Quantifying changes in climate variability and extremes: Pitfalls and their overcoming

Postprint version

Sippel, S., Zscheischler, J., Heimann, M., Otto, F. E., Peters, J., Mahecha, M. D.

Published in: **Geophysical Research Letters**

Reference: Sippel, S., Zscheischler, J., Heimann, M., Otto, F. E., Peters, J., Mahecha, M. D. (2015). Quantifying changes in climate variability and extremes: Pitfalls and their overcoming. Geophysical Research Letters, 42(22), 9990-9998. doi:10.1002/2015GL066307

Web link: http://onlinelibrary.wiley.com/wol1/doi/10.1002/2015GL066307/abstract

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 640176

Quantifying changes in climate variability and extremes: pitfalls and their overcoming

Sebastian Sippel,[1,2] Jakob Zscheischler,[1,2] Martin Heimann,[1] Friederike E.L. Otto,[3] Jonas Peters,[4] and Miguel D. Mahecha[1,5]

[1]Max Planck Institute for

Hot temperature extremes have increased substantially in frequency and magnitude over past decades. A widely used approach to quantify this phenomenon is standardizing temperature data relative to the local mean and variability of a reference period. Here we demonstrate that this conventional procedure leads to exaggerated estimates of increasing temperature variability and extremes. For example, the occurrence of '2-sigma extremes' would be overestimated by 48.2% compared to a given reference period of 30 years

Biogeochemistry, Hans-Knöll-Str. 10, 07745
Jena, Germany.

[2]Institute for Atmospheric and Climate
Science, ETH Zürich, 8075 Zürich,
Switzerland.

[3]Environmental Change Institute,
University of Oxford, South Parks Road,
Oxford OX1 3QY, UK.

[4]Max Planck Institute for Intelligent
Systems, Spemannstr. 38, 72076 Tübingen,
Germany.

[5]German Centre for Integrative
Biodiversity Research (iDiv),
Halle-Jena-Leipzig, 04103 Leipzig, Germany.

with time-invariant simulated Gaussian data. This corresponds to an increase from a 2.0% to 2.9% probability of such events. We derive an analytical correction revealing that these artifacts prevail in recent studies. Our analyses lead to a revision of earlier reports [e.g. *Huntingford et al.*, 2013]: For instance we show that there is no evidence for a recent increase in normalized temperature variability. In conclusion, we provide an analytical pathway to describe changes in variability and extremes in climate observations and model simulations.

1. Introduction

Quantifying to what extent the magnitude and frequency of extreme events are changing is a priority in climate change research [*IPCC*, 2012; *Seneviratne et al.*, 2014]. In recent years, unusually hot temperature extremes have occurred and these events are increasingly exceeding the range of historical variability [*Rahmstorf and Coumou*, 2011; *Mora et al.*, 2013]. Considerable scientific debate has sparked around whether present-day changes in extreme events are due to the shifting mean climatology, or whether we are also confronted with changing variability [*Hansen et al.*, 2012; *Huntingford et al.*, 2013; *Alexander and Perkins*, 2013; *Mora et al.*, 2013; *Seneviratne et al.*, 2014]. Of particular focus in this context are changes in temperature extremes, which have direct impacts upon human wellbeing and likewise affect ecosystem services and global biogeochemical cycles [*IPCC*, 2012; *Reichstein et al.*, 2013].

A widely used approach to address this question relies on normalizing climate data relative to a reference period [*Hansen et al.*, 2012; *Coumou and Robinson*, 2013; *Huntingford et al.*, 2013; *Kamae et al.*, 2014; *Curry et al.*, 2014] aiming to objectively compare temperature variability and extremes across space and time. This approach conventionally derives standardized anomalies by locally subtracting the mean (μ_{ref}) from and dividing the observations by the standard deviation (σ_{ref}) estimated from some reference period:

$$z = \frac{X - \mu_{ref}}{\sigma_{ref}} \tag{1}$$

The idea is to rank or count events based on departures from the local climatology (as defined by the reference period) in units of standard deviation (σ). Transformations of

this kind underpin studies of changes in the occurrence of monthly or seasonal temperature extremes [*Hansen et al.*, 2012; *Coumou and Robinson*, 2013; *Kamae et al.*, 2014; *Curry et al.*, 2014] and variability [*Huntingford et al.*, 2013]. Further, so-derived standardized anomalies have been used to determine continental-scale rankings of the most significant meteorological or geophysical extreme events [*Grumm and Hart*, 2001; *Hart and Grumm*, 2001; *Root et al.*, 2007; *Graham and Grumm*, 2010], and *Kodra and Ganguly* [2014] study asymmetry in the distributions of temperature extremes using a variant of this methodology.

In this paper, we demonstrate that this conventional normalization procedure inevitably leads to erroneous and exaggerated estimates of temperature extremes and variability outside a specified 'reference period'. Furthermore, we derive an analytical correction that accounts for these statistical artifacts and allows for an accurate quantification of large-scale climate variability and extremes.

2. Methodology and Results

2.1. Normalization-induced artefacts and an analytical correction for quantifying extremes

To test the suitability of the reference-period normalization, we conduct Monte-Carlo simulations with independent and identically distributed random variables drawn from a standard Gaussian distribution ($\mathcal{N}(\mu = 0, \sigma^2 = 1)$). This numerical experiment is set-up in analogy to investigations of monthly or seasonally standardized extremes [see *Hansen et al.*, 2012, for an example] in gridded temperature data with $k = 10^4$ time series ('grid cells') and $n = 60$ data points per time series ('years of data'), but consisting of

purely random Gaussian variables (i.i.d.). For each time series we generate anomalies and subsequently standardize these based on the conventional procedure (Eq. 1). Both mean ($\hat{\mu}_{ref}$) and standard deviation ($\hat{\sigma}_{ref}$) are estimated from each time series' first 30 values (i.e. $n_{ref} = 30$). The number of values exceeding σ extremes are counted at each time step in the original and normalized dataset (Figure 1, grey and red line, respectively).

Given that the statistical properties of the artificial data are time-invariant, there should be no change in the number of extremes across the dataset. However, in fact we find substantial increases in the number of extreme events outside the reference period along with a reduction in extremes within the reference period (Figure 1a, R code to reproduce these results in Text S1). A quantification of 2σ extremes across all grid cells in the artificial dataset leads to a considerable increase (red line in Figure 1a) in the out-of-base period relative to the reference period of about 48.2%. Considering only the out-of-base period the number of 2σ (3σ) events would be overestimated by 29.1% (131.0%) relative to the original Gaussian data (black line in Figure 1a), which corresponds to an increase from a 2.3% (1.3‰) chance to 2.9% (3.1‰). For illustration purposes, the distributions at a random time step inside and outside the reference period across all time series is shown in Figure 1b and 1c for anomalies and standardized variables, respectively. Overall, the artificial experiment reveals potentially severe artefacts in the widely applied reference period normalization. In the following paragraphs, we reveal the consequences of this conventional normalization and derive an analytical solution for the induced artefacts.

To understand the origin of the apparent increase in extremes we have to consider that the 'true' values for mean and variability are inherently unknown, which changes Eq. 1

to:

$$z = \frac{X - \hat{\mu}_{ref}}{\hat{\sigma}_{ref}}. \quad (2)$$

The estimates of the mean ($\hat{\mu}_{ref}$) and standard deviation ($\hat{\sigma}_{ref}$) are random variables with well-known statistical properties [*Von Storch and Zwiers*, 2001], drawn from an independent sample in case of analyzing the out-of-base period [*Zhang et al.*, 2005] (see Text S2 for a detailed statistical description), and subsequently pooled in space. Consequently, the biases between both periods are induced by a combination of two effects, firstly the generation of anomalies ($X_{anom} = X - \hat{\mu}_{ref}$), and secondly the standardization ($z = \frac{X_{anom}}{\hat{\sigma}_{ref}}$) (Figure 1b,c): The generation of anomalies systematically increases (decreases) the variance across grid cells in the out-of-base (reference) period [*Tingley*, 2012], but does not affect the underlying distribution (Text S2). However, the local standardization of each time series induces qualitative changes to the (spatial) distribution (for an analytical derivation see Text S2) such that heavier tails outside the reference period are induced (Figure 1c). This qualitative difference stems from the fact that any time point in the out-of-base period follows a t-distribution with $n_{ref} - 1$ degrees of freedom (Text S2). Hence, the heavier tails generated by the conventional standardization lead to a consistent and potentially severe overestimation of extreme events in the out-of-base period (Figure 1a) for relatively short, but in practice often used, sometimes unavoidable, reference periods. However, the distribution after normalization can be derived analytically (Text S2), and hence the biases can be rectified separately both for the reference and the out-of-base periods. Specifically, instead of counting 2σ (3σ) extremes in the out-of-base period, a search for the corresponding percentile threshold in the variance-adjusted t-distribution

96 (2.12σ (3.32σ), respectively, if $n = 30$) would allow for the detection of the correct number
97 of events (Figure 2a, Figure S1 for an illustration of the correction procedure). Further,
98 it is worth noting that even with an increasing number of samples in the reference period,
99 the convergence to small biases is slow. For autocorrelated data the artefacts are even
100 more pronounced owing to a smaller effective sample size (Figure S2a and Figure S2b,
101 respectively).

102 Before applying the proposed analytical correction we have to consider that tempera-
103 tures at monthly or seasonal time scales are typically non-stationary [*Ji et al.*, 2014], i.e.
104 simulated or observed time series might contain spatially and temporarily diverse trends.
105 Using Monte-Carlo type simulations of normalized Gaussian time series with changing
106 trends and variability we find that both exerts strong influence on the magnitude of the
107 biases (Text S3). Increasing (decreasing) trends or variability in the out-of-base period
108 severely deflates (inflates) the biases for the upper tail (Figure S2a,b). These insights are
109 equally applicable to the lower tail of the distribution if the sign of the trend is reversed.
110 To assess the issue of non-stationarity in more detail, we consider trends and changes in
111 variability in the artificial dataset introduced in Figure 1. First, random linear trends are
112 added in the out-of-base period to each random Gaussian time series, where the magni-
113 tudes of the trends at the last time step are drawn randomly for each grid cell from a
114 uniform distribution in the interval $[-1 \leq \delta \leq 1]$ in units of σ (Figure 2b). Second, we
115 investigate a trend in the out-of-base period coinciding with randomly assigned changes
116 in variability ($0.8 \leq \sigma \leq 1.2$, Figure 2c).

