

• Data Description Article •

HadISDH.extremes Part I: A Gridded Wet Bulb Temperature Extremes Index Product for Climate Monitoring[※]

Kate M. WILLETT*

Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, U.K.

(Received 18 November 2022; revised 17 April 2023; accepted 24 April 2023)

ABSTRACT

HadISDH.extremes is an annually updated global gridded monthly monitoring product of wet and dry bulb temperature-based extremes indices, from January 1973 to December 2022. Data quality, including spatial and temporal stability, is a key focus. The hourly data are quality controlled. Homogeneity is assessed on monthly means and used to score each gridbox according to its homogeneity rather than to apply adjustments. This enables user-specific screening for temporal stability and avoids errors from inferring adjustments from monthly means for the daily maximum values. For general use, a score (HQ Flag) of 0 to 6 is recommended. A range of indices are presented, aligning with existing standardised indices. Uniquely, provision of both wet and dry bulb indices allows exploration of heat event character — whether it is a “humid and hot”, “dry and hot” or “humid and warm” event. It is designed for analysis of long-term trends in regional features. HadISDH.extremes can be used to study local events, but given the greater vulnerability to errors of maximum compared to mean values, cross-validation with independent information is advised.

Key words: wet bulb temperature, heat extremes, climate monitoring

Citation: Willett, K. M., 2023: HadISDH.extremes Part I: A gridded wet bulb temperature extremes index product for climate monitoring. *Adv. Atmos. Sci.*, **40**(11), 1952–1967, <https://doi.org/10.1007/s00376-023-2347-8>.

Dataset Profile

Dataset Title	HadISDH.extremes: Met Office Hadley Centre led International Surface Dataset for Humidity heat extremes product
Time range	January 1973 to December 2022 – updated annually
Geographical scope	Global, 5° by 5° gridded
Data format	netCDF4
Data volume	5–9Mb from CEDA, 60–80 Mb from HadOBS
Data service system	HadOBS: www.metoffice.gov.uk/hadobs/hadisdh , CEDA: https://catalogue.ceda.ac.uk/uuid/2d1613955e1b4cd1b156e5f3edbd7e66
Sources of funding	This work and its contributors (Kate WILLETT) were supported by the UK–China Research & Innovation Partnership Fund through the Met Office Climate Science for Service Partnership (CSSP) China as part of the Newton Fund.
Dataset composition	The dataset contains a netCDF file for each heat index. Each file includes monthly gridbox fields of: actual values; anomalies relative to the 1991–2020 reference period; standard deviations of all station values within the gridbox; 1991–2020 climatologies and climatological standard deviations; climatological mean number of stations contributing to the gridbox; actual number of stations contributing to each gridbox month; and homogeneity quality (HQ) flags 1 to 8 and HQ score combining flags 1 to 7 for filtering out poorer quality gridboxes.

1. Background and summary

The wet bulb temperature (T_w) relates to heat stress (Sherwood and Huber, 2010; Schär, 2016) as a proxy for temperature regulation through evaporative cooling from sweating. The T_w is the air temperature cooled by evaporation. It

※ This paper is a contribution to the 2nd Special Issue on Climate Science for Service Partnership China.

* Corresponding author: Kate M. WILLETT
Email: kate.willett@metoffice.gov.uk

depends on the ambient temperature and amount of moisture. The less saturated the air is [<100 %rh relative humidity (RH)], the more evaporation can occur. This lowers the T_w , because evaporation removes energy. When the air is saturated (RH = 100 %rh) the T_w is equal to the dry bulb temperature (T). The difference between T_w and T represents the degree to which evaporative cooling can be effective. In terms of the human body, skin temperature is usually close to 35°C. Hence, when T_w is equal to 35°C, the air close to the skin is saturated and evaporative cooling from sweat can no longer occur. This is then a theoretical critical level above which humans cannot survive (Sherwood and Huber, 2010; Raymond et al., 2020). In fact, physiological tests have shown that the actual critical level can be much lower, even in healthy young adults undertaking moderate activity (Vecellio et al., 2022).

The T_w was historically observed as standard across the global network of weather stations. Since the 1980s, gathering pace in the early 2000s, sensors that measure RH via capacitance have become by far the most common type of humidity measurement (Ingleby et al., 2013). It is now the dewpoint temperature (T_d) or the RH that is reported. In order to have long-term, near-global records, the T_w must be back-calculated from T and T_d . This will be covered in more detail below.

Using T_w alone as a measure of heat stress would not indicate the dry heat events where T is high but T_w is not. It is highly unusual for extremely high T to coincide with extremely high T_w because it is often the lack of available moisture that enables T to increase so much. When T_w is high, the simultaneous T will likely be more moderate. Therefore, looking at T_w may allow for identification of additional “stealth heat events” that are driven by humidity more than temperature, and would otherwise go unidentified. While heat-related mortality appears to correlate as well or better with T compared to indices that include humidity (Armstrong et al., 2019), it is plausible that high humidity and moderate temperatures could reduce physical and mental productivity and lead to non-fatal, yet significant, health impacts.

In the field of climate extremes, there are now many well-established indices that have been devised to assess known elements of the climate that impact society. These are often called the ETCCDI (Expert Team on Climate Change Detection and Indices) indices, now under the remit of the Expert Team on Sector-Specific Climate Indices (ET-SCI; ET-SCI, 2022), and are documented within the Climact project (Climact, 2022) and the HadEX3 gridded extremes indices climate monitoring product (Dunn et al., 2020). These utilise daily maxima and minima T and daily precipitation totals. They characterise a wide range of hot, cold, wet, or dry extremes, durations of hot, cold, wet, or dry periods and exceedance of known impactful thresholds. None use humidity, yet humidity is important for human health and exacerbates stress relating to high temperatures. Several studies have presented T_w threshold exceedance values akin to these ET-SCI indices (Wang et al., 2019; Freychet

et al., 2020; Raymond et al., 2020; Yu et al., 2021) but none have yet provided this as an updating global monitoring product, alongside simultaneous indices for T . Therefore, herein is presented a new range of extremes indices utilising T_w and T , called HadISDH.extremes. Its purpose is the exploration of historical and current risk to human wellbeing and productivity from extreme humid heat. In addition, it supports model validation, which in turn provides confidence in modelled future risk. It is a gridded (5° by 5°) monthly product from January 1973 to December 2022 (at time of writing) and will be updated annually. This suite of humidity and temperature heat extremes indices is built within the framework of the existing Met Office Hadley Centre led International Surface Dataset for Humidity HadISDH.land (Willett et al., 2013, 2014), which is a long-term quality controlled, homogenised, gridded monthly mean land surface humidity monitoring product.

2. Methods

HadISDH.extremes follows the methodology for HadISDH.land (Willett et al., 2013, 2014). The data originate from the National Centers for Environmental Information’s Integrated Surface Dataset (ISD; Smith et al., 2011). These are sub-daily (hourly, 3 hourly, 4 hourly etc.) mandatory surface (1.5–2 m) in-situ observations from weather stations worldwide. These are then processed as part of the HadISD sub-daily monitoring product (Dunn et al., 2016; Dunn, 2019). HadISD selects stations from ISD that have some data between 1 January 1931 to present (March 2023 at time of writing, version 3.3.0.2022f), performs some merging where station series appear to be separated, and removes duplicates. A comprehensive quality control (QC) is conducted to remove apparent outlying values and clusters, repeated strings of values, overly frequent values and physically implausible values. This includes the removal of supersaturated values and persistent false saturation, which would affect the T_w extreme values. False saturation can be a pervasive problem when wet bulb thermometers are used because of the need for the wick surrounding the thermometer to be continually wetted. In dry climates and/or unmanned stations, it is difficult to ensure that the water reservoir does not dry out. The QC also includes a neighbour check to reinstate some of the failed values that have a regional signal and are therefore more likely to be real. Note that no QC can be perfect and that there will always be some good values removed while some bad values will remain. Indeed, the HadISD QC struggled with the record-smashing temperatures in Northwest North America in June/July 2021 and has subsequently been improved (Dunn, 2021a, b, c, d).

To minimise bias from intermittent temporal sampling, HadISDH.extremes concentrates on the relatively data-rich period of 1973 to present, with various checks on temporal sampling consistency as described in Table 1. These reduce the station count compared to HadISD and HadISDH.land. The hourly T and T_d in °C from HadISD are converted to

Table 1. Methodological steps for HadISDH.extremes processing. Note that HadEX3 missing data requirements are for a maximum of 3 days missing per month and 15 days missing per year (Dunn et al., 2020), but this results in ~200 fewer stations for HadISDH.extremes and far more intermittent monthly sampling.

