LETTER

On the effect of reference periods on trends in percentile-based extreme temperature indices

Robert J H Dunn and Colin P Morice

Met Office Hadley Centre, Fitz Roy Road, Exeter, EX1 3PB, United Kingdom

* Author to whom any correspondence should be addressed.

E-mail: robert.dunn@metoffice.gov.uk

Keywords: reference period, climate extremes indices, percentile-based indices, warming trends

Abstract

A number of studies have noted that the use of distinct reference periods when comparing indices measuring the frequency of days exceeding a particular temperature percentile threshold leads to apparently different behaviour. We show that these differences arise because of the interplay between the increasing temperatures and the choice of reference period. The time series of the indicators calculated using the different reference periods are offset, as expected, but also diverge. Linear trends calculated over the same period from the same underlying data but where different reference periods have been used are substantially different if a change in climatological conditions has occurred between the two reference periods. We show this not only occurs in our simple empirical approach, but also for the averages of gridded observational and reanalysis datasets and also at a station level. This has implications for data set comparisons using trends in temperature percentile indices that are based on different reference periods. It also has implications for updates to standard reference periods used to monitor the climate.

1. Introduction

Extreme temperature events cause some of the clearest impacts of climate change, affecting ourselves and our society, our infrastructure and of course, the natural world. To enable the intercomparison of extremes, the World Meteorological Organization (WMO) former Expert Team on Climate Change Detection and Indices (ETCCDI) developed a standard set of indices based on daily precipitation, and maximum and minimum temperatures for land surface observations (Karl et al 1999, Peterson et al 2001, Frich et al 2002, Peterson 2005, Zhang et al 2011). These 27 ETCCDI extremes indices, endorsed by the WMO, enabled the easy intercomparison of extremes across the globe in three ways. Firstly, the definitions were standardised and so users could be sure that the same quantities were being compared between regions. Secondly, a subset of the indices were specifically constructed to enable intercomparison between different regional climates (see below), something that extremes defined relative to fixed thresholds or as peak values cannot easily do (though these types of indices remain part of the ETCCDI family). Thirdly, derived data for the family of indices could be shared more freely than the underlying daily observations on which they were based which often have restrictive data policies (Thorne et al 2017, Alexander et al 2019).

Over time, a number of observation-based datasets have been developed in order to monitor changes in extremes, starting with HadEX (Alexander et al 2006), and most recently HadEX3 (Dunn et al 2020a) which now covers 1901–2018 at 1.25° × 1.875° latitude-longitude resolution.

Some of the extremes indices defined by the ETCCDI and used within datasets like HadEX3 use a reference period as part of their definition. For the temperature indices (TX90p, TX10p, TN90p and TN10p) thresholds are set by the 90th or 10th percentile values (for the maximum (TX) and minimum (TN) temperatures), as determined over a reference period. These thresholds are used to count the number of days where the maximum or minimum temperatures exceed them, e.g. TX90p (also known as the number of ‘warm days’, see table 1) is the number of days when the maximum temperature exceeds this 90th percentile threshold. In contrast to e.g. TXx (the
maximum TX measured within a month or year), these indices based on the counts exceeding the percentile values measure changes in climate extremes. There are also precipitation indices (R95p, R95pTOT, R99p, R99pTOT) which calculate the (fraction of) annual total precipitation falling in the wettest 5% or 1% of days, where again these thresholds are determined from the properties of the wet days during the reference period. The advantage of using percentile thresholds is that they adapt to the climate at each location, and so multiple locations can be more easily compared. Under a changing climate, this adaptation to the local climate still occurs; and by updating the percentile thresholds, different periods can also be compared. Other ETCCDI indices (e.g. SU, ‘summer days’, the number of days where TX > 25 °C) use a fixed threshold, which changes how comparisons work over time (and space). When using a fixed threshold (which is not dependent on the local climate), there are regions of the world which rarely/always exceed the chosen values, limiting their use in those places. However as the thresholds are fixed, the use of a reference period is not relevant and we do not consider these indices herein.

With a slowly changing network of stations, it is not sensible to perpetually use an older reference period in order to maintain consistency with earlier analyses. For stations to be included in any analysis, they must overlap in time with the reference period, which means that newer stations cannot be included if an older reference period is used. Furthermore, the effect of stations closing results in a decrease in coverage in the most recent times.

In the HadEX3 dataset of gridded extremes indices, for the subset of indices which use reference periods in their construction, two reference periods were used (1961–1990 and 1981–2010). This followed the then current WMO guidance, which states that 1961–1990 is to be retained for long-term climate change assessments (WMO 2007, 2017). However, the inclusion of the 1981–2010 period allowed assessment in regions where this was not possible when using 1961–1990. Dunn et al (2020a) compared the two versions of the dataset for each of the relevant indices using the global (land) average timeseries. They showed that there was a non-stationary difference between the two equivalent temperature indices based on different reference periods. We recreate this in figure 1 using HadEXv3.0.3, which demonstrates the non-stationary difference for TX90p for the two reference periods (for full details see Dunn et al (2020a)). The two versions of HadEX3 (using reference periods of 1961–1990 and 1981–2010) are masked to the same spatio-temporal coverage, but have slight differences in the underlying station network used to generate the gridded datasets. A non-stationary difference means that the linear trends using a simple least squares fit are different over the same period of record in the two versions of HadEX3. By using a later (warmer) reference period, the rate of change of TX90p (warm days) has been reduced compared to the version using an earlier (cooler) reference period. Furthermore, the non-linear nature of the exceedance counts measured by these temperature indices means it is not trivial to compare values of these indices when defined using different reference periods.

