Clim. Past Discuss., 11, 4569–4600, 2015
www.clim-past-discuss.net/11/4569/2015/
doi:10.5194/cpd-11-4569-2015
© Author(s) 2015. CC Attribution 3.0 License.

Climate of the Past Discussions
Open Access

This discussion paper is/has been under review for the journal Climate of the Past (CP). Please refer to the corresponding final paper in CP if available.

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

R. J. H. Dunn, K. M. Willett, D. E. Parker, and L. Mitchell

Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK

Received: 17 August 2015 – Accepted: 2 September 2015 – Published: 30 September 2015

Correspondence to: R. J. H. Dunn (robert.dunn@metoffice.gov.uk)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

We describe the first major update to the sub-daily station-based HadISD dataset. The temporal coverage of the dataset has been extended to 1931 to present, doubling the time range over which data are provided. Improvements made to the station selection and merging procedures result in 8113 stations being provided in version 2.0.0.2014f of this dataset. This station selection will be reassessed at every annual update, which is likely to result in increasing station numbers over time. The selection of stations to merge together making composites has also been improved and made more robust. The underlying structure of the quality control procedure is the same as for HadISD.1.0.x, but a number of improvements have been implemented in individual tests. Also, more detailed quality control tests for wind speed and direction have been added. The data will be made available as netCDF files at www.metoffice.gov.uk/hadobs/hadisd and updated annually.

1 Introduction

For observational datasets of climate data to remain current and useful for a wide set of potential applications, they require careful curation, nurturing and updating as the characteristics of, and issues with the dataset become known. Over time this results in a set of versions of a dataset, which can arise from something as simple as the inclusion of another year of observations, or be the output of a fundamentally new processing suite including many new and novel techniques. Datasets where this constant reassessment of their quality, coverage and purpose is not performed are likely to be superceded, and in some cases could give misleading results if used in an analysis.

The HadISD dataset (Dunn et al., 2012) took a subset of the station data held in the Integrated Surface Database (ISD) at the National Oceanic and Atmospheric Administration's National Centre for Environmental Information (NOAA/NCEI formerly the National Climatic Data Center (NCDC), Smith et al., 2011; Lott, 2004). These data

CPD
11, 4569–4600, 2015

Expanding HadISD

R. J. H. Dunn et al.

were subject to an objective, automated quality-control procedure which had particular attention paid to retaining true extreme values. The initial data release (v1.0.0.2011f) covered 1973–2011, with annual updates occurring during the early part of each calendar year; the latest update was to v1.0.3.2014f in April 2015. A homogeneity assessment was carried out on v1.0.2.2013f by Dunn et al. (2014) using the Pairwise Homogenisation Algorithm (PHA, Menne and Williams Jr, 2009). As HadISD contains sub-daily data, and the PHA assesses the homogeneity using monthly mean values, the adjustments returned by PHA were not applied to the data. Data files of the adjustment dates and magnitudes were provided, and these can be used to remove the stations with the most and largest inhomogeneities in any analysis. This homogeneity assessment is now part of the annual update process.

In this paper we outline the first major update to HadISD in which we extend the temporal coverage back to 1931 and also improve the station selection process as well as update some of the quality control tests. The overall procedure is very similar to the creation of HadISD.1.0.0 as outlined in Dunn et al. (2012). This new dataset, HadISD.2.0.0, is still a quality-controlled subset of the ~ 29k stations held in the ISD.

In Sect. 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 Sect. 3 with an overview in 4. The data provision is discussed in Sect. 5, with a summary in Sect. 7.

2 Updated station selection and merging

In HadISD.1.0.0 the stations included in the dataset were fixed at the first release, and no changes were made to this station list during the annual updates to the dataset. Therefore these annual updates to HadISD.1.0.x could not benefit from developments in the ISD made at NOAA/NCEI, for example updated station lists and improved coverage resulting from reprocessing. In HadISD.2.0.0 the station selection process becomes part of the general update. This means that each year the stations selected

CC BY

from the ISD may be different from the previous version, as different stations satisfy the selection criteria. As more data are added into the ISD archive and the length of record of meteorological stations grows then the number of stations selected for use in HadISD will also increase. However, it is also possible that improved knowledge of station moves over time will result in ISD station records being split, and hence no longer being of sufficient length to be included in HadISD.2.0.x.

Using the inventory files on the ISD FTP server (ftp://ftp.ncdc.noaa.gov/pub/data/noaa/), 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. The 14806 stations in this initial cut are investigated further using the detailed inventory file. Stations with a median observing interval of six hours or less as well as an equivalent amount of data present of 15 years of observations every six hours, with no requirement on continuity, are retained. This results in 8589 stations being taken forward for further processing. The methodology of this updated station selection procedure is shown in Fig. 1.

2.1 Merging stations

In HadISD.1.0.x, 934 of the final set of stations are composites, again using a static list of station matches. Therefore it is likely that a number of stations within these 8589 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 8589 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 above which stations are merged.

CPD
11, 4569–4600, 2015

In HadISD.1.0.x a hierarchical scoring system was adopted along with a detailed, manual comparison of the temperature anomalies from the ISD-Lite database (ftp://ftp.ncdc.noaa.gov/pub/data/noaa/isd-lite). For HadISD.2.0.0, our selection of merging candidates is based only on the latitude, longitude, elevation and station name. The Euclidean distance between the two stations is calculated using the latitude and longitude. Using an exponential decay with an *e*-folding distance of 25 km, a likelihood of similarity is derived from the station separation. A similar calculation is performed for the elevation, but using an *e*-folding distance of 100 m. The station names are compared using the Jaccard Index (Jaccard, 1901) as in the ISTI merging algorithm. This allows for slight differences in spelling between station names rather than requiring an identical match. If the product of these three probabilities is greater than 0.5, then the stations are deemed similar enough to merge. Using the horizontal and vertical separations and the station name ensures that large differences in any one of these three measures will result in no merger occurring. A reverse check is performed to ensure that a secondary station is not merged into two primary stations; only the primary station with the highest likelihood of a match is used.

Merging stations within the list of candidate stations results in a final list of 8113 stations, of which 2094 contain data from other station IDs which are in the full ISD archive. The increase in the data coverage by including stations from the full ISD holdings can be seen in Fig. 2. When the raw ISD data files are converted to NetCDF prior to processing, the primary stations are read in first, and then all secondaries are read in to fill in any gaps. The focus of HadISD at the moment is on temperature and dewpoint data and so observations are overwritten if those from a secondary station have both temperature and dewpoint in preference to the primary with only one of the two. If only one observation is available out of all stations, then temperature is preferred over dewpoint. Finally, observations closer to the top of the hour are preferred, but at lower importance than the temperature and dewpoint selection.

There are few stations prior to 1931 in the ISD archive, as shown in Fig. 2, hence our decision to only extend the dataset back to 1931. However, by checking in the full

CPD
11, 4569–4600, 2015

R. J. H. Dunn et al.

CC BY

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 distribution of stations can be seen in Fig. 3, and shows the expected high density in Europe and North America (especially the east coast). In HadISD.2.0.0 there are fewer stations in central and southern Africa and also South America. The distribution of merged stations is concentrated in those regions which have longer meteorological records (again Europe and North America, but also Australia). Station list of the final set of candidate stations and mergers are available on the HadISD website at www.metoffice.gov.uk/hadobs/hadisd.

