Computers and Geosciences 122 (2019) 35–44

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

GHCN-Daily: a treasure trove of climate data awaiting discovery

Jasmine B.D. Jaffrés[a,b,*,1]

[a] C&R Consulting, Townsville, Australia
[b] College of Science and Engineering, James Cook University, Townsville, Australia

ARTICLE INFO

Keywords:
GHCN-Daily database
Weather stations
Global climate data availability
Quality control
MATLAB
GNU Octave

ABSTRACT

International collaboration to create and maintain international, freely accessible datasets greatly facilitates research in many scientific fields. The Global Historical Climatology Network (GHCN)-Daily database provides access to a diverse range of daily weather station data, including precipitation and temperature variables. These data are supplied as individual, station-specific files and structured in a non-delimited format. Here, the GHCN-Daily data structure, spatio-temporal content and associated caveats are delineated. The regularly updated collection now features data from over 100 000 stations in 218 countries and territories. While rigorous quality tests are routinely applied for GHCN-Daily, the database excludes the original quality flags from the source agencies.

The extraction of climate variables from the GHCN-Daily database can be challenging for novice users and may thus dissuade from the uptake of this valuable dataset. Consequently, a user-friendly toolkit for MATLAB and GNU Octave is also provided to aid data retrieval from all relevant weather stations. The toolkit reformats the extracted GHCN-Daily data into a more accessible structure to facilitate data mining and research on a large scale.

1. Introduction

International, freely accessible databases are of great assistance to scientific research. A monthly version of the Global Historical Climatology Network (GHCN-Monthly) database was first released in the early 1990s, containing monthly temperature, precipitation and pressure data (Vose et al., 1992). The high popularity of GHCN-Monthly among the scientific community lead to progressive revisions and expansions of the database (Lawrimore et al., 2011; Peterson and Vose, 1997). More recently, a daily counterpart (GHCN-Daily) was released (Menne et al., 2012b), in recognition of higher data resolution requirements for some climate applications.

The GHCN-Daily database collates daily weather station data from across the world and is likely the most comprehensive global dataset for daily in situ climate measurements (Menne et al., 2012b). Since the overview by Menne et al. (2012b), an additional 38 countries and territories have been added to the database (Table 1). The collection now encompasses data from 218 countries and territories, with some records going back as far as 1763. GHCN-Daily is updated operationally, although the station record is source-dependent. A relatively dense network of up-to-date data is obtainable for Australia, Japan, Europe and North America. For some regions, most notably for Brazil, India and South Africa, historical data availability from GHCN-Daily is extensive, while updated data are only provided for a small subset of weather stations. Consequently, these regions exhibit a decline in data availability for the most recent years, as evident with weather stations collecting precipitation data between 1900 and 2017 (Fig. 1).

Access to the complete GHCN-Daily dataset, or the full temporal record of individual files, is provided in ASCII text format (*.dly* file extension, Fig. 2). Data access from these fixed-width text files can be challenging for novice users because of the unorthodox data format. Hence, a toolbox is presented here to extract and restructure the GHCN-Daily data in GNU Octave (Eaton et al., 2017) or MATLAB to facilitate data mining. The package is accessible through the GitHub (https://github.com/) and SourceForge (https://sourceforge.net/) file exchanges. The GHCN-Daily data structure, content and associated caveats are also examined in this paper.

2. Methods

2.1. Contents of GHCN-Daily

The GHCN-Daily database contains a wide range of variables. Five variables are classified as core elements in GHCN-Daily. These include

* C&R Consulting, Townsville, Australia.
E-mail address: jasmine@candrconsulting.com.au.
[1] Author Contributions: J.B.D. Jaffrés conceived, designed and executed all aspects of this publication and associated toolbox.

https://doi.org/10.1016/j.cageo.2018.07.003
Received 28 April 2018; Received in revised form 29 June 2018; Accepted 11 July 2018
Available online 02 August 2018
0098-3004/

Table 1
List of newly added countries and territories in GHCN-Daily since the overview by Menne et al. (2012b, their Table 5). [FIPS: Federal Information Processing Standard; UK: United Kingdom].

FIPS Code	Country	FIPS Code	Country
BG	Bangladesh	LO	Slovakia
BH	Belize	MB	Martinique (France)
BM	Burma	MC	Macau S.A.R.
BU	Bulgaria	MF	Mayotte (France)
BX	Brunei	MJ	Montenegro
CB	Cambodia	NE	Niue (New Zealand)
CG	Congo (Kinshasa)	NF	Norfolk Island (Australia)
CV	Cape Verde	NI	Nigeria
CW	Cook Islands (New Zealand)	NS	Suriname
DO	Dominica	NT	Netherlands Antilles (Netherlands)
EK	Equatorial Guinea	PU	Guinea-Bissau
EU	Europa Island (France)	QA	Qatar
FK	Falkland Islands (UK)	RE	Réunion (France)
GA	Gambia, The	RW	Rwanda
GH	Ghana	SB	Saint Pierre and Miquelon (France)
GI	Gibraltar (UK)	SN	Singapore
JO	Jordan	SX	South Georgia and the South Sandwich Islands (UK)
JU	Juan De Nova Island (France)	TE	Tromelin Island (France)
LE	Lebanon	WI	Western Sahara

