

Changes in Global and Regional Extreme Temperatures Over Land: An Extreme Value Theory Perspective

Adam Michael Bauer^{1,2}, Vikrant Sapkota^{1,2}, Ian Baxter³, Tiffany A. Shaw³, B. B. Cael^{1,2,3}

¹Climate Systems Engineering Initiative, University of Chicago, Chicago, Illinois

²Institute for Climate and Sustainable Growth, University of Chicago, Chicago, Illinois

³Department of the Geophysical Sciences, University of Chicago, Chicago, Illinois

Key Points:

- The generalized extreme value distribution cannot be rejected as an appropriate statistical model for extreme temperatures over land.
- We find no statistically significant trend in global extreme temperature relative to the mean; we do find significant regional trends.
- Variability contributes more to extreme temperature bias between CMIP6 models and reanalysis than the mean-state for large fractions of land.

Corresponding author: Adam Michael Bauer, ambauer@uchicago.edu

Abstract

Heat waves are among the most damaging consequences of climate change, yet how the distribution of extreme temperatures is changing remains unclear. Here, we show that the generalized extreme value distribution is fit-for-purpose to analyze extreme temperatures over land. We find that there is no statistically significant trend in global maximum temperatures relative to the mean in the reanalysis record. However, we do find statistically significant regional trends: about 14% of the land surface has a statistically significant positive trend relative to the mean, and about 17% of the land surface has a significant negative trend. We show that climate models struggle to reproduce the distributional characteristics of extreme temperatures found in reanalysis, other than the distributional mean. Finally, we find that biases in the variability, rather than the mean-state, of the extreme temperature distribution dominate biases in projections of extreme temperature risk over large fractions of the land surface.

Plain Language Summary

Heat waves have deleterious impacts on wellbeing, and these impacts are expected to grow with climate change. However, how and where the distribution of extreme temperatures is changing over time in response to increased greenhouse gas emissions remains unclear. In this paper, we develop a well-posed statistical framework, based on extreme value theory, to study changes in the extreme temperature distribution both at a global and regional level. Using our model, we argue that, in the reanalysis record, the global average heat wave is not growing faster than global average temperatures; however, this global perspective glosses over important regional heterogeneities, where in some locations the growth rate of the average heat wave is outpacing the growth rate of average temperatures. We then show that climate models struggle to reproduce the regional pattern of heat wave distribution characteristics outside of the distribution mean. Finally, we quantify the relative importance of variability and the mean-state in determining biases in extreme temperature risk between climate models and reanalysis. We find that, in both the historical record and when projecting extreme temperature risk into the future, bias in variability dominates the influence of bias in the mean-state over large fractions of the land surface.

1 Introduction

Heat waves are one of the most destructive weather events associated with climate impacts (Carleton & Hsiang, 2016; Rennert et al., 2022). As global temperatures rise from continued greenhouse gas emissions, heat waves become increasingly common, with potentially substantive impacts (Fischer et al., 2021). Heat waves can become more likely in a changing climate for two reasons: (1) the preindustrial distribution of temperatures is shifted to be warmer, leading to more heat waves despite no change in the shape of the underlying distribution, and (2) the upper tail of the temperature distribution is “stretching”, making extremes more likely even after taking distribution shifts into account. The former effect is often referred to as “mean-shifting”.

Prior work on trends in extreme temperatures in both the observational record and in climate models has highlighted both global and regional effects. McKinnon et al. (2024) leverage a non-parametric approach to argue that global maximum temperatures are not warming faster than the summertime median, implying only a “mean-shifting” effect and no “tail-stretching” of the warm tail. Regionally there has been both non-parametric statistical evidence (Kornhuber et al., 2024) and dynamical evidence (Patterson, 2023; Vautard et al., 2023) for “tail-stretching”, particularly in western Europe. However, McKinnon et al. (2024) found very scarce statistically significant regional evidence of “tail-stretching”. Many studies have further found that climate models tend to underestimate changes in extremes (Kornhuber et al., 2024; McKinnon et al., 2024) for reasons such as a too-weak meridional temperature gradient (Patterson, 2023) or a lacking representation of long-term aerosol changes (Schumacher et al., 2024).

Each of the statistical studies mentioned above uses non-parametric statistical techniques, a class of statistical methods that forego assuming a functional form for the distribution of the data that they analyze. This can be appealing as data sparsity can make determining what distributional model is appropriate difficult. However, if using a particular functional form is justified based on both (a) the process that generated the data and (b) its ability to reproduce the variability in the data it is used to model, a parametric approach can elucidate patterns in the data that a non-parametric approach misses. For instance, if the total measurement error for some observable quantity is the sum of many independent random processes, the Central Limit Theorem suggests that these errors should be Gaussian. If the residuals of a regression of those data are also Gaussian,

an Ordinary Least Squares (OLS) regression of those data may detect a significant slope that a rank-based regression may miss (due to the additional power of specifying a likelihood function implicitly in the OLS regression).

A conventional parametric way to analyze extreme statistics is extreme value theory, and in particular, the generalized extreme value (GEV) distribution. The GEV has been used in a wide variety of settings ranging from finance (Rocco, 2014) to engineering (Chowdhury et al., 1991), and has been utilized in a number of climate-related studies (Katz, 2010; Huang et al., 2016; Philip et al., 2020; Wehner et al., 2020; Kim et al., 2020; Fischer et al., 2021; Li et al., 2021; Gründemann et al., 2022; Wang et al., 2022; Thompson et al., 2023; Cael et al., 2024). Despite its use in many studies in the climate literature, however, the GEV distribution has not yet been verified as a fit-for-purpose model for the distribution of extreme temperatures over land.

Here, we show that the GEV distribution can be utilized to analyze extreme temperatures, and use it to understand changes in the distribution of extreme temperatures over land. Our paper makes three contributions.

The first is a rigorous statistical test that verifies the GEV distribution cannot be rejected as an appropriate parametric model for the distribution of temperature extremes over land. While many previous studies have utilized the GEV distribution to quantify climate risk, none of them carried out such a test to qualify their analysis. Our analysis rigorously verifies the approach of not only our study, but previous studies as well (e.g., Wehner et al., 2020; Kim et al., 2020; Thompson et al., 2023).

Our second contribution is to quantify the characteristics of the extreme temperature distribution in reanalysis and probe how and where these characteristics are changing. While our results agree with those of McKinnon et al. (2024) and indicate no statistically significant evidence of “tail-stretching” globally, we also uncover significant regional heterogeneity (Kornhuber et al., 2024). We find that some regions like western Europe (Vautard et al., 2023; Schumacher et al., 2024) have significant positive trends in distributional characteristics (14% of land), and some regions, like the midwestern United States (Singh et al., 2023), have negative trends (17% of land).

Finally, we quantify biases in climate model representations of extreme temperatures by comparing the distributional characteristics of the GEV when fit to climate model

output against the reanalysis. We find that climate models tend to reproduce the spatial pattern of the distributional mean quite well, though there is some bias in the intensity of the average maximum temperature. We further find that climate models struggle to reproduce the spatial pattern of variability in the extreme temperature distribution, as well as nonstationary features in the distribution. We diagnose the relative importance of these biases by comparing whether the mean-state or variability dominates biases in return levels for extreme heat, both in the historical record and in 2050. We find large areas where variability dominates bias, while in other areas the mean-state is more important, and comment on potential mechanisms that would give rise to these patterns.

2 Materials and Methods

2.1 Data

We utilize the $1^\circ \times 1^\circ$ gridded ERA5 reanalysis product for global data on two-meter air temperature (Hersbach et al., 2020). We use the reanalysis product to compute daily mean annual maxima and daily mean annual means, and apply the ERA5 land mask to exclude oceans. We analyze reanalysis data over the modern satellite era (1979–2024) owing to the sparsity of *in situ* measurements of temperature prior to 1979 (McKinnon et al., 2024). Our results are unchanged if 1950–1978 are included (*Supporting Information*).

We analyze three measures of extreme temperatures: the raw annual maximum temperature, the annual maximum temperature relative to the long-term trend in the annual mean temperature, and the annual maximum temperature relative to the annual mean temperature in the same year. Of course, there exist a number of additional definitions of extreme heat (Robinson, 2001; Klein Tank et al., 2009; Perkins & Alexander, 2013). We follow Thompson et al. (2023) in focusing on annual maximum temperature given that our study is global in scope, keeping in mind that for regional studies, other metrics might be better suited (Van Oldenborgh et al., 2018).

We utilize 29 models in the CMIP6 ensemble (Eyring et al., 2016) to test if the extreme value statistics of extreme temperatures in climate models agree with reanalysis. To ensure coverage over 1979–2024, we merge historical runs (providing data from 1979–2014) with SSP5–8.5 future projections (providing data for 2015–2024). Using SSP5–

8.5 is not expected to bias our results (relative to using, say, SSP2–4.5) as CO₂ concentration levels for each scenario do not significantly diverge over this period (the difference in 2024 is ≈ 2 ppm; Riahi et al., 2017). We use the same algorithm applied to the reanalysis dataset to process CMIP6 model output. In a sensitivity test, we utilize 36 AMIP simulations over 1979–2014 (Gates et al., 1999).

For our experiments that utilize each climate model in our ensemble ($n = 29$ for CMIP6 and $n = 36$ for AMIP), we select one model initialization per model to sample inter-model uncertainty. We also select each initialization of four models to sample internal climate variability. For a table of utilized models and member IDs, see the *Supporting Information*.

2.2 Statistical Theory

2.2.1 Kuiper Statistic

Our null hypothesis is that the GEV distribution is fit-for-purpose for analyzing temperature extremes over land. To test our null hypothesis, we utilize the Kuiper statistic, a generalization of the Kolmogorov–Smirnov statistic. The Kuiper statistic is defined as the sum of the maximum distance between the empirical cumulative distribution function (CDF) of a random variable x , $E(x)$, and the modeled CDF, $H(x)$, as taken from both sides of the CDFs,

$$V := \max [E(x) - H(x)] + \max [H(x) - E(x)]. \quad (1)$$

The primary benefit of using the Kuiper statistic is that it is as sensitive to discrepancies in the tails of the CDFs as it is sensitive to discrepancies in the medians (Press et al., 1992).

2.2.2 Generalized Extreme Value Distribution

We utilize the GEV distribution to analyze the block maxima of temperature in a given location. The GEV distribution is the only possible limit distribution for the properly normalized extrema of a sequence of independent and identically distributed random variables. In this way, the GEV is to extreme value theory what the Gaussian distribution is to classical statistics (via the Central Limit Theorem) (De Haan & Ferreira, 2006).

168 The probability distribution function of the GEV is given by

$$P(x, t) = \frac{1}{\sigma_t} u(x, t)^{\xi_t + 1} e^{-u(x, t)}, \quad (2)$$

169 where

$$u(x, t) := \begin{cases} \left(1 + \xi_t \left[\frac{x - \mu_t}{\sigma_t}\right]\right)^{-1/\xi_t} & \xi_t \neq 0 \\ \exp\left[-\frac{x - \mu_t}{\sigma_t}\right] & \xi_t = 0 \end{cases} \quad (3)$$

170 and x is a random variable, $\mu_t \in \mathbb{R}$ is the location parameter and roughly corresponds
 171 to the distribution mean, $\sigma_t > 0$ is the scale parameter and corresponds to the distri-
 172 bution variance, and $\xi_t \in \mathbb{R}$ is the shape parameter that controls the shape and heav-
 173 iness of the distribution tails. Treating each GEV parameter as a function of time is a
 174 natural way to incorporate nonstationarity in the model (see, e.g., Furrer et al., 2010;
 175 Katz, 2010). In the context of temperature extremes, theoretical papers (e.g., Stein, 2017)
 176 have argued that $\xi_t \leq 0$, and in practice ξ_t is never less than -0.5 , as this is the the-
 177 oretical lower bound for maximum likelihood estimation (MLE) algorithms to be well-
 178 posed (Coles, 2001).

179 We focus on two configurations of the GEV distribution. The first ignores nonsta-
 180 tionary effects and sets the shape parameter to zero, which corresponds to a stationary
 181 Gumbel distribution. This configuration limits the number of free parameters to avoid
 182 overfitting to a small sample of points and avoids the MLE estimator being ill-posed. In-
 183 deed, the shape parameter is constrained by the third moment or higher of the data (Coles,
 184 2001), and because we have a limited sample set, the MLE-estimated shape parameter
 185 is unlikely to be robust. To capture nonstationarity in the system, we leverage a second
 186 Gumbel configuration where the location parameter is allowed to vary linearly with time (Furrer
 187 et al., 2010; Cael et al., 2024),

$$\mu_t = \mu + \mu_{trend} t. \quad (4)$$

188 Note we do not fit a nonstationary trend in the scale parameter as previous studies have
 189 found no change in the size of variability over time (Rhines & Huybers, 2013). We es-
 190 timate GEV parameters $(\mu, \sigma, \mu_{trend})$ and their uncertainties at each grid cell using MLE
 191 and the Wald Hessian method (Coles, 2001), respectively (*Supporting Information*).

192 The interpretation of μ_{trend} depends on the data being fit. For the results in the
 193 main text, we use annual maximum temperatures relative to the annual mean, making
 194 μ_{trend} a test of pure “mean-shifting”: if annual maxima and means warm at the same

rate, the anomaly distribution is stationary and $\mu_{trend} = 0$. If $\mu_{trend} > 0$, then annual maxima are warming faster than annual means and there is “tail-stretching” in the distribution. (See the *Supporting Information* for a diagram demonstrating this point.)

2.3 Return Levels and Critical Return Periods

We can compute the return level $X_t(p)$ for a given return period $p > 0$ at time t using

$$X_t(p; \xi) = \mu_t + \sigma F(p; \xi), \quad (5)$$

where

$$F(p; \xi) = \begin{cases} -\ln(-\ln(1 - 1/p)), & \xi = 0 \\ -\frac{1}{\xi} \left[1 - (-\ln(1 - 1/p))^{-\xi} \right] & \xi \neq 0 \end{cases}. \quad (6)$$

We can exploit the linearity of (5) in μ_t and σ , that we set the shape parameter to be fixed, and that $F(p; \xi)$ is always increasing with p , to solve for the return period, p_t^* , such that for all $p > p_t^*$, misfit in the scale parameter has a greater influence on overall return level bias than misfit in the location parameter (*Supporting Information*). This p_t^* , which we will refer to as the critical return period, is given by

$$p_t^* = \begin{cases} [1 - \exp(-\exp[-\delta\mu_t/\delta\sigma])]^{-1} & \xi = 0 \\ [1 - \exp(-[1 + \xi\delta\mu_t/\delta\sigma]^{-1/\xi})]^{-1} & \xi \neq 0 \end{cases} \quad (7)$$

where $\delta\mu_t := \bar{\mu}_t^{CMIP} - \mu_t^{ERA5}$ and $\delta\sigma := \bar{\sigma}^{CMIP} - \sigma^{ERA5}$ are the biases between the CMIP6 ensemble mean of the location parameter and scale parameter, $\bar{\mu}_t^{CMIP}$ and $\bar{\sigma}^{CMIP}$, respectively, and the ERA5-estimated location and scale parameters, μ_t^{ERA5} and σ^{ERA5} , respectively. When p_t^* is small (e.g., 10 years), variability (as captured by the scale parameter) dominates biases in return levels; when p_t^* is large (e.g., 1000 years), the mean-state (as captured by the location parameter) dominates.

3 Results and Discussion

3.1 Can We Reject the GEV Distribution?

Discrepancies between reanalysis and the GEV can arise from: (1) limited sample size, (2) the GEV model being insufficient to describe the data (“model error”), and (3) the GEV parameters being chosen incorrectly (“parameter error”). To measure the size of these sources of error, we carry out two related tests.