Step	Method	Data completeness requirements
Initial station selection (4425 stations)	Assess data completeness at various levels of processing (all data, climatological months, months, days) and remove any station that fails	(1) > 24 000 observations over the period 1973 to present (2) > 1344 observations for each calendar month over the 1991–2020 climatological period (3) All climatological monthly means present where: (3a) ≥ 15 years of each calendar month with 1 in each decade of the climatological period (1991–2000, 2001–2010, 2011–2020) to make a climatological mean; where: (4a) ≥ 15 observations for each hour of the day over a month to make an hour-month mean, and (4b) ≥ 4 hour-month means with 1 in each 8-hour tercile (0000–0700, 0800–1500, 1600–2300 UTC) to make a monthly mean
Daily extremes	Identify maximum and minimum T and T_w for each day	≥ 4 observations per day with at least one in each 8-hour tercile (0000–0700, 0800–1500, 1600–2300 UTC)
Monthly extremes	Identify maximum, mean and minimum of the daily maxima and minima T and T_w for each month	(1) ≤ 6 days missing per month (2) ≤ 20 days missing per year and ≤ 3 missing months per year (3) ≥ 200 months per station
Monthly exceedances	Calculate percentage of days per month where the daily maximum T and T_w is \geq threshold (relative to days present only)	(1) ≤ 6 days missing per month (2) ≤ 20 days missing per year and ≤ 3 missing months per year (3) ≥ 200 months per station
Daily climatological N th percentiles	Calculate the N th percentile from 1991–2020 collection of daily maxima and minima T and T_w from 5 days spanning across candidate day (± 2 to 4 days used to collect a total of 5 days)	≥ 15 years of 5-day sets with at least 1 in each decade of the climatological period (1991–2000, 2001–2010, 2011–2020)
Monthly N th percentile exceedances	Calculate the percentage of days per month where the daily maxima and minima T and T_w exceed their daily climatological N th percentiles (relative to days present only) – following the methodology of Zhang et al. (2005)	(1) ≤ 6 days missing per month (2) ≤ 20 days missing per year and ≤ 3 missing months per year (3) ≥ 200 months per station
Monthly climatological N th percentiles	Calculate the mean of daily climatological N th percentiles (1991–2020)	(1) ≤ 6 daily climatological N th percentiles missing per month (2) 0 month climatological N th percentiles missing per year
Homogeneity testing	Use HadISDH.land monthly mean T and T_w homogenization information to create homogeneity quality scores for each gridbox (see text)	Stations with any large absolute adjustments are removed ($> 5^\circ\text{C}$ for T , $> 5^\circ\text{C}$ for T_d , $> 3 \text{ g kg}^{-1}$ for q and $> 15 \text{ \%rh}$ for RH), resulting in 4138 stations

hourly T_w in $^\circ\text{C}$ using the Stull (2011) formula:

$$T_w = \left[T \times \tan^{-1} \left(0.151977(\text{RH} + 8.313659)^{1/2} \right) + \tan^{-1} (T + \text{RH}) - \tan^{-1} (\text{RH} - 1.676331) + \left[0.00391838\text{RH}^{2/3} \times \tan^{-1} (0.023101\text{RH}) \right] - 4.686035 \right] \quad (1)$$

This equation is a close approximation to the true T_w , optimised for a surface pressure (P) of 1013 hPa, and valid over the range of RH = 5 to 99 %rh and $T = -20^\circ\text{C}$ to 50°C . Its reported error ranges from -1°C to 0.65°C , with negative

errors when $T < 20^\circ\text{C}$ and positive errors when $T > 20^\circ\text{C}$, generally. Stull (2011) showed that varying the pressure by 200 hPa makes negligible difference for RH $> 50 \text{ \%rh}$. Differences increase up to $\sim 1^\circ\text{C}$ for RH $\leq 10 \text{ \%rh}$. Equation (1) is not valid for very cold and low humidity conditions. Although its valid range does not extend to RH = 100 %rh, the bias here is very small at -0.2°C of T_w when $T = -20^\circ\text{C}$, and 0.1°C of T_w at $T = 30^\circ\text{C}$. It reduces the large positive biases in T_w considerably in the very low RH high T conditions that were found in HadISDH.landTw as part of developing the HadISDH.extremes dataset, where HadISDH.landTw used a different T_w formula (Jensen et al., 1990; Willett

et al., 2014).

The RH is calculated by first calculating the vapour pressure (e) and saturated vapour pressure (e_s) using the Buck (1981) formulae for e_w (with respect to water) or e_i (with respect to ice), in hPa:

$$e_w = 6.1121f \exp\left(\frac{\left(18.729 - \left(\frac{T_d}{227.3}\right)\right)T_d}{257.87 + T_d}\right), \quad (2)$$

where

$$f = 1 + 7 \times 10^{-4} + 3.46 \times 10^{-6}P,$$

$$e_i = 6.1115f \exp\left(\frac{\left(23.036 - \left(\frac{T_d}{333.7}\right)\right)T_d}{279.82 + T_d}\right), \quad (3)$$

in which

$$f = 1 + 3 \times 10^{-4} + 4.18 \times 10^{-6}P.$$

To calculate e_s , the same equations are used but substituting T for T_d . The 1991–2020 climatological mean surface P for the nearest (to the station) 1° by 1° gridbox monthly mean from the ERA5 reanalysis (Hersbach et al., 2020) is used for P , in hPa. Ideally, the simultaneously observed station P would be used but this is not always present nor of high enough quality. The contribution of P is so small that for the likely range of actual P values (not shown) this substitution makes negligible difference to the calculated e , e_s , RH and T_w . RH is calculated using the standard formula:

$$\text{RH} = \left(\frac{e}{e_s}\right) \times 100. \quad (4)$$

To determine whether to calculate e_i or e_w , Eq. (2) with respect to water is used first to calculate e and e_s , followed by a provisional RH using Eq. (4), and then a provisional T_w using Eq. (1). If the provisional $T_w \leq 0^\circ\text{C}$ then e and e_s are recalculated using Eq. (2) for e_i and then used to recalculate RH and T_w . Using e_w when an ice bulb should theoretically be present results in small positive biases in calculated e , RH and T_w .

It is unfortunate that the data must be converted from initial measurement (mostly RH or T_w) to reported value (T_d), and then again from reported value to T_w , as very small errors are likely to result from the process. It is not possible to associate individual observations with the instruments that made them at present and the assumption is made that these errors are random and very small compared to other sources of error (incorrect measurements, inhomogeneities etc.) and should therefore average out.

The hourly values are then processed into daily means, maxima and minima and various monthly statistics of daily extremes. The nature of extremes is quite different to that of the mean, and the values are far more sensitive to missing

data. For this reason, a more conservative approach on data completeness is necessary here than for HadISDH.land. It largely follows that of the HadEX3 datasets of indices (Dunn et al., 2020), with loosening to balance continuity of temporal coverage in the monthly values with biases stemming from incomplete daily sampling. Missing hours and days of observations can be more of a problem for humidity than temperature because the instruments are more prone to both instrument and reporting errors (Willett et al., 2013, 2014). Table 1 lists the dataset processing stages from hourly data to monthly extremes indices and the data completeness requirements along the way.

Importantly, HadISDH.extremes will most likely be an underestimate of the true daily maxima and minima to some degree. This is because it is based on hourly, and in many cases 3 hourly or less frequent observations. To minimise this error there must be at least one observation in each tercile of the day (0000–0759, 0800–1559, 1600–2359 UTC).

The calculation of the indices themselves again follows the HadEX3 methodology, specifically accounting for biases in exceedance of percentile thresholds within the climatological period by following the Zhang et al. (2005) bootstrapping methodology. The HadISDH.extremes data selection and processing narrows the dataset to 4460 stations. The geographical coverage and mean station density per gridbox month over the 1991–2020 climatological period are shown in Fig. 1.

HadISDH.extremes contains a wide suite of indices that largely follow, and are therefore comparable with, the ET-SCI temperature indices (Table 2). It is important to capture both the values of the extremes and how they are changing over different regions, as well as days of exposure to various high T_w levels. A range of set threshold exceedances are used from 25°C to 35°C that encompass those used by Raymond et al. (2020). The minimum T_w experienced across a month/region is also of interest as this relates to the ability to recover overnight from daytime maxima, so equivalent minimum T_w extremes are also included. A suite of equivalent T indices with a set of threshold exceedances from 25°C to 50°C are also included. Calculation of the percentile indices requires percentile thresholds for a climatological period, which here is 1991–2020. This period is consistent with HadISDH.land and compatible with the period of record. Unfortunately, this is not then directly comparable with indices from HadEX3, which uses the periods 1961–1990 and 1981–2010 [see Dunn and Morice (2022) for a discussion of the effect of reference period on the comparability of percentile-based indices]. Furthermore, note that HadEX3 TX (and related indices) mostly originate from the actual maximum daily temperature, which is a standard measurement. HadISDH.extremes TX is the maximum of the hourly (or in many cases 3 hourly or less frequent) observations.

Not all indices are applicable to the entire globe because many regions have never come close to the high-end extremes. For example, according to HadISDH.extremes data the UK has not yet exceeded a T_w above 31°C , while China has. However, regional acclimatisation is impor-

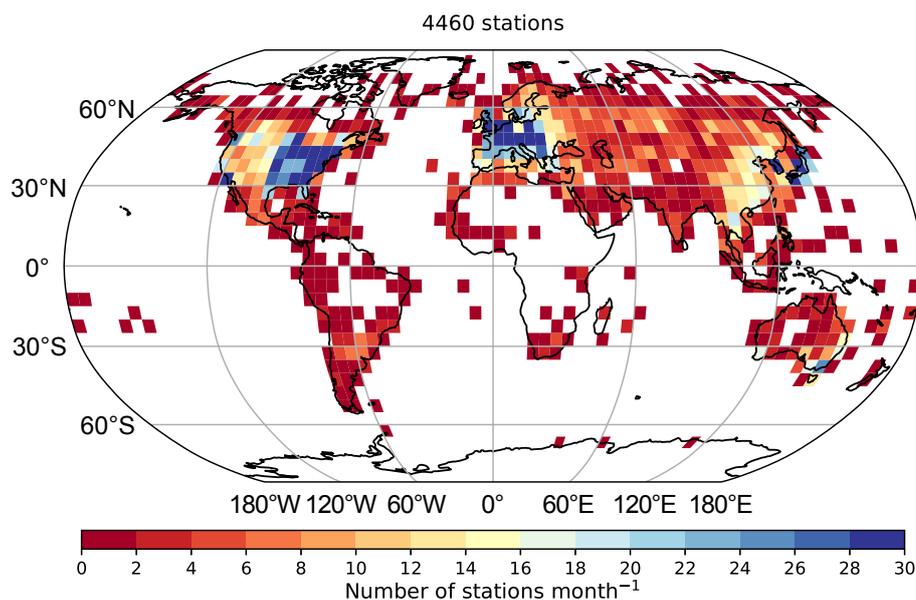


Fig. 1. Total number of stations contributing to the HadISDH.extremes dataset and their geographical location and mean station density per month over the 1991–2020 climatological period for each 5° by 5° gridbox.

tant for humans, animals and plants, so an increase in extreme values, even if well below the theoretically critical level of 35°C, could still have significant impacts. Acclimatisation has been less well studied for humidity compared to T but it has been found to be important (Shen and Zhu, 2015).

