HadISD.3.1.1: Product User Guide

Robert Dunn
robert.dunn@metoffice.gov.uk

December 8, 2020

Citation: Dunn, R. J. H., Willett, K. M., Parker, D. E., and Mitchell, L. (2016) Expanding HadISD: quality-controlled, sub-daily station data from 1931, Geosci. Instrum. Method. Data Syst., 5, 473491, https://doi.org/10.5194/gi-5-473-2016
Dunn, R. J. H. (2019) HadISD version 3, Hadley Centre Tech Note #103

[Alternatively, use the citation from the CEDA Archive: Met Office Hadley Centre; National Centers for Environmental Information - NOAA (2020): HadISD: Global sub-daily, surface meteorological station data, 1931-2019, v3.1.0.2019f. Centre for Environmental Data Analysis, date of citation.]

EMERGENCY ONE PAGE QUICK START GUIDE FOR HadISD.3.1.1

HadISD (HadISD.3.1.1) is a dataset of sub-daily *in-situ* observations for a number of meteorological variables.

What products are available?

All the HadISD products are available linked from the HadISD homepage at http://www.metoffice.gov.uk/hadobs/hadi

There are files for each station containing the observed variables and derived variables for humidity and heat stress. We also provide collections of stations, grouped by WMO number.

How do I obtain the data?

Data are available from http://www.metoffice.gov.uk/hadobs/hadisd/. You can download each station individually if you know the WMO numbers you require (a list of station IDs, locational metadata as well as station names is available). However, if you want a large number of stations, then we recommend downloading a `.tar.gz` archive of collections of stations (grouped by WMO number).

How do I read the HadISD.3.1.1 data?

The station data are stored in NetCDF format files. NetCDF files are a platform-independent, self-describing binary format and there are a number of common tools (Section 3.5) that can be used to access the data. Some basic python code is provided in Section 4 to show worked examples of reading the data and performing some simple calculations and processing.

What tools are available for these products?

Some basic Python code is provided in Section 4 to show worked examples of reading in the data and performing some simple calculations and processing.

How to cite the data set

Dunn, R. J. H., Willett, K. M., Parker, D. E., and Mitchell, L. (2016) Expanding HadISD: quality-controlled, sub-daily station data from 1931, Geosci. Instrum. Method. Data Syst., 5, 473491, https://doi.org/10.5194/gi-5-473-2016

Dunn, R. J. H. (2019) HadISD version 3, Hadley Centre Tech Note #103

Further citation information for e.g. the quality control tests [6], the homogeneity assessment [5] or the underlying Integrated Surface Dataset from NOAA (ISD, [16]) may also be necessary.

[Alternatively, use the citation from the CEDA Archive: Met Office Hadley Centre; National Centers for Environmental Information - NOAA (2020): HadISD: Global sub-daily, surface meteorological station data, 1931-2019, v3.1.0.2019f. Centre for Environmental Data Analysis, date of citation.]

Further information and contact

For further help please read the rest of the document. The papers describing the data set are the best place to find the technical details. Updated diagnostics are available from http://www.metoffice.gov.uk/hadobs/hadisd/. For further enquiries contact `robert.dunn@metoffice.gov.uk`. Data set updates will be tweeted from @metofficeHadOBS.

Contents

1 What is HadISD.3.1.1?

In a single sentence, HadISD.3.1.1 is a sub-daily, station-based, multivariate dataset of about 8500 stations spread around the globe. This means that each the data from all the stations are available as individual time-series files, rather than having been blended together as a space-filling representation (e.g. gridded) of the data. The data has a temporal resolution of hourly to 6-hourly (24 to 4 observations per day), and there are a number of meteorological variables present all within the same file.

2 Getting started with HadISD.3.1.1

2.1 How do I get the data?

You can download each station individually if you know the WMO numbers you require (a list of station IDs, locational metadata as well as station names is available). However, if you want a large number of stations, then we recommend downloading a `.tar.gz` archive of collections of stations (grouped by WMO number).

Collections of all these files by station ID (WMO number) are also available to make downloading easier.

On a Linux operating system, use `tar -xzf tarfile.tar.gz` to extract the archive. Under Windows, use whatever archive extraction software you have available, e.g. WinZip. This will place the station files in the local directory.

2.1.1 Primary station data

The primary product of integrated netCDF files are structured as one file per station. These files contain the primary climate variables (temperature; dew point temperature; sea level pressure; station level pressure; wind speed and direction; low, mid and high level cloud cover. There are also fields for the cloud base; past weather, the source station ID and also both the flag information and flagged values.

These data are available at http://www.metoffice.gov.uk/hadobs/hadisd/.

2.1.2 Derived station data (humidity & heat stress measures)

A supplementary product of pre-calculated humidity and heat-stress measures are available as netCDF files, again as one file per station. The humidity files contain again temperature, dew point temperature and sea level pressure as these were used to calculate the following humidity quantities: vapor pressure, saturation vapor pressure, wet bulb temperature, specific and relative humidity.

The heat stress measures are the temperature-humidity index, a pseudo wet bulb globe temperature, humidex, the apparent temperature and the heat index. Again, the temperature, dew point temperature and the wind speeds are provided as they were used in the calculations.

These data files are also available at http://www.metoffice.gov.uk/hadobs/hadisd/.