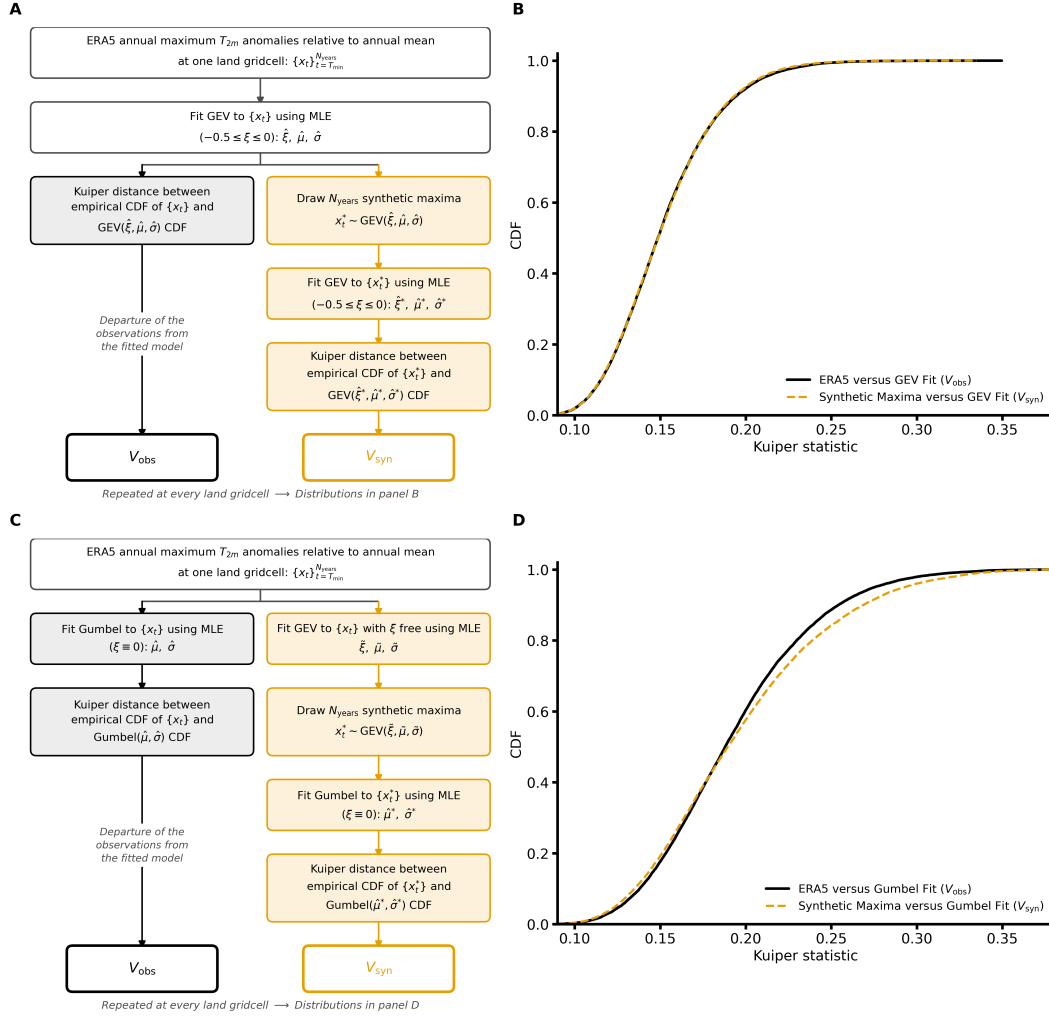


Figure 1. Schematic of Statistical Test of GEV Distribution and Pooled Kuiper Statistic Distribution. **A** and **C** show schematic outlines of our statistical tests to verify our statistical model. Note T_{min} is the first year in which we have data and N_{years} is the total number of years in our dataset ($N_{years} = 45$ when $T_{min} = 1979$). In **B**, black solid lines show the pooled Kuiper statistic CDF when the empirical CDF of ERA5 annual maximum temperature anomalies at a given location are compared with the fitted GEV distribution at a given location. The gold dashed line shows the Kuiper statistic CDF when we compare synthetic samples from a fitted GEV distribution against a GEV distribution that is fit to the synthetic data. In **D**, black solid lines show the pooled Kuiper statistic CDF when the empirical CDF of ERA5 annual maximum temperature anomalies at a given location are compared with the fitted Gumbel distribution at a given location. The gold dashed line shows the Kuiper statistic CDF when we compare synthetic samples from a fitted GEV distribution against a Gumbel distribution that is fit to the synthetic data. See *Can We Reject the GEV Distribution?* for more information on our statistical tests. Computations used anomalies in the annual maximum temperature relative to the annual mean temperature in the same year using reanalysis data from 1979–2024.

The first test compares errors (1) and (2) (Fig. 1A–B). First, we fit a GEV distribution at each grid cell to the reanalysis and compute the Kuiper statistic using (1). Second, we draw bootstrapped samples from the fitted GEV distribution at each grid cell, fit a GEV to the bootstrapped samples, and compute the Kuiper statistic between the samples and GEV fit to those samples. We take the same number of samples as we have years in the reanalysis record. We pool Kuiper statistics at each grid cell and report the cumulative distribution of Kuiper statistics in Fig. 1B.

Fig. 1B shows that small sample size explains nearly 100% of the errors between reanalysis and the GEV: the gold line in Fig. 1B represent errors in fitting a statistical model with no model or parameter errors, and the pooled errors are in near-perfect alignment with the black lines that are fit to the reanalysis. We therefore cannot reject our hypothesis that the GEV distribution is fit-for-purpose, justifying its use for the remainder of our analysis and in other studies (e.g., Wehner et al., 2020; Kim et al., 2020; Thompson et al., 2023).

However, in practice, we cannot be confident that the shape parameter, ξ , is robustly estimated when we have < 100 samples (Coles, 2001). The sign of the shape parameter is known to have substantial implications for risk assessment (Philip et al., 2020), and theoretical papers have argued that for temperature data the shape parameter is less than or equal to zero (Stein, 2017). We therefore must also test the validity of our assumption that a Gumbel distribution (which sets $\xi = 0$) describes the data well.

To this end, we first fit a Gumbel distribution to the reanalysis and compute the Kuiper statistic at each grid cell (black line, Fig. 1D). Second, we fit a *GEV* distribution to the reanalysis (allowing ξ to vary), draw bootstrapped samples from the fitted *GEV* distribution, fit a *Gumbel* distribution to the bootstrapped samples, and compute the Kuiper statistic (gold line, Fig. 1D). By first allowing ξ to be determined by MLE and then fitting a Gumbel distribution to samples drawn from the fitted distribution, we account for misspecification between the distribution we use to sample synthetic data (a GEV distribution with arbitrary shape parameter) and the model we fit to those samples (a Gumbel distribution).

As shown in Fig. 1D, parameter errors incurred by our choice of a Gumbel distribution introduce very small discrepancies between the two Kuiper CDFs, supporting our approximation that the data are Gumbel distributed. The Gumbel distribution fit each

type of anomaly we used for annual maximum temperature data best, while a Weibull distribution where $\xi = -0.25$ fit annual minimum data better than a Gumbel (*Supporting Information*). Our results therefore suggest that any residual misfit of our model at a given location arises primarily from small sample size rather than model or parameter errors in our statistical model, justifying its use for the remainder of our analysis.

3.2 Insights into the Extreme Temperature Distribution in Reanalysis

In Figs. 2A–B we show the location and scale parameters when nonstationary effects are excluded, and in Fig. 2C we show the trend in the location parameter when we include nonstationary effects. The spatial pattern of the location and scale parameters do not materially change when nonstationary effects are included (*Supporting Information*). We show equivalent plots for other anomaly types, annual minimum temperatures, and different GEV configurations in the *Supporting Information*.

We find that the spatial pattern of the stationary piece of the GEV location parameter and scale parameter in Fig. 2A–B are consistent with the spatial pattern of annual maximum temperature anomalies relative to the annual mean, with the average maximum temperature anomaly being larger at high latitudes relative to low latitudes. This can be explained by the larger seasonal cycle at high latitudes that leads the annual maximum temperature (normally occurring in summertime) to be higher than the annual mean (that is weighed down by wintertime temperatures).

The location parameter trend shown in Fig. 2C highlights the heterogeneity in the pattern of change for extreme temperatures over land. At the global level, we find that the area-weighted average location parameter trend is -0.2 ± 0.5 °C per decade. Our parametric statistical model therefore corroborates previous non-parametric statistical work arguing that there is no statistically significant trend (i.e., no “tail-stretching”) in global extreme temperatures relative to average temperatures (McKinnon et al., 2024).

However, we find that some notable regions have a positive trend in the location parameter and thus “tail-stretching”: the west coast of the United States, most of South America, western Europe, and southern Australia all have statistically significant positive trends, implying that the average maximum temperature is warming faster than the annual mean. The locations where we find positive “tail-stretching” make up about 14% of the global land surface.

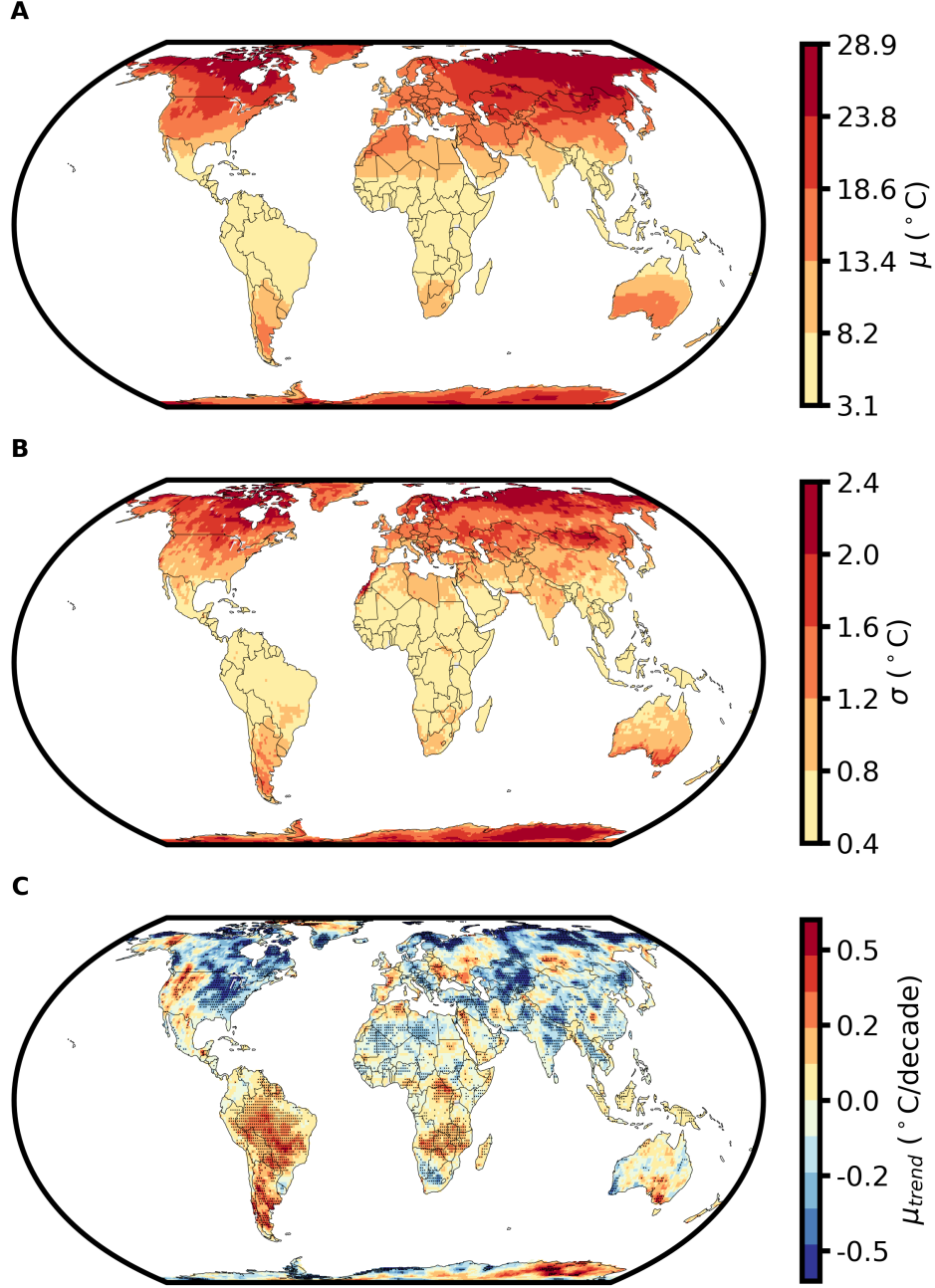


Figure 2. GEV Parameters. **A** shows the location parameter in each grid cell over land for the stationary Gumbel fit. **B** is as **A** but for the scale parameter. **C** shows the location parameter trend when we include nonstationary effects in the GEV fit. The location parameter corresponds to the distribution mean, the scale parameter corresponds to the distribution variance, and the location trend tests for “tail-stretching” of the distribution. Stippling shows where we have 95% confidence in the sign of the trend. Temperature anomalies used are annual maxima relative to the annual mean in the same year from 1979–2024 as computed using ERA5 reanalysis.

We also find negative trends across the eastern United States and Canada, as well as in Siberia and much of central Asia. Grid cells with a significant negative trend make up about 17% of the land surface. Note that a negative trend here does not necessarily mean that extreme temperature impacts are getting less troublesome in these regions, since other key drivers of heat stress may have changed and lead to a net increase in heat-related impacts. For example, Van Oldenborgh et al. (2018) show that, even though the annual maximum temperature trend over some parts of India is decreasing (owing to aerosol pollution and internal variability), the wet bulb temperature is increasing over the entire country. Both the positive trend over western Europe and the negative trend in the midwestern United States corroborate previous studies on heat extremes in these regions (Vautard et al., 2023; Singh et al., 2023; Schumacher et al., 2024; Kornhuber et al., 2024).

3.3 GEV Distribution for Extreme Temperatures in Climate Models

We test if climate models agree with reanalysis on the magnitude and trend of extreme temperatures by carrying out the same fitting process as ERA5 for one initialization of each climate model and for every initialization of MIROC6, NorCPM1, UKESM1-0-LL, and CNRM-ESM2-1 (*Materials and Methods*). We then compute (1) the median deviation of the climate model against ERA5 for each GEV parameter, where the median is taken over all grid cells, and (2) the correlation coefficient between the reanalysis GEV parameter and the climate model GEV parameter, with the regression taken over all grid cells. The median deviation provides a measure of bias, while the correlation coefficient indicates how well climate models reproduce the spatial pattern of reanalysis distribution parameters.

Focusing on the stationary component of the location parameter, we find that all simulations we consider have r -squared values above 0.65; while not perfect agreement, this signals that the spatial pattern of the location parameter stationary piece, and therefore the average seasonal cycle, is roughly aligned between models and reanalysis. However, despite solid correspondence in the spatial pattern, there can be large biases, indicating that even when there is agreement in the spatial pattern of the seasonal cycle, there can still exist discrepancies in the magnitude. These biases appear to be dominated by inter-model uncertainties (blue shadow in Fig. 3A) as opposed to internal variability (colored shadows in Fig. 3A). These results imply that climate models have some skill

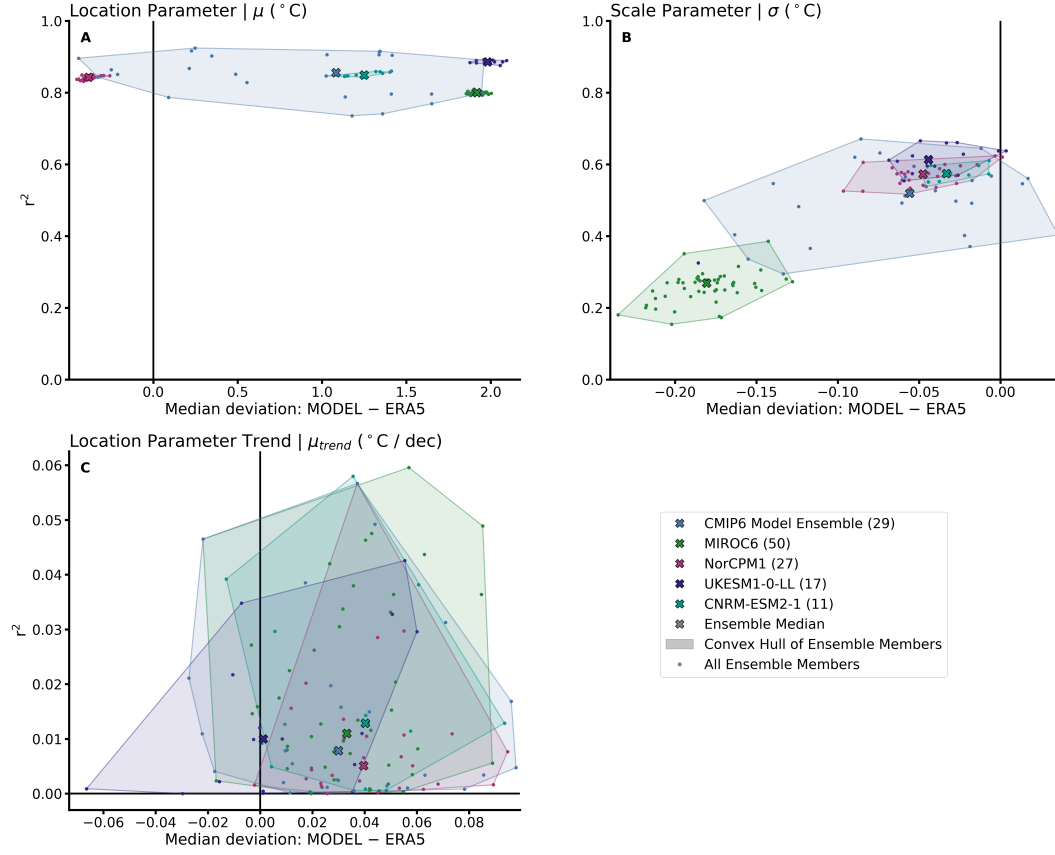


Figure 3. Correlation between CMIP6 and ERA5 GEV Parameters across CMIP6 Models. Each panel shows the correlation between a CMIP6 model and ERA5 reanalysis GEV parameter (see the title) in each grid box plotted against the median deviation between the model and reanalysis GEV parameter in each grid box. As in Figure 2, the location and scale parameters plotted in panels **A** and **B** show the results when no nonstationary effects are included, and panel **C** shows the trend in the location parameter when these effects are included. Note the number of ensemble members for each model are reported in parentheses in the legend. Outliers are excluded for visual clarity (CESM2-WACCM is not shown in all panels; in **A** KACE-1-0-G is not shown; in **B** MIROC-ES2L is not shown and one ensemble member of UKESM1-0-LL is not included in its convex hull). Temperature anomalies used are annual maxima relative to the annual mean in the same year from 1979–2024.

in reproducing the pattern of an *average maximum temperature* in the historical record, though they are still far from perfect.