The monthly extremes index values are then gridded by simple area averaging over each 5° by 5° gridbox. No interpolation is done to infill regions of missing data. Both anomaly and actual values of the extremes indices are gridded, along with climatological statistics. In most cases, the anomalies are the more robust variable because they are less biased by station location (elevation, local environment etc.) and more resilient to missing data. Using anomalies where possible rather than actual values is recommended. The T_w XX and TXX indices are not available as anomalies both because there is limited value in such a quantity and because these are the maximum daily maximum observation from within the gridbox rather than the mean over several stations, and therefore highly variable. These indices are highly uncertain and should be used with caution.

As with all long-term monitoring products, it is very important that there are no instabilities in the dataset due to changes in the observing system. This provides confidence that any trends present are signals of the climate rather than artefacts of observing system changes. Such inhomogeneity can arise when stations are moved or instrument types changed. For T_w , the move to RH and dewpoint sensors from the 1980s onwards is particularly problematic. This change was often made at a country level over a short period of time, and so can be difficult to detect with traditional neighbour comparison methods (Ingleby et al., 2013). In China, for example, the move from manual wet bulb ther-

mometers to automated RH sensors was made in the early 2000s (Freychet et al., 2020; Li et al., 2020). This also coincided with an increase from 3 hourly observations to hourly. Freychet et al. (2020) show that when unadjusted this results in a decrease in the RH over this period and therefore an underestimate in T_w trends. They note that this inhomogeneity is still apparent in ERA5 and the HadISDH.land monitoring product. HadISDH.land uses an automated neighbour comparison approach to homogenisation (Willett et al., 2014). Pairwise homogenisation (PHA; Menne and Williams, 2009) is performed on the monthly mean station values of T and dewpoint depression. A matrix of highly correlating neighbours is used to detect breakpoints and apply adjustments. The combined dates of detected breakpoints for each station from both variables are then used to apply further adjustments to T and T_w using the highly correlating neighbour networks as a reference series (Indirect PHA; Willett et al., 2014). Although the PHA approach can utilise known dates of changes to stations and observing practices, this information is not available digitally for the vast majority of stations and is more often not documented at all. Neither is a manual approach possible with 4000+ stations. This method has been found to be effective at removing the large inhomogeneities in the data but is less powerful when it comes to region-wide inhomogeneities.

Hourly or sub-daily resolution data are essential to producing extremes indices. However, detecting and adjusting for inhomogeneities, although not impossible (Brugnara et al., 2023), is very difficult to do on daily data, let alone hourly data. Inhomogeneities typically affect both the variability and the mean. Changes in station measurements can be different depending on the time of day or year or even background weather. Undertaking this for 4000+ stations and

Table 2. HadISDH.extremes index descriptions. *Italicised indices are those for which only actual values are provided. All others are available as actual and anomaly values.*

Index name	Index long name	Index description (NB: in all cases, climatological refers to the period 1991–2020)
T_wX	Maximum wet bulb temperature	Gridbox mean of station month maxima of daily maximum T_w
T_wN	Minimum wet bulb temperature	Gridbox mean of station month minima of daily minimum T_w
T_wX90p	90th percentile maximum wet bulb temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T_w exceeds the climatological 90th percentile of daily maxima
T_wM90p	90th percentile mean wet bulb temperature exceedance	Gridbox mean of station percentages of days per month where the daily mean T_w exceeds the climatological 90th percentile of daily means
T_wN10p	10th percentile minimum wet bulb temperature exceedance	Gridbox mean of station percentages of days per month where the daily minimum T_w is lower than the climatological 10th percentile of daily minima
T_wM10p	10th percentile mean wet bulb temperature exceedance	Gridbox mean of station percentages of days per month where the daily mean T_w is lower than the climatological 10th percentile of daily means
T_wX25	25°C maximum wet bulb temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T_w is equal to or exceeds 25°C
T_wX27	27°C maximum wet bulb temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T_w is equal to or exceeds 27°C
T_wX29	29°C maximum wet bulb temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T_w is equal to or exceeds 29°C
T_wX31	31°C maximum wet bulb temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T_w is equal to or exceeds 31°C
T_wX33	33°C maximum wet bulb temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T_w is equal to or exceeds 33°C
T_wX35	35°C maximum wet bulb temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T_w is equal to or exceeds 35°C
T_wXX	<i>Maximum maximum wet bulb temperature</i>	<i>Gridbox maximum of station month maxima of daily maximum T_w</i>
TX	Maximum temperature	Gridbox mean of station month maxima of daily maximum T
TN	Minimum temperature	Gridbox mean of station month minima of daily minimum T
TX90p	90th percentile maximum temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T exceeds the climatological 90th percentile of daily maxima
TM90p	90th percentile mean temperature exceedance	Gridbox mean of station percentages of days per month where the daily mean T exceeds the climatological 90th percentile of daily means
TN10p	10th percentile minimum temperature exceedance	Gridbox mean of station percentages of days per month where the daily minimum T is lower than the climatological 10th percentile of daily minima
TM10p	10th percentile mean temperature exceedance	Gridbox mean of station percentages of days per month where the daily mean T is lower than the climatological 10th percentile of daily means
TX25	25°C maximum temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T is equal to or exceeds 25°C
TX30	30°C maximum temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T is equal to or exceeds 30°C
TX35	35°C maximum temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T is equal to or exceeds 35°C
TX40	40°C maximum temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T is equal to or exceeds 40°C
TX45	45°C maximum temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T is equal to or exceeds 45°C
TX50	50°C maximum temperature exceedance	Gridbox mean of station percentages of days per month where the daily maximum T is equal to or exceeds 50°C
TN18	18°C minimum temperature exceedance	Gridbox mean of station percentages of days per month where the daily minimum T is equal to or exceeds 18°C
TXX	<i>Maximum maximum temperature</i>	<i>Gridbox maximum of station month maxima of daily maximum T</i>

ensuring that the process has not added more error than it has removed is a significant challenge. For HadISDH.extremes, a quality assurance approach is taken rather than a

homogeneity adjustment approach, whereby gridboxes are given a score depending on their likely level of inhomogeneity. This can then be used to screen or weight gridboxes

Table 3. Homogeneity quality (HQ) scoring system. All scores are applied at the individual gridbox month level. Bold text indicates the worst case for each score. Italicised text indicates critical scores where gridbox months are of extremely poor quality and exclusion is recommended.

Homogeneity variable	Description	Scoring detail	Worst case
HQ 1: Station count	Number of stations	1 station = 1 , >1 stations = 0	1 station only, moderate quality
HQ 2: Inhomogeneity density	Ratio of HadISDH.land inhomogeneity adjustments present per station: count of stations with adjustment present divided by total count of stations	1 per station = 2 , 0–1 per station = 1, 0 per station = 0	1 per station, poor quality
HQ 3: Small inhomogeneity density	Ratio of 0°C–0.5°C inhomogeneity adjustments present per station: count of stations with > 0°C and < 0.5°C absolute adjustment present divided by total count of stations	1 per station = 1 , 0–1 per station = 0, 0 per station = 0	1 per station, moderate quality
HQ 4: moderate inhomogeneity density	Ratio of 0.5°C–1°C inhomogeneity adjustments present per station: count of stations with $\geq 0.5^\circ\text{C}$ and < 1°C absolute adjustment present divided by total count of stations	1 per station = 3 , 0–1 per station = 1–2 (scaled linearly), 0 per station = 0	1 per station, poor quality
HQ 5: Large inhomogeneity density	Ratio of 1°C–2°C inhomogeneity adjustments present per station: count of stations with $\geq 1^\circ\text{C}$ and < 2°C absolute adjustment present divided by total count of stations	1 per station = 5 , 0–1 per station = 1–4 (scaled linearly), 0 per station = 0	1 per station, very poor quality
HQ 6: Very large inhomogeneity density	Ratio of > 2°C inhomogeneity adjustments present per station: count of stations with > 2°C absolute adjustment present divided by total count of stations	1 per station = 10 , 0–1 per station = 1–9 (scaled linearly), 0 per station = 0	<i>1 per station, extremely poor quality, exclusion recommended</i>
HQ 7: Mean adjustment magnitude	Mean adjustment per station: sum of adjustments present divided by total number of stations	$\geq \pm 2^\circ\text{C} = 10$, $\geq \pm 1^\circ\text{C} < \pm 2^\circ\text{C} = 4\text{--}9$ (scaled linearly), $\geq \pm 0.5^\circ\text{C} < \pm 1^\circ\text{C} = 2\text{--}3$ (scaled linearly), $\geq \pm 0^\circ\text{C} < \pm 0.5^\circ\text{C} = 1$, 0°C = 0	$\geq \pm 2^\circ\text{C}$, <i>extremely poor quality, exclusion recommended</i>
HQ 8: Mean absolute adjustment magnitude	Mean absolute adjustment per station: sum of absolute adjustments present divided by total number of stations	No score. Value is an extra indicator of poor quality.	Not applicable.
HQ Flag	Homogenisation Quality Flag: sum of HQ 1:7	$\geq 10 = \text{very poor quality}$, 7–9 = <i>poor quality</i> , 5–6 = moderate quality, 1–4 = good quality, 0 = excellent quality	≥ 7 , <i>very poor quality or worse, exclusion recommended</i>

accordingly for onward analysis. The homogenisation dates and adjustments for the HadISDH.land monthly mean T_w are used to provide homogeneity assurance information for each HadISDH.extremes gridbox month. These are described in Table 3. There are eight homogeneity quality scores and a combined quality flag (HQ Flag). Users can choose whether to use only the good (HQ Flag score = 0) data for analysis, or some other level of quality depending on their desired level of data completeness. This is explored further in section 4. Given the potentially sporadic nature of gridbox month exclusions, users should implement a continuity check or missing data threshold when computing long-term statistics like trends. Selection based on quality will always be a compromise over spatial and temporal completeness of the dataset.