Similarly, Yosef et al (2021) demonstrate a problem arising for temperature indices at individual stations when using two reference periods over Israel. They noted a discrepancy in the trends from two independent studies using different reference periods (Salameh et al 2019, Yosef et al 2019). By using a later (warmer, 1988–2017) reference period they showed that the decreasing trends of the cool (10th) percentile indices were amplified, and the increasing trends of the warm (90th) percentile indices were diminished compared to when using an earlier (cooler, 1961–1990) period; trends over the same temporal periods were different when using different reference periods, potentially leading to discrepant conclusions as to the change in extreme events.

In order to provide a reference to users of these indices and datasets, herein we outline how the reference periods, thresholds and indices interact under a changing climate from an empirical and practical perspective, to highlight perhaps unexpected behaviours that may be observed and build on the issues raised in Dunn et al (2020a), Yosef et al (2021).

| Index | Name       | Description                              | Units |
|-------|------------|------------------------------------------|-------|
| TN10p | Cool nights| Number of days when daily min temperature < 10th percentile | days   |
| TX10p | Cool days  | Number of days when daily max temperature < 10th percentile | days   |
| TN90p | Warm nights| Number of days when daily min temperature > 90th percentile | days   |
| TX90p | Warm days  | Number of days when daily max temperature > 90th percentile | days   |
2. Relative motion of thresholds

We start very simply, in order to set the scene and also familiarise readers with the indices which are being investigated. In figure 2(a) we show the probability distribution function (PDF) of (e.g. maximum) temperatures over a 30-year reference period, represented by a Gaussian function, $\phi(x)$, of $\mu = 0^\circC$ and $\sigma = 1^\circC$. Although idealised, this PDF is a reasonable representation of the distribution of temperatures when averaged over time and space (see e.g. Dunn et al (2019) and references therein). The thresholds determined from the 10th and 90th percentiles are also shown in figure 2 along with the integrated area of the curve which exceeds these (blue and red shaded regions). These areas are the quantities measured by the TX10p and TX90p ETCCDI extremes indices, and can be also represented by a shift in the cumulative distribution function (CDF, see appendix figure 7).

As the climate warms, the distribution of temperatures will shift to higher temperatures, which we represent in this example as a shift in the mean parameter of the PDF from $\mu = 0.0^\circC$ to $\mu = 0.5^\circC$ (figure 2(b)). We assume there is no change in the shape of the distribution (e.g. variance, skewness). By using the same temperature threshold as before, the number of low exceedences falls and the number of high exceedences increases, as indicated by the changes in size of the shaded areas, and so these indices can monitor changes in the number of these extremes.

An equivalent action to this shift in the mean of the Gaussian distribution against fixed temperature thresholds, is for the distribution to remain stationary, but the temperature thresholds to move. This is because a change in the mean parameter has the same effect as a change in a temperature threshold for the CDF of a Gaussian. Hence, in the frame of the warmer Gaussian, the lower threshold is now at the 4th and the higher threshold at the 78th percentile respectively, as indicated by the second x-axis in figure 2(b) and in figure 7. Therefore, under a warming climate, using an earlier (cooler) reference period means that for the most recent years, the index is not counting events with the same level of ‘extremity’ for the instantaneous climate at the time (see also Dunn et al (2019)).

The shift in the temperature curve does not result in a symmetric or monotonic change in percentile space, with a 6 percent change in one and a 12 percent change in the other in this example, arising from the non-linear shape of the Gaussian. It is this which leads to the difference in the timeseries curves seen in figure 1, which we now investigate in more detail.

3. Influence on timeseries

To investigate the interplay of a warming climate with thresholds determined over specified reference periods we simulate a set of daily data (365 samples representing the days in a year) drawn from a Gaussian distribution ($\mu = 0^\circC$ at 1900, $\sigma = 2^\circC$). By applying an underlying linear trend of $0.02^\circC$ year$^{-1}$ we model a simplistically warming climate, and repeat the simulation of the daily data for each year from 1900 to 2020. We then calculate values for TX90p...
(counts of days where the temperatures are over the 90th percentile) using two reference periods to define the percentile boundaries, 1961–1990 and 1981–2010 following HadEX3, and also the WMO guidance (WMO 2007, 2017). Studies into the distribution of the underlying daily maximum and minimum temperatures from which the indices are calculated found that a Gaussian was a reasonable representation (Donat and Alexander 2012).