The skill of climate models in reproducing the reanalysis diminishes substantially for higher order moments of the distribution. For the scale parameter, the r^2 maxes out at 0.65, reflecting discrepancies between climate models and reanalysis in producing the variability of annual maximum temperatures. ERA5 also tends to have a larger scale parameter than CMIP6 models, suggesting that CMIP6 models have weaker variability in extremes than the reanalysis, corroborating dynamical studies (e.g., Vautard et al., 2023). The ability of models to reproduce the spatial pattern of the trend in the location parameter is also poor (Fig. 3C). For the scale parameter, inter-model uncertainties dominate internal variability, while for the location parameter trend both uncertainties have similar magnitudes.

The primary implication of the inability for climate models to reproduce the scale parameter of the reanalysis is that predictions about the *hottest extreme temperature* will be very inconsistent across models. Likewise, mismatches between reanalysis and climate model location parameter trends will result in inconsistent projections of *how the distribution of extreme temperatures will change* over time.

We find that these conclusions are robust for other anomaly types and annual minimum temperatures, as well as when we use AMIP model output instead of CMIP6 model output in the analysis (*Supporting Information*). We therefore expect that results are driven by differences in land-atmosphere coupling, model parametrization, or other differences; we can, however, rule out atmosphere-land-ocean coupling from driving these results. This conclusion is supported by the fact that a separate analysis focused only on the ocean found the multi-model average r^2 for the scale parameter to be 0.87 (Cael et al., 2024), whereas the multi-model average in the present study is 0.50.

How do the differences between the climate model-estimated and reanalysis-derived distributions of extremes propagate into estimates of extreme temperature risk? Previous studies have suggested that biases in the location parameter are the leading-order contributor to biases in return levels between climate models and reanalysis (Kim et al., 2020). However, Kim et al. (2020) only consider the 20-year return level, which, following (7), could have underestimated the role of the scale parameter by focusing on small return periods. Indeed, in the *Supporting Information* we show that the marginal sen-

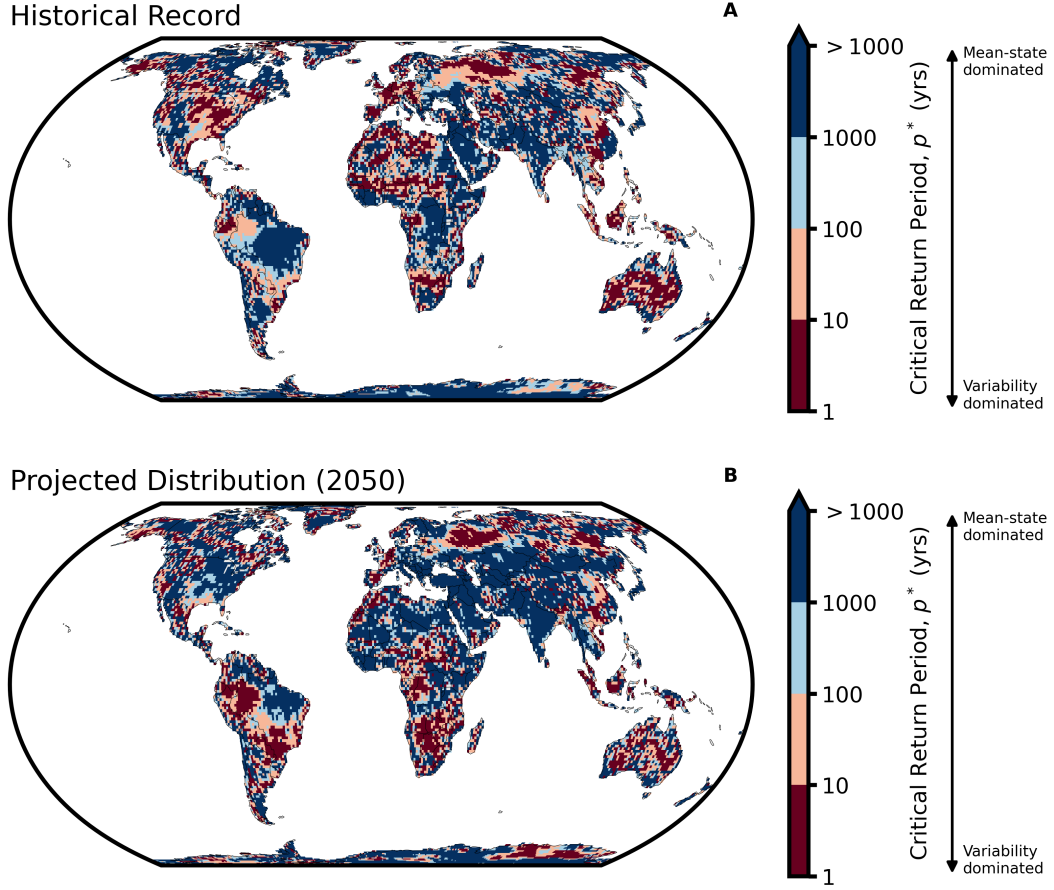


Figure 4. Critical Return Periods. In **A**, we show the critical return period as computed with (7) in the historical record. The critical return period is the period p^* such that for all $p > p^*$, biases in variability dominate biases in the mean-state in determining overall return level bias. Small (large, resp.) values indicate bias in return levels in CMIP6 as compared to ERA5 are dominated by the variability (the mean-state, resp.). **B** is as **A**, but projecting the extreme temperature distribution out to 2050.

sitivity of return levels to changes in the scale parameter is larger than that of the location parameter for all return periods greater than ≈ 3 years when using a Gumbel distribution. This finding is congruent with previous work suggesting that variability is more important than the mean in determining extreme temperature risk in similar extreme value theory-based models (Katz & Brown, 1992). It remains to determine if the total bias in return levels (that is, the sensitivity multiplied by misfit) is dominated by biases in variability or the mean-state.

To this end, in Figure 4 we compute the critical return period, as given by (7), in the historical record (when $\mu_{trend} = 0$, Fig. 4A) and when we project extreme temperatures out to 2050 (Fig. 4B). In the historical record (Fig. 4A), we find that the critical return period is < 10 years for 19% of the land area, and the critical return period is < 100 years for 40% of the land surface. In these locations, it is therefore the variability, not the mean, that is the leading order determinant of biases in return levels and thus projections of climate risk.

In particular, we find that the eastern and midwestern United States, most of Europe, northwestern South America, Australia, Indonesia, and parts of Russia and China are variability-dominated. Previous dynamics-based studies focused on the midwestern United States and western Europe have argued that climate models underestimate the impact of atmospheric circulation on heat extremes (Patterson, 2023; Singh et al., 2023; Vautard et al., 2023), as well as misrepresent changes in aerosol forcing (Schumacher et al., 2024). Similar mechanisms could be leading-order factors in the remaining regions. In contrast, central South America and central Asia are mean-dominated in the historical record.

Projecting the extreme temperature distribution out to 2050, we find similar results in terms of total land area: about 21% of the land surface has critical return periods of < 10 years and 40% has critical return periods of < 100 years (Fig. 4B). Western Europe persists as being variability dominated along with Indonesia and Australia, while central Asia remains mean-state dominated. However, the spatial patterns are not identical between the historical record and in 2050: the midwestern and eastern United States, as well as eastern Europe, flip from being variability dominated in the historical record to being mean-dominated in 2050. This could imply that circulation-driven changes in extreme temperatures found in the historical record are primarily a result of

internal variability, meaning that as extreme temperatures are projected into the future, thermodynamic effects (i.e., mean-state) will begin to dominate circulation effects (i.e., variability). Southern South America and southern Africa transition from being mean-state dominated in the historical record to being variability dominated; future research should focus on possible mechanisms in these regions that could explain this change.

3.4 Conclusion

The extreme value theory-based approach used in this paper unifies two recent findings in the extreme temperatures literature in a single statistical model: that, globally, there is no statistically significant warming trend in extremes relative to the mean (McKinnon et al., 2024), while regionally, there are locations with statistically significant positive and negative trends (Kornhuber et al., 2024). We further found that there are large regions where variability, rather than the mean-state, dominates biases between climate model and reanalysis projections of extreme temperature risk, contrary to previous findings (Kim et al., 2020). Both the regional heterogeneity, as well as the central role of variability, highlighted throughout this study in determining the distribution of rare, high-impact events heightens the need for additional regional models of temperature extremes that span the full model hierarchy (Shaw & Stevens, 2025) to better inform decision-makers on relevant levels of climate risk.

Open Research Section

All data used in this study are public. Code to reproduce this analysis can be found at: <https://github.com/adam-bauer-34/gev-heat-waves>. Code and data will be frozen in Zenodo upon acceptance.

Conflict of Interest declaration

No authors have conflicts of interest to disclose.

Acknowledgments

We are grateful for helpful comments from B. B. Cael’s research group members. AMB is supported by the University of Chicago’s Institute for Climate and Sustainable Growth.

References

- Cael, B. B., Burger, F. A., Henson, S. A., Britten, G. L., & Frölicher, T. L. (2024, January). Historical and future maximum sea surface temperatures. *Science Advances*, 10(4), eadj5569. Retrieved 2026-03-12, from <https://www.science.org/doi/10.1126/sciadv.adj5569> doi: 10.1126/sciadv.adj5569
- Carleton, T. A., & Hsiang, S. M. (2016, September). Social and economic impacts of climate. *Science*, 353(6304), aad9837. Retrieved 2026-03-27, from <https://www.science.org/doi/10.1126/science.aad9837> doi: 10.1126/science.aad9837
- Chowdhury, J. U., Stedinger, J. R., & Lu, L. (1991, July). Goodness-of-fit tests for regional generalized extreme value flood distributions. *Water Resources Research*, 27(7), 1765–1776. Retrieved 2026-03-27, from <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/91WR00077> doi: 10.1029/91WR00077
- Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London: Springer London. Retrieved 2026-02-24, from <http://link.springer.com/10.1007/978-1-4471-3675-0> doi: 10.1007/978-1-4471-3675-0
- De Haan, L., & Ferreira, A. (2006). *Extreme Value Theory* (T. V. Mikosch, S. M. Robinson, & S. I. Resnick, Eds.). New York, NY: Springer New York. Retrieved 2026-03-12, from <http://link.springer.com/10.1007/0-387-34471-3> doi: 10.1007/0-387-34471-3
- Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016, May). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5), 1937–1958. Retrieved 2026-03-12, from <https://gmd.copernicus.org/articles/9/1937/2016/> doi: 10.5194/gmd-9-1937-2016
- Fischer, E. M., Sippel, S., & Knutti, R. (2021, August). Increasing probability of record-shattering climate extremes. *Nature Climate Change*, 11(8), 689–695. Retrieved 2026-03-27, from <https://www.nature.com/articles/s41558-021-01092-9> doi: 10.1038/s41558-021-01092-9
- Furrer, E. M., Katz, R. W., Walter, M. D., & Furrer, R. (2010, October). Statistical modeling of hot spells and heat waves. *Climate Research*, 43(3), 191–205.

- Retrieved 2026-02-24, from <http://www.int-res.com/abstracts/cr/v43/n3/p191-205/> doi: 10.3354/cr00924
- Gates, W. L., Boyle, J. S., Covey, C., Dease, C. G., Doutriaux, C. M., Drach, R. S., ... Williams, D. N. (1999, January). An Overview of the Results of the Atmospheric Model Intercomparison Project (AMIP I). *Bulletin of the American Meteorological Society*, 80(1), 29–55. Retrieved 2026-05-22, from [http://journals.ametsoc.org/doi/10.1175/1520-0477\(1999\)080<0029:A00TRO>2.0.CO;2](http://journals.ametsoc.org/doi/10.1175/1520-0477(1999)080<0029:A00TRO>2.0.CO;2) doi: 10.1175/1520-0477(1999)080<0029:A00TRO>2.0.CO;2
- Gründemann, G. J., Van De Giesen, N., Brunner, L., & Van Der Ent, R. (2022, October). Rarest rainfall events will see the greatest relative increase in magnitude under future climate change. *Communications Earth & Environment*, 3(1), 235. Retrieved 2026-03-27, from <https://www.nature.com/articles/s43247-022-00558-8> doi: 10.1038/s43247-022-00558-8
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J. (2020, July). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049. Retrieved 2026-03-12, from <https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3803> doi: 10.1002/qj.3803
- Huang, W. K., Stein, M. L., McInerney, D. J., Sun, S., & Moyer, E. J. (2016, July). Estimating changes in temperature extremes from millennial-scale climate simulations using generalized extreme value (GEV) distributions. *Advances in Statistical Climatology, Meteorology and Oceanography*, 2(1), 79–103. Retrieved 2026-03-27, from <https://ascmo.copernicus.org/articles/2/79/2016/> doi: 10.5194/ascmo-2-79-2016
- Katz, R. W. (2010, May). Statistics of extremes in climate change. *Climatic Change*, 100(1), 71–76. Retrieved 2026-02-23, from <http://link.springer.com/10.1007/s10584-010-9834-5> doi: 10.1007/s10584-010-9834-5
- Katz, R. W., & Brown, B. G. (1992, July). Extreme events in a changing climate: Variability is more important than averages. *Climatic Change*, 21(3), 289–302. Retrieved 2026-03-23, from <http://link.springer.com/10.1007/BF00139728> doi: 10.1007/BF00139728
- Kim, Y.-H., Min, S.-K., Zhang, X., Sillmann, J., & Sandstad, M. (2020, September). Evaluation of the CMIP6 multi-model ensemble for climate extreme indices.