2.2 Extra processing for specific countries

Since the release of HadISD.1.0.0 a number of issues have come to light about countries which have specific problems with the data held in ISD. For two of these, Germany and Canada, we have been able to carry out some extra processing to increase the quality of the station records.

2.2.1 Germany

The stations in Germany have station identifying numbers in the ISD that start with 09 and 10. However, it is the remaining 4 digits of the ID number that uniquely identify the station within Germany (A. Becker, personal communication, 2012). Therefore, we have been able to explicitly merge the 09 stations into the 10 stations. We still perform the merging checks outlined above to ensure that no spurious mergers are performed. This results in 44 stations being merged together prior to the station selection criteria being applied.

2.2.2 Canada

Only 1000 WMO numbers have been assigned for use in Canada, and as a result, many have been re-used when old stations have closed, and new ones opened. In

CPD
11, 4569–4600, 2015
Expanding HadISD
R. J. H. Dunn et al.

CC BY

some cases, this has resulted in apparent station moves in the ISD record. Using a list kindly supplied by Environment Canada (L. Cudlip, personal communication, 2014) we have been able to assess some of the Canadian stations in the ISD record. The list contained information for 994 stations which could be categorised as follows (the number of stations in each is given in parentheses):

- Single – stations which appeared in the list only once (529).
- On/Off – stations which had an "active" and "inactive" status indicating the start and end dates of operation (47).
- Good Station Moves – stations which showed a change in location, with dates showing the end of reporting at the previous location, and the start in the new location (216).
- Overlap Moves – similarly to good station moves, but the start of reporting in the new location occurs before the end of reporting at the old (15).
- Possible Homogeneity issues – multiple dates at a single location, perhaps indicating changes in instrumentation (92).
- Questionable Moves – location changes with no dates given showing the end at one or the beginning at another location (33).
- Dates – cases where "active" and "inactive" statuses occurred at the same time, so the final status could not be determined (49).
- Other – more complex sets of start and end dates that could not be categorised easily (13).

In the ISD, there are more than 1000 stations listed as being in Canada. We selected those which were likely to correspond to the WMO stations (those which have ISD IDs

CPD
11, 4569–4600, 2015

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀ ▶
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

cc BY

that match 71???0–99999). This resulted in 934 stations which we could compare to the Environment Canada list.

Stations which appeared in the Single, On/Off and Possible Homogeneity issues categories were retained in the candidate station list (668). Those from the Questionable Moves, Dates, Overlap moves and Other were rejected from the station list (110).

The 216 stations in the Good Moves list were processed further. Using the station details in the ISD list, the period of time when the station was in this location as determined from the Environment Canada list was extracted. Usually this was the most recent location. The start and end times of the station were adjusted as appropriate to ensure that only the period in the location as given in the full ISD station list was used when further selecting stations. In many cases this will result in the station not being selected for inclusion with HadISD.

Of the 934 Canadian stations we were able to assess, 800 were kept for processing by further selection criteria, in 30 the station names were sufficiently different to reduce the probability of a good merger below the threshold and 104 were rejected. There are other stations which are located in Canada (which do not match the psuedo-WMO IDs used by ISD) which we could not process. These, along with the 30 which were not in the Environment Canada list, were retained in the station selection procedure as we have no information indicating that there are problems with them.

3 Updating the quality control tests

As part of this update we took the chance to re-write the quality control software from IDL into Python, as this language is becoming more commonly used and is also Open Source. All the code used to create HadISD.2.0.x is written in Python, and will be made available alongside the dataset from www.metoffice/gov/uk/hadobs/hadisd/.

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 others, this was

CPD
11, 4569–4600, 2015

CC BY

not possible, primarily those where curve-fitting was used to determine critical values. We have also used this opportunity to improve the functionality of some of the tests. We outline the changes made and the tests where differences exist between the two code versions in the Appendix, but the quality control checks where more substantive changes have been made are detailed below.

3.1 Distributional gap

In HadISD.1.0.x, the second part of the distributional gap test takes all observations within a calendar month (over all years), and by fitting a Gaussian to this distribution determined threshold values. Going outwards from the centre, the distribution is scanned for gaps beyond this threshold value, and any observations occurring beyoud the gap are flagged.

In a number of cases it has come to light that a simple gaussian is not a good fit to the bulk of the observations, resulting in thresholds that are too high. We therefore have increased the complexity of the fitted gaussian by allowing for non-zero skew and kurtosis. This allows the thresholds (as calculated when the fitted curve drops below $y = 0.1$) to occur closer to the bulk of the distribution. In Fig. 4 the asymmetrical nature of the underlying distribution of pressure observations from Durango (764230–99999) can be clearly seen. The closer fit of the Gaussian with skew and kurtosis allows the small set of clearly erroneous observations with an IQR-offset of −4 to be flagged.

3.2 Streaks

The threshold values for straight repeated strings in HadISD.1.0.x were fixed, but dependent on the reporting resolution of the station (see Dunn et al., 2012 Table 4). To allow these thresholds to be calulated dynamically, the distribution of repeated values is analysed. Using an inverse decay curve a new threshold is proposed when this curve falls below 0.1. This threshold is modified by finding the next empty bin to ensure the entire main distribution is retained (see Fig. 5). However, if this dynamically calculated

CPD
11, 4569–4600, 2015

CC BY

threshold is larger than what was used in HadISD.1.0.x, then the old value from Dunn et al. (2012) Table 4 is retained.

3.3 Spike

In HadISD.1.0.x the critical values for determining whether a first-difference may be a spike were determined from the IQR of the first-differences. Similarly to the updated repeated streak check, the updated critical values are calculated from the distribution of first-difference magnitudes. This distribution is again fitted with an inverse decay curve to obtain a first guess at the critical values, which is then modified by finding the next empty bin. This threshold is used if it is smaller than that obtained from the IQR of the first differences.

This test has also been made symmetric, so that the jump down out of the spike has to be greater than the critical value (as opposed to half the critical value as used in HadISD.1.0.x, see Fig. 11 in Dunn et al., 2012).

3.4 Unusual variance check

This test includes a section to select periods in the sea-level pressure which are likely to be the result of intense (tropical) storms. The extreme low pressure at locations which usually have very uniform pressure values increases their monthly variance and so could result in erroneous flagging. Previously the minimum pressure and the maximum wind speed within a calendar month were assessed for contemporaneity and that they were at least 4.5 median absolute deviations (MAD) from the median value. Now, all time periods within a month where both the wind speed and SLP exceed 4 MAD from the median are used when checking for storm signals in case two storms occur within the same calendar month.

CPD
11, 4569–4600, 2015

CC BY

3.5 Winds

The level of quality control applied to the wind speed and direction observations in HadISD.1.0.x was not as high as for temperature, dewpoint temperature and sea-level pressure. Therefore, in HadISD.2.0.0 we have added in a set of logical checks for wind speed and direction as well as testing for the year-to-year consistency of the wind rose for the station.

The logical checks are based on those outlined in DeGaetano (1997) Table 2. By convention, if the wind speed is $0\,\mathrm{m\,s^{-1}}$ then the direction is recorded at 0°, and a northerly wind is recorded as 360°. In ISD, the wind direction has been recorded as missing for calm periods, and so we use these logical checks to set the wind direction as 0° when the speed is $0\,\mathrm{m\,s^{-1}}$. In the remaining four cases shown in Table 1, the observations are flagged.

To quality control the distribution of the wind speed and direction, we use the method outlined in Lucio-Eceiza et al. (2015) to assess rotations between wind roses. Their work focusses on the homogeneity of the wind record, with the aim to adjust erroneous years. In this instance we just remove years where the wind rose is very different to all others.