2.2 How do I use the data?

The station data are stored in NetCDF format files. NetCDF files are a platform-independent, self-describing binary format and there are a number of common tools (Section 3.5) that can be used to access the data. They are commonly used across climate science. Furthermore, the data files are CF compliant meaning that the metadata in the files is in a standardised format. The structure of the data files is described in more detail in Section 3. Some basic python code is provided in Section 4 to show worked examples of reading the data and performing some simple calculations and processing.

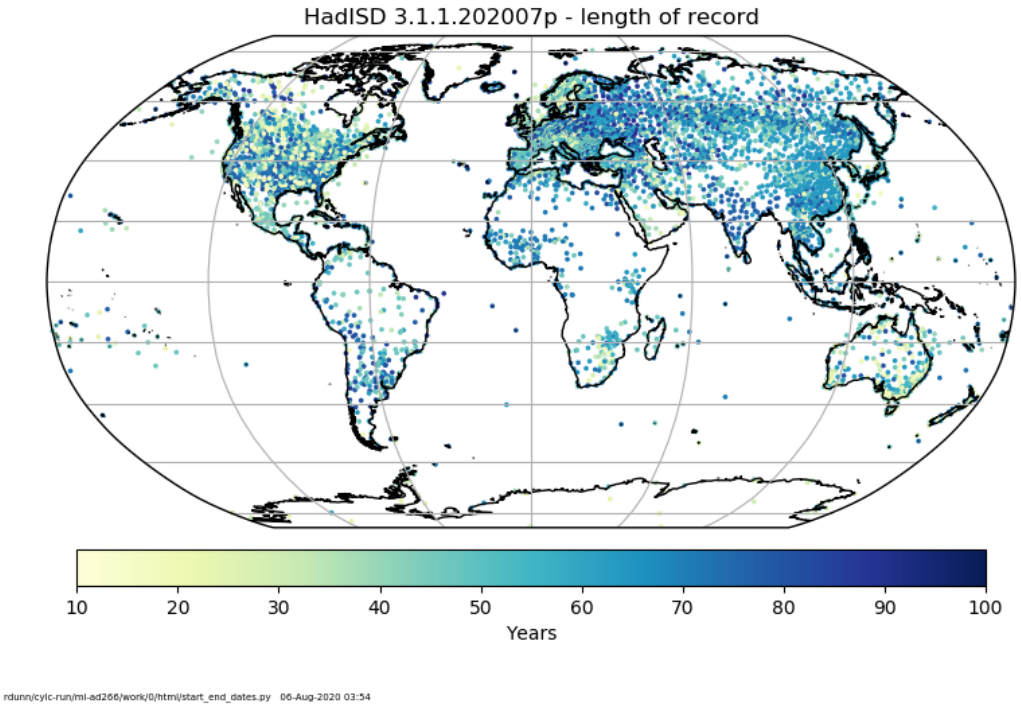

Figure 1: The distribution of the HadISD stations (v3.1.1.202007p) and their length of record determined from the first and last observation available for each.

Within each file for each station there are some metadata about the dataset as a whole, about the properties (location etc.) of the station, and then a number of fields (one per meteorological variable). The time axis has been compressed, so only hours where there are reported (either valid or flagged) observations for at least one variable are stored.

2.2.1 Do's and Don't's of using the data

Do - send us feedback when you use the data to robert.dunn@metoffice.gov.uk.

Don't - assume that the timeseries are homogeneous. Homogeneity information is available, but the timeseries have not been adjusted.

Do - check the data policy of HadISD (available here).

2.3 Basic data statistics

The distribution of the stations over the world is shown in Figure 1. This shows that although there is good coverage in North America and Eurasia, there is poorer coverage in South America, Africa and central parts of Australia.

Figure 2 also shows the length of record available for each station (determined from the start and end of the record only). Note that a station may have one or more extended periods where reporting ceases within the start and end dates.

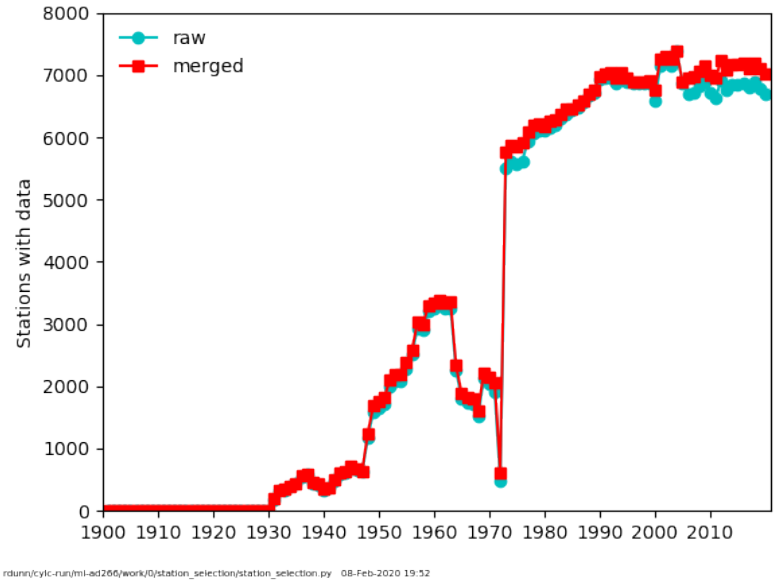

Figure 2: The distribution of the HadISD stations (v3.1.1.202001p) over time. The two lines show the raw (using stations as they are in the ISD) and merged (where we look for locational matches to augment records). This process is done once per year when reselecting the stations in HadISD.

