Clim. Past, 10, 2171–2199, 2014
www.clim-past.net/10/2171/2014/
doi:10.5194/cp-10-2171-2014

Climate of the Past
Open Access

Investigating uncertainties in global gridded data sets of climate extremes

R. J. H. Dunn[1], M. G. Donat[2], and L. V. Alexander[2]

[1]Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK
[2]ARC Centre of Excellence for Climate System Science and Climate Change Research Centre, UNSW, Sydney, NSW 2052, Australia

Correspondence to: R. J. H. Dunn (robert.dunn@metoffice.gov.uk)

Received: 3 April 2014 – Published in Clim. Past Discuss.: 13 May 2014
Revised: 10 November 2014 – Accepted: 11 November 2014 – Published: 12 December 2014

Abstract. We assess the effects of different methodological choices made during the construction of gridded data sets of climate extremes, focusing primarily on HadEX2. Using global land-surface time series of the indices and their coverage, as well as uncertainty maps, we show that the choices which have the greatest effect are those relating to the station network used or that drastically change the values for individual grid boxes. The latter are most affected by the number of stations required in or around a grid box and the gridding method used. Most parametric changes have a small impact, on global and on grid box scales, whereas structural changes to the methods or input station networks may have large effects. On grid box scales, trends in temperature indices are very robust to most choices, especially in areas which have high station density (e.g. North America, Europe and Asia). The precipitation indices, being less spatially correlated, can be more susceptible to methodological choices, but coherent changes are still clear in regions of high station density. Regional trends from all indices derived from areas with few stations should be treated with care. On a global scale, the linear trends over 1951–2010 from almost all choices fall within the 5–95th percentile range of trends from HadEX2. This demonstrates the robust nature of HadEX2 and related data sets to choices in the creation method.

1 Introduction

Understanding the uncertainties present within a data set can enable better decision making, as well as enhancing further research applications (Matthews et al., 2013). Uncertainties arise from a variety of sources, including unknowns in the underlying data, parameters chosen when processing them, and differences between the methods used to do the processing. In some cases the uncertainties can be calculated and combined to give final ranges in the data product. In many cases the processing is too complex for this to be achievable, and so ensemble data sets have been produced to sample the range of possible plausible final outcomes (e.g. HadCRUT4; Morice et al., 2012). To quantify the effect of processing methods on the underlying data, benchmark data sets have been used (e.g. USHCN; Williams et al., 2012), and in some cases the methods themselves have been assessed in this way (e.g. COST-HOME; Venema et al., 2012). In other cases, multiple data products have been produced in a number of institutions that sample the methodological (structural) uncertainties, e.g. the global surface temperature record from HadCRUT4, MLOST (Smith et al., 2008), NASA-GISS (Hansen et al., 2010) and Berkeley (Rohde et al., 2013). If the results obtained from different methods and parameter choices remain similar, they can be considered robust and conclusions can be drawn with high confidence.

In recent years a number of gridded data sets of climate extremes indices have been released. The first global data set to contain all 27 indices recommended by the World Meteorological Organization (WMO)/CLIVAR/JCOMM Expert Team on Climate Change Detection and Indices (ETCCDI) was HadEX (Alexander et al., 2006), which was based on work by Frich et al. (2002). HadEX covered the period 1951–2003 and was static, i.e. not updated. Over recent years an international effort aimed to bring the data set up to date. This work resulted in HadEX2 (Donat et al., 2013a), which

Published by Copernicus Publications on behalf of the European Geosciences Union.

substantially increases the time span covered to 1901–2010 and also has a greater surface coverage of the globe.

A number of partner data sets to HadEX2 also exist which follow a similar methodology. GHCNDEX (Donat et al., 2013b) uses only the Global Historical Climate Network (GHCN)-Daily data set as input, rather than the mix of large collections of data and country-based inputs to HadEX2. As GHCNDEX is continually updated in near-real time, it can also be used for climate monitoring. The gridded HadGHCND data set has also been used to calculate these indices (Donat et al., 2013b) using temperature data only.

On the whole, for the temperature-based indices, the different extremes indices data sets agree, both for global averages and regional trends (Donat et al., 2014). Some systematic differences were found for the values of the absolute indices: as expected, those calculated from the daily gridded data (HadGHCND) were less extreme (lower maxima and higher minima; Donat et al., 2013b). However for the precipitation indices, the agreement between HadEX and GHCNDEX is less robust (Donat et al., 2013b), though the agreement improves when the data sets are masked to have the same coverage. This demonstrates the large effect that the spatial coverage of the data sets has, especially for the precipitation indices. The precipitation indices have more complex spatial changes, and hence the global averages are more sensitive to changes in coverage than the temperature indices (see also Wan et al., 2013).

With the increasing use of these data sets to assess regional changes in extremes, we need to assess whether these results are unduly sensitive to any choices in the methods. An assessment of this nature probes the parametric and structural uncertainties of the data set. The Intergovernmental Panel on Climate Change (IPCC) define the parametric uncertainty as that coming from choices of parameters within the analysis scheme and the structural uncertainty as from choices of the scheme itself (Hartmann et al., 2013, Box 2.1), and these are the definitions we shall use here. For a complex method, choices of the order of a calculation (e.g. calculating grids of annual extremes or extremes from daily grids) or values of a threshold parameter may have unexpected consequences for the final outcomes.

Both HadEX2 and GHCNDEX calculate the extremes indices for each station, and then average the stations to form a gridded data set. Extreme (temperature) indices were also calculated from daily temperature grids provided through the HadGHCND data set (Caesar et al., 2006). This helped to assess whether scaling issues related to the order of processing when calculating grids of annual extremes (i.e. calculate grids from annual extremes vs. annual extremes from daily grids) may affect analyses of global trends. Such comparison is relevant, e.g. when using the gridded data sets of observed extremes as reference for climate model evaluation (for which extremes are generally calculated from daily output fields). Donat et al. (2014) documented that these different approaches in the calculation of extremes grids lead to differences in the actual values, but trend estimates appeared to be largely robust for temperature indices. No such comparison has yet been performed for precipitation extremes, due to the lack of long-term global daily grids of observed precipitation.

Recently, Sillmann et al. (2013) compared extremes indices calculated from state-of-the-art global climate models participating in the CMIP5 to four reanalysis data sets, and Donat et al. (2014) compared the three in situ data sets described above to five reanalysis products. Furthermore, Yin et al. (2014) compare five data sets of a subset of these indices over China. Assessing whether areas of disagreement between the models, reanalyses and observational data can be reduced by more fully understanding the uncertainties associated with the observational data sets is an important step in their use. But more generally, having an estimate of the uncertainties in the observational data sets is vital for their accurate use. With the advent of coordinated efforts to provide climate services, policy and planning decisions are increasingly being made using insight from observational data sets combined with model analyses. As a result it can be highlighted where results from data sets are robust and, moreover, where they are not.

Here we will focus specifically on HadEX2, but, as noted above, the results are applicable to all data sets which follow a similar calculation method. We first outline the HadEX2 methods relevant to this work in Sect. 2 and the effect of the completeness requirement used when plotting global average time series (Sect. 3). The individual methodological choices are presented in Sect. 4. We discuss our findings in Sect. 5, and Sect. 6 summarises the study.

2 HadEX2 methods

In this section we will outline, in particular, the methodological choices (parametric and structural) which will be assessed below. For a full description of the methods used to create HadEX2 see Alexander et al. (2006) and Donat et al. (2013a).

Between 6500 and 7500 stations are available for each temperature index, and around 11500 stations for the precipitation indices. The geographical distribution of the stations can be seen in Fig. 1 of Donat et al. (2013a). For both temperature and precipitation, the Amazon region, large parts of Africa and the southern Arabian Peninsula have no stations. The western Australian desert and the high-latitude regions of Russia are also relatively undersampled. We note that all these extremes index data sets are restricted to the land surface, and that their coverage varies between time steps (monthly or annually). Therefore, although we follow Donat et al. (2013a) and Alexander et al. (2006) and describe the time series as “global average”, they are not truly global.

To perform the gridding of the station data, HadEX2 uses a modified form of Shepard’s angular distance weighting (ADW) scheme (Shepard, 1968). It was initially chosen

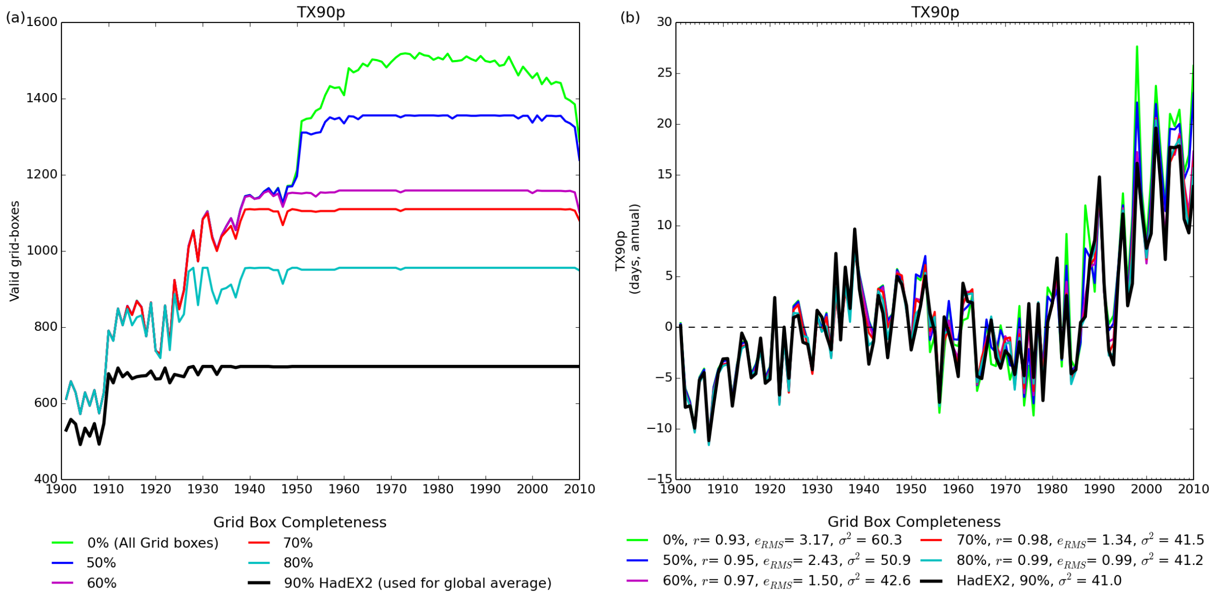

Figure 1. (a) Numbers of grid boxes covered given different completeness requirements, for TX90p ($T_{\max} > 90$th percentile). The total coverage is shown in light green, whereas the coverage used when calculating global averages is in black, with a range of other completeness percentages shown in between. **(b)** The time series for global average of TX90p of HadEX2 (black) and the completeness percentages from panel **(a)**. Also shown are the comparisons of other completeness criteria to that used in HadEX2 (90 %) using the correlation coefficient, r, the root-mean-square error e_{RMS} and the variance, σ^2.

when creating HadEX because it had been shown to be an appropriate method for gridding irregularly spaced data (New et al., 2000). For each index, the correlation coefficients between all station pairs are calculated and are plotted against the distance between the stations. The correlation coefficients decay with distance, and averaging over 100 km bins, this decay is fitted with second-order polynomial (see Fig. A1 of Alexander et al., 2006). We assume that the bin at zero distance has perfect correlation (with a data point at (0,1)), but the best-fit line is not forced to pass through this point. Small instrumental and siting effects (on the scale of metres and so far below the scale of the bins used) are likely to result in a non-unity correlation at zero distance. Using this polynomial fit, the distance at which the correlation has fallen by a factor of $1/e$ is obtained. This distance is the decorrelation length scale (DLS, also known as the correlation decay distance; see Caesar et al., 2006; Jones et al., 1997). Furthermore, as part of the ADW scheme, there is a weighting parameter, m, which determines the steepness of the decay with distance. In HadEX2 (and the other related data sets), this weighting function has been chosen to be $m = 4$ (it is the effect of parametric choices like this that are investigated in the course of this study). A DLS is calculated for each index individually at the timescale of the index (monthly or annually) and for five separate latitude bands: four 30° bands between 30° S and 90° N and one 60° band between 30 and 90° S, where there are few stations. The DLS values are then linearly interpolated to avoid discontinuities at the band boundaries.

A $3.75° \times 2.5°$ grid is used for HadEX2. For each grid box centre, all the stations within a DLS are combined using ADW to obtain the value for the grid box in HadEX2. There have to be a minimum of three stations within a radius of one DLS for a grid box value to be calculated. This means that there exist grid boxes which themselves do not contain any stations but are just in sufficient proximity to three stations to have a value assigned. These annual (and monthly in the case of some indices) grid box values make up the final data set.

HadEX2 contains a total of 29 extremes indices, 17 temperature-based and 12 precipitation-based, which are defined in Table 1. Some indices are calculated in a similar way to others, and thus fall naturally into categories. For temperature there are percentile-based ones (TX10p, TX90p, TN10p, TN90p), block maxima/minima (TXx, TXn, TNx, TNn), duration-based indices (CSDI, WSDI, GSL), exceedance frequency of fixed threshold values (SU, TR, FD, ID) and others which do not fit into any of the above categories (DTR, ETR); for precipitation there are exceedance frequency of fixed threshold values (R10mm, R20mm), precipitation totals (PRCPTOT, R95p, R99p), block maxima (Rx1day, Rx5day), percentile-based precipitation totals (R95pTOT, R99pTOT, SDII) and duration-based indices (CDD, CWD). For this study we use the version of HadEX2 as of May 2013.

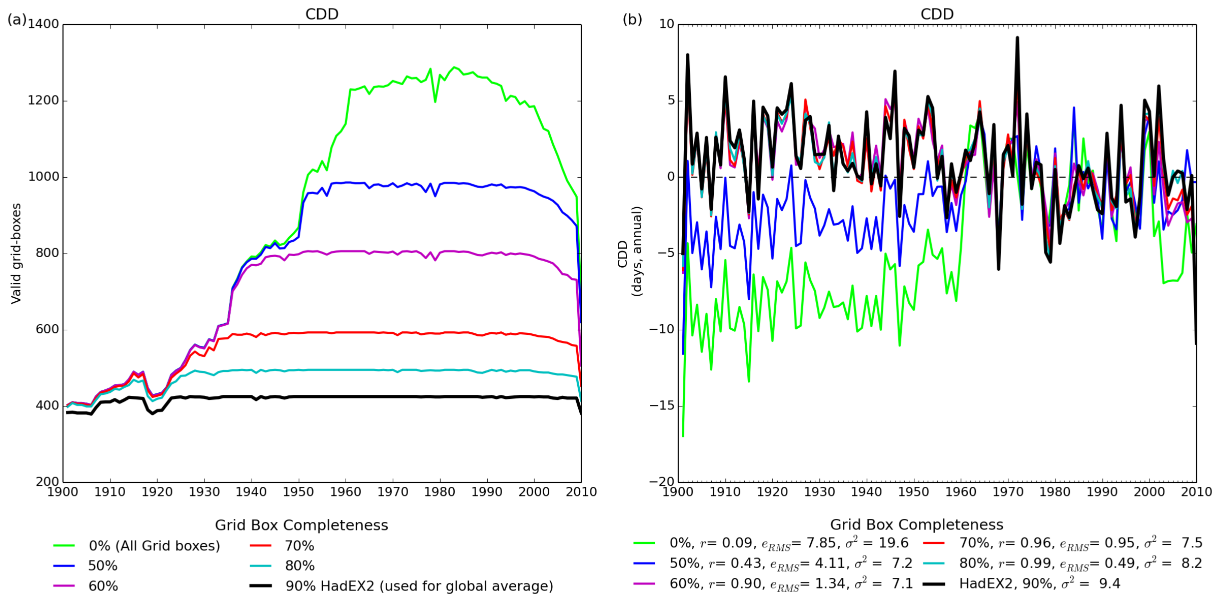

Figure 2. (a) Numbers of grid boxes covered given different completeness requirements, for CDD (consecutive dry days). **(b)** The time series for global average of TX90p of HadEX2 and different completeness requirements. For further details see Fig. 1.

3 Grid box completeness

In the calculation of the global average, Donat et al. (2013a) use only those boxes which have at least 90 % of data during the period, i.e. 99 years over 1901–2010. This is to reduce the effect of varying coverage on the global average, which would otherwise change drastically as regions start and stop contributing to this summary measure. However this requirement adds a large restriction onto the areas of the globe which can contribute to the global land-surface average. We investigate this effect separately from the rest of the methodological changes as it influences them all when calculating the global time series. However, it is more of a presentational choice when calculating the global average summary time series. We note that all of the global time series plots have been centred over the period 1961–1990 as per Donat et al. (2013a). This does have the effect of appearing to reduce the spread during the climatological period used for the centring process.

As can be seen in Fig. 2 of Donat et al. (2013a), there is a steady climb in the number of grid boxes with data until around 1960, where the curve levels off, before beginning to fall again in the early 2000s. When the number of grid boxes is restricted to those which have 90 % completeness over the period of the data set, the number of grid boxes is much reduced but is more constant over the period. For TX90p ($T_{\max} > 90$th percentile), the number of grid boxes with 90 % completeness remains steady at around 700 for most of the period (see Fig. 1a), but depends on the index used and is generally lower for the precipitation indices, e.g. CDD (consecutive dry days, Fig. 2).

In Fig. 1 we also show the root-mean-square error (e_{RMS}) and the variance (σ^2) for each of the completeness criteria. The variance is calculated from the global mean time series using the full record. The root-mean-square error is calculated on the difference of the time series with that from HadEX2, again over the full record. These two quantities can be used in combination with a visual assessment of the global time series to determine how similar they are.

When using a percentile-based temperature index, e.g. TX90p, there is only a relatively small effect of the completeness requirement on the global trends (see Fig. 1b). The global average time series are highly correlated with HadEX2 for all completeness values, and only for very low completeness criteria in time (less than around 50 % of years present for a grid box to be included) are there somewhat larger deviations, and even these are only apparent in the most recent decades of the time series where the grid box numbers diverge by a large amount. The correlation coefficients remain above 0.9, indicating close agreement between the time series, which can be clearly seen in Fig. 1b.