To perform this assessment of the wind rose, we calculate the root-mean-square error (rmse) for each annual wind rose when compared to that calculated for the entire record. These rmse values are fitted with a Rician distribution (appropriate for rmse values). As in the distributional gap check, we use the location where this fitted curve falls below 0.01 as a proposed threshold, and search outwards for the first empty bin which is used as the final threshold. Any years where the rmse is larger than this are flagged. This test does flag whole years at a time, but will highlight and remove those years where the distribution of wind directions is radically different to the average, identifying possible undocumented station moves.

CPD
11, 4569–4600, 2015

CC BY

In HadISD.2.0.0, wind speeds are now also checked for unusual variance, as well as the odd cluster, streak and record checks which were processed in HadISD.1.0.x. In all these cases the wind direction is now also flagged synergistically.

3.6 Neighbour checks

By increasing the span of the dataset, the selection of neighbours needed to be improved. If the selection method of HadISD.1.0.x had been retained, then it is likely that during the early record, stations would be compared to neighbours that have no data during that time. The new procedure is as follows.

The closest 20 neighbours within the limits of 500 m elevation and 300 km 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. These cycles are removed by first calculating the daily mean, and subtracting that from the data. Then the means for each of the 24 h are calculated over all days, and also removed. Therefore anomalous hours and days will stand out. The linear combination of the correlation coefficient and overlap fraction is used to rank the neighbours, and up to the best ten neighbours are chosen, requiring that at least two occur within each quadrant if possible.

Using these updated neighbouring stations, the remainder of the test is very similar as for HadISD.1.0.x. However, the inter-quartile range of the difference series is calculated for each calendar month separately, rather than for the entire record. For widely separated neighbours, the variations in the station climatology over the annual cycle may result in inter-station differences that are on average larger in some months than others.

During the neighbour checks, some of the intra-station checks are un-done, as documented for HadISD.1.0.x in Dunn et al. (2012). Although this is retained for the odd cluster, climatological, gap and dew-point depression checks, it is no longer performed on the spike check, as a visual inspection showed that the flags on many true spikes were being removed.

CPD
11, 4569–4600, 2015

CC BY

4 Overview of HadISD.2.0.0.2014f and comparison to HadISD.1.0.3.2014f

The summary of the fraction of observations removed for each of the three main variables are shown in Fig. 6. The values for each variable and test are shown in Table 2. As in HadISD.1.0.x, the majority of stations have very low flagging rates, with less than 1 % of observations removed. There are some regional and country-scale patterns that emerge in the flagging rates. For temperature the large regions which have the highest proportion of flagged observations are the eastern and northern North America and western and central Europe. On average the removal rates are higher for the dewpoint temperature than for temperature, but with similar regions showing higher than average removal rates. The majority of stations have comparatively few sea-level pressure observations removed, but the cluster of Mexican stations is still present, but now joined by Japan and parts of the Phillipines. The wind observations show relatively high proportion of flags compared to the other variables, with relatively many stations having more than 5 % of observations removed.

Comparing Fig. 6 to Fig. 20 of Dunn et al., 2012 the patterns of flagging are very similar, despite the different station selection and increase temporal coverage. Similarly, the fraction of stations with a certain percentage of observations removed by a given test (Tables 2 and A2) show very similar patterns of removal to those in Tables 6 and 9 of Dunn et al. (2012). There are, however, some differences. The proportion of stations where repeated values are identified and removed has increased; the result of setting the thresholds dynamically for each station as outlined in Sect. 3.2. Similarly fewer stations have large numbers of spikes identified (Sect. 3.3). The correction of the unusual variance check (Table A1) has increased the fractions of stations with observations removed by this check.

In HadISD.2.0.0 we continue to perform the homogenisation assessment started for HadISD.1.0.2.2013f by Dunn et al. (2014). This uses the Pairwise Homogenisation Algorithm from Menne and Williams Jr (2009) with monthly-mean values as well as monthly-mean diurnal ranges (temperatures and dewpoint temperatures) or monthly-

CPD
11, 4569–4600, 2015

maximum values (wind speeds) calculated from the sub-daily data. The information about the change point locations and magnitudes will be made available along with the dataset, and updated annually. Examples of the distribution of inhomogeneity sizes and their distribution in time are shown in Fig. 7. The distribution of inhomogeneities are very similar to those found for HadISD.1.0.2.2013f in Dunn et al. (2014). Change points are also found in the extended portion of this dataset, before 1973, where fewer of the 8113 stations contain data.

Hence, not only the length of record and quality of the station data, but also the number and size of inhomogeneities are important when assessing stations that are suitable for climate monitoring. Therefore we do not perform a selection on these lines as the requirements for this will differ between applications. We encourage users to make their own assessment as which stations are suitable for their particular investigation.

5 Data provision

HadISD.2.0.0 is provided as Network Common Data Format version 4 files (NetCDF4) at www.metoffice.gov.uk/hadobs/hadisd/. We have moved from NetCDF3 files as used in HadISD.1.0.x to NetCDF4. This format allows for internal compression, and so results in smaller file sizes on disc, which will hopefully make them easier to process and download. The inventory files, log-files of the processing and also summary plots will also be made available alongside the updated data files. A list of the fields available in each NetCDF file are given in Table 3. Of note is that wind gust, past significant weather and the precipitation variables have not been quality controlled.

The versioning scheme will be the same as for HadISD.1.0.x, with annual updates occurring at the beginning of each calendar year. To ensure that as much data from the previous year is included in the updates, these are carried out in a two stage process. A preliminary dataset will be released early in the year (for example v2.0.1.2015p in

CPD
11, 4569–4600, 2015

January 2016) with a final version (e.g. v2.0.1.2015f) a few months later to ensure that late-arriving data are included.

6 Derived hourly quantities: humidity and heat stress

The HadISDH.2.0.0 dataset (Willett et al., 2014) of monthly humidity measures is based on the HadISD.1.0.x observations. The sub-daily observations are converted to monthly measures and homogenised to enable long-term climate monitoring of land-surface humidity. In HadISD.2.0.0 we also release data files containing sub-daily humidity and heat-health measures. These are calculated directly from the sub-daily observations of temperature, dewpoint temperature and pressure.

The formulae we use are the same as in HadISDH (see Willett et al., 2014 for full details) but we give the method here with the specific formulae in Table 4. Firstly the sub-daily sea-level pressure values provided in HadISD are converted to station-level pressure using the formula from List (1963). This is different to HadISDH, where the climatological monthly mean sea-level pressure values from the 20th Century Reanalysis V2 (Compo et al., 2011) were used.

The temperature, dewpoint temperature and station pressure are then used to calculate the vapour pressure with respect to water. This is used to calculate the wet-bulb temperature. If this wet-bulb temperature is below 0° C then the process is repeated using the formulae with respect to ice. The resulting vapour pressure values are used to obtain the specific and relative humidities.

On top of this, these humidity values are used to derive a number of heat-stress metrics on an hourly basis. These are outlined in Table 5. These will allow the study of individual heat wave events not only through meteorological variables but also those which capture the impact on human heat-health.

Neither of these two sets of variables have been quality controlled or homogenised separately, and will inherit any remaining data issues present within the input variables drawn from HadISD. However the homogeneity information from the temperatures and

CPD
11, 4569–4600, 2015

CC BY

dewpoint temperatures will be suitable to select stations with few and small inhomogeneities.