- 470 *Weather and Climate Extremes*, 29, 100269. Retrieved 2026-06-11, from
 471 <https://linkinghub.elsevier.com/retrieve/pii/S2212094719302439>
 472 doi: 10.1016/j.wace.2020.100269
- 473 Klein Tank, A. M. G., Zwiers, F. W., & Zhang, X. (2009). *Guidelines on Analy-*
 474 *sis of Extremes in a Changing Climate in Support of Informed Decisions for*
 475 *Adaptation* (Climate Data and Monitoring WCDMP No. 72). Switzerland:
 476 World Meteorological Organization. Retrieved from [https://www.ecad.eu/](https://www.ecad.eu/documents/WCDMP_72_TD_1500_en_1.pdf)
 477 [documents/WCDMP_72_TD_1500_en_1.pdf](https://www.ecad.eu/documents/WCDMP_72_TD_1500_en_1.pdf)
- 478 Kornhuber, K., Bartusek, S., Seager, R., Schellnhuber, H. J., & Ting, M. (2024,
 479 December). Global emergence of regional heatwave hotspots outpaces climate
 480 model simulations. *Proceedings of the National Academy of Sciences*, 121(49),
 481 e2411258121. Retrieved 2026-06-23, from [https://pnas.org/doi/10.1073/](https://pnas.org/doi/10.1073/pnas.2411258121)
 482 [pnas.2411258121](https://pnas.org/doi/10.1073/pnas.2411258121) doi: 10.1073/pnas.2411258121
- 483 Li, C., Zwiers, F., Zhang, X., Li, G., Sun, Y., & Wehner, M. (2021, May).
 484 Changes in Annual Extremes of Daily Temperature and Precipitation in
 485 CMIP6 Models. *Journal of Climate*, 34(9), 3441–3460. Retrieved 2026-
 486 06-11, from [https://journals.ametsoc.org/view/journals/clim/34/9/](https://journals.ametsoc.org/view/journals/clim/34/9/JCLI-D-19-1013.1.xml)
 487 [JCLI-D-19-1013.1.xml](https://journals.ametsoc.org/view/journals/clim/34/9/JCLI-D-19-1013.1.xml) doi: 10.1175/JCLI-D-19-1013.1
- 488 McKinnon, K. A., Simpson, I. R., & Williams, A. P. (2024, October). The
 489 pace of change of summertime temperature extremes. *Proceedings of the*
 490 *National Academy of Sciences*, 121(42), e2406143121. Retrieved 2026-
 491 03-23, from <https://pnas.org/doi/10.1073/pnas.2406143121> doi:
 492 10.1073/pnas.2406143121
- 493 Patterson, M. (2023, May). North-West Europe Hottest Days Are Warm-
 494 ing Twice as Fast as Mean Summer Days. *Geophysical Research Let-*
 495 *ters*, 50(10), e2023GL102757. Retrieved 2026-07-10, from [https://](https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL102757)
 496 agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL102757 doi:
 497 10.1029/2023GL102757
- 498 Perkins, S. E., & Alexander, L. V. (2013, July). On the Measurement of Heat
 499 Waves. *Journal of Climate*, 26(13), 4500–4517. Retrieved 2026-06-11, from
 500 <http://journals.ametsoc.org/doi/10.1175/JCLI-D-12-00383.1> doi:
 501 10.1175/JCLI-D-12-00383.1
- 502 Philip, S., Kew, S., Van Oldenborgh, G. J., Otto, F., Vautard, R., Van Der Wiel,

- 503 K., ... Van Aalst, M. (2020, November). A protocol for probabilis-
 504 tic extreme event attribution analyses. *Advances in Statistical Climatol-
 505 ogy, Meteorology and Oceanography*, 6(2), 177–203. Retrieved 2026-06-
 506 11, from <https://ascmo.copernicus.org/articles/6/177/2020/> doi:
 507 10.5194/ascmo-6-177-2020
- 508 Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numeri-
 509 cal recipes in C: the art of scientific computing* (2nd ed ed.). Cambridge: Cam-
 510 bridge university press.
- 511 Rennert, K., Errickson, F., Prest, B. C., Rennels, L., Newell, R. G., Pizer, W.,
 512 ... Anthoff, D. (2022, October). Comprehensive evidence implies a higher
 513 social cost of CO₂. *Nature*, 610(7933), 687–692. Retrieved 2026-03-27,
 514 from <https://www.nature.com/articles/s41586-022-05224-9> doi:
 515 10.1038/s41586-022-05224-9
- 516 Rhines, A., & Huybers, P. (2013, February). Frequent summer temperature extremes
 517 reflect changes in the mean, not the variance. *Proceedings of the National
 518 Academy of Sciences*, 110(7). Retrieved 2026-07-14, from [https://pnas.org/](https://pnas.org/doi/full/10.1073/pnas.1218748110)
 519 doi/full/10.1073/pnas.1218748110 doi: 10.1073/pnas.1218748110
- 520 Riahi, K., Van Vuuren, D. P., Kriegler, E., Edmonds, J., O'Neill, B. C., Fuji-
 521 mori, S., ... Tavoni, M. (2017, January). The Shared Socioeconomic
 522 Pathways and their energy, land use, and greenhouse gas emissions impli-
 523 cations: An overview. *Global Environmental Change*, 42, 153–168. Re-
 524 trieved 2026-03-12, from [https://linkinghub.elsevier.com/retrieve/](https://linkinghub.elsevier.com/retrieve/pii/S0959378016300681)
 525 pii/S0959378016300681 doi: 10.1016/j.gloenvcha.2016.05.009
- 526 Robinson, P. J. (2001, April). On the Definition of a Heat Wave. *Journal of Applied
 527 Meteorology*, 40(4), 762–775. Retrieved 2026-06-11, from [http://journals](http://journals.ametsoc.org/doi/10.1175/1520-0450(2001)040<0762:OTDOAH>2.0.CO;2)
 528 [.ametsoc.org/doi/10.1175/1520-0450\(2001\)040<0762:OTDOAH>2.0.CO;2](http://journals.ametsoc.org/doi/10.1175/1520-0450(2001)040<0762:OTDOAH>2.0.CO;2)
 529 doi: 10.1175/1520-0450(2001)040(0762:OTDOAH)2.0.CO;2
- 530 Rocco, M. (2014, February). Extreme value theory in finance: A survey. *Jour-
 531 nal of Economic Surveys*, 28(1), 82–108. Retrieved 2026-03-27, from [https://](https://onlinelibrary.wiley.com/doi/10.1111/j.1467-6419.2012.00744.x)
 532 onlinelibrary.wiley.com/doi/10.1111/j.1467-6419.2012.00744.x doi:
 533 10.1111/j.1467-6419.2012.00744.x
- 534 Schumacher, D. L., Singh, J., Hauser, M., Fischer, E. M., Wild, M., & Senevi-
 535 ratne, S. I. (2024, April). Exacerbated summer European warming not

- 536 captured by climate models neglecting long-term aerosol changes. *Com-*
 537 *munications Earth & Environment*, 5(1), 182. Retrieved 2026-06-11,
 538 from <https://www.nature.com/articles/s43247-024-01332-8> doi:
 539 10.1038/s43247-024-01332-8
- 540 Shaw, T. A., & Stevens, B. (2025, March). The other climate crisis. *Nature*,
 541 639(8056), 877–887. Retrieved 2026-03-24, from [https://www.nature.com/](https://www.nature.com/articles/s41586-025-08680-1)
 542 [articles/s41586-025-08680-1](https://www.nature.com/articles/s41586-025-08680-1) doi: 10.1038/s41586-025-08680-1
- 543 Singh, J., Sippel, S., & Fischer, E. M. (2023, November). Circulation dampened heat
 544 extremes intensification over the Midwest USA and amplified over Western
 545 Europe. *Communications Earth & Environment*, 4(1), 432. Retrieved 2026-
 546 06-11, from <https://www.nature.com/articles/s43247-023-01096-7> doi:
 547 10.1038/s43247-023-01096-7
- 548 Stein, M. L. (2017, March). Should annual maximum temperatures follow a gen-
 549 eralized extreme value distribution? *Biometrika*, 104(1), 1–16. Retrieved
 550 2026-02-23, from [https://academic.oup.com/biomet/article-lookup/doi/](https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/asw070)
 551 [10.1093/biomet/asw070](https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/asw070) doi: 10.1093/biomet/asw070
- 552 Thompson, V., Mitchell, D., Hegerl, G. C., Collins, M., Leach, N. J., & Slingo,
 553 J. M. (2023, April). The most at-risk regions in the world for high-impact
 554 heatwaves. *Nature Communications*, 14(1), 2152. Retrieved 2026-06-
 555 11, from <https://www.nature.com/articles/s41467-023-37554-1> doi:
 556 10.1038/s41467-023-37554-1
- 557 Van Oldenborgh, G. J., Philip, S., Kew, S., Van Weele, M., Uhe, P., Otto, F., ...
 558 AchutaRao, K. (2018, January). Extreme heat in India and anthropogenic
 559 climate change. *Natural Hazards and Earth System Sciences*, 18(1), 365–381.
 560 Retrieved 2026-06-12, from [https://nhess.copernicus.org/articles/18/](https://nhess.copernicus.org/articles/18/365/2018/)
 561 [365/2018/](https://nhess.copernicus.org/articles/18/365/2018/) doi: 10.5194/nhess-18-365-2018
- 562 Vautard, R., Cattiaux, J., Hap  , T., Singh, J., Bonnet, R., Cassou, C., ... Yiou,
 563 P. (2023, October). Heat extremes in Western Europe increasing faster than
 564 simulated due to atmospheric circulation trends. *Nature Communications*,
 565 14(1), 6803. Retrieved 2026-06-11, from [https://www.nature.com/articles/](https://www.nature.com/articles/s41467-023-42143-3)
 566 [s41467-023-42143-3](https://www.nature.com/articles/s41467-023-42143-3) doi: 10.1038/s41467-023-42143-3
- 567 Wang, H., Zuo, Z., Qiao, L., Zhang, K., Sun, C., Xiao, D., ... Zhang, R. (2022,
 568 November). Frequency of the winter temperature extremes over Siberia

569 dominated by the Atlantic Meridional Overturning Circulation. *npj Cli-*
570 *mate and Atmospheric Science*, 5(1), 84. Retrieved 2026-03-27, from
571 <https://www.nature.com/articles/s41612-022-00307-w> doi: 10.1038/
572 s41612-022-00307-w

573 Wehner, M., Gleckler, P., & Lee, J. (2020, December). Characterization of long
574 period return values of extreme daily temperature and precipitation in the
575 CMIP6 models: Part 1, model evaluation. *Weather and Climate Extremes*,
576 30, 100283. Retrieved 2026-06-11, from [https://linkinghub.elsevier.com/
577 retrieve/pii/S2212094719302440](https://linkinghub.elsevier.com/retrieve/pii/S2212094719302440) doi: 10.1016/j.wace.2020.100283

Supporting Information for: “Changes in Global and Regional Extreme Temperatures Over Land: An Extreme Value Theory Perspective”

Adam Michael Bauer^{1,2}, Vikrant Sapkota^{1,2}, Ian Baxter³, Tiffany A. Shaw³,

B. B. Cael^{1,2,3}

¹Climate Systems Engineering Initiative, University of Chicago, Chicago, Illinois

²Institute for Climate and Sustainable Growth, University of Chicago, Chicago, Illinois

³Department of the Geophysical Sciences, University of Chicago, Chicago, Illinois

Contents of this file

1. Text S1 to S3
2. Figures S1 to S16
3. Table S1

Introduction. This Supporting Information gives additional exposition on our maximum likelihood estimation of GEV parameters, a detailed derivation of the critical return period, and a proof of the relative marginal sensitivity of GEV return levels to bias in the location and scale parameters.

Table S1 details the climate models used in our analysis.

Figure S1 graphically shows the proof given in Text S3. Figure S2 gives a schematic of how $\mu_{trend} > 0$ describes tail stretching. Figures S3-S5 are robustness tests for our Kuiper

statistic test. Figures S6-S11 are as Figure 2 in the main text but for different extreme temperature anomalies and GEV specifications. Figures S12-S16 are as Figure 3 in the main text but for different extreme temperature anomalies.

Text S1: Maximum Likelihood Estimation of GEV Parameters. This section briefly outlines the maximum likelihood estimation (MLE) algorithm for computing GEV parameters and the computational resources utilized to greatly decrease runtime. The MLE algorithm can be written as

$$\begin{aligned} & \max_{(\mu_t, \sigma_t, \xi_t)} \sum_{x_t \in X_t} \ln p_t(x_t; \mu_t, \sigma_t, \xi_t), \\ & \text{Subject to : } \sigma_t > 0 \end{aligned} \quad (1)$$

where T is the maximum time in the interval of analysis, X_t is a set of data in a given year t that we fit the distribution to (e.g., annual maximum temperature in a given grid cell in 1993) and $p_t(\cdot)$ is the GEV probability distribution function described in the main text at a given time t .

We transform the equations in (1) to ensure numerical stability, a common issue when carrying out nonstationary MLE. Namely, we transform the time variable t from $t \in \{1979, 1980, \dots, 2024\}$ to $\tilde{t} \in [0, 1]$ using standard normalization. This allows the trend terms in the GEV parameters to be well-behaved and nearly guarantee stability. After this transformation, our MLE algorithm converged for 100% of grid cells on average for each anomaly type we considered. We utilized the SLSQP algorithm to carry out the maximization (Nocedal & Wright, 2006).

MLE estimation at each grid cell for three types of anomalies is computationally costly; we estimated that doing so serially for three types of anomalies, ERA5, CMIP6, and AMIP would take up to 115 hours on a laptop computer. We utilized `openMPI` (Message Passing Interface Forum, 2021) and `dask` (Dask Development Team, 2016) to carry out

MLE at each grid cell for each anomaly type nearly simultaneously, allowing the process to take under 3 hours of total wall time.

Finally, we use the Hessian to estimate the standard errors for each estimated parameter; if the optimal vector of GEV parameters is $\boldsymbol{\theta}$, then the standard error, σ_{θ_i} , of GEV parameter $\theta_i \in \boldsymbol{\theta}$ is given by diagonal element of the inverse Hessian corresponding to θ_i , i.e.,

$$\sigma_{\theta_i}^2 = -H(\boldsymbol{\theta})_{i,i}^{-1}. \quad (2)$$

The Hessian was computed analytically using Mathematica (Wolfram Research Inc., n.d.) and coded up by hand.

Text S2: Critical Return Periods. In this section, we derive Eqn. (2.7) in the main text for the critical return period p_t^* . Consider the definition of the return level for return period p at time t given by

$$X_t(p) = \underbrace{\mu_t}_{=:X_t^{mean-state}} + \underbrace{\sigma F(p; \xi)}_{=:X^{variability}(p)} \quad (3)$$

where

$$F(p; \xi) = \begin{cases} -\ln(-\ln(1 - 1/p)), & \xi = 0 \\ -\frac{1}{\xi} \left[1 - (-\ln(1 - 1/p))^{-\xi} \right] & \xi \neq 0 \end{cases} \quad (4)$$

as in the main text. In our setup, ξ is treated as a constant prior to carrying out MLE, so $F(p; \xi)$ is a multiplier to σ for all return periods p . This allows us to cleanly define the mean-state contribution to return levels, $X_t^{mean-state}$ and the variability contribution to return levels, $X^{variability}(p)$ in (3). Note that $X_t^{mean-state}$ has no dependence on the return period while $X^{variability}(p)$ has no time dependence, as we treat the variability as constant in time (Rhines & Huybers, 2013).

We can use (3) to compute the average return level at time t for return period p in the CMIP6 ensemble, as

$$\bar{X}_t^{CMIP}(p) = \bar{\mu}_t^{CMIP} + \bar{\sigma}^{CMIP} F(p; \xi), \quad (5)$$

where bars represent ensemble means. We carry out this calculation at each grid cell but neglect lat/lon subscripts on GEV parameters and return levels for clarity. We can carry out the same calculation in ERA5 to compute

$$X_t^{ERA5}(p) = \mu_t^{ERA5} + \sigma^{ERA5} F(p; \xi). \quad (6)$$

The difference between (5) and (6) defines the bias in return levels, $\delta X_t(p)$, which we write as

$$\begin{aligned}\delta X_t(p) &:= \bar{X}_t^{CMIP}(p) - X_t^{ERA5}(p) = \underbrace{\bar{\mu}_t^{CMIP} - \mu_t^{ERA5}}_{=: \delta\mu_t} + \underbrace{(\bar{\sigma}^{CMIP} - \sigma^{ERA5})}_{=: \delta\sigma} F(p; \xi) \\ &= \delta\mu_t + \delta\sigma F(p; \xi)\end{aligned}\tag{7}$$

where we have defined $\delta\mu_t$ and $\delta\sigma$ as in the main text. By comparing (7) and the definitions in (3) we can define

$$\delta X^{variability}(p) := \delta\sigma F(p; \xi)\tag{8}$$

$$\delta X_t^{mean-state} := \delta\mu_t\tag{9}$$

as the variability and mean-state contributions to return level bias, respectively.

We are interested in solving for the p_t^* such that

$$\delta X^{variability}(p) > \delta X_t^{mean-state}.\tag{10}$$

for all $p > p_t^*$. Using (8), (9), and (4), one can show that $\delta X^{variability}(p) > \delta X_t^{mean-state}$

for all $p > p_t^*$ when we write p_t^* as

$$p_t^* = \begin{cases} [1 - \exp(-\exp[-\delta\mu_t/\delta\sigma])]^{-1} & \xi = 0 \\ \left[1 - \exp\left(-[1 + \xi\delta\mu_t/\delta\sigma]^{-1/\xi}\right)\right]^{-1} & \xi \neq 0 \end{cases}\tag{11}$$

after some algebraic effort in inverting $F(p; \xi)$ to solve for p , which completes the derivation.

