period` and `precip1_depth`.

Despite these data not being subjected to any quality control, a number of users have taken these precipitation data for their assessments. Also, it has become clear that the naming of the NetCDF fields could be confusing for users, as it arises from a convention in the parent ISD dataset rather than anything to do with the time resolution of the HadISD[2]. Therefore the precipitation accumulation included in the HadISD files could be from a range of accumulation periods in the record as different report types were combined in the ISD (accessible from the `precip1_depth` field). Finally, as there are four precipitation fields, it could be that there are reports in fields two to four of the ISD even if that in the first is empty, flagged or has zero accumulation, and hence users may not have had a true picture of the precipitation amounts.

Now all four precipitation fields are extracted from the ISD to create new NetCDF variables for all of the possible accumulation values that are present in the ISD. As well as the standard accumulation periods (24, 12, 6, 3 and hourly), there are also some 2-hourly, 9-hourly, 15-hourly and 18-hourly values present in the ISD files, so these have also been retained.

To do a complete quality control check on these values at the same level as the temperature or pressure observations, for example, is a large task. The INTENSE project[3] are working on sub-daily precipitation, with a focus on hourly accumulations. As part of creating their database (which includes all precipitation information from the ISD) they are developing quality control software. Therefore users who wish to do more detailed quality control are encouraged to contact the INTENSE team directly.

[1] The ISD files use string fields to hold the data, with the first three characters forming an identifying code (`AA1` to `AA4` for precipitation). Therefore, documentation relating to the ISD and HadISD has at times used the term “precipitation fields” to refer to these entries.

[2] For a more detailed description of this issue, see http://hadisd.blogspot.com/2018/03/precipitation-in-hadisd.html

[3] https://research.ncl.ac.uk/intense/

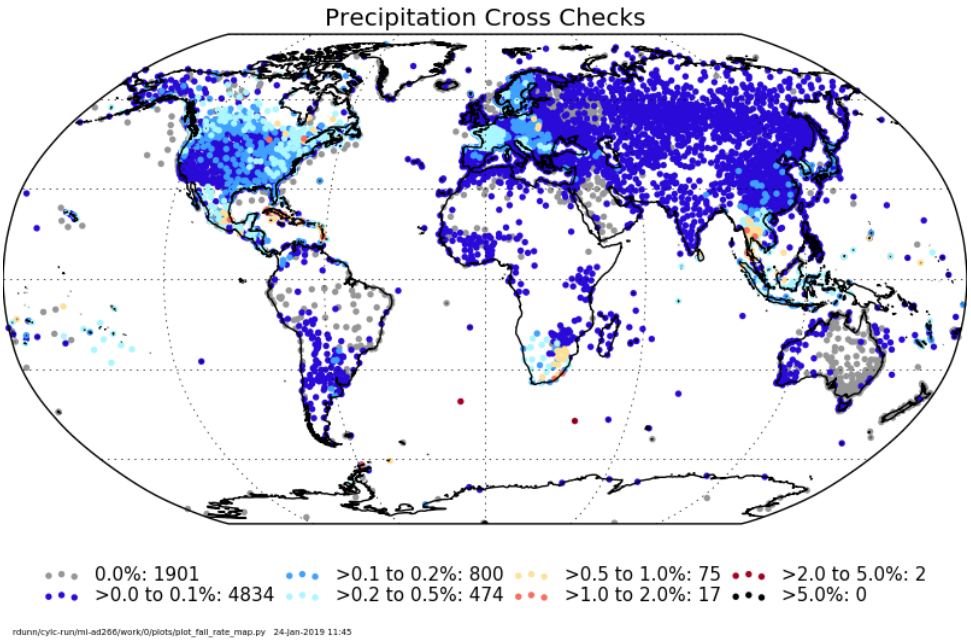

Figure 1: The fraction of timestamps with precipitation accumulation values flagged by the precipitation consistency check in HadISD v3.0.0.2018f.

However, one consistency check is performed in HadISD, which tests to ensure that shorter accumulation periods have smaller accumulation depths at a single timestamp. This presumes that the precipitation accumulations reported on a given timestamp were all taken at that timestamp, rather than being a delayed report for an earlier period. All accumulation periods which have non-missing values are tested, to ensure that ordering of the accumulation depths is the same as the accumulation periods (ascending order). All values are flagged if any one accumulation depth is larger than one from a longer accumulation period (see Fig. 1 for the distribution of the flags).

3.2 Station Level Pressure

A user had asked for station level pressure (separate to the sea-level pressure) for the stations where this was reported. In fact, of course all stations measure station level pressure, but this is then converted to a hypothetical value at sea level for that location during initial data processing. In the ISD, the sea-level pressure field is already populated, is more complete than the station-level pressure field, and is used directly when creating HadISD.