Following the solution for stationary time series outlined above, we offer an analytical correction that allows handling of the additional artefacts induced by non-stationarities (Text S4). In essence, normalizing non-stationary data induces a non-central version of Student's t-distribution. This analytical distribution can be used to avoid normalization-induced biases entirely if changes in the trend or variability are known (Figure 2b,c). Likewise, estimating the trend and/or changes in variability largely allowes for removing the biases (Figure 2b,c). As above, σ-extremes are counted based on the biased estimate of the conventional procedure (red line), and based on the application of the suggested correction procedure using known (blue) and estimated (green) trends and changes in variability. Throughout this paper, Singular Spectrum Analysis (SSA), a non-linear spectral decomposition methodology [*Golyandina and Zhigljavsky*, 2013; *von Buttlar et al.*, 2014] is used to estimate trend components, before the analytical correction procedure based on the noncentral t-distribution is applied. Trends are extracted as 31-year and larger components using a 45-year SSA window length ($L = 45$).

2.2. Quantifying extremes in Earth observation data

In this subsection, we assess how monthly temperature extremes on land have changed over the second half of the 20th century in the Northern hemisphere up to present by applying the statistical approach outlined above. In order to avoid potential inhomogeneities related to gridded observations, we analyze the state-of-the-art Twentieth Century Reanalysis dataset [*Compo et al.*, 2011] (Version 2). The reanalysis dataset assimilates only surface pressure measurements and monthly sea surface temperatures into an atmosphere and land general circulation model [*Compo et al.*, 2011] and is hence independent from

station temperature measurements. The dataset has been specifically designed to assess climate variability and extremes statistics 'spanning the instrumental record', and has been demonstrated to reproduce the observed temperature trends and variability to a very large extent [*Compo et al.*, 2011].

In our analysis, we first interpolate the dataset to a 2° x 2° regular latitude-longitude grid, and mask ocean pixels. Second, we estimate separately for each month and grid cell the trend component, local mean and (non-detrended and detrended) standard deviation in two different reference periods (1921-1950 and 1951-1980). Thirdly, each pixel time series is normalized using both reference periods and the detrended and non-detrended σ_{ref} estimates. For each month we calculate the area affected by 2σ and 3σ extremes, using the conventional normalization approach and our correction. We use the trend estimates for our correction, but assume an approximately unchanged variance over the past decades [*Huntingford et al.*, 2013]. Lastly, we derive seasonal averages of the 'area affected by extremes' for Northern hemisphere summer (JJA, Figure 3).

Our analysis reveals that the exceedance of monthly 2σ and 3σ temperature extremes in summer has indeed increased substantially over the Northern hemisphere (Figure 3a,b for land areas in the NH outer tropics). However, the bias-adjusted time series show a consistently slower and smoother increase as compared to the conventionally applied uncorrected normalization procedure. A break point analysis using piecewise linear regression [*Toms and Lesperance*, 2003] based on our revised figures indicates that the recent rapid increase in hot summer months in the Northern hemisphere (2σ and 3σ events) started to emerge around the late 1980s or early 1990s (Figure 3b).

The magnitude of the biases and the discontinuities at the reference and out-of-base period are robust across different reference periods, and also hold if trends are subtracted before estimating local variability [*Coumou and Robinson*, 2013] (Figure S3 and Figure S4). Increases in extremes relative to local variability show a clear zonal pattern (Figure 3c) with the largest increases in the tropics and subtropics. Therefore, biases induced by the normalization are largest in areas where the trend is relatively small compared to local variability (Figure 3d). However, it is worth noting that peculiarities of the station-based observational record such as urban heat islands or local land-use changes are not accounted for in the 20th Century Reanalysis [*Parker*, 2011]. In addition, the availability of pressure observations varies through time [*Compo et al.*, 2011]. As such, the main purpose of the present analysis is to illustrate the potential biases induced by reference period standardization in spatio-temporal datasets.

2.3. Implications for large-scale assessments of variability and asymmetry

Normalization-induced biases are not only relevant for assessments of extremes, but a careful consideration of such statistical pre-processing techniques is equally important for analysis of variability and asymmetry in spatio-temporal datasets. An example is provided by a recent study that investigated whether temperature variability has changed over the second half of the 20th century on global and continental scales [*Huntingford et al.*, 2013]. The authors argue that annual temperatures in low-variance regions have become more variable over the past decades, whilst global temperature variability has remained near constant. This explanation stems from the authors' observation that normalized variability has increased more than absolute (spatial) variability (16% vs. 2% increases between

1963-1980 and 1981-1996). Using the 20th Century reanalysis dataset we reproduce the increases in the annual, global, area-weighted standard deviation (12.9% vs. 1.8% increases, when using the conventional data processing scheme [*Huntingford et al.*, 2013], Figure 4).

However, an artificial experiment in analogy to the previous subsection shows that the conventional normalization procedure changes the standard deviation of the data (Figure 4a), and in particular yields an increase in standard deviation between the reference and the out-of-base period. Therefore, we correct the conventionally normalized standard deviation of annual temperatures in the 20th Century Reanalysis dataset empirically and analytically. The former is achieved by simulating the reduction in standard deviation in artificial Gaussian data (Fig. 4a), whereas the latter is achieved by using an earlier reference period (1921-1950) and the application of our analytical correction. The empirical and analytical corrections reduce the increase in normalized variability from 12.9% to 5.6% and 6.0%, respectively (see Fig. 4b). A permutation-based significance test [*Fay and Shaw*, 2010] shows that the increases in mean corrected normalized standard deviation between both periods are not significant ($p_{empirical} = 0.147$ and $p_{analytical} = 0.110$), whereas conventional normalization yields a highly significant increase ($p_{conventional} = 0.004$). Hence, the relatively small and non-significant difference between the increases in standardized and absolute variability might indeed be due to the explanation offered previously [*Huntingford et al.*, 2013], and potentially related to major El-Niño events in the latter period [*Fedorov and Philander*, 2000]. If the periods before and after 1980 are extended to derive a larger sample, this reduces the increase in normalized vari-

ability to only 2% (1981-2006 vs. 1955-1980). Thus, based on our proposed normalization we cannot confirm that changes across low-variance regions have occurred over the past decades. Nonetheless, our results underpin that global temperature variability has not changed [*Huntingford et al.*, 2013], and additionally show that this finding holds both in absolute and normalized terms.

Finally, another recent study [*Kodra and Ganguly*, 2014] reports that asymmetry in temperature distributions of seasonal extreme values at daily time scale (both minima and maxima, i.e. the hottest and coldest day per season) is strongly increasing towards both the cold and hot tails in model projections of future climate conditions relative to a recent period. As a pre-processing step, the authors derive 'anomalies' of seasonal extremes by subtracting the mean of the recent (historical) climatology of seasonal extremes from both periods. This procedure leads to narrower distributions in the reference period and a broader distribution in the future (independent) period (see Text S2). This variance inflation in skewed extreme value distributions leads to the observed effect even in stationary time series, and should hence be interpreted with caution (Figure S5, and Text S6).

3. Outlook and Conclusion

The observation that a commonly used normalization of temperature data is inappropriate for assessing changes in variability, extremes, and asymmetry is of general validity and should also be considered in investigations of other climatological and Earth observations. The steadily growing archives of Earth observations derived from both ground based as well as satellite remote sensing data requires reconsidering conventional data an-

alytic approaches such as standardization. For instance, extremes in gridded standardized anomalies of rainfall and storms [*Grumm and Hart*, 2001; *Hart and Grumm*, 2001; *Root et al.*, 2007; *Graham and Grumm*, 2010; *Curry et al.*, 2014] have been studied using varieties of the conventional standardization procedure and are potentially distorted by the artefacts discussed in this paper. Further, our results might facilitate the interpretation of single climatic extreme events or trends that are frequently characterized in terms of standardized departure from climatology, both inside and/or outside the climatological reference period [*Schär et al.*, 2004; *Barriopedro et al.*, 2011; *Xu et al.*, 2012; *Ramos et al.*, 2014; *Cook et al.*, 2015]. Although our analytical treatment using the t-distribution is confined to distributions that can be approximated as Gaussian, we emphasize that the induction of biases in the tails due to dependent/independent estimators of location and scale are fundamental and hold indeed across a wide range of distributions. Furthermore, because temperature extremes are bounded [*Nogaj et al.*, 2006], approximations of temperature values by distributions with infinite tails (such as Gaussian and the t-distribution) might poorly estimate the most extreme temperatures. Here we offer a correction which adjusts biases in variability and extremes induced by a widely used data preprocessing approach. Alternatively, statistically more advanced but readily available tools, such as the theory of extreme values [*Katz et al.*, 2013; *Nogaj et al.*, 2006] offer complementary approaches to quantify extreme events under non-stationary conditions that are not affected by the statistical issues reported in this paper.

In conclusion, data normalization for the detection of changes in extremes or variability has to be applied with caution: otherwise there is a risk to arbitrarily inflate both extremes

and variability in the time periods under scrutiny. Our study demostrates how to avoid biases of this kind. However, our analyses do not call into question the major qualitative results that were outlined in previous studies [*Hansen et al.*, 2012; *Seneviratne et al.*, 2014]: hot temperature extremes have increased considerably on the global scale, a trend which is most likely to continue throughout the 21st century [*Coumou and Robinson*, 2013; *Sillmann et al.*, 2013].

Acknowledgments. We thank Fabian Gans, Jannis von Buttlar, Jens Schumacher, Markus Reichstein, Sonia Seneviratne, Myles Allen and two anonymous referees for helpful comments that strongly helped to improve the manuscript; and we are grateful to Silvana Schott for graphical support. This study was supported by the European Space Agency ESA support to science element CAB-LAB: Coupled Atmosphere Biosphere Virtual LABoratory and the European Commission project BACI 'Detecting changes in essential ecosystem and biodiversity properties towards a Biosphere Atmosphere Change Index' (grant ID 640176). S.S. is grateful to the 'German National Academic Foundation' (Studienstiftung des Deutschen Volkes) for support and the International Max Planck Research School for Global Biogeochemical Cycles (IMPRS-gBGC) for training. Support for the Twentieth Century Reanalysis Project dataset (`http://www.esrl.noaa.gov/psd/data/gridded/data.20thC_ReanV2.html`) is provided by the U.S. Department of Energy, Office of Science Innovative and Novel Computational Impact on Theory and Experiment (DOE INCITE) program, and Office of Biological and Environmental Research (BER), and by the National Oceanic and Atmospheric Administration Climate Program Office.

References

Alexander, L., and S. Perkins (2013), Debate heating up over changes in climate variability, *Environ. Res. Lett.*, *8*(4), 041,001.