Text S3: Marginal Sensitivities of Return Levels to GEV Parameters. In this section, we show that there exists a set of return periods $P \subset \mathbb{R}^+$ where the marginal sensitivity of the return level X_p for $p \in P$ is more sensitive to changes in the GEV scale parameter, σ , than the location parameter, μ . We focus only on the case where $\xi_t = 0$; extending this proof to the case where $\xi_t \neq 0$ is straightforward. We show the content of this proof graphically in Figure S1.

Proof. Consider the definition of the return level

$$X_p = \begin{cases} \mu_t - \sigma_t \ln(-\ln(1 - 1/p)), & \xi_t = 0 \\ \mu_t - \frac{\sigma_t}{\xi_t} \left[1 - (-\ln(1/p))^{-\xi_t} \right] & \xi_t \neq 0 \end{cases} \quad (12)$$

To address the sensitivities of the return levels to each parameter, we take the partial derivative of X_p with respect to the location and scale parameters. Doing so, we find that

$$\frac{\partial X_p}{\partial \mu} = 1, \quad (13)$$

and

$$\frac{\partial X_p}{\partial \sigma} = -\ln(-\ln(1 - 1/p)). \quad (14)$$

It remains to solve for the range of p where $\partial X_p / \partial \sigma > \partial X_p / \partial \mu$. Using (13) and (14), we can show that

$$p > \frac{1}{1 - e^{-1/e}} \approx 3.24887. \quad (15)$$

Therefore, for all return periods p in the set $P := \{p : p > (1 - e^{-1/e})^{-1}\}$, the marginal sensitivity of the return level of period p to small changes in the scale parameter is greater than the marginal sensitivity to small changes in the location parameter. $\square$

Table S1: **CMIP6 and AMIP Model Configuration.**

Model	CMIP6 Primary Member	AMIP Primary Member
ACCESS-CM2	N/A	r1i1p1f1
ACCESS-ESM1-5	N/A	r10i1p1f1
AWI-CM-1-1-MR	r1i1p1f1	r1i1p1f1
AWI-ESM-1-1-LR	N/A	r1i1p1f1
BCC-CSM2-MR	r1i1p1f1	N/A
BCC-ESM1	N/A	r1i1p1f1
CAMS-CSM1-0	r1i1p1f1	r1i1p1f1
CESM2-FV2	N/A	r1i1p1f1
CESM2-WACCM	r1i1p1f1	N/A
CESM2-WACCM-FV2	N/A	r1i1p1f1
CMCC-CM2-HR4	N/A	r1i1p1f1
CMCC-CM2-SR5	r1i1p1f1	r1i1p1f1
CMCC-ESM2	r1i1p1f1	r1i1p1f1
CNRM-CM6-1-HR	r1i1p1f1	r1i1p1f1
CNRM-ESM2-1	r10i1p1f2	r10i1p1f2
CanESM5	N/A	r10i1p1f1
EC-Earth3-AerChem	N/A	r1i1p1f1
EC-Earth3-CC	r1i1p1f1	r1i1p1f1
EC-Earth3-Veg	r10i1p1f1	N/A
EC-Earth3-Veg-LR	r1i1p1f1	r1i1p1f1
FGOALS-f3-L	N/A	r1i1p1f1
FGOALS-g3	r1i1p1f1	N/A
GFDL-CM4	r1i1p1f1	r1i1p1f1
GFDL-ESM4	r1i1p1f1	r1i1p1f1
HadGEM3-GC31-LL	r1i1p1f3	r1i1p1f3
HadGEM3-GC31-MM	r1i1p1f3	r1i1p1f3
IITM-ESM	r1i1p1f1	r1i1p1f1
INM-CM4-8	r1i1p1f1	r1i1p1f1
INM-CM5-0	r10i1p1f1	r10i1p1f1
IPSL-CM5A2-INCA	N/A	r1i1p1f1
KACE-1-0-G	r1i1p1f1	r1i1p1f1
KIOST-ESM	r1i1p1f1	r1i1p1f1
MIROC-ES2L	r1i1p1f2	N/A
MIROC6	r10i1p1f1	r10i1p1f1
MPI-ESM-1-2-HAM	N/A	r1i1p1f1
MPI-ESM1-2-HR	r10i1p1f1	r10i1p1f1
MPI-ESM1-2-LR	r10i1p1f1	r10i1p1f1
MRI-ESM2-0	N/A	r1i1p1f1
NorCPM1	r2i1p1f1	N/A
NorESM2-LM	r1i1p1f1	r1i1p1f1
SAM0-UNICON	N/A	r1i1p1f1

Model	CMIP6 Primary Member	AMIP Primary Member
TaiESM1	r1i1p1f1	r1i1p1f1
UKESM1-0-LL	r10i1p1f2	N/A

References

- Dask Development Team. (2016). Dask: Library for dynamic task scheduling [Computer software manual]. Retrieved from <http://dask.pydata.org>
- Message Passing Interface Forum. (2021, June). MPI: A message-passing interface standard version 4.0 [Computer software manual]. Retrieved from <https://www.mpi-forum.org/docs/mpi-4.0/mpi40-report.pdf>
- Nocedal, J., & Wright, S. J. (2006). *Numerical Optimization*. Springer New York. Retrieved 2026-03-23, from <http://link.springer.com/10.1007/978-0-387-40065-5> doi: 10.1007/978-0-387-40065-5
- Rhines, A., & Huybers, P. (2013, February). Frequent summer temperature extremes reflect changes in the mean, not the variance. *Proceedings of the National Academy of Sciences*, 110(7). Retrieved 2026-07-14, from <https://pnas.org/doi/full/10.1073/pnas.1218748110> doi: 10.1073/pnas.1218748110
- Wolfram Research Inc. (n.d.). *Mathematica, Version 14.3*. Retrieved from <https://www.wolfram.com/mathematica> (Champaign, IL, 2025)

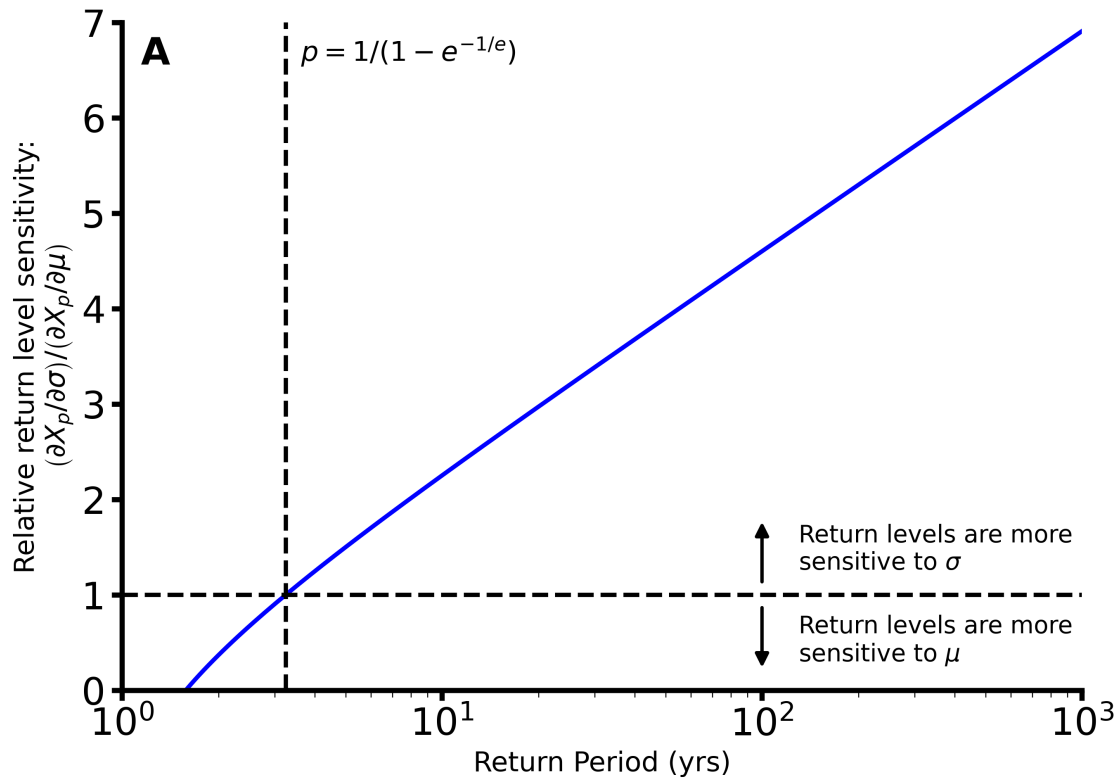


Figure S1. Proof of Relative Sensitivity of Return Levels to Scale and Location Parameter. We show the relative sensitivity of return levels to changes in the scale parameter and location parameter plotted against the return period (blue line). When the blue line is larger than (less than, resp.) one, return levels are more sensitive to changes in the scale parameter (location parameter, resp.). The return period where the relative sensitivity is equal to unity is shown by the vertical line.

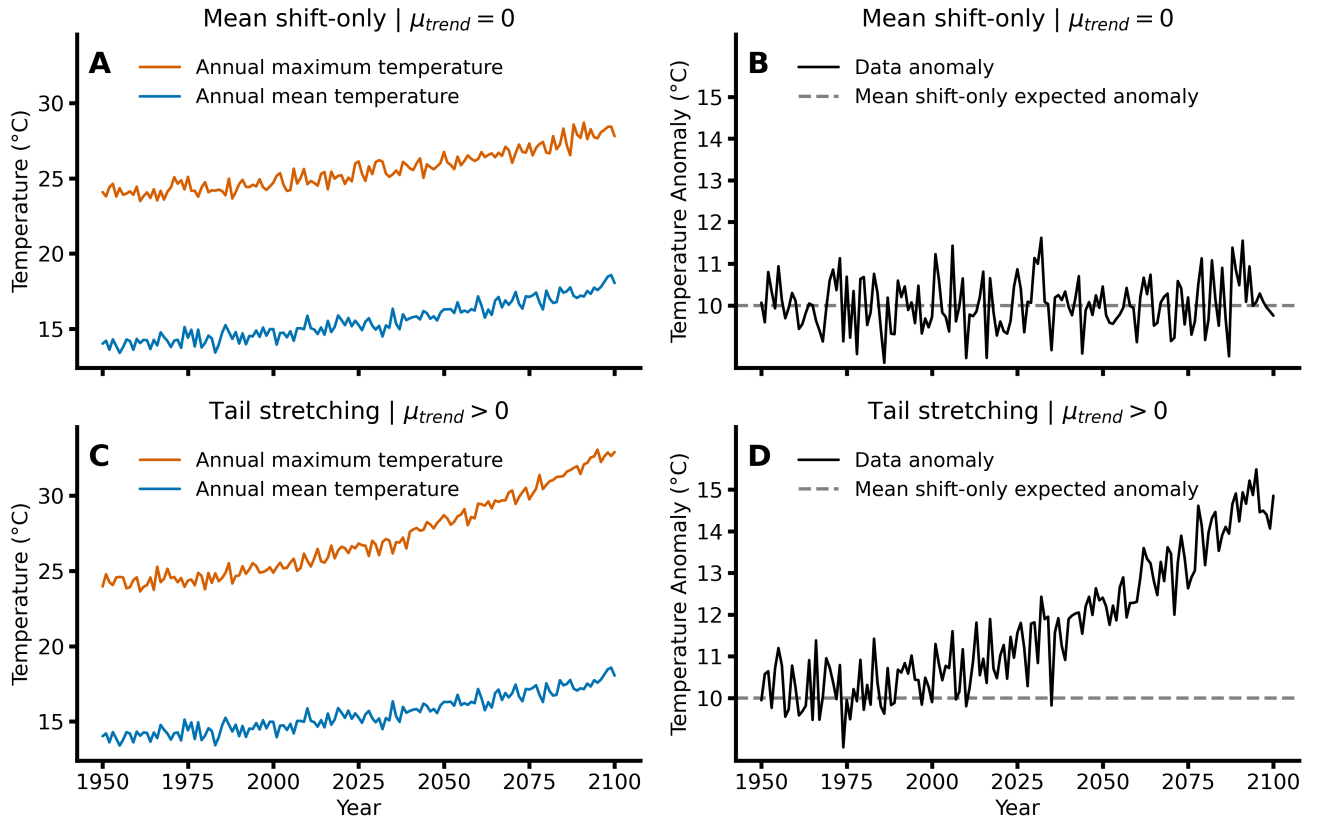


Figure S2. Diagram demonstrating mean shifting versus tail stretching in temperature anomalies. In panel **A**, we show time series of annual maximum temperature (orange line) and annual mean temperature (blue line). Both lines grow at the same rate over time and hence only mean-shifting is demonstrated. In **B**, we compute the anomaly of annual maxima subtracted from the annual mean. Notice that $\mu_{trend} = 0$ as there is no time trend in the anomalies. **C** is as **A**, but now annual maximum temperatures are warming faster than the annual mean, i.e., there is tail stretching. **D** translates this fact into anomaly space, where we observe a time trend in the anomalies. μ_{trend} captures this trend in our analysis whenever $\mu_{trend} > 0$ and the trend is statistically significant. All data are synthetic and meant to be illustrative.

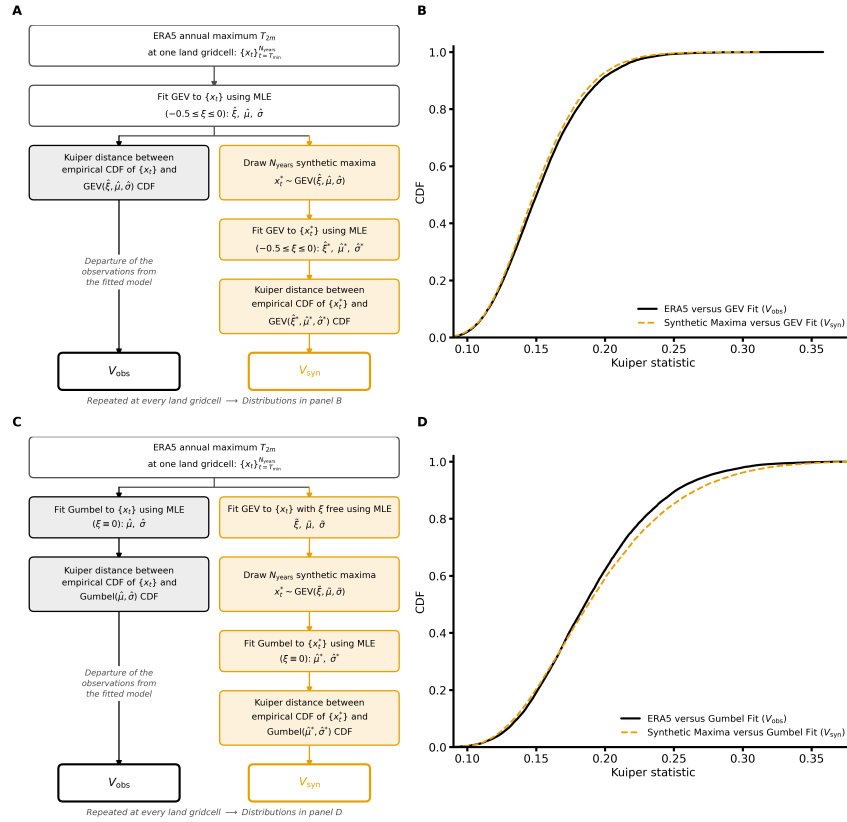


Figure S3. Schematic of Statistical Test of GEV Distribution and Pooled

Kuiper Statistic Distribution (Raw Values). **A** and **C** show schematic outlines of

our statistical tests to verify our statistical model. Note T_{min} is the first year in which we have data and N_{years} is the total number of years in our dataset ($N_{years} = 45$ when

$T_{min} = 1979$). In **B**, black solid lines show the pooled Kuiper statistic CDF when the empirical CDF of ERA5 annual maximum temperatures at a given location are compared

with the fitted GEV distribution at a given location. The gold dashed line shows the Kuiper statistic CDF when we compare synthetic samples from a fitted GEV distribution

against a GEV distribution that is fit to the synthetic data. In **D**, black solid lines show the pooled Kuiper statistic CDF when the empirical CDF of ERA5 annual maximum

temperatures at a given location are compared with the fitted Gumbel distribution at a given location. The gold dashed line shows the Kuiper statistic CDF when we compare

synthetic samples from a fitted GEV distribution against a Gumbel distribution that is fit to the synthetic data. See *Can We Reject the GEV Distribution?* in the main text

for more information on our statistical tests. Computations used raw annual maximum temperatures from 1979-2024.