However, when using another index, e.g. CDD (see Fig. 2), completeness requirements have clear systematic effects on the results. Although year-to-year variations are very similar, the global averages before 1960 are smaller for low completeness criteria. By including extra-short-term grid boxes, the period post 1960 has been biased upwards, which, combined with the climatology period of 1961–1990, results in the apparent low values prior to this. Specifically for CDD, the results when selecting only long-term stations (see Sect. 4.4) indicate that stations in central and eastern Asia report mainly in the later period of the data set and HadEX2

Table 1. The abbreviations, definitions and units of all the indices assessed in this work.

Index	Name	Definition	Unit
Temperature			
TXx	Max T_{max}	Warmest daily maximum temperature	°C
TXn	Min T_{max}	Coldest daily maximum temperature	°C
TNx	Max T_{min}	Warmest daily minimum temperature	°C
TNn	Min T_{min}	Coldest daily minimum temperature	°C
DTR	Diurnal temperature range	Mean difference between daily maximum and daily minimum temperature	°C
ETR	Extreme temperature range	Difference between monthly maximum and minimum temperature	°C
GSL	Growing season length	Annual number of days between the first occurrence of 6 consecutive days with $T_{mean} > 5$°C and the first occurrence of 6 consecutive days with $T_{mean} < 5$°C. For the Northern (Southern) Hemisphere this is calculated between 1 January and 31 December (1 July to 30 June).	Days
CSDI	Cold spell duration indicator	Annual number of days with at least 6 consecutive days when $T_{min} <$ 10th percentile	Days
WSDI	Warm spell duration indicator	Annual number of days with at least 6 consecutive days when $T_{max} >$ 90th percentile	Days
TX10p	Cool days	Percentage of days when $T_{max} <$ 10th percentile	% of days
TX90p	Warm days	Percentage of days when $T_{max} >$ 90th percentile	% of days
TN10p	Cool nights	Percentage of days when $T_{min} <$ 10th percentile	% of days
TN90p	Warm nights	Percentage of days when $T_{min} >$ 90th percentile	% of days
FD	Frost days	Annual number of days when $T_{min} < 0$°C	Days
ID	Ice days	Annual number of days when $T_{max} < 0$°C	Days
SU	Summer days	Annual number of days when $T_{max} > 25$°C	Days
TR	Tropical nights	Annual number of days when $T_{min} > 20$°C	Days
Precipitation			
Rx1day	Maximum 1-day precipitation	Maximum 1-day precipitation total	mm
Rx5day	Maximum 5-day precipitation	Maximum 5-day precipitation total	mm
PRCPTOT	Annual contribution from wet days	Annual sum of daily precipitation ≥ 1.0 mm	mm
SDII	Simple daily intensity index	Annual total precipitation divided by the number of wet days (when total precipitation ≥ 1.0 mm)	mm day^{-1}
R95p	Annual contribution from very wet days	Annual sum of daily precipitation > 95th percentile	mm
R95pTOT	Fraction from very wet days	R95p × 100/PRCPTOT	%
R99p	Annual contribution from extremely wet days	Annual sum of daily precipitation > 99th percentile	mm
R99pTOT	Fraction from extremely wet days	R99p × 100/PRCPTOT	%
CWD	Consecutive wet days	Maximum annual number of consecutive wet days (when precipitation ≥ 1.0 mm)	Days
CDD	Consecutive dry days	Maximum annual number of consecutive dry days (when precipitation < 1.0 mm)	Days
R10mm	Heavy precipitation days	Annual number of days when precipitation > 10 mm	Days
R20mm	Very heavy precipitation days	Annual number of days when precipitation > 20 mm	Days

shows high values of CDD in western China (Xinjiang and the Tibetan Plateau), which supports this reasoning. This highlights the importance of requiring a high completeness when constructing the global average time series. Many other indices (both temperature and precipitation) show large differences between the time series of HadEX2 and versions when the completeness is ≤ 50 %, especially outside of the climatology period (1960–1990). The indices which show no large changes tend to be the ones based on percentiles (e.g. TX90p; R99p: annual sum of daily precipitation > 99th percentile).

The correlation coefficients between HadEX2 and the versions using different completeness requirements for both CDD and TX90p remain above 0.9 until the completeness drops below around 60 % of years, and similar results are seen in the plots for the remaining indices (see Supplement).

Therefore, if studying the full time span of the HadEX2 data set (1901–2010), then a completeness criterion of at least 60 % would be required (66 years present out of 110 for the grid box to be included). However, the higher the completeness threshold is set, the more reliable the results will be, albeit for a smaller fraction of the globe. This has less of an effect in the later period when more data are available. The recommendation for masking the data based on the completeness of the grid boxes over time when assessing time series of area averages was also highlighted by Donat et al. (2013b).

However, although the effect of selecting only the grid boxes which have at least 90 % of years with non-missing data is large, we will compare the remaining methodological choices with this restriction in place rather than reducing the requirement to 60 %. Plots showing the number of grid boxes filled by each methodological choice will be shown without this restriction in place in order to demonstrate changes in this quantity more clearly. Any differences to this approach will be noted in the text and figure captions.

4 Uncertainty investigations

As HadEX2 contains 29 indices, in around 7–8 categories (Sect. 2), and as we are studying a range of methodological choices, we only show representative results judged likely to be of greatest interest. In the majority of cases, the uncertainties will be very similar for indices in the same category. All plots for each of the indices will be provided in the Supplement.

We will present the small, parametric changes first as these are likely to have the smallest effect. Then we will focus on larger changes to the source data or methods: "structural changes". In all cases, the time series have been centred relative to the period 1961–1990. We do not assess the effect of the quality of the station data during this work. HadEX2 has been created from a number of different input data sources (see Donat et al., 2013a, for more details) and therefore has differing levels of quality control applied between stations and not all of the raw data are available for independent quality control assessments to be made. For those indices where monthly values are available, in this work we only assess the effects on the annual values.

4.1 Weighting function (parametric)

In the ADW scheme used in HadEX, a parameter, m, determines the steepness of the decay of the weighting function with distance (Eq. A2 in Alexander et al., 2006). The weighting parameter has been set to $m = 4$ in both HadEX and HadEX2, as this provided a reasonable compromise between reducing the root-mean-square error (e_{RMS}) between the gridded and station data and spatial smoothing (Donat et al., 2013a; Caesar et al., 2006). Donat et al. (2013a) note that the results are almost identical when using values between 1 and 10 for m.

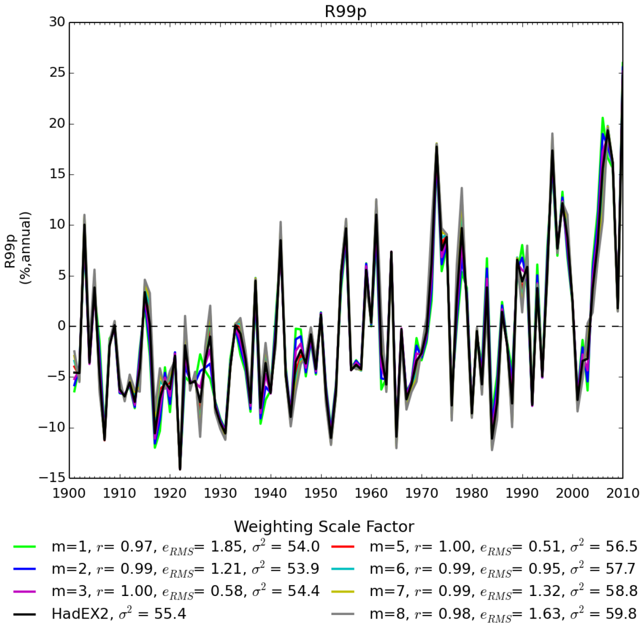

Figure 3. The time series for global average R99p (annual sum of daily precipitation > 99th percentile) showing the different curves for the different weighting values (colours are indicated below the plot). The figure also shows the correlation coefficient (r), variance (σ^2) and root-mean-square error (e_{RMS}) between each series and HadEX2. All choices have been centred relative to the climatology period 1961–1990, reducing the spread during this time.

We vary this weighting parameter from one to eight, and show the results in Fig. 3 for the R99p index (annual sum of daily precipitation > 99th percentile). Indeed, the changes are very minor, and are almost imperceptible on the global time series plot. For a large m, the decay of the weighting function is steep, which would lead to stronger gradients locally. A small m results in a slower decay, weaker gradients and hence smoother variation across grid boxes. Hence changes in m may lead to small local differences, but as seen in Fig. 3, global-scale results are almost identical for all values of m. This is probably also valid for (sub)continental averages.

For all other indices, there are also only very small changes to the global time series, and the correlation coefficients rarely differ from $r = 1$. Any differences that do exist tend to be at earlier times when the coverage is lower. The coverage (grid boxes with non-missing values, not shown) is also identical to that of HadEX2.

4.2 Stations within a DLS (parametric)

In HadEX2, for a grid box to have a value, there have to be at least three stations within a radius of one DLS of the grid box centre. For the percentile-based temperature indices this

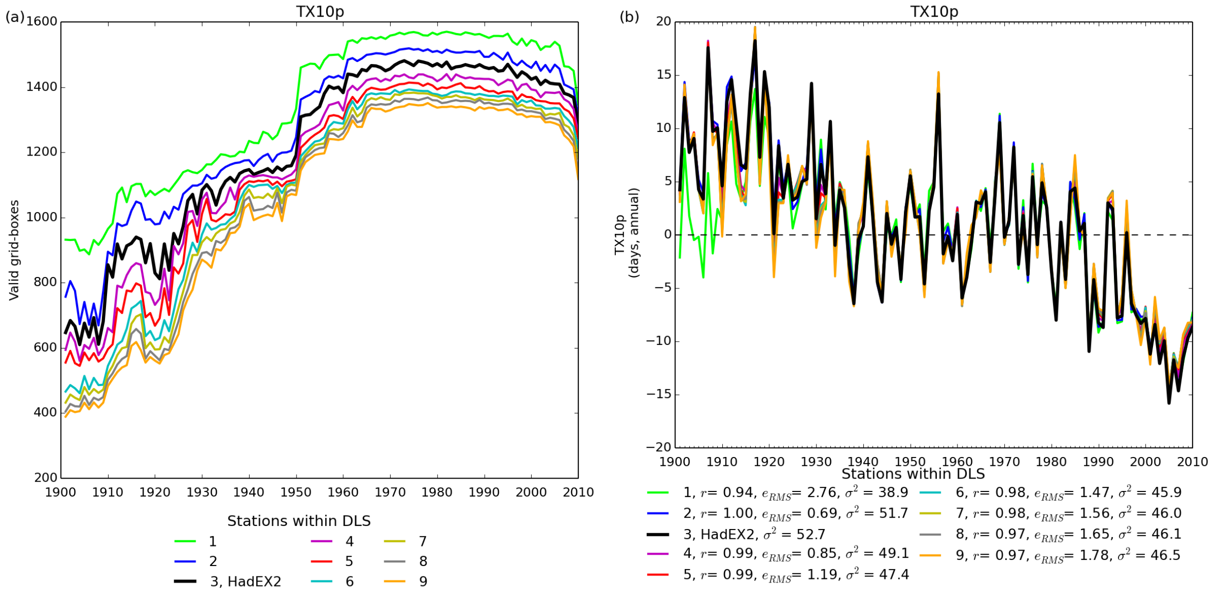

Figure 4. **(a)** The number of non-missing grid boxes for TX10p ($T_{\max} <$ 10th percentile) for the different numbers of stations required within a DLS of the grid box centre. All grid boxes with sufficient stations are shown in the coverage series. **(b)** The time series for global average TX10p for the different numbers of stations required within a DLS of the grid box centre, given the overall 90 % completeness requirement. All choices have been centred relative to the climatology period 1961–1990, reducing the spread during this time.

DLS can be of the order of 1000 km, but it is much shorter for the precipitation-based indices. We vary this parametric choice between one and nine stations within one DLS of the grid box centre. The resulting global average coverage and time series for TX10p ($T_{\max} <$ 10th percentile) are shown in Fig. 4. Again, the changes between choices are very minor in this index for the later period, though larger than for the choices of the weighting parameter. The correlation between the different time series remains very high, but as the number of stations required within a grid box rises, so does the root-mean-square error (e_{RMS} in Fig. 4b). Pre-1950 the global average with only one station required within a DLS results in lower estimates for the indices in almost all cases (see Supplement) notably changing the long-term behaviour.

For some of the precipitation indices (e.g. R99p, annual sum of precipitation > 99th percentile) the correlation with HadEX2 also decreases as more stations are required per DLS. The long-term behaviour, which in many indices shows no strong non-zero trend, is unchanged, but the year-to-year variability changes. As discussed below, this arises because of the severe reduction in coverage as the number of stations required per DLS increases (from ∼ 1100 grid boxes to ∼ 200 for this index).

When masking all of the nine choices by the coverage of the version requiring nine stations within one DLS (the most restrictive choice), all the global time series curves are identical. This means that the changes in the global time series curves seen in Fig. 4b are driven by the coverage and not by changes in the grid box values. It is highly likely that there are changes in individual grid box values or on regional scales. However, as the ADW gridding method gives the highest weight to stations which are nearest to the grid box centre with a decay to more remote stations, these changes are expected to be small (especially in dense networks). Therefore, on a global average, there are no apparent changes resulting from changing the DLS but keeping the same coverage.

Figure 5 shows how choices that affect regional level results can cancel out when considering the global (land-surface) average. The correlation coefficient (r) of the local detrended time series with HadEX2 for each of the eight other possible choices is calculated over the 1951–2010 period. The mean of these for each grid box is shown in Fig. 5a. We use the correlation coefficient of the detrended time series in order to pick out the short-timescale variability rather than any long-term trend that may dominate in some indices for r values of the raw time series. The linear trend used to detrend the data was determined using a median of pairwise slopes (Theil–Sen) estimator (Theil, 1950; Sen, 1968; Lanzante, 1996) over the period 1951–2010. We require at least two-thirds (40 out of 60) of years with valid data for a trend to be calculated.

In some grid boxes only one of the possible versions results in a value; these have been shaded in grey. They arise from the choice which only requires one station per DLS as this is the least restrictive. In the case of TX10p, the correlation coefficients between the different versions are very high for most of North and South America, Europe, Asia and

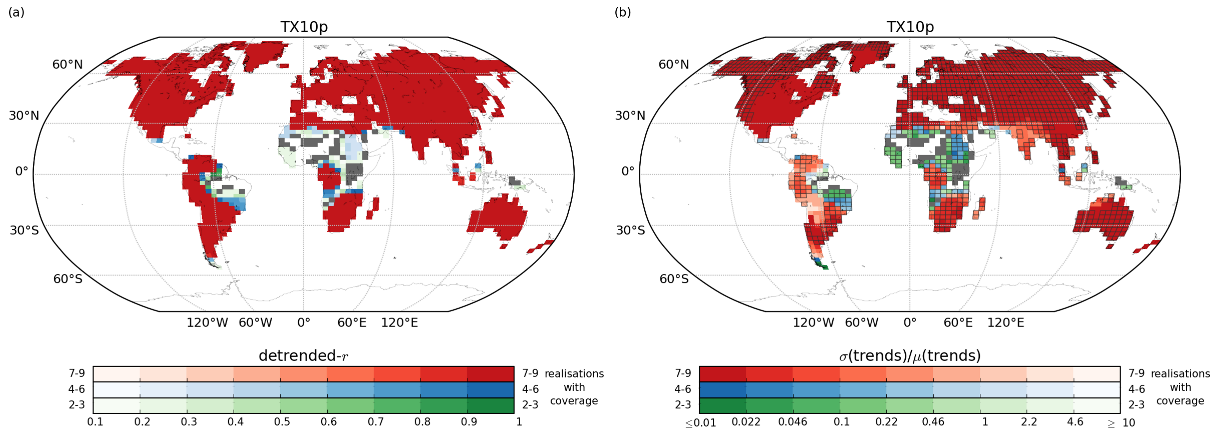

Figure 5. (a) The mean correlation coefficient of the detrended time series, r with HadEX2 for all the grid boxes and **(b)** the standard deviation of the trends divided by the mean trend (σ/μ) calculated for the period 1951 to 2010 for TX10p. Grey grid boxes are those where only one of the nine possible options for the number of stations within the DLS results in a value, green where two to three, blue where four to six, and red where seven to nine. In the right panel, boxes which have been outlined are those where there is high confidence in a non-zero trend in HadEX2.

Australia, and southern Africa, though in some cases not all of the choices result in a value for a specific grid box. The only large regions which have low correlation values or few choices which fill the grid boxes are around central and Saharan Africa, the Amazonian region and Indonesia. This picture is largely unchanged for the other temperature indices. In some indices there are fewer choices which result in a value in the high latitudes, parts of central Asia and central Australia, as well as the regions mentioned above. For other indices, almost the entire globe is covered. Precipitation indices have a much smaller area where all choices result in a value (see R99p in Fig. 6a) as a consequence of the much smaller DLS value in most cases.

As the representation of long-term trends is also important, we show in Fig. 5b the standard deviation of the linear trends divided by the mean of the trends (σ/μ) calculated over 1951–2010. Linear trends for all methods were calculated using the median of pairwise slopes method, and from the resulting distribution of trends the standard deviation and mean were obtained. If the value of σ/μ is small, then there is little variation in the value of the trends compared to the size of the mean trend, and so the trends are robust to the different choices. However, if the value of σ/μ is large, then there are large variations in the trends, and so they are not robust. This is more likely if the value of the mean trend is small. On the Indian subcontinent and also in South America, although the mean r values of the detrended time series were high, there is a higher variation in σ/μ than in North America, Europe and Asia.

The confidence of the sign of a trend is also determined from the median of pairwise slopes method by requiring that both the 5th and 95th percentiles of the slopes have the same sign. In this way we can be confident that a non-zero trend exists, and have some indication of its magnitude. Grid boxes where this is the case in HadEX2 are highlighted with a solid surround. This is different to the way trend significance was calculated in Donat et al. (2013a), who used the Mann–Kendall test.

