The Sensitivity of Daily Temperature Variability and Extremes to Dataset Choice

MIA H. GROSS, MARKUS G. DONAT, AND LISA V. ALEXANDER

*Climate Change Research Centre and ARC Centre of Excellence for Climate System Science,
University of New South Wales, Sydney,
New South Wales, Australia*

SCOTT A. SISSON

*School of Mathematics and Statistics, University of New South Wales, Sydney,
New South Wales, Australia*

(Manuscript received 12 April 2017, in final form 20 September 2017)

ABSTRACT

Robust conclusions regarding changes in the temperature distribution rely on the accuracy and reliability of the input datasets used. Differences between methodologies and datasets in previous studies add uncertainty when comparing and quantifying findings. Here, the authors investigate the sensitivity of assessing global and regional temperature variability and extremes over 1980–2014 in gridded datasets of daily temperature anomalies. A gridded in situ–based dataset, Hadley Centre Global Historical Climatology Network–Daily (HadGHCND), is compared against several commonly used reanalysis products by assessing both the entire distribution and the tails of the distribution. Empirical probability distribution functions show sensitivity to the input dataset when estimating aspects such as standard deviation and skewness, with the mean showing robust results for most regions, irrespective of dataset choice. Standard deviation is especially sensitive, with larger disagreements between datasets for some regions more than others, such as Africa and the Mediterranean region, and with larger differences in minimum temperatures compared with maximum temperatures. Estimates of extreme parameters also show sensitivity to dataset choice, particularly in the lower tails and for daily minimum temperature anomalies. Comparing changes in the means and the extremes of the temperature distributions, the cold extremes in the lower tails have been warming at a faster rate than the mean of the entire distribution for much of the Northern Hemisphere extratropics, with warm extremes warming at a faster rate than the mean in some subtropical regions. These documented sensitivities call for caution when assessing changes in temperature variability and extremes, as dataset choice can have substantial effects on results.

1. Introduction

Temperature extremes represent one of the most obvious impacts of climate change on society. For example, the impacts from heatwaves can range from increases in human mortality and morbidity to effects on agriculture and infrastructure (IPCC 2012). Variability within temperature can affect the probability and frequency of extreme events (Mearns et al. 1984; Katz and Brown 1992; Rahmstorf and Coumou 2011; McKinnon

et al. 2016), making it critical to understand how different aspects of the temperature distribution are changing and how they might change in the future. For this reason, many studies have investigated how extremes might be affected by changes in the distribution of temperature due to climate change (e.g., Rahmstorf and Coumou 2011; Donat and Alexander 2012; Hansen et al. 2012; Rhines and Huybers 2013; McKinnon et al. 2016). Despite the abundance of such studies, there remain some contested issues regarding changes in temperature variability. Mainly, while some studies have found increases in global temperature variability (e.g., Hansen et al. 2012), others have concluded that extremes are shifting toward hotter temperatures along with the mean, with little or no change in global variability (e.g., Brown et al. 2008; Donat and Alexander

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JCLI-D-17-0243.s1>.

Corresponding author: Mia H. Gross, mia.gross@unsw.edu.au

2012; Huntingford et al. 2013; Rhines and Huybers 2013; Tingley and Huybers 2013). This suggests that while there is consensus in terms of changes in the mean temperature, there remains considerable uncertainty in changes of other aspects of the distribution.

The issues and uncertainties surrounding previous studies are likely related to the different methods and datasets used for investigating changes in the distribution of temperature, which make these studies difficult to compare and quantify. Most studies to date have used monthly or seasonal datasets (e.g., Hansen et al. 2012; Coumou and Robinson 2013; Rhines and Huybers 2013). However, the temporal aggregation of these datasets can smooth out the individual events that can occur on daily time scales. This makes it critical to use daily data to detect changes in the temperature distribution, as it is the characteristics of extremes (such as the frequency, intensity, and duration) that are most likely to impact society (Alexander and Perkins 2013). This requires long-term, continuous, and consistent high-quality datasets (Klein Tank et al. 2009).

The limited studies that have used daily data to investigate changes in the temperature distribution have predominantly used three main statistical methods. This includes assessing statistical moments of the probability density function (PDF), such as the mean, variance, and skewness (e.g., Donat and Alexander 2012), using extreme value theory (EVT) to explicitly characterize the tails of the distribution (e.g., Kharin and Zwiers 2005; Brown et al. 2008; Christidis et al. 2011), and examining changes in different percentiles (e.g., Robeson 2004; Simolo et al. 2010; McKinnon et al. 2016). Some of these studies (e.g., Brown et al. 2008; Donat and Alexander 2012) have used Hadley Centre Global Historical Climatology Network–Daily (HadGHCND) (<https://www.metoffice.gov.uk/hadobs/hadghcnd/>), a quasi-global daily gridded temperature dataset (Caesar et al. 2006). However, such an in situ–based daily dataset still lacks data for certain regions, such as parts of South America and Africa. Reanalysis products that use assimilated observational data for the globe are therefore commonly used to investigate global changes in variability and extremes (e.g., Huntingford et al. 2013; Donat et al. 2014). Additionally, because of their global completeness and areal structure, reanalyses might be considered advantageous when evaluating climate models against historical observation-based data (e.g., Kharin et al. 2013; Sillmann et al. 2013; Donat et al. 2016). However, it is still not clear if assessments of changes in temperature variability and extremes are sensitive to these different types of input data, a factor that might affect a study's conclusions.

A crucial first step in gaining a more comprehensive understanding of how temperature variability and

extremes are changing requires a sensitivity analysis of some of the most commonly used datasets of daily temperatures. Further, a systematic and holistic approach would require a combination of assessing both the PDFs and the tails of the distribution using EVT (Katz et al. 2013; Sardeshmukh et al. 2015). Using both methods in parallel can help address the uncertainties in assessing the tail alone, as well as inferring changes in extremes relative to the mean of the entire distribution (Sardeshmukh et al. 2015). Here, we aim to use these two approaches to compare commonly used reanalysis products with the HadGHCND dataset. Our aim is not to make any judgments about the quality of the observational data itself, but rather to determine if analyses are sensitive to the input dataset. We use a mathematically consistent approach and consider both global and regional temperature distributions, to provide critical information for future studies that wish to use these types of datasets to investigate changes in temperature variability and extremes.