Barriopedro, D., E. Fischer, J. Luterbacher, R. Trigo, and R. García-Herrera (2011), The hot summer of 2010: Redrawing the temperature record map of Europe, *Science*, *332*(6026), 220–224.

Compo, G. P., J. S. Whitaker, P. D. Sardeshmukh, N. Matsui, R. J. Allan, X. Yin, B. E. Gleason, R. S. Vose, G. Rutledge, P. Bessemoulin, et al. (2011), The Twentieth Century Reanalysis Project, *Q. J. R. Meteorolog. Soc.*, *137*(654), 1–28.

Cook, B., T. Ault, and J. Smerdon (2015), Unprecedented 21st century drought risk in the American Southwest and Central Plains, *Sci. Adv.*, *1*(1), e1400,082.

Coumou, D., and A. Robinson (2013), Historic and future increase in the global land area affected by monthly heat extremes, *Environ. Res. Lett.*, *8*(3), 034,018.

Curry, C., J. Sillmann, D. Bronaugh, K. Alterskjaer, J. Cole, D. Ji, B. Kravitz, J. Kristjánsson, J. Moore, H. Muri, et al. (2014), A multimodel examination of climate extremes in an idealized geoengineering experiment, *J. Geophys. Res. Atmos.*, *119*(7), 3900–3923.

Fay, M. P., and P. A. Shaw (2010), Exact and asymptotic weighted logrank tests for interval censored data: the interval R package, *J. Stat. Softw.*, *36*(2).

Fedorov, A. V., and S. G. Philander (2000), Is El Niño changing?, *Science*, *288*(5473), 1997–2002.

Golyandina, N., and A. Zhigljavsky (2013), *Singular Spectrum Analysis for time series*, Springer.

Graham, R. A., and R. H. Grumm (2010), Utilizing normalized anomalies to assess synoptic-scale weather events in the Western United States, *Weather Forecasting*, *25*(2), 428–445.

Grumm, R. H., and R. Hart (2001), Standardized anomalies applied to significant cold season weather events: Preliminary findings, *Weather Forecasting*, *16*(6), 736–754.

Hansen, J., M. Sato, and R. Ruedy (2012), Perception of climate change, *Proc. Natl Acad. Sci. USA*, *109*(37), E2415–E2423.

Hart, R. E., and R. H. Grumm (2001), Using normalized climatological anomalies to rank synoptic-scale events objectively, *Mon. Weather Rev.*, *129*(9), 2426–2442.

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*(7462), 327–331.

IPCC (2012), Summary for policymakers, in *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*, edited by C. B. Field, V. Barros, T. Stocker, Q. Dahe, D. J. Dokken, K. L. Ebi, M. D. Mastrandrea, K. J. Mach, G. K. Plattner, S. K. Allen, M. Tignor, and P. M. Midgley, Cambridge University Press.

Ji, F., Z. Wu, J. Huang, and E. P. Chassignet (2014), Evolution of land surface air temperature trend, *Nature Clim. Change*.

Kamae, Y., H. Shiogama, M. Watanabe, and M. Kimoto (2014), Attributing the increase in Northern Hemisphere hot summers since the late 20th century, *Geophys. Res. Lett.*, *41*(14), 5192–5199.

Katz, R. W., P. F. Craigmile, P. Guttorp, M. Haran, B. Sanso, and M. L. Stein (2013), Uncertainty analysis in climate change assessments, *Nature Clim. Change*, *3*(9), 769–771.

Kodra, E., and A. R. Ganguly (2014), Asymmetry of projected increases in extreme temperature distributions, *Sci. Rep.*, *4*, 5884.

Mora, C., A. G. Frazier, R. J. Longman, R. S. Dacks, M. M. Walton, E. J. Tong, J. J. Sanchez, L. R. Kaiser, Y. O. Stender, J. M. Anderson, et al. (2013), The projected timing of climate departure from recent variability, *Nature*, *502*(7470), 183–187.

Nogaj, M., P. Yiou, S. Parey, F. Malek, and P. Naveau (2006), Amplitude and frequency of temperature extremes over the north atlantic region, *Geophys. Res. Lett.*, *33*(10).

Parker, D. E. (2011), Recent land surface air temperature trends assessed using the 20th Century Reanalysis, *J. Geophys. Res. Atmos.*, *116*(D20).

Rahmstorf, S., and D. Coumou (2011), Increase of extreme events in a warming world, *Proc. Natl Acad. Sci. USA*, *108*(44), 17,905–17,909.

Ramos, A. M., R. M. Trigo, and M. L. R. Liberato (2014), A ranking of high-resolution daily precipitation extreme events for the Iberian Peninsula, *Atmos. Sci. Lett.*, *15*(4), 328–334.

Reichstein, M., M. Bahn, P. Ciais, D. Frank, M. D. Mahecha, S. I. Seneviratne, J. Zscheischler, C. Beer, N. Buchmann, D. C. Frank, D. Papale, A. Rammig, P. Smith, K. Thon-

330 icke, M. van der Velde, S. Vicca, A. Walz, and M. Wattenbach (2013), Climate extremes
331 and the carbon cycle, *Nature*, *500*(7462), 287–295.

332 Root, B., P. Knight, G. Young, S. Greybush, R. Grumm, R. Holmes, and J. Ross (2007), A
333 fingerprinting technique for major weather events, *J. Appl. Meteorol.*, *46*(7), 1053–1066.

334 Schär, C., P. Vidale, D. Lüthi, C. Frei, C. Häberli, M. Liniger, and C. Appenzeller (2004),
335 The role of increasing temperature variability in European summer heatwaves, *Nature*,
336 *427*(6972), 332–336.

337 Seneviratne, S. I., M. G. Donat, B. Mueller, and L. V. Alexander (2014), No pause in the
338 increase of hot temperature extremes, *Nature Clim. Change*, *4*(3), 161–163.

339 Sillmann, J., V. Kharin, F. Zwiers, X. Zhang, and D. Bronaugh (2013), Climate extremes
340 indices in the CMIP5 multimodel ensemble: Part 2. future climate projections, *J. Geo-*
341 *phys. Res. Atmos.*, *118*(6), 2473–2493.

342 Tingley, M. P. (2012), A Bayesian ANOVA scheme for calculating climate anomalies, with
343 applications to the instrumental temperature record, *J. Clim.*, *25*(2), 777–791.

344 Toms, J., and M. Lesperance (2003), Piecewise regression: a tool for identifying ecological
345 thresholds, *Ecology*, *84*(8), 2034–2041.

346 von Buttlar, J., J. Zscheischler, and M. Mahecha (2014), An extended approach for spa-
347 tiotemporal gapfilling: dealing with large and systematic gaps in geoscientific datasets,
348 *Nonlinear Processes Geophys.*, *21*, 203–215.

349 Von Storch, H., and F. W. Zwiers (2001), *Statistical Analysis in Climate Research*, Cam-
350 bridge University Press.

Xu, X. T., S. L. Piao, X. H. Wang, A. P. Chen, P. Ciais, and R. B. Myneni (2012), Spatio-temporal patterns of the area experiencing negative vegetation growth anomalies in china over the last three decades, *Environ. Res. Lett.*, *7*(3).

Zhang, X. B., G. Hegerl, F. W. Zwiers, and J. Kenyon (2005), Avoiding inhomogeneity in percentile-based indices of temperature extremes, *J. Clim.*, *18*(11), 1641–1651.

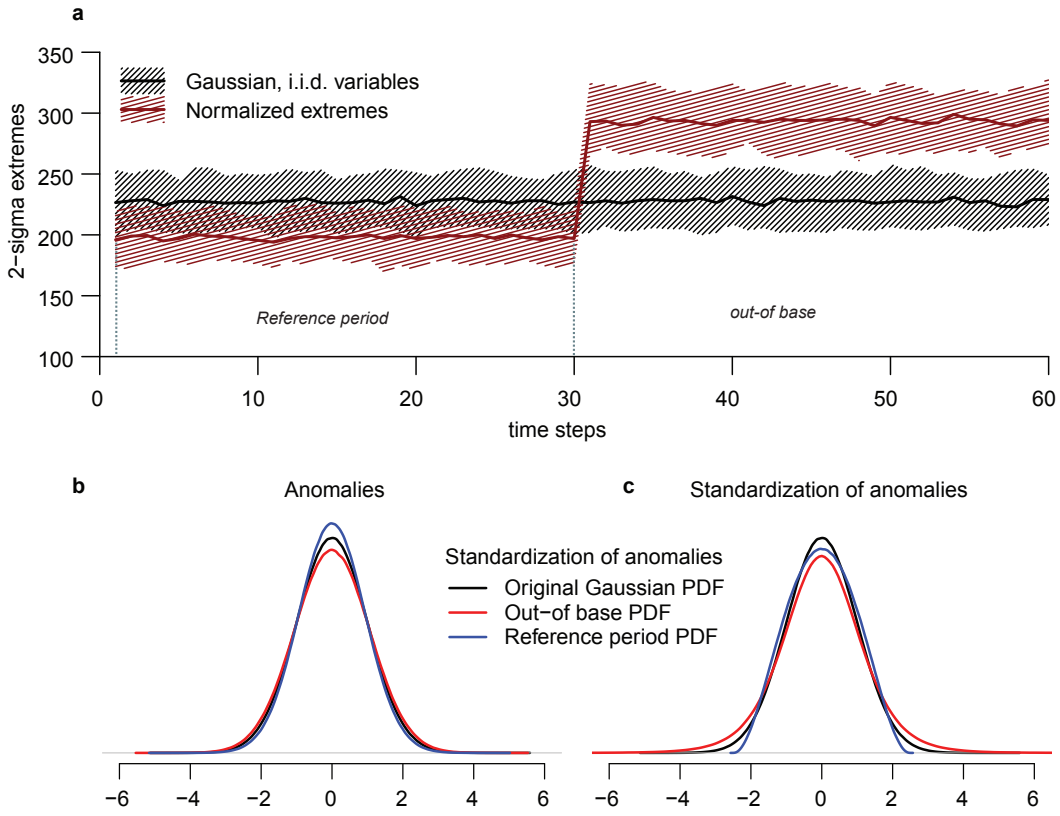

Figure 1. Biases in the detection of extreme events in stationary and independent Gaussian data induced by normalization. a) Occurrences of positive 2-sigma extremes in artificial Gaussian time series based on 10,000 replicates over 60 time-points before normalizing the data (black line), and after normalizing each replicate using the first 30 samples as reference period. b) Illustration of variance inflation and reduction through the generation of anomalies in the out-of-base (blue) vs. reference period (red) PDF ($n_{ref} = 8$ for illustration). c) Changing tails in normalized (i.e., divided by the SD estimate) Gaussian variables ($n_{ref} = 8$ for illustration). Coloured shading in (a) indicates the 5th to 95th percentile in repeated simulations.