We note that in many cases a linear trend is not a good descriptor for the long-term behaviour of the indices. Therefore we also provide figures in the Supplement which use the difference between the early (1951–1970) and late (1991–2010) periods of this data set. In these figures, a change is significant if the ranges of the median-absolute deviations from the two periods do not overlap. The dominant differences between these plots and those from the linear trends are the grid boxes which are assessed to have significant changes, with fewer occurring when using the differences.

However, linear trends are a simple and well-understood way of summarising the long-term behaviour of a time series (Hartmann et al., 2013, Box 2.2). Also, by restricting the period to post-1951, we exclude the early period for which trends in the global mean are more dependent on coverage (see Sect. 3). By using a linear trend we focus on one (easily understood) measure of the change in the indices since 1951, especially as we are assessing the similarity of the low-frequency variability between the versions rather than trying to remove it. More complex analysis methods, for example change-point detection methods, could be used to identify dramatic changes in the behaviour of the indices between different versions of the data set. However, as in this work we are assessing the effect of parametric and structural choices of the method on the behaviour of the indices rather than the behaviour of the indices over time, a linear trend provides a simple and easily understood way of summarising the long-term changes for this regional analysis.

For some of the fixed threshold indices (FD: frost days; ID: ice days; SU: summer days; TR: tropical nights) there

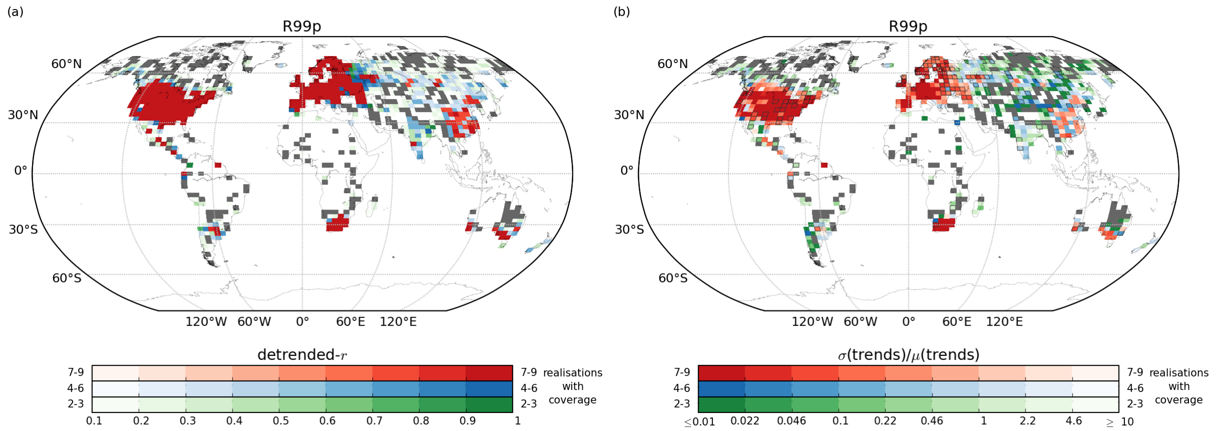

Figure 6. **(a)** The mean correlation coefficient of the detrended time series, r with HadEX2 for all the grid boxes and **(b)** the standard deviation of the trends divided by the mean trend (σ/μ) calculated for the period 1951 to 2010 for R99p. For further details see Fig. 5.

are some very high correlation values and very low variances, especially in regions of the globe where these thresholds are always or never exceeded. However, in these regions these indices are of limited value. Correlations are not calculated if no data are present, so that the coverage can look different to other indices. Also, for some stations, these indices were not calculated at regional workshops[1] when the fixed thresholds were irrelevant for a specific climatic region.

The other temperature-based indices show similar features in both the time series and the maps, albeit with slightly different coverage resulting from the differences in the DLS for each index. North America, Europe and Asia have most of the choices resulting in most boxes having a value, whereas South America, southern Africa and Australia have more variable coverage. In most of these areas, the r values of the detrended time series are high, but the reduced trend variance (σ/μ) also shows high values in South America, Africa, India, central Asia and high latitudes for some of the indices. Most temperature indices also show clear non-zero trends in the global average. The high level of agreement between the choices for the reduced variances (darker colours) indicates that these trends are robust to this methodological choice in most regions.

The DLSs for most of the precipitation indices are much smaller than for the temperature indices. Hence the areas which consistently have most of the choices resulting in a grid box value are much smaller: only the densely instrumented parts of North America and Australia, Europe, South Africa, and parts of South America, India and China (compare Fig. 6 with Fig. 5) have seven or more choices resulting in a grid box value. However, even in some of these areas, the reduced variances are quite high. This is in some cases due to the long-term behaviour showing no strong non-zero trend in some of these regions and indices. Hence, here is a high sensitivity to changes arising from the reduction in coverage as the number of stations required within a DLS is increased.

There is no clear correspondence between the location of grid boxes where there is high confidence in a non-zero trend in HadEX2 and those regions where the σ/μ is small for the precipitation indices. Most of the precipitation indices have few regions where there is a notable non-zero trend (PRCPTOT is an exception). In fact, some areas with high σ/μ have high confidence in non-zero trends in HadEX2. Hence, even if there is strong evidence for a non-zero trend, further investigation would be required to ensure its significance as it may well be extrapolated from distant stations.

4.3 Stations within a grid box (parametric)

The ADW gridding method of HadEX2 does not require that a given grid box contain any stations within it but merely that there be at least three stations within one DLS of the grid box centre (see Sect. 4.2). Thus there could be a number of grid boxes without any actual nearby stations that have values because of the ADW interpolation. We now require that there be stations within the grid box itself (ranging from one to five). This will assess the complementary effects of the robustness of a grid box average against the robustness of a regional average. The weighting used to calculate the grid box value is the same as in the ADW scheme, but boxes are only filled if there are the requisite number of stations inside them. Requiring even just one station within the grid box has a drastic effect on the coverage (see Fig. 7a).

For the global time series, the requirement that a grid box contain stations has little effect on the post-1940 trends (Fig. 7b). There are, however, large differences prior to this. In regions with relatively high station density (and hence also more in the most recent periods), the grid box value will still

[1] When creating HadEX2, some data came from regional workshops. During these workshops, indices were calculated from the daily data to overcome some of the concerns about data sharing and exchange. For further details see Alexander et al. (2006) and Donat et al. (2013a).

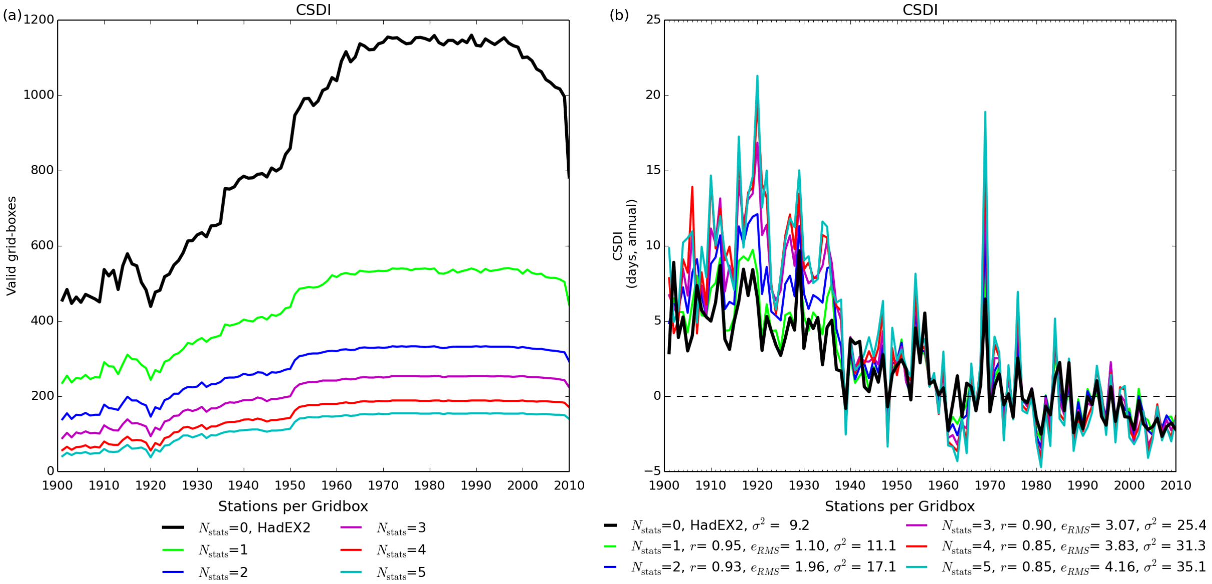

Figure 7. (a) The number of non-missing grid boxes and **(b)** the time series for global average CSDI (cold spell duration indicator) for the different numbers of stations within a grid box. Note that the global average only takes grid boxes which have 90 % completeness or more, whereas all grid boxes are shown in the coverage series. All choices have been centred relative to the climatology period 1961–1990, reducing the spread during this time.

be dominated by the stations within the box. Therefore finding only few differences in the later period is expected. Although HadEX2 has a greater coverage through ADW interpolation, the number of stations that are included are the same. The extra coverage in HadEX2 is extrapolated from data from these stations, and so it would be very surprising if there were drastic changes in the global average trends.

By using the correlation information between the observed stations when calculating the DLS, we can be confident that this correlation is still valid in grid boxes where no station is found (North et al., 2011). Hence using the ADW gridding scheme which interpolates into regions without stations is reasonable. However, the result of this is that the globally averaged time series of the interpolated and un-interpolated version are very similar as no new information has been added during the interpolation. For the earlier period, requiring a greater and greater number of stations to be present in the grid boxes has a larger effect as there are fewer stations and the coverage becomes very small. If we were to calculate decadal averages of the indices, the DLS of these time averages would likely be larger than the DLS used for the construction of HadEX2. Therefore decadal averages of the indices may be representative of a larger area than just that covered in HadEX2. One note of caution is that this is likely only true for decadal averages, but not for decadal versions of some of the indices, i.e. decadally averaged TXx would be expected to be representative of a larger spatial area than the DLS values calculated for the monthly and annual grids; however, this would not be the case for the maximum TX over the entire decade.

Although the long-timescale trends for all the choices are similar for the latter period, what is apparent is that some individual short-term spikes become more pronounced as the number of stations required increases. If these features in the global average arise from small geographical regions with a high station density, then these can more easily dominate the global average as the number of stations per grid box increases and there are fewer and fewer grid boxes which have valid values.

The uncertainty maps for this set of methodological choices are shown for CSDI (cold spell duration indicator) in Fig. 8. There are many regions where only one choice gives a value, which are shown in grey, and these are from the HadEX2 data set. The more stations required to be within a grid box, the higher the mean correlation of the grid box time series.

Where only two of the choices result in grid box having a value, there must be one station in the grid box (the other choice being HadEX2). As this station dominates the value of the grid box, the time series can be noisy, and hence the correlation with HadEX2 can be low. If more stations are required, the time series will become less noisy as each station contributes less to the grid box average time series. Concurrently, stations within a grid box are in close proximity, and hence are likely to correlate well. Therefore the mean correlation between all the series will improve with an increase in the required number of stations, which is what is observed in Fig. 8a (boxes coloured green are very pale, blue less so, and red ones are the most intense). For grid boxes where not

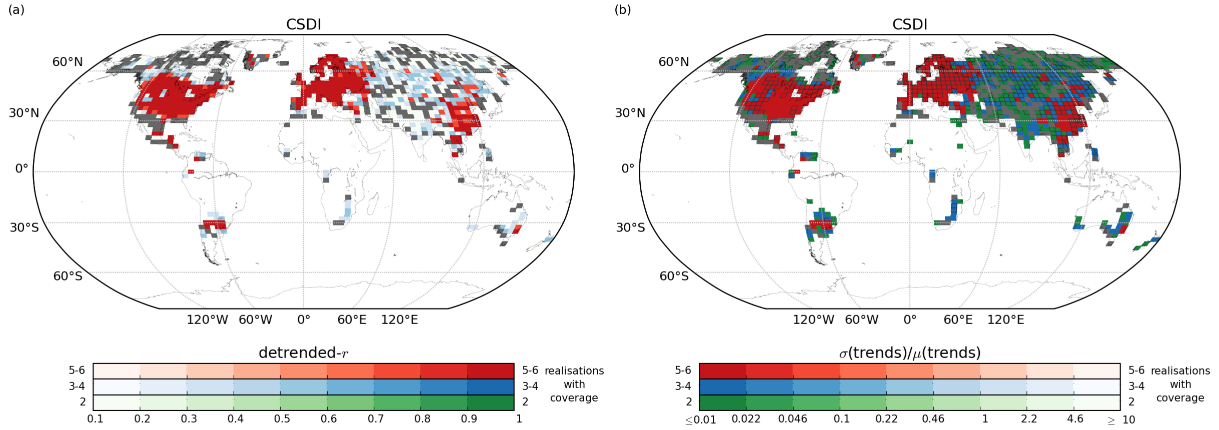

Figure 8. (a) The mean correlation coefficient of the detrended time series, r with HadEX2 for all the grid boxes and **(b)** the standard deviation of the trends divided by the mean trend (σ/μ) calculated for the period 1951–2010 for CSDI and the number of stations per grid box. Green grid boxes are those where two of the six possible choices of stations per grid box result in a value, blue where three to four, and red where five to six. Note that the HadEX2 method does not require any stations to be present within a grid box for a value to be calculated. For further details see Fig. 5.

all the choices result in a value, there is a greater range in differences than in the trends (less intense colours).

For all indices the coverage reduces to North America, Europe and China (and also India and South Africa for precipitation) when five or more stations are required, with only central Asia being a large area filled in when this restriction is relaxed down to one station (for temperature indices). This highlights the limitations in the available data, and the effect of the DLS in increasing the apparent coverage of HadEX2. For indices where the DLS is large (e.g. TN90p, $T_{\min} > 90$th percentile), although most of the land surface is covered in HadEX2, only those areas listed above remain when the restriction on the number of stations is imposed. For indices which have a small DLS (e.g. Rx1day) the effect is less pronounced as the size of the DLS already limits valid grid boxes to those with stations.

4.4 Long-term stations (parametric)

Most stations in HadEX2 do not report for the full 1901–2010 period. Stations dropping in and out could cause inhomogeneities and changes in the coverage which may feed through into the global average. As shown in Sect. 3, selecting grid boxes which have 90 % completeness results in much smoother global average time series compared to when selecting all grid boxes.

We therefore select stations which have reported for a long period of time and see how only using these effects the coverage and time series. Stations can either be selected requiring that they report for greater than a given number of years or that they have a start date before a given year (and an end date after a different year, if desired), with the latter highlighting areas which only report during more recent times. The difference in which stations were selected changes the value of the DLS. Hence the coverage can be higher than for HadEX2, especially in the early part of the series. In the maps, trends and correlations have been calculated over the 1951–2010 period.

Stations were selected which reported for more than 40 to more than 80 years out of the total of 110, in 10-year increments. As can be seen for TXx (maximum $T_{\max}$) in Fig. 9a, this has a large effect on the number of grid boxes available, but there are few differences in the global time series (Fig. 9b). Other indices behave very similarly, and those where there are larger deviations from HadEX2 occur mainly in the early part of the data set (which may be partly because the longest records are concentrated in limited regions). This stability is, in part, likely due to the grid box completeness criterion when calculating the global time series (Sect. 3), though selecting only stations with very long records will not result in the same grid boxes contributing to global average time series as selecting grid boxes directly with long records. The choices with shorter record lengths (40 and 50 years) have higher correlations between the time series, lower e_{RMS} and more similar variances to HadEX2 than those choices with longer record lengths (70 and 80 years). This is because HadEX2 does not place any restriction on the length of record a station must have for it to be included. GHCNDEX is restricted to stations which have at least 40 years of data after 1951 to minimise inhomogeneities arising from a variable station network, but HadEX2 does include data with some short time series in parts. The long-term behaviour of most of the indices is unaffected by the selection of these long-term stations. For some indices (FD, frost days; GSL, growing season length; SU, summer days; TR, tropical nights) there are large differences over the first decade (1900–1910). However the value of the global average during this period is also very different from the following decades.

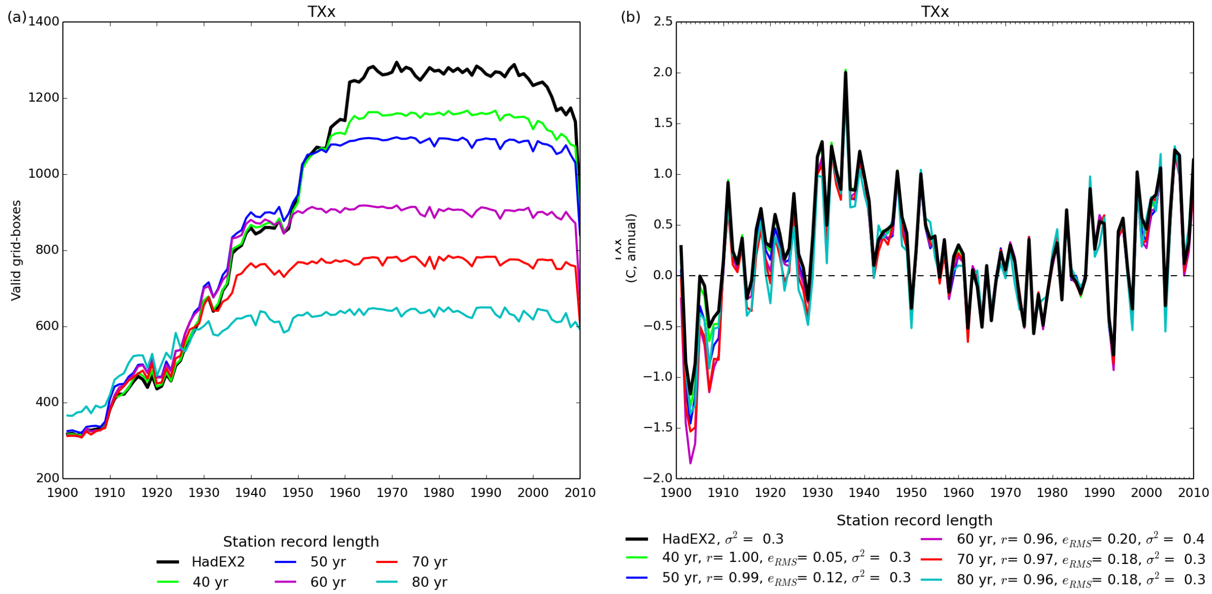

Figure 9. **(a)** The number of non-missing grid boxes and **(b)** the time series for global average TXx (maximum $T_{\max}$) for the different station record lengths. Note that the global average only takes grid boxes which have 90 % completeness or more, whereas all grid boxes are shown in the coverage series. All choices have been centred relative to the climatology period 1961–1990, reducing the spread during this time.