2. Data and methods

a. Observational data

We use HadGHCND as the base dataset to compare against reanalysis products. Since it is quasi-global over land and is based only on in situ daily maximum and minimum temperatures, this gives mostly independent information with which to compare against other products. It is available from 1950 and, at the time of analysis, ended in 2014. HadGHCND uses approximately 2500 stations that are interpolated onto a 2.5° latitude $\times$ 3.75° longitude grid using an angular distance weighting technique (Shepard 1968; Caesar et al. 2006).

b. Reanalysis data

Four reanalysis products are selected for intercomparison: ERA-Interim (Dee et al. 2011), NCEP–DOE Reanalysis-2 (NCEP2) (Kanamitsu et al. 2002), the Japanese 55-year Reanalysis (JRA-55) (Kobayashi et al. 2015), and the Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2) (Bosilovich et al. 2015; Gelaro et al. 2017). These were chosen primarily based on their prevalence in the literature for investigating temperature variability and extremes, as well as their comparability in terms of dataset length. We purposely did not include some commonly used first-generation reanalyses, because for NCEP–NCAR Reanalysis-1 (NCEP1) (Kalnay et al. 1996; Kistler et al. 2001) inhomogeneities in the representation of warm extremes have been documented (e.g., Donat et al. 2014). Other products, such as the Twentieth Century Reanalysis (20CR; Compo et al. 2011) and ERA-40 (Uppala et al. 2005), were excluded as

FIG. 1. The 26 regions specified in the IPCC SREX, and the spatial distribution of stations used in HadGHCND at the time of interpolation. An asterisk (*) indicates regions that are used for analysis of time series (i.e., Figs. 4–13). Regions 7, 8, 9, 15, and 16 are excluded from analyses based on not fulfilling the completeness criteria specified in the methods section (section 2).

data for recent years covered by the chosen reanalyses are not available.

To test the sensitivities of analyses to the different input datasets, we focus on the common time period of 1980–2014. While this is a relatively short period of time to investigate long-term climate changes, it is sufficient for the purpose of assessing robustness between different datasets. All reanalysis products are regridded using a bilinear remapping technique and masked to the grid cell size and spatial coverage of HadGHCND. Other regridding techniques were tested, such as conservative remapping techniques (Jones 1999), and results were not sensitive to the regridding method used. We also only include grid cells that have at least 80% of temporal completeness in HadGHCND, thereby excluding data-sparse regions. In addition to meeting this criterion, analysis is only performed when there are at least 50% of data available for both the first and last 10 years of HadGHCND data. This ensures there are enough data points to robustly assess changes over a 35-yr period. While the regridding itself is likely to add some uncertainty to the intercomparison of products, it has been shown that this effect is small compared to the structural differences between datasets (e.g., Loikith et al. 2015). We use daily maximum and minimum temperature anomalies that were calculated relative to a mean annual cycle over the entire period of investigation (i.e., 1980–2014).

c. Statistical methods

Two approaches are used to assess dataset sensitivity for investigating changes in temperature variability and extremes. First, we examine PDFs for the entire temperature distribution over the period of record chosen,

and second, we use EVT for direct analysis of the tails of the distribution and to understand how extremes are changing relative to the mean of the distribution.

We analyze PDFs for the 26 regions specified in the IPCC Special Report on Extremes (SREX) (IPCC 2012) (Fig. 1). These were chosen for this study as they indicate regions of common climates and enable sufficient data points for each region for reliable statistical analysis. Figure 1 shows the spatial distribution of all included 2500 stations used in the interpolation of HadGHCND. Furthermore, we show the number of stations per SREX region, and number of grid cells with data in each region in Table 1. Empirical PDFs are calculated by pooling all time steps and grid cells within a region to produce a distribution of temperature over the period 1980–2014. We note that different ways of how grid cells are spatially aggregated may affect estimates of distribution changes (Rhines and Huybers 2013; Director and Bornn 2015). The focus of our study, however, is on dataset agreement, and we treat all datasets the same. Therefore, such sensitivities regarding spatially averaged grid cells are less relevant here. We used 140 equally spaced bins, ranging from -45° to 45°C , to calculate the PDFs, as this choice most clearly captured the features of the dataset. We repeated our analyses for other bin width settings, but our conclusions were not sensitive to bin size choices for calculating the PDFs.

SREX regions are only included in the analysis if at least three grid cells within that region fulfill the completeness criteria specified in section 2b. While this criterion does not necessarily ensure spatial representativeness, it increases statistical power in terms of sample size for estimating the regional PDFs. For the

TABLE 1. Number of HadGHCND stations within each of the 26 SREX regions, and number of grid cells with observational data per region for daily maximum and daily minimum temperature anomalies. Counts of stations only consider the number of used stations available at the time of interpolation for HadGHCND, and so variations of stations over time are not considered. Counts of grid cells with available data per region are calculated after a criterion is applied where grid cells must have at least 80% temporal completeness in HadGHCND, as well as at least 50% of data available in the first and last 10 years of HadGHCND data.

SREX region	Number of stations per region	Number of grid cells with data per region (Tmax)	Number of grid cells with data per region (Tmin)
1. Alaska/northwest Canada	79	77	77
2. Canada/Greenland/Iceland	142	134	114
3. Western North America	1795	68	68
4. Central North America	1976	39	39
5. Eastern North America	1435	23	23
6. Central America/Mexico	182	4	4
7. Amazon	7	0	2
8. North-East Brazil	0	0	0
9. West coast of South America	1	0	0
10. Southeastern South America	25	15	8
11. Northern Europe	489	37	37
12. Central Europe	219	46	46
13. Mediterranean	138	39	39
14. Sahara/North Africa	16	14	15
15. West Africa	1	0	0
16. East Africa	18	0	0
17. Southern Africa	40	10	9
18. Northern Asia/Russia	375	245	245
19. West Asia	61	41	43
20. Central Asia	22	33	33
21. Tibetan Plateau	41	47	47
22. East Asia	350	86	86
23. South Asia	5	13	14
24. Southeast Asia	61	25	14
25. Northern Australia	123	50	50
26. Southern Australia/New Zealand	276	22	22

time series analysis in the main text, we only show selected regions that represent different levels of data coverage and different climatic zones, those being regions 3 (western North America), 10 (southeastern South America), 13 (Mediterranean region), and 18 (northern Asia/Russia). Figures for the remaining regions are included as online supplemental material (<https://doi.org/10.1175/JCLI-D-17-0243.s1>). We also include a global analysis, calculated by aggregating all grid cells that fulfill the completeness criteria.

The sample mean, standard deviation, and skewness are calculated for each region and dataset. To investigate temporal changes in these quantities, time series of annual statistics are plotted over the investigation period (see Figs. 4–8). We calculated trends using both a linear regression and Sen's trend estimator (Sen 1968), estimating trend significance using the Mann–Kendall test (Kendall 1975). Autocorrelation in the time series was accounted for by adjusting the sample size to the equivalent number of independent values, or equivalent sample size (Zwiers and von Storch 1995). Results were found to be insensitive to the test used, and so we only

show linear regression trends, with significance calculated at the 5% level using a t test.

