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

Robert J. H. Dunn[1], Kate M. Willett[1], David E. Parker[1], and Lorna Mitchell[1]

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

Correspondence to: robert.dunn@metoffice.gov.uk

Abstract. HadISD is a sub-daily, station-based, quality-controlled dataset designed to study past extremes of temperature, pressure and humidity and allow comparisons to future projections. Herein we describe the first major update to the 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 7677 stations being provided in

5   version 2.0.0.2015p of this dataset. 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.

10   ## 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

15   constant reassessment of their quality, coverage and purpose is not performed are likely to be superseded, and in some cases could give misleading results if used in an analysis.

The HadISD dataset (Dunn et al., 2012) used 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 were subject to an objective,

20   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 v1.0.4.2015f in July 2016. 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.

The initial release of HadISD.1.0.0 started in 1973 because of poor data availability in ISD in 1972 owing to a change in the Global Telecommunications System (GTS). Prior to 1972 there are more records in the ISD, but not as many as after this year.

5   Therefore HadISD.1.0.0 and subsequent versions concentrated on the period which had the greatest number of records. Now it is worthwhile re-addressing the station selection procedure (which had been static since around 2010) and at the same time increasing the timespan of this dataset. This will enable the dataset to be more easily used for the study of long-term changes, as the ∼40 years of HadISD.1.0.x was limiting in this regard.

In this paper we outline the first major update to HadISD in which we extend the temporal coverage back to 1931 and also

10   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 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 with an overview in Section 4. The data provision is discussed

15   in Section 5, with a note on derived humidity and heat-stress quantities newly provided in HadISD.2.0.0 in Section 6, and a summary in Section 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

20   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 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

25   in ISD station records being split, and hence no longer being of sufficient length to be included in HadISD2.0.0.

Using the inventory files on the ISD FTP server (ftp.ncdc.noaa.gov/pub/data/noaa/), stations are selected on the basis of a number of requirements. The first stage is to process stations within Germany and Canada separately to account for known issues with the station IDs in these countries. This process is fully described in Section 2.2, as these processes use the merging algorithm described therein. These updated station lists are used for these countries.

30   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 to pass the first cut. The (currently) 14957 stations in this initial cut are investigated further using the detailed inventory file. The median reporting interval is checked to ensure that it is six hourly or less (yielding a median of at least 120 reports per month overall), and also on a calendar-month basis so that each calendar month has a median of at least
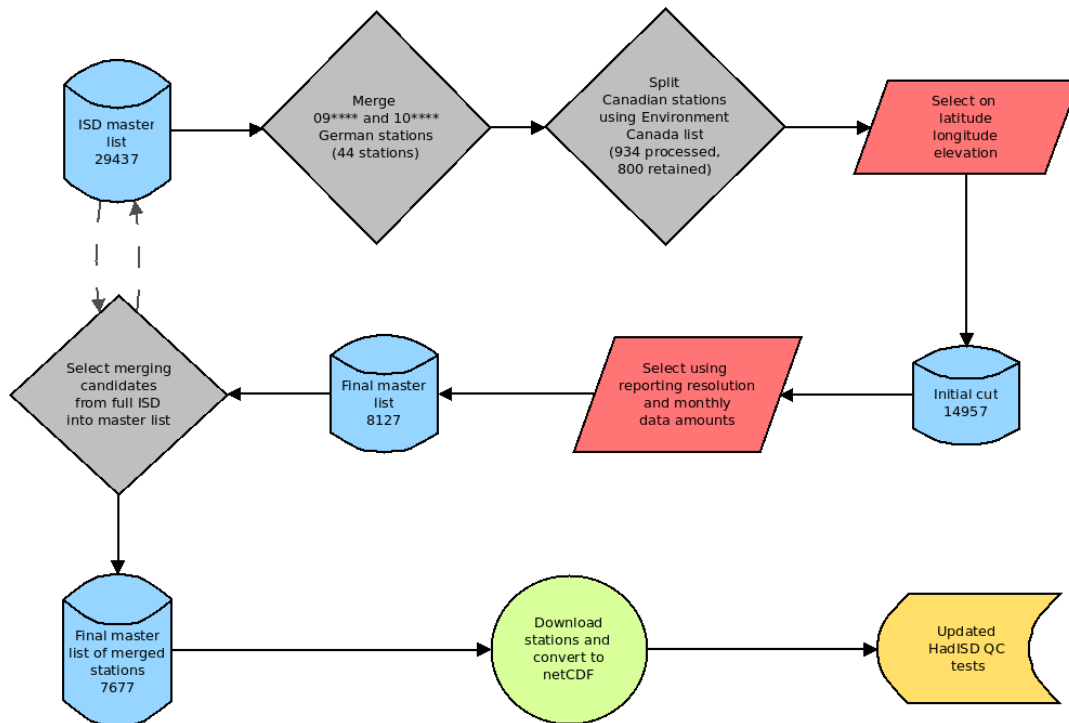
Figure 1. The process used for the station selection and merging in HadISD.2.0.0

120 observations. Finally, there have to be 15 years worth of months which have the equivalent amount of data for observations every six hours. Stations which pass these three checks, with no requirement on continuity, are retained. This results in 8127 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, using its static list of station matches. Therefore it is likely that a number of stations within the 8127 taken forward for HadISD.2.0.0 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 8127 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 with a station-name similarity algorithm and a threshold set above which stations are merged.

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.ncdc.noaa.gov/pub/data/noaa/isd-lite). Merged stations could also be composited out of many short records, which on their own would have had insufficient data to be included in the final station list.

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 latitudes and longitudes of the stations are only stored to a precision of a single decimal place, which in the worst case can result in an accuracy of only ∼5 km. In cases where distinct stations are close (urban stations, for example) but have very similar names, then this may result in an erroneous merge.

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, arising for example from different spellings of the same station in countries with multiple languages or non-Roman alphabets.

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 preclude merging. 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 7677 stations, of which 1993 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. These secondary stations have no requirements on record length or median reporting interval.

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. This figure also shows why HadISD.1.0.x was chosen to start in 1973. However, by checking in the full ISD catalogue for stations to merge with, the coverage has been improved back to 1950, as well as smaller improvements at other times. The distribution of stations in the first, last and every 20th year is shown in the Appendix (Fig. 9). Although in the very early years, only small parts of the globe have any coverage, the increase in station numbers from the late 1940s to 1960 results in a much more comprehensive coverage of the globe.

The distribution of all 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 the new station selection and merging scheme has yielded fewer stations in central and southern Africa and also South America than there are in HadISD.1.0.x. The distribution of merged stations is concentrated in those regions which have longer meteorological records (again Europe and North America, but also Australia). Station lists
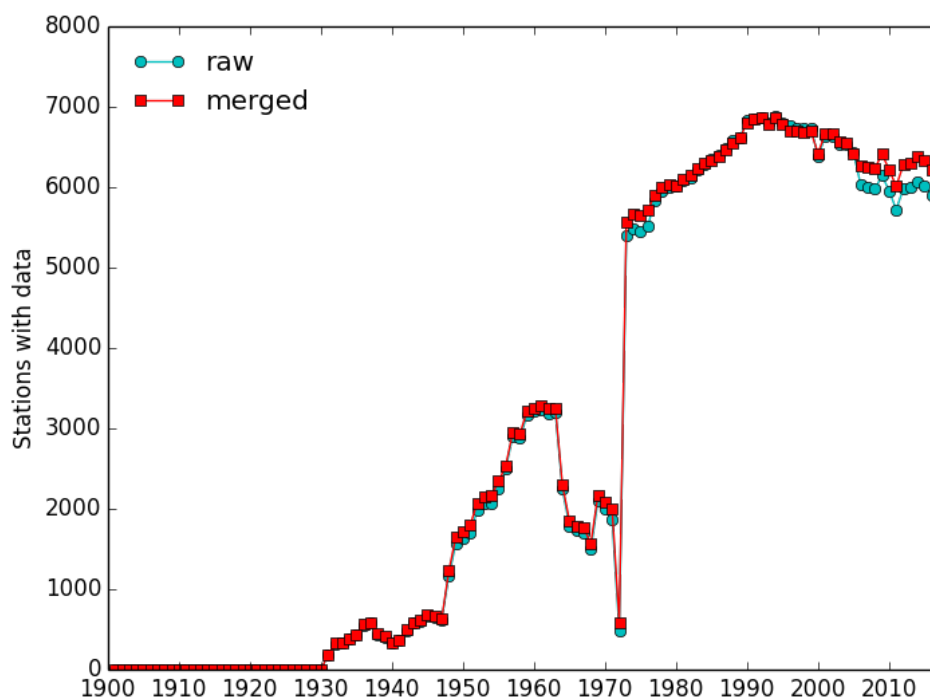
Figure 2. Time sequence of the number of stations before (cyan circles) and after (red squares) merging.