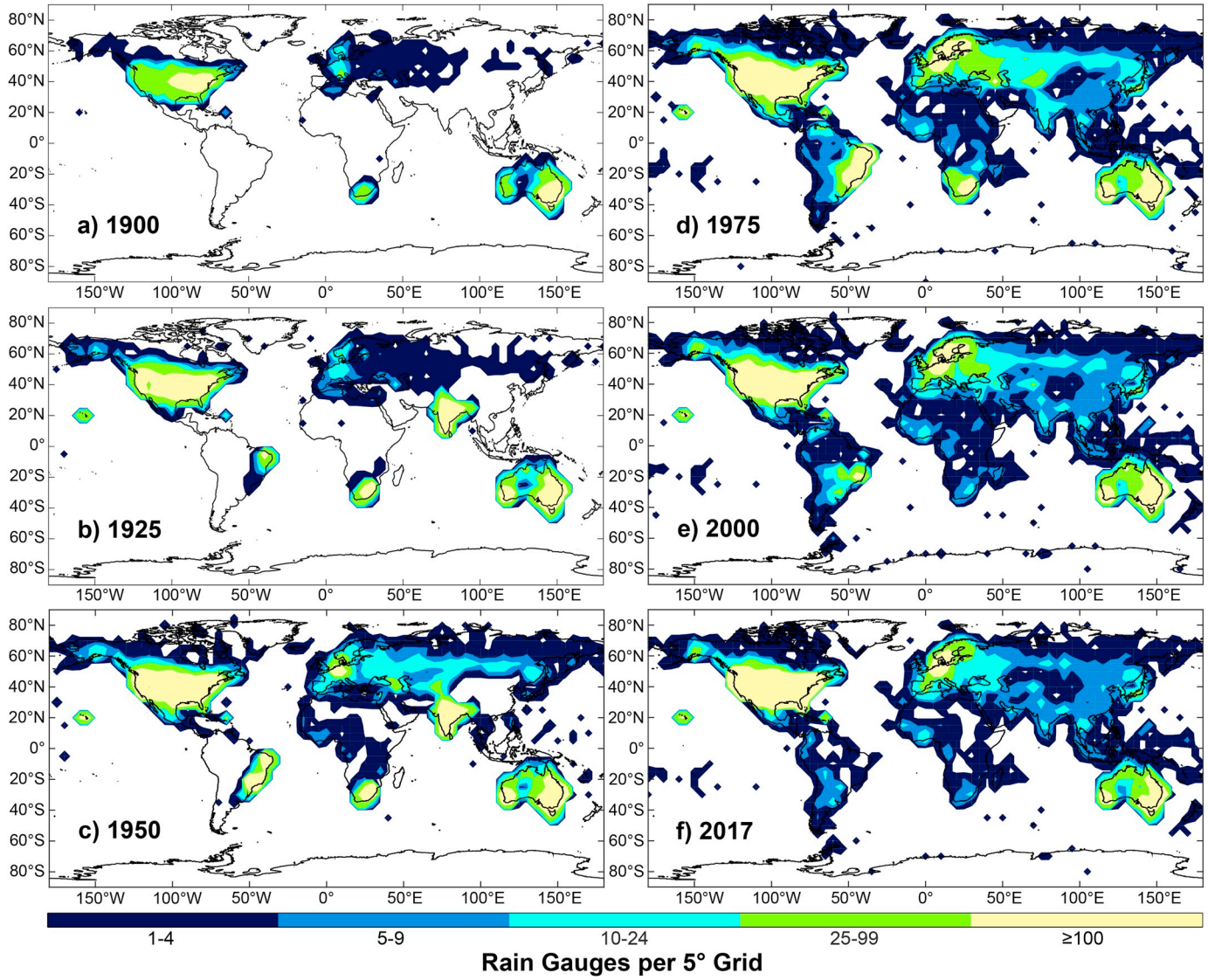

Fig. 1. Number of operational weather stations within the GHCN-Daily database recording precipitation data per 5° grid cell in a) 1900, b) 1925, c) 1950, d) 1975, e) 2000 and f) 2017.

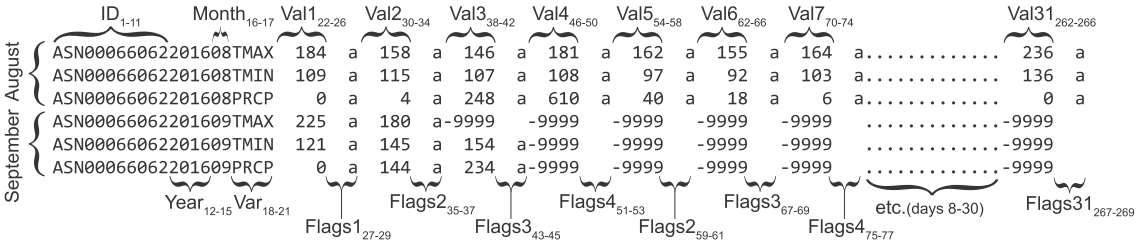

Fig. 2. Sample of GHCN-Daily *.dly* file structure. Subscripts refer to the column location. Missing variable values are denoted by −9999. [ID: weather station identifier; Var: variable; Val1: variable value on day 1; Flags1: three flags for day 1 - measurement flag, quality flag and source flag; Val2: variable value on day 2; Flags2: three flags for day 2; etc.].

precipitation (PRCP), minimum and maximum temperature (TMIN and TMAX, respectively), snowfall (SNOW) and snow depth (SNWD). The full list of available variables is provided in Table 2. Each *.dly* file comprises data from an individual weather station that recorded at least one type of variable. Missing values are denoted by −9999. Metadata contain three flags for every data value, including a measurement flag, a quality flag and a source flag (Table 3). The full GHCN-Daily archive (version 3.23-upd-2018031804) was accessed (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/) for this study, encompassing data from more than 100 000 weather stations around the globe (Menne et al., 2012a).

2.2. Extraction of GHCN-Daily data

Data retrieval from global databases is not always trivial, especially for large-scale studies and novice users. GHCN-Daily data are retrieved from separate text files (with *.dly* extension) for each weather station. These text files are not in a delimited format (Fig. 2) and new users may therefore have difficulty extracting the required data. GHCN-Daily data are listed in chronological order. Each row contains data for one month (31 days, variables for excess days are assigned −9999) and one variable (Fig. 2). If several variables were recorded by the weather station, each variable is listed consecutively on separate rows for that month.

A GHCN-Daily data extraction routine is provided to facilitate GHCN-Daily data mining. The toolkit (*ghcnd_access*) is compatible with both MATLAB (R2017b) and free open-source alternative GNU Octave (v4.2.0; Eaton et al., 2017). The script was purposely written generically to allow retrieval and transformation of any data contained within the database. Several existing functions were tested during the process of writing the script. Functions *fread* and *fgetl* are both suitable for data extraction from *.dly* files, although the *fread* function was found to be superior, as *fgetl* is performing the extraction at significantly higher computational expense (> 5 times longer run time).

Table 2
List of variables contained in GHCN-Daily, along with their associated abbreviation and unit (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt).

Abbreviation	Unit	Variable	Abbreviation	Unit	Variable
PRCP	mm/10	precipitation	MDWM	km	multiday wind movement
SNOW	mm	snowfall	MNPN	°C/10	daily minimum temperature of evaporation pan water
SNWD	mm	snow depth	MXPN	°C/10	daily maximum temperature of evaporation pan water
TMAX	°C/10	maximum temperature	PGTM	HHMM [a]	peak gust time
TMIN	°C/10	minimum temperature	PSUN	%	daily percent of possible sunshine
ACMC	%	24-hour average cloudiness (30-second ceilometer data)	SN*# [b]	°C/10	minimum soil temperature
ACMH	%	24-hour average cloudiness (manual observations)	SX*# [b]	°C/10	maximum soil temperature
ACSC	%	average cloudiness sunrise to sunset (30-second ceilometer data)	TAVG	°C/10	average temperature [c]
ACSH	%	average cloudiness sunrise to sunset (manual observations)	THIC	mm/10	thickness of ice on water
AWDR	°	average wind direction	TOBS	°C/10	temperature at the time of observation
AWND	(m s^{-1})/10	average wind speed	TSUN	min	daily total sunshine
DAEV		number of days included in MDEV	WDF1	°	direction of fastest 1-minute wind
DAPR		number of days included in MDPR	WDF2	°	direction of fastest 2-minute wind
DASF		number of days included in MDSF	WDF5	°	direction of fastest 5-minute wind
DATN		number of days included in MDTN	WDFG	°	direction of peak wind gust
DATX		number of days included in MDTX	WDFI	°	direction of highest instantaneous wind
DAWM		number of days included in MDWM	WDFM	°	fastest mile wind direction
DWPR		number of days with non-zero precipitation included in MDPR	WDMV	km	24-hour wind movement
EVAP	mm/10	pan evaporation	WESD	mm/10	water equivalent of snow on the ground
FMTM	HHMM [a]	time of fastest mile or fastest 1-minute wind	WESF	mm/10	water equivalent of snowfall
FRGB	cm	base of frozen ground layer	WSF1	(m s^{-1})/10	fastest 1-minute wind speed
FRGT	cm	top of frozen ground layer	WSF2	(m s^{-1})/10	fastest 2-minute wind speed
FRTH	cm	thickness of frozen ground layer	WSF5	(m s^{-1})/10	fastest 5-minute wind speed
GAHT	cm	difference between river and gauge height	WSFG	(m s^{-1})/10	peak gust wind speed
MDEV	mm/10	multiday evaporation total (use with DAEV)	WSFI	(m s^{-1})/10	highest instantaneous wind speed
MDPR	mm/10	multiday precipitation total (use with DAPR and DWPR, if available)	WSFM	(m s^{-1})/10	fastest mile wind speed
MDSF	mm	multiday snowfall total (use with DASF)	WT** [d]		weather type
MDTN	°C/10	multiday minimum temperature (use with DATN)	WV** [d]		weather in the vicinity
MDTX	°C/10	multiday maximum temperature (use with DATX)			