We use EVT to determine if analyses of the extremes are sensitive to dataset choice and assess changes in the extremes explicitly. In particular, a nonstationary point process (PP) model is fitted to the data to describe annual exceedances above and below the upper and lower 1.5%, respectively, as in Brown et al. (2008). We additionally tested different thresholds, such as the upper and lower 1% and 2% of data, and found the results were not sensitive to the choice of threshold. This model corresponds to a Poisson process for the inter-exceedance times, with a generalized Pareto distribution for the size of the exceedances (e.g., Coles 2001). Similar to the methods of Brown et al. (2008), we model nonstationarity of the process through a time-dependent threshold and location parameter (explicitly, a linear trend in the location parameter) over the period of investigation, as implemented in the “extRemes” package in R (Gilleland and Katz 2011). Nonlinearity in scale and shape parameters was also tested but found not to be present, and the fidelity of the PP model to the data

was assessed using standard goodness-of-fit procedures, including quantile–quantile (QQ) plots. As in [Brown et al. \(2008\)](#), we calculate the difference of this location trend with the trend in the mean for all available grid cells to investigate how the magnitude of the extremes are changing relative to the mean, hereinafter referred to as “excess trends.” Local significance of the excess trends was calculated for each grid cell using a Mann–Kendall test at the 5% level.

In addition, we plotted time series of the location, scale, and shape parameters estimated from a stationary PP model, as shown in [section 3c](#). These are calculated from decadal moving windows, starting in 1980 and ending in 2014. Trends in the PP fit parameters are similarly calculated using the same methods as those used for the moments of the entire PDF.

3. Results

a. Global and regional PDFs

PDFs for SREX regions that adhere to our completeness criteria, as well as global PDFs, are shown for maximum temperature anomalies (Tmax; [Fig. 2](#)) and minimum temperature anomalies (Tmin; [Fig. 3](#)). Dashed vertical lines of different colors represent the threshold of the top and bottom 1.5% of the data (“extremes”) for each dataset. Plots of the respective standard deviation and skewness values are shown in the online supplemental material ([Fig. S1](#)).

As expected, higher-latitude regions generally have broader PDFs across all datasets, compared with lower latitudes for both maximum and minimum temperature anomalies, indicative of higher variability. For most regions, PDFs of the reanalyses are wider compared to the HadGHCND-derived PDFs for both the globe and most regions. This separation might be due to more spatial smoothing in HadGHCND, where the correlation length scale (CLS) used for interpolation is based on monthly mean temperatures ([Caesar et al. 2006](#)). This leads to search radii of several hundred to thousand kilometers for stations to include when calculating the gridbox values. It is possible that this monthly CLS is larger than the correlation of daily temperature values would be, potentially causing overly strong smoothing in the HadGHCND dataset. Broader PDFs in the reanalyses are especially apparent in the Mediterranean region (region 13) and some subtropical regions, such as northern Africa (region 14) and southern Africa (region 17). PDFs for Northern Hemisphere extratropical regions are also wider in the reanalyses compared with HadGHCND, although less so for parts of North America (e.g., regions 3–5) and northern Asia/Russia

(region 18). For all regions and the globe, there are fewer similarities between the PDFs of the different datasets for Tmin compared to Tmax. Skewness differences are larger between datasets for Tmin, with NCEP2 showing a more negatively skewed PDF for most regions compared with the other datasets.

The PDFs also show notable differences in the extremes of the distribution. That is, the 1.5% thresholds that define the tails of the distribution occur at varying temperatures depending on the dataset used, with greater differences found between datasets for Tmin compared to Tmax. For many regions, the thresholds in the reanalysis products occur at more extreme temperatures compared to HadGHCND. For instance, in western North America (region 3) the cold extremes threshold for Tmin is just above -18°C for NCEP2, while it is only just below -12°C for HadGHCND. For most regions as well as the globe, NCEP2-derived PDFs consistently show a bias toward more extreme thresholds, compared with the other datasets. This is particularly apparent in the cold tails compared with the warm tails of the distribution. Although less obvious, PDFs for ERA-Interim, JRA-55, and MERRA-2 also show a bias toward more extreme temperatures thresholds compared with those shown in HadGHCND.

In addition to the higher sensitivities found in Tmin compared to Tmax, the tails of the PDFs show sensitivity to the input dataset, particularly in the cold tails. Overall, the NCEP2-derived PDFs are the most different from HadGHCND; however, the PDFs show sensitivity to all four reanalysis products, particularly for assessing standard deviation and skewness.

b. Time series of the statistical moments

To explore temporal changes and trends in the datasets, [Figs. 4–8](#) show time series of the annual mean, standard deviation, and skewness for Tmax and Tmin for the globe and for selected regions (regions 3, 10, 13, and 18; see [Fig. 1](#)). The remaining regions and their respective trends are included in the supplemental material ([Figs. S2–S7](#), [Tables S1–S3](#)). These time series describe changes over 1980–2014. Decadal trends and their significance for each dataset are shown for each plot.

Results for the global mean are relatively robust, irrespective of the dataset used ([Fig. 4](#)), with a significant increasing trend found in all datasets for both Tmax and Tmin. Aside from some small differences in NCEP2, the most notable differences occur toward the end of the time series, particularly in MERRA-2 from around 2007–09. This discrepancy in MERRA-2 is a documented issue ([Sánchez-Lugo et al. 2016](#); [Simmons et al. 2017](#)).

Tmax

FIG. 2. Empirical PDFs of daily maximum temperature anomalies for all SREX regions that fulfill the completeness criteria, and for the globe. The PDFs are calculated from the data that are pooled over the period 1980–2014, with different colored PDFs representing the different datasets. Vertical long-dashed lines indicate the upper and lower 1.5% thresholds. An asterisk (*) indicates a different y-axis scale used for regions 13, 14, 23, and 24.

Tmin

FIG. 3. As in Fig. 2, but for daily minimum temperature anomalies.

Global

FIG. 4. Time series of global average daily maximum and minimum temperature anomalies for the first three statistical moments annually, that is, the mean, variance, and skewness. Decadal trends are shown for each dataset in respective colors. An asterisk (*) indicates significant trends at the 5% level.

Unlike the mean, changes in global standard deviation and skewness (Fig. 4) appear sensitive to the input dataset for assessing trends over the past 35 years; however, there is reasonable temporal correlation between the different time series. There is a noticeably larger spread in standard deviation across the datasets, compared with the mean or skewness. This is particularly apparent in Tmin and NCEP2, where NCEP2 shows a more substantial decreasing trend in global standard deviation in Tmin of -0.16°C per decade, compared with -0.03°C per decade in HadGHCND. This difference in trends is likely a result of a step change in NCEP2 that occurs around 1998 and is not evident in the other datasets. This inhomogeneity is likely a signature of a step change related to only specific regions, as discussed in subsequent paragraphs. The time series for global skewness show variation, and display mostly nonsignificant trends that are close to zero. Again, the largest differences are shown in NCEP2 for Tmin, with the discussed step change in NCEP2 likely contributing to this.

While the global mean is similar across datasets, differences become apparent for certain regions more than

others. For instance, the mean for western North America (region 3; Fig. 5) and northern Asia/Russia (region 18; Fig. 8) is mostly robust regardless of dataset choice, whereas in contrast southeastern South America (region 10; Fig. 6) and the Mediterranean (region 13; Fig. 7) display more sensitivity. This could be indicative of greater uncertainty within data-sparse regions compared to those that are more data rich. For example, regions 3 and 18 include 1795 and 375 station observations respectively; while regions 10 and 13 only have 25 and 138 observations respectively (see Table 1). With regard to the Mediterranean region, most available stations are located in southern Europe, with few stations in North Africa (see Fig. 1). Pooling the grid cells within each SREX region allows us to include region 13 in our analysis; however, the sensitivity of the Mediterranean discussed above might be a reflection of the unequal distribution of observations with data in this region.