2.4 Contact us

If you have further questions about the HadISD.3.1.1 data set, please contact us at robert.dunn@metoffice.gov.uk.

2.5 FAQ

2.5.1 Is HadISD.3.1.1 the data set for me?

Yes - if you are looking for station time series data

Yes - if you are looking for sub-daily observations

No - if you are looking for a space-filling representation of past weather and climate

Maybe - if you are looking for homogenised information

Alternatives to the HadISD.3.1.1 are the ISD at NOAA or a number of reanalysis products (though these are gridded, rather than station based).

2.5.2 How does the versioning schema work?

The dataset numbering is of the form HadISD.x.y.z.YYYYMMi, and is outlined in more detail in [6, 4]. The major version indicator is `X` and indicates a major change and would be accompanied by a peer-reviewed paper. `Y` is a more minor change, e.g. in one of the QC tests and would be described in a tech-note. Finally, `Z` is a small change, for example addition or changes to data in the past. The last complete year and month of the dataset is given by YYYYMM, and the final character shows if the dataset is f-final or p-preliminary. Therefore HadISD.3.1.0.2019f would be the final version of the dataset

containing data up to the end of 2019. Also there was a software change between versions 3.0.0 and 3.1.0. Therefore, HadISD.3.1.1.202007p is the preliminary version for 2020, and contains data up to the end of July.

2.5.3 How is the flag information structured?

In the netCDF files, there is a field which contains all the flags set by each of the tests. This 71-column array has columns corresponding to each of the tests run in the creation of HadISD as shown in Table 1[1].

Plots showing the summaries of these (percentages of each variable flagged by each test) are available on the website (see Online Material. There are also plots showing the combined rates for all tests and all tests looking for repeated strings.

Each version also has a file summarising the number and fraction of stations with certain flagging rates for each of these tests and the two sets of summaries (strings and total) [all_fails_summary_VERSION.dat].

The 71-column array stored in the qc_flags field has the values as shown in Table 2.

Values which have failed one or more tests are removed from the main data fields, however they are retained in a new field (flagged_obs) in case a user wishes to use these for some reason. These are stored in a 19 column array, where the columns are ordered as shown in Table 3.

2.5.4 How are flags indicated in the files?

The netCDF structure automatically has defined a missing data indicator (MDI) which has been set to -1×10^{30}. Where the QC tests have set flags and those values have been removed, they are replaced by a flagged data indicator (FDI) of -2×10^{30}.

2.5.5 Where do the observations come from?

The observations in HadISD are all drawn from the Integrated Surface Database (ISD) held by NOAA/NCEI and available from https://www.ncdc.noaa.gov/isd. For further information, do have a read of the papers outlining the ISD [16, 12, 11] and other references available on the ISD website.

2.5.6 How does HadISD.3.1.1 differ from ISD and HadISDH?

HadISD comprises a subset of the ISD stations (around 9000 compared to the 35,000) which are available in netCDF form (rather than fixed field files) and have undergone additional merging and quality control routines.

The HadISDH is a gridded, monthly humidity product based (for the land component) on the HadISD [18, 19]. HadISD is a sub-daily, station based product, and so to create HadISDH, firstly the humidity quantities are calculated on the sub-daily observations, aggregated to monthly values, homogenised and then blended to make gridded fields.

2.5.7 How are stations selected for HadISD?

To be included in HadISD, stations in the ISD need to have a known latitude, longitude and elevation. They also need to cover at least 15 years between the first and last observations. This preliminary station list is then further filterd. The median reporting interval needs to be 6 hourly (on average). Finally, there need to be at least 15 years worth of months which have the equivalent of 6 hourly data (equivalent of 120 observations). Further details on these selection criteria are in [7].

[1] We use zero-indexing for these columns as our scripts are written in Python which is zero-indexed.

Column	Test Code	Test
0	DUP	Duplicate
1	TFV	Frequent Value - Temperature
2	DFV	Frequent Value - Dew point
3	SFV	Frequent Value - SLP
4	DNL	Diurnal Cycle
5	TGP	Gap - Temperature
6	DGP	Gap - Dew point
7	SGP	Gap - SLP
8	TRC	Record - Temperature
9	DRC	Record - Dewpoint
10	WRC	Record - Wind speed
11	PRC	Record - SLP
12	TSS	Straight string - Temperature
13	DSS	Straight string - Dew point
14	WSS	Straight string - Wind Speed
15	PSS	Straight string - SLP
16	HTS	Hour string - Temperature
17	HDS	Hour string - Dew point
18	HWS	Hour string - Wind speed
19	HPS	Hour string - SLP
20	DTS	Day string - Temperature
21	DDS	Day string - Dew point
22	DWS	Day string - Wind speed
23	DPS	Day string - SLP
24	TCM	Climatological Outlier - Temperature
25	DCM	Climatological Outlier - Dew point
26	PCM	Climatological Outlier - SLP
27	TSP	Spike - Temperature
28	DSP	Spike - Dew point
29	PSP	Spike - SLP