[a] Hours (HH) and minutes (MM).
[b] See Table 4 for code specification (* and #).
[c] Note: TAVG from source 'S' corresponds to an average for the period ending at 2400 UTC rather than local midnight (cf. Table 3).
[d] See Table 5 for code specification (**).

Table 3
Flag specifications for GHCN-Daily data values (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt).

Flag	Measurement	Flag	Quality	Flag	Source
Blank	no measurement information applicable	Blank	did not fail any quality assurance check	Blank	no source (i.e. data value missing)
				0	U.S. Cooperative Summary of the Day (NCDC DSI-3200)
				6	CDMP Cooperative Summary of the Day (NCDC DSI-3206)
				7	U.S. Cooperative Summary of the Day - transmitted via WxCoder3 (NCDC DSI-3207)
				A	U.S. Automated Surface Observing System (ASOS) real-time data
				a	Australian Bureau of Meteorology (BOM)
B	precipitation total formed from two 12-hour totals			B	U.S. ASOS data for October 2000-December 2005 (NCDC DSI-3211)
				b	Belarus update
				C	Environment Canada
D	precipitation total formed from four six-hour totals	D	failed duplicate check	D	Short-time delay US National Weather Service CF6 daily summaries provided by the High Plains Regional Climate Center
				E	European Climate Assessment and Dataset
				F	U.S. Fort data
		G	failed gap check	G	Official Global Climate Observing System or other government-supplied data
H	highest or lowest hourly temperature (TMAX or TMIN) or the average of hourly values (TAVG)			H	High Plains Regional Climate Center real-time data
		I	failed internal consistency check	I	Non-U.S. data received through personal contacts
K	converted from knots	K	failed streak/frequent-value check	K	U.S. Cooperative Summary of the Day data digitised from paper observer forms
L	temperature appears to be lagged with respect to reported hour of observation	L	failed check on length of multiday period		
		M	failed megaconsistency check	M	Monthly METAR Extract (additional ASOS data)
		N	failed naught check	N	Community Collaborative Rain, Hail and Snow Network
O	converted from oktas	O	failed climatological outlier check		
P	identified as “missing presumed zero” in DSI 3200 and 3206				
				Q	Quarantined African data - withheld from public release until permission was granted from the respective meteorological services
		R	failed lagged range check	R	NCEI Reference Network Database
				r	All-Russian Research Institute of Hydrometeorological Information-World Data Center
		S	failed spatial consistency check	S	Global Summary of the Day (NCDC DSI-9618) NOTE: These derived daily values may differ significantly from “true” daily data, particularly for precipitation (use with caution).
				s	China Meteorological Administration
T	trace of precipitation, snowfall or snow depth	T	failed temporal consistency check	T	Snowpack telemetry data from the U.S. Department of Agriculture's Natural Resources Conservation Service
				U	Remote automatic weather station data from the Western Regional Climate Center
				u	Ukraine update
W	converted from 16-point WBAN (Weather Bureau Army Navy) code (for wind direction)	W	temperature too warm for snow	W	WBAN/ASOS Summary of the Day from NCDC's Integrated Surface Data (ISD)
		X	failed bounds check	X	U.S. First-Order Summary of the Day (NCDC DSI-3210)
		Z	flagged as a result of an official Datzilla investigation	Z	Datzilla official additions or replacements
				z	Uzbekistan update

Table 4
Ground cover (*) and soil depth (#) codes for variables SN*# and SX*# in GHCN-Daily (cf. Table 2).

Code (*)	Ground Cover	Code (#)	Soil Depth (cm)
0	unknown		
1	grass	1	5
2	fallow	2	10
3	bare ground	3	20
4	brome grass	4	50
5	sod	5	100
6	straw mulch	6	150
7	grass muck	7	180
8	bare muck		

2.3. User inputs

The toolkit allows for several user inputs. For each code execution, the user is requested to select the target variable (*var_target*) to be retrieved. The full list of variables (Table 2) can also be accessed on the GHCN-Daily website (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt). The term *in_dir* specifies the path to the GHCN-Daily *.dly* files, while *out_dir* provides the location of the output file (*.mat* file, containing extracted and transformed variable). *dly_which* lists the weather station files to be considered for access, which may include all or a subset of available files.

GHCN-Daily supplies some data in non-standard metric units by applying a scaling factor (e.g. tenths of mm, tenths of °C, tenths of m/s). By default, the toolkit converts these variables into a more regular format (mm for precipitation, °C for temperature, etc.), although the option is provided to maintain the original units.

Optional conversions into monthly values (e.g. average temperature or total precipitation) are available for most variables, with an additional matrix (*count_month_obs*) created that lists the number of days with data per month and weather station. The matrix *count_month_obs* permits the elimination of scarcely sampled months, without re-extracting the data from the *.dly* files.

The toolbox enables the retrieval of the chosen variable from all weather station files located in *in_dir*. The script may require several hours to assemble all data if the complete GHCN-Daily database is accessed (in excess of 100 000 individual weather station files), depending on the selected variable. Several options are provided to decrease the number of data files accessed for compilation. An inventory file (ghcnd-inventory.txt) is issued with the GHCN-Daily database (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-inventory.txt). This inventory file permits targeted file access by limiting data retrieval to weather stations that contain data for *var_target*. Further, data can be extracted for the full period (1763 – now), a reduced range or a specific year (*yr_options*). Additional user input is requested when the reduced range (start and end year) or the specific year option is selected.