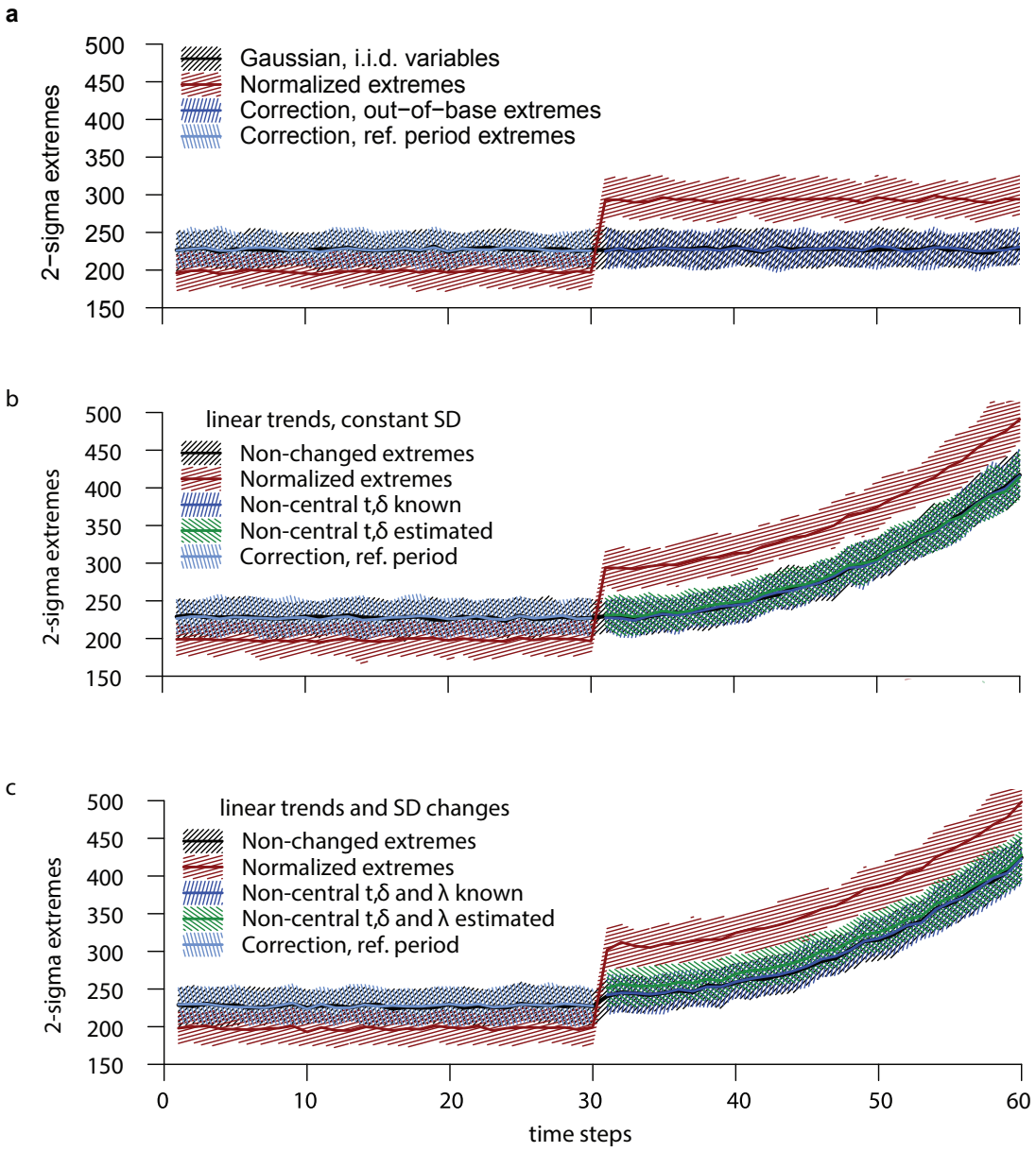

Figure 2. Correction of normalization-induced biases in stationary and non-stationary time series consisting of independent random variables. Detecting 2-sigma extreme events in a) Stationary Gaussian time series, b) Gaussian time series with random linear trends added in the out-of-base period ($-1 < \delta_{t=60} < 1$, in units of σ), c) Gaussian time series with random linear trends ($-1 < \delta_{t=60} < 1$, in units of σ) and changing variance ($0.8\sigma_{ref} < \lambda\sigma_{ref} < 1.2\sigma_{ref}$) in the out-of-base period. In each panel, coloured shading indicates the 5th to 95th percentile in repeated simulations ($k = 10^4$ simulated time series in all panels).

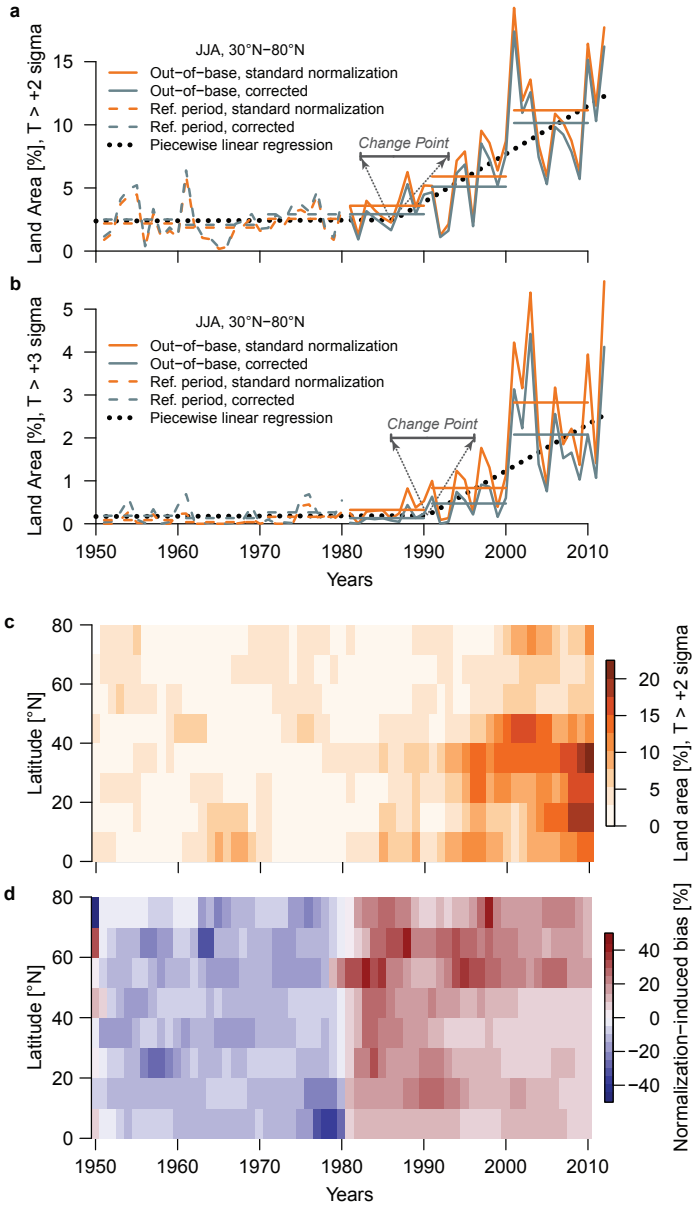

Figure 3. Increase in normalized hot temperature extremes in a spatio-temporal dataset (20th Century Reanalysis [*Compo et al.*, 2011]). a,b) Time series of fraction of extratropical Northern hemisphere land area covered by positive monthly 2σ (a) and 3σ (b) events in summer (reference period: 1951-1980). Horizontal lines indicate decadal averages for the conventional normalization procedure (light blue) and our proposed correction (orange). c) Zonal evolution of fraction of land area covered by monthly positive 2σ extremes in Northern hemisphere summer. d) Zonal evolution of relative biases induced by the conventional normalization approach.

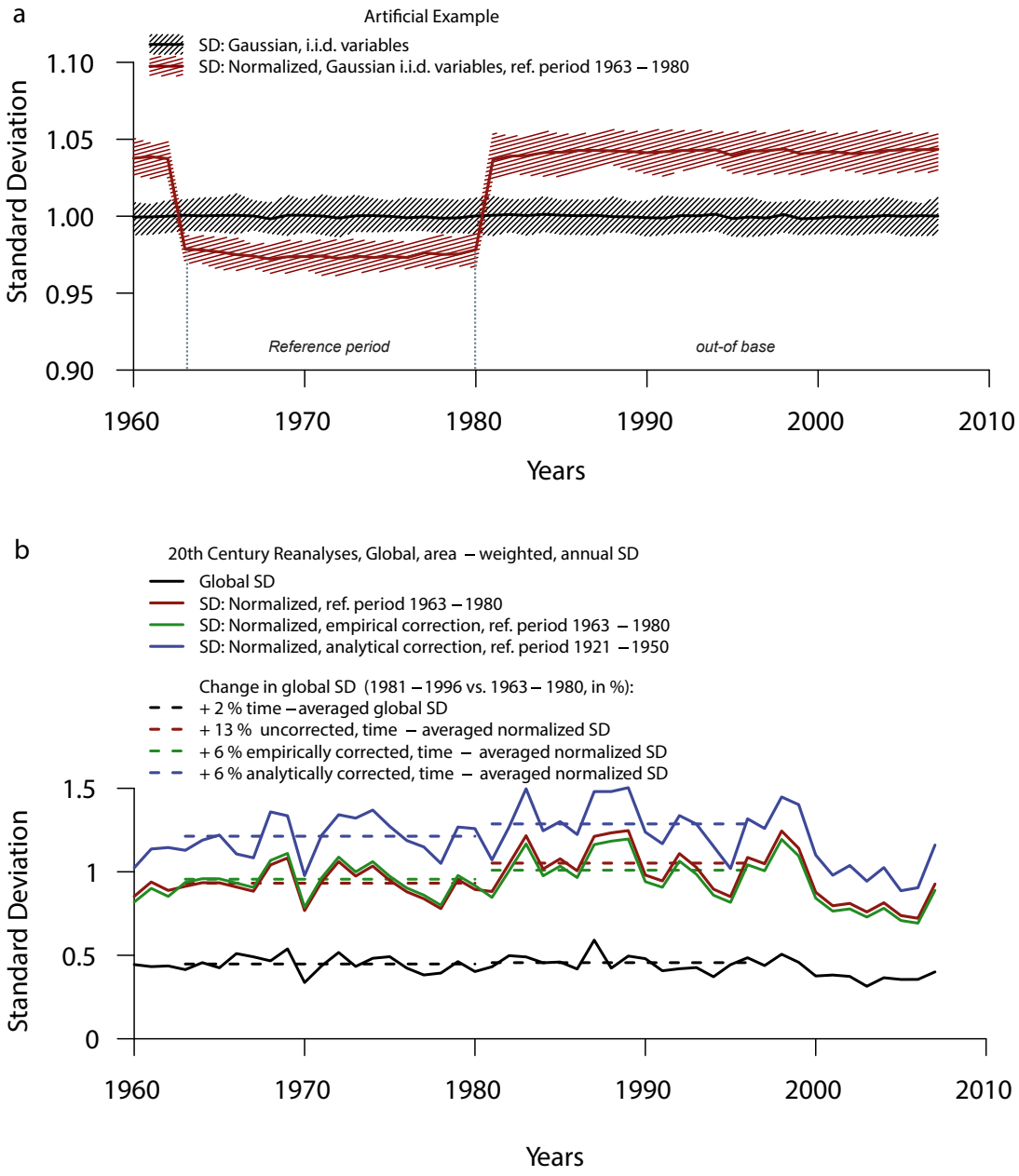

Figure 4. Normalization-induced changes in variability. a,b) Time series of normalized variability following the data processing scheme of *Huntingford et al.* [2013] in an artificial example ($k = 10^4$ time series) with i.i.d. Gaussian variables (a) and in the 20th Century Reanalysis dataset (b).