7 Summary

We present the first major update to the sub-daily station-based HadISD dataset where the temporal coverage has been extended back to 1931. As part of this the station selection and merging algorithms have been updated, and will be run as part of the annual update cycle. HadISD.2.0.0.2014f contains 8113 stations of which 2094 are composites resulting from the merging procedure. The quality control tests have been adjusted to account for the increased length of record, but also improved to take advantage of our increased knowledge of the dataset and the extremes within it. More detailed quality control tests have been applied to the wind speed and direction observations. The temperature and dewpoint observations have been used to create sub-daily humidity and heat-stress datasets. All data and Supplement files will be made available at www.metoffice.gov.uk/hadobs/hadisd.

Appendix

Here we detail the changes in the quality control tests that have occurred on conversion to Python.

Acknowledgements. We thank Andreas Becker (DWD) and Lee Cudlip (EC) for their help and suggestions for improving the records in Germany and Canada respectively.

This work was partly funded by the European Union under the 7th Framework Programme Collaborative Project ERA-CLIM2, Grant Agreement Number 607029 and also the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). This work is distributed under the Creative Commons Attribution 3.0 License together with an author copyright. This license does not conflict with the regulations of the Crown Copyright.

References

ACSM: Prevention of thermal injuries during distance running, Med. Sci. Sport. Exer., 16, iv–xiv, 1984. 4591

Buck, A. L.: New equations for computing vapor pressure and enhancement factor, J. Appl. Meteorol., 20, 1527–1532, 1981. 4590

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D. and Worley, S. J.: The twentieth century reanalysis project, Q. J. Roy. Meteor. Soc., 137, 1–28, 2011. 4583

DeGaetano, A. T.: A quality-control routine for hourly wind observations, J. Atmos. Ocean. Tech., 14, 308–317, 1997. 4579, 4587

Dikmen, S. and Hansen, P.: Is the temperature-humidity index the best indicator of heat stress in lactating dairy cows in a subtropical environment?, J. Dairy Sci., 92, 109–116, 2009. 4591

Dunn, R. J. H., Willett, K. M., Thorne, P. W., Woolley, E. V., Durre, I., Dai, A., Parker, D. E., and Vose, R. S.: HadISD: a quality-controlled global synoptic report database for selected variables at long-term stations from 1973–2011, Clim. Past, 8, 1649–1679, doi:10.5194/cp-8-1649-2012, 2012. 4570, 4571, 4577, 4578, 4580, 4581

Dunn, R. J. H., Willett, K. M., Morice, C. P., and Parker, D. E.: Pairwise homogeneity assessment of HadISD, Clim. Past, 10, 1501–1522, doi:10.5194/cp-10-1501-2014, 2014. 4571, 4581, 4582, 4600

El Fadli, K. I., Cerveny, R. S., Burt, C. C., Eden, P., Parker, D., Brunet, M., Peterson, T. C., Mordacchini, G., Pelino, V., Bessemoulin, P., Stella, J. L., Driouech, F., Wahab, M. M. A., and Pace, M. B.: World Meteorological organization assessMent of the PurPorted World record 58 c teMPerature extreMe at el azizia, libya (13 sePteMber 1922), B. Am. Meteorol. Soc., 94, 199–204, 2013. 4592

Jaccard, P.: Etude comparative de la distribution florale dans une portion des Alpes et du Jura, vol. 37, Impr. Corbaz, 1901. 4573

Jensen, M. E., Burman, R. D., and Allen, R. G.: Evapotranspiration and Irrigation Water Requirements, ASCE, 1990. 4590

CC BY

List, R. J.: Smithsonian Meteorological Tables, Vol. 114, 6th Edn., Smithsonian Institution, Washington DC, 268 pp., 1963. 4583, 4590

Lott, J. N.: The quality control of the integrated surface hourly database, 84th American Meteorological Society Annual Meeting, 2004, Seattle, WA, American Meteorological Society, Boston, MA, 7.8, 2004. 4570

Lucio-Eceiza, E. E., González-Rouco, J. F., Navarro, J., Beltrami, H., Hidalgo, A., and Conte, J.: Quality control of surface wind observations in north eastern North America. Part II: Measurement errors, J. Atmos. Ocean. Tech., submitted, 2015. 4579

Masterton, J. and Richardson, F.: Humidex: A method of quantifying human discomfort due to excessive heat and humidity, Report No., CL1 1-79. Downsview, Ontario: Atmospheric Environment Service, Environment Canada, 1979. 4591

Menne, M. J. and Williams Jr, C. N.: Homogenization of temperature series via pairwise comparisons, J. Climate, 22, 1700–1717, 2009. 4571, 4581

Peixoto, J. and Oort, A. H.: The climatology of relative humidity in the atmosphere, J. Climate, 9, 3443–3463, 1996. 4590

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, 1, 75–102, 2014. 4572

Rothfusz, L. P.: The heat index "equation" or more than you ever wanted to know about heat index: National Weather Service Southern Region technical attachment SR/SSD 90-23, Fort Worth: National Weather Service 1990. 4591

Smith, A., Lott, N., and Vose, R.: The integrated surface database: recent developments and partnerships, B. Am. Meteorol. Soc., 92, 704–708, 2011. 4570

Steadman, R. G.: Norms of apparent temperature in Australia, Aust. Meteorol. Mag., 43, 1–16, 1994. 4591

Willett, K. M., Dunn, R. J. H., Thorne, P. W., Bell, S., de Podesta, M., Parker, D. E., Jones, P. D., and Williams Jr., C. N.: HadISDH land surface multi-variable humidity and temperature record for climate monitoring, Clim. Past Discuss., 10, 2717–2766, doi:10.5194/cpd-10-2717-2014, 2014. 4583

CPD
11, 4569–4600, 2015
Expanding HadISD
R. J. H. Dunn et al.

Table 1. Logical Wind Checks used in HadISD.2.0.0, adapted from (DeGaetano, 1997).

1	Speed $< 0\,\mathrm{m\,s^{-1}}$
2	Direction $< 0^\circ$ or $> 360^\circ$
3	If direction $= 0^\circ$, speed $\neq 0\,\mathrm{m\,s^{-1}}$
4	If speed $= 0\,\mathrm{m\,s^{-1}}$, direction $\neq 0^\circ$

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

CPD
11, 4569–4600, 2015

Expanding HadISD

R. J. H. Dunn et al.

Table 2. Summary of removal of data from individual stations by the different tests for the 8113 stations considered in detailed analysis.