The differences in the mean between datasets for regions 10 and 13 are most notable in NCEP2. As for the globe, variance and skewness show a greater sensitivity to dataset choice compared with the mean. Regions 3 and

FIG. 5. As in Fig. 4, but for SREX region 3, located in western North America.

18, both located in the Northern Hemisphere extratropics, show a clear step change in NCEP2 in standard deviation in Tmin occurring around 1998. After examination of all regions, the step change is found in most regions in this latitude band, as well as in region 10. This inhomogeneity in NCEP2 is dominant enough to have a signature in the global standard deviation. We assume that this step change is artificial, but a metadata search of NCEP2 around this time has not determined the cause. Further discussion on this issue is provided in section 4. While NCEP2 shows the largest differences across the regions, disagreements with HadGHCND are also found in ERA-Interim, JRA-55, and MERRA-2. For example, for Tmin there is a decreasing trend of -0.04°C in standard deviation for region 10 in HadGHCND, while the other datasets, excluding NCEP2, indicate an increasing trend. In terms of changes in skewness, trends are mostly nonsignificant and centered around zero. While in region 3 skewness values are similar between all datasets for Tmax, other regions show distinctions, particularly in NCEP2 and for Tmin.

In summary, irrespective of dataset choice, changes in mean Tmax and Tmin are robust for many regions, and a significant increasing trend is found for the globe. For the other statistical moments, a higher sensitivity to

dataset choice is clear. The degree of sensitivity, however, is shown to be dependent on the region, with regions of known high-quality data being more similar between datasets, with the exception of NCEP2.

c. Time series of decadal extreme value fits

To investigate changes in the extremes, Figs. 9–13 show time series of the location, scale, and shape parameters that have been estimated from stationary PP fits over 10-yr moving windows. The extremal point process model is fitted to the high and low tails of both daily maximum and minimum temperature anomalies ($T_{\text{max,high}}$, $T_{\text{max,low}}$, $T_{\text{min,high}}$, and $T_{\text{min,low}}$). As before, extreme values are defined as those data points that exceed a threshold of 1.5% (in the upper or lower tail). We use running decadal windows to ensure sufficient data points are available for a robust statistical analysis.

For the globe (Fig. 9), significant increasing trends in the location parameter are shown for the high tails of both Tmax and Tmin across all datasets. Here, a trend of 0.21 per decade ($T_{\text{max,high}}$) and 0.31 per decade ($T_{\text{min,high}}$) is calculated for HadGHCND, with similar trends in most reanalyses. Mostly significant decreasing trends are found for the low tails of Tmax and Tmin,

Region 10

FIG. 6. As in Fig. 4, but for SREX region 10, located in southeastern South America.

excluding nonsignificant trends in HadGHCND and MERRA-2 ($T_{\max_{\text{low}}}$). The time series of NCEP2 for $T_{\min_{\text{low}}}$ again shows a step change occurring around 1998, where the location parameter decreases substantially more than for the other datasets. Although smoother than the step change in NCEP2 shown for global variance (Fig. 4), because of the use of decadal windows compared with annual data, this suggests that the inhomogeneity in NCEP2 is related to the cold tail of the distribution of minimum temperature anomalies. Section 4 discusses this in more detail. Overall, the location parameter for the globe indicates increasing trends in the high tails, along with decreasing trends in the low tails. This indicates a shifting distribution toward warmer conditions, with both and cold extremes becoming warmer.

Trends in the global-scale parameter (Fig. 9) are close to zero for all datasets and temperature variables, although slightly decreasing significant decadal trends in HadGHCND are shown for all variables except $T_{\min_{\text{high}}}$. There is some sensitivity in scale to the input dataset, with $T_{\min_{\text{low}}}$ showing the most differences between datasets compared with the other temperature variables. As for the location parameter, NCEP2

shows a sudden decrease in scale for $T_{\min_{\text{low}}}$ starting with the 1994–2003 window that is not apparent in the other datasets. This is also shown to a lesser extent in $T_{\max_{\text{low}}}$. Other notable differences include a steeper decreasing trend in JRA-55 and MERRA-2 compared with HadGHCND in both low tails, and a slightly increasing significant trend in ERA-Interim in $T_{\min_{\text{low}}}$. Trends in global shape are mostly nonsignificant and close to zero, excluding a significant increasing trend in ERA-Interim and MERRA-2 ($T_{\max_{\text{high}}}$). For all temperature variables, the reanalyses show different temporal patterns to HadGHCND. Again, this is particularly evident in Tmin, and in the low tails.

As shown for the statistical moments in the previous subsection, the sensitivities in the time series analyses differ between both region and temperature variable. For the location parameter, NCEP2 consistently stands out as being the most different across all regions (including those in the supplemental material; see Figs. S8–S19 and Tables S4–S9). The step change shown in global location for $T_{\min_{\text{low}}}$ is present for both western North America (region 3; Fig. 10) and northern Asia/Russia (region 18; Fig. 13), and most other Northern Hemisphere extratropical regions, but is not a feature in the

Region 13

FIG. 7. As in Fig. 4, but for SREX region 13, located in the Mediterranean region.

time series for southeastern South America (region 10; Fig. 11) and the Mediterranean (region 13; Fig. 12). This suggests that the apparent inhomogeneity in the global time series of NCEP2 is likely related to the cold extremes for Northern Hemisphere extratropical regions. We note that we do not further discuss long-term trends in NCEP2 due to this inhomogeneity. Overall, the location parameter shows sensitivity to all reanalysis products for all regions, with more data-sparse regions showing the least consistency between datasets. For example, for region 10, HadGHCND shows a steeper decrease in location toward the end of the time series for $T_{\max_{\text{low}}}$, compared to the other datasets. Region 23, another data-sparse region located in South Asia, shows a substantial step change in NCEP2 in location for $T_{\max_{\text{low}}}$ that is not present in the other datasets (see Fig. S10). Trends in scale and shape also differ between datasets. For example, the reanalyses show clear deviances from HadGHCND in scale and shape in the cold tails for region 10 and in more data-rich regions, such as regions 3 and 18. JRA-55 shows a more substantial decreasing trend in scale for $T_{\min_{\text{low}}}$, compared with the other datasets (excluding NCEP2, which is affected by the discussed inhomogeneity). Other key differences are found for region 18 for $T_{\max_{\text{low}}}$. Here, HadGHCND

increases in scale from 1988 to 1997, during which time JRA-55 shows a decreasing trend in scale. ERA-Interim and MERRA-2 capture the temporal pattern of HadGHCND here, although to a lesser extent. However, differences between these products and HadGHCND are apparent for other regions (e.g., changes in shape for $T_{\min_{\text{low}}}$ in region 13). In general, the reanalyses are distinct from HadGHCND in both scale and shape.