We note that for the ETCCDI indices as used in e.g. HadEX3, the percentiles for each calendar day are calculated from the observations centered on a 5-day window for the reference period. Furthermore, to avoid possible inhomogeneity across the in-reference and out-reference periods, a bootstrapping procedure is used as described in Zhang et al. (2005). We have not included an annual cycle in our simulation, and so using separate thresholds for each calendar day is not required.

As expected, the two curves shown in figure 3(a) have similar behaviour in the year-to-year variation towards the end of this simulated record, but similarity is much reduced at the beginning of the record due to the stochastic nature of the threshold exceedences in any given year. What is of importance here is that the difference between the two curves is not stationary and grows over time (figure 3(c)), as was noted between the two versions of HadEX3 using different reference periods shown in figure 1 and in Dunn et al. (2020a).

For the HadEX3 dataset, the timespan of the timeseries and difference curve is limited by the availability of the observational data (1901–2018). We increase the time span of this simulation to 1800–2300 in figures 3(b) and (d), where the shapes of both the timeseries and difference curve are clear. This highlights how the non-linearities in the timeseries and difference curves (figures 3(a)–(d)) are not apparent over short timescales, but are revealed over a number of centuries. The end point (2300) has been chosen arbitrarily, though is also often used as the end point for long climate simulations for e.g. sea level rise. By this time the two timeseries converge as sufficient warming has taken place that almost all days are warmer than the threshold.

In this simple example, the overall shape of the timeseries (more clearly shown in figure 3(b)) will be that of the CDF of the underlying Gaussian distribution of temperatures, arising as the Gaussian distribution slides past the stationary threshold. The CDF of a Gaussian distribution, \( \Phi(x) \), mean \( \mu \), variance \( \sigma^2 \) is given by the integral from the threshold, \( z \), to infinity:

\[
\Phi(z, \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \int_z^\infty \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx.
\]

Using the simple model described above, we calculate this CDF, and show the resulting empirically expected values for TX90p in figure 3, which appear as the smoother lines. From these CDFs we also derive the empirical shape of the difference curve (Gaussian) and also show that in panels (c) and (d).

By using this simple example (and we certainly do not expect the use of a reference period set over a century in the past for the purpose of calculating these ETCCDI indices) we can see how the underlying change in temperatures (linear) along with their distribution (Gaussian) at any point in time combine to drive the shape of the difference curves and their divergence. The exact nature of the effect is

1 Usually, this integral is defined from \(-\infty\) to \(z\), but in the case where \(z\) is negative, this is the same as from a positive \(z\) to \(\infty\).
dependent on the shape of the data distribution and how this affects the change in the integral above the threshold as the location parameter changes over time. In section 5 and the appendix we show similar effects to the ones discussed here occurring in simulated examples for other distributions.

In HadEX3, the underlying change in daily temperatures is not constant over the 1901–2018 period, as demonstrated by any of the commonly available monitoring products of global temperature (e.g. see www.metoffice.gov.uk/hadobs/monitoring/dashboard.html). Therefore the shape of the difference between the timeseries of threshold driven metrics in the two versions of HadEX3 reflects the shape of the variation in observed maximum temperatures during the 20th and 21st centuries (see figure 1(b) and figure 10 in Dunn et al (2020a)). Over this comparatively short time period and close to the reference periods, the greatest change is driven by the long-term variation in temperatures, rather than short-term, year-to-year variability.

Also apparent in figure 3(d) is an increase in the variance of the difference curve, peaking during the period of maximum difference. This is a natural consequence of the interaction of the sampling, the parent distribution and the threshold. When the temperature thresholds from both reference periods are in the upper tail of the temperature distribution, then for a given sample, the number of days which exceed the lower of these two thresholds will be very similar to the number which exceed the higher. Most days for that sample year will lie below both thresholds, and hence the difference is small between the two indices is small. When the thresholds are close to the mean of the parent distribution, there is a greater chance for a day to be above the lower threshold, but below the higher one. Hence year-to-year variability in the samples has a greater effect on the difference when the distribution has moved so that the thresholds are close to the middle.

The timeseries from HadEX3 shown in figure 1 is derived by taking a quasi-global (land) average for each year in the dataset (‘quasi’ because the spatio-temporal coverage is not complete). HadEX3 blends together the thousands of underlying stations using an Angular Distance Weighting routine (Sheppard 1968), which smooths the local differences in station values. By taking a global average, this naturally smooths out some of the year-to-year variations, and so the non-stationary difference between the two versions can stand out clearly. In contrast, when using a single station, the year-to-year variation is larger in comparison with long term trends. We now show how these effects are also present at the station level and in an independent dataset.

Yosef et al (2021) showed that trends in these indices at a number of stations in Israel were not the same when using different reference periods. Here we use the European Climate Assessment & Dataset (Klein Tank et al 2002), which were input data.
to HadEX3, and take a long running station (de Bilt, NL, Station ID: 162) to explore the behaviour of temperature extremes indicators at this example station. We show in figure 4 the behaviour for TN90p (which is analogous to TX90p but instead using the minimum temperature and most clearly demonstrates the effects). In this case the temperature thresholds are determined on a daily basis aggregated over the 30-year reference period in a 5-day sliding window (150 values). For ease, we remove February 29th from this analysis. In comparison to HadEX3 (figure 1), the year-to-year variability is larger for this single station. However, the difference in the curves for the two reference periods is stable until around 1980, and starts to rise thereafter, which corresponds to the period with the greatest increase in the index values. So, although not as prominent, the non-linearity in the difference is also found at the station level.