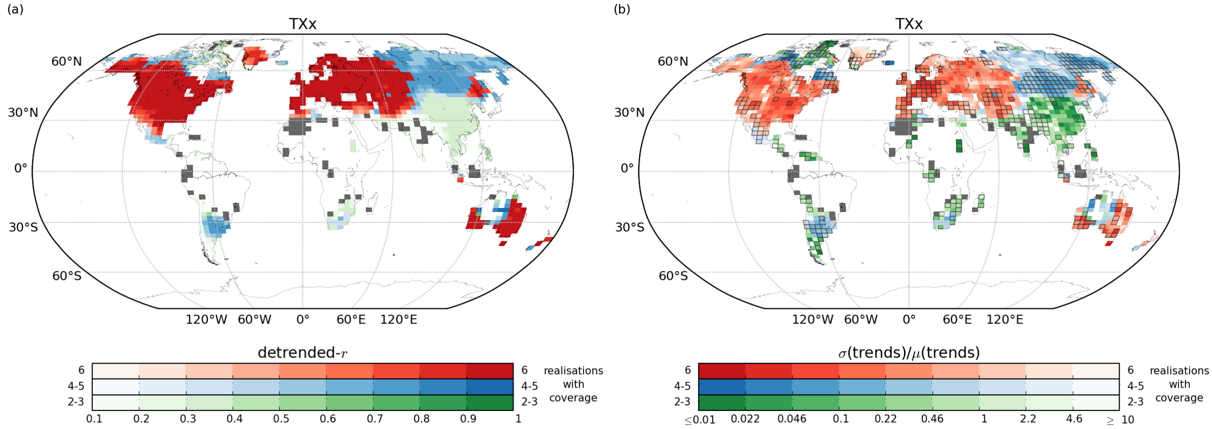

Figure 10. **(a)** The mean correlation coefficient of the detrended time series, r with HadEX2 for all the grid boxes and **(b)** the standard deviation of the trends divided by the mean trend (σ/μ) calculated for the period 1951–2010 for TXx and station record length. Grey grid boxes are those where only one of the six possible station reporting length choices results in a value, green where two to three, blue when four to five, and red where six. In the right-hand plot, boxes which have been outlined are ones where the trend was significant in the HadEX2 version. For further details see Fig. 5 and text.

This inhomogeneity results from there being no data from Australia and South America for the first 10–11 years of HadEX2 for these indices (which is allowed by the completeness requirement of 99 years of 110 present).

The uncertainty maps in Fig. 10 show clearly how different parts of the globe have different station record lengths for TXx. The regions where all six choices for the length of station record result in filled grid boxes are those where sufficient stations have very long record lengths, e.g. North America, Europe and parts of Australia for all indices, and these have high average correlation values. Regions with 4–5 choices resulting in filled grid boxes, e.g. eastern Russia and parts of South America for the temperature indices, have shorter records. China, along with parts of Africa and South America, stands out as having shorter records in the HadEX2 data set and therefore only has grid box values when selecting stations with > 40 years of data. For most precipitation indices the coverage is much more restricted, with large parts of the globe having no data or few realisations with coverage. However, as can be seen in the time series in the Supplement,

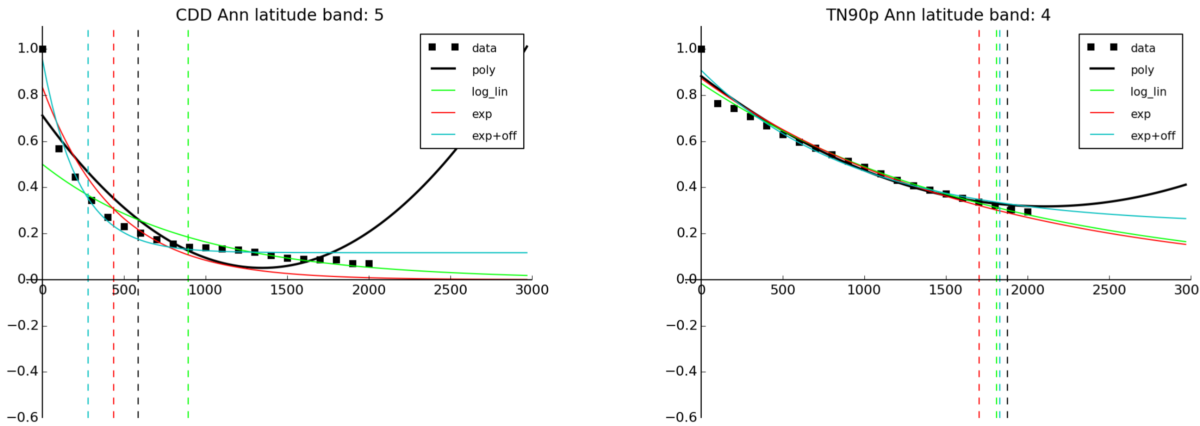

Figure 11. The binned inter-station correlation coefficients against distance for left CDD (consecutive dry days) and right TN90p ($T_{\mathrm{min}} >$ 90th percentile) along with the curves from all four fitting methods. The vertical dashed lines are the derived decorrelation length scales.

the effect of this reduction in coverage on the global average time series is small as the long-term behaviour is reproduced.

There is some correspondence between those grid boxes which have confidence in non-zero trends and those with low σ/μ values. However, this low σ/μ does not necessarily correspond to the number of choices which fill that particular grid box, i.e. highlighted boxes and/or dark shading in Fig. 10b are/is not always red. A very similar pattern is observed in the figures using the differences between the early and late periods (see Supplement).

When taking stations which report over a specific set of years we use the following five time periods: 1950–2000, 1940–2000, 1930–2000, 1920–2000 and 1910–2000. Again, the global time series look very similar (see Supplement) for all indices, with the long-term behaviour retained for all five periods. The uncertainty maps also appear reasonably similar, but with some important differences. Firstly, China, India and parts of central Africa and South America appear in grey, indicating that only one of the choices (presumably the least restrictive one, HadEX2) results in a valid grid box in these locations. Also, there is an area in eastern Russia where all six choices result in valid grid box values but in the station record length was only filled in four or five cases. This indicates that there are some large gaps in the station records in this region.

4.5 DLS fitting methods (parametric)

The method by which the DLS is found in HadEX2 is described in Sect. 2. In some cases the decline in the correlation values is smooth, but in others (and this can be seen in CDD, Fig. 11, but more clearly for other indices in the Supplement) there are bumps and wiggles in the decay curve at high separations. As the DLS is just an input to the gridding scheme, we have classed this as a parametric uncertainty. A polynomial expansion was used to fit the decay curve and obtain the DLS. As using a polynomial is an approximation to an exponential in this case[2], a number of different curves were fitted to the binned correlation values to obtain the DLS. By taking the logarithm of the correlations (y values) a linear function was fitted, which was then converted back to the exponential form (labelled “log_lin” in Fig. 11). This has the advantage of fitting a straight line, but it places more weight on the correlation values at larger distances which are not necessarily of interest. A true exponential was fitted, along with one which allowed for a non-zero offset (i.e. $ae^{bx}+c$). Although more complex curves could have been used in addition to these four, using an explicit exponential function already results in an improvement of the fit over the polynomial approximation used in HadEX2. More complex functions would be able to fit the bumps and wiggles seen at high separations, but in most cases these occur at separations larger than the DLS and so are likely to have little impact. Therefore for simplicity in this analysis we just use these four[3].

All four of these fitting methods are shown in Fig. 11 for the CDD (consecutive dry days) and TN90p ($T_{\mathrm{min}} >$ 90th percentile) indices for the latitude bands running from 60–90 and 30–60° N, respectively. The DLS is the location where the curve has dropped to $1/e$ of its value at zero separation. As theoretically it is expected that there is perfect correlation at zero distance, we place a point at (0,1). But, as in HadEX2, we do not force the curve to pass through this point, as perfect correlation is not generally assumed to occur because of, for example, instrumental and microclimate effects, especially for the indices measuring actual values. Studies have shown that even for instruments on the same site or even in the same

[2] The polynomial is a Taylor expansion of the exponential function to second order.

[3] The fitting of all methods bar the polynomial one failed for the latitude band 30–0° S for FD (frost days) and ID (ice days). This is because no stations measured any ice or frost days in this latitude band.

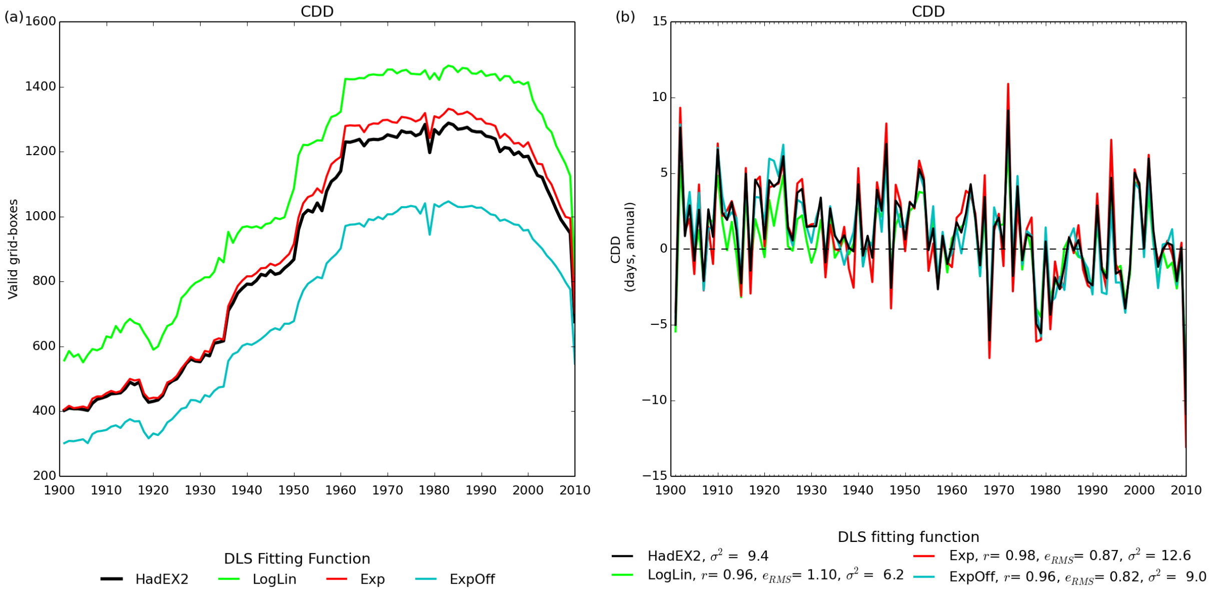

Figure 12. **(a)** The number of non-missing grid boxes and **(b)** the time series for global average CDD for the different DLS calculation methods. Note that the global average only takes grid boxes which have 90 % completeness or more, whereas all grid boxes are shown in the coverage series. All choices have been centred relative to the climatology period 1961–1990, reducing the spread during this time.

screen differences remain in the measured values (e.g. Clark et al., 2014; Sun et al., 2005), and these differences will fold through into the indices.

Imperfect correlation at zero distance is a manifestation of the "nugget effect" (e.g. Cressie, 1993; Journel and Huijbregts, 1978). This nugget is clearly visible in Fig. 11 for TN90p, where there is a difference in slope in the data for separations less than and greater than 100 km. Although other studies which use the DLS do force the fitted curve to pass through (0,1) (e.g. Jones et al., 1997; Hofstra and New, 2009), other studies clearly show nuggets in, for example, global temperature data (Rohde et al., 2013), the variance of air–sea fluxes (Lindau, 2003) and other meteorological variables reported by voluntary observing ships (Kent et al., 1999).

The two indices shown in Fig. 11 show some of the range of possible fits from the different model curves. For the CDD, the polynomial fit is clearly the worst fitting curve, and in this case overestimates the DLS compared to the best fitting curves (exponential). The polynomial fit is very poor at large distances, but the important part is the section during which the curve falls by $1/e$. The log_lin curve is also a poor fit, except at the larger distances, and hence results in a larger DLS value than the other methods. However, for the TN90p, there is very little difference between the four different methods. Again the DLS is largest from the polynomial fit, but the difference between the DLS values is smaller in TN90p than in CDD. For most indices and latitude bands, the exponential methods (with or without offset) result in the closest fit to the data. The "log_lin" method is worst at capturing the curvature at small separations, especially for precipitation indices where the correlations drop very rapidly over the first few bins.

There can be quite a range in the DLS obtained from the different methods, as shown in Fig. 11 for CDD with a range of 500 to 1000 km. When the values of the DLS are small, this can make a large difference to which stations contribute to a grid box value. For indices which have a large DLS (e.g. TN90p), the change in the DLS value will have a more limited impact on the grid box value.

Overall, most of the temperature indices show good agreement between all four methods (including the polynomial fit) and result in small differences in the DLSs obtained. The precipitation indices show larger differences in the accuracy of the fits across the four methods, resulting in larger differences in the calculated DLSs (see Supplement). The DLS from the polynomial in these cases tends to be larger than that from the exponential fits.

In the global time series, there are differences between HadEX2 and the alternative fitting methods for CDD (Fig. 12), but no perceptible difference on a global scale for TN90p (Fig. 13). For CDD, the long-term behaviour is unchanged, and the small-scale peaks and troughs in the global time series usually occur at the same times, but the values of the global averages are different between the DLS calculation methods. Consequently, the correlations between HadEX2 and the exponential model versions are still very high (0.96 and 0.98). As was mentioned in the discussion about the number of stations within a grid box (Sect. 4.3), the stations present in HadEX2 remain the same, and as the

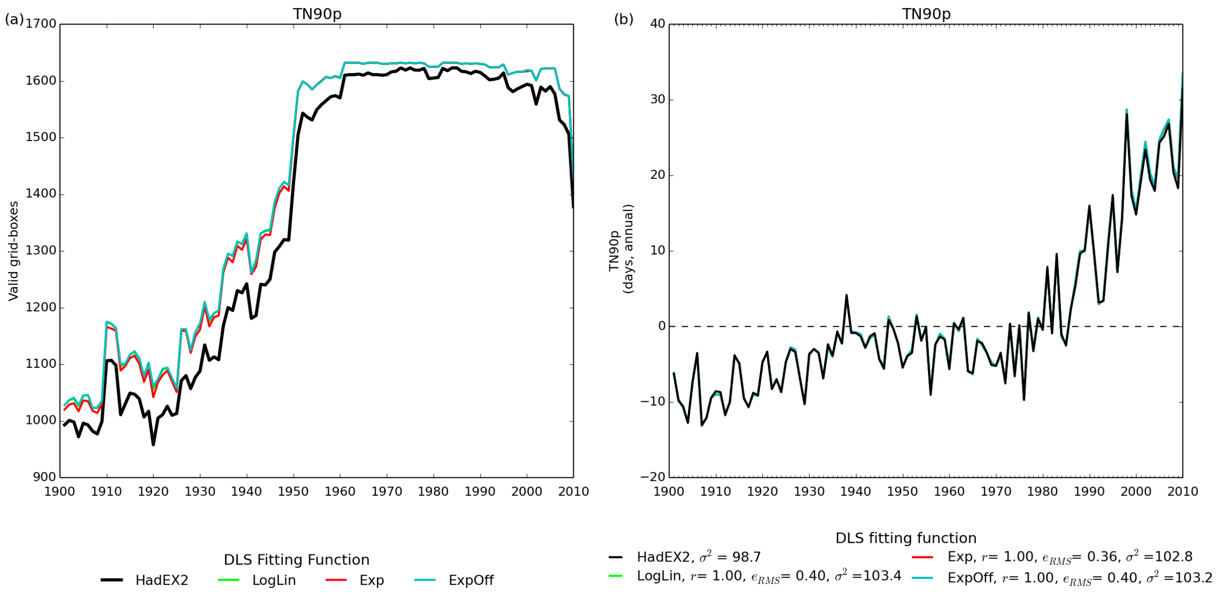

Figure 13. **(a)** The number of non-missing grid boxes and **(b)** the time series for global average TN90p for the different DLS calculation methods. Note that the global average only takes grid boxes which have 90 % completeness or more, whereas all grid boxes are shown in the coverage series. All choices have been centred relative to the climatology period 1961–1990, reducing the spread during this time.

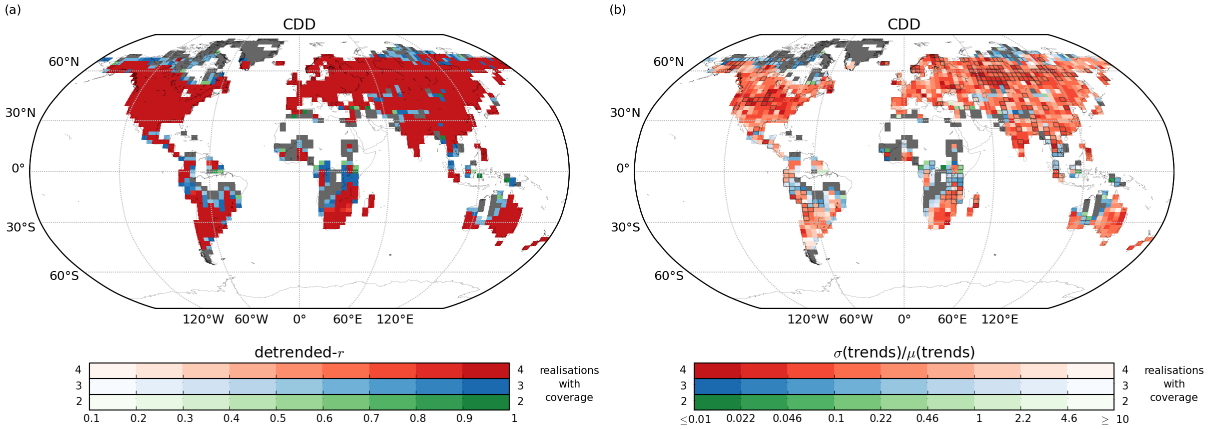

Figure 14. **(a)** The mean correlation coefficient of the detrended time series, r with HadEX2 for all the grid boxes and **(b)** the standard deviation of the trends divided by the mean trend (σ/μ) calculated for the period 1951 to 2010 for CDD and the number of stations per grid box. Grey grid boxes are those where only one of the four possible DLS calculation choices results in a value, green when two, blue when three, and red when all four. For further details see Fig. 5.