Test (Number)	Variable	Number of stations in each detection rate band (as % of total original observations removed) 0	0–0.1	0.1–0.2	0.2–0.5	0.5–1.0	1.0–2.0	2.0–5.0	> 5.0
Duplicate months check	All	8098	0	0	0	0	1	0	14
Odd cluster check	T	2098	4681	334	154	16	5	0	15
	Td	2782	4736	313	178	21	3	4	76
	SLP	2070	4074	623	424	122	58	48	694
	ws	1849	4509	840	734	172	6	0	3
Frequent values check	T	7966	105	10	13	4	6	6	3
	Td	7925	109	13	23	12	10	13	8
	SLP	7942	35	16	22	11	11	12	64
Diurnal cycle check	All	7575	12	83	203	102	57	37	44
Distributional gap check	T	2053	5319	248	223	117	72	53	28
	Td	1043	5929	479	353	134	94	59	22
	SLP	2991	3748	423	420	200	122	99	110
Known records check	T	8032	81	0	0	0	0	0	0
	Td	8113	0	0	0	0	0	0	0
	SLP	6870	1115	22	24	20	30	25	7
	ws	8113	0	0	0	0	0	0	0
Repeated streaks/unusual spell frequency check	T	4567	2016	340	397	284	300	188	21
	Td	4041	1950	323	547	435	475	309	33
	SLP	7370	613	51	33	18	10	13	5
	ws	5645	1080	352	403	288	213	104	28
Climatological outliers check	T	1201	6078	449	217	91	35	29	13
	Td	828	6265	519	324	106	39	26	6
Spike check	T	2669	5304	79	43	11	5	2	0
	Td	828	6885	81	42	6	2	2	0
	SLP	2838	5193	40	29	6	4	3	0
T and Td cross-check: Supersaturation	T, Td	8113	0	0	0	0	0	0	0
T and Td cross-check: Wet bulb drying	Td	4406	2649	346	352	163	98	74	25
T and Td cross-check: Wet bulb cutoffs	Td	5822	395	427	615	338	249	184	83
Cloud Clean-up	c	420	772	403	817	1049	1641	1986	1025
Unusual variance check	T	5933	76	494	980	404	153	52	21
	Td	5883	44	491	957	441	203	77	7
	SLP	6873	25	284	501	282	104	38	6
	ws	5443	205	618	1000	461	272	102	12
Nearest neighbour data check	T	1740	6085	97	76	60	24	21	10
	Td	1553	6194	162	105	54	22	19	4
	SLP	2758	4876	249	139	38	25	17	11
Station clean up	T	1533	2662	886	1590	856	342	133	111
	Td	1228	1842	928	1691	1101	693	366	264
	SLP	1696	2243	578	816	633	470	575	1102
	ws	1613	3520	903	1243	634	243	99	88
Logical Wind	wd	5735	1500	231	314	176	106	41	10
Wind Rose	ws	4354	1810	131	205	215	335	621	442

CPD
11, 4569–4600, 2015
Expanding HadISD
R. J. H. Dunn et al.

CC BY

Table 3. Variables present within the NetCDF files in HadISD.2.0.0. The second column indicates whether the value is an instantaneous measure or a time averaged quantity. The third column shows the subset that we quality controlled.

Variable	Instantaneous (I) or past period (P) measurement	Subsequent QC
Temperature	I	Y
Dewpoint	I	Y
SLP	I	Y
Total cloud cover	I	Y
High cloud cover	I	Y
Medium cloud cover	I	Y
Low cloud cover	I	Y
Cloud base	I	N
Wind speed	I	Y
Wind direction	I	Y
Wind gust	I	N
Past significant weather #1	P	N
Precipitation depth #1	P	N
Precipitation period #1	P	N
True Input Station	–	–
QC flags	–	–
Flagged observations	–	–

CPD
11, 4569–4600, 2015

CC BY

Table 4. Humidity formulae used in HadISD.2.0.0, as used in HadISDH.2.0.0 (Willett et al, 2014).

Variable	Equation	Source	Notes
Specific humidity (q) in g kg^{-1}	$q = 1000\left(\frac{0.622e}{P_{\mathrm{mst}} - ((1-0.622)e)}\right)$	Peixoto and Oort (1996)	
Relative humidity (RH) in %rh	$\mathrm{RH} = 100\left(\frac{e}{e_s}\right)$		
Vapour Pressure (e) with respect to water in HPa (when $T_w > 0^{\deg}$)	$e = 6.1121 \times f_w \times \exp\left(\frac{\left(18.729 - \left(\frac{T_d}{227.3}\right)\right)T_d}{257.87 + T_d}\right)$ $f_w = 1 + 7\times10^{-4} + 3.46\times10^{-6}P_{\mathrm{mst}}$	Buck (1981)	Substitute T for T_d to give the saturation vapour pressure e_s
Vapour Pressure (e) with respect to ice in HPa (when $T_w \le 0^{\deg}$ C)	$e = 6.1115 \times f_w \times \exp\left(\frac{\left(23.036 - \left(\frac{T_d}{333.7}\right)\right)T_d}{279.82 + T_d}\right)$ $f_w = 1 + 3\times10^{-4} + 4.18\times10^{-6}P_{\mathrm{mst}}$	Buck (1981)	
Wet bulb temperature (T_w) in $^{\deg}$ C	$T_w = \frac{aT + bT_d}{a+b}$ $a = 6.6\times10^{-5}P_{\mathrm{mst}}$ $b = \frac{409.8e}{(T_d + 237.3)^2}$	Jensen et al. (1990)	
Station Pressure in hPa	$P_{\mathrm{mst}} = P_{\mathrm{msl}}\left(\frac{T}{T + 0.0065Z}\right)^{5.625}$	List (1963)	Temperature T, station height Z in metres

CPD
11, 4569–4600, 2015

Table 5. Heat stress measures calcualted in HadISD.2.0.0.

Variable	Equation	Source	Notes
Temperature–Humidity Index (THI)	$\mathrm{THI} = (1.8T + 32) - (0.55 - 0.0055\mathrm{RH})(1.8T - 26))$	Dikmen and Hansen (2009)	
Pseudo Wet-bulb Globe Temperature (WBGT)	$\mathrm{WBGT} = (0.567T) + (0.393e_v) + 3.94$	ACSM (1984)	
Humidex	$h = T + (0.5555(e_v - 10))$	Masterton and Richardson (1979)	
Apparent Temperature	$T_a = T + (0.33e_v) - (0.7w) - 4$	Steadman (1994)	
Heat Index	$\mathrm{HI} = -42.379 + 2.04901523T_f + 10.14333127\mathrm{RH} - 0.22475541T_f\mathrm{RH} - 0.0068378837T_f^2 - 0.05481717\mathrm{RH}^2 + 0.001228747T_f^2\mathrm{RH} + 8.5282 \times 10^{-4}T_f\mathrm{RH}^2 - 1.99 \times 10^{-6}T_f^2\mathrm{RH}^2$ $\mathrm{adj}_1 = \frac{13\mathrm{RH}}{4}\sqrt{\frac{17\mathrm{abs}(T_f - 95)}{17}}$ $\mathrm{adj}_2 = \frac{\mathrm{RH} - 85}{10} \times \frac{87 - T_f}{5}$ $\mathrm{HI} = 0.5(T_f + 61 + 1.2(T_f - 68) + 0.094\mathrm{RH})$	Rothfusz (1990)	Where T_f is the temperature in Fahreneit. If $\mathrm{RH} < 13$ and $80 \le T_f \le 112$, adj_1 is subtracted from HI; if $\mathrm{RH} > 85$ and $80 \le T_f \le 87$ adj_2 is added to HI. Furthermore, if these calculations would result in a $\mathrm{HI} < 80$, then the simpler formula is used.

CPD
11, 4569–4600, 2015
Expanding HadISD
R. J. H. Dunn et al.

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

CC BY

Table A1. Summary of changes in tests.