We also use the Twentieth Century Reanalysis (20CR) (Slivinski et al 2019) to show how these effects are present in another gridded dataset. This reanalysis only assimilates surface pressure observations in order to reconstruct the weather patterns of the past, and so is independent of the temperature observations used in the HadEX3 or ECA&D examples given in figures 1 and 4. We calculate the TX90p index from the daily maximum and minimum temperature fields using the two reference periods, and calculate the land-surface average time series for both, along with their difference and linear trends (figure 5). As for HadEX3 and de Bilt, the difference in the time series increases during the recent period of rapid global warming. Hence the effects described in time series derived from our simple model are also found in a number of different data sets.

4. Influence on trends in these temperature indices

Although the differences in time series are noteworthy and important to understand, the influence on the linear trends of changes in these extremes indices is more critical. Linear trends have been calculated using a simple least-squares algorithm on the time series from both references periods and shown in figures 1, 4, and 5 in order to demonstrate the impact. As noted in Dunn et al (2020a), it is not expected that trends calculated over time should be linear, but using linear trend values provides a useful summary of changes.

It is clear that for HadEX3 (figure 1), for a single station (figure 4) and for the 20CR reanalysis (figure 5) linear trends in these percentile-based temperature extremes indices calculated over the same time period from the similar/same input data but using different reference periods are substantially different. In fact the difference between the trend values is a sizeable fraction of the magnitude of the trend.

This difference in trends was also noted in observational data by Yosef et al (2021) where they highlight the difficulties when comparing trends using different reference periods in two previous studies (Salameh et al 2019, Yosef et al 2019) with very similar underlying data. They found that the negative trends in cool indices (TN10p, TX10p) were enhanced when using a later (warmer) reference period, and the positive trends in warm indices (TX90p, TN90p) were
enhanced when using an earlier (cooler) reference period. We have shown for HadEX3, 20CR and de Bilt that the positive trends in TX90p/TN90p are enhanced when using the 1961–1990 reference period over 1981–2010.

To investigate these changing trends further, and over longer timescales with greater warming, we return to the simulated data. The shape of the timeseries over long periods (figure 3) is clearly non-linear, and hence trends calculated over a set period or with a set starting year will change over time. In figures 3(e) and (f) we show the least-squares 50-year trend for the simulated indices using both reference periods and plotted at the end of the trend period, which clearly demonstrates this non-linearity.

It is possible to calculate the theoretically expected behaviour of these trends from the parameters of the changing distribution. The value of the simulated index is the area of the Gaussian distribution exceeding the temperature threshold (see figure 2). Were the distribution to move to higher or lower temperatures, the instantaneous rate of change of this area is proportional to the value of the Gaussian at that temperature threshold. To put it another way, the shape of the theoretically expected timeseries is that of the cumulative probability distribution of the Gaussian. The trend of this timeseries at a given moment is given by the gradient of the CDF. As the CDF is the finite integral of the Gaussian distribution, then the differential of the CDF is the original Gaussian distribution, evaluated at the threshold $z$.

$$\frac{d}{dx} \Phi = \phi |_{z}. \quad (2)$$

For the empirical distribution, it is of course possible to calculate the instantaneous trend, rather than needing data over multiple decades. The empirical curves are the value of the Gaussian distribution evaluated at the midpoint of the trend period, but plotted at the end. So the values plotted at 1950 in figures 3(e) and (f) are those calculated in 1925.

Clearly, a feature of the way these ETCCDI temperature extremes indices have been constructed is that as the global climate warms, trends in these indices will be non-linear even if the underlying change in daily temperatures is linear (as in our model). Although the effect is likely to be small in comparison with that of the year-to-year variability for land observations, it is something users should be aware of. Especially as it was noted by Klein Tank et al (2009) that a change in reference period would only have ‘a small impact on the results for the changes in the indices over time.’ Yosef et al (2021) suggest that this is still true for a slowly changing climate, but under the more rapid changes experienced in the last few decades, this assertion does not appear to hold.

The important consequence of the non-stationary difference shown in figures 1, 3(c) and (d), 4 and 5 is that for the same time period, identical input observations and dataset methodology, users would determine different trends from versions with different reference periods. For this example, given the underlying trend in the temperatures is linear, the trend values for the 1981–2010 reference period lag those for the 1961–1990 reference period by 20 years (figures 3(e) and (f)). Of course, between datasets that use different methodologies, other inconsistencies are likely to be introduced that also contribute to different trends,
but if the reference period does not match, then this effect is essential to include. Users could conclude that there were true underlying differences or even inconsistencies between the datasets. However, this is not necessarily correct and it is at least partly, if not wholly, the result of the way these indices have been constructed and their interaction with our rapidly changing climate (Yosef et al 2021).