DLS increases, the extra coverage is extrapolated from the same data. It would therefore be surprising if there were large changes in the behaviour of the global average time series. A similar result is found for all other indices, with smaller differences between the time series for the temperature indices than for the precipitation ones. Differences tend to be larger in the earlier part of the record when the coverage is lower and hence the size of the DLS used has a greater effect. In many cases across all indices, the correlations between global averages are above 0.9, but the log_lin model is usually the lowest. The index measuring tropical nights (TR) has a large change pre-1950 which has a notable impact on the long-term behaviour. Given the change in coverage occurring at the same time, different DLS values combined with this is the likely cause of the difference from HadEX2.

Geographical differences, as shown in Fig. 14, will depend on whether the DLSs determined using the four different methods are similar or very different. We have chosen to show an index (CDD) where the DLS changes by a large amount, especially in the high-latitude regions (Fig. 11). Focusing on the Canadian Arctic in Fig. 14, a gradation can be seen from the areas where all four methods result in filled grid boxes to those where only one method does (Greenland). Areas where all four methods result in a value are likely to have a high station density, and the further from these regions the grid box is, the fewer of the methods result in a value for that box. Also, although the σ/μ may be small, the mean correlations are also reduced outside of the areas with high station density. CDD does not have particularly strong trends (either positive or negative) in these mid-high-latitude regions (see Fig. 8 in Donat et al., 2013a). Therefore as the DLS decreases, local variations become more prominent as dense station networks in the vicinity are no longer able to smooth them out. This results in local differences between the time series, and a small correlation coefficient in these high-latitude regions (northern Canada and Russia). There is a large variation from index to index as to which regions are filled using all four fitting methods to those with only one. Temperature indices tend to have North America, Europe and Asia filled using any of the fitting methods, as the DLSs are large regardless of method used. Only in the regions where the station density is lower do changes in the calculated DLS make a difference: sub-Saharan Africa, parts of South America and Australia. For the precipitation indices, larger areas of the globe are only filled by one of the realisations. The DLSs are on the whole smaller for these indices, so even small changes can make a large difference to the spatial coverage. A very similar pattern is observed in the maps showing the range in differences between the early and late periods (see Supplement).

Grid boxes in which there is some confidence that a trend is non-zero are predominantly found in areas where all four of the DLS calculations result in a grid box value. However, the converse is not true, especially for precipitation indices. For indices which have large DLS values, there are few areas where one or more methods exclude the grid box (see, for example, TN10p, $T_{\min} < 10$th percentile, in the Supplement). The changes in coverage are affected more by the correlation decay for a given index than by changes in the fitting method used.

4.6 Gridding methods (structural)

The gridding method used in HadEX2 is an adapted version of the angular distance weighting scheme (Shepard, 1968). This accounts for the angular distribution of the stations as well as their distance from the grid box centre. It also interpolates into empty grid boxes which are close to stations.

Three alternative gridding methods are outlined below: the climate anomaly method (CAM; Jones, 1994), the reference station method (RSM; Hansen and Lebedeff, 1987) and the first-difference method (FDM; Peterson et al., 1998).

4.6.1 Climate anomaly method

The CAM has a long history of being used in other gridded data sets, for example the HadCRUT series of surface temperature climate anomalies[4] (see e.g. Jones, 1994; Morice et al., 2012). Climate anomalies are calculated from a common reference period and then these anomalies are combined. Usually a 30-year reference period is used, with either some requirements on the number of years present within that period or using, for example, neighbouring stations to estimate the effect of the missing data on the full period value.

In this analysis, the climatological reference period has been chosen to be 1961–1990 to match that used elsewhere throughout this paper. At least 25 out of the 30 years had to have valid data for a climatology to be calculated. A simple mean across all stations in the grid box resulted in each annual value.

By the nature of this method, for a grid box to have a value, there must be at least one station present within it (with sufficient data in the period 1961–1990). If there is only one station present, this has been assumed to be representative of the entire grid box. As this may not be a valid assumption, should this method be used, it may be prudent to require at least two or three stations per grid box, but this will result in a further decrease in the coverage (see Sect. 4.3).

4.6.2 Reference station method

The RSM described by Hansen and Lebedeff (1987) starts by selecting the station with the longest record within the area of influence of the grid box centre. Hansen and Lebedeff (1987) used a fixed distance of 1200 km, which was derived from the decay of the correlation between stations. In this study we use the DLS appropriate for the latitude of the grid box as in the ADW for HadEX2. Having selected the station with the longest record within one DLS of the grid box centre, successively shorter stations are processed. The new stations' temperature records are adjusted so that their mean over the common period is the same as the composite of all stations that have been processed so far. Then, distance-weighted averages are re-calculated to obtain the new composite station. This process is repeated until all stations within one DLS of the grid box centre have been included into the composite station.

The advantages of this method are that a common reference period is not required. However, stations are still required to have at least 20 years of overlap with the composite station, and so stations with very short records are still not included. However, stations which report in early times, but not during a specified reference period, can be accommodated by this method. Peterson et al. (1998) point out that this method relies on the reference station (the one with the longest record) for providing an accurate representation of climatic changes, uncorrupted by biases arising from station moves, instrument changes or other changes in observing procedures.

[4] Climate anomalies are the differences from a mean or median without division by the standard deviation.

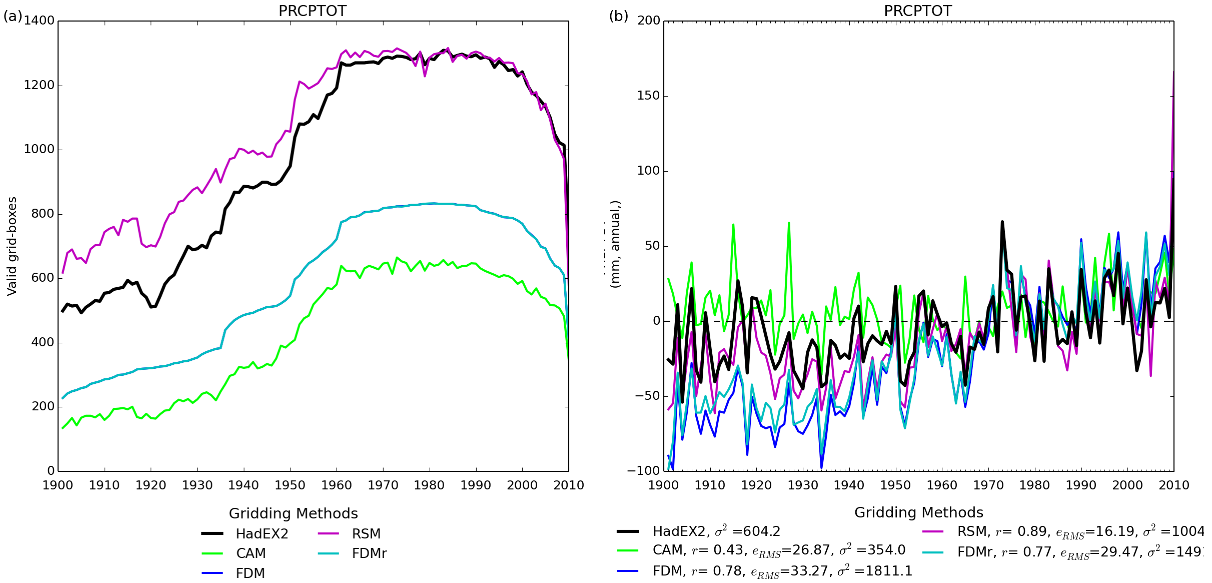

Figure 15. (a) The number of non-missing grid boxes and **(b)** the time series for global average PRCPTOT (total annual precipitation) for the five different gridding methods. Note that the global average only takes grid boxes which have 90 % completeness or more, whereas all grid boxes are shown in the coverage series. All choices have been centred relative to the climatology period 1961–1990, reducing the spread during this time.

If there are non-climatic inhomogeneities in the reference station, these will feed through into the final grid box value. In the creation of HadEX2, quality control and homogeneity checks varied from country to country, but in most cases data have been checked by researchers from the country of origin (e.g. regional workshops) or passed through automated quality control procedures (e.g. GHCN-Daily (Menne et al., 2012), ECAD (Klok and Klein Tank, 2009)) (Donat et al., 2013a). No extra homogeneity assessment using, for example, the techniques of Menne and Williams Jr. (2009), Mestre et al. (2011), Domonkos (2011) and Toreti et al. (2012) has been performed on the data. No consistent quality control has been applied to the raw daily data as this was not possible given the way the data have been collected (Alexander et al., 2006; Donat et al., 2013a), and so any inhomogeneities present could feed through into the final grid box values.

4.6.3 First-difference method (FDM)

The FDM was proposed by Peterson et al. (1998) as an alternative to the previous two methods, as it did not require either a common reference period or a reference station, and therefore could use all available data. This method does, however, suffer if there are frequent gaps in the data, and is sensitive to outliers if they occur at the beginning or end of the time series.

In this method, the annual values are converted to a series of first differences, with the value for the first year of the series being zero. We average the first differences from different stations within the grid box to obtain the grid box value. The final step is to reconstruct the centred time series by performing a cumulative sum.

Using a simple first difference is probably most suited to temperature indices, whereas something more complex, e.g. a ratio, may be more appropriate for precipitation. For simplicity, we used the simple first difference for all indices, and did not find any blatantly spurious results. Years with missing data within the time series were filled with the average first difference for that station, except years before the start and after the end of the station's reporting period. In this method the errors accumulate as the cumulative sum is calculated. The errors are likely to be smaller in the most recent period, but larger in the past, when the station networks were sparser. Working forwards in time carries these larger errors into the most recent period. Therefore we also run a version of the first differencing in reverse (FDMr).

4.6.4 Gridding methods results

The results from the five gridding methods for PRCPTOT (total annual precipitation) can be seen in Fig. 15. The coverage of the RSM, which uses the DLS to find stations within a region to merge together, is similar to that of the ADW method, and is actually larger in early times. For the CAM and the FDM/FDMr methods, the coverage is smaller, as these methods require that stations be present within a grid box. The CAM is more restrictive on which stations it includes when calculating grid box average values, and thus has the lower coverage of the two.

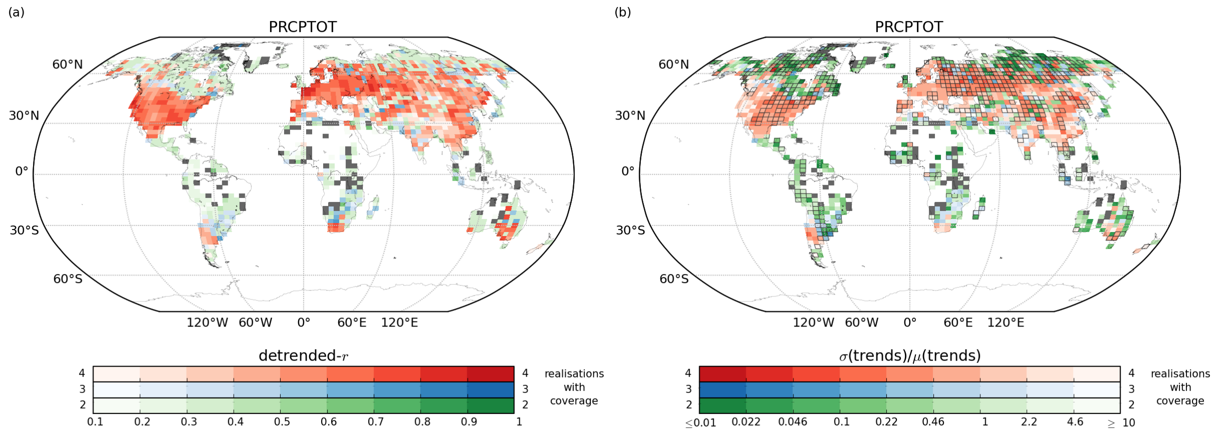

Figure 16. **(a)** The mean correlation coefficient of the detrended time series, r with HadEX2 for all the grid boxes and **(b)** the standard deviation of the trends divided by the mean trend (σ/μ) calculated for the period 1951–2010 for PRCPTOT and the four gridding methods. Grey grid boxes are those where only one of the five possible gridding methods results in a value, green where two, blue where three, and red where all four of the possible choices. For further details see Fig. 5.

Reversing the order of the cumulative sum in the FDM (FDMr) has resulted in no changes in the coverage for any of the indices. In the global time series there are only small differences in the year-to-year values between the FDM and the FDMr versions, with the long-term behaviour being virtually identical. The two versions of the FDM are shown in the time series plots (Fig. 15) but only the standard version (FDM) in the maps (Fig. 16).

The correspondence between the different gridding methods on the time series is relatively good for the post-1950 period. Prior to this time, however, there are large differences in the global average, which lead to different long-term behaviours over the entire period of the data. Although the sizes of the short-timescale variations do not always match, they occur at the same time and in the same direction (i.e. local year-to-year differences have the same sign, if not the same magnitude). Sometimes the magnitudes of these short-timescale variations are larger than in HadEX2. The RSM has the highest correlation coefficients with the HadEX2 time series, owing to the extrapolation, smoothing and coverage that this method has in common with the ADW scheme of HadEX2.

Compared to all the previous uncertainties, the gridding methods have a much larger effect on both the short- and long-term global averages. Most of the indices, not just the precipitation ones, show large differences to HadEX2, especially, but not only, at early times. Indices in which there are comparatively small changes resulting from the gridding methods are CSDI and WSDI (cold and warm spell duration indicator), CDD (consecutive dry days), GSL (growing season length), SU (summer days) and TX90p ($T_{\max} > 90$th percentile). Both GSL and SU have a discontinuity pre-1910 (discussed in Sect. 4.4), which is present in the ADW and RSM interpolating methods, but not in the other two (CAM, FDM). The absence of the discontinuity results in a large decrease in the variance.

Other indices have changes in the short-term variability, but the long-term behaviour is roughly the same as HadEX2 (DTR, ETR: diurnal and extreme temperature range; FD: frost days; R10mm, R20mm: days when $P > 10$ mm or $P > 20$ mm; R95pTOT, R99pTOT: R95p/PRCPTOT, R99p/PRCPTOT; SDII: simple daily intensity index; TNn: minimum $T_{\min}$; TXn: minimum $T_{\max}$). For TN90p ($T_{\min} >$ 90th percentile) and TX10p ($T_{\max} <$ 10th percentile), the CAM and FDM change only the short-timescale variations; the RSM, however, reduces the amplitude of the long-term behaviour. In the remaining precipitation indices (PRCPTOT; R95p, R99p: sum of precipitation $>$ 95th/99th percentile; Rx1day, Rx5day: maximum 1- and 5-day precipitation total), the FDM results in stronger (increasing totals) long-term trends than in HadEX2; in PRCPTOT the CAM method results in a weakening trend (decreasing totals). For the remaining temperature indices, TNx stands out as the short-term variability in the global average is greatly increased when using the RSM. However, the other indices also have changes in their long-term trends; in TN10p ($T_{\min} <$ 10th percentile) and TXx (maximum $T_{\max}$) the RSM method reduces the long-term trend in the first half of the period. The FDM results in a much stronger positive trend than HadEX2 in the tropical nights index (TR).

There is no clear pattern to these results, either by index type or by gridding method. In some cases long-term trends or the short-term variability are enhanced and strengthened by one of the methods, whereas in other cases they are reduced. This shows how sensitive some of the HadEX2 indices can be to the gridding method used.

To illustrate the regional influences of the gridding methods, we show the corresponding uncertainty maps for

PRCPTOT in Fig. 16. Note that the results from the FDMr are not included as they are very similar to the FDM results, and these two methods are not fully independent. Only a few regions have high r values from the detrended time series, on the whole corresponding to those regions with dense station networks. Many grid boxes in high latitudes, South America, Africa, Australia and parts of Asia are only filled by two of the methods (likely to be RSM and ADW). In these regions, as the grid box values have been interpolated from surrounding stations, the differences in the two methods have resulted in differences in the short-timescale variations, and hence low r values. In regions where the station density is high, and all four methods fill the boxes, the grid box average is driven by stations within the box, and so short-term variations match between the four methods more often, resulting in higher r values. However for some indices, even regions with a high station density have low r values (most precipitation indices, TNx, TXx).

The values of σ/μ indicate that there is often a large spread in the values of the trend compared to the mean trend. Indices which have strong long-term trends (usually temperature-based ones) have smaller relative spreads than those which have weak long-term trends (primarily precipitation indices). However, some grid boxes where only two methods fill the box have a very small spread in the trends compared to regions where all four methods fill the box. These areas are likely to be only filled by the RSM and ADW methods, which both interpolate using neighbouring stations, and therefore have similar sets of stations combining to form a grid box value, resulting in similar trends. When all four methods fill a grid box, there is a greater range in methodological choices and thus a likely greater range in the trend magnitudes and a larger relative spread. A similar pattern is observed in the maps for the differences between the early and late period (see Supplement). However, overall, a larger spread in differences is observed, especially where only two methods result in a value for the grid box.

4.7 Station network (sub-sampling)

Changes in the station network have been shown to cause significant changes in global analyses, especially for precipitation-based indices (Wan et al., 2013; Trenberth et al., 2014). To assess the effect of the station network on the global and regional trends, we perform a sub-sampling experiment. We sub-sample the parent population (in this case, the set of HadEX2 stations) without replacement and re-run the entire creation process of the data set. To sample the effect of fewer stations within the network, 100 iterations each were run using random selections of 25, 50 and 75 % of the total station number for each index. These iterations recalculated the DLS and gridding for each run. Of course, with most stations being found in North America, Europe and eastern Asia, even random selections will have, on average, similar distributions and coverage to those using the full network.