3. Data records

The HadISDH.extremes dataset v1.0.0.2022f is available in netCDF format under the Open Government Licence

(<https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>). The data can be accessed from the Centre for Environmental Data Analysis (CEDA) Archive (Willett, 2023a; <https://catalogue.ceda.ac.uk/uuid/2d1613955e1b4cd1b156e5f3edbd7e66>). It is also available alongside other related material and information from the Met Office Hadley Centre climate monitoring observations portal, HadOBS: www.metoffice.gov.uk/hadobs/hadisdh/.

Additionally, the land-based Met Office Hadley Centre's Integrated Surface Dataset of Humidity HadISDH.land.4.5.1.2022f (Willett et al., 2023) was used within this paper. This is the latest version available at time of writing. It is downloadable as netCDF files from www.metoffice.gov.uk/hadobs/hadisdh under the Open Government Licence (<https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>). It is also available from the CEDA Archive: <https://catalogue.ceda.ac.uk/uuid/8956cf9e31334914ab4991796f0f645a>.

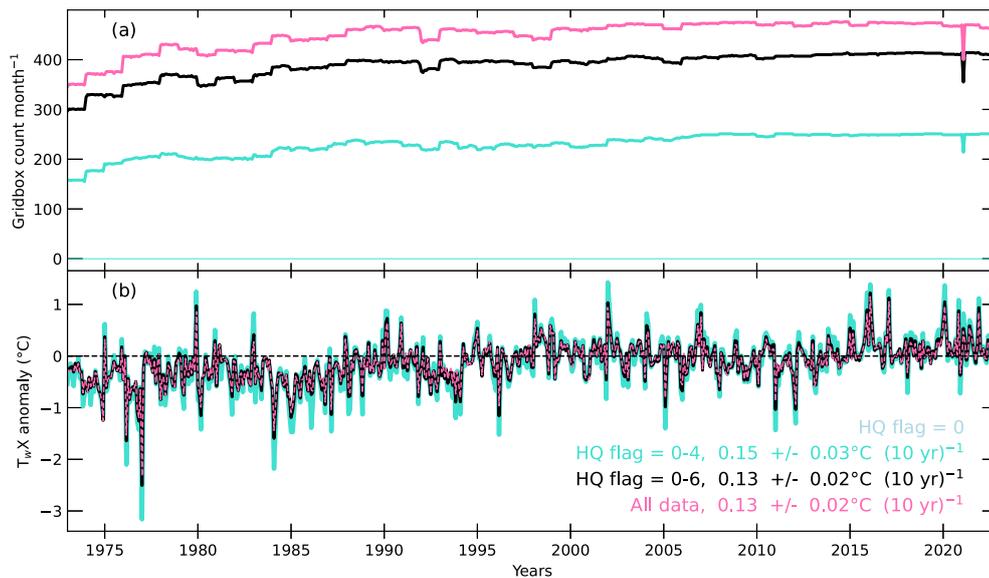


Fig. 2. Comparison of monthly global average time series of (a) gridbox counts and (b) mean T_{wX} anomalies (1991–2020 baseline) for different homogenisation quality score (HQ Flag) filters. Gridboxes have been screened to remove those with lower than 70% completeness over the 600-month record. Global averages are computed using cosine weighting based on the latitude at the centre of each gridbox. Decadal trends in T_{wX} anomalies are shown on panel (b), fitted using ordinary least-squares regression with 90th percentile confidence intervals corrected for AR(1) following Santer et al. (2008). Gridbox count time series for when the HQ Flag = 0 is constantly zero because the subsequent 70% completeness screening removes all gridboxes.

4. Technical validation

Underpinning any climate monitoring product is its long-term stability and “quality”. As discussed in section 2, inhomogeneity in HadISDH.extremes is addressed by allowing the exclusion of gridboxes where inhomogeneity is likely to be an issue. Figure 2 shows monthly gridbox counts and global monthly mean T_{wX} anomalies over time under different selections of HQ Flag scores (Table 3), with additional screening to remove any gridboxes with less than 70% temporal completeness. The “All data” time series (pink) does not exclude any gridboxes based on their underlying inhomogeneity. It shows a significant trend of $0.13 \pm 0.02^{\circ}\text{C} (10 \text{ yr})^{-1}$ for T_{wX} anomalies. The number of gridboxes present per month steadily increases over time until December 2020. The sudden drop in early 2021 is due to a region-wide missing few days of data over Europe and Asia, particularly affecting Northeast Russia. This is in combination with strict missing data criteria for HadISDH.extremes, where more than six missing days results in the entire month being removed. Although the increasing data density over time coincides with the long-term trend in T_{wX} , the two are unlikely to be related for a number of reasons. Firstly, the trend magnitudes of the three different HQ Flag selections shown in Fig. 2 are essentially indistinguishable, given that they are within each other’s 90th percentile confidence ranges. Secondly, the magnitude and variability of the monthly anomalies are very similar, despite the “HQ Flag = 0–4” beginning with almost 50% fewer gridboxes than the others. Thirdly, there is no discernible feature in T_{wX} during 2021 that indi-

cates an impact of the loss of almost 15% of gridboxes. Finally, the trend is in line with other evidence in support of increasing near-surface water vapour content, such as ERA5 trends in near surface specific humidity (Simmons et al., 2021), other reanalyses (Willett et al., 2022) and total column water vapour over land (Mears et al., 2022).

To ensure data quality, excluding gridboxes with an HQ Flag of 7 or greater is recommended. The resulting moderate quality data (HQ Flag = 0–6; black) has around 28 000 fewer gridboxes in total, which is around 50 fewer each month. These data show an almost identical time series and trend [$0.13 \pm 0.02^{\circ}\text{C} (10 \text{ yr})^{-1}$] to the “All data” selection. The largest inhomogeneities are usually the easiest to detect and so there is reasonable confidence that this selection is removing the worst gridboxes. In fact, this makes little difference to the global average. Regional and seasonal averages may be more sensitive. This is explored further below.

A more conservative choice may be to exclude gridbox months where the HQ Flag is 5 or greater (HQ Flag = 0–4; turquoise). This high-quality selection approximately halves the number of gridboxes initially. Gridbox numbers steadily increase over time. Despite the large difference in gridbox numbers, the time series and trends are very similar to the “HQ Flag = 0–6” and “All data” selections, with a trend of $0.15 \pm 0.03^{\circ}\text{C} (10 \text{ yr})^{-1}$ for T_{wX} anomalies. However, given the very low numbers of gridboxes, and tendency for larger variability, this selection is less robust. The gridbox count for an “HQ Flag = 0” selection is zero once gridboxes are subsequently screened for 70% temporal completeness. This is because inhomogeneity is a ubiquitous issue and, addition-