Test	T	Td	SLP	ws	wd	clouds	Changes and Notes
			Applies to				
Intra-station							
Duplicate months check	X	X	X	X	X	X	No Change
Odd cluster check	X	X	X	X	X		Wind direction flagged using wind speed
Frequent values check	X	X	X				Bug which prevented DJF from being correctly processed fixed
Diurnal cycle check	X	X	X	X	X	X	No Change
Distributional gap check	X	X	X				Threshold values calculated from Gaussian allowing for non-zero skew and kurtosis.
Known record check	X	X	X	X	X		Values updated to account for El Fadli et al. (2013). Wind direction flagged using wind speed.
Repeated streaks/unusual spell frequency check	X	X	X	X	X		Threshold calculated from distribution of length of runs of repeated values. Wind direction flagged using wind speed
Climatological outliers check	X	X					Threshold values can change because of differences in the fitted Gaussian curve
Spike check	X	X	X				Bug arising from single and double precision values fixed. Threshold calculated from distribution of first differences. Changes resulting from the way missing/flagged values are handled when calculating first differences. Test now symmetric.
T and Td cross-check: Supersaturation		X					No Change
T and Td cross-check: Wet bulb drying		X					No Change
T and Td cross-check: Wet bulb cutoffs		X					Improved calculation of reporting frequencies results in minor changes.
Cloud coverage logical checks						X	No Change
Unusual variance check	X	X	X	X	X		Bug fixed so test applies to all observations not just the unflagged ones
Wind checks				X	X		Logical and wind-rose check added
Inter-station							
Nearest neighbour data check	X	X	X				Neighbours selected using correlation and data-overlap values. Distributions of differences calculated on monthly basis. Unflagging of Odd Cluster check improved, but removed for the Spike Check as it was retaining obvious spikes.
Station clean up	X	X	X	X	X		

CPD
11, 4569–4600, 2015
Expanding HadISD
R. J. H. Dunn et al.

Table A2. As Table 2 but in %.

Test (Number)	Variable	Stations with detection rate band (% of total original observations) 0	0–0.1	0.1–0.2	0.2–0.5	0.5–1.0	1.0–2.0	2.0–5.0	> 5.0
Duplicate months check	All	99.8	0.0	0.0	0.0	0.0	0.0	0.0	0.2
Odd cluster check	T	35.8	57.7	4.1	1.9	0.2	0.1	0.0	0.2
	Td	34.3	58.4	3.9	2.2	0.3	0.0	0.0	0.9
	SLP	25.5	50.2	7.7	5.2	1.5	0.7	0.6	8.6
	ws	22.8	55.6	10.4	9.0	2.1	0.1	0.0	0.0
Frequent values check	T	98.2	1.3	0.1	0.2	0.0	0.1	0.1	0.0
	Td	97.7	1.3	0.2	0.3	0.1	0.1	0.2	0.1
	SLP	97.9	0.4	0.2	0.3	0.1	0.1	0.1	0.8
Diurnal cycle check	All	93.4	0.1	1.0	2.5	1.3	0.7	0.5	0.5
Distributional gap check	T	25.3	65.6	3.1	2.7	1.4	0.9	0.7	0.3
	Td	12.9	73.1	5.9	4.4	1.7	1.2	0.7	0.3
	SLP	36.9	46.2	5.2	5.2	2.5	1.5	1.2	1.4
Known records check	T	99.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
	Td	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SLP	84.7	13.7	0.3	0.3	0.2	0.4	0.3	0.1
	ws	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Repeated streaks/unusual spell frequency check	T	56.3	24.8	4.2	4.9	3.5	3.7	2.3	0.3
	Td	49.8	24.0	4.0	6.7	5.4	5.9	3.8	0.4
	SLP	90.8	7.6	0.6	0.4	0.2	0.1	0.2	0.1
	ws	69.6	13.3	4.3	5.0	3.5	2.6	1.3	0.3
Climatological outliers check	T	14.8	74.9	5.5	2.7	1.1	0.4	0.4	0.2
	Td	10.2	77.2	6.4	4.0	1.3	0.5	0.3	0.1
Spike check	T	32.9	65.4	1.0	0.5	0.1	0.1	0.0	0.0
	Td	13.5	84.9	1.0	0.5	0.1	0.1	0.0	0.0
	SLP	35.0	64.0	0.5	0.4	0.1	0.0	0.0	0.0
T and Td cross-check: Supersaturation	T, Td	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
T and Td cross-check: Wet bulb drying	Td	54.3	32.7	4.3	4.3	2.0	1.2	0.9	0.3
T and Td cross-check: Wet bulb cutoffs	Td	71.8	4.9	5.3	7.6	4.2	3.1	2.3	1.0
Cloud Clean-up	c	5.2	9.5	5.0	10.1	12.9	20.2	24.5	12.6
Unusual variance check	T	73.1	0.9	6.1	12.1	5.0	1.9	0.6	0.3
	Td	72.5	0.5	6.1	11.9	5.4	2.5	0.9	0.1
	SLP	84.7	0.3	3.5	6.2	3.5	1.3	0.5	0.1
	ws	67.1	2.5	7.6	12.3	5.7	3.4	1.3	0.1
Nearest neighbour data check	T	21.4	75.0	1.2	0.9	0.7	0.3	0.3	0.1
	Td	19.1	76.3	2.0	1.3	0.7	0.3	0.2	0.0
	SLP	34.0	60.1	3.1	1.7	0.5	0.3	0.2	0.1
Station clean up)	T	18.9	32.8	10.9	19.6	10.6	4.2	1.6	1.4
	Td	15.1	22.7	11.4	20.8	13.6	8.5	4.5	3.3
	SLP	20.9	27.6	7.1	10.1	7.8	5.8	7.1	13.6
	ws	19.9	34.6	8.1	11.7	7.7	5.5	7.6	4.9
Logical Wind	wd	70.7	18.5	2.8	3.9	2.2	1.3	0.5	0.1
Wind Rose	ws	53.7	22.3	1.6	2.5	2.7	4.1	7.7	5.4

CPD
11, 4569–4600, 2015

Expanding HadISD

R. J. H. Dunn et al.

CC BY

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

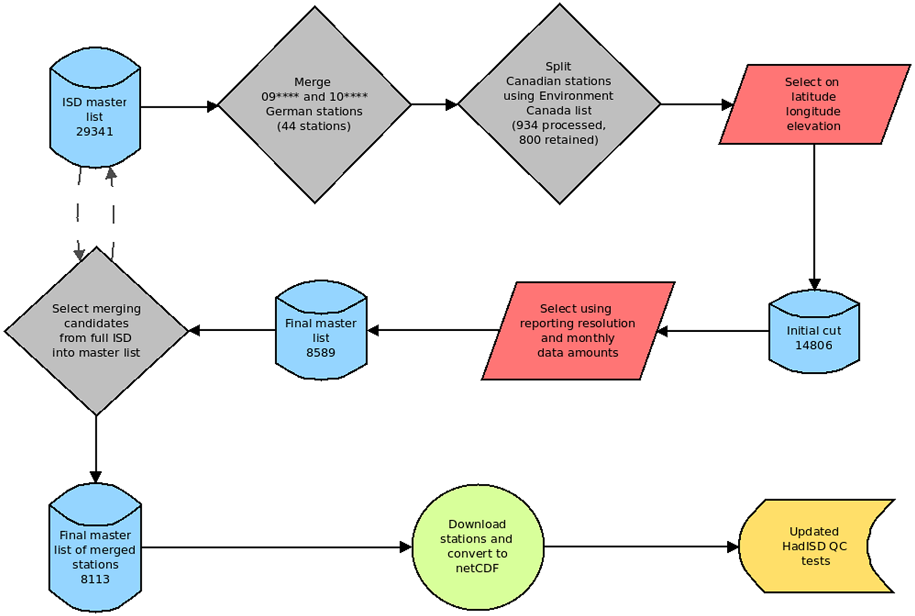

Figure 1. The process used for the station selection and merging in HadISD.2.0.0.