Supporting Information for "Quantifying changes in climate variability and extremes: pitfalls and their overcoming"

Sebastian Sippel,[1,2] Jakob Zscheischler,[1,2] Martin Heimann,[1] Friederike E.L. Otto,[3] Jonas Peters,[4] and Miguel D. Mahecha[1,5]

Corresponding author: S. Sippel, Max Planck Institute for Biogeochemistry, 07745 Jena, Germany. (ssippel@bgc-jena.mpg.de)

[1]Max Planck Institute for Biogeochemistry, Hans-Knöll-Str. 10, 07745 Jena, Germany.

[2]Institute for Atmospheric and Climate Science, ETH Zürich, 8075 Zürich, Switzerland.

[3]Environmental Change Institute, University of Oxford, South Parks Road, Oxford OX1 3QY, UK.

[4]Max Planck Institute for Intelligent Systems, Spemannstr. 38, 72076 Tübingen, Germany.

Contents of this file

1. Text S1 to S6

2. Figures S1 to S5

[5]German Centre for Integrative

Biodiversity Research (iDiv),

Halle-Jena-Leipzig, 04103 Leipzig, Germany.

Text S1. Guide to the artificial normalization example

We provide the original source code that was used to carry out the artificial normalization example shown in Figure 1 in a step-by-step guide using the R Statistical Programming Environment [*R Core Team*, 2013]. We first generate an artificial dataset containing 10,000 time series, where each time series consists of $n = 60$ independent and identically distributed Gaussian variables. As stated in the main text, this can be understood as an analogy to a spatio-temporal temperature dataset that comprises 60 years of data across 10,000 geographical grid cells. Subsequently, each time series is centered and scaled with estimates of the mean and standard deviation as derived from a reference period of length $n_{ref} = 30$ (here, the first 30 values of each time series are chosen). For each time point t, we then count the number of σ-extremes in the original Gaussian data and the normalized data (Figure 1 in the main paper). Lastly, the proposed correction (for a formal derivation see Text S2) leads to the corrected normalized time series shown in Figure 2a.

```
# Define parameters for normalization example:
nref = 30;          # Length of reference period
ngridcells = 10000; # Number of independent grid cells
sigma = 2;          # Sigma threshold
```

```
# Generate Gaussian time series each of which consists of 60 values:
data.orig = sapply(1:ngridcells, FUN=function(x) rnorm(60));
# Estimate the mean and standard deviation of each time series
# based on the reference period (first 30 values):
mean.estimate = sapply(1:ngridcells, FUN=function(x) mean(data.orig[1:nref,x]));
sd.estimate = sapply(1:ngridcells, FUN=function(x) sd(data.orig[1:nref,x]));
# Generate anomalies, and normalize each time series with its sample mean
```

```
# and sample standard deviation:
data.anom = sapply(1:ngridcells, FUN=function(x) data.orig[,x]-mean.estimate[x]);
data.norm = sapply(1:ngridcells, FUN=function(x) data.anom[,x]/sd.estimate[x]);
```

```
# count +2sigma events throughout each time series and at each time step, for the
# original and normalized data:
data.orig.2sigma.extremes = apply(X=data.orig, 1, function(x) length(which(x > 2)));
data.norm.2sigma.extremes = apply(X=data.norm, 1, function(x) length(which(x > 2)));
# Compute the corrected number of sigma extremes:
# Out-of-base period:
data.norm.2sigma.extremes.obase.cor = apply(X=data.norm, MARGIN=c(1),
FUN=function(x) length(which((x / sqrt(1+1/nref)) > qt(pnorm(sigma), df=nref-1))));
# Reference period:
data.norm.2sigma.extremes.ibase.cor = apply(X=data.norm, MARGIN=c(1),
FUN=function(x) length(which(((x*x)*nref/((nref-1)*(nref-1))) > qbeta(pnorm(sigma),
shape1 = 0.5, shape2 = nref/2-1))));
```

```
# Plot the number of sigma extremes:
plot(data.norm.2sigma.extremes, col='darkred', pch=8)
points(data.orig.2sigma.extremes, col='black', pch=8)
points(x = 1:nref, data.norm.2sigma.extremes.ibase.cor[1:nref], col='darkblue',
pch=8)
points(x = c(1:60)[-(1:nref)], data.norm.2sigma.extremes.obase.cor[-(1:nref)],
col='darkgreen', pch=8)
legend('topleft', c('Conventional normalization', 'i.i.d. Gaussian variables',
 'Normalization + correction, reference period', 'Normalization + correction,
 out-of-base'), col=c('darkred', 'black', 'darkblue', 'darkgreen'), pch=8)
```

Text S2. Normalization-induced changes to stationary and independent Gaussian time series

At any grid cell i, time series of the form $X_{t,i}; t = 1, ..., n; i = 1, ..., k$ are normalized to yield standardized 'z-scores' with respect to a defined reference period as a subset of the full record:

$$z_{t,i} = \frac{X_{t,i} - \hat{\mu}_{ref,i}}{\hat{\sigma}_{ref,i}} \quad . \tag{1}$$

In this example, each sample in each time series $X_{t,i}$ is drawn independently from a Gaussian distribution with the expected value $E[X_{t,i}] = \mu_i$ and the variance given by $Var(X_{t,i}) = \sigma_i^2$. Thus, the estimators $\hat{\mu}_i$ for the mean μ_i and the estimator $\hat{\sigma}_i^2$ for the variance σ_i^2 satisfy [*Von Storch and Zwiers*, 2001] in each grid cell

$$\hat{\mu}_i = \frac{1}{n}\sum_{t=1}^{n} X_{t,i} \sim \mathcal{N}(\mu_i, \frac{\sigma_i^2}{n}) \quad \text{and} \tag{2}$$

$$\hat{\sigma}_i^2 = \frac{1}{n-1}\sum_{t=1}^{n}(X_{t,i} - \hat{\mu}_i)^2 \sim \sigma_i^2 \chi_{n-1}^2 \frac{1}{n-1} \quad . \tag{3}$$

Hence, the collection of sample means $\hat{\mu}_{ref,i}$ follows a normal distribution with expected value $E[\hat{\mu}_{ref,i}] = \mu_i$ and variance $Var(\hat{\mu}_{ref,i}) = \frac{\sigma_i^2}{n_{ref}}$ (Eq. 2) across grid cells. Here we show that this widely used normalization approach changes the statistical properties of the distribution across grid cells. This extends an issue previously discussed [*Zhang et al.*, 2005], but here we are not confined to percentile-based estimates of temperature extremes. In the following subsections we distinguish normalization in the reference period (where

the estimators are dependent on the samples) from the normalization in the out-of-base period, where the estimators are independent from the samples.

In the following sections we consider each grid cell independently. In order to improve readability, we therefore omit the index i for the grid cells and simply write X_t.

Text S2a. Normalization in the out-of-base period

At any time t in the (independent) out-of-base period, the anomalies are given by the random variable

$$X_{anom,t} = X_t - \hat{\mu}_{ref} \quad , \tag{4}$$

with different realizations across grid cells. Consequently, anomalies that are generated by subtracting the reference period (that is, independent) sample mean follow again a Gaussian distribution, because the difference between two Gaussian variables $X = X_1 - X_2$ is Gaussian distributed [*Johnson et al.*, 1994] with $\mu = \mu_1 - \mu_2$ and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2$, i.e.,

$$X_{anom,t} \sim \mathcal{N}(0, \sigma^2(1 + \frac{1}{n_{ref}})) \quad . \tag{5}$$

Please note that the increase in variance caused by deriving anomalies and implied by Eq. 5 holds for any distribution with finite variances, i.e. not only Gaussian distributions.

Dividing anomalies by the estimated standard deviation ('standardizing') yields standardized 'z-scores':

$$z_t = \frac{X_{anom,t}}{\hat{\sigma}_{ref}} \quad . \tag{6}$$

Following Eq. 3, the 'z-scores' are characterized by Student's t-distribution with $\nu = n-1$ degrees of freedom (cf. the definition of the t-distribution [*Fisher*, 1925]), which is scaled by the variance inflation given in Eq. 5:

$$z_t \sim \sqrt{1 + \frac{1}{n_{ref}}} \cdot t(n_{ref} - 1) \quad . \tag{7}$$

Hence, after normalization, we expect the grid cell values at any given time step t in the out-of-base period to follow a scaled t-distribution (Eq. 7), rather than the Gaussian distribution as implied in earlier reports [*Hansen et al.*, 2012; *Coumou and Robinson*, 2013]. Although the t-distribution converges against the Gaussian distribution for a large number of degrees of freedom (i.e. increasing n_{ref}, see Figure 1 and Figure S1), its tails are considerably heavier even for a relatively large number of degrees of freedom. This well-known distribution allows us to derive a correction based on quantiles for normalized z-scores that can be constructed by adjusting the 'σ-extreme' of interest using Eq. 7 (see Figure S1 for an illustration). For example, the probability of a 2σ-extreme in a Gaussian distribution corresponds to a 2.12σ event in the scaled t-distribution (for $n_{ref} = 30$, Section S1).