of the final set of candidate stations and mergers are available on the HadISD website at www.metoffice.gov.uk/hadobs/hadisd. At each annual update we will also make available lists of stations that are newly included into HadISD.2.0.x and those that are no longer included (for example, they have been merged or removed from the ISD). These additional lists should enable users to work with a dynamic station list more easily.

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 for which specific problems affect 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 (Andreas Becker, personal communication). Therefore, we have been able to explicitly merge the 09 stations into the 10 stations. We still perform the merging checks

Station distribution for 6 hourly reporting frequency
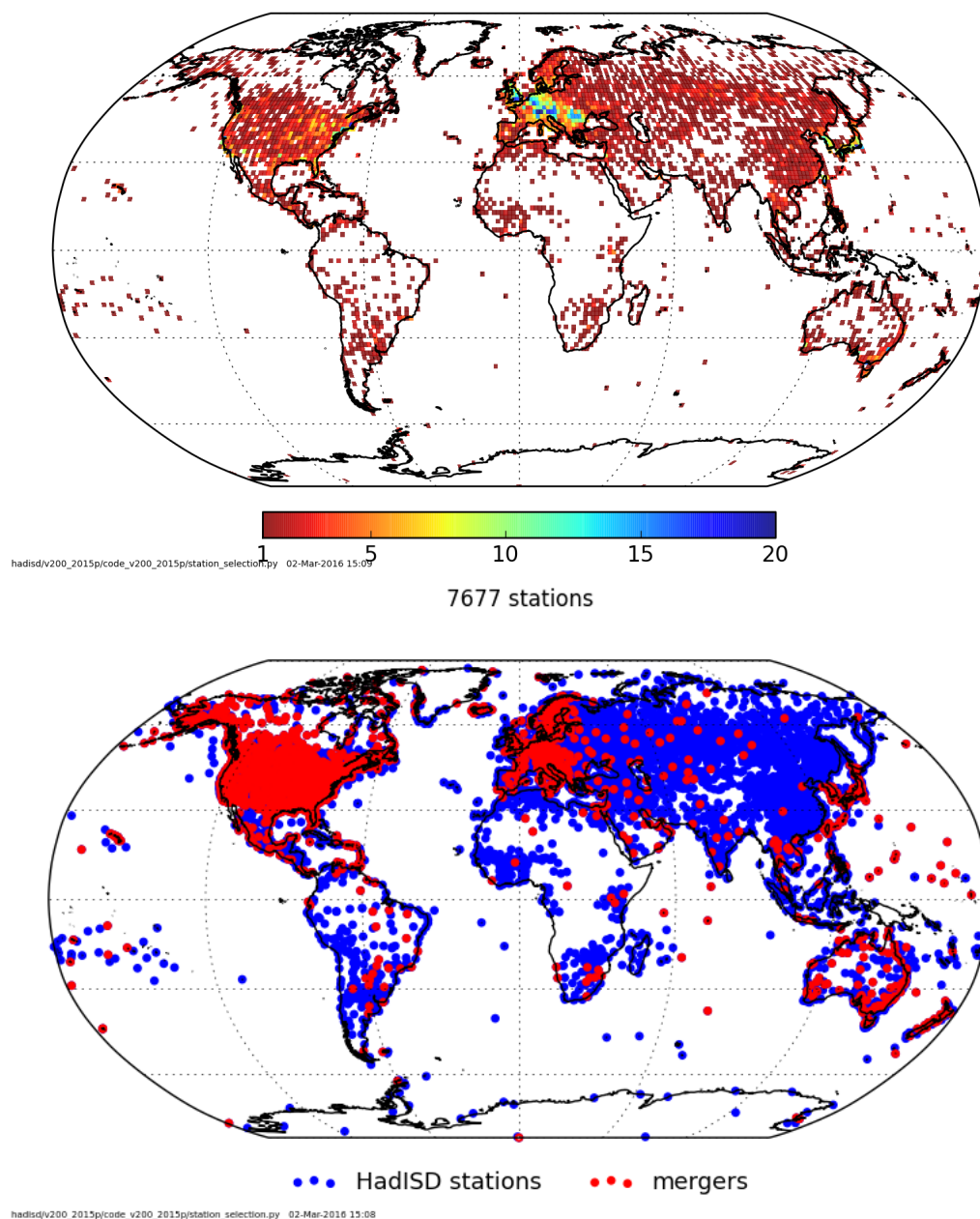
Figure 3. TOP The location of the final set of stations. For presentational purposes we show the number of stations within $1° \times 1°$ grid boxes. BOTTOM The locations of the 1993 stations which are composites (red) with all non-composites shown in blue. The composite stations are plotted on top of the non-composites and so dominate the area in locations with high station densities, e.g. North America and Europe.

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 some cases, this has resulted in apparent station moves in the ISD record. Using a list kindly supplied by Environment Canada (Lee Cudlip, personal communication) 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 that match 71xxx0-99999). 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 Overlap moves, Questionable Moves, Dates 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 resulted in the station not being selected for inclusion with HadISD.

Of the 934 Canadian stations we were able to assess, 798 were kept for processing by further selection criteria: in 32 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 pseudo-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.2.0.0 is written in Python (and run using version 2.7.6), and will be made available alongside the dataset at 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 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

The distributional gap check in HadISD.1.0.x had two parts. The first part worked on a monthly level, comparing the anomalies of the monthly median values. By standardising against the inter-quartile range (IQR), and comparing step-wise from the middle of the distribution outwards, asymmetries were identified and flagged if severe enough. For more details see Dunn et al. (2012).

The second part of the distributional gap test compares all observations within a calendar month (over all years). A histogram is created from all observations within a calendar month (e.g. all Januaries), and a gaussian distribution fitted. Threshold values are determined by using the positions where this fitted frequency falls below $y = 0.1$ and rounding outwards to the next integer-plus-one. Going outwards from the center, the distribution is scanned for gaps which occur beyond this threshold value, and any observations occurring beyond the gap are flagged.

In a number of cases it has come to light that a simple gaussian distribution 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 distribution by allowing for non-zero skew and kurtosis by using a Gauss-Hermite series[1]. The updated thresholds (as calculated when the fitted curve drops below $y = 0.1$) then occur closer to the bulk of the distribution then when using a plain gaussian curve.