As can be seen in Fig. 17a for the ETR (extreme temperature range) index, the grid box coverage is unsurprisingly lower for the runs with only 25 % of stations, compared to those with 50 or 75 % or the complete set of HadEX2 stations. The scatter in the coverage also increases as the number of stations decreases, as would be expected. As the DLS is recalculated for each sub-sampling run, there will be a spread in values obtained across the iterations. This may result in better coverage than for the full HadEX2 station list, but in many cases the coverage will be smaller. From the global average time series (Fig. 17b), it is not clear whether a single sub-sampling run is biased to high or low values. The three different sub-sampling run sets form a band around the HadEX2 series, but at individual short-timescale peaks and troughs are more extreme, particularly in the 25 % runs, and especially at early times. The width of the band also increases towards the start of the data set, which corresponds to the decrease in coverage observed before around 1950. The larger range in values at earlier periods results in greater uncertainty in the overall long-term behaviour of this index. For indices which have global, long-term linear trends over the entire period clearly different from zero, the sub-sampling runs do not change this (e.g. TN10p). But for indices with no strong global trend, or a non-linear behaviour over the full period, different slopes can arise from run to run.

Some other indices show a similar close relationship with HadEX2. The temperature percentile indices show very little variation. However the precipitation indices are more variable: see, for example, Fig. 18 for R20mm (sum of days with > 20 mm precipitation). The runs with 75 % of stations capture the year-to-year peaks and troughs, whereas those with 25 % of stations are much noisier. Also, the long-term behaviour could be very different, in the extreme ranging from a notable positive to a notable negative trend (increasing number of days to decreasing number). Changing the number of input stations when calculating precipitation-based indices can have a large impact on their globally averaged behaviour (see Wan et al. (2013) and Trenberth et al. (2014) for a study on drought-based indices). Sub-sampling the station network demonstrates the large effect the underlying station coverage (in space and time) has on the final global result, especially for the precipitation indices.

Most of the Northern Hemisphere grid boxes are filled by over two-thirds of the 25 % runs in ETR as shown in Fig. 19, and a similar pattern is found for the other temperature indices, and also for the maps using the differences between the early and late periods (see Supplement). This indicates that in the regions of high station density grid boxes are almost always filled by the HadEX2 creation process even with a much reduced station density. In the Southern Hemisphere, however, there are large areas where less than one-third of the runs fill the grid boxes. Although the grid boxes are filled by most of the runs, the range in trends is not uniform. In North America, the ETR shows a consistent small range in

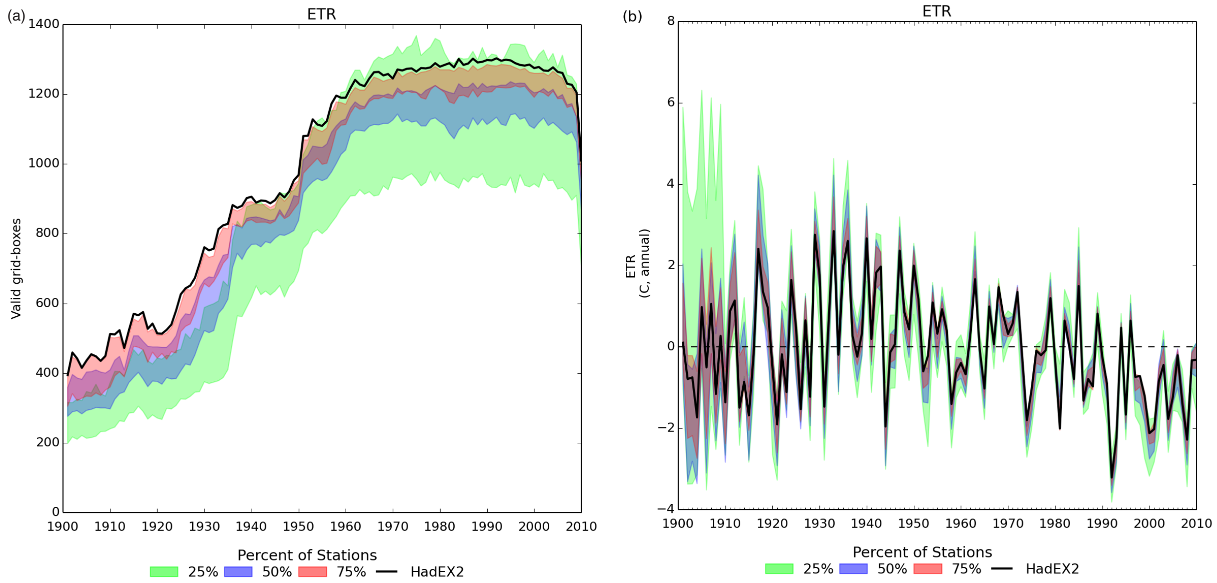

Figure 17. **(a)** The number of non-missing grid boxes and **(b)** the time series for the global average of ETR (extreme temperature range) for the sub-sampling runs. Each colour shows the maximum range of the time series for each of the three sets of runs: 25 % of stations in green, 50 % in blue and 75 % in red. These are transparent, so purple is the overlap of the 50 and 75 % ranges. The solid black line shows the HadEX2 results in both panels. Note that the global average only takes grid boxes which have 90 % completeness or more, whereas all grid boxes are shown in the coverage series. All choices have been centred relative to the climatology period 1961–1990, reducing the spread during this time.

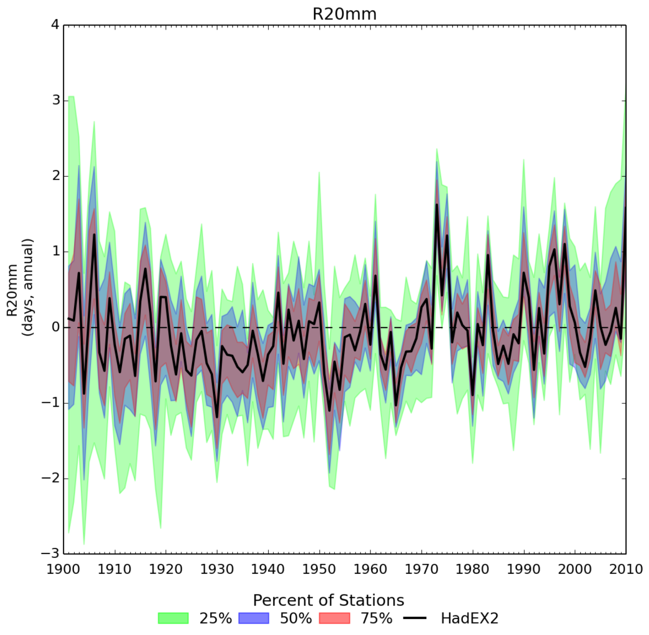

Figure 18. As for Fig. 17b but showing the time series for the global average of R20mm for the sub-sampling runs. All choices have been centred relative to the climatology period 1961–1990, reducing the spread during this time.

the trends, but parts of central Europe and central Asia exhibit a large variance in the trends.

For all the precipitation indices apart from CDD, PRCPTOT and R10mm, only the USA and Europe, with small regions elsewhere, are filled by more than two-thirds of the 25 % runs. In large parts of the globe, less than one-third of the runs result in grid boxes having a value. The colours are also generally much less intense, showing a wider range in linear trends over the latter part of the data set across the sub-sampling runs.

5 Discussion

5.1 Taylor diagrams

In order to make the different methodological choices easier to interpret we use a presentation method common in climate model evaluation analyses – the Taylor diagram (Taylor, 2001). These diagrams are a way of showing graphically how well two patterns (in this case time series) match. However, by using just the global (land-surface) average time series, we lose the regional information, which can to some extent cancel out. One diagram for each of the categories of indices outlined in Sect. 2 are shown in Fig. 20.

The x and y axes are the standard deviation of the time series, with the reference data set (HadEX2 in this case) being plotted on the x axis. The polar axis represents the correlation between the time series of the two data sets ranging from zero at 0° to one at 90° as calculated over the entire period. The advantage of this diagram is that it also shows the

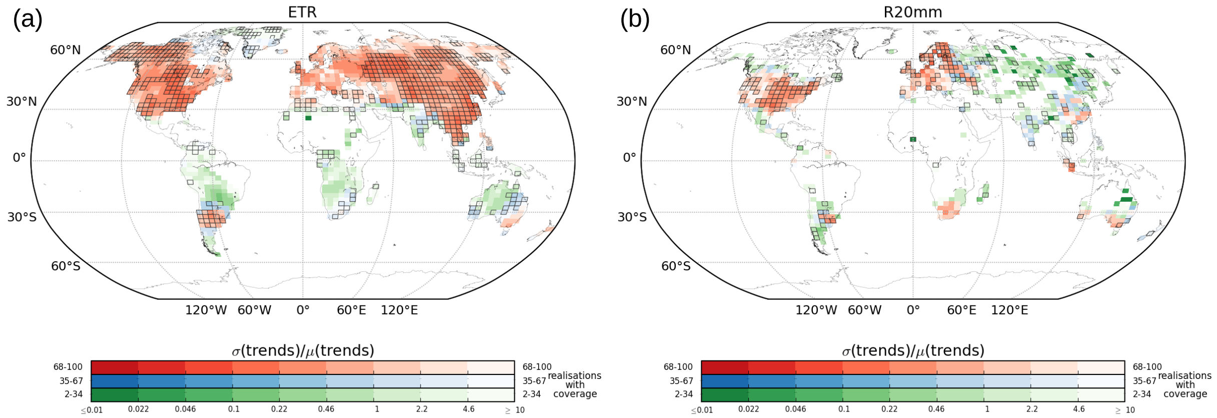

Figure 19. The standard deviation of the trends divided by the mean trend (σ/μ) calculated for the period 1951–2010 for **(a)** ETR and **(b)** R20mm using the 25 % sub-sampling runs. Green grid boxes are those where 2–34 runs result in a value, blue for 35–67 and red for 68–100. For further details see Fig. 5.

root-mean-square error (e_{RMS}), shown by the grey semicircles centred on the reference data set in Fig. 20. In this way all of the time series shown in the previous sections can be compared to the reference series calculated from HadEX2, hence providing a summary of the changes each choice has on the global time series for each index.

This diagram allows for the comparison of both the long-term and short-term variation between the parametric and structural choices and is used in the discussion below. The standard deviation of the global average time series gives some level of the internal noise and variability of the time series. If the different structural and parametric choices result in large changes to the short-term variability characteristics, then this will stand out along this axis. The correlation coefficient will show both how well long-term trends have been captured by the different versions of HadEX2 and how well short-term changes agree in time.

In general, the further a point sits from the reference series (HadEX2), the worse the agreement between the two time series, either with differences in the internal variability or the correlation. We show the effect of the completeness criterion, as discussed in Sect. 3, on the diagrams despite this not being a parametric or structural uncertainty in the data set but rather a criterion when creating the summary time series. In most cases, the largest changes are seen in the mean correlation of the grid box time series rather than the standard deviation (which can also be noted in the legends of the different time series). As noted in Sect. 3, there are large differences between HadEX2 and those global average time series calculated when the completeness restriction drops to below 60 % for many indices. These differences are especially clear at early times when there is less data coverage. In some indices these differences can change the long-term trend.

For many of the indices shown in Fig. 20 this has one of the largest effects on the agreement between the time series and HadEX2. This demonstrates the importance of the completeness requirement when calculating global averages. Regional time series may be less affected, but this will be dependent on the index as well as location and size of a specific region.

5.2 Parametric uncertainties

The smallest effects on global trends and variances were observed when investigating the parametric uncertainties. The weighting (scale) parameter of the ADW scheme had almost zero effect, and these time series sit very close to HadEX2 in Fig. 20 for all indices. Selecting only long-term stations has an impact via the station coverage, but this is muted because HadEX2 stipulates $\gtrsim$90 % temporal completeness for each grid box. For the precipitation indices this resulted in a reduction in the correlation between the time series, but the level of internal variability remains the same, and for most temperature indices it has a relatively small effect. However, for the number of frost days (FD), this uncertainty results in a large increase in the internal variability. This increase has arisen from the inhomogeneity in the time series pre-1910, as discussed in Sect. 4.4. The restriction on long-term stations has a large effect on the value of the global average during this period, increasing the overall variance.

Relaxing the criterion for at least three stations within a DLS allows values for each index to be calculated for more land-surface grid boxes, thus apparently increasing the coverage. This is unlikely to change any of the results in areas with a high station network density, but decreases the correlation with HadEX2 for the global average as more regions are included. Conversely, increasing the number of stations required within a DLS reduces the coverage to just those areas which have a high station density and also reduces the correlation with HadEX2 for the global average. Therefore the regions which have few stations have the greatest effect on the global average when changing the number of stations within a DLS. The effect of this choice is

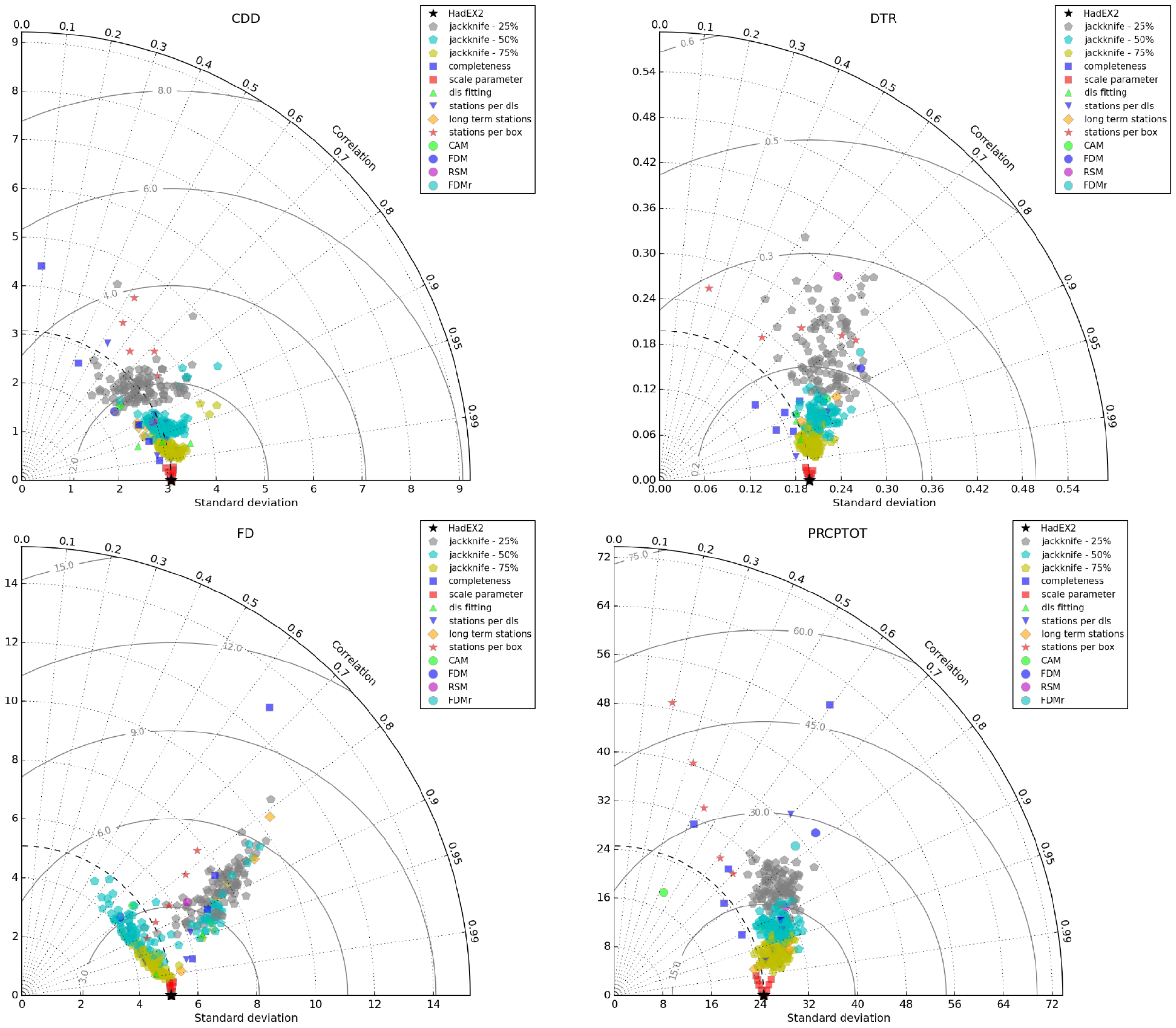

Figure 20. The Taylor Diagram using the global average timeseries for each choice. Each of the different methodological choices in the previous sections are shown using a different symbol and colour as indicated in the legend. Diagrams for CDD (Consecutive Dry Days), DTR (Diurnal Temperature Range), FD (Frost Days) and PRCPTOT (Total Annual Precipitation) are shown.

larger in the precipitation indices than in the temperature ones (see Fig. 20), with the correlation decreasing but only small changes in σ. Also, many of the precipitation indices show only small or no long-term signal, and so the correlation coefficient of the time series with HadEX2 is dominated by the short-timescale variations. As the regions that can contribute to the global average change, these short-timescale variations change location, but their magnitude remains roughly the same. Hence, the correlation decreases but the standard deviation remains approximately constant in the Taylor diagrams.

When the minimum number of stations per grid box increases, the coverage decreases but the global average trends do not change by much over the recent period. In the early part of the data set, there can be large differences between HadEX2 and the different choices. This manifests itself in reductions in the correlations as well as increases in the standard deviation for most indices as the short-timescale variations change. As the coverage is reduced by restricting the regions included to those with higher station densities, these dominate to a greater extent, resulting in the larger short-timescale variations and lower correlations.