Table 1: Link between columns in qc_flags field and the QC tests applied. The test codes are also available in the tests_codes.txt file. [Continued]

2.5.8 How to cite the data?

Dunn, R. J. H., Willett, K. M., Parker, D. E., and Mitchell, L. (2016) Expanding HadISD: quality-controlled, sub-daily station data from 1931, Geosci. Instrum. Method. Data Syst., 5, 473491, https://doi.org/10.5194/gi-5-473-2016

Dunn, R. J. H. (2019) HadISD version 3, Hadley Centre Tech Note #103

Alternatively, use the citation from the CEDA Archive: Met Office Hadley Centre; National Centers for Environmental Information - NOAA (2020): HadISD: Global sub-daily, surface meteorological station data, 1931-2019, v3.1.0.2019f. Centre for Environmental Data Analysis, date of citation. CEDA link

Column	Test Code	Test
30	SSS	Supersaturation
31	DPD	Dew point depression
32	DCF	Dew point cutoff
33	CUOT	Total cloud unobservable
34	CUOL	Low cloud unobservable
35	CUOM	Mid cloud unobservable
36	CUOH	High cloud unobservable
37	CST	Cloud - small total
38	FLW	Full low cloud
39	FMC	Full mid cloud
40	NGC	Negative cloud value
41	TOT	Neighbour outlier - Temperature
42	DOT	Neighbour outlier - Dew point
43	SOT	Neighbour outlier - SLP
44	TMB	Month clean up - Temperature
45	DMB	Month clean up - Dew point
46	SMB	Month clean up - SLP
47	WMB	Month clean up - Wind speed
48	BBB	Month clean up - Wind direction
49	CMB	Month clean up - Total cloud
50	LMB	Month clean up - Low cloud
51	MMB	Month clean up - Mid cloud
52	HMB	Month clean up - High cloud
53	BMB	Month clean up - Cloud base
54	OCT	Odd Cluster - Temperature
55	OCD	Odd Cluster - Dew point
56	OCW	Odd Cluster - Wind speed
57	OCS	Odd Cluster - SLP
58	TVR	Variance - Temperature
59	DVR	Variance - Dew point
60	SVR	Variance - SLP
61	WVR	Variance - Wind speed

Table 1: Continued

2.5.9 There are gaps in the data, what can I do?

If there gaps in the data as indicated by the missing data indicator value (-1×10^{30}) then unfortunately these data were not present in the ISD (our parent dataset). Whether these data are available anywhere else, we are not sure, and would advise you to look in other databases or contact the National Hydrological and Meteorological Service who would be the maintainer of that station.

However, if the gap is indicated by the flagged data indicator (-2×10^{30}) then these values are available in the `flagged_obs` field.

2.5.10 What versions are archived?

Through the CEDA archive (https://catalogue.ceda.ac.uk/) we have archived most of the recent annual updates and will continue to do this for future annual updates. Internally, we archive all annual

Column	Test Code	Test
62	WSL	Wind Logic - Wind speed
63	WDL	Wind Logic - Wind direction
64	WRS	Wind Rose
65	WSP	Spike - Wind speed
66	RSS	Straight string - Wind speed
67	HRS	Hour string - Wind speed
68	DRS	Day string - Wind speed
69	STNLP	Station level pressure
70	PPTN	Precipitation Checks

Table 1: Continued

Flag Value	Meaning
0	No flag set, this test passed
1	Flag set, this test failed, value removed
2	Tentative flag set, to be resolved during neighbour check*
-1	Insufficient neighbours to run neighbour outlier check

Table 2: Flag values. * Note, these tentative flags should all have been set to 0 or 1 during the neighbour checks, but are included here for completeness in case this process failed for some reason.

and at the moment monthly updates as well. In the future, we may need to remove archives of older monthly updates if storage space becomes a limitation.

2.5.11 How does the time axis work?

HadISD is a sub-daily dataset. Some stations report data every hour, some every 3 hours, some every 6 hours. In many cases the reporting frequency for individual stations changes over their record, with longer intervals in the early period, and shorter intervals in the later period (as more stations become automated). We measure the time in hours (integers) starting at 1931-January-1 at 00:00UTC.

The time axis is compressed, which means that the data files do not include those hourly timestamps which have no observational data (flagged or unflagged) associated with them. Only a fraction of the stations within the HadISD have their first observational value in the 1930s, most start in the early 1970s. This approach reduces the on-disc size of these files and hence reduces the download time.

3 Using the time series files

3.1 File names

The filenames for the station files follow a simple pattern

`hadisd.version_start-end_station[_suffix].nc`.

We'll walk through this using the example below:

`hadisd.3.1.1.202007p_19310101-20200801_100040-99999.nc.gz`

hadisd	Dataset identifier (in line with other HadOBS products)
3.1.1.202007p	Version, with increments as outlined in the FAQ
19310101	Start of dataset (January 1, 1931 at 00h00)
20200801	End of dataset (August 1, 2020 at 00h00)
100040-99999	Station identifier (WMO-WBAN))

Column	Test Code
0	Temperature
1	Dew point
2	Sea level pressure
3	Station level pressure
4	Wind speeds
5	Wind direction
6	Total cloud cover
7	Low cloud cover
8	Mid cloud cover
9	High cloud cover
10	1h preciptation total
11	2h preciptation total
12	3h preciptation total
13	6h preciptation total
14	9h preciptation total
15	12h preciptation total
16	15h preciptation total
17	18h preciptation total
18	24h preciptation total

Table 3: Order of columns in the flagged observation field.