2.4. Data requirements

The toolkit requires for at least one GHCN-Daily *.dly* file to be present in the input directory. Further, the GHCN-Daily inventory (ghcnd-inventory.txt) and station (ghcnd-stations.txt; ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.txt) files need to be located in the same input folder. The inventory file allows the elimination of unwanted *.dly* file, while the stations file is utilised to extract and store the geographical position of all relevant weather stations.

2.5. Data extraction, transformation and storage

The *ghcnd_access* toolkit extracts and transforms the requested GHCN-Daily data based on the selected user inputs. All data are aggregated in several matrices, as described in the user's guide provided with the *ghcnd_access* toolkit. Fig. 3 shows the matrix structure for monthly data output, with additional matrices containing date and weather station information. The matrix structure for daily data closely resembles Fig. 3. The main difference consists in the omission of data gaps in daily outputs for improved storage efficiency, with individual weather station records and associated date vectors saved in a cell array. By default, the extracted data are stored in mat-file format, although users can alter the script to save data in additional file formats, such as ASCII text files for small datasets or netCDF for larger data files. A post-processing script is provided to change the compiled daily data from *.mat* to *.xlsx* format.

3. Results - data availability

The GHCN-Daily database contains weather station data from the year 1763 onwards. Data availability in GHCN-Daily varies significantly depending on data type and region. Most GHCN-Daily data (54.4%) derive from the United States, followed by Australia (16.2%), Canada (8.2%) and Brazil (5.7%) (Fig. 4). Seventy-six nations and territories are

Table 5
Codes for variables Weather Type (WT**) and Weather in the Vicinity (WV**) in GHCN-Daily (cf. Table 2).

Code (**)	Weather Type	Code (**)	Weather in the Vicinity
01	fog, ice fog or freezing fog (may include heavy fog)	01	fog, ice fog or freezing fog (may include heavy fog)
02	heavy fog or heaving freezing fog (not always distinguished from fog)		
03	thunder	03	thunder
04	ice pellets, sleet, snow pellets or small hail		
05	hail (may include small hail)		
06	glaze or rime		
07	dust, volcanic ash, blowing dust, blowing sand or blowing obstruction	07	ash, dust, sand or other blowing obstruction
08	smoke or haze		
09	blowing or drifting snow		
10	tornado, waterspout or funnel cloud		
11	high or damaging winds		
12	blowing spray		
13	mist		
14	drizzle		
15	freezing drizzle		
16	rain (may include freezing rain, drizzle, and freezing drizzle)		
17	freezing rain		
18	snow, snow pellets, snow grains or ice crystals	18	snow or ice crystals
19	unknown source of precipitation		
		20	rain or snow shower
21	ground fog		
22	ice fog or freezing fog		

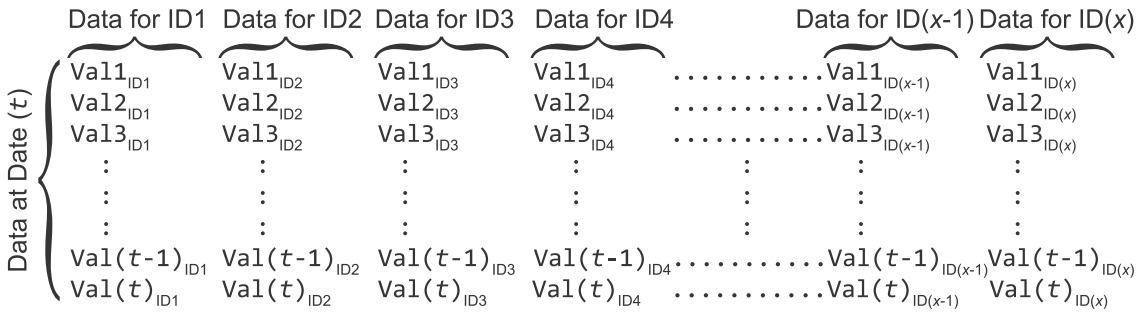

Fig. 3. Matrix structure for monthly output. Subscripts refer to the column location. $\mathrm{Val}(t)_{\mathrm{ID}(x)}$ refers to the variable at date (t) for weather station number (x). Missing variable values are denoted by NaN.

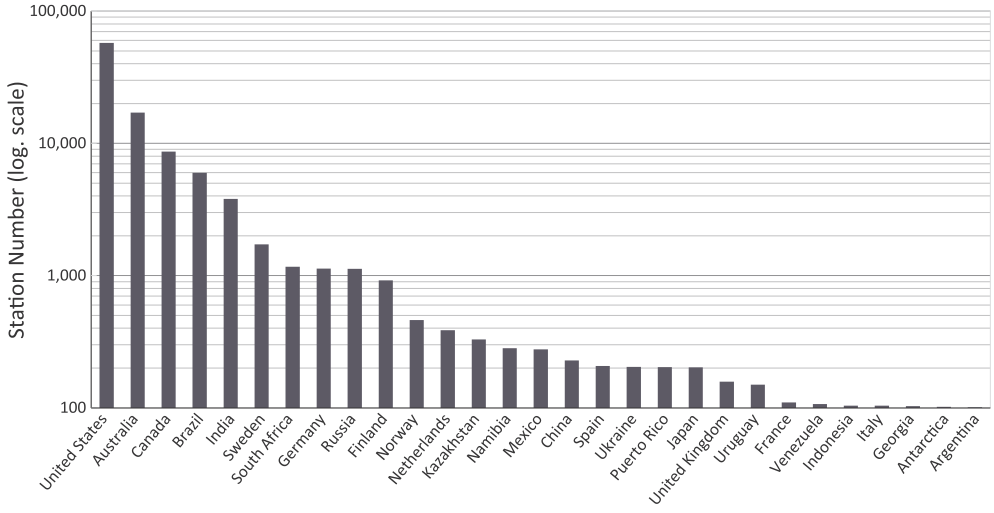

Fig. 4. Total number of weather stations per country or territory contained in the GHCN-Daily database. All sources with at least 100 weather stations are listed. Station numbers are shown in logarithmic scale.