In summary, changes in the tails of the distribution are sensitive to dataset choice. As for the analyses of the entire distribution, more overall differences in the datasets are shown for minimum temperature anomalies, with differences particularly evident in the scale and shape parameters. Analyses of changes in the extremes correspond with these findings and, more specifically, show that the largest dataset inconsistencies occur in the cold tails of the distribution of minimum temperatures.

d. Changes in extremes relative to changes in the mean

Using the above methods together (i.e., jointly assessing both the entire distribution as well as the tails) can provide insight into how extremes are changing with respect to the mean. Figures 14 and 15 show the difference between the trend in the nonstationary PP fit of the location parameter and the trend in the annual mean

Region 18

FIG. 8. As in Fig. 4, but for SREX region 18, located in northern Asia/Russia.

temperature anomalies. NCEP2 is not included due to the inhomogeneity present in the time series, making it inappropriate to assess trends over a 35-yr period. The trend differences (calculated per decade for the period 1980–2014) represent so-called excess trends and are a useful way of describing regional differences in the rates of warming between extremes and the mean (Brown et al. 2008). As in the subsequent figures, when extremes are warming faster (slower) than the mean, the excess trend is positive (negative). Stippling indicates grid boxes that are significant at the 5% level. For additional robustness, we also calculated the difference between the 98.5th quantile (1.5th quantile for the cold tails) and the mean (see Fig. S20). These results are very similar to the excess trends shown here.

For the warm extremes ($T_{\max_{\text{high}}}$; Fig. 14), most Northern Hemisphere high-latitude regions indicate that the mean has been warming faster than the extremes, irrespective of the product used. This is slightly underestimated by JRA-55. The datasets agree that warm extremes have been warming significantly faster than the mean in the Mediterranean region, southwestern Asia, East Asia, South America, Africa, and parts of Australia. For some of these regions, such as

the Mediterranean, the reanalyses show smaller magnitude excess trends in the extremes compared to HadGHCND, with rates of around $0.1^{\circ}\text{--}0.2^{\circ}\text{C}$ (reanalyses) compared to $0.4^{\circ}\text{--}0.6^{\circ}\text{C}$ (HadGHCND). Other regions, such as South America, show greater magnitude positive excess trends in the reanalyses compared with HadGHCND. Differences are also found for parts of northern Canada, where a positive excess trend is shown in HadGHCND. This same area is opposite in sign for ERA-Interim, and smaller and nonsignificant in JRA-55 and MERRA-2.

Excess trends in $T_{\min_{\text{high}}}$ (Fig. 14) are similar to those of the warm extremes, in terms of both spatial patterns and sign of trend, as well as dataset agreement, although positive excess trends are generally of a smaller magnitude to those of the warm extremes. Some differences in the datasets are shown for Australia, for example, where ERA-Interim shows greater positive excess trends in parts of eastern Australia compared with HadGHCND, while MERRA-2 and JRA-55 tend to show smaller magnitude excess trends than HadGHCND for much of the country.

The datasets differ the most for excess trends in the low tails (Fig. 15). Positive excess trends in extremes are shown

FIG. 9. Global time series using running decadal windows of the extreme value parameters, including location, scale, and shape, calculated for the high tails of maximum and minimum temperature anomalies ($T_{\max_{\text{high}}}$ and $T_{\min_{\text{high}}}$, respectively), and for the low tails of maximum and minimum temperature anomalies ($T_{\max_{\text{low}}}$ and $T_{\min_{\text{low}}}$, respectively). The x axis shows years, starting with the 1980–89 window, and ending with 2005–14, and parameter estimates are indicated on the y axis. Decadal trends are indicated in the respective dataset colors for each parameter and temperature variable. An asterisk (*) indicates significant trends at the 5% level.

for much of the Northern Hemisphere extratropics, particularly in North America. Negative excess trends are found for parts of Europe and Asia, while small negative excess trends are shown for some tropical and Southern Hemisphere regions. For $T_{\max_{\text{low}}}$, most regions with positive excess trends show larger magnitude trends in the reanalyses than for HadGHCND. Conversely, regions showing negative excess trends in HadGHCND show smaller magnitude excess trends in the reanalyses, such as in parts of Europe and western Asia. Broadly, JRA-55 shows the most differences to HadGHCND, with larger significant positive excess trends for much of the Northern Hemisphere extratropics and smaller negative trends in Europe. For example, positive excess trends greater than 0.7°C per decade are shown for some parts of Russia in JRA-55, with the same area showing negative excess trends between -0.2° and -0.3°C in HadGHCND. ERA-Interim and MERRA-2 also differ from HadGHCND in this area, showing mixed excess trends.

For the cold extremes ($T_{\min_{\text{low}}}$; Fig. 15), the largest disagreements occur over Russia and the European continent. As for $T_{\max_{\text{low}}}$, JRA-55 shows greater positive excess trends over Russia, and smaller negative excess trends over western Asia and central Europe, compared to the other datasets.

Overall, for the warm extremes, some subtropical regions show that the extremes have been warming at a faster rate than the mean over 1980–2014, while for the cold extremes positive excess trends are found for much of the Northern Hemisphere extratropics. Spatially, all datasets tend to agree on this; however, the magnitude of the excess trend differs. The warm tails of the distribution tend to be similar between datasets across the globe compared with the cold tails, which show greater sensitivity to dataset choice.

4. Discussion

The choice of dataset for investigating changes in temperature variability and extremes can affect conclusions regarding changes in the temperature distribution. The use of different datasets in previous work along with some differences in conclusions provides motivation for a systematic approach to determining the sensitivities of an analysis to the input datasets used. We acknowledge that nonclimatic artifacts can potentially affect a spatially averaged dataset such as HadGHCND, including changes to the station network, data quality, and homogeneity issues, and gridding uncertainties in terms of spatial averaging of point data, particularly for analyzing extremes (Donat et al. 2014; Dunn et al. 2014; Director and Bornn 2015). In addition, HadGHCND uses a large CLS and is gridded using a relatively coarse

resolution, so resultant values tend to be spatially smoothed (Caesar et al. 2006), further adding to the inherent uncertainties of using this type of dataset. However, given that it is interpolated using long-term in situ-based data, we are more willing to assume that there are fewer inhomogeneities than those introduced from assimilation using a highly variable network of data in the reanalyses.

Irrespective of what product is used, trends in the mean temperature are mostly robust. This suggests that changes in the mean are not particularly sensitive to dataset choice. However, some datasets show more differences in the mean than others, depending on the region, for example, in the Mediterranean region, and in more data-sparse regions such as parts of South America. Differences in the mean for these regions might be due to uncertainties in the observational data itself, rather than the ability of reanalysis data to reproduce the observations. It is also possible that the assimilation is poorly constrained by observations. Other statistical parameters aside from the mean, including the standard deviation and skewness, as well as those related explicitly to the extremes, show more sensitivity to the input dataset. As in the mean, regions with higher data uncertainty due to sparse observations, such as South America, show more disagreement between datasets than regions that are known to have high-quality data, as in North America.