Furthermore, as can be seen from figure 6, if the reference periods were never updated, then the scenario could arise that the entire distribution were such that almost no days were cooler than the 10th percentile. And taking this example to the extreme, that all days were warmer than the 90th percentile. In the case of the 10th percentile, the difference between the two versions (using different reference periods) becomes less and less, as the stationary threshold moves further out in the tail of the distribution. In this case, the trends are becoming smaller as the distribution continues to warm, asymptotically reaching zero. Although in this case we do not expect users to conclude that there have been no changes in the cool extremes. However, for locations which have low annual temperature variability and hence relatively narrow distributions, a given shift in the mean temperature will have a larger effect than locations with high variability.

The effects of the reduction in the power of the 10th percentile to monitor extremes can already be seen in the timeseries of TN10p shown in Dunn et al (2020a) but also in annual monitoring reports using GHCNDex (Donat et al 2013), e.g. Dunn et al (2020b). The number of cool nights has reduced by almost 20 days since the 1961–1990 reference period. Should the warming continue at the present rate, it may be only a matter of a few decades before only a few nights per year are cooler than this threshold. As above, we do not expect the 1961–1990 reference period to still be in use by then for this purpose, but we use this as an example of how using different (older) reference periods (as is still done for HadEX3 and GHCNDex for consistency with earlier products) can affect trend intercomparisons.

5. Discussion

Using a simple model we show that under a linearly warming climate, the choice of a reference period has an impact on both the timeseries of threshold-dependent temperature extremes indices and their linear trends calculated over a given period. The non-stationary difference in the time series curves, as noted in Dunn et al (2020a), means that the comparisons between analyses of threshold-dependent metrics defined using different reference periods are not simple. This is in contrast to e.g. using a different periods to calculate anomalies, where the absolute values of curves can be corrected for straightforward comparison.

Most importantly, the linear trends calculated over the same period and even using the same data will be different when using different reference periods for the extremes indices (figures 1 and 3). We show that this results in differences in the estimates of the speed at which extremes are changing. This is important not only for studies using observational data to investigate the recent changes in extremes, but also for climate model projections. In our simple model, which uses a linear warming trend of 0.02 °C per year, the difference in trends for the maximum temperature indices (defined using different reference periods) is around 1 day per decade for 1900–1950, rising to two days per decade by 1970–2020 (figure 3(e)). As can be seen more clearly in figure 3(f) the change in trends when using the 1981–2010 reference period lags that when using 1961–90 by 20 years, the temporal separation of the two reference periods. The shape of the Gaussian results in a non-stationary difference between the timeseries from the two reference periods. This gives grounds for two recommendations: (a) that trends in temperature percentile exceedance statistics are not used interchangeably when derived from different reference periods and (b) that the reference period is always stated.

On a global average for HadEX3 we find a difference of 1 day per decade (a difference of around 25 percent, figure 1) over 1961–2018, for 20CR a difference of 2.4 days (a difference of over 30 percent, figure 5), and at the station at De Bilt (NL) where temperature has warmed relatively quicker than the global average, we find a difference of 3 days per decade for 1960–2020 (a difference of around 50 percent, figure 4). Although as shown in figure 3(d), this difference in trends is not linear, the magnitude of this effect can be substantial even when using a comparatively long baseline period for calculating the trends. Conversely, for the minimum indices, under a warming climate the difference in trends decreases (figure 6), but should our climate start cooling as a result of reductions in greenhouse gas concentrations, the opposite would occur.

Under a stable climate, the distributions and thresholds are unchanged between the two reference periods, and the effects we describe herein are not present. It is the recent, rapid rise in temperatures across the globe which has resulted in changes to the distribution of temperatures in recent decades, and hence the features we describe in this study. These would be enhanced in regions warming faster than the average. However, we note that when shifting baselines by a decade, there may be some regions where the reverse effect happens due to regional variability.
We also note that under a warming climate, the effect of using two reference periods to set thresholds using the same percentiles is the equivalent of choosing two different fixed thresholds or two different percentiles for the same reference period. In the example of the indices defined by the WMO Expert Team on Sector-Specific Climate Indices (ET-SCI), this could be using the number of days where e.g. $\text{TX} > 30^\circ\text{C}$ ($\text{TXge}30$) and $\text{TX} > 35^\circ\text{C}$ ($\text{TXge}35$). However, in cases like these it is clearer to the user/reader that two different quantities are being compared, whereas the change in the reference period on a single index is not.

As can be seen from the expected behaviour of the trends over time in figures 3 and 6, the trends change in a non-linear way. For the upper percentile under a warming climate and Gaussian distribution, this initially manifests as an accelerating increase even though the change in underlying temperature distribution is linear. Although this is a theoretical example, figures 3 and 6 demonstrate that care needs to be taken in interpreting apparent accelerations of extremes (indices).