Text S2b. Normalization in the reference period

In the reference period, the estimators of mean and variance are not independent from the samples. This fact causes the underestimation of extremes in the reference period, as illustrated for instance in Figure 1 in the main paper. In this subsection, we first discuss the changes induced to the distribution by deriving anomalies (i.e. Eq. 4), and secondly demonstrate how changes induced by normalization according to Eq. 6 in the reference period can be analytically corrected.

The generation of anomalies in the reference period in analogy to Eq. 4 reduces the variability across grid cells to $Var(X_{anom,t}) = \sigma^2(1 - \frac{1}{n_{ref}})$. Note that this result does not only hold for the Gaussian distribution but for any distribution with finite second moments:

$$\begin{aligned} Var(X_{anom,t}) &= Var(X_t - \hat{\mu}_{ref}(X)) \\ &= Var(X_t) - 2Cov(X_t, \hat{\mu}_{ref}(X_t)) + \frac{Var(X_t)}{n_{ref}} \\ &= Var(X_t) - 2\frac{1}{n_{ref}} \sum_{s=1}^{n_{ref}} Cov(X_t, X_s) + \frac{Var(X_t)}{n_{ref}} \\ &= \sigma^2 - 2\frac{\sigma^2}{n_{ref}} + \frac{\sigma^2}{n_{ref}} \\ &= \sigma^2(1 - \frac{1}{n_{ref}}) \quad . \end{aligned}$$

A subsequent standardization of anomalies following Eq. 6 in the reference period changes the sample distribution across grid cells qualitatively to a non-Gaussian distribution. The resulting distribution follows a beta-distribution [*Thompson*, 1935; *Johnson et al.*, 1995]

$$(\frac{X_{anom,t}}{\hat{\sigma}_{ref}})^2 \sim n_{ref} Beta(0.5, \frac{n_{ref} - 1}{2}) \quad . \tag{8}$$

Alternatively, the distribution of standardized anomalies within the reference period has been described as a '*tau*-distribution' [*Thompson*, 1935], where τ is defined as $\tau = \frac{X_{anom,t}}{\hat{\sigma}_{ref}}$. Here, *tau* is related to a t-distribution with $\nu = n_{ref} - 2$ degrees of freedom by $\tau = t_\nu \sqrt{\frac{n_{ref}-1}{n_{ref}-2+t_\nu^2}}$. Similarly to above, the transformation given by Eq. 8 can be used to adjust the detection of normalized extremes within the reference period by quantile adjustments (see Figure S1). From the quantile-quantile plots shown in Figure S1

it becomes obvious that a normalization across time-invariant Gaussian data yields an underestimation of extremes in the reference period (a), and an overestimation in the out-of-base (independent) period (b).

Text S3. Monte Carlo simulations

In order to test how specific features that are present in climatic data might affect the biases in normalized tails in the detection of spatially aggregated extremes, we conduct a variety of Monte-Carlo type simulations.

Each simulation is set up as follows:

• Generate $k = 10^5$ time series, each of which with $n = 130$ data points, drawn independently from a Gaussian distribution (exception: autocorrelated time series, see below).

• Define a reference period length of $n_{ref} = 30$, which has been used in climatological studies [*Hansen et al.*, 2012] (exception: experiment using a variable reference period length, see below).

• Define remaining 100 data points in each time series as the out of base period.

• Detect extremes by counting 'σ extremes' in normalized and original time series for each time step t.

• Calculate the biases in the tails as relative differences (in percent) between the conventionally normalized time series (Eq. 2 in the main text) and the original time series (i.e. without normalization).

First, we test how the length of the reference period influences biases in the tails. It can be seen from the analytical argument put forward in section S2 that the biases in the

normalized tails are a function of sample size in the reference period. To illustrate this, we vary the length of the reference period (Figure S2a). The biases are decreasing for longer reference periods. However, in practical terms relatively large sample sizes in the reference period are needed in order to detect relatively rare events with small biases if the conventional normalization scheme is used.

Second, we assess the effect of autocorrelation on the biases in the normalized tails. Autocorrelation is a feature frequently present in climatic time series [*Zwiers and von Storch*, 1995], and hence should be accounted for in statistical analyses. We simulate time series from an AR(1) process as

$$X_{AR1}(t) = \alpha X_{AR1}(t-1) + Z(t), \tag{9}$$

with white noise innovations $Z \sim \mathcal{N}(0, \tau^2)$. The model's parameter α determines the strength of the autocorrelation and is varied in the range $0 \leq \alpha \leq 0.9$. The overestimation of extremes strongly increases for autocorrelated data, which urges for caution in using a normalization scheme in such time series. The reason for the stronger overestimation compared to the standard normalization procedure is three-fold: Firstly, the variance of the sample mean of autocorrelated data [*Zieba*, 2010] is larger as compared to Eq. 2:

$$\hat{\mu}_{X_{AR1},ref} = \mathcal{N}(0, [n + 2\sum_{k=1}^{n-1}(n-k)\rho_k]\frac{\sigma^2}{n^2}), \tag{10}$$

where ρ_k denotes the autocorrelation coefficient of the AR(1) model.

Secondly, the standard variance estimator (Eq. 3) is biased for autocorrelated data [*Bayley and Hammersley*, 1946]. The construction of an unbiased variance estimator is possible [*Zieba*, 2010], but requires the autocorrelation structure to be known exactly.

Thirdly, the normalized distributions follow Student's t-distribution (as above), if the variance and mean estimates are derived from an independent sample. Hence, these three issues are causing the drastically increasing biases seen in Figure S2b for autocorrelated data.

Furthermore, trends and changing variance are common features in climatic time series [*Ji et al.*, 2014; *Huntingford et al.*, 2013; *Screen*, 2014]. We test empirically how changes in the mean or variance in the independent period are changing the detection biases in normalized extremes. To do so, we add various offsets in the range $-1 \leq \delta \leq +2[\sigma]$. Similarly, we change the variance in the out-of-base period to $0.5 \leq \sigma \leq 2$. Subsequently, the relative difference between the standard normalization scheme and the true number of extremes is calculated (Figure S2c). Our Monte-Carlo simulations reveal that normalization biases (as discussed in the main text of this paper) are not constant under changes of the mean and variance of the time series. Although an analytical treatment is possible (see Section S4), this empirical exercise allows to illustrate the sensitivity of the biases to both sign and magnitude of trends and changes in variance. Positive changes in the mean or variance are reducing the observed biases in the upper tail of the distribution, because any positive σ extreme would 'shift' towards the center of the distribution in this case. However, negative trends or changes in variance would induce the opposite effect and lead to a drastic overestimation in the upper tail. These results are equally applicable to the lower tail if the sign of the trend is reversed. We conclude that any assessment of extremes or the tails of normalized climatic data across different spatial or temporal domains needs to take potential non-stationarities into account.

Text S4. Normalization bias in non-stationary and independent time series

This section is motivated by the fact that normalization-induced biases are sensitive to trends or changes in variance (see Section S3). Here, we outline a correction method that takes such non-stationarities into account. Consider any random variable $X_{orig} \sim \mathcal{N}(\mu_{ref}, \sigma^2_{ref})$, from which $\hat{\mu}_{ref}$ and $\hat{\sigma}^2_{ref}$ are estimated. Assume that at any time t outside the reference period the mean changes to $\mu_{t,obase} = \mu_{ref} + \delta_t$ and the standard deviation changes to $\sigma_{obase} = \lambda \cdot \sigma_{ref}$.

Non-stationarity in the out-of-base period would change the Gaussian distribution to

$$X_t \sim \mathcal{N}(\mu + \delta_t, \lambda^2 \cdot \sigma^2). \tag{11}$$

The generation of anomalies for Gaussian data is given in Eq. 4 and the sample means follow Eq. 2. Put together, this yields a distribution of anomalies across grid cells given by

$$X_{anom,t} \sim \mathcal{N}(\delta_t, \sigma^2(\lambda^2 + \frac{1}{n_{ref}}). \tag{12}$$

Accordingly, and similar to Eq. 5, the spatial aggregation for the detection of extremes in the tails would result in a broader (but qualitatively unchanged) distribution. A search for non-adjusted σ extremes becomes hence inadequate.

However, the subsequent standardization of non-stationary and independent time series is more important for biases in the tails. A generalization of Student's t-distribution is the non-central t-distribution [*Johnson et al.*, 1995], which is skewed and results from Eq. 6, if $X_{anom,t}$ is replaced by a random Gaussian variable with non-zero mean [*Von Storch and Zwiers*, 2001]. Hence, a standardization of non-stationary Gaussian time series based

on Eq. 6 yields a spatial distribution of

$$\frac{X_{anom,t}}{\hat{\sigma}_{ref}} = \sqrt{\lambda^2 + \frac{1}{n_{ref}}} \cdot \frac{[\frac{X_{anom}-\delta_t}{\sqrt{\lambda^2+\frac{1}{n_{ref}}}} + \frac{\delta_t}{\sqrt{\lambda^2+\frac{1}{n_{ref}}}}]}{\hat{\sigma}_{ref}} \tag{13}$$

$$\Rightarrow z = \frac{X_{anom,t}}{\hat{\sigma}_{ref}} \sim \sqrt{\lambda^2 + \frac{1}{n_{ref}}} \cdot t'(\nu = n-1, ncp = \frac{\delta_t}{\sqrt{\lambda^2 + \frac{1}{n_{ref}}}}) \quad . \tag{14}$$

This can be seen as a centering and scaling of the enumerator in Eq. 13 to yield a unit normal variable and an additive non-centrality-parameter. Hence, the division by the estimates of the standard deviation $\hat{\sigma}_{ref}$ yields a scaled version of the non-central t-distribution (Eq. 14), implying $k = n_{ref} - 1$ degrees of freedom. Therefore, an analytical correction similar to Section S2 can be constructed if the change in location and scale outside the reference period can be estimated (see also Figure 2, main text). However, since estimates of trends or variance changes are made on relatively short time series, and because these are not independent from the estimated mean or variability, some minor biases remain (Figure 2, main text). These biases are negligible if only the mean has changed, and they are much smaller than biases in the tails induced by an uncorrected normalization procedure if variance changes are estimated as well. Nevertheless, we argue for some caution if very rare events are to be detected based on the application of a normalization transformation.