The dataset end time is exclusive, the dataset start time is inclusive. So for the example above the first timestamp is January 1, 1901 at 00:00Z (midnight), and the last is July 31, 2020 at 23:00Z (as the midnight observation is assigned to the following day).

The "_suffix" in the above pattern is used to denote the humidity or heat stress files, e.g.:
`hadisd.3.1.1.202007p_19310101-20200801_100040-99999_heat_stress.nc.gz`

3.2 Station identifiers

The station identifiers are taken from the ISD, who use two sets. The first 6-digit one is a quasi-WMO number (from the time when the World Meteorological Organisation [WMO] assigned ranges of numbers to each country. For example, the UK is 03000 to 03899 and the Republic of Ireland is 03900 to 03999.) In the ISD, when the final (sixth) digit is a zero (0), then the first five are the WMO number. If it is not zero, then the first five are a number of the WMO form correct for the geo-political location, but it is not an official WMO number. At the time of writing (August 2020) the WMO numbers are being replaced with a new system.

The last five digits of the station identifier are a WBAN system used by the USAF and NOAA internally for US administered stations only. These should be "99999" for all other stations.

3.3 Contents of data files

We show the field information for the primary NetCDF files in Table 4. These metadata are all available through the NetCDF header information.

Field Name (var_name)	Standard Name	Size	Units	QC	Description
longitude	longitude	[1]	degrees	NA	Station longitude
latitude	latitude	[1]	degrees	NA	Station latitude
elevation	surface_altitude	[1]	m	NA	Station longitude
station_id	-	[12]	-	NA	Station ID number
temperatures	surface_temperature	[t]	C	Y	Dry bulb air temperature at screen height (∼ 2m)
dewpoints	dew_point_temperature	[t]	C	Y	Dew point temperature at screen height (∼ 2m)
slp	air_pressure_at_sea_level	[t]	hPa	Y	Sea level pressure at screen height (∼ 2m)
stnlp	surface_air_pressure	[t]	hPa	Y	Station level pressure at screen height (∼ 2m)
windspeeds	wind_speed	[t]	m/s	Y	Wind speed at mast height (∼ 10m)
winddirs	wind_from_direction	[t]	degrees	Y	Wind direction at mast height (∼ 10m)
total_cloud_cover	cloud_area_fraction	[t]	oktas	Y	Total cloud cover fraction
low_cloud_cover	low_type_cloud_area_fraction	[t]	oktas	Y	Low level cloud cover fraction
mid_cloud_cover	medium_type_cloud_area_fraction	[t]	oktas	Y	Mid level cloud cover fraction
high_cloud_cover	high_type_cloud_area_fraction	[t]	oktas	Y	High cloud cover fraction

Table 4: Fields present in the primary NetCDF files. In the size column, 't' indicates an array dimension of length corresponding to the time axis. [Continued]

3.4 Time Axis

In order to save disc space and reduce the download size for users, the time axis of these files has been compressed. This means that hours which have no data for any of the variables are not retained in the NetCDF files. If there are flagged values, then these time stamps are still available.

Our axis units work in "hours since 1931-01-01 00:00" with integer numbers of hours. Example code to expand this time axis to the full range is given in Section 4, but in essence the approach we recommend is as follows. Using the start and end times given for the dataset version (present in each file name), make an array for the complete set of hours. Then match the ones present in the file to this complete set.

3.5 Tools that can be used to work with data files

A list of software tools that work with NetCDF files is maintained by UCAR https://www.unidata.ucar.edu/software/netcdf/software.html). Some simple tools for viewing and manipulating NetCDF files in Linux include:

- ncdump: provided with the NetCDF library, produces a text rendering of a NetCDF file (Unidata at UCAR).

Field Name (var_name)	Standard Name	Size	Units	QC	Description
precip1_depth	lwe_thickness_of_precipitation_amount	[t]	mm	Y	Depth of precipitation reported in 1 hour
precip2_depth	lwe_thickness_of_precipitation_amount	[t]	mm	Y	Depth of precipitation reported in 2 hours
precip3_depth	lwe_thickness_of_precipitation_amount	[t]	mm	Y	Depth of precipitation reported in 3 hours
precip6_depth	lwe_thickness_of_precipitation_amount	[t]	mm	Y	Depth of precipitation reported in 6 hours
precip9_depth	lwe_thickness_of_precipitation_amount	[t]	mm	Y	Depth of precipitation reported in 9 hours
precip12_depth	lwe_thickness_of_precipitation_amount	[t]	mm	Y	Depth of precipitation reported in 12 hours
precip15_depth	lwe_thickness_of_precipitation_amount	[t]	mm	Y	Depth of precipitation reported in 15 hours
precip18_depth	lwe_thickness_of_precipitation_amount	[t]	mm	Y	Depth of precipitation reported in 18 hours
precip24_depth	lwe_thickness_of_precipitation_amount	[t]	mm	Y	Depth of precipitation reported in 24 hours
cloud_base	cloud_base_altitude	[t]	m	N	Cloud base height of lowest layer
wind_gust	wind_speed_of_gust	[t]	m/s	N	Wind gust speed at mast height ($\sim$ 10m)
past_sigwx1	-	[t]	-	N	Past significant weather phenomena (refer to ISD documentation)
time	time	[t]	-	NA	Time of measurement
input_station_id	-	[t,12]	-	NA	Input station ID for each observation (for merged stations)
quality_control_flags	-	[t,71]	-	NA	QC flags for each test on each variable
flagged_obs	-	[t,19]	-	NA	Values removed from each variable by the QC tests
reporting_stats	-	[-]	-	-	Not currently used

Table 4: Continued.