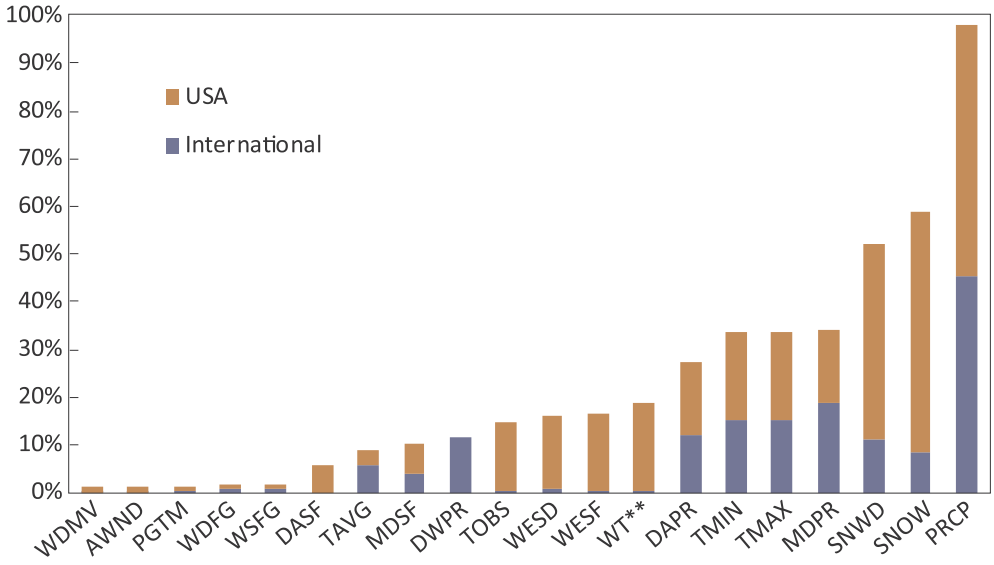

Fig. 5. Prevalence of a subset of GHCN-Daily variables (%) in weather station records within the United States and globally. All variables contained by at least 1% of weather stations are shown.

represented by fewer than five weather stations.

The GHCN-Daily database comprises 57 variables (Table 2). The prevalence of variables monitored by at least 1% of the stations is given in Fig. 5. Most variables (42) are recorded by less than 2% of weather stations contained within GHCN-Daily. Precipitation is the most commonly registered variable (98.1%) in the GHCN-Daily database, while about 33.4% gather TMIN and TMAX (Fig. 5). Fellow core variables (cf. Section 2.1) SNOW and SNWD are logged by more than half the weather stations. However, the vast majority of these observations are limited to the United States (Fig. 6), where SNOW and SNWD are

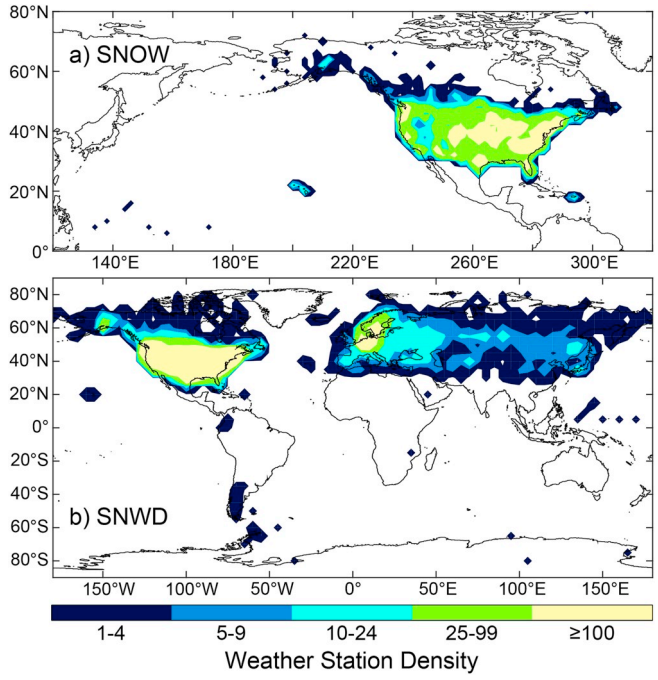

Fig. 6. GHCN-Daily weather station network density in 2017 for a) SNOW (2° grid) and b) SNWD (5° grid) data.

measured by 93.2% and 75.5% of their weather stations, respectively (Fig. 5).

Very limited data are available prior to 1900, with recordings increasing significantly from about 1890 onwards (Fig. 7a–c). These early records predominantly derive from North America, Australia, Europe and South Africa (Fig. 1a), while weather station data for other regions are generally restricted to the post-1950 period (Fig. 1c–f). The record length for most weather stations and data types tends to be less than 30 years (Fig. 7d–f). Manually observed day-time (ACSH) and 24-hour (ACMH) cloudiness time series have the highest median record length (26.8 and 21.8 years, respectively), followed by temperature variables TAVG (average temperature: 22.6 years), TMIN (21.1 years) and TMAX (20.0 years). In comparison, the median for PRCP is 12.8 years of activity.

The average record length of supplied weather station data exhibits notable regional variability, as shown for PRCP data (Fig. 8). Asian archives tend to have relatively lengthy time series (≥ 30 years, Fig. 8d–f), while records in the Americas are more commonly of short duration (< 30 years, Fig. 8a–c). Nevertheless, countries including the United States and Brazil boast numerous time series that exceed 50 years of activity, along with Australia, India, Russia, South Africa and northern European countries. The spatial pattern of Russia's long-term weather stations closely matches the route of the Trans-Siberian Railway (Fig. 8e).

4. Discussion

While elaborate internal quality checks are routinely conducted for the GHCN-Daily database (Menne et al., 2012b), these are not as extensive as quality checks performed by some of the source agencies (e.g. Australia's Bureau of Meteorology [BOM]) that also benefit from local knowledge. As the original quality flags from the source agencies are not included in the GHCN-Daily database (Jaffrés et al., 2018), it is likely that some of the GHCN-Daily data are of poor quality even when no quality flag (Table 3) was attributed to those data. For instance, BOM flagged all TMIN data until 1863 of station ASN00074128 as suspect, with many values below the official record low of their respective month. Conversely, these observations do not include a quality flag in GHCN-Daily. The incorrect assumption of accurate data may significantly impact on research results, especially when affected data cover an extended period and are consistently biased. Systematic instrument errors, changes in measurement practises and station relocation (Jarraud, 2008; Torok and Nicholls, 1996) are among well-known issues impacting on data quality and consistency (homogeneity) when undetected or disregarded. Thus, diligent review of any unexpected data patterns (e.g. trends, shifts and outliers) is vital to increase confidence that results are genuine and not artefacts resulting from flawed data.

Ideally, the GHCN-Daily database should also incorporate all original quality flags and their description. However, a uniform incorporation of pre-existing quality information is unfeasible, as the rigour of quality assessments and the definitions and detail of quality flags are very diverse among the various suppliers to GHCN-Daily. Realistically, a more simplistic approach is likely necessary. For instance, an additional flag value could be introduced to the GHCN-Daily quality section (Table 3) that identifies any data found suspect by the source agency (but passed all internal GHCN-Daily tests). As such, the dimensions of the individual *.dly* files would remain unchanged. However, the inclusion of such information would entail the collaboration of the data suppliers. This external quality flag provision demands a uniform format for the seamless integration with the GHCN-Daily database (e.g. a value of 1 if the data raised any concerns). Hence, a substantial rise in resources would be required from data providers with digital quality records, as 1) flagged data need to be identified, 2) flags require conversion to a standardised value, and 3) data transmission volumes would increase to include the quality flags. Not all agencies have the means to provide quality information with their data. In addition, the rigour of external data evaluation is very disparate. Hence, the incorporation of external quality flags might introduce additional bias because of the heterogeneity (availability and thoroughness) of these original assessments.