CPD
11, 4569–4600, 2015
Expanding HadISD
R. J. H. Dunn et al.

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

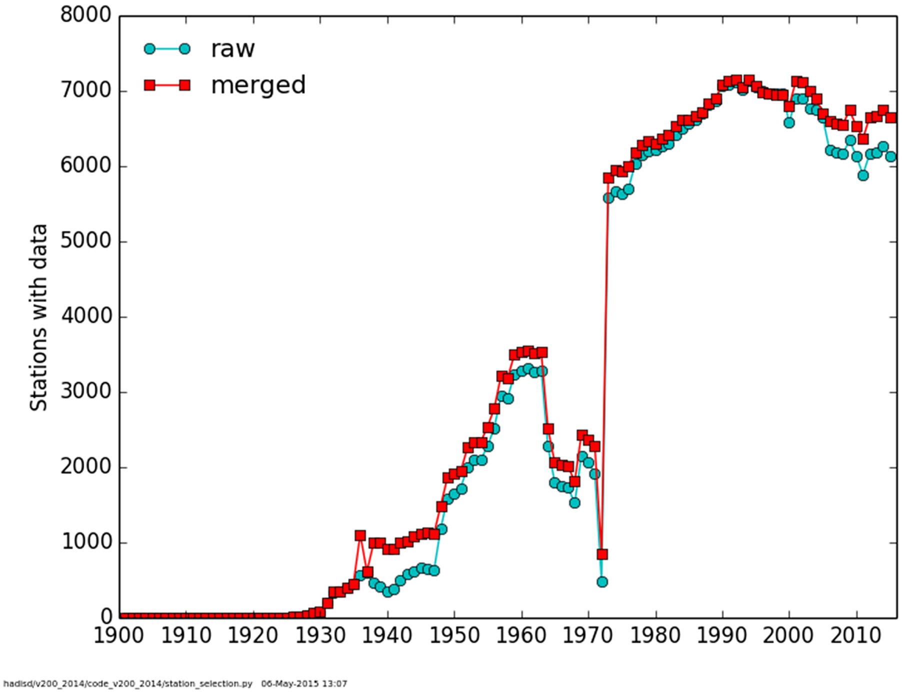

Figure 2. The distribution of stations with time before (cyan circles) and after (red squares) merging.

CPD
11, 4569–4600, 2015

CC BY

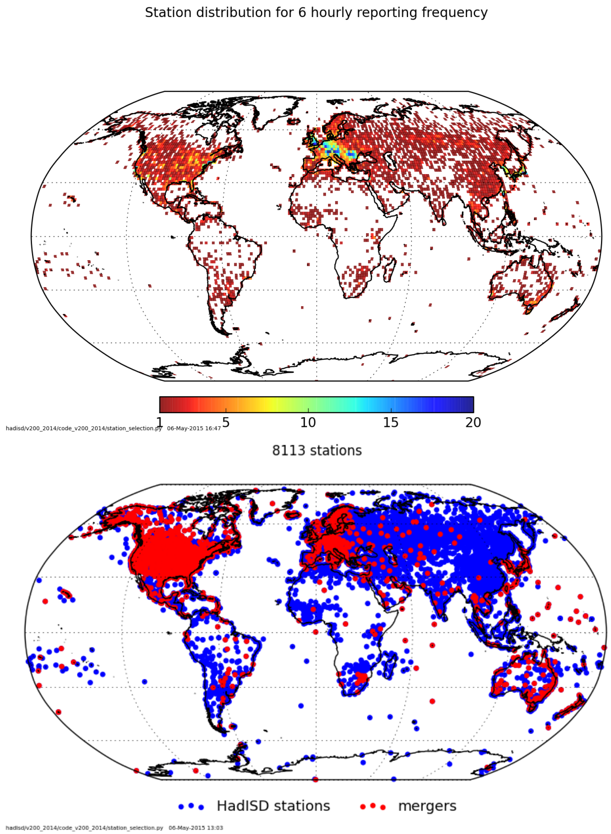

Figure 3. Top: the location of the final set of stations. For presentational purposes we show the number of stations within 1° × 1° grid boxes. Bottom: the locations of the 2094 stations which are composites.

CPD
11, 4569–4600, 2015

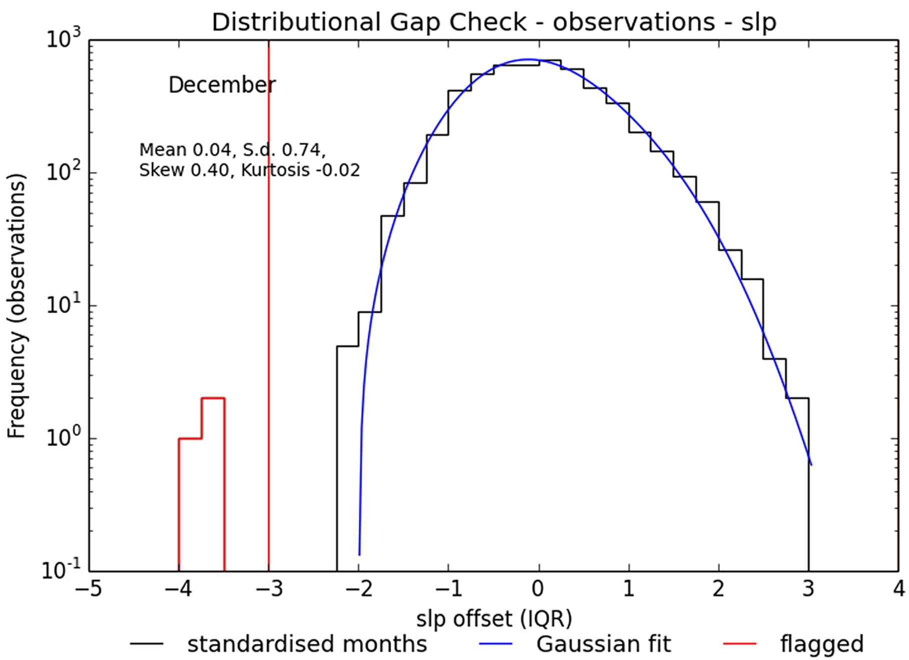

Figure 4. The improved distributional gap check working on SLP data from 764230–99999 Durango (24.06° N, 104.60° W, 1872 m). Using a Gaussian without skew and kurtosis may have included cluster of observations at around −4IQR which are removed in this upgraded test.