---

[1]http://www.astro.rug.nl/software/kapteyn-beta/kmpfittutorial.html?highlight=gauss#fitting-gauss-hermite-series
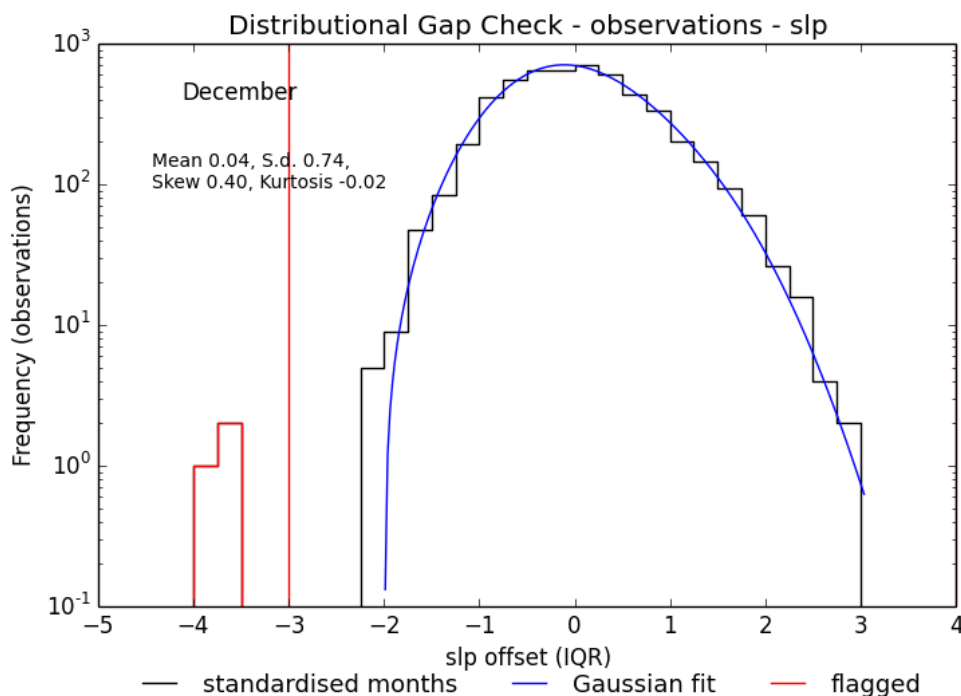
Figure 4. The improved distributional gap check working on SLP data from 764230-99999 Durango (24.06N, 104.60W, 1872m). Using a gaussian distribution without skew and kurtosis may have included cluster of observations at around -4IQR which, being beyond the threshold indicated by the vertical red line, are removed in this upgraded test.

In Figure 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 distribution with skew and kurtosis allows the small set of clearly erroneous observations with an IQR-offset of -4 to be flagged.

The additional checks for low sea-level pressure observations in this test are still used (see Dunn et al., 2012 Section 4.1.9).

3.2 Streaks

In HadISD.1.0.x this test searches for consecutive observation replication, replication at the same hour over a number of days or whole day replication for a run of days. These are dependent on the reporting resolution of the station (see Dunn et al., 2012 Table 4).

The updated version of this test allows these thresholds to be calculated dynamically. Firstly the length of each string of repeated values is obtained, and then distribution of these strings lengths is analysed. An inverse decay curve is fitted to this distribution (blue line in Figure 5), and a threshold is set when this curve falls below $y = 0.1$ (red line). This threshold
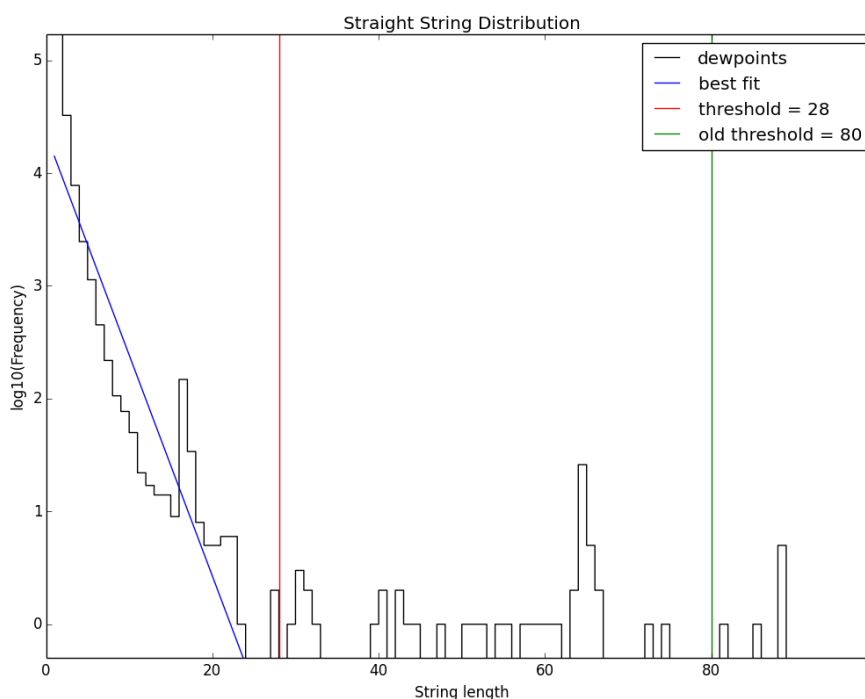
Figure 5. The dynamic threshold assignment from the improved streak check on dewpoint temperature data from 724750-99999 Milford Municipal Airport (38.4N, 113.0W, 1536m). Note the logarithmic y-axis, and hence the linear inverse decay curve in blue. The threshold used in HadISD.1.0.0 (green) retained a large number of streaks of repeated values which are now removed from this station when using the new threshold (red).

is modified by finding the next empty bin to ensure the entire main distribution is retained (see Figure 5). However, if this dynamically calculated 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. This ensures that this test can only result in more stringent removal of repeated streaks rather than fewer.

3.3 Spike

5 In HadISD.1.0.0 the spike check compared the first-difference values between observations to critical values to determined whether a spike had occurred or not. These critical values were determined from the IQR of the first-differences. Similarly to the updated repeated streak check, in HadISD.2.0.0 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
10 the first differences.

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

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

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).

3.4 Unusual Variance Check

This test identifies whole months where the within-month variance of normalised anomalies is sufficiently greater than the
5   median variance over the full station series for that month. It includes an algorithm to identify periods in the sea-level pressure which are likely to be the result of intense (usually tropical) storms (low pressure systems), and hence need to be retained. Locations which experience tropical storms usually have very uniform pressure values. Therefore, the extreme low pressure values occuring during the passage of a storm increases their monthly variance and so could result in erroneous flagging.

In HadISD.1.0.x the minimum pressure and the maximum wind speed within a calendar month were assessed for contempo-
10   raneity 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.

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
15   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.

By general convention, if the wind speed is 0 m s^{-1} (calm) then the direction should be recorded at $0°$. Conversely, a non-zero northerly wind speed should be recorded with a direction of $360°$. 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 to $0°$ when the speed has been recorded
20   as 0 m s^{-1}.

The logical checks used in HadISD.2.0.0 are based on those outlined in Table 2 of DeGaetano (1997) and are summarised in Table 1. This results in the removal of negative wind speeds, and also directions outside of the range $0 \to 360°$ inclusive. Observations with a non-zero wind speed but a direction of $0°$ and calm periods with a direction $\neq 0°$ are also removed.

To quality control the distribution of the wind speed and direction, we use the method outlined in Lucio-Eceiza et al. (2015)
25   to assess rotations between wind roses. Their work focuses 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 frequency curve falls below $y = 0.01$ as the 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 or changes in instrumentation.

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

In HadISD.1.0.x the closest 10 stations within 50 m elevation and 300 km distance were selected as neighbours, and where possible, these would be evenly distributed around four quadrants (northeast, southeast, southwest and northwest).

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, correlation between the neighbour and target is obtained after removing the annual and diurnal cycles. These cycles are removed by first calculating the daily long-term mean, and subtracting that from the data. Then the relative means for each of the 24 hours 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. For the 356 stations where fewer than three neighbours were found, then this test was not run.

Using these updated neighbouring stations, the remainder of the test is very similar as for HadISD.1.0.x. The station-neighbour difference series are calculated. However, the inter-quartile range of the difference series is calculated for each calendar month separately, rather than for the entire record. 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. Any observation associated with a difference exceeding 5IQR of the whole difference series is flagged.