Fischer et al. (2021) demonstrate how ‘record-shattering’ extremes (ones which surpass previous records by a large amount) are becoming increasingly probable. They use the 'TX7d' index, the hottest week of TX values, which does not have a threshold component. In their assessment extremes are compared to the previous record event rather than a climatological baseline, which is in contrast to the use of a stationary threshold as investigated here. They show that it is the warming rate which drives their observed changes in probability; in particular an accelerating warming rate after periods of little or no warming. In this work we have used a linear warming rate and show that even this can result in ‘accelerations’ of specific extremes (indices).

The WMO currently state that 1961–1990 is be retained for long-term climate change assessments (WMO 2007, 2017). However, especially for observation based analyses more data may be available over a later period. In an assessment of the future behaviour of climate extremes indices in the CMIP5 multimodel projections, Sillmann et al. (2013) included the temperature indices discussed here using the 1961–1990 reference period. They do not comment on the non-linearity in trends as the representative concentration pathways (RCPs) models used have a variety of shapes and hence a variety of global temperature responses. RCP8.5 (high emissions) is the scenario which results in global temperature changes which are the closest to linear over the 21st century. Figure 8 of Sillmann et al. (2013) shows that for the low percentiles (TN10p and TX10p) these asymptotically approach 0% by 2100, similar behaviour to our simulations in figure 6.

This asymptotic approach to 0% could be taken as an indication that there are limitations for the usefulness of these indices when combined with reference periods from a long time ago. However, firstly...
this shows that updating the reference periods is important under a changing climate when monitoring changes in the recent past, or projecting them into the near future. This would be valuable even though we have shown how comparisons between studies using different reference periods is complex and that the translation between extremes indices derived from different reference periods is non-trivial. But secondly, there will still be some use in using the older reference periods (e.g. 1961–1990 in this example), even if there are very few cool days or nights as measured using that threshold, as this will help illustrate how much the climate has changed. This is similar to the way that it is common to link changes in global temperatures back to a pre-industrial value, a climate which no living person has ever experienced (though plenty of long-lived inhabitants of Earth, and also some of our infrastructure has).

There is an analogy which can be drawn between the use of different reference periods and ‘shifting baseline syndrome’ as used in ecological studies. This effect refers to how human perceptions of biological systems change as we lose the experience and knowledge about past conditions (Kahn Jr and Friedman 1995, Pauly 1995). Under a changing biosphere, people either update their own perception of what is normal, or they base it on what was normal from their own experience at e.g. the beginning of their careers. Therefore, as observers leave (retire), much earlier conditions are forgotten. In our study here, by updating a reference period, we ‘forget’ what warm or cool days were in the earlier period, and base all our assessments on the new one. Hence, although using 1961–1990 for a study which looks at cool days out to 2100 could find that almost no days are classed as cool (Stillmann et al 2013) and so could seem a poor choice, that change in the climate is still captured compared to an alternative which updated the reference period to be based on e.g 2061–2090. Hence, it is important to be aware the effect that choices of particular reference periods will have on the conclusions able to be drawn. The WMO guidance allows for the use of different reference periods for different purposes, but as we show here, comparing results from studies using different reference periods is non-trivial (WMO 2007, 2017).

We have focussed here on four specific ETCCDI indices, which were designed to help understand and monitor changes in land-based warm and cool periods. Further indices (WSDI and CSDI) add in a duration (minimum six days) where temperatures exceed the percentile-defined thresholds. We also see the effect of using different reference periods in these indices (see figure S63 of Dunn et al (2020a)), where using a later (warmer) reference period of 1981–2010 results in more cool spells being identified in the early twentieth century, but fewer warm spells in recent years, when compared to 1961–1990, in agreement with Yosef et al (2021). We also note that marine heatwaves are defined very similarly to the WSDI and CSDI indices (Hobday et al 2016), using a seasonally-varying threshold for at least five successive days. As this threshold is usually the 90th percentile, then if the reference period for its calculation is also changed, similar impacts as to those presented here could be observed.

There are also ETCCDI precipitation indices which use a reference period, but these are not constructed in an identical way to the temperature indices presented here. For example, R95p is the total of the rain falling in the wettest 5% of days. This is also available in a normalised form, R95pTOT, which is the same calculation but divided by total annual precipitation. In the appendix we show how our investigations change if we use a Gamma distribution to emulate rainfall, as used by many studies (see table 1 in Ye et al (2018)), with ‘RX10p’ and ‘RX90p’ indices (figure 8). Indices calculated using this distribution show similar differences in the time series and also trends over time to those shown when using the Gaussian. Furthermore, all these features also remain when using a triangular distribution (figure 9), as they are the impact of the interaction of the distribution as it slides past the threshold under a changing climate, rather than being intrinsic to a particular type of distribution.

Yosef et al (2021) note larger changes between their two reference periods in warmer months than cooler ones, implying an asymmetric change and a non-Gaussian shape to the underlying distribution. As a future study it would be interesting to see how a change in shape (e.g. non-stationary standard deviation or skew), or a non-linear change in the mean (to match the varying rate of global warming over the 20th century) would alter these features.