CC BY

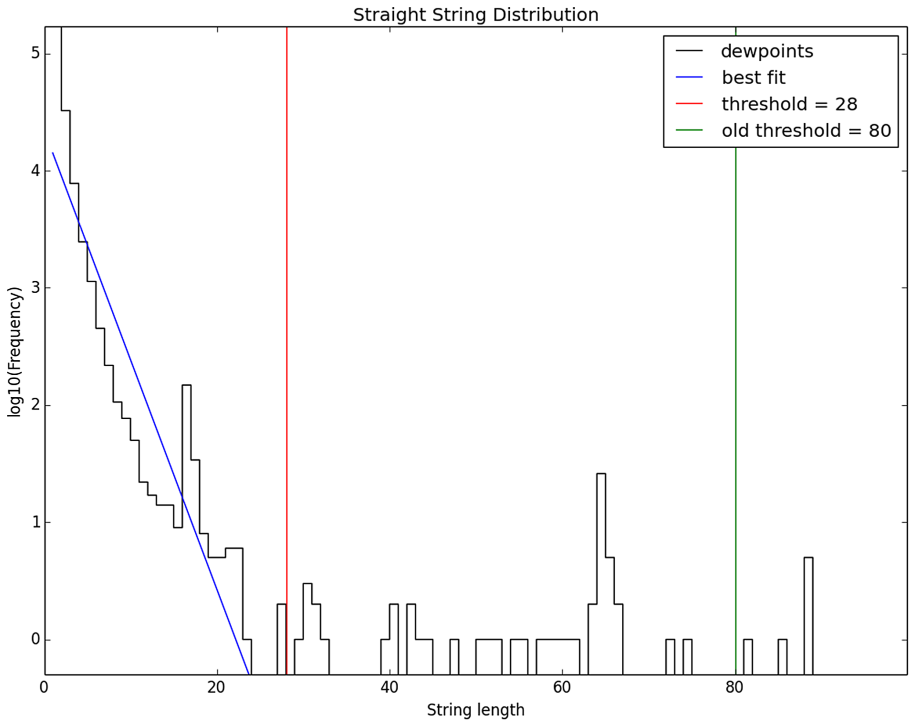

Figure 5. The dynamic threshold assignment from the improved streak check on dewpoint temperature data from 724750–99999 Milford Municipal Airport (38.4° N, 113.0° W, 1536 m). The threshold used in HadISD.1.0.0 retained a large number of streaks of repeated values which are now removed from this station.

CPD
11, 4569–4600, 2015

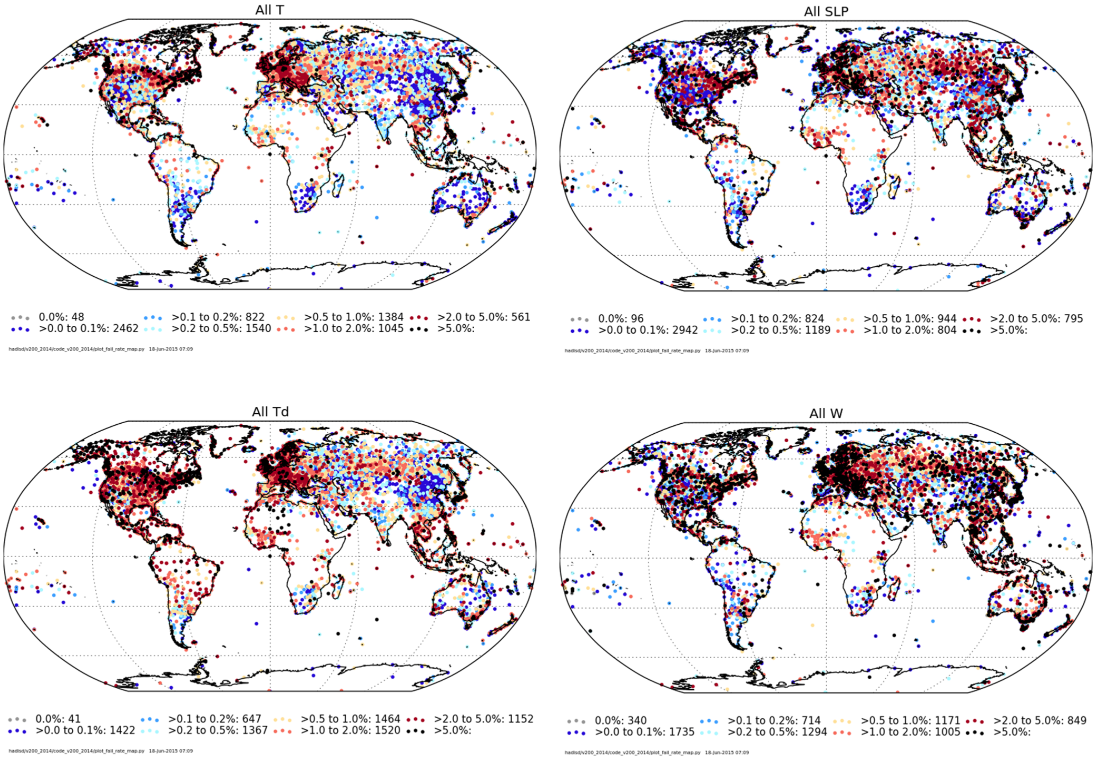

Figure 6. Rejection rates by variable for each station showing the temperature, dewpoint temperature, sea-level pressure and wind speed. Different rejection rates are show by different colours, and the legend also shows the number of stations in each band. The stations with a greater proportion of observations flagged are plotted on top.

CPD
11, 4569–4600, 2015

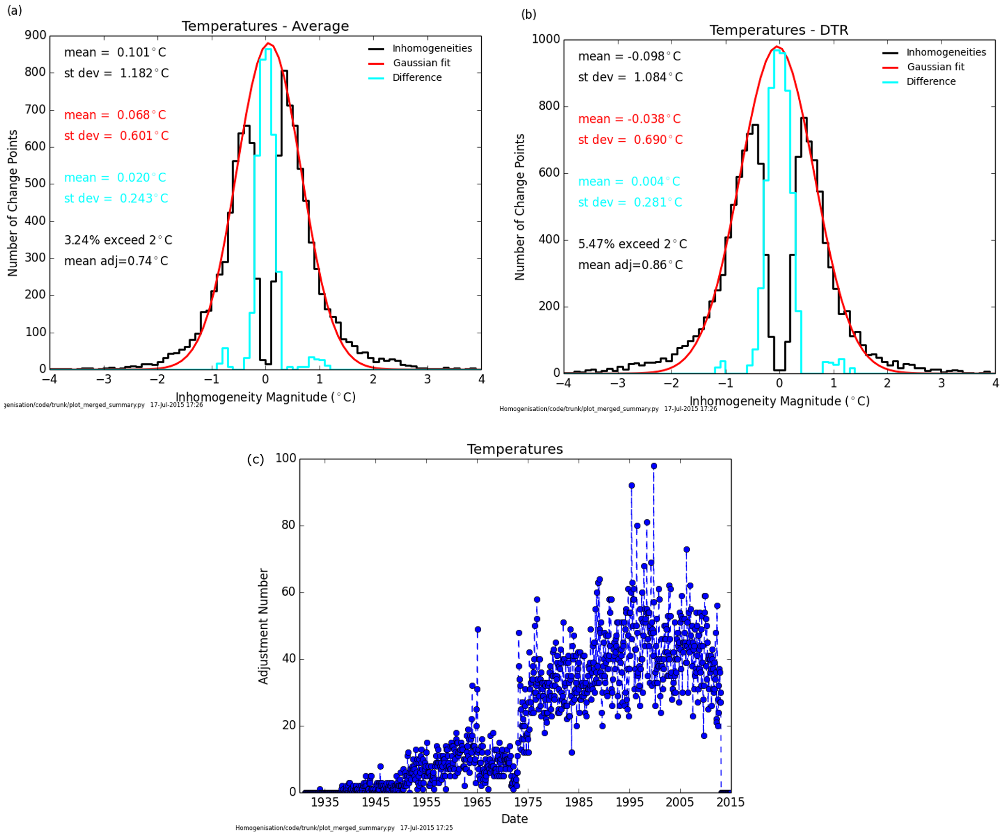

Figure 7. The distribution of inhomogeneities using the monthly-mean (top) temperatures and (middle) diurnal temperature range. Bottom: the number of change points found in each year from both the calculation methods (see Dunn et al., 2014 for full details).

CPD
11, 4569–4600, 2015