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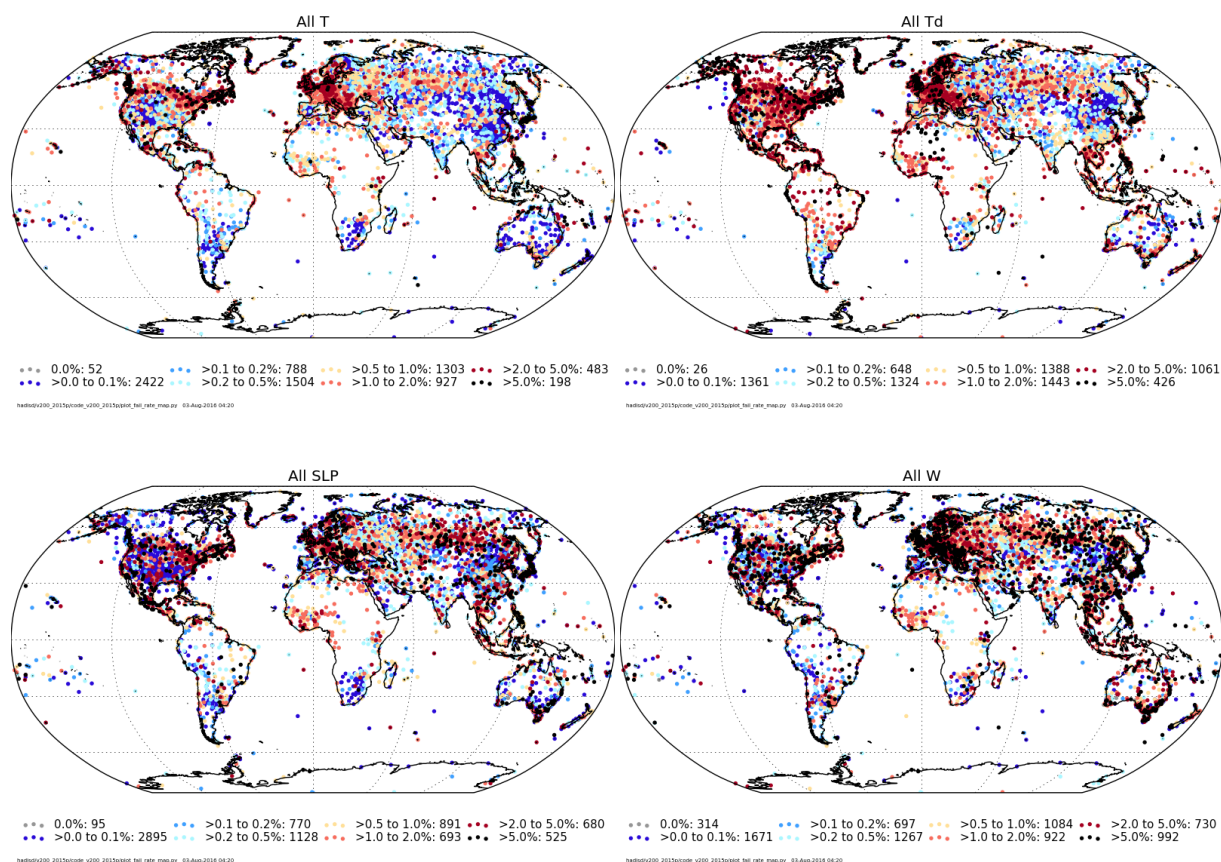
Figure 6. Rejection rates by variable for each station showing the temperature, dewpoint temperature, sea-level pressure and wind speed. Different rejection rates are shown by different colours, and the legend also shows the number of stations in each category. The stations with a greater proportion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

4   Overview of HadISD.2.0.0.2015p and comparison to HadISD.1.0.4.2015p

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 3. As in HadISD.1.0.x, the majority of stations have very low flagging rates, with less than 1 per cent of observations removed. There are some regional and country-scale patterns that emerge in the flagging

5   rates. For temperature the large regions which have the highest proportion of flagged observations are 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 Philippines. The wind observations show a relatively high proportion of flags compared to the other variables, with relatively many stations having more than 5 per cent of observations removed.

Comparing Fig. 6 and 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

5   by a given test (Tables 3 and 8) 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 Section 3.2. Similarly fewer stations have large numbers of spikes identified (Section 3.3) and fewer observations are removed as a result of the gap check (Section 3.1). The correction of the unusual variance check (Table 7) has increased the fractions of stations with observations removed by this

10   check. There are also increased removals by the neighbour check (Section 3.6) and also the clean-up.

In HadISD.2.0.0.2015p 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-maximum values (wind speeds) calculated from the sub-daily data. The information about the change point locations and magnitudes will be made available

15   along with the dataset, and updated annually. Examples of the distribution of inhomogeneity sizes and their distribution in time are shown in Fig. 7, with further details of the numbers of breaks found in Table 2. These histograms show the distribution of inhomogeneity sizes in black, along with a best fit Gaussian curve in red. Under the assumption that a Gaussian curve is appropriate for the size of inhomogeneities found in HadISD, then the numerous small inhomogeneities which are not able to be found using this automated method are shown in cyan. The distribution of inhomogeneities are very similar to those found

20   for HadISD.1.0.2.2013f in Dunn et al. (2014). The bottom panel if Fig. 7 shows the number of change points present in each year. Despite the fewer stations present before 1973, change points are still found with this smaller network of stations.

We also show in Fig. 8 the number of break points detected in the temperature record for each station as well as the distribution of stations with record length and number of breaks (equivalent plots for dewpoint, sea-level pressure and wind speeds are show in the Appendix, Figs 10 and 11). Not only the length of record and quality of the station data, but also the

25   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/

30   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

	Diagnostic	Temperature	Dewpoint	SLP	Wind Speeds
		Number of Stations			
input station number		7677	7677	7677	7677
Not Tested		132	259	1127	870
Tested		7545	7418	6550	6807
No Change Points		1073	581	2717	491
With Change Points		6472	6837	3833	6316
		Number of Change Points Detected			
	mean	10828	16103	9025	19012
	DR	12741	14490	-	-
	maximum	-	-	-	18437
Total Combined Change points		20883	26710	8897	27982
Change Points/Station		3.24	3.93	2.31	4.42
Mean Adjustment Magnitude*	mean	0.733	0.968	0.770	0.567
	DR	0.858	0.725	-	-
	maximum	-	-	-	0.871

Table 2. Update of Table 1 in Dunn et al. (2014). The number of stations used and the number which had too few neighbouring stations for each of the four variables. The number of change points detected for the calculation methods used for each variable, along with the number per station (excluding those with too few neighbours). * Although no adjustments were made, the values were still extracted. The Diagnostic column shows values for homogenisation assessments using the daily mean, the diurnal range (DR) and the daily maximum values (wind speed only).

the fields available in each NetCDF file are given in Table 4. Of note is that the 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 v2.0.1.2016p in January 2017) with a final version (e.g. v2.0.1.2016f) a few months later to ensure that late-arriving data are included.

Updates to the dataset will be made public on the Twitter handle @MetOfficeHadOBS and also at the blog http://hadisd. blogspot.co.uk/. We encourage users of this dataset to contact the authors if they find any issues, for example observations which they believe have been erroneously flagged.