6. Summary

The ETCCDI extremes indices are widely used for the monitoring of current and past extreme events, as well as in comparison with future projections. We have presented a short investigation into how the use of different reference periods when defining a subset of the temperature indices can result in apparently inconsistent behaviour. We have shown that under a warming climate, linear trends calculated over the same period, using the same underlying data will be substantially different when using two separate 30-year reference periods (1961–1990 and 1981–2010). We have simulated this using a simple model, but these effects are also present in a number of separate datasets. The difference in trends for the indices from the upper part of the distribution (TX90p, TN90p) can be large, up to 50 percent larger from the earlier reference period. The direction of change of the trends will depend on which tail of the
distribution the index is assessing, with differences in trends for the indices from the lower end of the distribution (TX10p, TN10p) becoming less.

In essence, the thresholds set under a warming climate by the two reference periods are independent, and trends in day counts exceeding these should not be expected to be the same. Therefore it is important that (a) trends in temperature percentile exceedance statistics are not used interchangeably when derived from different reference periods and (b) the reference period is always stated.

Data availability statement

The data that support the findings of this study are openly available and were obtained from the following sources.

HadEX3.0.3 - www.metoffice.gov.uk/hadobs/hadex3
ECA&D - www.ecad.eu
20CR - https://psl.noaa.gov/data/gridded/data.20thC_ReanV3.monolevel.html

Acknowledgments

We acknowledge the data providers in the ECA&D project, and useful comments from Nick Rayner, John Kennedy and Dan Hollis. Robert Dunn and Colin Morice were supported by the Met Office Hadley Centre Climate Programme funded by BEIS.

Appendix

In figure 7 we show how a shift in the distribution affects the percentiles from the perspective of the CDF. With a warming of 0.5 °C, the shift of the CDF is clear, and by using the same threshold values, their new percentile equivalents are obvious.

In figure 8 we show the effect of changing the underlying distribution from a Gaussian to a Gamma, in order to more closely represent the distribution of daily precipitation. We use a shape parameter of 2, and the trend in the location parameter is also 0.02 mm year$^{-1}$. Similar effects occur with non-stationary differences in the time series and trends over the same period being different when using the two reference periods. In this case the difference curve and trend differences also follow a Gamma distribution, and are comparatively smaller in magnitude.

Finally, in figure 9 we show the effect of changing the underlying distribution from a Gaussian to a Triangular one, then similar effects occur, though in this case the difference curve and trend differences are also triangular.
Figure 7. The definitions of threshold-sensitive ETCCDI indices, and what a changing climate does to these from the perspective of the cumulative distribution function (CDF). The CDF pictured in panel (a) is from a Gaussian with $\mu = 0$ and $\sigma = 1$, with that in (b) $\mu = 0.5$ and $\sigma = 1$. In panel (b) we show the x-axis for the new distribution centred on the new mean value, with the original shown in the dashed grey lines. The changes in the percentile values arising from the thresholds have also been highlighted.

Figure 8. ((a), (b)) Curves of simulated RX10p and RX90p using two different reference periods (1961–1990 and 1981–2010 (RP 61–90 and RP 81–10 respectively), shown with the shaded regions). The expected value over the reference period (36.5 days) is shown by the dashed horizontal line. ((c), (d)) The difference between the two curves plotted in the top panel. ((e) and (f)) The 50-year linear trend calculated at each year for both reference periods. In each panel the empirical curves derived from the parameters of the simulation are also shown, which are the smooth lines.
**Figure 9.** (a), (b) Curves of simulated TX10p and TX90p using two different reference periods (1961–1990 and 1981–2010 (RP 61–90 and RP 81–10 respectively), shown with the shaded regions). The expected value over the reference period (36.5 days) is shown by the dashed horizontal line. (c) and (d) The difference between the two curves plotted in the top panel. (e) and (f) The 50-year linear trend calculated at each year for both reference periods. In each panel the empirical curves derived from the parameters of the simulation are also shown, which are the smooth lines.

**ORCID iDs**

Robert J H Dunn [https://orcid.org/0000-0003-2469-5989](https://orcid.org/0000-0003-2469-5989)

Colin P Morice [https://orcid.org/0000-0002-5656-1021](https://orcid.org/0000-0002-5656-1021)

**References**

Alexander L V, Fowler H J, Bador M, Behrangi A, Donat M, Dunn R and Venugopal V 2019 On the use of indices to study extreme precipitation on sub-daily and daily timescales  *Environ. Res. Lett.* 14 125008

Alexander L V, Zhang X, Peterson T C, Caesar J, Gleason B, Klein Tank A M G and Vazquez-Aguirre J L 2006 Global observed changes in daily climate extremes of temperature and precipitation  *J. Geophys. Res.: Atmos.* 111 D05109

Donat M G and Alexander L V 2012 The shifting probability distribution of global daytime and night-time temperatures  *Geophys. Res. Lett.* 39 L14707

Donat M G, Alexander L V, Yang H, Durre I, Vose R and Caesar J 2013 Global land-based datasets for monitoring climatic extremes  *Bull. Am. Meteorol. Soc.* 94 1097–1106