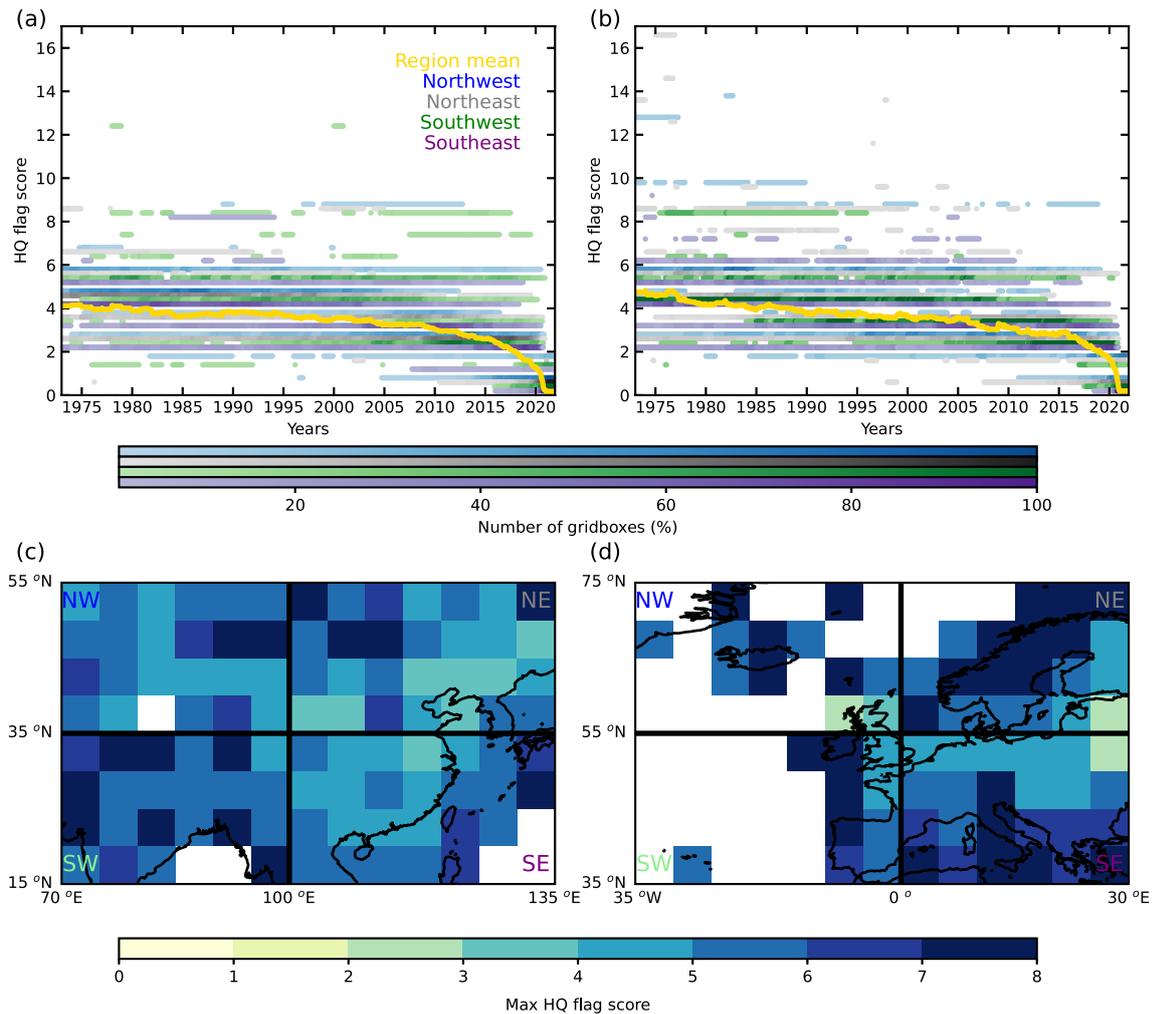


Fig. 3. Regional comparison of HQ Flag scores across the (a, c) China/East Asia region and (b, d) UK/Europe region. Panels (a) and (b) show the percentage of gridboxes with each HQ Flag score for each subregion over time, colour-coded by quadrant (NW, NE, SW, SE), and the regional mean HQ Flag score over time. Each quadrant's set of gridboxes is slightly offset on the y-axis for visibility. Panels (c) and (d) show the maximum HQ Flag score for each gridbox across the region and the quadrant division. Region extents for China/East Asia are (15° – 55° N, 70° – 135° E), and for the UK/Europe are (35° – 70° N, 35° W– 30° E).

ally, the HQ Flag penalises gridboxes with only one station (HQ 1 score, Table 3) because this makes it harder to detect inhomogeneity and presents greater susceptibility to individual station bias and error. Clearly, an “HQ Flag = 0” selection is not advisable. From this analysis, filtering HadISDH. extremes to use only those gridboxes with an HQ Flag = 0–6 appears to be a sensible choice in terms of data density, long-term stability and robustness.

4.1. Regional sensitivity to different HQ Flag score selections

The change from wet bulb thermometers to RH sensors, and from manual to automated observations, has been found to have caused considerable inhomogeneity over China in the early 2000s (Freychet et al., 2020; Li et al., 2020). This is likely true for many other regions. Such region-wide changes are less easily detected with the automated HadISDH.land homogenisation methodology. Figure 3

explores the temporal features of the HQ Flag scores across the regions of China/East Asia and UK/Europe to identify whether any particular time period stands out in terms of identified inhomogeneity. Gridboxes containing stations with a large inhomogeneity detected receive high HQ Flag scores, especially where there is only a single station or all stations have a large inhomogeneity of the same sign (positive or negative). There are several, but not many, occurrences of gridbox months with an HQ Flag score of 7 or greater over the China/East Asia region, mostly in the southwest quadrant, and mostly pre-2000. Partly, this is coming from the gridboxes that also include stations from countries to the southwest (e.g., India and Nepal). There is no noticeable peak in HQ Flag scores over the early 2000s. This supports the idea that HadISDH.land T_w monthly mean homogenisation may be unable to detect the region-wide change in observing practices over China. Visual exploration of several HadISDH.land T_w and RH station time series and their adjustments

does not show any obvious peak or feature around the early 2000s (not shown). Therefore, the HQ Flag score screening of HadISDH.extremes may be less useful over China and users should keep this in mind.

For the UK/Europe region, there are several occurrences of gridbox months with an HQ Flag score of 7 or greater from all quadrants, particularly in the early record and in the north. The HQ Flag score regional mean is very similar to that of the China/East Asia region, decreasing from around 4 over time to near 0 in 2021. There are virtually no gridboxes

with an HQ Flag score of 0 for either the China/East Asia or UK/Europe region until 2005. This demonstrates the pervasiveness of inhomogeneity. Figure 3 further supports excluding gridboxes with an HQ Flag score of 7 or greater, removing the poorest quality observations while maintaining data density.

To explore the effects of HQ Flag selection on regional and seasonal time series, Figs. 4 and 5 show the China/East Asia and UK/Europe regions' JJA seasonal mean time series for maximum daily maximum T_w (T_{wX}) and T (TX)

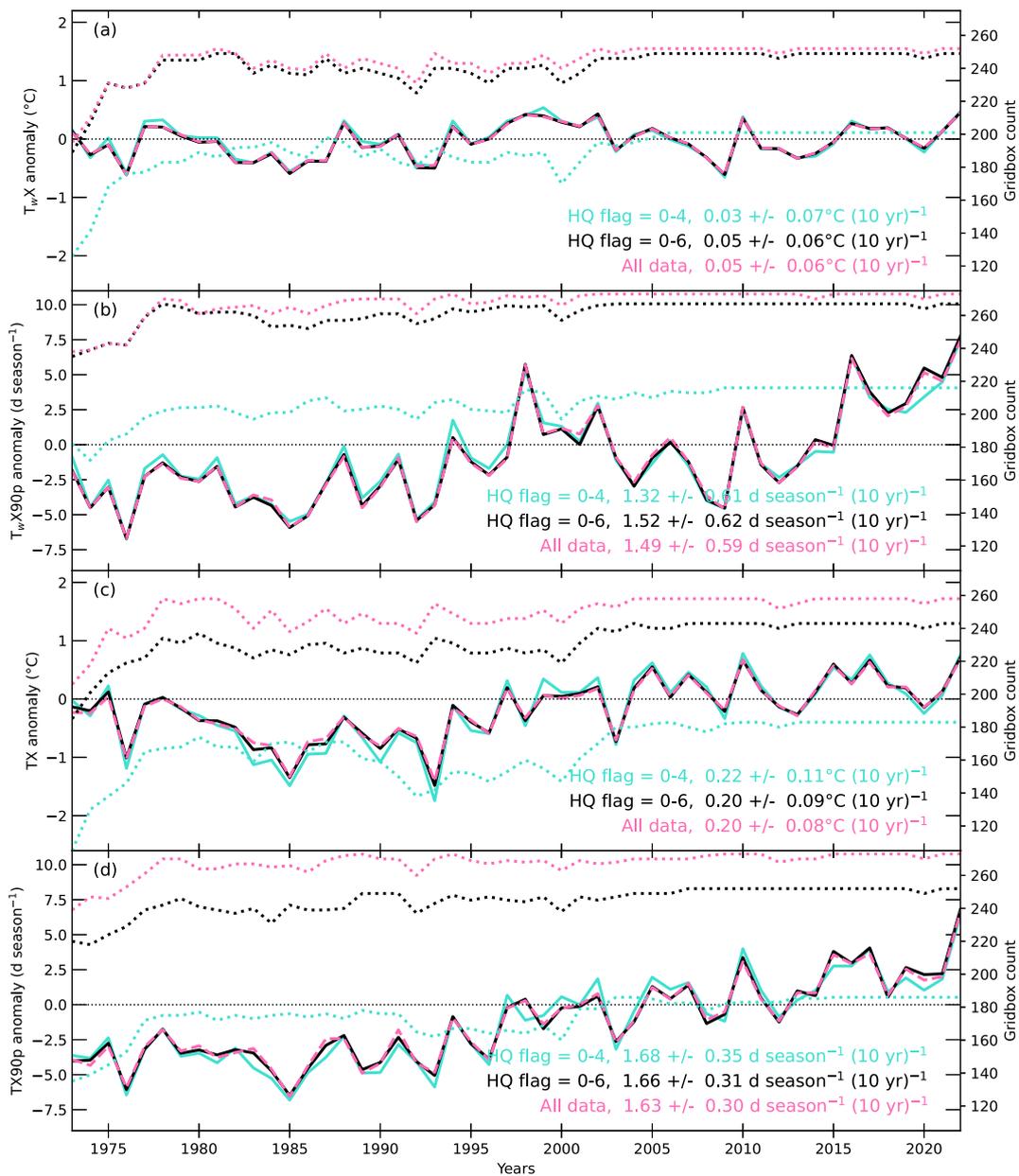


Fig. 4. China/East Asia regional mean JJA mean (a) T_{wX} and (c) TX anomalies in $^\circ\text{C}$, and JJA total (b) T_{wX90p} and (d) TX90p anomalies in d season^{-1} from HadISDH.extremes for different HQ Flag selections. Dotted time series show the annual gridbox counts for each selection relative to the right-hand y-axes. All anomalies are relative to 1991–2020. Regional extents are as presented in Fig. 3 but with gridboxes screened to remove those with lower than 70% completeness over the 600-month record. Decadal trends in anomalies are shown for each HQ Flag selection, fitted using ordinary-least squares regression with 90th percentile confidence intervals corrected for AR(1) following Santer et al. (2008).