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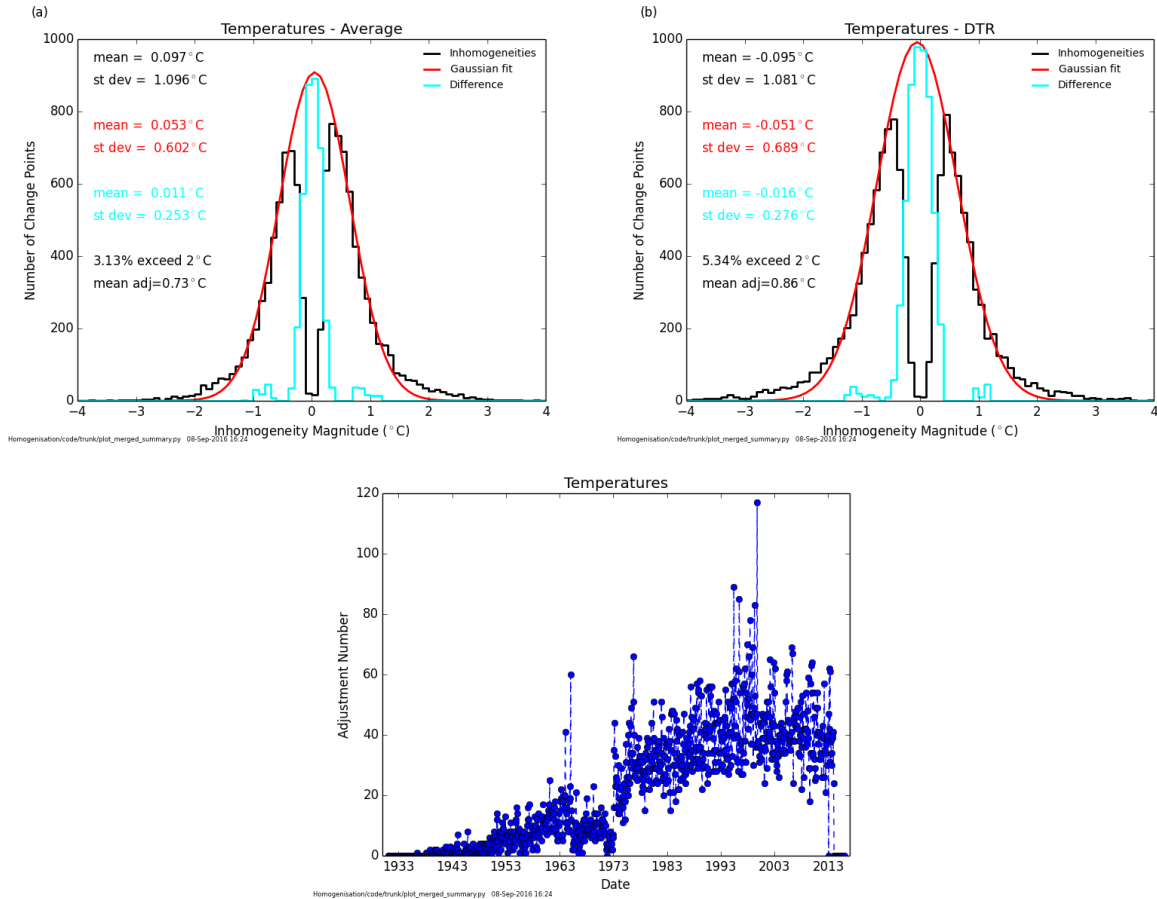
Figure 7. The distribution of inhomogeneities using the monthly-mean TOP temperatures and MIDDLE diurnal temperature range. BOTTOM The number of change points found in each year from both the calculation methods (see Dunn et al. 2014 for full details).

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 5. Firstly the sub-daily sea-level pressure values provided in HadISD are converted to
5   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. Also, this means that if there are no pressure values in the HadISD, then no humidity or heat stress variables have been calculated.

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
10   repeated using the formulae with respect to ice. The resulting vapour pressure values are used to obtain the specific and relative humidities.
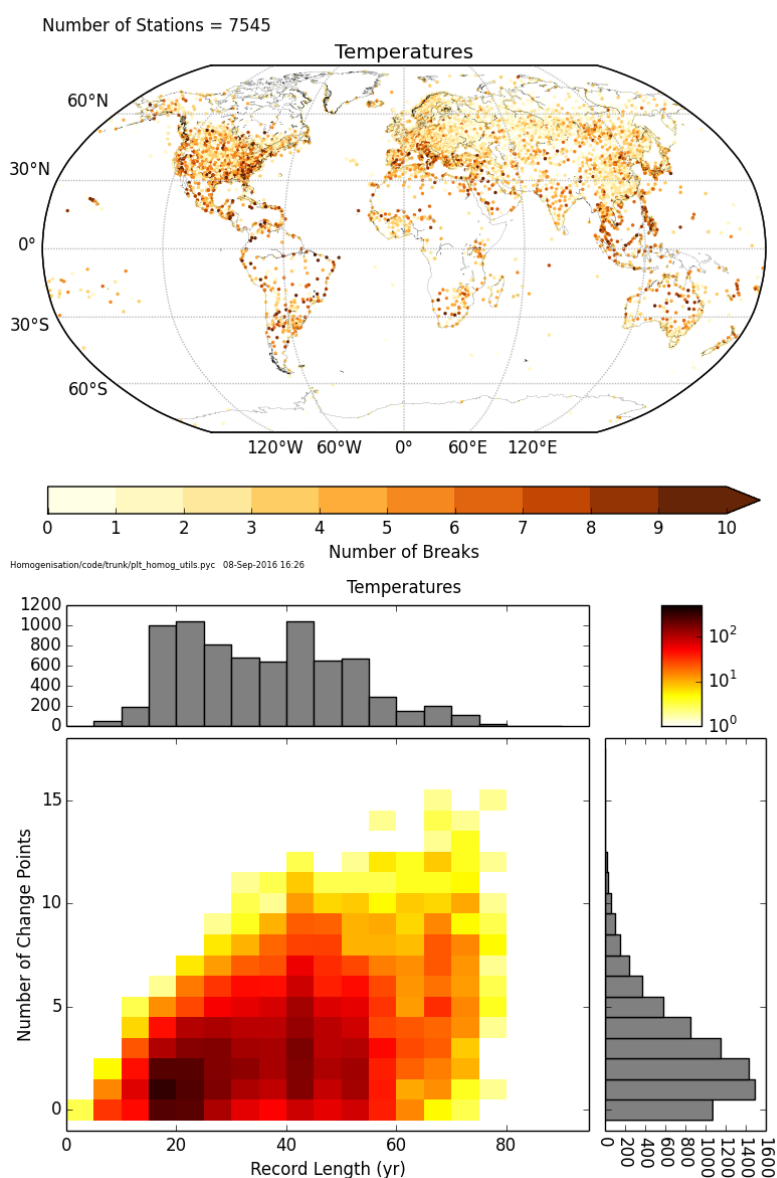
Figure 8. TOP The number of change points detected for each station. BOTTOM The distribution of stations with numbers of change points and record length. Histograms show the distribution of stations with record length (above) and with change point number (right), which are the projections onto the x and y axes of the main panel respectively. The colour bar (top right) is on a logarithmic scale.

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	7665	0	0	0	0	1	0	11
Odd cluster check	T	2851	4404	282	117	13	2	0	8
	Td	2723	4481	271	131	15	1	4	51
	SLP	2025	3884	579	376	111	48	51	603
	ws	1829	4336	781	611	117	1	0	2
Frequent values check	T	7543	101	9	11	5	3	4	1
	Td	7493	109	11	22	14	10	12	6
	SLP	7506	37	15	20	12	10	14	63
Diurnal cycle check	All	7153	13	88	201	92	49	35	46
Distributional gap check	T	1960	5095	200	191	104	66	42	19
	Td	976	5681	434	315	131	79	43	18
	SLP	2816	3614	385	375	181	111	94	101
Known records check	T	7602	75	0	0	0	0	0	0
	Td	7677	0	0	0	0	0	0	0
	SLP	6511	1046	23	25	18	29	20	5
	ws	7677	0	0	0	0	0	0	0
Repeated streaks/unusual spell frequency check	T	4238	1947	318	394	270	293	195	22
	Td	3722	1857	302	534	425	459	336	42
	SLP	6941	569	68	40	21	15	12	11
	ws	5161	1081	361	393	307	218	123	33
Climatological outliers check	T	1162	5789	400	188	80	29	25	4
	Td	741	6016	476	3283	101	36	20	4
Spike check	T	2414	5138	78	34	7	4	2	0
	Td	894	6662	79	33	6	1	2	0
	SLP	2582	5019	38	27	4	4	3	0
T and Td cross-check: Super-saturation	T, Td	7677	0	0	0	0	0	0	0
T and Td cross-check: Wet bulb drying	Td	4100	2593	331	336	150	87	57	23
T and Td cross-check: Wet bulb cutoffs	Td	5461	409	424	594	324	243	157	65

Table 3. Summary of removal of data from individual stations by the different tests for the 7677 stations considered in detailed analysis.