Dunn R J H, Alexander L V, Donat M G, Zhang X, Bador M, Harold N and Bin Hj Yussof M N 2020a Development of an updated global land in situ-based data set of temperature and precipitation extremes: Hadex3  *J. Geophys. Res.: Atmos.* 115 D05109

Dunn R J H, Perkins-Kirkpatrick S, Schlegel R W and Donat M G 2020b Temperature extremes in the state of the climate in 2019  *Bull. Am. Meteorol. Soc.* 101 S28–S30

Dunn R J H, Willett K M and Parker D E 2019 Changes in statistical distributions of sub-daily surface temperatures and wind speed  *Earth Syst. Dyn.* 10 765–88

Fischer E, Sippel S and Knutti R 2021 Increasing probability of record-shattering climate extremes  *Nat. Clim. Change* 11 689–95

Frich P, Alexander L V, Della-Marta P, Gleason B, Haylock M, Tank A K and Peterson T 2002 Observed coherent changes in climatic extremes during the second half of the twentieth century  *Clim. Res.* 19 193–212

Hobday A J et al 2016 A hierarchical approach to defining marine heatwaves  *Prog. Oceanogr.* 141 227–38

Kahn Jr P H and Friedman B 1995 Environmental views and values of children in an inner-city black community  *Child Dev.* 66 1403–17

Karl T R, Nicholls N and Ghazi A 1999 Clivar/GCOS/WMO workshop on indices and indicators for climate extremes  *Weather and Climate Extremes* (Berlin: Springer) pp 3–7

Klein Tank A M G, Wijngaard J B, Knnen G P, Bhm R, Demare G, Gocheva A and Petrovic P 2002 Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment  *Int. J. Climatol.* 22 1441–53

Klein Tank A, Zwiers F and Zhang X 2009 Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation (WMO-TD No. 1500)

Pauly D 1995 Anecdotes and the shifting baseline syndrome of fisheries  *Trends Ecol. Evol.* 10 430

Peterson T C 2005 Climate change indices  *World Meteorol. Organ. Bull.* 54 83–6

Peterson T, Folland C, Gruzda G, Hogg W, Moksitt A and Plummer N 2001  *Report on the Activities of the Working Group on Climate Change Detection and Related Rapporteurs* (Geneva: World Meteorological Organization)

Salameh A A M, Gmiz-Fortis S R, Castro-Dez Y, Abu Hammad A and Esteban-Parra M J 2019 Spatio-temporal analysis for extreme temperature indices over the levant region  *Int. J. Climatol.* 39 5556–82
Shepard D 1968 A two-dimensional interpolation function for irregularly-spaced data Proc. 1968 23rd ACM Conf. pp 517–24
Sillmann J, Kharin V V, Zwiers F W, Zhang X and Bronaugh D 2013 Climate extremes indices in the CMIP5 multimodel ensemble: part 2. Future climate projections J. Geophys. Res.: Atmos. 118 2473–93
Sillmann J, Kharin V V, Zwiers F W, Zhang X and Bronaugh D 2013 Climate extremes indices in the CMIP5 multimodel ensemble: part 2. Future climate projections J. Geophys. Res.: Atmos. 118 2473–93
Sillmann J, Kharin V V, Zwiers F W, Zhang X and Bronaugh D 2013 Climate extremes indices in the CMIP5 multimodel ensemble: part 2. Future climate projections J. Geophys. Res.: Atmos. 118 2473–93
Slivinski L C et al 2019 Towards a more reliable historical reanalysis: improvements for version 3 of the twentieth century reanalysis system Q. J. R. Meteorol. Soc. 145 2876–908
Thorne P W, Allan R J, Ashcroft L, Brohan P, Dunn R J H, Menne M J and Worley S J 2017 Toward an integrated set of surface meteorological observations for climate science and applications Bull. Am. Meteorol. Soc. 98 2689–702
WMO 2007 The role of climatological normals in a changing climate WMO/td-No. 1377
WMO 2017 WMO guidelines on the calculation of climate normals WMO Technical Report WMO-No. 1203
Ye L, Hanson L S, Ding P, Wang D and Vogel R M 2018 The probability distribution of daily precipitation at the point and catchment scales in the United States Hydrol. Earth Syst. Sci. 22 6519–31
Yosef Y, Aguilar E and Alpert P 2019 Changes in extreme temperature and precipitation indices: using an innovative daily homogenized database in Israel Int. J. Climatol. 39 5022–45
Yosef Y, Aguilar E and Alpert P 2021 Is it possible to fit extreme climate change indices together seamlessly in the era of accelerated warming? Int. J. Climatol. 41 E952–63
Zhang X, Alexander L, Hegerl G C, Jones P, Tank A K, Peterson T C and Zwiers F W 2011 Indices for monitoring changes in extremes based on daily temperature and precipitation data Wiley Interdiscip. Rev.: Clim. Change 2 651–70
Zhang X, Hegerl G, Zwiers F W and Kenyon J 2005 Avoiding inhomogeneity in percentile-based indices of temperature extremes J. Clim. 18 1641–51