September 9, 2026, 10:38pm

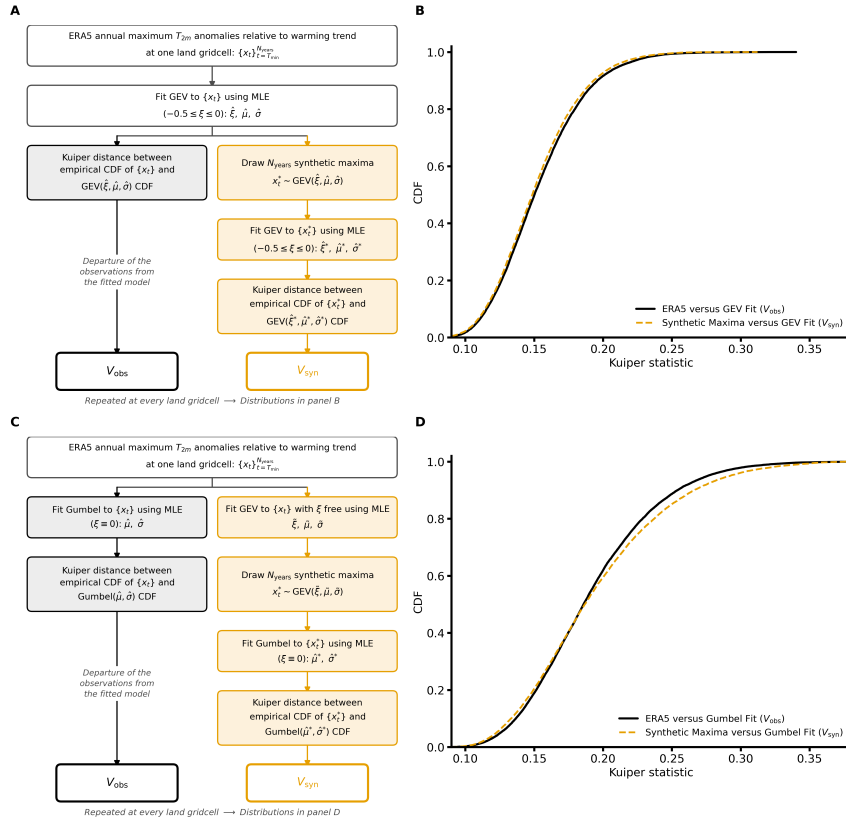


Figure S4. Schematic of Statistical Test of GEV Distribution and Pooled

Kuiper Statistic Distribution (Anomalies Relative to the Long-term Warming

Trend). **A** and **C** show schematic outlines of our statistical tests to verify our statistical

model. Note T_{min} is the first year in which we have data and N_{years} is the total number

of years in our dataset ($N_{years} = 45$ when $T_{min} = 1979$). In **B**, black solid lines show

the pooled Kuiper statistic CDF when the empirical CDF of ERA5 annual maximum

temperature anomalies at a given location are compared with the fitted GEV distribution

at a given location. The gold dashed line shows the Kuiper statistic CDF when we compare

synthetic samples from a fitted GEV distribution against a GEV distribution that is fit to

the synthetic data. In **D**, black solid lines show the pooled Kuiper statistic CDF when the

empirical CDF of ERA5 annual maximum temperature anomalies at a given location are

compared with the fitted Gumbel distribution at a given location. The gold dashed line

shows the Kuiper statistic CDF when we compare synthetic samples from a fitted GEV

distribution against a Gumbel distribution that is fit to the synthetic data. See *Can We*

Reject the GEV Distribution? in the main text for more information on our statistical

tests. Computations used anomalies in the annual maximum temperatures relative to the

long-term warming trend using data from 1979-2024.

September 9, 2026, 10:38pm

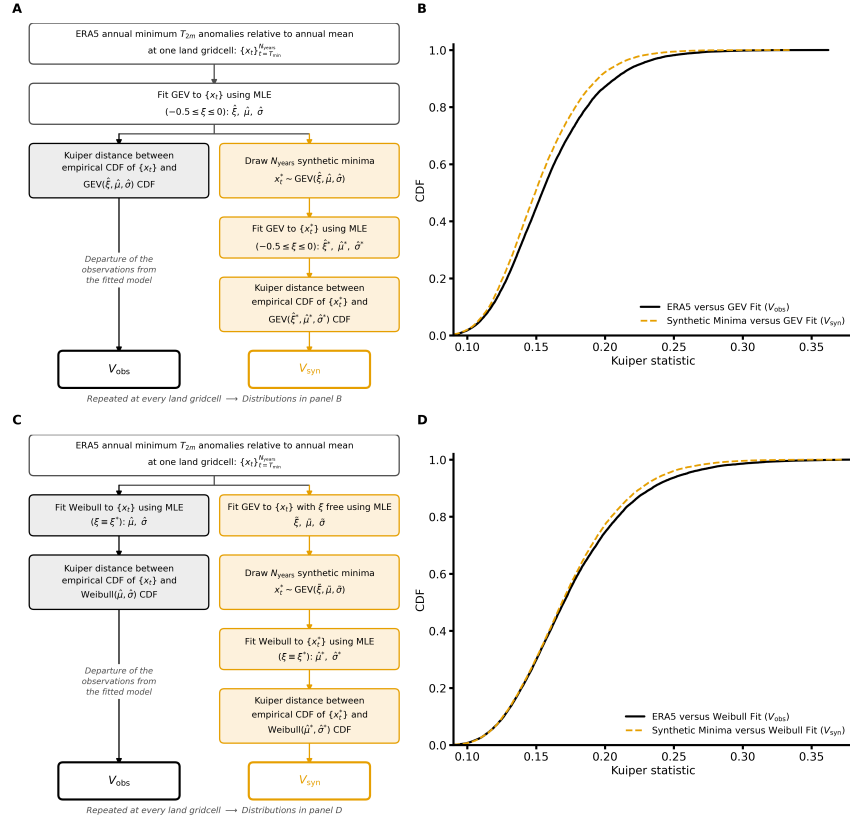


Figure S5. Pooled Kuiper Statistic Cumulative Distribution Functions

(CDFs) for Annual Minimum Temperature Anomalies over Land. A and C

show schematic outlines of our statistical tests to verify our statistical model. Note T_{min}

is the first year in which we have data and N_{years} is the total number of years in our

dataset ($N_{years} = 45$ when $T_{min} = 1979$). In B, black solid lines show the pooled Kuiper

statistic CDF when the empirical CDF of ERA5 annual minimum temperature anomalies

at a given location are compared with the fitted GEV distribution at a given location.

The gold dashed line shows the Kuiper statistic CDF when we compare synthetic samples

from a fitted GEV distribution against a GEV distribution that is fit to the synthetic

data. In D, black solid lines show the pooled Kuiper statistic CDF when the empirical

CDF of ERA5 annual minimum temperature anomalies at a given location are compared

with the fitted Weibull distribution at a given location with $\xi = -1/4$. The gold dashed

line shows the Kuiper statistic CDF when we compare synthetic samples from a fitted

GEV distribution against a Weibull distribution that is fit to the synthetic data. See *Can*

We Reject the GEV Distribution? in the main text for more information on our statistical

tests. Computations used anomalies in the annual minimum temperatures relative to the

minimum mean temperature in the same year using data from 1979-2024.

September 9, 2026, 10:38pm

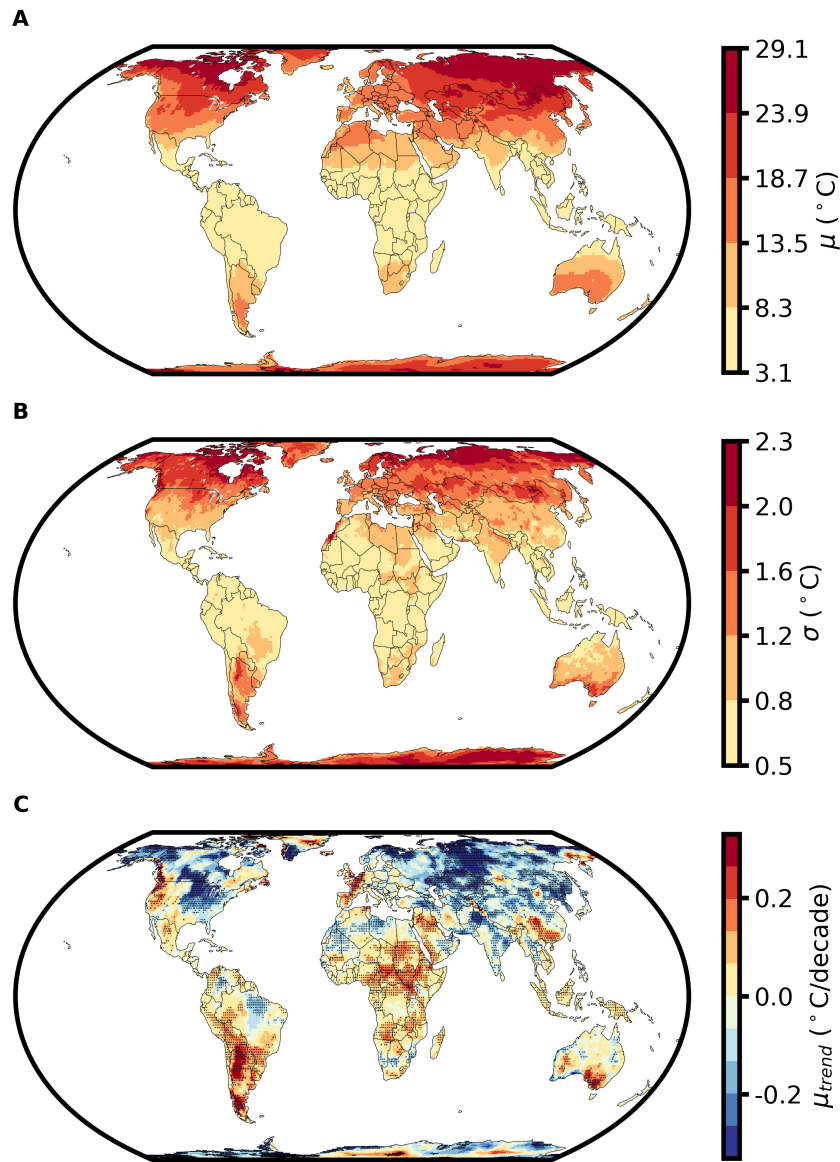


Figure S6. GEV Parameters (1950-2024 Analysis Period). **A** shows the location parameter in each grid cell over land for the stationary Gumbel fit. **B** is as **A** but for the scale parameter. **C** shows the location parameter trend when we include nonstationary effects in the GEV fit. The location parameter corresponds to the distribution mean, the scale parameter corresponds to the distribution variance, and the location trend tests for “tail-stretching” of the distribution. Stippling shows where we have 95% confidence in the sign of the trend. Temperature anomalies used are annual maxima relative to the annual mean in the same year from 1950-2024.

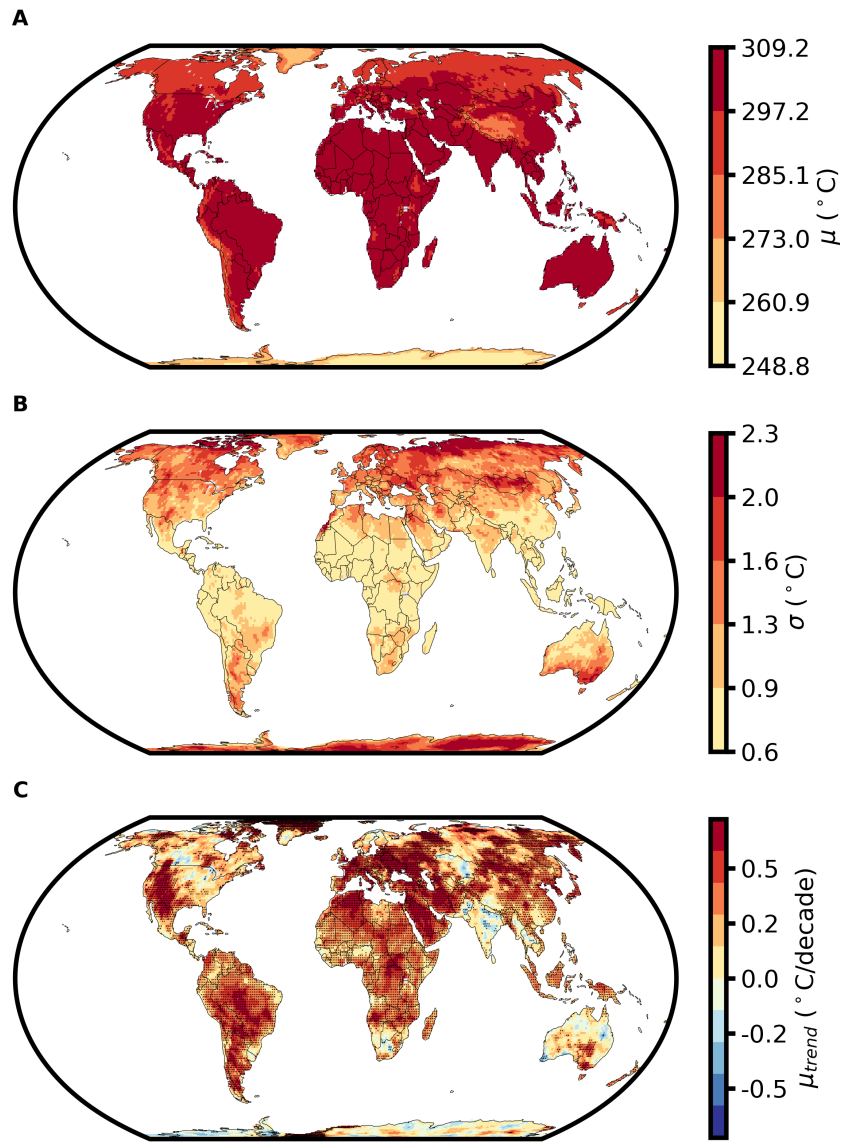


Figure S7. GEV Parameters (Raw Data). **A** shows the location parameter in each grid cell over land for the stationary Gumbel fit. **B** is as **A** but for the scale parameter. **C** shows the location parameter trend when we include nonstationary effects in the GEV fit. The location parameter corresponds to the distribution mean, the scale parameter corresponds to the distribution variance, and the location trend tests for “tail-stretching” of the distribution. Stippling shows where we have 95% confidence in the sign of the trend. Temperature data used are raw temperature values from 1979-2024.

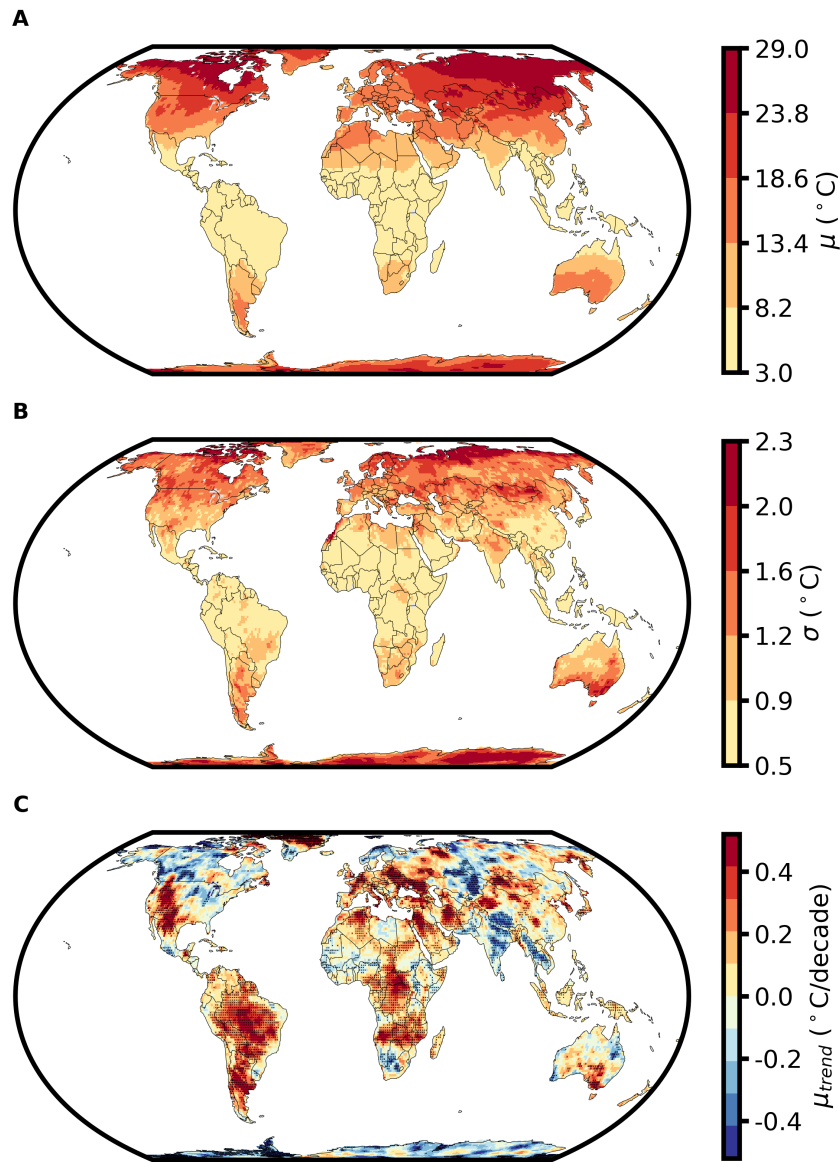


Figure S8. GEV Parameters (Anomalies Relative to Long-term Warming Trend). **A** shows the location parameter in each grid cell over land for the stationary Gumbel fit. **B** is as **A** but for the scale parameter. **C** shows the location parameter trend when we include nonstationary effects in the GEV fit. The location parameter corresponds to the distribution mean, the scale parameter corresponds to the distribution variance, and the location trend tests for “tail-stretching” of the distribution. Stippling shows where we have 95% confidence in the sign of the trend. Temperature anomalies used are annual maxima relative to the long-term warming trend from 1979-2024.