Test	Variable	Number of stations in each detection rate band (as % of total original observations removed)							
Cloud Clean-up	c	364	673	405	772	926	1576	1995	966
Unusual variance check	T	5674	75	522	924	314	114	39	15
	Td	5554	51	507	978	356	166	60	5
	SLP	6471	23	288	482	280	93	32	8
	ws	5162	225	623	951	401	221	81	13
Logical Wind	wd	4546	2026	294	390	230	132	45	14
Wind Rose	ws	4234	1678	115	176	182	254	599	439
Station clean up	T	1549	2599	883	1505	737	247	78	79
	Td	1206	1823	893	1666	1020	586	288	195
	SLP	1668	2212	547	772	567	364	515	1032
	ws	1603	2734	609	890	525	331	565	420
Nearest neighbour data check	T	1619	5821	74	69	53	21	14	6
	Td	1401	5957	132	104	45	23	13	2
	SLP	2600	4683	214	97	34	22	16	11

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 6. 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.

The cleaned HadISD.2.0.0 data are used to derive these variables. However, neither of these two sets of variables have undergone further quality control or homogenisation processes. Therefore they will inherit any remaining data issues present within the input variables drawn from HadISD. However the homogeneity information from the temperatures and 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.2.0.0.2015p contains 7677 stations of which 1993 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 files and supplementary material will be made available at www.metoffice.gov.uk/hadobs/hadisd.

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	-	-

Table 4. 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.

Acknowledgements

We thank Andreas Becker (DWD) and Lee Cudlip (EC) for their help and suggestions for improving the records in Germany and Canada respectively. We also thank Blair Trewin and an anonymous reviewer for their helpful comments during our submission to Climate of the Past along with the editor, Marie-France Loutre.

5   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 BEIS/Defra Met Office Hadley Centre Climate Programme (GA01101).

Variable	Equation	Source	Notes
Specific humidity (q) in g kg^{-1}	$q = 1000 \left(\dfrac{0.622e}{P_{\mathrm{mst}} - ((1 - 0.622)e)} \right)$	Peixoto and Oort (1996)	
Relative humidity (RH) in %rh	$RH = 100 \left(\dfrac{e}{e_{\mathrm{s}}} \right)$		
Vapour Pressure (e) with respect to water in HPa (when $T_{\mathrm{w}} > 0^{\deg}$)	$e = 6.1121.f_{\mathrm{w}}.\exp \left(\dfrac{18.729 - \left(\dfrac{T_{\mathrm{d}}}{227.3} \right) T_{\mathrm{d}}}{257.87 + T_{\mathrm{d}}} \right)$ $f_{\mathrm{w}} = 1 + 7 \times 10^{-4} + 3.46 \times 10^{-6} P_{\mathrm{mst}}$	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 $T_{\mathrm{w}} \leq 0^{\deg}$ C)	$e = 6.1115.f_{\mathrm{w}}.\exp \left(\dfrac{23.036 - \left(\dfrac{T_{\mathrm{d}}}{333.7} \right) T_{\mathrm{d}}}{279.82 + T_{\mathrm{d}}} \right)$ $f_{\mathrm{w}} = 1 + 3 \times 10^{-4} + 4.18 \times 10^{-6} P_{\mathrm{mst}}$	Buck (1981)	
Wet bulb temperature (T_{w}) in $^{\deg}$ C	$T_{\mathrm{w}} = \dfrac{aT + bT_{\mathrm{d}}}{a + b}$ $a = 6.6 \times 10^{-5} P_{\mathrm{mst}}$ $b = \dfrac{409.8e}{(T_{\mathrm{d}} + 237.3)^2}$	Jensen et al. (1990)	
Station Pressure in hPa	$P_{\mathrm{mst}} = P_{\mathrm{msl}} \left(\dfrac{T}{T + 0.0065Z} \right)^{5.625}$	List (1963)	Temperature T, station height Z in metres

Table 5. Humidity formulae used in HadISD v2.0.0. as in HadISDH v2.0.0 Willett et al. (2014).

References

ACSM: Prevention of thermal injuries during distance running, Med. Sci. Sports Exerc., 16, iv–xiv, 1984.

Buck, A. L.: New equations for computing vapor pressure and enhancement factor, Journal of applied meteorology, 20, 1527–1532, 1981.

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., et al.: The twentieth century reanalysis project, Quarterly Journal of the Royal Meteorological Society, 137, 1–28, 2011.

DeGaetano, A. T.: A quality-control routine for hourly wind observations, Journal of Atmospheric and Oceanic Technology, 14, 308–317, 1997.

Variable	Equation	Source	Notes
Temperature-Humidity Index (THI)	$THI = (1.8T + 32) - (0.55 - 0.0055RH)(1.8T - 26))$	Dikmen and Hansen (2009)	
Pseudo Wet-bulb Globe Temperature (WBGT)	$WBGT = (0.567T) + (0.393e_v) + 3.94$	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	$HI = -42.379$ $+2.04901523T_f + 10.14333127RH$ $-0.22475541T_f RH - 0.006837837T_f^2$ $-0.05481717RH^2 + 0.001228747T_f^2 RH$ $+8.5282 \times 10^{-4} T_f RH^2$ $-1.99 \times 10^{-6} T_f^2 RH^2$ $\text{adj}_1 = \frac{13RH}{4}\sqrt{\frac{17\text{abs}(T_f - 95)}{17}}$ $\text{adj}_2 = \frac{RH - 85}{10} \cdot \frac{87 - T_f}{5}$ $HI = 0.5(T_f + 61 + 1.2(T_f - 68) + 0.094RH)$	Rothfusz (1990)	Where T_f is the temperature in Fahrenheit. If $RH < 13$ and $80 \leq T_f \leq 112$, adj_1 is subtracted from HI; if $RH > 85$ and $80 \leq T_f \leq 87$ adj_2 is added to HI. Furthermore, if these calculations would result in a $HI < 80$, then the simpler formula is used.

Table 6. Heat stress measures calculated in HadISD v2.0.0.

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

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.

5   Dunn, R., Willett, K., Morice, C., and Parker, D.: Pairwise homogeneity assessment of HadISD, Climate of the Past, 10, 1501–1522, 2014.

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., et al.: World Meteorological organization assessMent of the PurPorted World record 58 c teMPerature extreMe at el azizia, libya (13 sePteMber 1922), Bulletin of the American Meteorological Society, 94, 199–204, 2013.

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

10   Jensen, M. E., Burman, R. D., and Allen, R. G.: Evapotranspiration and irrigation water requirements, ASCE, 1990.

List, R. J.: Smithsonian Meteorological Tables, Quarterly Journal of the Royal Meteorological Society, 114, 1963.

Lott, J. N.: 7.8 THE QUALITY CONTROL OF THE INTEGRATED SURFACE HOURLY DATABASE, 2004.

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, Journal of Atmospheric and Oceanic Technology, p. submitted, 2015.

5   Masterton, J. and Richardson, F.: Humidex: a method of quantifying human discomfort due to excessive heat and humidity, Downsview, Ont.: Atmospheric Environment, 1979.

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

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

10   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.

Rothfusz, L. P.: The heat index equation (or, more than you ever wanted to know about heat index), Fort Worth, Texas: National Oceanic and Atmospheric Administration, National Weather Service, Office of Meteorology, pp. 90–23, 1990.

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

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

Willett, K., Dunn, R., Thorne, P., Bell, S., de Podesta, M., Parker, D., Jones, P., and Williams Jr, C.: HadISDH land surface multi-variable humidity and temperature record for climate monitoring, Climate of the Past, 10, 1983–2006, 2014.