- Climate Data Operators (CDO)s: a set of command line utilities for performing operations on NetCDF files including concatenation, editing and mathematics (https://code.mpimet.mpg.de/projects/cdo).
- ncview: a program to produce graphical displays of the contents of NetCDF files. More information can be found here. A more complete list can be found here.

In addition, packages are available in most commonly-used scientific programming languages for reading and working with NetCDF files. For example, in Python there are numerous packages including:

- netCDF4 - this basic package provides functionality to read NetCDF files and extract metadata.

- Iris - developed by the Met Office, Iris provides functionality to read, write and process files in a variety of formats including NetCDF. Some example snippets of code using Iris to process HadISD.3.1.1 are provided in Section 4.

There are a number of packages in R that can be used to process NetCDF files:

- ncdf4: https://cran.r-project.org/web/packages/ncdf4/index.html
- raster: https://cran.r-project.org/web/packages/raster/index.html
- rcdo: https://github.com/r4ecology/rcdo
- RNetCDF: https://cran.r-project.org/web/packages/RNetCDF/index.html
- CM SAF R tools: https://www.mdpi.com/2220-9964/8/3/109

4 Worked examples

4.1 Inspection of temperature series

Read in the temperature data and plot the resulting timeseries using the Iris package in Python.

```
import os
import iris
import iris.quickplot as qplt
import matplotlib.pyplot as plt

DATALOC = "/data/local/"

temperature_cube = iris.load_cube(os.path.join(DATALOC, \
           'hadisd.3.1.1.202007p_19310101-20200801_100040-99999.nc'), "surface_temperature")

qplt.plot(temperature_cube, ".")
plt.show()
```

Note that this uses the `standard_name` value, rather than the `var_name` attribute used in this guide. An alternate option is given below.

```
import os
import iris
import iris.quickplot as qplt
import matplotlib.pyplot as plt

DATALOC = "/data/local/"

cubes = iris.load(os.path.join(DATALOC,
             'hadisd.3.1.1.202007p_19310101-20200801_100040-99999.nc'))

for cube in cubes:
    if cube.var_name == "temperatures":
        temperature_cube = cube
        break

qplt.plot(temperature_cube, ".")
plt.show()
```

4.2 Expansion of time axis

If not using the Iris utilities, then to uncompress the time axis the following can be adopted.

```
import os
import iris
import numpy as np
import datetime as dt
import matplotlib.pyplot as plt

DATALOC = "/data/local/"

temperature_cube = iris.load_cube(os.path.join(DATALOC, \
          'hadisd.3.1.1.202007p_19310101-20200801_100040-99999.nc'), "surface_temperature")

times = temperature_cube.coord('time').points
temperatures = temperature_cube.data

dataspan = dt.datetime(2020, 8, 1) - dt.datetime(1931, 1, 1)
full_times = np.arange(dataspan.days * 24)

mask = np.in1d(full_times, times, assume_unique=True)

full_temperatures = np.ma.zeros(full_times.shape)
full_temperatures.mask = np.ones(full_times.shape)
full_temperatures.fill_value(-1.e30)
full_temperatures[mask] = temperatures

plt.scatter(full_times, full_temperatures, ".")
plt.xlabel("Hours since 1931-01-01 00:00")
plt.ylabel("Temperature (C)")

plt.show()
```

5 Dataset Characteristics

5.1 Homogeneity

We outline in [5] the steps we have taken to make a homogeneity assessment of HadISD (temperature, dew point temperature, sea level pressure and wind speed variables only).

This assessment is repeated in January every year once that version of the dataset is final. The results are made available on http://www.metoffice.gov.uk/hadobs/hadisd/.

5.2 Comparison to other long records

HadISD is not homogenised, although the homogeneity assessment identifies both break-point locations in time and their magnitudes. This means that for long-term trend analysis careful further investigation and preparation may be necessary. The outcomes of the homogeneity analysis can be used to select stations which have few or small magnitude breaks (the most homogeneous ones) for use in these kinds of assessments. We show an example of this in [5], where we use CRUTEM [9] as an independent dataset for surface temperature over land. We selected stations with fewer or smaller inhomogeneities for a range of criteria, and hence were able to show the effect of removing the stations with the largest or most persistent breaks, and also the resulting change to the station network.

However, the data themselves have not been homogenised, in that no adjustments have been applied.

6 Appendix

We show in Tables 5 and 6 the contents of the humidity and heat stress NetCDF files respectively.