Text S5. Subtraction of trend components before computing standard deviation estimates

Several previous papers have used detrending procedures before estimating the standard deviation in a reference period [*Coumou and Robinson*, 2013; *Huntingford et al.*, 2013]. This data preprocessing step is assumed to avoid an overestimation of variability due to

potential trends in time series in the (arbitrarily chosen) reference period. Others have used the period 1951-1980 as the reference, because this period is widely assumed to be associated with largely stationary temperatures [*Hansen et al.*, 2012]. The removal of trends before computing the standard deviation of each time series reveals only very minor changes both in terms of the overall increase in extremes and the preprocessing-induced biases. We estimate trends in each time series using Singular Spectrum Analysis as described in the Methodology section of the main paper, but other methodologies are likewise applicable. Next, we standardize each time series with the standard deviation estimates computed from detrended series and reproduce Figure 3 from the main paper (Figure S3).

To test the sensitivity of the biases and extremes to the choice of reference period, we repeat the previous analysis by normalizing the data based on mean and detrended SD estimates calculated for 1921-1950 (Figure 4). Although the choice of reference period influences the absolute number of σ extremes (because 1951-1980 had been warmer than 1921-1950), the biases that are induced by the normalization procedure are still in a similar magnitude (Figure 4).

Text S6. Asymmetry in temperature distributions

Another important question to test is whether recent estimates of asymmetry [*Kodra and Ganguly*, 2014] in seasonal extreme value distributions might be affected by subtracting a 'historical climatology', estimated from each time series. For this purpose, we follow the methodology of an earlier study [*Kodra and Ganguly*, 2014] but with i.i.d. Gaussian variables:

• We generate 60 seasons with each 90 days in $k = 10,000$ time series (that is, in analogy to spatial replicates)

• For each season, we only retain the maximum value. This procedure yields a distribution that can be approximated by a Weibull type extreme value distribution [*Coles et al.*, 2001]

• Now, each time series is split into a historical and future period (first and second half of the time series, respectively)

• Following *Kodra and Ganguly* [2014], we compute the mean of the 'historical' period and subtract it from each times series.

• Subsequently, percentiles of the future and historical period are computed across all time series, and percentile-wise differences between the future and historical period are analyzed (Figure 5)

• We compare the so-derived percentile-wise changes to simply generating the differences between future and historical percentiles without the previous transformation (Figure S5a)

As shown in Section S2, this procedure invariably leads to an inflation (reduction) of the variance in the surrogate 'future' ('historical') period. Hence, the upper tail of the 'future' extreme value distribution has increased, whereas the lower tail has decreased relative to untransformed changes (see red and grey lines in Figure S5a). However, since extreme value distributions are skewed, the change in variability also explains the observation of increased assymetry, if the changes in both tails are compared (Figure S5b). This increased asymmetry is not observed if the analysis is conducted without subtracting historical means (grey line in Figure 5b). These results are shown for extreme value distributions generated by retaining the highest value in each season, but would apply equally if only seasonal minima were retained (but with reversed changes in asymmetry).

References

Bayley, G. V., and J. M. Hammersley (1946), The" effective" number of independent observations in an autocorrelated time series, *Supplement to the Journal of the Royal Statistical Society*, pp. 184–197.

Coles, S., J. Bawa, L. Trenner, and P. Dorazio (2001), *An introduction to statistical modeling of extreme values*, vol. 208, Springer.

Coumou, D., and A. Robinson (2013), Historic and future increase in the global land area affected by monthly heat extremes, *Environ. Res. Lett.*, *8*(3), 034,018.

Fisher, R. A. (1925), Applications of students distribution, *Metron*, *5*(3), 90–104.

Hansen, J., M. Sato, and R. Ruedy (2012), Perception of climate change, *Proc. Natl Acad. Sci. USA*, *109*(37), E2415–E2423.

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*(7462), 327–331.

Ji, F., Z. Wu, J. Huang, and E. P. Chassignet (2014), Evolution of land surface air temperature trend, *Nature Clim. Change*.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1994), *Continuous Univariate Distributions*, vol. 1, Wiley & Sons.

Johnson, N. L., K. S., and N. Balakrishnan (1995), *Continuous Univariate Distributions*, vol. 2, Wiley & Sons.

Kodra, E., and A. R. Ganguly (2014), Asymmetry of projected increases in extreme temperature distributions, *Sci. Rep.*, *4*, 5884.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Screen, J. A. (2014), Arctic amplification decreases temperature variance in northern mid-to high-latitudes, *Nature Clim. Change*, *4*, 577–582.

Thompson, W. R. (1935), On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation, *The Annals of Mathematical Statistics*, *6*(4), 214–219.

Von Storch, H., and F. W. Zwiers (2001), *Statistical Analysis in Climate Research*, Cambridge University Press.

Zhang, X. B., G. Hegerl, F. W. Zwiers, and J. Kenyon (2005), Avoiding inhomogeneity in percentile-based indices of temperature extremes, *J. Clim.*, *18*(11), 1641–1651.

Zieba, A. (2010), Effective number of observations and unbiased estimators of variance for autocorrelated data an - overview, *Metrology and Measurement Systems*, *17*(1), 3–16.

Zwiers, F. W., and H. von Storch (1995), Taking serial correlation into account in tests of the mean, *J. Clim.*, *8*(2), 336–351.

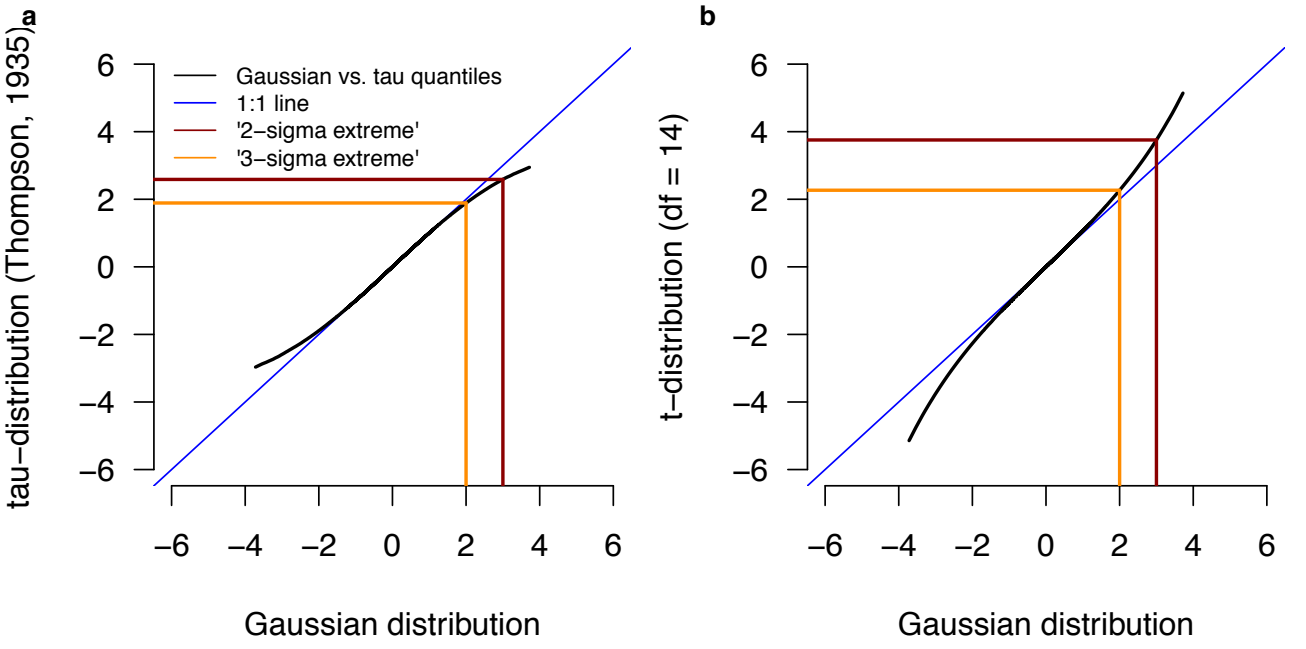

Figure S1. Proposed analytical correction for normalization-induced artefacts. Quantile-quantile plots of original Gaussian distributions vs. a) *tau*-distribution and b) the corresponding t-distribution after normalization. The reference period length was chosen as $n = 15$ for illustration purposes. The simple quantile correction proposed is illustrated for the normalization within a reference period (a) and in the out-of-base (independent) period (b) for 2σ and 3σ extremes.

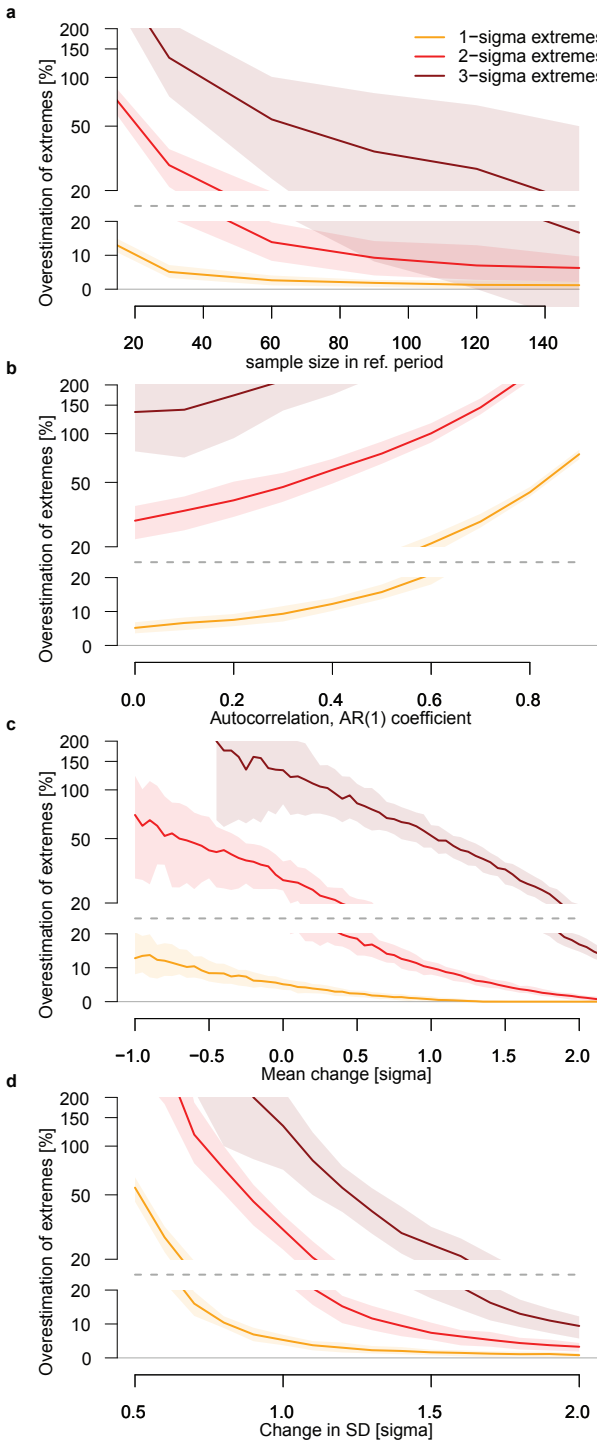

Figure S2. Sensitivity tests of normalization-induced biases in the tails. Monte-Carlo type simulations are conducted to show how the biases in the upper tail are affected by a) varying sample size, b) different degrees of autocorrelation, c,d) trends and changing variance in the out-of-base period, respectively.