The station level pressure is now extracted from the ISD data files and included in the final netCDF files in the `stnlp` field. An additional test is now performed to ensure that these observations have reasonable quality. Using the timestamps where both the sea and station level pressure are measured, difference between these two quantities is calculated. This difference should be relatively stable over the entire station record, and any outliers are flagged. Locations where the difference is more than 4.5

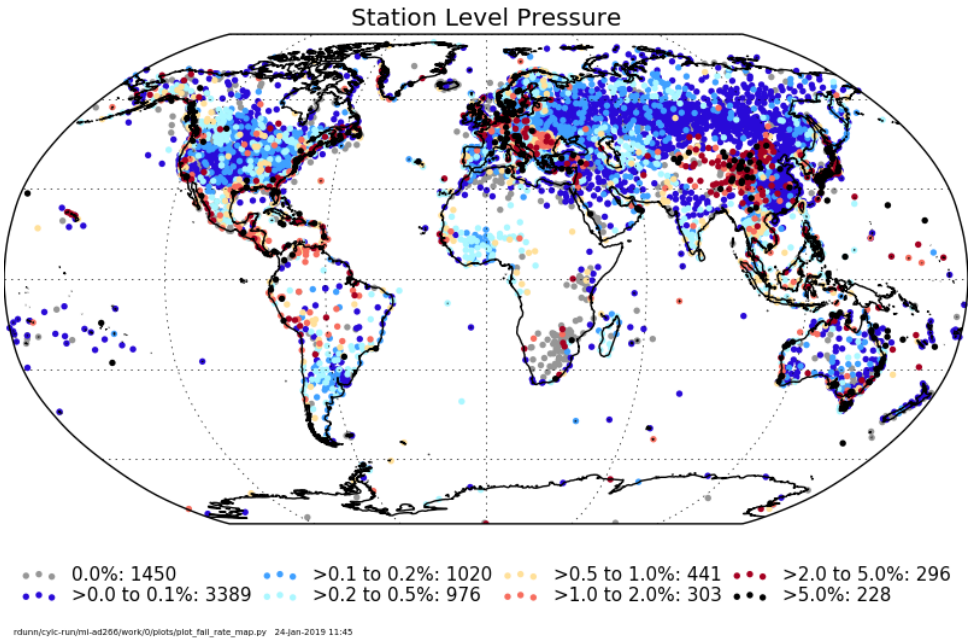

Figure 2: The fraction of timestamps with station level pressure values flagged by the pressure consistency check in HadISD v3.0.0.2018f.

median absolute deviations away from the median difference are flagged and removed.

As this check uses the median of the pressure differences, to ensure that this is stable during monthly updates, the median is calculated from all data up to the end of the last complete calendar year (as for other tests, see Section 2.1).

Any flags set on the sea-level pressure from the following tests are applied to the station-level pressure: odd cluster[4], frequent values, distributional gap, world record[5], repeated streaks, spike and excess variance (see Section 2.1).

As can be seen in Fig. 2, although only few station have no flags set from the pressure consistency check, most have a very low flagging rate. A cluster of stations with greater flagging rates can be found in the upland areas near to the Himalaya and regions to the north. A manual check revealed that in many stations, flags usually occurred early in the station record, for isolated observations. However it is clear that there may be a more persistent issue for some of the stations where the flagging rate is higher.

4 Summary

The quality control routines which assess and flag the ISD observations to form HadISD have been adapted to allow for monthly updates. Distributions and threshold values are set from completed calen-

[4]This test identifies short, isolated clusters of observations which cannot easily be compared with those immediately adjacent in the timeseries.

[5]The WMO recognised world records of meteorological variables are used to identify spurious observations.

dar years of data only, and so the appending of months does not affect these. Updated data from all months of the in-progress calendar year will be downloaded each month, and so there is the possibility of data in the early part of the year to change from month-to-month.

The way that precipitation data in HadISD are handled and provided to users has been improved. The different accumulation periods from the four precipitation fields in the ISD are dis-aggregated and the precipitation depth from each accumulation period is provided separately. Logical consistency between the set of accumulation periods available at each timestamp is assessed. Station level pressure is also extracted from the ISD and timestamps where the difference between this and the sea-level pressure value exceeds the range from data up to the end of the last full calendar year are flagged.

Monthly data updates are intended to be run in the first 10 days of a calendar month, depending on the availability of data in the ISD and staff resources.

Acknowledgements

RJHD thanks Kate Willett and David Parker for discussions and insight throughout the development of this dataset, Lizzie Good and Nick Rayner for helpful comments on this text, and Gil Compo (UCAR/CIRES) for the suggestion of including the station-level pressure.

This work was funded by the Joint BEIS/Defra Met Office Hadley Centre Climate Programme (GA01101).

References

Dunn, R., Willett, K., Morice, C., and Parker, D. (2014). Pairwise homogeneity assessment of hadisd. *Climate of the Past*, 10(4):1501–1522.

Dunn, R., Willett, K., Thorne, P., Woolley, E., Durre, I., Dai, A., Parker, D., and Vose, R. (2012). Hadisd: a quality-controlled global synoptic report database for selected variables at long-term stations from 1973–2011. *Climate of the Past*, 8(5):1649–1679.

Dunn, R. J. H., Willett, K. M., Parker, D. E., and Mitchell, L. (2016). Expanding hadisd: quality-controlled, sub-daily station data from 1931. *Geoscientific Instrumentation, Methods and Data Systems*, 5(2):473–491.

Lott, J. N. (2004). 7.8 the quality control of the integrated surface hourly database.

Menne, M. J. and Williams Jr, C. N. (2009). Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22(7):1700–1717.

Smith, A., Lott, N., and Vose, R. (2011). The integrated surface database: Recent developments and partnerships. *Bulletin of the American Meteorological Society*, 92(6):704–708.

Met Office
FitzRoy Road
Exeter
Devon
EX1 3PB
United Kingdom

Produced with the Met Office LaTeX template v3.0.0