Field Name (var_name)	Standard Name	Size	Units	QC	Description
longitude	longitude	[1]	degrees	NA	Station longitude
latitude	latitude	[1]	degrees	NA	Station latitude
elevation	surface_altitude	[1]	m	NA	Station longitude
station_id	-	[12]	-	NA	Station ID number
temperatures	surface_temperature	[t]	C	Y	Dry bulb air temperature at screen height (∼ 2m)
dewpoints	dew_point_temperature	[t]	C	Y	Dew point temperature at screen height (∼ 2m)
slp	air_pressure_at_sea_level	[t]	hPa	Y	Sea level pressure at screen height (∼ 2m)
vapor_pressure	water_vapor_pressure	[t]	hPa	NA	Vapor pressure calculated w.r.t. water
saturation_vapor_pressure	water_vapor_pressure	[t]	hPa	NA	Saturation vapor pressure calculated w.r.t. water
wet_bulb_temperature	wet_bulb_temperature	[t]	C	NA	Wet bulb temperature at screen height (∼ 2m)
specific_humidity	specific_humidity	[t]	g/kg	NA	Specific humidity at screen height (∼ 2m)
relative_humidity	relative_humidity	[t]	%rh	NA	Relative humidity at screen height (∼ 2m)
time	time	[t]	-	NA	Time of measurement
input_station_id	-	[t,12]	-	NA	Input station ID for each observation (for merged stations)

Table 5: Fields present in the humidity NetCDF files. In the size column, '*t*' indicates an array dimension of length corresponding to the time axis.

We show in Tables 7 and 8 the calculations used for the humidity and heat stress metrics, reproduced from the HadISD2 paper [7].

References

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

[2] Arden L Buck. New equations for computing vapor pressure and enhancement factor. *Journal of applied meteorology*, 20(12):1527–1532, 1981.

[3] S Dikmen and PJ Hansen. Is the temperature-humidity index the best indicator of heat stress in lactating dairy cows in a subtropical environment? *Journal of dairy science*, 92(1):109–116, 2009.

[4] RJH Dunn. Hadisd v3: monthly updates. *Hadley Centre Tech Note*, (103), 2019.

Field Name (var_name)	Standard Name	Size	Units	QC	Description
longitude	longitude	[1]	degrees	NA	Station longitude
latitude	latitude	[1]	degrees	NA	Station latitude
elevation	surface_altitude	[1]	m	NA	Station longitude
station_id	-	[12]	-	NA	Station ID number
temperatures	surface_temperature	[t]	C	Y	Dry bulb air temperature at screen height (∼ 2m)
dewpoints	dew_point_temperature	[t]	C	Y	Dew point temperature at screen height (∼ 2m)
windspeeds	wind_speed	[t]	m/s	Y	Wind speed at mast height (∼ 10m)
temperature_humidity_index	temperature_humidity_index	[t]		NA	Temperature humidity index (THI) at screen height (∼ 2m)
wet_bulb_globe_temperature	wet_bulb_globe_temperature	[t]		NA	Wet bulb globe temperature (WBGT) at screen height (∼ 2m)
humidex	humidex	[t]		NA	Humidex at screen height (∼ 2m)
apparent_temperature	apparent_temperature	[t]		NA	Apparent temperature at screen height (∼ 2m)
heat_index	heat_index	[t]		NA	Heat index at screen height (∼ 2m)
time	time	[t]	-	NA	Time of measurement
input_station_id	-	[t,12]	-	NA	Input station ID for each observation (for merged stations)

Table 6: Fields present in the heat stress NetCDF files. In the size column, '*t*' indicates an array dimension of length corresponding to the time axis.

[5] RJH Dunn, KM Willett, CP Morice, and DE Parker. Pairwise homogeneity assessment of hadisd. *Climate of the Past*, 10(4):1501–1522, 2014.

[6] RJH Dunn, KM Willett, PW Thorne, EV Woolley, I Durre, A Dai, DE Parker, and RS Vose. Hadisd: a quality-controlled global synoptic report database for selected variables at long-term stations from 1973–2011. *Climate of the Past*, 8(5):1649–1679, 2012.

[7] Robert JH Dunn, Kate M Willett, David E Parker, and Lorna Mitchell. Expanding hadisd: Quality-controlled, sub-daily station data from 1931. *Geoscientific Instrumentation, Methods and Data Systems*, 5(2):473, 2016.

[8] Marvin Eli Jensen, Robert D Burman, and Rick G Allen. Evapotranspiration and irrigation water requirements. ASCE, 1990.

[9] PD Jones, DH Lister, TJ Osborn, C Harpham, M Salmon, and CP Morice. Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research: Atmospheres*, 117(D5), 2012.