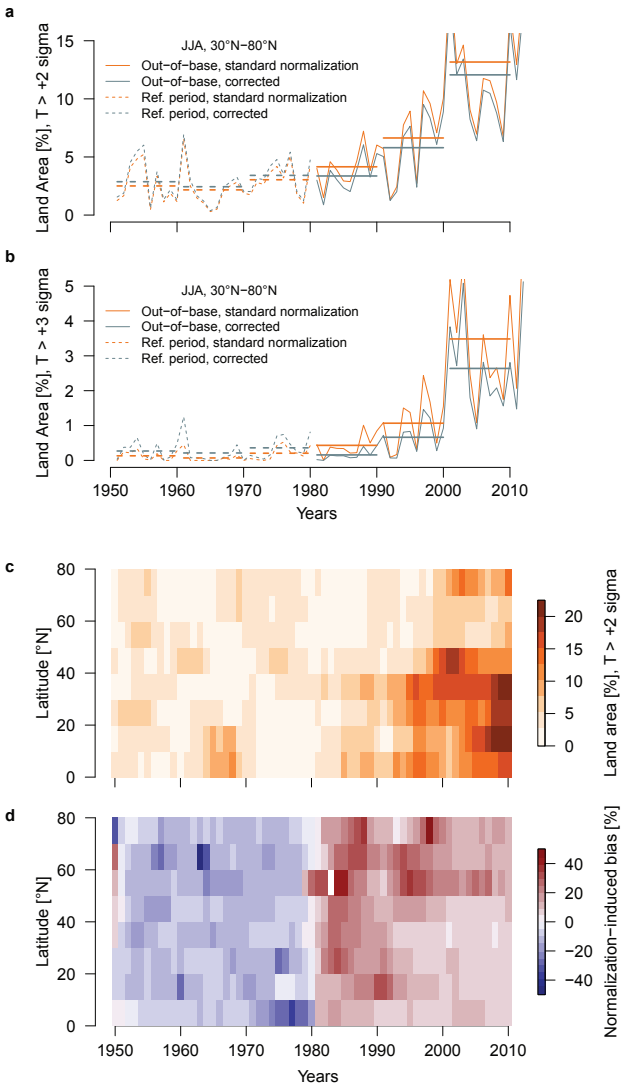

Figure S3. Increase in normalized hot temperature extremes in a spatio-temporal dataset (20th Century Reanalysis). a,b) Time series of fraction of extratropical Northern hemisphere land area covered by positive monthly 2σ (a) and 3σ (b) extremes in summer (reference period: 1951-1980). Horizontal lines indicate decadal averages for the conventional normalization procedure (light blue) and our proposed correction (orange). c) Zonal evolution of fraction of land area covered by monthly positive 2σ extremes in Northern hemisphere summer. d) Zonal evolution of relative biases induced by the conventional normalization approach. In all panels, the time series have been detrended before estimating the estimate of the standard deviation in the reference period (1951-1980).

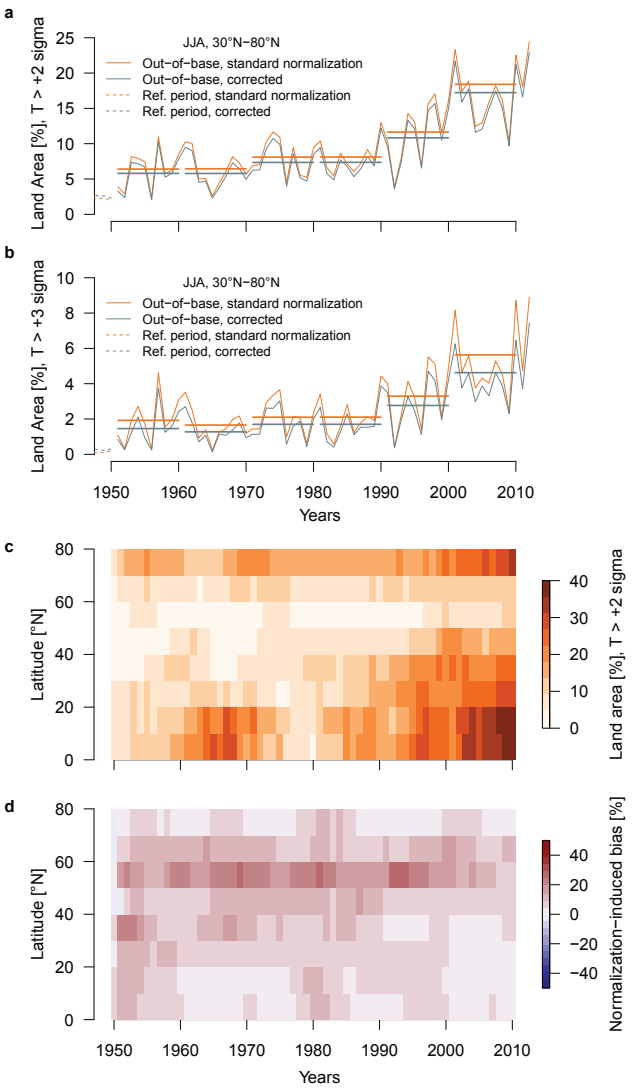

Figure S4. Increase in normalized hot temperature extremes in a spatio-temporal dataset (20th Century Reanalysis). a,b) Time series of fraction of extratropical Northern hemisphere land area covered by positive monthly 2σ (a) and 3σ (b) extremes in summer (reference period: 1951-1980). Horizontal lines indicate decadal averages for the conventional normalization procedure (light blue) and our proposed correction (orange). c) Zonal evolution of fraction of land area covered by monthly positive 2σ extremes in Northern hemisphere summer. d) Zonal evolution of relative biases induced by the conventional normalization approach. In all panels, the time series have been detrended before estimating the estimate of the standard deviation in the reference period (1921-1950).

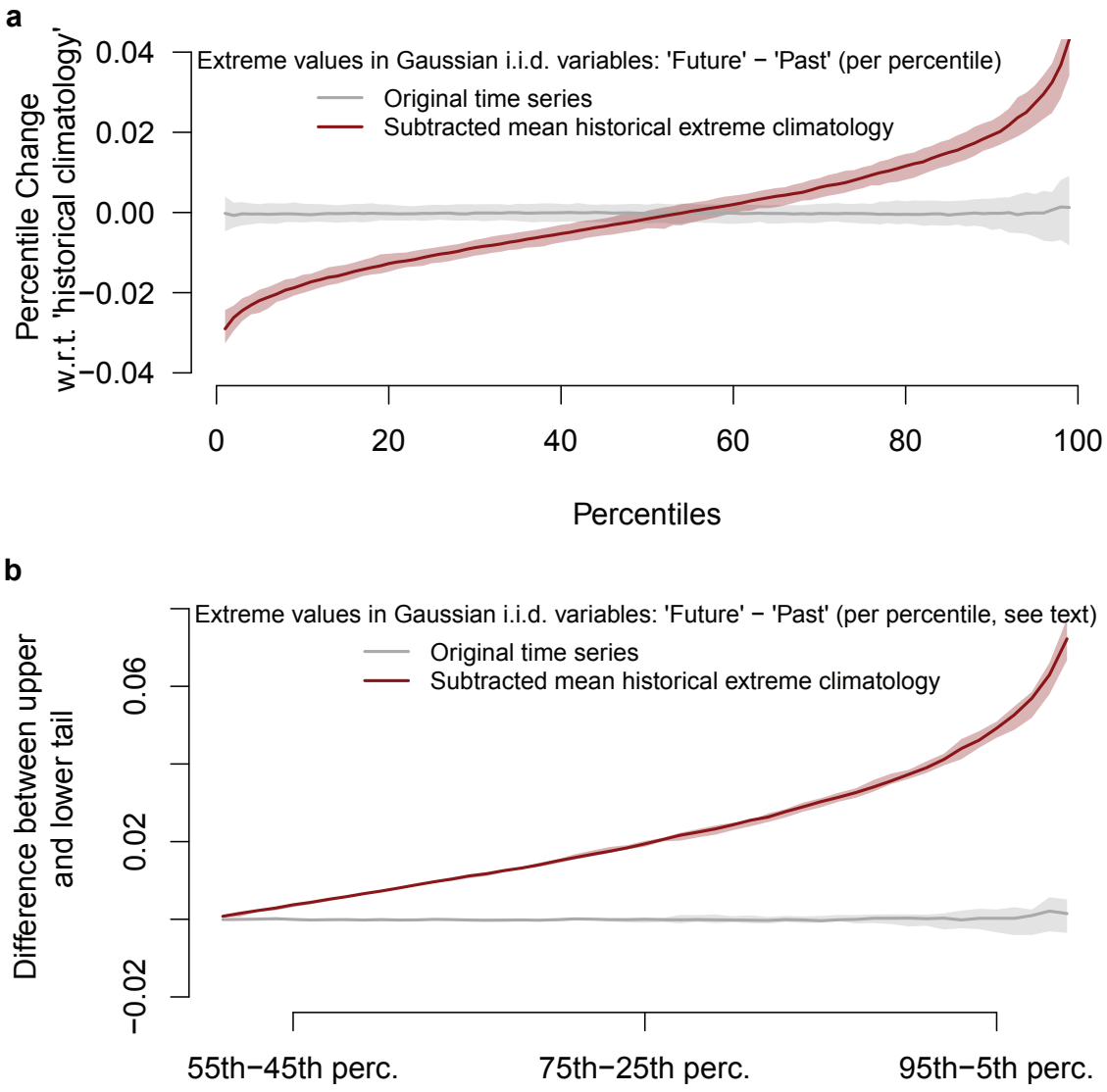

Figure S5. Spurious increase in asymmetry due to data pre-processing. a) Percentile-wise changes across a large number of time series, expressed as the difference between a 'historical' and 'future' period. Induction of asymmetry occurs only if a historical mean climatology is estimated and subtracted from each time series. b) Like above, but differences in symmetric percentiles between the upper and lower tail, further illustrating induced asymmetry in the upper tail. Results are likewise applicable to the lower tail (with reversed asymmetry), if extreme value distribution are generated from minimum values.