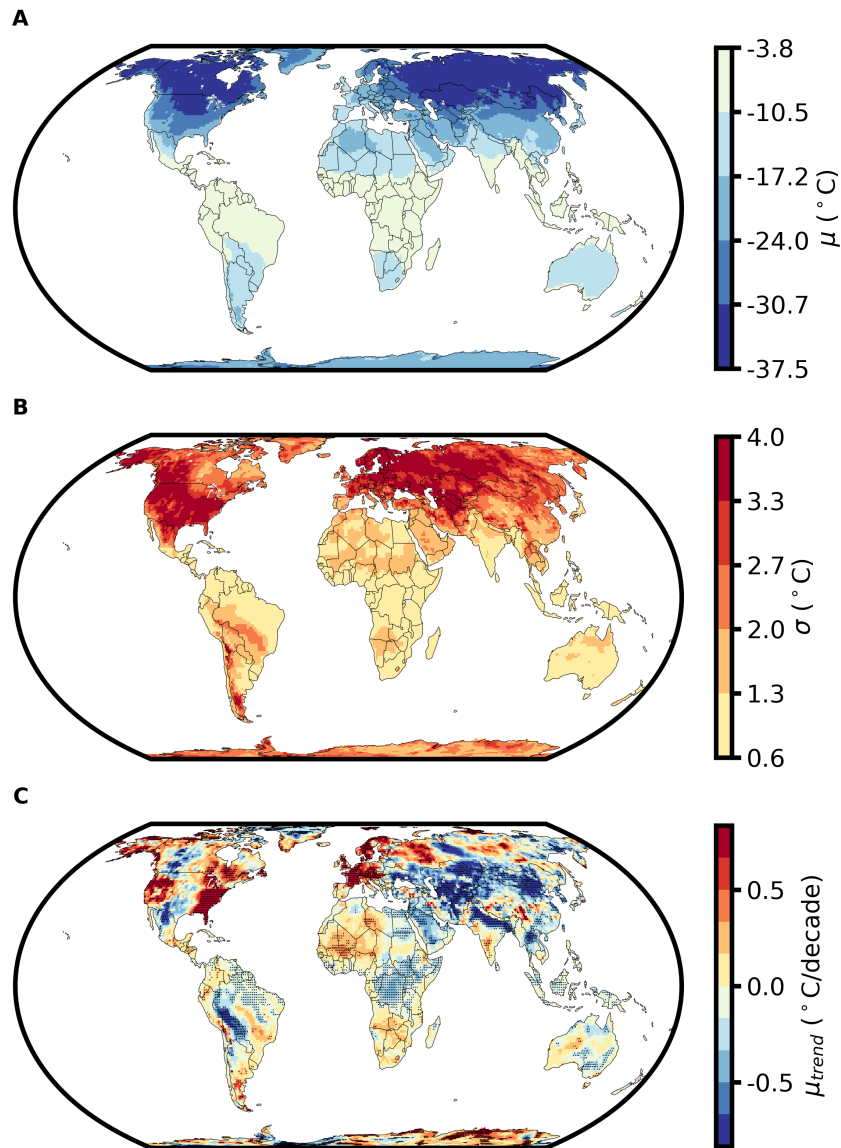


Figure S9. GEV Parameters (Annual Minimum Temperatures). **A** shows the location parameter in each grid cell over land for the stationary Gumbel fit. **B** is as **A** but for the scale parameter. **C** shows the location parameter trend when we include nonstationary effects in the GEV fit. The location parameter corresponds to the distribution mean, the scale parameter corresponds to the distribution variance, and the location trend tests for “tail-stretching” of the distribution. Stippling shows where we have 95% confidence in the sign of the trend. Temperature anomalies used are annual minima relative to the annual mean in the same year from 1979-2024.

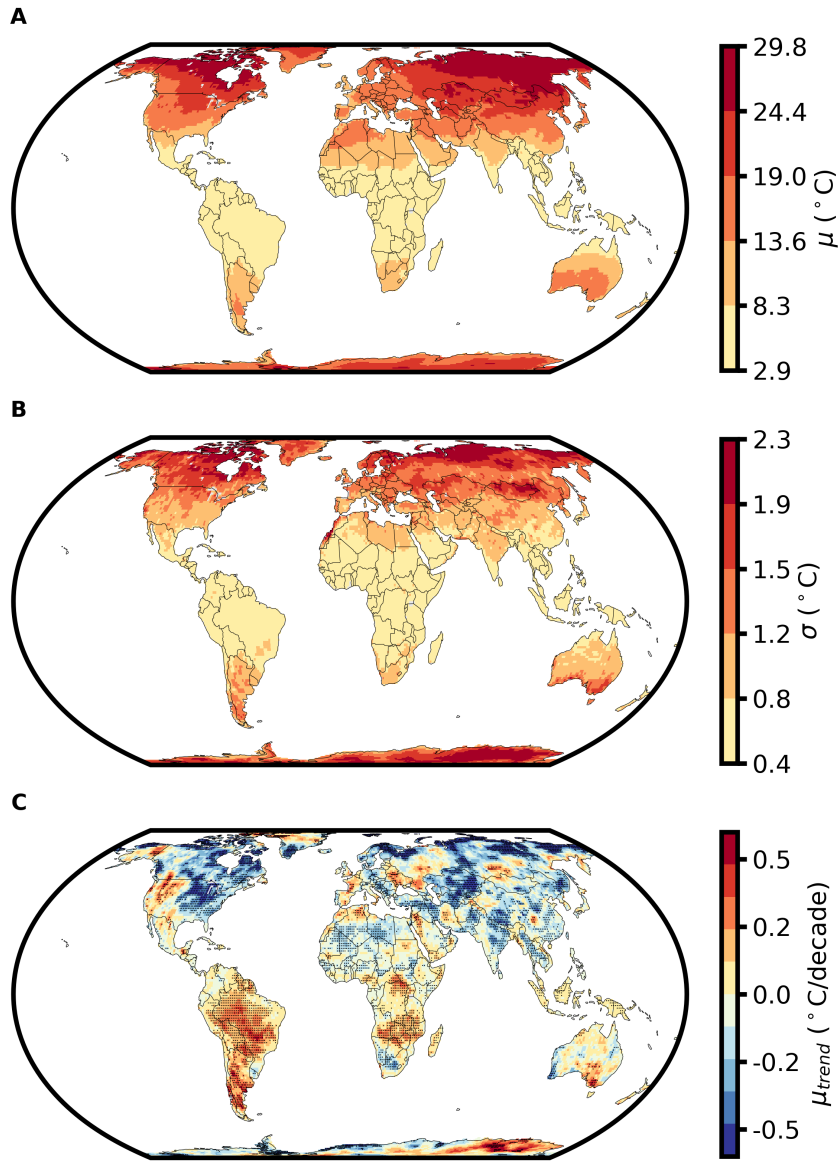


Figure S10. GEV Parameters (Anomalies Relative to Annual Mean, Non-stationary Fit Only). A, B, and C show the location parameter, scale parameter, and location parameter trend when we include nonstationary effects. Results are not materially impacted relative to Figure 2 in the main text. The location parameter corresponds to the distribution mean, the scale parameter corresponds to the distribution variance, and the location trend tests for “tail-stretching” of the distribution. Stippling shows where we have 95% confidence in the sign of the trend. Temperature anomalies used are annual maxima relative to the annual mean in the same year from 1979-2024.

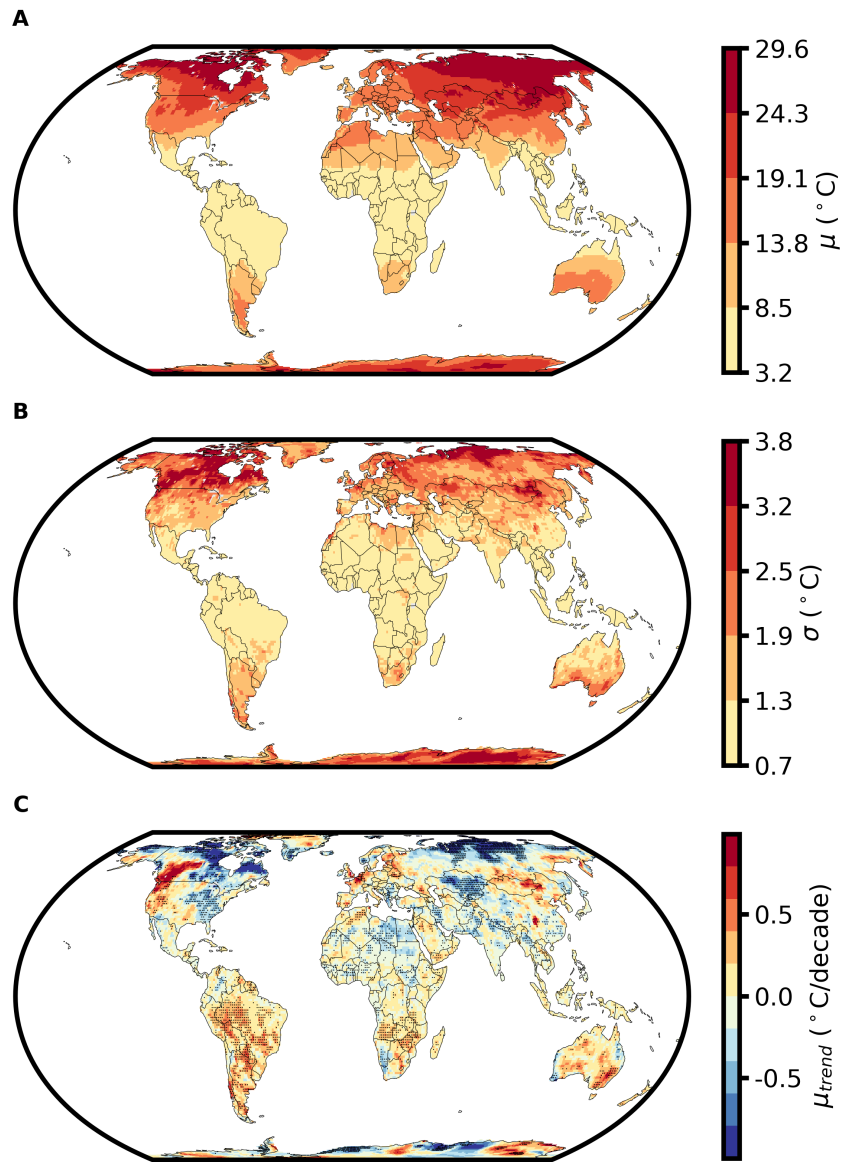


Figure S11. GEV Parameters ($\xi = -0.5$). **A** shows the location parameter in each grid cell over land for the stationary fit where $\xi = -0.5$. **B** is as **A** but for the scale parameter. **C** shows the location parameter trend when we include nonstationary effects in the GEV fit. The location parameter corresponds to the distribution mean, the scale parameter corresponds to the distribution variance, and the location trend tests for “tail-stretching” of the distribution. Stippling shows where we have 95% confidence in the sign of the trend. Temperature anomalies used are annual maxima relative to the annual mean in the same year from 1979-2024.

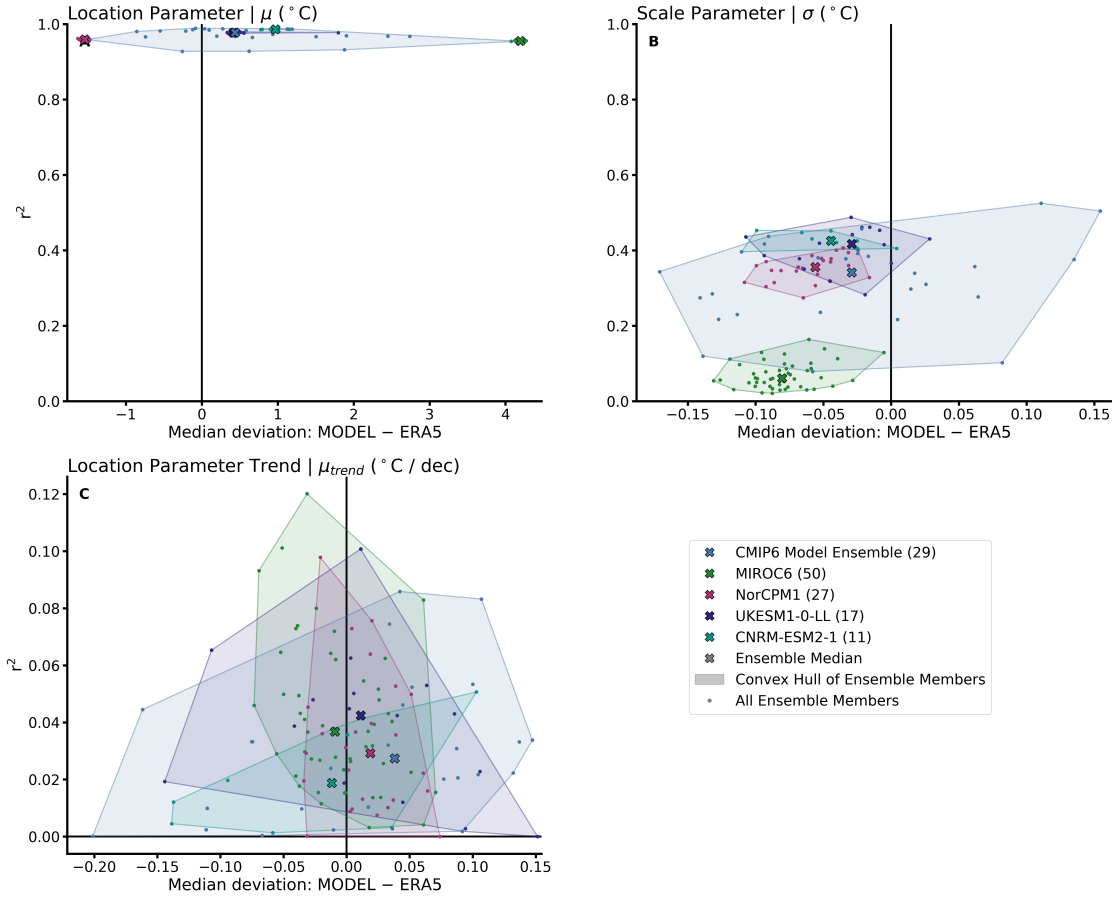


Figure S12. Correlation between CMIP and ERA5 GEV Parameters across CMIP Models (Raw Data). Each panel shows the correlation between a CMIP6 models and ERA5 reanalysis GEV parameter (see the title) in each grid box plotted against the median deviation between the model and reanalysis GEV parameter in each grid box. As in Figure 2 in the main text, the location and scale parameters plotted in panels **A** and **B** show the results when no nonstationary effects are included, and panel **C** shows the trend in the location parameter when these effects are included. Note we use one CMIP model realization for each data point. Outliers are excluded for visual clarity (CESM-WACCM is not shown in all panels; in **b** MIROC-ES2L is not shown; in **C** HadGEM3-GC31-LL is not shown). Temperature anomalies used are raw values from 1979-2024.