Variable	Equation	Source	Notes
Specific humidity (q) in g kg^{-1}	$q = 1000\left(\frac{0.622e}{P_{\mathrm{mst}} - ((1-0.622)e)}\right)$	[14]	
Relative humidity (RH) in %rh	$RH = 100\left(\frac{e}{e_{\mathrm{s}}}\right)$		
Vapour Pressure (e) with respect to water in HPa (when $T_{\mathrm{w}} > 0°\mathrm{C}$)	$e = 6.1121.f_{\mathrm{w}}.\exp\left(\frac{18.729 - \left(\frac{T_{\mathrm{d}}}{227.3}\right)T_{\mathrm{d}}}{257.87 + T_{\mathrm{d}}}\right)$ $f_{\mathrm{w}} = 1 + 7\times10^{-4} + 3.46\times10^{-6}P_{\mathrm{mst}}$	[2]	Substitute T for T_{d} to give the saturation vapour pressure e_{s}
Vapour Pressure (e) with respect to ice in HPa (when $T_{\mathrm{w}} \leq 0°\mathrm{C}$)	$e = 6.1115.f_{\mathrm{w}}.\exp\left(\frac{23.036 - \left(\frac{T_{\mathrm{d}}}{333.7}\right)T_{\mathrm{d}}}{279.82 + T_{\mathrm{d}}}\right)$ $f_{\mathrm{w}} = 1 + 3\times10^{-4} + 4.18\times10^{-6}P_{\mathrm{mst}}$	[2]	
Wet bulb temperature (T_{w}) in °C	$T_{\mathrm{w}} = \frac{aT + bT_{\mathrm{d}}}{a+b}$ $a = 6.6\times10^{-5}P_{\mathrm{mst}}$ $b = \frac{409.8e}{(T_{\mathrm{d}}+237.3)^2}$	[8]	
Station Pressure in hPa	$P_{\mathrm{mst}} = P_{\mathrm{msl}}\left(\frac{T}{T + 0.0065Z}\right)^{5.625}$	[10]	Temperature T, station height Z in metres

Table 7: Humidity formulae used in HadISD v2.0.0. as in HadISDH v2.0.0 [19].

[10] R J List. Smithsonian meteorological tables. *Quarterly Journal of the Royal Meteorological Society*, 114, 1963.

[11] J Neal Lott. 7.8 the quality control of the integrated surface hourly database. 2004. https://ams.confex.com/ams/84Annual/webprogram/Paper71929.html.

[12] Neal Lott, R Vose, SA Del Greco, TF Ross, S Worley, and JL Comeaux. The integrated surface database: Partnerships and progress. In *Extended Ab-*

Variable	Equation	Source	Notes
Temperature-Humidity Index (THI)	$THI = (1.8T + 32) - (0.55 - 0.0055RH)(1.8T - 26))$	[3]	
Pseudo Wet-bulb Globe Temperature (WBGT)	$WBGT = (0.567T) + (0.393e_{\mathrm{v}}) + 3.94$	[1]	
Humidex	$h = T + (0.5555(e_{\mathrm{v}} - 10))$	[13]	
Apparent Temperature	$T_{\mathrm{a}} = T + (0.33e) - (0.7w) - 4$	[17]	
Heat Index	$HI = -42.379 + 2.04901523T_{\mathrm{f}} + 10.14333127RH - 0.22475541T_{\mathrm{f}}RH - 0.006837837T_{\mathrm{f}}^2 - 0.05481717RH^2 + 0.001228747T_{\mathrm{f}}^2RH + 8.5282 \times 10^{-4}T_{\mathrm{f}}RH^2 - 1.99 \times 10^{-6}T_{\mathrm{f}}^2RH^2$ $\mathrm{adj}_1 = \frac{13RH}{4}\sqrt{\frac{17\mathrm{abs}(T_{\mathrm{f}} - 95)}{17}}$ $\mathrm{adj}_2 = \frac{RH - 85}{10}.\frac{87 - T_{\mathrm{f}}}{5}$ $HI = 0.5(T_{\mathrm{f}} + 61 + 1.2(T_{\mathrm{f}} - 68) + 0.094RH)$	[15]	Where T_{f} is the temperature in Fahreneit. If $RH < 13$ and $80 \leq T_{\mathrm{f}} \leq 112$, adj_1 is subtracted from HI; if $RH > 85$ and $80 \leq T_{\mathrm{f}} \leq 87$ adj_2 is added to HI. Furthermore, if these calculations would result in a $HI < 80$, then the simpler formula is used.

Table 8: Heat stress measures calculated in HadISD v2.0.0. T is dry bulb temperature, RH is the relative humidity, e is the vapor pressure, and w the wind speed.

stracts, 24th Conf. on Interactive Information and Processing Systems, 2008. https://ams.confex.com/ams/88Annual/webprogram/Paper131387.html.

[13] JM Masterton and FA Richardson. *Humidex: a method of quantifying human discomfort due to excessive heat and humidity*. Downsview, Ont.: Atmospheric Environment, 1979.

[14] JoséP Peixoto and Abraham H Oort. The climatology of relative humidity in the atmosphere. *Journal of climate*, 9(12):3443–3463, 1996.

[15] Lans P Rothfusz. 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*, pages 90–23, 1990.

[16] Adam Smith, Neal Lott, and Russ Vose. The integrated surface database: Recent developments and partnerships. *Bulletin of the American Meteorological Society*, 92(6):704–708, 2011.

[17] Robert G Steadman. Norms of apparent temperature in australia. *Aust. Met. Mag*, 43:1–16, 1994.

[18] Kate M Willett, CN Williams, Robert JH Dunn, Peter W Thorne, Stephanie Bell, Michael de Podesta, Phil D Jones, and David E Parker. Hadisdh: an updateable land surface specific humidity product for climate monitoring. *Climate of the Past*, 9(2), 2013.

[19] KM Willett, RJH Dunn, PW Thorne, S Bell, M de Podesta, DE Parker, PD Jones, and CN Williams Jr. Hadisdh land surface multi-variable humidity and temperature record for climate monitoring. *Climate of the Past*, 10(6):1983–2006, 2014.