High quality records with limited temporal gaps are also crucial for many research fields. For instance, uninterrupted or near-complete daily data are required for evaluation of long-term climate patterns (Rajah et al., 2014; Wang et al., 2017), bushfire risk analysis (Lucas, 2010) and antecedent soil moisture estimation for streamflow studies (Woldemeskel and Sharma, 2016). In GHCN-Daily, the median temporal continuity of PRCP data is 90.7%.

A diverse range of techniques is available to infill temporal data gaps (e.g. Creutin and Obled, 1982; Eischeid et al., 2000). These can be broadly categorised into 1) methods that only consider data within the temporal record and 2) procedures that incorporate auxiliary observations from nearby sites. Among the simplest approaches is linear interpolation based on the values immediately preceding and following the gap. Several factors need to be considered when selecting a suitable gap-filling method to ensure that derived values closely represent reality: Appropriate treatment of missing data depends on the variable type. For example, the heterogeneity of precipitation commonly warrants different management compared to spatially and temporally more uniform variables. Further, the nature (duration and quantity) of the temporal data gaps, climate (geographic position and seasonal characteristics) and availability of ancillary data sources may also influence the choice of procedure.

GHCN-Daily encompasses a diverse range of variables and regions. However, only a subset of all daily, land-based weather station data is contained in GHCN-Daily. Various reasons are given for limited participation by potential source agencies in making data freely available and contributing to projects like GHCN. These motives include technological, financial and proprietary access restrictions (Martin et al., 2015; Page et al., 2004; Thorne et al., 2017). Among the issues hampering contribution to global data repositories are non-digitised data (Martin et al., 2015; Page et al., 2004). Except for the five core elements (Section 2.1), associated multi-day variables and TAVG, sources for most variables are restricted to agencies affiliated with the United

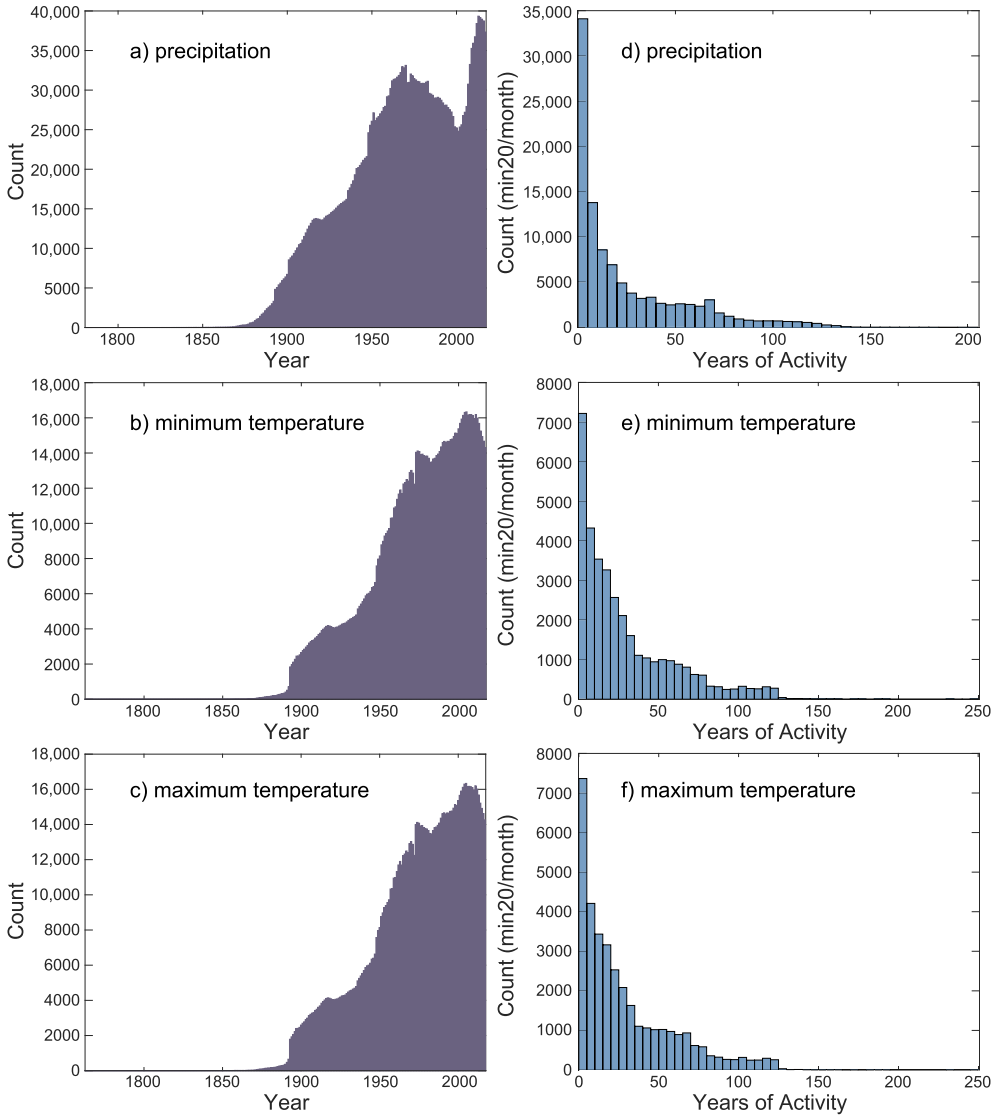

Fig. 7. GHCN-Daily data availability of a,d) precipitation, b,e) minimum and c,f) maximum temperature. Figures a-c list the number of active weather stations per year. Figures d-f tally the years of activity of individual weather stations based on the number of months with a minimum of twenty observations (min20/month).

States. Interestingly, GHCN-Daily also comprises snow variables from regions that do not experience any snowfall (i.e. constant 0 mm snow), including Pacific island nations (Federated States of Micronesia, Palau and Marshall Islands) that are freely associated with the United States.

Several variables are conspicuous by their complete absence in GHCN-Daily (e.g. relative humidity and pressure measurements), even though a substantial number of weather stations do observe these variables. For instance, while most Australian weather stations only record precipitation, in agreement with the data availability pattern seen in GHCN-Daily, some stations register additional variables, including relative humidity, wind speed, mean sea level pressure and evaporation (Jaffrés et al., 2018). Yet, supply to GHCN-Daily from Australia's BOM, one of the largest and most regular contributors to the database, is restricted to daily and multi-daily temperature and precipitation variables.

Both water vapour (humidity) and pressure are among surface data classified as essential climate variables (Bojinski et al., 2014; GCOS, 2016). Alternative sources for these variables exist, although not necessarily at the daily timescale and in point-based (non-gridded) format. The International Surface Pressure Databank (ISPD) constitutes the world's largest collection of pressure observations (Cram et al., 2015). ISPD includes data from marine and terrestrial stations, as well as tropical cyclone best track pressure reports (Cram et al., 2015). However, ISPD is not updated as frequently as GHCN-Daily. A major data source for ISPD is the Integrated Surface Database (ISD; Smith et al., 2011). ISD is a global archive containing hourly weather station data and constitutes one of the data sources for GHCN-Daily (source flag W; Table 3). A further ISD-derived product is HadISD, an annually updated dataset providing sub-daily temperature, pressure, humidity, heat stress, cloud cover and wind variables based on a quality-controlled subset of long-term weather stations (Dunn et al., 2016). Efforts are underway to create a more comprehensive land-based international meteorological observation databank (CLIMOD; Thorne et al., 2017) that would enable diverse data access from a centralised source.