Much of the recent literature regarding changes in temperature variability has highlighted the importance in understanding differences in regional changes in temperature compared with the global mean. It is already clear that some regions are warming at different rates compared to the global average (e.g., Seneviratne et al. 2016; Sutton et al. 2016). Here, we show that the globally averaged (where data are sufficiently complete) maximum temperature trend in HadGHCND is 0.36°C per decade over 1980–2014, whereas, for example, southeastern South America shows a trend of 0.67°C per decade (Figs. 4 and 6, respectively). In addition to the mean, we also see regional differences in the trend of extremes relative to the local mean. We show, for example, that cold extremes are warming faster than the mean for much of the Northern Hemisphere, consistent with Arctic amplification (Screen 2014; Rhines et al. 2017). For the warm extremes, some subtropical regions show a faster trend in the extremes compared with the mean. Additionally, excess trends in the cold extremes are generally greater in magnitude over North America and parts of northern Asia/Russia than they are for warm extremes, consistent with studies that have found stronger increases cold extremes compared with warm extremes (e.g., Donat et al. 2013).

Region 3

FIG. 10. As in Fig. 9, but for SREX region 3 (western North America).

Region 10

FIG. 11. As in Fig. 9, but for SREX region 10 (southeastern South America).

Region 13

FIG. 12. As in Fig. 9, but for SREX region 13 (Mediterranean).

Region 18

FIG. 13. As in Fig. 9, but for SREX region 18 (northern Asia/Russia).

FIG. 14. Excess trends, that is, the difference between the trend in the location parameter and the trend in the annual mean for HadGHCND and the reanalyses, for the high tails of maximum and minimum temperature anomalies ($T_{\max_{\text{high}}}$ and $T_{\min_{\text{high}}}$, respectively). Positive excess trends represent regions where the extremes are warming faster than the mean, while negative excess trends show regions where the mean is warming faster than the extremes. Decadal trends ($^{\circ}\text{C}$) are calculated over the period 1980–2014. Stippling indicates grid boxes that are significant at the 5% level.

This is robust for all datasets; however, the magnitude of excess trend differs substantially depending on the variable and region. This is a significant finding, as regions are disproportionately impacted by extremes due to differences in socioeconomic factors that can increase a region's vulnerability to extreme events (IPCC 2012). The consequences of higher-magnitude extremes can potentially further exacerbate the unequal distribution of impacts on regional scales.

Overall, ERA-Interim appears to most closely resemble HadGHCND. This is consistent with other studies. For example, Donat et al. (2014) found that

ERA-Interim closely resembled trends and interannual variations in temperature extremes shown in interpolated observational datasets for the period 1980–2010, noting the higher dataset consistency in the most recent three decades, in contrast with the large disagreements between datasets in the presatellite era. Excluding an inhomogeneity in MERRA-2 in the global mean for the most recent years (see Fig. 4), MERRA-2 and JRA-55 resemble mean trends in the observations reasonably well, even for more data-sparse regions (see Fig. 6). However, when assessing aspects of temperature other than the mean, there are still distinct differences

FIG. 15. As in Fig. 14, but for the low tails of maximum and minimum temperature anomalies ($T_{\max_{\text{low}}}$ and $T_{\min_{\text{low}}}$, respectively).

between the observation-based product and the reanalyses, so caution should be exercised if using these products to investigate changes in variability and extremes.

NCEP2 is clearly the most different from the other datasets, caused partly by the inhomogeneity shown to occur around 1998, but also shown in more short-term variations, such as in the mean for southeastern South America (Fig. 6). Previous studies have noted the differences in NCEP2 compared with other reanalyses (e.g., Kharin et al. 2007; Kharin et al. 2013; Sillmann et al. 2013). Being a low-resolution, first-generation product (Kanamitsu et al. 2002; NCAR 2016), this might be expected, but perhaps not to the extent shown here. While many errors were rectified from the original NCEP1 product, some suggest a poor representation of

the Southern Hemisphere in NCEP2 (NCAR 2016). Here, we show large differences in many regions around the globe. For example, the step change around 1998 shown for the globe is particularly noteworthy in some Northern Hemisphere high-latitude regions. Although we have attempted to explain this, we have yet to discover any documented reason. However, potential explanations could be related to changes in the assimilated data affecting snow cover, as these step changes are inconsistent with other datasets and are particularly notable for the cold extremes. Further investigation is warranted here.

The similarities or differences between the reanalysis products themselves as well as with HadGHCND are affected by their independence of in situ-based measurements. For example, NCEP2 does not incorporate

any screen-level data to assimilate near-surface temperature (Simmons et al. 2004), unlike the other reanalyses used here. ERA-Interim and JRA-55 both interpolate between in situ data at the screen level, and the atmospheric levels of the reanalysis model (Dee et al. 2011; Kobayashi et al. 2015). As such, ERA-Interim and JRA-55 are not independent of surface measurements, and so it may be expected that they are more similar to one another than to NCEP2. As in NCEP2, MERRA-2 does not assimilate using near-surface temperatures (Bosilovich et al. 2015). Despite MERRA-2 being the latest product used in this study, ERA-Interim and JRA-55 show more overall consistency between themselves as well as HadGHCND in many instances, as is similarly found in other comparison studies (e.g., Simmons et al. 2017). However, short-term variations, especially in the mean for MERRA-2, show close similarities to both ERA-Interim and HadGHCND.

5. Conclusions

Global daily maximum and minimum temperatures are increasing. This conclusion is robust regardless of the dataset used. Regional increases in the mean are mostly robust to the dataset; however, there is a slightly greater sensitivity to the input dataset than for global assessments. Other characteristics of the temperature distribution show substantial sensitivity to the dataset used, highlighting some of the uncertainties involved in addressing changes in temperature variability. Assessing temperature extremes also displays sensitivity to the dataset choice, with results showing sometimes substantially different magnitude changes in extremes.

Irrespective of the approach used, differences between datasets regarding distribution changes are the greatest in the cold tails of the distribution and for daily minimum temperature anomalies. This suggests potentially greater uncertainties for this part of the distribution that need to be considered for future work investigating changes in the distribution of temperature. Further investigation is required to understand why greater sensitivities are shown in the cold tails. For all temperature variables, NCEP2 has the largest inconsistencies, compared with the gridded in situ-based observational dataset, HadGHCND. However, the “higher generation” reanalysis products, that is, ERA-Interim, JRA-55, and MERRA-2, also still show distinct differences in the higher-order moments. Dataset disagreement is generally largest for regions that are more data sparse, such as southeastern South America and southern Africa, while better agreement between datasets is found for regions that are data rich and known to have higher-quality data, such as North America.

Despite inconsistencies in the results depending on the dataset, all products show that cold extremes have been warming at a faster rate than the mean for much of the Northern Hemisphere extratropics, while warm extremes have been warming faster for many subtropical regions. We must be able to make robust conclusions regarding the regional differences in rates of warming of extremes and the local mean, as future planning and local adaptation strategies rely on it. Future work in this field would benefit from using datasets of longer time scales to provide more robust trend estimates.