Although the DLS fitting method used in HadEX2 is relatively simple, it proves to be effective in capturing the geographical coherence of the data for most of the temperature indices. For the precipitation indices, which have short DLS values, this functional form does not do well at large separations. For most indices, the polynomial fit overestimates the DLS, which has a greater effect for the precipitation indices with smaller DLS values than for the temperature ones. For some indices one or more of these differing fitting methods have a large effect (e.g. CWD: consecutive wet days; TR:

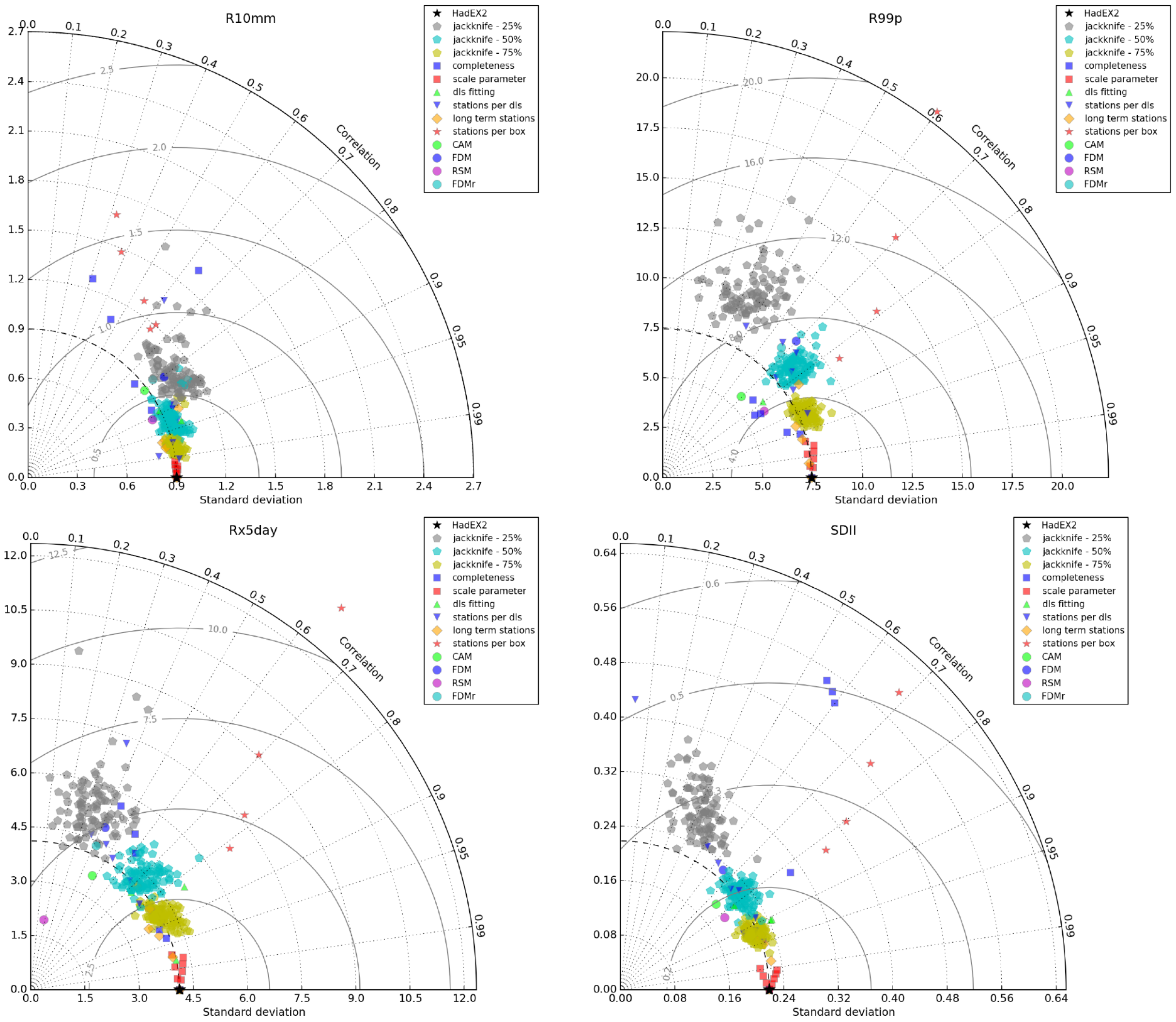

Figure 20. Continued. Diagrams for R10mm (Number of days with precipitation > 10 mm), R99p (annual sum of daily precipitation $> 99^{th}$ percentile), Rx5day (maximum 5 day precipitation total) and SDII (Simple Daily Intensity Index) are shown.

tropical nights). But in most cases the changes in the DLS fitting method do not stand out as drastically changing the global average when compared to other sources of uncertainty.

5.3 Structural uncertainties

The structural uncertainties often had a larger impact than most of the parametric changes. By changing the gridding scheme, the weighting given to different stations changes. This has one of the largest effects on the global trends, and more at a regional level. Two classes of scheme were tested – those which interpolated (RSM and ADW) and those which did not (FDM and CAM). Hence, grid boxes mainly had values from only the two interpolation methods (ADW and RSM) or from all methods. The spread of linear trends is likely to have been undersampled for grid boxes where only two methods resulted in values. Hence the trends in these regions have a small spread, but have a larger spread in the areas where all five methods fill a grid box. Also, the similarity between the two pairs of methods generally resulted in higher correlations and lower reduced variances when only two methods filled a grid box than when all five methods did.

Regions which had a high station density were less susceptible to the changes brought about by the different gridding methods. In these regions, the grid box value depends more on stations within the grid box than those outside. Therefore the effect of interpolation and weighting is smaller.

The indices which were identified in Sect. 4.6.4 as having very little difference between the five methods can also be identified from the Taylor diagrams (CDD, CSDI, GSL, SU, TX90p and WSDI). GSL and SU have a large decrease in the variance, which stems from the non-interpolating methods not being affected by the discontinuity pre-1910.

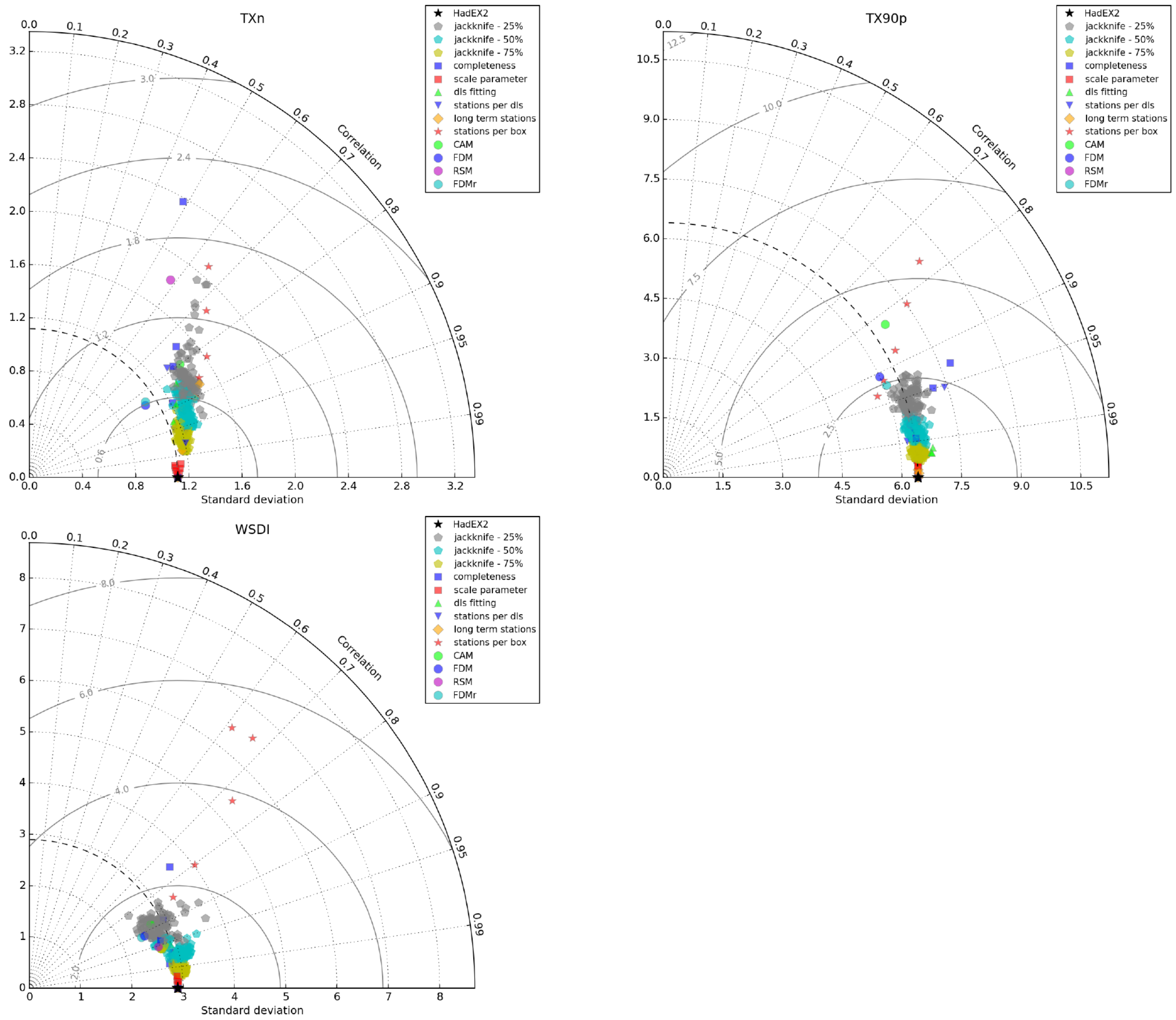

Figure 20. Continued. Diagrams for TXn (Minimum Tmax), TX90p (Tmax $<$ 90th percentile) and WSDI (Warm Spell Duration Indicator) are shown.

Those indices which predominantly have changes in the short-term variability (DTR, ETR, FD, R10mm, R20mm, R95pTOT, R99pTOT, SDII, TNn and TXn) mainly show changes in the correlation. The effect of the large increase in amplitude of short-timescale variations when using the RSM method for DTR is also very clear. TXn shows the previously mentioned discontinuity pre-1910 for RSM, which results in large changes in the correlation and the variance. The change in the long-term behaviour of TN90p and TX10p in by the RSM method can also be seen by the comparatively large change in correlation. The remaining indices tend to show differences in the long-term behaviour (PRCPTOT, R95p, R99p, Rx1day, Rx5day) leading to changes both in correlation and the internal variability.

For many indices, changing the input station network had the largest effect on the global time series. When using only 25 % of the HadEX2 stations, the coverage could be larger than when using the full station list because of changes in the DLS. However the small set of stations resulted in large deviations, especially at early times when the stations contributing to the global average are further reduced. On a regional level, however, areas of high station density still have consistent and robust trends for most of the temperature and some of the precipitation indices (see Fig. 19). As seen in the Taylor diagrams, it is mainly the correlation that is affected, which can reduce by a large amount for some of the 25 % runs. There is also a lesser tendency for the standard deviation to increase, which matches the increase in internal variability observed for other sources of uncertainty where the coverage is reduced. This analysis shows how the underlying input station network can have a large impact on the final global behaviour, and of course regionally as well (Wan et al., 2013; Trenberth et al., 2014).

Some of the indices show changes in the variability which appear to follow two regimes (FD, GSL, SU, TR), and these are all indices where the discontinuity pre-1910 has been observed. As mentioned above, this arises from the lack of Australian and South American data during the first few years of the data set. In the sub-sampling experiment, the few stations which contribute to these regions will not always be selected, hence resulting in different behaviour on a global average compared to when these stations are included.

5.4 Combined uncertainties

The extremes indices assessed in this study are regularly used for the monitoring of changes in the occurrence of extremes across the globe. Therefore the overall uncertainties are vital to ensure that the reliability of the trends can be assessed. Additionally, providing quantified uncertainties to data sets is an important and required step in making these fit for purpose in the current era.

To summarise the results from each of the parametric and structural uncertainties over all the indices we show the linear trend from HadEX2 calculated over 1951–2010 using the median of pairwise slopes estimator in the Table 2. We also show the statistical range in slopes as the 5th to 95th percentile. For each of the methodological choices we show the range in linear trend calculated for each choice, as well as how many of the trends fall within the statistical range of HadEX2. This only gives a global overview and also only for the latter part of the data set, when the coverage effects are smaller.

For most of the percentile and block maxima temperature indices, almost all the choices fall within the statistical range of trends of HadEX2, and similarly for some of the duration-based ones (CSDI, WSDI: cold and warm spell duration indicators). This indicates that the statistical uncertainty in the linear trends is larger than the structural and parametric uncertainties from the different methodological choices. This can also be seen in the Taylor diagrams for these indices, where the points from all the different sources of uncertainty investigated cluster relatively tightly around HadEX2.

GSL (growing season length) and ID (ice days) have the worst agreement, but this is only apparent in a few of the choices. These, along with the other threshold-based temperature indices, have a larger spread in the Taylor diagrams, showing the decrease in correlation over both short and long timescales as well as changes in the short-term variability. The precipitation indices are on the whole only slightly less robust than the temperature indices, as more of the range of trends from HadEX2 fall outside of the envelope set by the different methodological choices. This may be in part due to their relatively large trend uncertainties for the precipitation indices in HadEX2 on a global scale. The heavy precipitation totals and duration indices appear to be the least robust, which matches the difference in spreads observed in the Taylor diagrams. However, in all indices, for most choices more than half the global trends fall within the statistical range of HadEX2. As many of the precipitation indices have no strong long-term linear trend on a global scale, this measure is less useful than for the temperature indices.

The comparison between the gridding scheme results (change in the values of individual grid boxes) and the minimum number of stations within a grid box (change in the coverage) demonstrates the two main sources of methodological uncertainty within HadEX2 and its related data sets. Most of the stations in HadEX2 and its related data sets are found in North America, Europe, Asia and Australia. Methodological choices which affect coverage generally have small effects on the global averages because these averages are dominated by those well-observed areas. Also, there are only small effects on grid box values in those areas, as the correlations are high and the spread in trends low (Fig. 8). Changing the gridding method not only changes the coverage but also affects how the stations are blended together, changing local grid box values. This results in lower correlations and a greater spread in linear trends for individual grid boxes (Fig. 16). Other larger sensitivities are observed when changing the network of input stations, through the sub-sampling experiments, or when calculating the global average itself using different data completeness criteria for the individual grid boxes (Wan et al., 2013).

In this analysis we have assessed a number of methodological choices made during the creation of the HadEX2 data set to assess the sensitivity of global and regional trends to these parameters. There are now a range of data sets available which follow the HadEX2 method, as mentioned in Sect. 1, and all of these are very likely to have similar sensitivities to the choices assessed here. This family of data sets does, however, probe the effect of completely independent station networks, in a way that the sub-sampling experiments run in this analysis (Sect. 4.7) cannot do (see Donat et al., 2013b). As expected, many of the parametric uncertainties have the smallest effect on the results, whereas the structural ones have the largest effect. Methodological choices which drastically change the grid box value or the coverage are those which have the greatest effects: the gridding method, station network (sub-sampling) and stations within a grid box. Comparing the global time series of all methodological choices together, HadEX2 seems reasonable and generally lies towards the centre of the range of variation exhibited between the different choices. It also optimises the spatial coverage by interpolating and hence providing information for data-sparse regions based on the correlation structure of the data.

In the course of this study we have not been able to assess all sources of uncertainty. The quality of the station data has not been investigated. Although quality control procedures have been applied during the creation of parent data sets (e.g. GHCN-Daily, ECAD), by national meteorological services and also at regional workshops, these are unlikely to be consistent for all stations over the entire span of the data set. When taking global averages, we have not investigated the

Table 2. A summary of all the uncertainties assessed for each index in this work. For HadEX2, the linear trends and their uncertainties (5th to 95th percentile range) have been calculated over 1951–2010 using the median of pairwise slopes estimator (Sen, 1968; Lanzante, 1996) and are per decade. For the six different choices investigated in this study, the range obtained is shown as well as the number of choices which fall within the statistical range of the HadEX2 slopes.