5. Conclusions

The GHCN-Daily database offers a vast, rapidly expanding and diverse dataset from globally distributed weather stations. However, the database's data structure (Fig. 2) may potentially discourage the exploration of GHCN-Daily data by novice users. The user-friendly, flexible *ghcnd_access* toolkit allows the straightforward extraction and transformation of any data contained within the GHCN-Daily database using either MATLAB or free open-source software GNU Octave. A

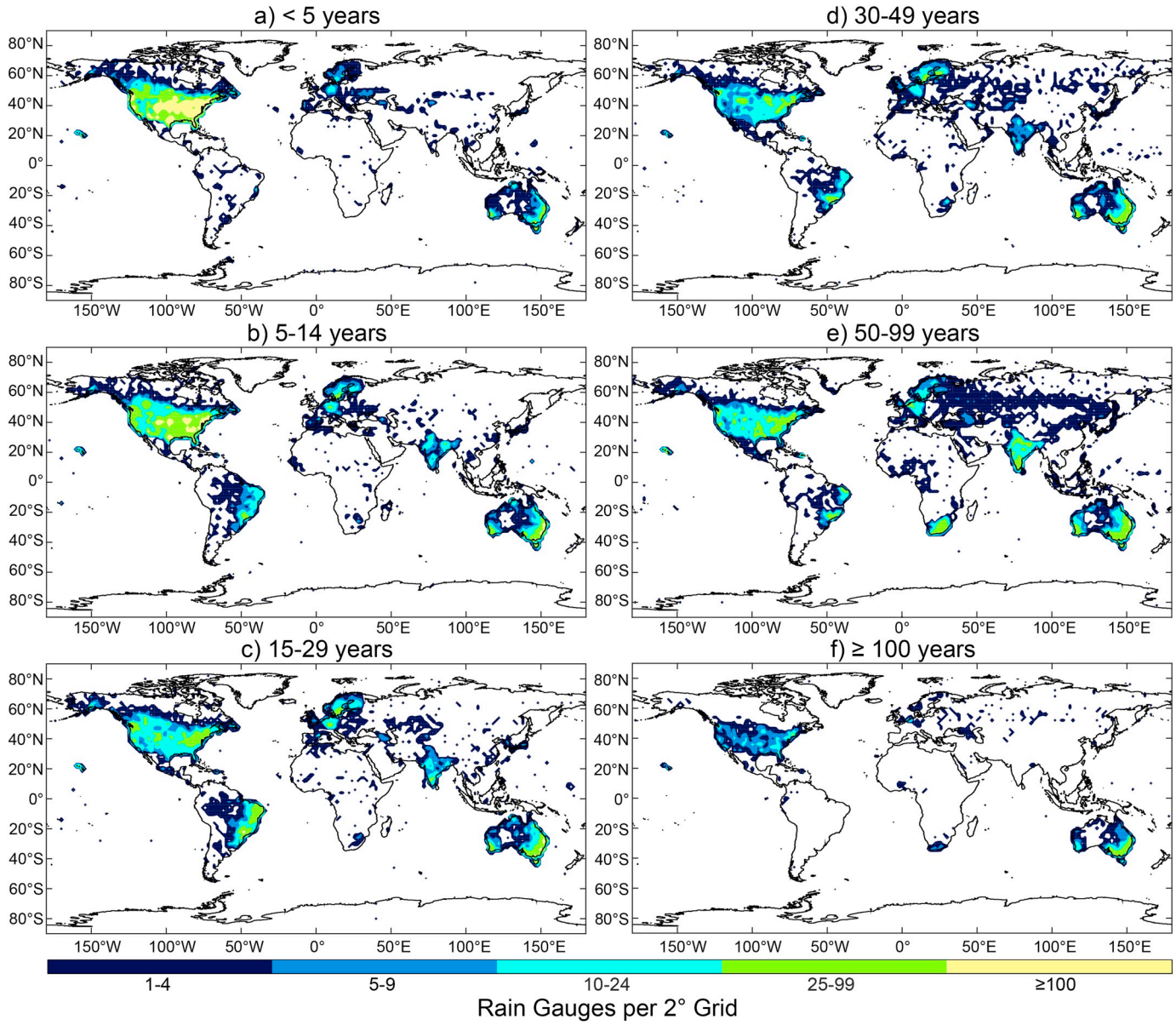

Fig. 8. Number of weather stations per 2° grid cell within the GHCN-Daily database that recorded precipitation for a) less than five years, b) over a period of 5–14 years, c) 15–29 years, d) 30–49 years, e) 50–99 years and f) ≥100 years. The total record length for individual weather stations was calculated based on the number of months with a minimum of twenty observations.

user's guide is supplied with the *ghcnd_access* toolkit that describes the retrieved data format in detail.

The GHCN-Daily database provides significant opportunity to further our understanding of weather and climate variability based on in situ observations. Potential applications are numerous and research topics include 1) variability on multi-decadal timescales and its impact on seasonal and long-term forecasting, 2) teleconnection between remote marine and atmospheric climate signals and terrestrial climate patterns and 3) intricacies of localised weather patterns (e.g. orographic effects). Long-term data records are required to decipher between the various types of variability (e.g. natural vs anthropogenic, multi-decadal vs century-scale) and to evaluate long-term trends. A comprehensive understanding of meteorological data patterns is crucial for effective management of climate variability and change, including drought alleviation, flood risk reduction and agricultural development. In addition to studies focussing on in situ data, the GHCN-Daily database may also be applied in a supportive capacity (e.g. for model calibration and validation).

Computer code availability

The *ghcnd_access* toolbox is utilised to extract data from the Global Historical Climatology Network (GHCN)-Daily database (Menne et al., 2012b) and to change the *.dly* file format into a more accessible structure. The toolbox can be run in either MATLAB or open source alternative GNU Octave. The *ghcnd_access* package is accessible through the GitHub (https://github.com/jjaffres/ghcnd_access/) and SourceForge (https://sourceforge.net/projects/ghcnd-access/) file exchanges. The *ghcnd_access* package includes a comprehensive user's guide (User's Guide.pdf) and a basic guide (readme_toolbox.txt).

Acknowledgements

Global climate data contained within GHCN-Daily, and information on associated weather stations, were obtained from ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/. Coastlines in Figs. 1, 6 and 8 were derived from the Digital Bathymetric Data Base, an ongoing project of the Naval

Research Laboratory. The constructive comments from three anonymous reviewers have greatly contributed to improving the quality of this manuscript.