This paper provides a first step in documenting the inconsistencies regarding changes in temperature variability and extremes and, for example, will help in making the best dataset choices for model evaluation moving forward. In addition, by understanding these preliminary data issues, more confident and robust conclusions can be made to understand changes in the characteristics of the temperature distribution, and therefore provide critical information for future planning for extremes.

Acknowledgments. This study was supported by the Australian Research Council (ARC) Centre of Excellence for Climate System Science (Grant CE110001028). MGD received funding from the ARC (Grant DE150100456), and SAS from the ARC Centre of Excellence for Mathematical and Statistical Frontiers (Grant CE140100049). We thank John Caesar from the UK Met Office for providing information on the number and spatial distribution of stations available in HadGHCND. All data used are freely available. HadGHCND data can be downloaded from <http://www.metoffice.gov.uk/hadobs/hadghcnd/download.html> and were downloaded for this study on February 4, 2016. ERA-Interim data are available with registration from <http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/>, NCEP2 data can be downloaded from <https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis2.gaussian.html>, JRA-55 data are available with registration from http://jra.kishou.go.jp/JRA-55/index_en.html#download, and MERRA-2 data can be downloaded from <https://disc.sci.gsfc.nasa.gov/datasets?page=1&keywords=MERRA-2>.

REFERENCES

- Alexander, L., and S. Perkins, 2013: Debate heating up over changes in climate variability. *Environ. Res. Lett.*, **8**, 041001, <https://doi.org/10.1088/1748-9326/8/4/041001>.
- Bosilovich, M. G., and Coauthors, 2015: MERRA-2: Initial evaluation of the climate. Technical Report Series on Global Modeling and Data Assimilation, Vol. 43, NASA Tech. Rep.

- NASA/TM-2015-104606, 139 pp., <https://gmao.gsfc.nasa.gov/pubs/docs/Bosilovich803.pdf>.
- Brown, S. J., J. Caesar, and C. A. T. Ferro, 2008: Global changes in extreme daily temperature since 1950. *J. Geophys. Res.*, **113**, D05115, <https://doi.org/10.1029/2006JD008091>.
- Caesar, J., L. Alexander, and R. Vose, 2006: Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set. *J. Geophys. Res. Atmos.*, **111**, D05101, <https://doi.org/10.1029/2005JD006280>.
- Christidis, N., P. A. Stott, and S. J. Brown, 2011: The role of human activity in the recent warming of extremely warm daytime temperatures. *J. Climate*, **24**, 1922–1930, <https://doi.org/10.1175/2011JCLI4150.1>.
- Coles, S., 2001: *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics, Springer, 208 pp.
- Compo, G. P., and Coauthors, 2011: The Twentieth Century Reanalysis Project. *Quart. J. Roy. Meteor. Soc.*, **137**, 1–28, <https://doi.org/10.1002/qj.776>.
- Coumou, D., and A. Robinson, 2013: Historic and future increases in the global land area affected by monthly heat extremes. *Environ. Res. Lett.*, **8**, 034018, <https://doi.org/10.1088/1748-9326/8/3/034018>.
- Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, <https://doi.org/10.1002/qj.828>.
- Director, H., and L. Bornn, 2015: Connecting point-level and gridded moments in the analysis of climate data. *J. Climate*, **28**, 3496–3510, <https://doi.org/10.1175/JCLI-D-14-00571.1>.
- Donat, M. G., and L. V. Alexander, 2012: The shifting probability distribution of global daytime and night-time temperatures. *Geophys. Res. Lett.*, **39**, L14707, <https://doi.org/10.1029/2012GL052459>.
- , and Coauthors, 2013: 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, <https://doi.org/10.1002/jgrd.50150>.
- , J. Sillmann, S. Wild, L. V. Alexander, T. Lippmann, and F. W. Zwiers, 2014: Consistency of temperature and precipitation extremes across various global gridded in situ and reanalysis dataset. *J. Climate*, **27**, 5019–5035, <https://doi.org/10.1175/JCLI-D-13-00405.1>.
- , L. V. Alexander, N. Herold, and A. J. Dittus, 2016: Temperature and precipitation extremes in century-long gridded observations, reanalyses, and atmospheric model simulations. *J. Geophys. Res. Atmos.*, **121**, 11 174–11 189, <https://doi.org/10.1002/2016JD025480>.
- Dunn, R. J. H., M. G. Donat, and L. V. Alexander, 2014: Investigating uncertainties in global gridded datasets of climate extremes. *Climate Past*, **10**, 2171–2199, <https://doi.org/10.5194/cp-10-2171-2014>.
- Gelaro, R., and Coauthors, 2017: The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *J. Climate*, **30**, 5419–5454, <https://doi.org/10.1175/JCLI-D-16-0758.1>.
- Gilleland, E., and R. W. Katz, 2011: New software to analyze how extremes change over time. *Eos, Trans. Amer. Geophys. Union*, **92**, 13–20, <https://doi.org/10.1029/2011EO020001>.
- Hansen, J., M. Sato, and R. Ruedy, 2012: Perception of climate change. *Proc. Natl. Acad. Sci. USA*, **109**, E2415–E2423, <https://doi.org/10.1073/pnas.1205276109>.
- Huntingford, C., P. D. Jones, V. N. Livina, T. M. Lenton, and P. M. Cox, 2013: No increase in global temperature variability despite changing regional patterns. *Nature*, **500**, 327–330, <https://doi.org/10.1038/nature12310>.
- IPCC, 2012: Summary for policymakers. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*, C. B. Field et al., Eds., Cambridge University Press, 3–21.
- Jones, P. W., 1999: First- and second-order conservative remapping schemes for grids in spherical coordinates. *Mon. Wea. Rev.*, **127**, 2204–2210, [https://doi.org/10.1175/1520-0493\(1999\)127<2204:FASOCR>2.0.CO;2](https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2).
- Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, [https://doi.org/10.1175/1520-0477\(1996\)077<0437:TNYRP>2.0.CO;2](https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2).
- Kanamitsu, M., W. Abisuzaki, J. Woollen, S.-K. Yang, J. J. Hnilo, M. Fiorino, and G. L. Potter, 2002: NCEP–DOE AMIP-II Reanalysis (R-2). *Bull. Amer. Meteor. Soc.*, **83**, 1631–1643, <https://doi.org/10.1175/BAMS-83-11-1631>.
- Katz, R. W., and B. Brown, 1992: Extreme events in a changing climate: Variability is more important than averages. *Climatic Change*, **21**, 289–302, <https://doi.org/10.1007/BF00139728>.
- , P. F. Craigmile, P. Guttorp, M. Haran, B. Sansó, and M. L. Stein, 2013: Uncertainty analysis in climate change assessments. *Nat. Climate Change*, **3**, 769–771, <https://doi.org/10.1038/nclimate1980>.
- Kendall, M. G., 1975: *Rank Correlation Methods*. Charles Griffin, 272 pp.
- Kharin, V. V., and F. W. Zwiers, 2005: Estimating extremes in transient climate change simulations. *J. Climate*, **18**, 1156–1173, <https://doi.org/10.1175/JCLI3320.1>.
- , —, X. Zhang, and G. C. Hegerl, 2007: Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. *J. Climate*, **20**, 1419–1444, <https://doi.org/10.1175/JCLI4066.1>.
- , —, —, and M. Wehner, 2013: Changes in temperature and precipitation extremes in the CMIP5 ensemble. *Climatic Change*, **119**, 345–357, <https://doi.org/10.1007/s10584-013-0705-8>.
- Kistler, R., and Coauthors, 2001: The NCEP–NCAR 50-Year Reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteor. Soc.*, **82**, 247–267, [https://doi.org/10.1175/1520-0477\(2001\)082<0247:TNNYRM>2.3.CO;2](https://doi.org/10.1175/1520-0477(2001)082<0247:TNNYRM>2.3.CO;2).
- Klein Tank, A. M. G., F. W. Zwiers, and X. Zhang, 2009: Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation. WMO, WCDMP-72, WMO-TD 1500, 56 pp., www.ecad.eu/documents/WCDMP_72_TD_1500_en_1.pdf.
- Kobayashi, S., and Coauthors, 2015: The JRA-55 reanalysis: General specifications and basic characteristics. *J. Meteor. Soc. Japan*, **93**, 5–48, <https://doi.org/10.2151/jmsj.2015-001>.
- Loikith, P. C., and Coauthors, 2015: Surface temperature probability distributions in the NARCCAP hindcast experiment: Evaluation methodology, metrics, and results. *J. Climate*, **28**, 978–997, <https://doi.org/10.1175/JCLI-D-13-00457.1>.
- McKinnon, K. A., A. Rhines, M. P. Tingley, and P. Huybers, 2016: The changing shape of Northern Hemisphere summer temperature distributions. *J. Geophys. Res. Atmos.*, **121**, 8849–8868, <https://doi.org/10.1002/2016JD025292>.
- Mearns, L. O., R. W. Katz, and S. H. Schneider, 1984: Extreme high temperature events: Changes in their probabilities with changes in mean temperature. *J. Climate Appl. Meteor.*, **23**,