20   **Appendix**

Here we detail the changes in the quality control tests that have occurred on conversion to Python (Table 7), as well as a version of Table 3 but showing the percentage of stations rather than the actual numbers (Table 8).

We also show in Fig. 9 the distributions of stations across the globe in six example years, as well as the number of break points detected and the distribution with record length for dewpoint, sea-level pressure and wind speeds (Figs. 10 and 11

25   respectively)
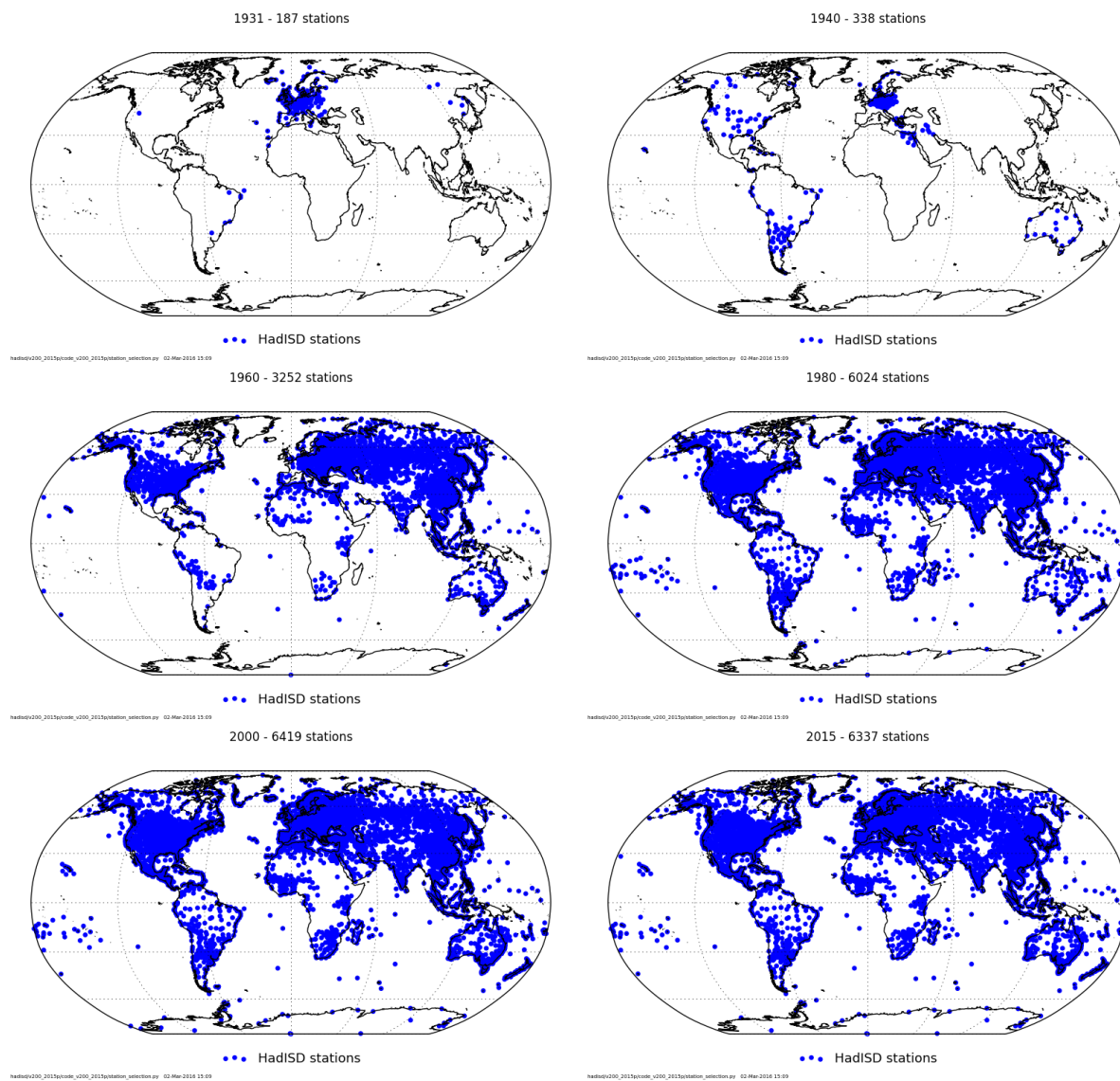
Figure 9. The distribution of stations across the globe in six years (1931, 1940, 1960, 1980, 2000 and 2015).
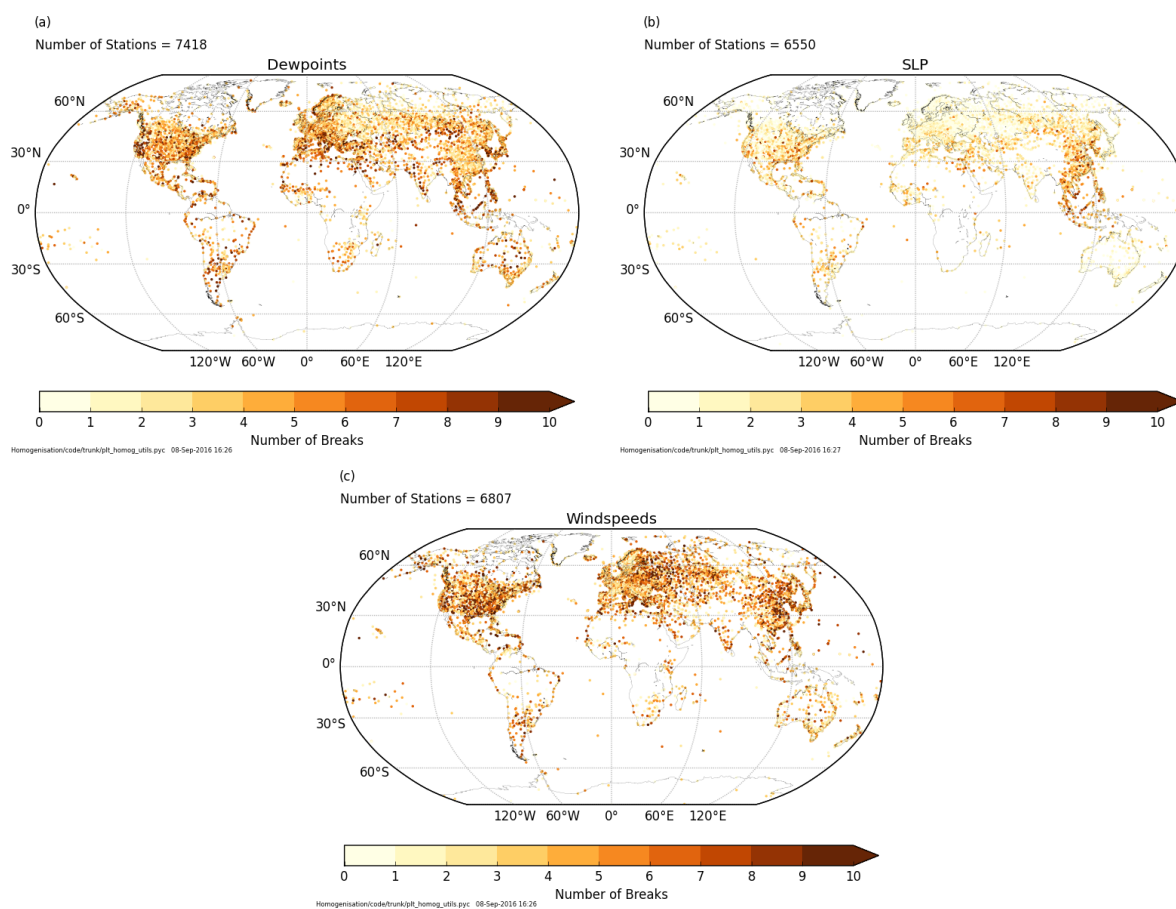
Figure 10. The number of change points detected for each station for dewpoints, SLP and wind speed.