References

Bojinski, S., Verstraete, M., Peterson, T.C., Richter, C., Simmons, A., Zemp, M., 2014. The concept of essential climate variables in support of climate research, applications, and policy. Bull. Am. Meteorol. Soc. 95 (9), 1431–1443.

Cram, T.A., Compo, G.P., Yin, X., Allan, R.J., McColl, C., Vose, R.S., Whitaker, J.S., Matsui, N., Ashcroft, L., Auchmann, R., Bessemoulin, P., Brandsma, T., Brohan, P., Brunet, M., Comeaux, J., Crouthamel, R., Gleason, B.E., Groisman, P.Y., Hersbach, H., Jones, P.D., Jónsson, T., Jourdain, S., Kelly, G., Knapp, K.R., Kruger, A., Kubota, H., Lentini, G., Lorrey, A., Lott, N., Lubker, S.J., Luterbacher, J., Marshall, G.J., Maugeri, M., Mock, C.J., Mok, H.Y., Nordli, Ø., Rodwell, M.J., Ross, T.F., Schuster, D., Srnec, L., Valente, M.A., Vizi, Z., Wang, X.L., Westcott, N., Woollen, J.S., Worley, S.J., 2015. The international surface pressure databank version 2. Geoscience Data J. 2 (1), 31–46.

Creutin, J.D., Obled, C., 1982. Objective analyses and mapping techniques for rainfall fields: an objective comparison. Water Resour. Res. 18 (2), 413–431.

Dunn, R.J.H., Willett, K.M., Parker, D.E., Mitchell, L., 2016. Expanding HadISD: quality-controlled, sub-daily station data from 1931. Geosci. Instrum. Method. Data Syst. 5 (2), 473–491.

Eaton, J.W., Bateman, D., Hauberg, S., Wehbring, R., 2017. GNU Octave Version 4.2.1 Manual: a High-level Interactive Language for Numerical Computation.

Eischeid, J.K., Pasteris, P.A., Diaz, H.F., Plantico, M.S., Lott, N.J., 2000. Creating a serially complete, national daily time series of temperature and precipitation for the western United States. J. Appl. Meteorol. 39 (9), 1580–1591.

GCOS, 2016. The Global Observing System for Climate: Implementation Needs. World Meteorological Organization 315 pp.

Jaffrés, J.B.D., Cuff, C., Rasmussen, C., Hesson, A.S., 2018. Teleconnection of atmospheric and oceanic climate anomalies with Australian weather patterns: a review of data availability. Earth Sci. Rev. 176, 117–146.

Jarraud, M., 2008. Guide to Meteorological Instruments and Methods of Observation (WMO-no. 8). World Meteorological Organisation, Geneva, Switzerland 681 pp.

Lawrimore, J.H., Menne, M.J., Gleason, B.E., Williams, C.N., Wuertz, D.B., Vose, R.S., Rennie, J., 2011. An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. J. Geophys. Res.: Atmosphere 116 (D19).

Lucas, C., 2010. On developing a historical fire weather data-set for Australia. Aust. Meteorol. Oceanogr. J. 60, 1–14.

Martin, D.J., Howard, A., Hutchinson, R., McGree, S., Jones, D.A., 2015. Development and implementation of a climate data management system for western Pacific small island developing states. Meteorol. Appl. 22 (2), 273–287.

Menne, M.J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R.S., Gleason, B.E., Houston, T.G., 2012a. Global Historical Climatology Network - Daily (GHCN-daily) (v3.23-upd-2018031804). NOAA National Climatic Data Center, https://doi.org/10.7289/V5D21VHZ.

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E., Houston, T.G., 2012b. An overview of the global historical climatology network-daily database. J. Atmos. Ocean. Technol. 29 (7), 897–910.

Page, C.M., Nicholls, N., Plummer, N., Trewin, B., Manton, M., Alexander, L., Chambers, L.E., Choi, Y., Collins, D.A., Gosai, A., Della-Marta, P., Haylock, M.R., Inape, K., Laurent, V., Maitrepierre, L., Makmur, E.E.P., Nakamigawa, H., Ouprasitwong, N., McGree, S., Pahalad, J., Salinger, M.J., Tibig, L., Tran, T.D., Vediapan, K., Zhai, P., 2004. Data rescue in the southeast Asia and south Pacific region: challenges and opportunities. Bull. Am. Meteorol. Soc. 85 (10), 1483–1490.

Peterson, T.C., Vose, R.S., 1997. An overview of the global historical climatology network temperature database. Bull. Am. Meteorol. Soc. 78 (12), 2837–2850.

Rajah, K., O'Leary, T., Turner, A., Petrakis, G., Leonard, M., Westra, S., 2014. Changes to the temporal distribution of daily precipitation. Geophys. Res. Lett. 41 (24), 8887–8894.

Smith, A., Lott, N., Vose, R., 2011. The integrated surface database: recent developments and partnerships. Bull. Am. Meteorol. Soc. 92 (6), 704–708.

Thorne, P.W., Allan, R.J., Ashcroft, L., Brohan, P., Dunn, R.J.H., Menne, M.J., Pearce, P.R., Picas, J., Willett, K.M., Benoy, M., Bronnimann, S., Canziani, P.O., Coll, J., Crouthamel, R., Compo, G.P., Cuppett, D., Curley, M., Duffy, C., Gillespie, I., Guijarro, J., Jourdain, S., Kent, E.C., Kubota, H., Legg, T.P., Li, Q., Matsumoto, J., Murphy, C., Rayner, N.A., Rennie, J.J., Rustemeier, E., Slivinski, L.C., Slonosky, V., Squintu, A., Tinz, B., Valente, M.A., Walsh, S., Wang, X.L., Westcott, N., Wood, K., Woodruff, S.D., Worley, S.J., 2017. Towards an integrated set of surface meteorological observations for climate science and applications. Bull. Am. Meteorol. Soc. 98 (12), 2689–2702.

Torok, S.J., Nicholls, N., 1996. A historical annual temperature dataset for Australia. Aust. Meteorol. Mag. 45 (4), 251–260.

Vose, R.S., Schmoyer, R.L., Steurer, P.M., Peterson, T.C., Heim, R., Karl, T.R., Eischeid, J.K., 1992. The Global Historical Climatology Network: Long-term Monthly Temperature, Precipitation, Sea Level Pressure, and Station Pressure Data, Oak Ridge National Lab., TN (United States). Carbon Dioxide Information Analysis Center, Oak Ridge, TN.

Wang, K., Zhang, T., Zhang, X., Clow, G.D., Jafarov, E.E., Overeem, I., Romanovsky, V., Peng, X., Cao, B., 2017. Continuously amplified warming in the Alaskan Arctic: implications for estimating global warming hiatus. Geophys. Res. Lett. 44 (17), 9029–9038.

Woldemeskel, F., Sharma, A., 2016. Should flood regimes change in a warming climate? The role of antecedent moisture conditions. Geophys. Res. Lett. 43 (14), 7556–7563.