- 1601–1613, [https://doi.org/10.1175/1520-0450\(1984\)023<1601:EHTECI>2.0.CO;2](https://doi.org/10.1175/1520-0450(1984)023<1601:EHTECI>2.0.CO;2).
- NCAR, 2016: The climate data guide: NCEP Reanalysis (R2). Accessed 26 October 2016, <https://climatedataguide.ucar.edu/climate-data/ncep-reanalysis-r2>.
- Rahmstorf, S., and D. Coumou, 2011: Increase in extreme events in a warming world. *Proc. Natl. Acad. Sci. USA*, **108**, 17 905–17 909, <https://doi.org/10.1073/pnas.1101766108>.
- Rhines, A., and P. Huybers, 2013: Frequent summer temperature extremes reflect changes in the mean, not the variance. *Proc. Natl. Acad. Sci. USA*, **110**, E546, <https://doi.org/10.1073/pnas.1218748110>.
- , K. A. McKinnon, M. P. Tingley, and P. Huybers, 2017: Seasonally resolved distributional trends of North American temperatures show contraction of winter variability. *J. Climate*, **30**, 1139–1157, <https://doi.org/10.1175/JCLI-D-16-0363.1>.
- Robeson, S. M., 2004: Trends in time-varying percentiles of daily minimum and maximum temperature over North America. *Geophys. Res. Lett.*, **31**, L04203, <https://doi.org/10.1029/2003GL019019>.
- Sánchez-Lugo, A., C. Morice, and P. Berrisford, 2016: Surface temperature [in “State of the Climate in 2015”]. *Bull. Amer. Meteor. Soc.*, **97** (8), S12–S13.
- Sardeshmukh, P. D., G. P. Compo, and C. Penland, 2015: Need for caution in interpreting extreme weather statistics. *J. Climate*, **28**, 9166–9187, <https://doi.org/10.1175/JCLI-D-15-0020.1>.
- Screen, J. A., 2014: Arctic amplification decreases temperature variance in northern mid- to high-latitudes. *Nat. Climate Change*, **4**, 577–582, <https://doi.org/10.1038/nclimate2268>.
- Sen, P. K., 1968: Estimates of the regression coefficient based on Kendall's tau. *J. Amer. Stat. Assoc.*, **63**, 1379–1389, <https://doi.org/10.1080/01621459.1968.10480934>.
- Seneviratne, S. I., M. G. Donat, A. J. Pitman, R. Knutti, and R. L. Wilby, 2016: Allowable CO₂ emissions based on regional and impact-related climate targets. *Nature*, **529**, 477–483, <https://doi.org/10.1038/nature16542>.
- Shepard, D., 1968: A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the 1968 23rd ACM National Conference, ACM Publ. P-68, Brandon/Systems Press, 517–524.
- Sillmann, J., V. V. Kharin, X. Zhang, F. W. Zwiers, and D. Bronaugh, 2013: Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *J. Geophys. Res. Atmos.*, **118**, 1716–1733, <https://doi.org/10.1002/jgrd.50203>.
- Simmons, A. J., and Coauthors, 2004: Comparison of trends and low-frequency variability in CRU, ERA-40, and NCEP/NCAR analysis of surface air temperature. *J. Geophys. Res.*, **109**, D24115, <https://doi.org/10.1029/2004JD005306>.
- , P. Berrisford, D. P. Dee, H. Hersbach, S. Hirahara, and J.-N. Thépaut, 2017: A reassessment of temperature variations and trends from global reanalyses and monthly surface climatological datasets. *Quart. J. Roy. Meteor. Soc.*, **143**, 101–119, <https://doi.org/10.1002/qj.2949>.
- Simolo, C., M. Brunetti, M. Maugeri, T. Nanni, and A. Speranza, 2010: Understanding climate change-induced variations in daily temperature distributions over Italy. *J. Geophys. Res.*, **115**, D22110, <https://doi.org/10.1029/2010JD014088>.
- Sutton, R., E. Suckling, and E. Hawkins, 2016: What does global mean temperature tell us about local climate? *Philos. Trans. Roy. Soc.*, **373A**, 20140426, <https://doi.org/10.1098/rsta.2014.0426>.
- Tingley, M., and P. Huybers, 2013: Recent temperature extremes at high northern latitudes unprecedented in the past 600 years. *Nature*, **496**, 201–205, <https://doi.org/10.1038/nature11969>.
- Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961–3012, <https://doi.org/10.1256/qj.04.176>.
- Zwiers, F. W., and H. von Storch, 1995: Taking serial correlation into account in tests of the mean. *J. Climate*, **8**, 336–351, [https://doi.org/10.1175/1520-0442\(1995\)008<0336:TSCIAI>2.0.CO;2](https://doi.org/10.1175/1520-0442(1995)008<0336:TSCIAI>2.0.CO;2).