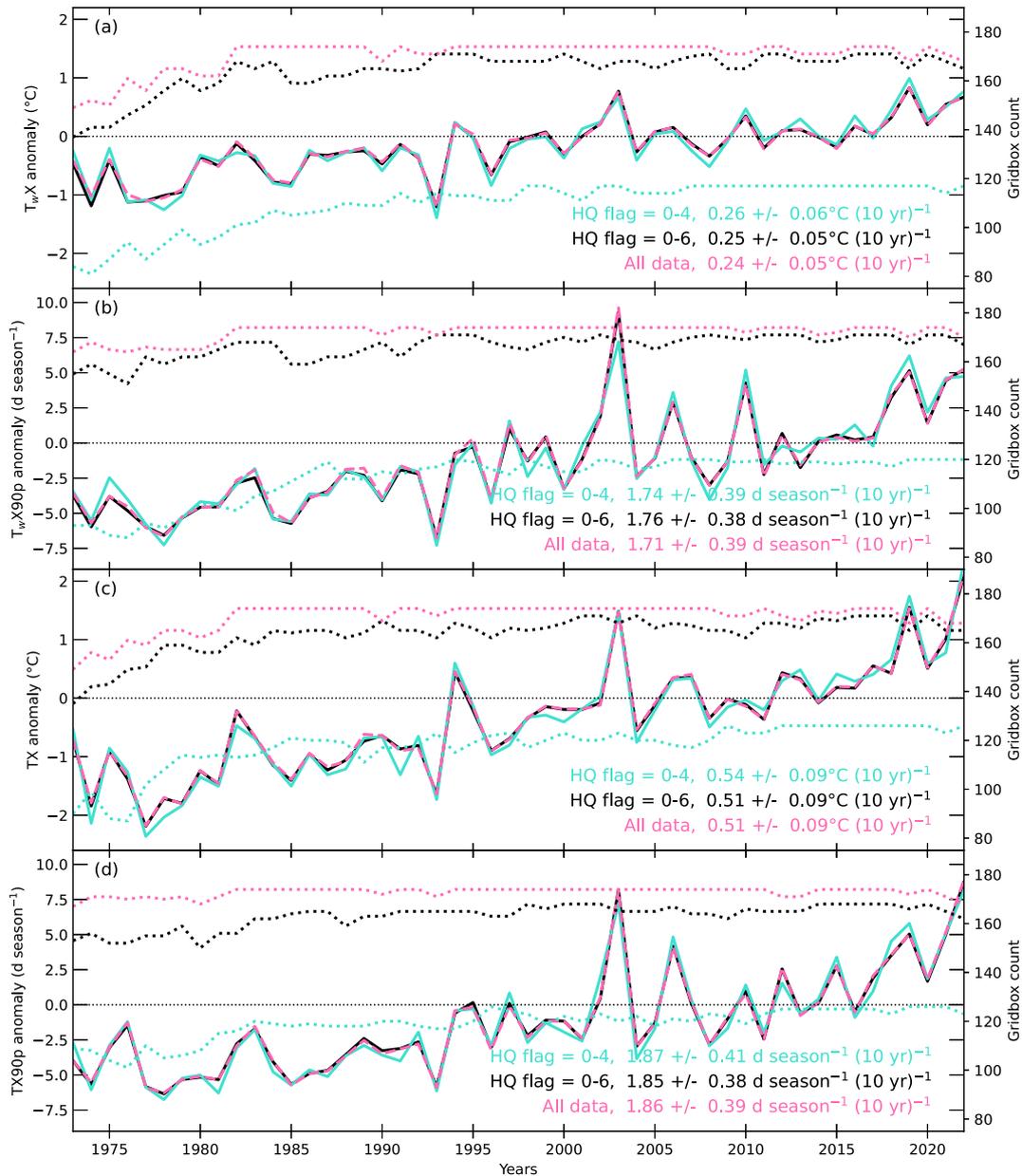


Fig. 5. As in Fig. 4 but for the UK/Europe region.

and the seasonal total days exceeding the 90th percentile of daily maximum T_w (T_wX90p) and T ($TX90p$). These are a measure of the magnitude of peak extremes and frequency of moderate extremes, respectively. For both regions, the selection of gridboxes based on HQ Flag scores, despite resulting in 25%–30% differences in gridbox counts, makes very little difference to the regional mean time series, even at this seasonal level. The “HQ Flag = 0–4” selection results in far poorer coverage and leads to slightly larger variability because the regional mean is then more sensitive to outliers.

The UK/Europe region, of which around a third is ocean, has fewer contributing gridboxes than the China/East Asia region. Although the HQ Flag selection removes fewer gridboxes over the UK/Europe region, the expectation is that a smaller region with fewer gridboxes might be more

affected by reduced coverage. This does not appear to be the case at these scales. However, should analysis be conducted over smaller spatial scales, this is likely to become an issue.

The “All data” selection tends to produce the smallest trends, with the exception of T_wX and T_wX90p over the China/East Asia region and $TX90p$ over the UK/Europe region. The “HQ Flag = 0–4” selection tends to have the largest 90th percentile confidence intervals, which is indicative of its greater variability. However, decadal trends and confidence intervals are indistinguishable across the three selections as they are within a few hundredths of a decimal place of each other.

Although at these scales, from a regional mean perspective, there is little perceived advantage of selecting “HQ

Flag = 0–6” over “All data”, it is still prudent to remove those gridboxes known to contain large uncertainty related to inhomogeneity. The “HQ Flag = 0–6” selection balances removal of the largest inhomogeneities versus maintaining sufficient gridbox density.

To explore the issue of instrument-related inhomogeneity over the China/East Asia region (Freychet et al., 2020) further, the regional mean JJA RH and T_w time series from HadISDH.land, which has been homogenised, are shown in Fig. 6. Freychet et al. (2020) showed that the uncorrected ERA5 JJA mean T_w (their Fig. 4) trend is less than $0.05^\circ\text{C} (10 \text{ yr})^{-1}$ over the period 1979–2017. Corrected T_w trends in ERA5 and their Chinese observational dataset are close to $0.2^\circ\text{C} (10 \text{ yr})^{-1}$. They infer, logically, that uncorrected trends in the extremes of T_w will also be underestimated, and their trends from corrected observational data and ERA5 are indeed significantly positive, also at close to $0.2^\circ\text{C} (10 \text{ yr})^{-1}$. The HadISDH.land T_w trend (Fig. 6a), albeit for a wider region and longer time period, is comparable at $0.19 \pm 0.04^\circ\text{C} (10 \text{ yr})^{-1}$. For comparison, the trend over the period 1979–2017 is indistinguishable at $0.18 \pm 0.06^\circ\text{C} (10 \text{ yr})^{-1}$. However, the HadISDH.extremes T_w X trend over the China/East Asia region (Fig. 4a) is only $0.05 \pm 0.06^\circ\text{C} (10 \text{ yr})^{-1}$, and $0.06 \pm 0.11^\circ\text{C} (10 \text{ yr})^{-1}$ over the period 1979–2017. This is comparable with the uncorrected ERA5 trends from Freychet et al. (2020), which suggests that the

“HQ Flag = 0–6” selection is not sufficient to remove the effect of inhomogeneity over this region. This is unsurprising given the fact that the Chinese inhomogeneity over the early 2000s cannot be detected by a peak in HQ Flag scores (Fig. 3), although HadISDH.land does appear to be detecting and adjusting for this inhomogeneity to some extent.

Freychet et al. (2020) showed that when the observations are homogenised there is no trend in JJA RH over China. Conversely, HadISDH.land shows a trend of $-0.22 \pm 0.18 \text{ \%rh} (10 \text{ yr})^{-1}$ (Fig. 6b), becoming even more negative over the period 1979–2017 at $-0.41 \pm 0.25 \text{ \%rh} (10 \text{ yr})^{-1}$. Clearly, HadISDH.land overestimates decreasing trends in RH over the China/East Asia region compared to the Freychet et al. (2020) analyses, despite being comparable for T_w . The China/East Asia HadISDH.land time series is very similar to the unhomogenised observations, ERA5 and especially the ERA-Interim time series presented in Fig. 1 in Freychet et al. (2020), with a peak in 1993 and then a marked decline until 2009. Correspondingly, the China/East Asia region mean JJA T_w timeseries (Fig. 6a) shows a flatter period between 1998 and 2015, overlapping the period of sharply decreasing RH. For comparison, the UK/Europe region’s mean JJA RH (Fig. 6b) shows a steady decrease over the entire period, while the JJA T_w (Fig. 6a) shows a steady increase. Both the UK/Europe region’s mean JJA RH and T_w trends are greater in absolute magnitude than the China/

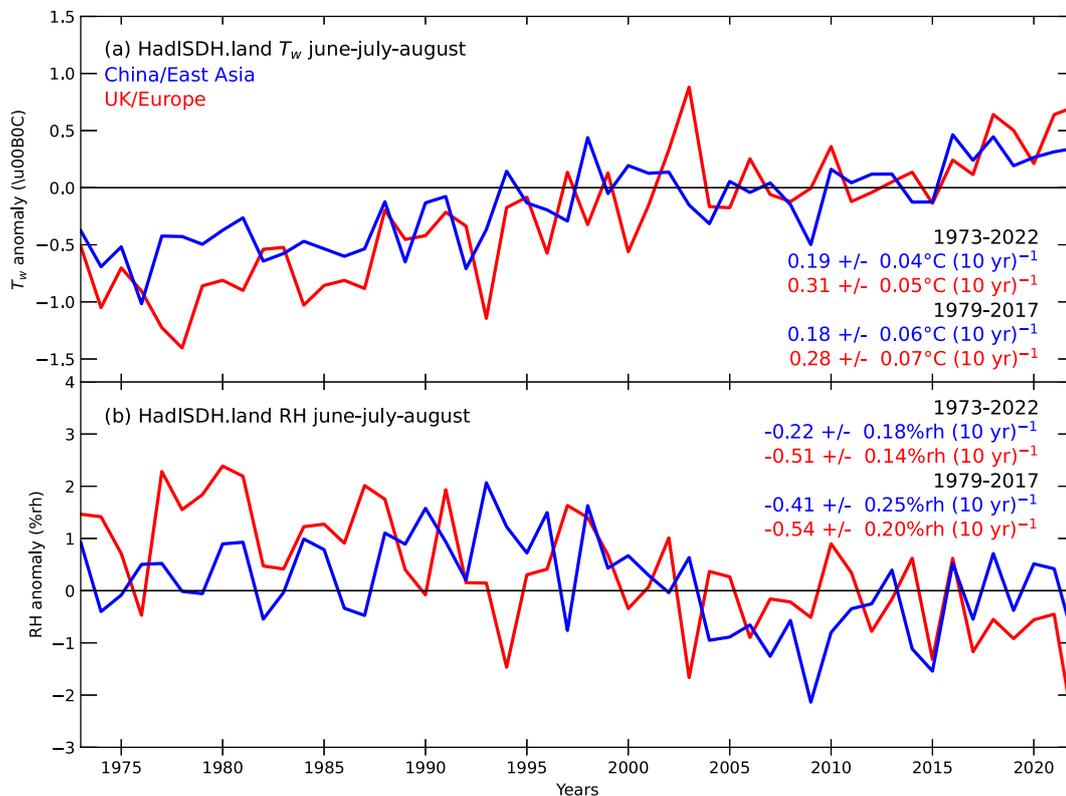


Fig. 6. Regional average JJA mean anomaly (relative to 1991–2020) time series for (a) T_w and (b) RH from HadISDH.land. Gridboxes have been screened to remove those with lower than 70% completeness over the 600-month record. Decadal trends in anomalies shown were fitted using ordinary least-squares regression with 90th percentile confidence intervals corrected for AR(1) following Santer et al. (2008).