Index	Range of linear trend of global average time series						
	HadEX2	Weighting	Stations/DLS	Stations/grid box	Long stations	DLS methods	Gridding
				Temperature			
TXx	0.11 (0.04 → 0.17)	0.11 → 0.11 (8/8)	0.10 → 0.22 (9/10)	0.03 → 0.11 (5/6)	0.10 → 0.11 (6/6)	0.08 → 0.11 (4/4)	0.07 → 0.12 (5/5)
TXn	0.33 (0.21 → 0.45)	0.33 → 0.33 (8/8)	0.29 → 0.39 (10/10)	0.13 → 0.33 (3/6)	0.29 → 0.38 (6/6)	0.27 → 0.33 (4/4)	0.24 → 0.36 (5/5)
TNx	0.17 (0.11 → 0.23)	0.17 → 0.17 (8/8)	0.17 → 0.27 (9/10)	0.11 → 0.17 (6/6)	0.16 → 0.18 (6/6)	0.12 → 0.18 (4/4)	0.16 → 0.23 (4/5)
TNn	0.43 (0.32 → 0.53)	0.42 → 0.43 (8/8)	0.33 → 0.48 (10/10)	0.37 → 0.43 (6/6)	0.40 → 0.47 (6/6)	0.29 → 0.43 (3/4)	0.38 → 0.51 (5/5)
DTR	−0.07 (−0.09 → −0.06)	−0.07 → −0.07 (8/8)	−0.07 → −0.06 (9/10)	−0.09 → −0.07 (6/6)	−0.09 → −0.05 (5/6)	−0.07 → −0.06 (4/4)	−0.13 → −0.07 (1/5)
ETR	−0.38 (−0.49 → −0.28)	−0.38 → −0.37 (8/8)	−0.41 → −0.06 (9/10)	−0.38 → −0.30 (6/6)	−0.38 → −0.31 (6/6)	−0.42 → −0.31 (4/4)	−0.44 → −0.23 (4/5)
GSL	0.83 (0.34 → 1.29)	0.79 → 0.84 (8/8)	0.83 → 1.81 (6/10)	0.83 → 1.31 (5/6)	0.83 → 1.47 (3/6)	0.78 → 1.08 (4/4)	0.83 → 1.88 (1/5)
CSDI	−0.66 (−0.84 → −0.47)	−0.67 → −0.64 (8/8)	−0.70 → −0.50 (10/10)	−0.66 → −0.52 (6/6)	−0.71 → −0.51 (6/6)	−0.69 → −0.54 (4/4)	−0.81 → −0.66 (5/5)
WSDI	1.32 (0.85 → 1.69)	1.27 → 1.32 (8/8)	1.16 → 1.62 (10/10)	0.89 → 1.32 (6/6)	0.89 → 1.32 (6/6)	1.16 → 1.72 (3/4)	1.02 → 1.32 (5/5)
TX10p	−2.49 (−3.09 → −1.93)	−2.50 → −2.49 (8/8)	−2.52 → −2.41 (10/10)	−2.49 → −1.66 (4/6)	−2.57 → −2.45 (6/6)	−2.53 → −2.49 (4/4)	−2.49 → −2.41 (5/5)
TX90p	3.25 (2.18 → 4.26)	3.19 → 3.25 (8/8)	2.85 → 3.60 (10/10)	2.14 → 3.25 (5/6)	2.77 → 3.29 (6/6)	3.25 → 3.89 (4/4)	2.88 → 3.25 (5/5)
TN10p	−4.19 (−4.84 → −3.57)	−4.23 → −4.17 (8/8)	−4.22 → −3.96 (10/10)	−4.19 → −3.55 (5/6)	−4.22 → −3.67 (6/6)	−4.19 → −4.12 (4/4)	−4.29 → −4.18 (5/5)
TN90p	5.84 (4.66 → 7.07)	5.76 → 6.02 (8/8)	5.24 → 5.99 (10/10)	4.24 → 5.84 (4/6)	4.48 → 5.86 (4/6)	5.84 → 5.97 (4/4)	4.68 → 5.84 (5/5)
FD	−1.75 (−2.23 → −1.30)	−1.77 → −1.73 (8/8)	−2.09 → −1.14 (9/10)	−1.75 → −1.43 (6/6)	−1.75 → −1.34 (6/6)	−1.82 → −1.07 (3/4)	−2.09 → −1.75 (5/5)
ID	−0.70 (−1.00 → −0.42)	−0.71 → −0.68 (8/8)	−1.14 → −0.56 (5/10)	−0.95 → −0.66 (6/6)	−1.61 → −0.70 (4/6)	−0.93 → −0.70 (4/4)	−1.35 → −0.70 (1/5)
SU	1.07 (0.69 → 1.42)	1.06 → 1.11 (8/8)	0.90 → 1.39 (10/10)	0.36 → 1.07 (3/6)	0.85 → 1.30 (6/6)	0.51 → 1.07 (3/4)	1.02 → 1.38 (5/5)
TR	1.24 (0.95 → 1.52)	1.16 → 1.27 (8/8)	0.96 → 1.94 (9/10)	0.69 → 1.24 (2/6)	0.78 → 1.24 (5/6)	1.12 → 1.74 (2/4)	1.08 → 1.54 (4/5)
				Precipitation			
Rx1day	0.42 (0.18 → 0.69)	0.40 → 0.46 (8/8)	0.33 → 0.82 (7/10)	0.42 → 0.62 (6/6)	0.42 → 0.63 (6/6)	0.11 → 0.47 (3/4)	0.13 → 0.54 (4/5)
Rx5day	0.49 (−0.03 → 1.03)	0.40 → 0.54 (8/8)	0.39 → 1.27 (7/10)	0.48 → 0.76 (6/6)	0.49 → 0.82 (6/6)	0.23 → 0.49 (4/4)	0.25 → 0.87 (5/5)
PRCPTOT	4.50 (1.66 → 7.19)	4.30 → 4.68 (8/8)	2.10 → 5.29 (10/10)	3.51 → 5.70 (6/6)	4.50 → 9.58 (3/6)	4.50 → 6.22 (4/4)	4.50 → 8.85 (3/5)
SDII	0.05 (0.03 → 0.07)	0.05 → 0.05 (8/8)	0.03 → 0.08 (5/10)	0.05 → 0.07 (5/6)	0.04 → 0.06 (6/6)	0.04 → 0.05 (4/4)	0.04 → 0.07 (3/5)
R95p	3.29 (2.08 → 4.66)	3.08 → 3.37 (8/8)	2.69 → 5.24 (6/10)	3.29 → 5.20 (4/6)	3.18 → 4.31 (6/6)	2.99 → 3.36 (4/4)	2.92 → 5.43 (3/5)
R95pTOT	0.30 (0.18 → 0.42)	0.29 → 0.31 (8/8)	0.26 → 0.45 (6/10)	0.30 → 0.45 (4/6)	0.28 → 0.34 (6/6)	0.23 → 0.31 (4/4)	0.21 → 0.30 (5/5)
R99p	1.60 (0.81 → 2.44)	1.50 → 1.74 (8/8)	1.38 → 2.73 (6/10)	1.60 → 2.39 (6/6)	1.60 → 1.97 (6/6)	1.37 → 1.60 (4/4)	1.54 → 2.79 (4/4)
R99pTOT	0.14 (0.06 → 0.22)	0.13 → 0.15 (8/8)	0.12 → 0.23 (8/10)	0.14 → 0.19 (6/6)	0.12 → 0.14 (6/6)	0.10 → 0.14 (4/4)	0.12 → 0.17 (5/5)
CWD	−0.01 (−0.03 → 0.01)	−0.01 → −0.01 (8/8)	−0.01 → 0.02 (6/10)	−0.01 → 0.01 (6/6)	−0.01 → 0.02 (3/6)	−0.01 → 0.03 (3/4)	−0.01 → 0.02 (4/5)
CDD	0.24 (−0.10 → 0.59)	0.21 → 0.31 (8/8)	−0.54 → 0.28 (6/10)	−0.12 → 0.24 (4/6)	−0.61 → 0.24 (1/6)	−0.08 → 0.32 (4/4)	−0.42 → 0.24 (2/5)
R10mm	0.14 (0.05 → 0.20)	0.13 → 0.15 (8/8)	0.04 → 0.14 (9/10)	0.12 → 0.19 (6/6)	0.13 → 0.27 (3/6)	0.10 → 0.14 (4/4)	0.13 → 0.23 (2/5)
R20m	0.04 (−0.01 → 0.10)	0.04 → 0.05 (8/8)	0.04 → 0.15 (6/10)	0.04 → 0.10 (5/6)	0.04 → 0.13 (3/6)	0.04 → 0.08 (4/4)	0.04 → 0.11 (3/5)

effect of the averaging algorithm, nor the latitude weighting, and we have also only used one method for obtaining linear trends (median of pairwise slopes), and in some cases a linear trend is unlikely to be the most applicable. These unassessed sources of uncertainty will also have an impact on conclusions drawn from the data sets. Further work is required to pull through the observational uncertainties into gridded data sets of climate extremes.

The indices which have strong global trends in HadEX2 continue to have strong global trends under all of the model choices assessed here, though in some cases with reduced amplitude or increased short-term variability. On the whole, these are the temperature-based indices. Regionally, the areas with a high station density are also more robust to the different methodological choices, with high correlations between the different choices and small variances in the trends for each grid box. Those areas which have a lower station density are more susceptible to local changes in the trends and short-timescale behaviour arising from the effect of the methodological choices. Users should therefore be cautious when using these data sets for small regions and be aware of the coverage of the data when doing assessments. We note that in many cases it would be possible to perform a more in-depth analysis than that presented here, especially at a regional level. To allow this, all of the data files for each index of the methodological choices presented here will be made available for this purpose at www.metoffice.gov.uk/hadobs/hadex2/. However, by focusing on the global scales, we have aimed to assess the uncertainties related to the main applications of the data sets: the investigation of long-term trends and the inter-annual variability of the ETCCDI indices.

The main limitation to the data set and its cousins is the availability of the data. All the different choices and experiments run in this assessment take larger or smaller fractions of the available station data and process them in similar (albeit slightly different) ways. Therefore the global time series for indices which have a strong trend are very similar, as all grid boxes are interpolated from the same parent data. For indices which have no strong trend or are inherently more variable, the changes in methods rarely introduce a strong trend or a drastic change into the variability either.

6 Summary

We have assessed the effects of a number of methodological choices, both parametric and structural, which were made during the creation of the HadEX2 data set of gridded extremes indices. This allows for the quantification of some of the uncertainties present within the HadEX2 data set. The largest effects on global average time series come from those methodological choices which make large changes to the final spatial coverage of the data set or to the grid box values themselves. The main choices which result in these kinds of changes are changes in the station network (sub-sampling experiments), the gridding method used and the requirement to have a certain number of stations within the grid box or DLS. When comparing the global average time series with those from HadEX2 using a Taylor diagram, these choices have the lowest correlations and the largest difference in standard deviation. Trends and variances are most robust in North America, Europe and Asia as well as the southern tip of Africa and eastern Australia. These are also the areas which have the highest station network density. High-latitude regions and the majority of South America and Africa often have lower agreement for all indices, as these have low station densities and thus are more susceptible to changes in coverage and local grid box values. Temperature indices are to be more coherent and resistant to changes in the methods than precipitation indices. In regions with high station density, and for indices which have a clear non-zero trend over 1950–2010, the linear trends from almost all choices fall within the statistical range of trends from HadEX2, indicating that the structural and parametric uncertainties of the linear trends are smaller than the statistical uncertainties for these predominantly temperature indices. For these, HadEX2 and its related data sets is robust to choices in the creation method. The precipitation indices show more variation as a result of the different parametric and structural choices, but for the later period, there is also high consistency between them. For indices that have no strong non-zero trend (predominantly precipitation indices), the long-term behaviour can be different for each of the choices, especially for the early period.

The Supplement related to this article is available online at doi:10.5194/cp-10-2171-2014-supplement.

Acknowledgements. We thank the four anonymous referees, whose comments and suggestions improved the manuscript.

R. J. H. Dunn was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101) and thanks the Australian Research Council's Centre of Excellence for Climate System Science (UNSW, Sydney) for support and hospitality during the final stages of this project. M. G. Donat and L. V. Alexander were supported by Australian Research Council grants CE110001028 and LP100200690. We thank David Parker and Kate Willett for suggestions which improved the final manuscript. R. J. H. Dunn thanks Gareth Jones for suggesting the use of a Taylor diagram and Colin Morice and John Kennedy for useful discussions about gridding methods.

Edited by: J. Luterbacher

References

Alexander, L., Zhang, X., Peterson, T., Caesar, J., Gleason, B., Klein Tank, A., Haylock, M., Collins, D., Trewin, B., Rahimzade h, F., Tagipour, A., Rupa Kumar, K., Revadekar, J., Griffiths, G., Vincent, L., Stephenson, D. B., Burn, J., and Aguilar, E., Brunet, M., Taylor, M., New, M., Zhai, P., Rusticucci M., Vazquez-Aguirre, J. L.: Global observed changes in daily climate extremes of temperature and pr ecipitation, J. Geophys. Res.-Atmospheres (1984–2012), 111, doi:10.1029/2005JD006290, 2006.

Caesar, J., Alexander, L., and Vose, R.: Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set, J. Geophys. Res., 111, D05101, doi:10.1029/2005JD006280, 2006.

Clark, M. R., Lee, D. S., and Legg, T. P.: A comparison of screen temperature as measured by two Met Office observing systems, Internat. J. Climatol., 34, 2269–2277, 2014.

Cressie, N.: Statistics for Spatial Data: Wiley Series in Probability and Statistics, 1993.

Domonkos, P.: Adapted Caussinus-Mestre Algorithm for Networks of Temperature series (ACMANT), Int. J. Geosci., 2, 293–309, 2011.

Donat, M., Alexander, L., Yang, H., Durre, I., Vose, R., Dunn, R., Willett, K., Aguilar, E., Brunet, M., Caesar, J., Hewitson, B., Jack, C., Klein Tank, A. M. G., Kruger, A. C., Marengo, J., Peterson, T. C., Renom, M., Rojas, C. O., Rusticucci, M., Salinger, J., Elrayah, A. S., Sekele, S. S., Srivastava, A. K., Trewin, B., Villarroel, C., Vincent, L. A., Zhai, P., Zhang, X., and Kitching, S.: Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset, J. Geophys. Res. Atmos., 118, 2098–2118, 2013a.

Donat, M. G., Alexander, L. V., Yang, H., Durre, I., Vose, R., and Caesar, J.: Global land-based datasets for monitoring climatic extremes, Bull. Am. Meteorol. Soc., 94, 997–1006, 2013b.

Donat, M. G., Sillmann, J., Wild, S., Alexander, L. V., Lippmann, T., and Zwiers, F. W.: Consistency of temperature and precipitation extremes across various global gridded in situ and reanalysis data sets, J. Climate, 27, 5019–5035, 2014.

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

Hansen, J. and Lebedeff, S.: Global trends of measured surface air temperature, J. Geophys. Res.-Atmospheres (1984–2012), 92, 13345–13372, 1987.

Hansen, J., Ruedy, R., Sato, M., and Lo, K.: Global surface temperature change, Rev. Geophys., 48, doi:10.1029/2010RG000345, 2010.

Hartmann, D. L., Tank, A. M. G. K., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., Soden, B. J., Thorne, P. W., Wild, M., and Zhai, P. M.: Observations: Atmosphere and Surface, in: Climate Change 2013: The Physical Science Basis, in: Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 159–254, 2013.

Hofstra, N. and New, M.: Spatial variability in correlation decay distance and influence on angular-distance weighting interpolation of daily precipitation over Europe, Internat. J. Climatol., 29, 1872–1880, 2009.

Jones, P., Osborn, T., and Briffa, K.: Estimating sampling errors in large-scale temperature averages, J. Climate, 10, 2548–2568, 1997.

Jones, P. D.: Hemispheric surface air temperature variations: a reanalysis and an update to 1993, J. Climate, 7, 1794–1802, 1994.

Journel, A. G. and Huijbregts, C. J.: Mining geostatistics, Academic press, 1978.

Kent, E. C., Challenor, P. G., and Taylor, P. K.: A statistical determination of the random observational errors present in voluntary observing ships meteorological reports, J. Atmos. Oc. Technol., 16, 905–914, 1999.

Klok, E. and Klein Tank, A.: Updated and extended European dataset of daily climate observations, Internat. J. Climatol., 29, 1182–1191, 2009.

Lanzante, J. R.: Resistant, Robust and Non-Parametric techniques for the analysis of Climate Data: Theory and Examples, including Applications to Historical Radiosonde Station Data, Internat. J. Climatol., 16, 1197–1226, 1996.

Lindau, R.: Errors of Atlantic air-sea fluxes derived from ship observations, J. Climate, 16, 783–788, 2003.

Matthews, J. L., Mannshardt, E., and Gremaud, P.: Uncertainty Quantification for Climate Observations, Bull. Am. Meteorol. Soc., 94, ES21–ES25, 2013.

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

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

Mestre, O., Gruber, C., Prieur, C., Caussinus, H., and Jourdain, S.: SPLIDHOM: A method for homogenization of daily temperature observations, J. Appl. Meteorol. Climatol., 50, 2343–2358, 2011.

Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, J. Geophys. Res.-Atmospheres (1984–2012), 117, doi:10.1029/2011JD017187, 2012.

New, M., Hulme, M., and Jones, P.: Representing twentieth-century space-time climate variability – Part II: Development of 1901–1996 monthly grids of terrestrial surface climate, J. Climate, 13, 2217–2238, 2000.

North, G. R., Wang, J., and Genton, M. G.: Correlation models for temperature fields, J. Climate, 24, 5850–5862, 2011.

Peterson, T. C., Karl, T. R., Jamason, P. F., Knight, R., and Easterling, D. R.: First difference method: Maximizing station density for the calculation of long-term global temperature change, J. Geophys. Res.-Atmospheres (1984–2012), 103, 25967–25974, 1998.

Rohde, R., Muller, R., Jacobsen, R., Permutter, S., Rosenfeld, A., Wurtele, J., Curry, J., Wickham, C., and Mosher, S.: Berkeley Earth Temperature Averaging Process, Geoinfor Geostat: An Overview, 1, doi:10.4172/2327-4581.1000103, 2013.

Sen, P. K.: Estimates of the Regression Coefficient Based on Kendall's Tau, J. Am. Statist. Assoc., 63, 1379–1389, 1968.

Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data, in: Proceedings of the 1968 23rd ACM national conference, 517–524, ACM, 1968.

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., and Bronaugh, D.: Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate, J. Geophys. Res.-Atmospheres, 118, 1716–1733, 2013.

Smith, T. M., Reynolds, R. W., Peterson, T. C., and Lawrimore, J.: Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006), J. Climate, 21, 2283–2296, 2008.

Sun, B., Baker, C. B., Karl, T. R., and Gifford, M. D.: A comparative study of ASOS and USCRN temperature measurements, J. Atmos. Oc. Technol., 22, 679–686, 2005.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res.-Atmospheres (1984–2012), 106, 7183–7192, 2001.

Theil, H.: A rank-invariant method of linear and polynomial regression analysis. I, II, III, Nederl. Akad. Wetensch., Proc., 53, 386–392, 521–525, 1397–1412, 1950.

Toreti, A., Kuglitsch, F. G., Xoplaki, E., and Luterbacher, J.: A novel approach for the detection of inhomogeneities affecting climate time series, J. Appl. Meteorol. Climatol., 51, 317–326, 2012.

Trenberth, K. E., Dai, A., van der Schrier, G., Jones, P. D., Barichivich, J., Briffa, K. R., and Sheffield, J.: Global warming and changes in drought, Nat. Clim. Change, 4, 17–22, 2014.

Venema, V. K. C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J. A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C. N., Menne, M. J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Prohom Duran, M., Likso, T., Esteban, P., and Brandsma, T.: Benchmarking homogenization algorithms for monthly data, Clim. Past, 8, 89–115, doi:10.5194/cp-8-89-2012, 2012.

Wan, H., Zhang, X., Zwiers, F. W., and Shiogama, H.: Effect of data coverage on the estimation of mean and variability of precipitation at global and regional scales, J. Geophys. Res.-Atmos., 118, 534–546, 2013.

Williams, C. N., Menne, M. J., and Thorne, P. W.: Benchmarking the performance of pairwise homogenization of surface temperatures in the United States, J. Geophys. Res.-Atmospheres (1984–2012), 117, doi:10.1029/2011JD016761, 2012.

Yin, H., Donat, M. G., Alexander, L. V., and Sun, Y.: Multi-dataset comparison of gridded observed temperature and precipitation extremes over China, Internat. J. Climatol., doi:10.1002/joc.4174, 2014.