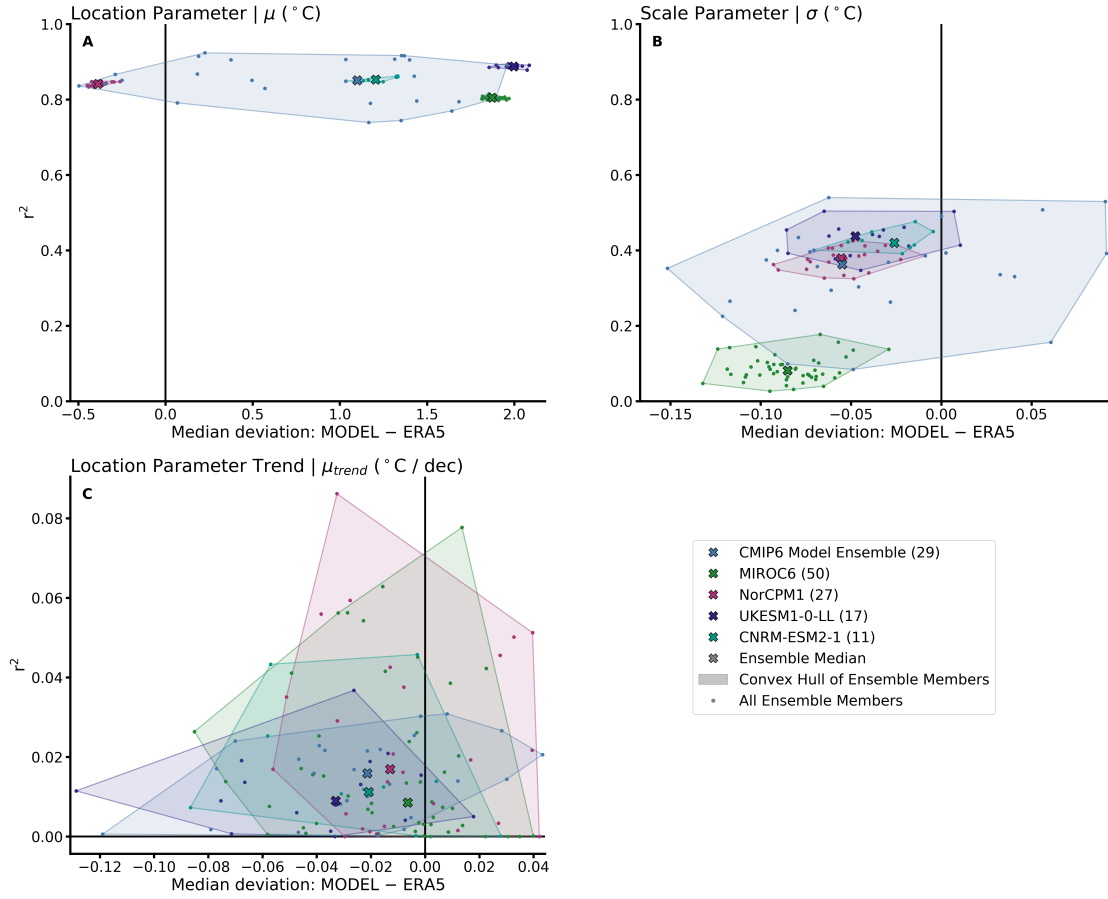


Figure S13. Correlation between CMIP and ERA5 GEV Parameters across CMIP Models (Anomalies Relative to Long-term Warming Trend). Each panel shows the correlation between a CMIP6 models and ERA5 reanalysis GEV parameter (see the title) in each grid box plotted against the median deviation between the model and reanalysis GEV parameter in each grid box. As in Figure 2 in the main text, the location and scale parameters plotted in panels **A** and **B** show the results when no nonstationary effects are included, and panel **C** shows the trend in the location parameter when these effects are included. Note we use one CMIP model realization for each data point. Outliers are excluded for visual clarity (CESM-WACCM is not shown in all panels; in **A** KACE-1-0-G is not shown; in **B** MIROC-ES2L is not shown; in **C** EC-Earth3-Veg is not shown). Temperature anomalies used are annual maxima relative to the long-term warming trend from 1979-2024.

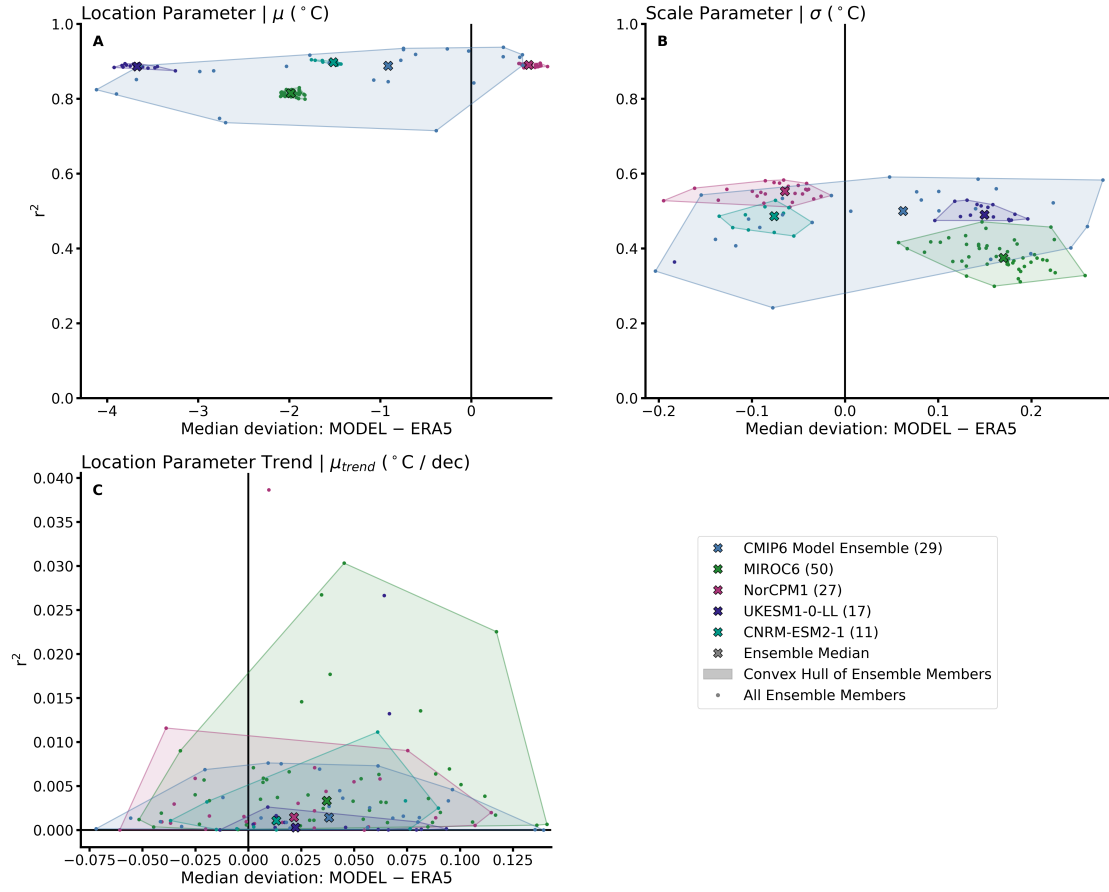


Figure S14. Correlation between CMIP and ERA5 GEV Parameters across CMIP Models (Annual Minima Relative to Annual Mean). Each panel shows the correlation between a CMIP6 models and ERA5 reanalysis GEV parameter (see the title) in each grid box plotted against the median deviation between the model and reanalysis GEV parameter in each grid box. As in Figure 2 in the main text, the location and scale parameters plotted in panels **A** and **B** show the results when no nonstationary effects are included, and panel **C** shows the trend in the location parameter when these effects are included. Note we use one CMIP model realization for each data point. Outliers are excluded for visual clarity (in **A** KIOST-ESM is not shown; in **B** IITM-ESM and KIOST-ESM are not shown and one member of UKESM1-0-LL is not included in the convex hull; in **C** 2 UKESM1-0-LL ensemble members and one NorCPM1 member is not included in the convex hull). Temperature anomalies used are annual minima relative to the annual mean in the same year from 1979-2024.

September 9, 2026, 10:38pm

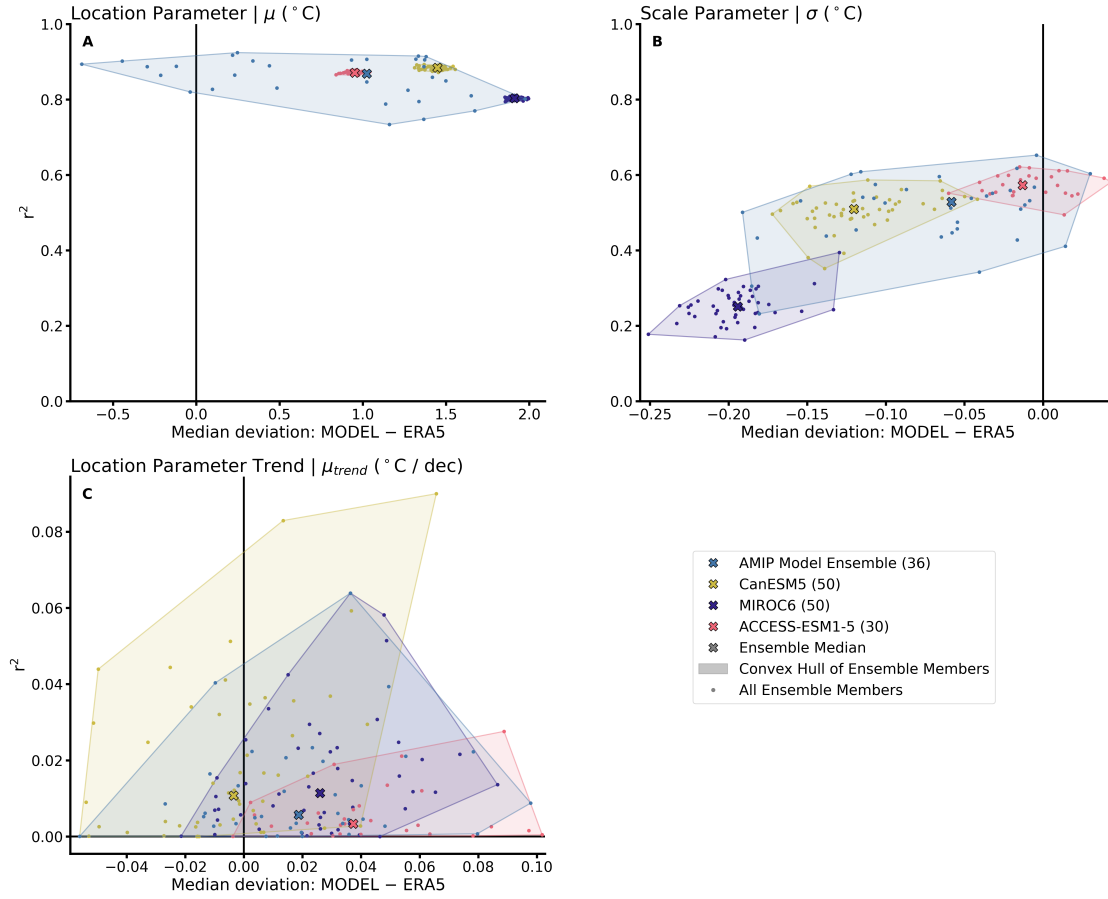


Figure S15. Correlation between AMIP and ERA5 GEV Parameters across AMIP Models. Each panel shows the correlation between a AMIP models and ERA5 reanalysis GEV parameter (see the title) in each grid box plotted against the median deviation between the model and reanalysis GEV parameter in each grid box. As in Figure 2 in the main text, the location and scale parameters plotted in panels **A** and **B** show the results when no nonstationary effects are included, and panel **C** shows the trend in the location parameter when these effects are included. Outliers are excluded for visual clarity (in **A** KACE-1-0-G, MPI-ESM-1-2-HAM, FGOALS-f3-L, and KOIST-ESM are not shown; in **B** GFDL-ESM4 and IPSL-CM5A2-INCA are not shown and one member of the ACCESS-ESM1-5 ensemble is not included in the convex hull; in **C** one member of the ACCESS-ESM1-5 ensemble is not included in the convex hull). Temperature anomalies used are annual maxima relative to the annual mean in the same year from 1979-2024.

September 9, 2026, 10:38pm

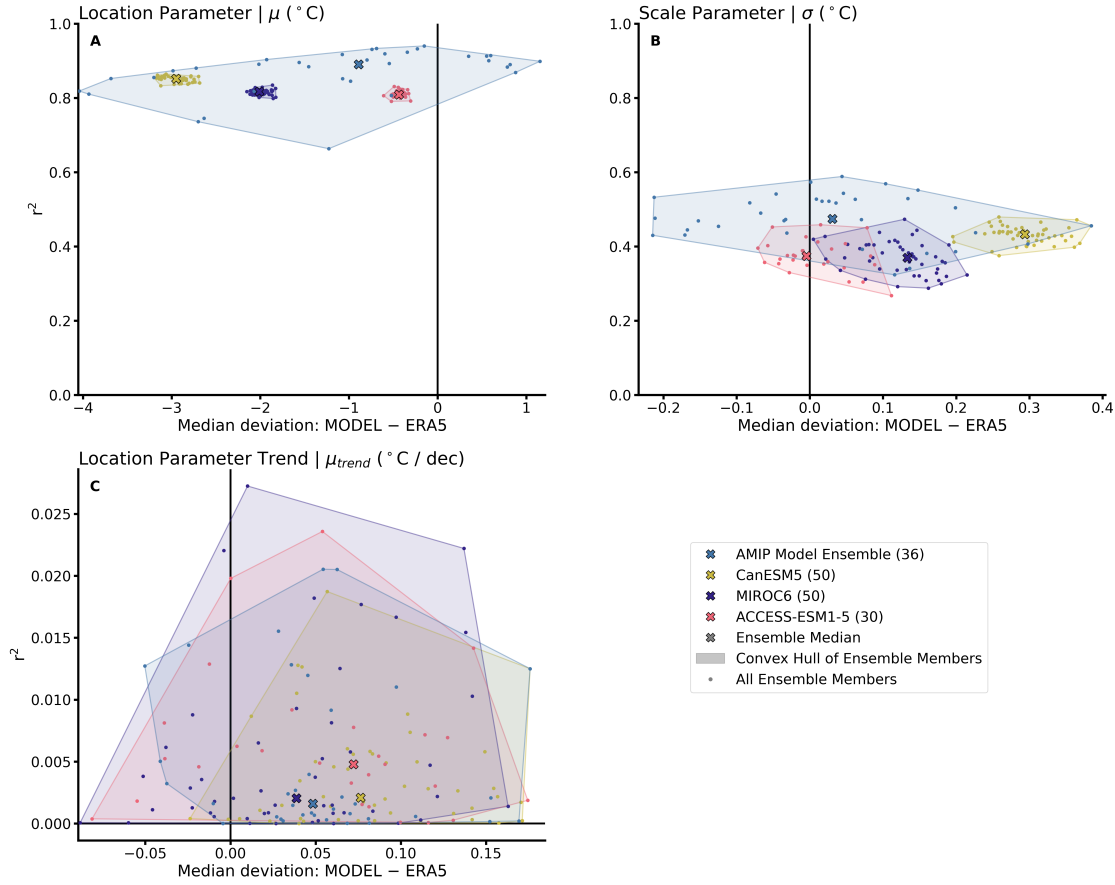


Figure S16. Correlation between AMIP and ERA5 GEV Parameters across AMIP Models (Annual Minima Relative to Annual Mean). Each panel shows the correlation between a AMIP models and ERA5 reanalysis GEV parameter (see the title) in each grid box plotted against the median deviation between the model and reanalysis GEV parameter in each grid box. As in Figure 2 in the main text, the location and scale parameters plotted in panels **A** and **B** show the results when no nonstationary effects are included, and panel **C** shows the trend in the location parameter when these effects are included. Outliers are excluded for visual clarity (in **A** KOIST-ESM and MPI-ESM-1-2-HAM is not shown; in **B** IITM-ESM and KOIST-ESM are not shown; in **C** one MIROC6 and one ACCESS-ESM1-5 members are not included in the convex hulls). Temperature anomalies used are annual minima relative to the annual mean in the same year from 1979-2024.