Table 7. Summary of Changes in Tests

Test	T	Td	SLP	ws	wd	clouds	Changes and Notes
				Applies to			**Changes and Notes**
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 a gaussian distribution 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 distribution 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. Test limit now 4 median absolute deviations
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		

Table 8. As Table 3 but in per cent. As a result of rounding, rows may not sum to exactly 100.0 per cent.

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	0.1
Odd cluster check	T	37.1	57.4	3.7	1.5	0.2	0.0	0.0	0.1
	Td	35.5	58.4	3.5	1.7	0.2	0.0	0.1	0.7
	SLP	26.4	50.6	7.5	4.9	1.4	0.6	0.7	7.9
	ws	23.8	56.5	10.2	8.0	1.5	0.0	0.0	0.0
Frequent values check	T	98.3	1.3	0.1	0.1	0.1	0.0	0.1	0.0
	Td	97.6	1.4	0.1	0.3	0.2	0.1	0.2	0.1
	SLP	97.8	0.5	0.2	0.3	0.2	0.1	0.2	0.8
Diurnal cycle check	All	93.2	0.2	1.1	2.6	1.2	0.6	0.5	0.6
Distributional gap check	T	25.5	66.4	2.6	2.5	1.4	0.9	0.5	0.2
	Td	12.7	74.0	5.7	4.1	1.7	1.0	0.6	0.2
	SLP	36.7	47.1	5.0	4.9	2.4	1.4	1.2	1.3
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	84.8	13.6	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	55.2	25.4	4.1	5.1	3.5	3.8	2.5	0.3
	Td	48.5	24.2	3.9	7.0	5.5	6.0	4.4	0.5
	SLP	90.4	7.4	0.9	0.5	0.3	0.2	0.2	0.1
	ws	67.2	14.1	4.7	5.1	4.0	2.8	1.6	0.4
Climatological outliers check	T	15.1	75.4	5.2	2.4	1.0	0.4	0.3	0.1
	Td	9.7	78.4	6.2	3.7	1.3	0.5	0.3	0.1
Spike check	T	31.4	66.9	1.0	0.4	0.1	0.1	0.0	0.0
	Td	11.6	86.8	1.0	0.4	0.1	0.0	0.0	0.0
	SLP	33.6	65.4	0.5	0.4	0.1	0.1	0.0	0.0
T and Td cross-check: Super-saturation	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	53.4	33.8	4.3	4.4	2.0	1.1	0.7	0.3
T and Td cross-check: Wet bulb cutoffs	Td	71.1	5.3	5.5	7.7	4.2	3.2	2.0	0.9
Cloud Clean-up	c	4.7	8.8	5.3	10.1	12.1	20.5	26.0	12.6

Test	Variable	Stations with detection rate band (% of total original observations)							
(Number)		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
Unusual variance check	T	73.9	1.0	6.8	12.0	4.1	1.5	0.5	0.2
	Td	72.3	0.7	6.6	12.7	4.6	2.2	0.8	0.1
	SLP	84.3	0.3	3.8	6.3	3.6	1.2	0.4	0.1
	ws	67.2	2.9	8.1	12.4	5.2	2.9	1.1	0.2
Logical Wind	wd	59.2	26.4	3.8	5.1	3.0	1.7	0.6	0.2
Wind Rose	ws	55.2	21.9	1.5	2.3	2.4	3.3	7.8	5.7
Station clean up	T	20.2	33.9	11.5	19.6	9.6	3.2	1.0	1.0
	Td	15.7	23.7	11.6	21.7	13.3	7.6	3.8	2.5
	SLP	21.7	28.8	7.1	10.1	7.4	4.7	6.7	13.4
	ws	20.9	35.6	7.9	11.6	6.8	4.3	7.4	5.5
Nearest neighbour data check	T	21.1	75.8	1.0	0.9	0.7	0.3	0.2	0.1
	Td	18.2	77.6	1.7	1.4	0.6	0.3	0.2	0.0
	SLP	33.9	61.0	2.8	1.3	0.4	0.3	0.2	0.1
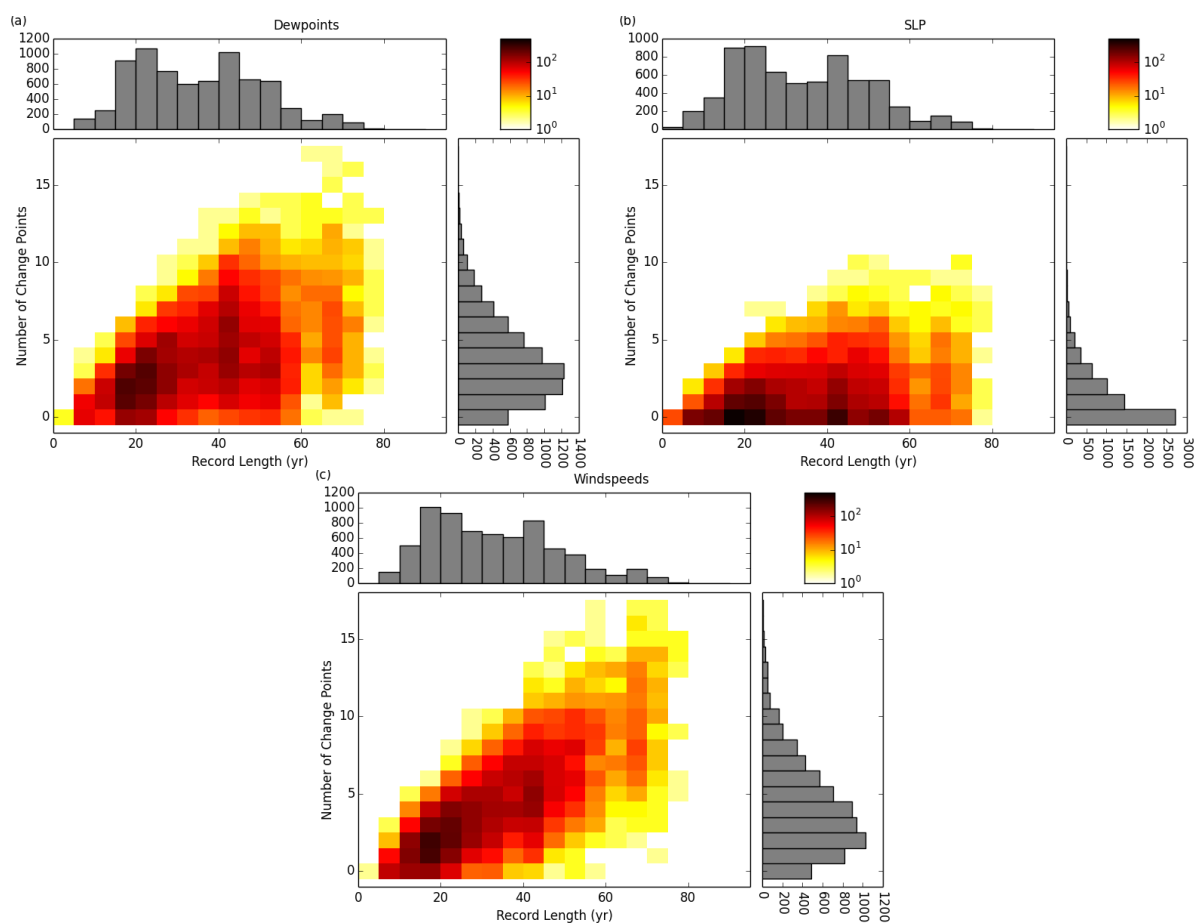
Figure 11. The distribution of stations with numbers of change points and record length for dewpoints, SLP and wind speed. Histograms show the distribution of stations with record length (above) and with change point number (right), which are the projections onto the x and y axes of the main panel respectively. The colour bar (top right) is on a logarithmic scale