East Asia region's mean trends.

The 1979–2017 global mean annual RH trend from HadISDH.land is $-0.22 \pm 0.17 \text{ \%rh (10 yr)}^{-1}$, and the JJA trend is $-0.30 \pm 0.13 \text{ \%rh (10 yr)}^{-1}$. Declining trends in HadISDH.land and ERA-Interim reanalysis RH have been shown to be widespread in spatial extent, especially across the midlatitudes (Weber, 2022). This global drying is well supported by theoretical understanding of thermodynamic drivers (Joshi et al., 2008; Simmons et al., 2010; Berg et al., 2016; Chadwick et al., 2016; Weber, 2022). Essentially, the land has warmed faster than the oceans. As a result, there is insufficient moisture available, from the relatively dry land surface and from the moisture advected from over the slower warming oceans, to maintain constant RH near the faster warming land surface. At regional scales, alternative drivers are also likely important – namely, physiological and dynamical drivers (Weber, 2022). Physiological drivers comprise the reduced evapotranspiration resulting from increased stomatal conductance efficiency in the presence of elevated CO₂ levels, and changes to land cover and land use. Dynamical drivers relate to changes in modes of variability or atmospheric circulation. Therefore, the decreasing HadISDH.land RH over the China/East Asia region appears plausible. However, given the disagreement with the nationally corrected Chinese RH observations and the agreement with the uncorrected ERA5 T_wX, it remains a region of large uncertainty, which should be kept in mind by users of HadISDH.extremes. The presence of inhomogeneities causing increasing trends in T_wX and related indices to be underestimated cannot be ruled out.

The JJA regional mean behaviour in HadISDH.land RH and T_w for the UK/Europe is larger in absolute terms than over China/East Asia. The UK/Europe region's air is becoming less saturated, with RH decreasing significantly at $-0.51 \pm 0.14 \text{ \%rh (10 yr)}^{-1}$. Simultaneously, the T_w is increasing significantly at $0.31 \pm 0.05^\circ\text{C (10 yr)}^{-1}$, so despite the decrease in saturation, the heat content of the air from its temperature and water vapour is increasing. Therefore, the risk to human and livestock health is increasing. This result is in broad agreement with global trends in RH and specific humidity (Willett et al., 2014, 2022; Simmons et al., 2021). Overall, trends in T_wX, T_wX90p, TX and TX90p are also larger than over the China/East Asia region. The presence of inhomogeneities causing increasing trends in T_w to be underestimated and decreasing trends in RH to be overestimated cannot be ruled out.

In conclusion, changes related to humidity over the China/East Asia region remain uncertain. The results and dataset presented here should be interpreted over China/East Asia with caution, keeping in mind the possibility that HadISDH.extremes trends in T_w extremes may be underestimated.

5. Usage notes

Herein is presented the methodology and data for the

Met Office Hadley Centre Integrated Surface Database for Humidity extremes indices product HadISDH.extremes.1.0.0.2022f. This is a dataset containing 27 near-surface wet bulb and dry bulb temperature-based extremes indices (Table 2) designed for assessment of large-scale, long-term features of humid and dry heat extremes. It is a near-global monthly gridded (5° by 5°) product beginning in January 1973 and continuing until December 2022 (at time of writing), with annual updates envisaged.

Data quality and long-term stability have been a key focus for this dataset, with stations having to meet strict data completeness requirements and undergo a suite of quality control tests. In addition, each gridbox month has an accompanying HQ Flag score, which can be used to exclude gridboxes most affected by inhomogeneity. Selecting gridboxes with an HQ Flag = 0–6 is recommended as a sensible choice that balances spatiotemporal coverage with long-term stability of the data. For global and large-scale regional seasonal means, this selection made negligible differences to long-term trends and anomaly magnitudes.

Despite these efforts, the nature of extremes means that they are more susceptible to error than monthly means. This is because the indices are based on counts of single hourly values for each day or for T_wX a single hourly value within the month. There is no averaging whereby random error can be muted and missing days have a much greater potential to result in error. Using gridbox mean anomalies where possible can reduce error but only where there are two or more stations contributing to that gridbox. Furthermore, the HadISDH.extremes (T_wX and TX) will always likely be an underestimate of the true value because the original observations are discrete in time, sampling at most hourly and often 3 hourly or less frequently. For these reasons, when using this dataset to study small-scale features, cross-validation against independent information such as national datasets and reports should be undertaken. It is also strongly recommended that users read the accompanying analysis and validation paper (Willett, 2023b).

The known issue of a region-wide shift from manual wet bulb thermometers to automated RH sensors over China is likely still an issue for HadISDH.extremes. Despite the provision of HQ Flag scores to screen out inhomogeneous data, the necessary automation of the inhomogeneity detection process means that region-wide changes are not easily detected. Hence, the HQ Flag scores over China may not fully identify poor quality gridboxes. Trends in wet bulb extremes are considerably smaller for the HadISDH.extremes China/East Asia region compared to those from the homogenised Chinese dataset (Freychet et al., 2020). However, significant positive trends are detectable in HadISDH.extremes in the percentage of days equal to or exceeding specific thresholds, and in days exceeding the 90th percentile of daily maximum T_w.

A product such as HadISDH.extremes will always contain some level of uncertainty. Herein is presented an approach to minimise the effect from systematic biases (inhomogeneity) by using the HQ Flag scores. It is noted that the

data have been “quality controlled” to remove the detectable random errors, recognising that inevitably some remain. Stations are selected and the data from hourly to monthly indices are processed in a manner that prioritises long, near-continuous records, suitable for analysis of long-term trends. The final product provides gridbox mean anomalies to reduce the effect of uneven spatial sampling over heterogeneous topography. Despite these efforts, it is recommended that users of HadISDH.extremes consider using the following measures to further reduce the effect of potential biases:

- **Spatial consistency:** Temperature and humidity anomalies often have good (a few 100 km) spatial correlation over a few 100 km. Users should ensure there are similar signals across at least neighbouring gridboxes, avoiding drawing conclusions from individual gridboxes that are outliers compared to their neighbours. Users can have greater confidence in widespread consistent features.

- **Safety in numbers:** Data-dense regions are less susceptible to individual station errors. Users can be more confident in analysis over data-dense regions (as shown in Fig. 1 and by the station count information provided in HadISDH.extreme data files) and should be cautious where gridboxes contain only one station.

- **Comparison with other supporting evidence:** There is a general expectation that high humidity extremes will increase over many regions given the widespread increase in temperature related to anthropogenic emissions and the resulting increased ability for warmer air to hold more water as a vapour. Users could compare trends in mean temperature and humidity over similar regions from other products, taking into account their potential uncertainties too.

- **Validation with national datasets and key events:** National meteorological services may make some of their own holdings of station observations available for comparison with small-scale features in HadISDH.extremes, and users can look for evidence of known extreme heat events within HadISDH.extremes.

- **Screening for continuity:** Analysis of long-term trends and analysis summing counts of exceedance becomes less robust where there are gaps in the data. Users could screen the gridboxes prior to analysis to remove those with lower than desired (e.g., 70 %) completeness over the temporal record and/or add a year completeness check to remove those years with any missing months.

- **Asking appropriate questions of the data:** HadISDH.extremes is better for assessment of large-scale, long-term features rather than specific records associated with a locale. When looking at correlations and potential drivers, users should have a theory- and understanding-driven approach.

Acknowledgements. This work and its contributors (Kate WILLETT) were supported by the UK–China Research & Innovation Partnership Fund through the Met Office Climate Science for Service Partnership (CSSP) China as part of the Newton Fund.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing,

adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

REFERENCES

- Armstrong, B., and Coauthors, 2019: The role of humidity in associations of high temperature with mortality: A multicountry, multicity study. *Environmental Health Perspectives*, **127**(9), 097007, <https://doi.org/10.1289/EHP5430>.
- Berg, A., and Coauthors, 2016: Land–atmosphere feedbacks amplify aridity increase over land under global warming. *Nature Clim. Change*, **6**, 869–874, <https://doi.org/10.1038/nclimate3029>.
- Brugnara, Y., M. P. McCarthy, K. M. Willett, and N. A. Rayner, 2023: Homogenization of daily temperature and humidity series in the UK. *International Journal of Climatology*, **43**(4), 1693–1709, <https://doi.org/10.1002/joc.7941>.
- Buck, A. L., 1981: New equations for computing vapor pressure and enhancement factor. *J. Appl. Meteorol.*, **20**, 1527–1532, [https://doi.org/10.1175/1520-0450\(1981\)020<1527:NEFCVP>2.0.CO;2](https://doi.org/10.1175/1520-0450(1981)020<1527:NEFCVP>2.0.CO;2).
- Chadwick, R., P. Good, and K. Willett, 2016: A Simple Moisture Advection Model of Specific Humidity Change over Land in Response to SST Warming. *J. Climate*, **29**, 7613–7632, <https://doi.org/10.1175/JCLI-D-16-0241.1>.
- Climpact, 2022: Climpact open source R software for calculating sector-specific extremes indices. Available from <https://climpact-sci.org/>.
- Dunn, R. J. H., 2019: HadISD version 3: Monthly updates, Hadley Centre Technical Note 103, 10 pp. Available from <https://www.metoffice.gov.uk/research/library-and-archive/publications/science/climate-science-technical-notes>.
- Dunn, R. J. H., 2021a: Adapting the QC to account for the June 2021 North American Heatwave (part 3). Available from https://hadisd.blogspot.com/2021/09/adapting-qc-to-account-for-june-2021_20.html.
- Dunn, R. J. H., 2021b: Adapting the QC to account for the June 2021 North American Heatwave (part 2). [Available online from <https://hadisd.blogspot.com/2021/09/adapting-qc-to-account-for-june-2021.html>]
- Dunn, R. J. H., 2021c: Adapting the QC to account for the June 2021 North American Heatwave (part 1). Available from <https://hadisd.blogspot.com/2021/07/adapting-qc-to-account-for-june-2021.html>.
- Dunn, R. J. H., 2021d: The June 2021 North American Heatwave and v3. 1. 2. 202106p. Available from <https://hadisd.blogspot.com/2021/07/the-june-2021-north-american-heat-wave.html>.
- Dunn, R. J. H., and C. P. Morice, 2022: On the effect of reference periods on trends in percentile-based extreme temperature indices. *Environmental Research Letters*, **17**(3), 034026,

- <https://doi.org/10.1088/1748-9326/ac52c8>.
- Dunn, R. J. H., K. M. Willett, D. E. Parker, and L. Mitchell, 2016: Expanding HadISD: Quality-controlled, sub-daily station data from 1931. *Geoscientific Instrumentation, Methods and Data Systems*, **5**, 473–491, <https://doi.org/10.5194/gi-5-473-2016>.
- Dunn, R. J. H., and Coauthors, 2020: Development of an updated global land in situ-based data set of temperature and precipitation extremes: HadEX3. *J. Geophys. Res.: Atmos.*, **125**, e2019JD032263, <https://doi.org/10.1029/2019JD032263>.
- ET-SCI, 2022: Expert Team for Sector-Specific Indices webpage. Available from <https://public.wmo.int/en/events/meetings/expert-team-sector-specific-climate-indices-et-sci>.
- Freychet, N., S. F. B. Tett, Z. Yan, and Z. Li, 2020: Underestimated change of wet-bulb temperatures over East and South China. *Geophys. Res. Lett.*, **47**, e2019GL086140, <https://doi.org/10.1029/2019GL086140>.
- Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, <https://doi.org/10.1002/qj.3803>.
- Ingleby, B., D. Moore, C. Sloan, and R. Dunn, 2013: Evolution and accuracy of surface humidity reports. *J. Atmos. Oceanic Technol.*, **30**(9), 2025–2043, <https://doi.org/10.1175/JTECH-D-12-00232.1>.
- Jensen, M. E., R. D. Burman, and R. G. Allen (ed). 1990: Evapotranspiration and Irrigation Water Requirements. ASCE Man. and Rep. on Engineering Pract. No. 70, New York, 332 p.
- Joshi, M. M., J. M. Gregory, M. J. Webb, D. M. H. Sexton, and T. C. Johns, 2008: Mechanisms for the land/sea warming contrast exhibited by simulations of climate change. *Climate Dyn.*, **30**, 5, p. 455–465 11 p.
- Li, Z., Z. W. Yan, Y. N. Zhu, N. Freychet, and S. Tett, 2020: Homogenized daily relative humidity series in China during 1960–2017. *Adv. Atmos. Sci.*, **37**(4), 318–327, <https://doi.org/10.1007/s00376-020-9180-0>.
- Mears, C. A., J. P. Nicolas, O. Bock, S. P. Ho, and X. Zhou, 2022: [Global Climate] Total Column Water Vapour [in “State of the Climate in 2021”]. *Bull. Amer. Meteor. Soc.*, **103**(8), S52–S54.
- Menne, M. J., and C. N. Williams Jr., 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate*, **22**(7), 1700–1717, <https://doi.org/10.1175/2008JCLI2263.1>.
- Raymond, C., T. Matthews, and R. M. Horton, 2020: The emergence of heat and humidity too severe for human tolerance. *Science Advances*, **6**(19), eaaw1838, <https://doi.org/10.1126/sciadv.aaw1838>.
- Santer, B. D., and Coauthors, 2008: Consistency of modelled and observed temperature trends in the tropical troposphere. *International Journal of Climatology*, **28**, 1703–1722, <https://doi.org/10.1002/joc.1756>.
- Schär, C., 2016: Climate extremes: The worst heat waves to come. *Nature Climate Change*, **6**, 128–129, <https://doi.org/10.1038/nclimate2864>.
- Shen, D. D., and N. Zhu, 2015: Influence of the temperature and relative humidity on human heat acclimatization during training in extremely hot environments. *Building and Environment*, **94**, 1–11, <https://doi.org/10.1016/j.buildenv.2015.07.023>.
- Sherwood, S. C., and M. Huber, 2010: An adaptability limit to climate change due to heat stress. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 9552–9555, <https://doi.org/10.1073/pnas.0913352107>.
- Simmons, A. J., K. M. Willett, P. D. Jones, P. W. Thorne, and D. P. Dee, 2010: Low-frequency variations in surface atmospheric humidity, temperature, and precipitation: Inferences from reanalyses and monthly gridded observational data sets. *J. Geophys. Res.: Atmos.*, **115**, D01110, <https://doi.org/10.1029/2009JD012442>.
- Simmons, A. J., and Coauthors, 2021: Low frequency variability and trends in surface air temperature and humidity from ERA5 and other datasets. European Centre for Medium-range Weather Forecasting (ECMWF) Technical Memoranda 881, ECMWF, Shinfield Park, Reading, UK, 97 pp, <https://doi.org/10.21957/ly5vbtbfd>.
- Smith, A., N. Lott, and R. Vose, 2011: The integrated surface database: Recent developments and partnerships. *Bull. Amer. Meteor. Soc.*, **92**, 704–708, <https://doi.org/10.1175/2011BAMS3015.1>.
- Stull, R., 2011: Wet-bulb temperature from relative humidity and air temperature. *J. Appl. Meteorol. Climatol.*, **50**(11), 2267–2269, <https://doi.org/10.1175/JAMC-D-11-0143.1>.
- Vecellio, D. J., S. T. Wolf, R. M. Cottle, and W. L. Kenney, 2022: Evaluating the 35°C wet-bulb temperature adaptability threshold for young, healthy subjects (PSU HEAT Project). *Journal of Applied Physiology*, **132**(2), 340–345, <https://doi.org/10.1152/jappphysiol.00738.2021>.
- Wang, P. Y., L. R. Leung, J. Lu, F. F. Song, and J. P. Tang, 2019: Extreme wet-bulb temperatures in China: The significant role of moisture. *J. Geophys. Res.: Atmos.*, **124**, 11 944–11 960, <https://doi.org/10.1029/2019JD031477>.
- Weber, K. M. F., 2022: Why is the atmosphere over land becoming drier? Exploring the roles of atmospheric and land-surface processes on relative humidity. PhD dissertation, University of Sheffield.
- Willett, K. M., 2023a: HadISDH.extremes: Gridded global monthly land surface wet bulb and dry bulb temperature extremes index data version 1.0.0.2022f. NERC EDS Centre for Environmental Data Analysis, <https://doi.org/10.5285/2d1613955e1b4cd1b156e5f3edbd7e66>.
- Willett, K. M., 2023b: HadISDH.extremes Part II: Exploring humid heat extremes using wet bulb temperature indices. *Adv. Atmos. Sci.*, <https://doi.org/10.1007/s00376-023-2348-7>.
- Willett, K. M., D. A. Lavers, M. Bosilovich, and A. J. Simmons, 2022: [Global Climate] Surface Humidity [in “State of the Climate in 2021”]. *Bull. Amer. Meteor. Soc.*, **103**(8), S50–S52.
- Willett, K. M., C. N. Williams Jr., R. J. H. Dunn, P. W. Thorne, S. Bell, M. de Podesta, P. D. Jones, and D. E. Parker, 2013: HadISDH: An updateable land surface specific humidity product for climate monitoring. *Climate of the Past*, **9**, 657–677, <https://doi.org/10.5194/cp-9-657-2013>.
- Willett, K. M., R. J. H. Dunn, P. W. Thorne, S. Bell, M. de Podesta, D. E. Parker, P. D. Jones, and C. N. Williams Jr., 2014: HadISDH land surface multi-variable humidity and temperature record for climate monitoring. *Climate of the Past*, **10**, 1983–2006, <https://doi.org/10.5194/cp-10-1983-2014>.
- Willett, K. M., R. J. H. Dunn, P. W. Thorne, S. Bell, M. de

- Podesta, D. E. Parker, P. D. Jones, and C. N. Williams Jr., 2023: HadISDH.land: Gridded global monthly land surface humidity data version 4. 5. 1.2022f. NERC EDS Centre for Environmental Data Analysis, <https://doi.org/10.5285/8956cf9e31334914ab4991796f0f645a>.
- Yu, S., S. F. B. Tett, N. Freychet, and Z. W. Yan, 2021: Changes in regional wet heatwave in Eurasia during summer (1979–2017). *Environmental Research Letters*, **16**, 064094, <https://doi.org/10.1088/1748-9326/ac0745>.
- Zhang, X. B., G. Hegerl, F. W. Zwiers, and J. Kenyon, 2005: Avoiding inhomogeneity in percentile-based indices of temperature extremes. *J. Climate*, **18**(11), 1641–1651, <https://doi.org/10.1175/JCLI3366.1>.