Environmental Research: Climate

Crossmark

RECEIVED
31 Oct 2025

REVISED
dd Month yyyy

PAPER

When Climate Datasets Are Too Short: A Practical Guide for Bias and Uncertainty in Determining the Probability of Rare Events using GEV Distributions

Ronak N. Patel[1*] and Tapio Schneider[1]

[1]Department of Environmental Science and Engineering, California Institute of Technology, Pasadena, CA, USA
[*]Author to whom any correspondence should be addressed.

E-mail: ronak@caltech.edu

Keywords: climate extremes, GEV distribution, sampling uncertainty, climate risk assessment

Abstract

Quantifying the probability of rare climate extremes is essential for risk assessment, infrastructure design, and adaptation planning. Generalized extreme value (GEV) distributions provide a widely used statistical framework for estimating the return levels of low-probability, high-impact events, but such estimates are often derived from limited observational or simulated records. Using 13,000 years of idealized aquaplanet simulations, we show that short time series (<30 years) generally lead to substantial underestimation of extreme precipitation and temperature return levels (through underestimation of the GEV shape parameter ξ). The minimum length required for unbiased estimation varies by variable and latitude, but generally lies between 30 and 50 years. Beyond this threshold, ensembles of multiple segments can be used to precisely and unbiasedly quantify uncertainty in rare-event statistics. However, even when total data volume is large, ensembles composed of shorter segments retain substantial bias. Applying this framework to a 500-year GFDL-ESM4 simulation reveals similar behavior, with the shape and scale parameters as key predictors of return level uncertainty and bias. By linking theoretical bias behavior to physically interpretable climate features, this work provides practical guidance for designing and analyzing datasets used in climate hazard assessment. Long-duration ensemble members or observational segments are necessary for robust GEV-based estimates of low-probability, high-impact events.

1 Introduction

Extreme precipitation and temperature events pose significant risks to human lives, infrastructure, agriculture, and broader economic systems. Uncertainty in the likelihood of such extremes has direct implications for climate adaptation, flood and disaster risk management, insurance, and resilient infrastructure planning (1). Given the limited length of observational records, estimating the frequency of low-probability, high-impact events remains a considerable challenge, particularly in regions where such events have yet to be observed due to short data records.

Generalized Extreme Value (GEV) models (2) are widely used to study extreme events across a range of applications from heavy precipitation (e.g., 3; 4; 5; 6) and extreme heat (e.g., 7; 8), to hail (e.g., 9) and even financial risk modeling (e.g., 10). The GEV framework offers a theoretical basis for estimating both the frequency and magnitude of rare events, enabling extrapolation into the far tails of a distribution beyond what has been historically observed (11). Hydrologists and climate scientists frequently use GEV distributions to estimate return periods for precipitation and temperature extremes and to assess how their frequencies may change under future warming (5). For risk managers more broadly, it provides a quantitative foundation for anticipating rare but high-impact events when data are limited. Yet, despite their widespread use, the reliability of GEV-based estimates depends critically on the amount of data available and the length of each continuous time series segment used to estimate GEV parameters.

Previous studies have highlighted that short observational records often produce systematic biases, particularly in the shape parameter ξ, which governs tail behavior and thus the assessed probability of extreme events (e.g., 12; 13; 14). Long model simulations have also been used to

empirically assess the spread in return level estimates arising from limited samples; for example, (15) used 1000-year runs to examine uncertainties in 100-year return level estimates of temperature based on 20- and 50-year subsets. While such studies have demonstrated the existence and direction of these biases, they have neither explicitly connected them to the theoretical framework describing how block maxima converge to their asymptotic distribution nor examined how they vary systematically across latitude and different climate relevant variables.

Here, we show that the minimum segment length required for unbiased inference depends on climate regime and variable, but generally requires more than 30-50 years in a record. We diagnose how finite-sample biases in GEV parameters vary across latitude for both temperature and precipitation, link these patterns to theoretical expectations, and assess how long of a block size is needed to obtain stable and unbiased return level estimates. This bridges the gap between statistical theory and applied climate risk modeling, offering both conceptual and practical insights for how GEV-based estimates can misrepresent risk when data are limited.

We use long, stationary simulations of temperature and precipitation from two climate models: a 250-year aquaplanet simulation with an idealized atmosphere model, and a 500-year preindustrial control run of the GFDL-ESM4 earth system model. The aquaplanet's zonal symmetry allows us to effectively create multi-millennial datasets, enabling quantification of sampling biases in GEV parameters and their propagation into return level estimates. The GFDL-ESM4 control simulation provides a more realistic testbed of these GEV parameter influences in the presence of more realistic internal climate variability and spatial heterogeneity.

Our results demonstrate that short, stationary records (fewer than ∼30 years) systematically underestimate the GEV shape parameter for many types of variables, leading to underestimation of the 100 and 1000-year return levels. Once record lengths exceed approximately 30 years, biases in the shape parameter diminish, return level estimates are unbiased, and bootstrap methods or ensembles can provide a robust way to estimate return level uncertainty since the resulting ensemble will also be unbiased.

By quantifying the relationship between record length, parameter bias, and return level uncertainty, this study provides actionable guidance for practitioners estimating rare events from short datasets. We show that temperature extremes are generally underestimated while precipitation extremes can be overestimated in short records, reflecting the influence of the shape parameter. Negative ξ values (bounded tails) tend to produce underestimation, whereas positive ξ values (heavy tails) can lead to overestimation. Although these directional patterns may be expected from the mathematical literature (e.g., 16; 17; 18), connecting these theoretical expectations to empirical results across idealized and complex models provides a physically interpretable framework for understanding how and why GEV-based estimates can fail. Recognizing how the sign of ξ shapes the biases allows practitioners to anticipate and correct for systematic errors, supporting more robust infrastructure design, hazard planning, and financial risk management under limited data conditions.

2 Data and Methods

We evaluate return levels for extreme temperature and precipitation using a long-run idealized aquaplanet simulation as well as data from the GFDL-ESM4 preindustrial control (picontrol) simulation (19) to determine the effect of short time series on the estimation of rare event probabilities.

2.1 Model Data

2.1.1 Idealized Aquaplanet Model We use an idealized general circulation model (GCM) similar to that described in (20), which involves an aquaplanet simulation at T42 resolution (approximately 2.8 degree resolution). We modify the model in (20) by using a 30 m slab ocean and 10 vertical levels. The model is based on the Geophysical Fluid Dynamics Laboratory (GFDL) Flexible Modeling System (FMS) software framework.

The model is initialized from an isothermal rest state and integrated for 250 years, with the first 2 years discarded as spin-up. The remaining 248 years are used for analysis. During this period, the model reaches a statistically stationary state, meaning the daily precipitation and temperature time series exhibit stable mean, variance, and autocorrelation over time. To approximate present-day climate conditions, the global mean temperature at the lowest model level (954 hPa) is 284 K, consistent with Earth's current near-surface climate. The total precipitation field used in this study is defined as the sum of parameterized and grid-scale precipitation. The updated model's global-mean precipitation rate of 4.1 mm day^{-1} closely matches the 4.3 mm day^{-1} reported in (20), despite slight differences in model configuration.

To robustly estimate probabilities of extreme events without incurring excessive computational costs, we take advantage of the aquaplanet's symmetry and statistical properties to effectively extend the available data record. Temperature and precipitation fields in this idealized model exhibit zonal and hemispheric symmetry, enabling us to pool independent data from both hemispheres and multiple longitudes to better estimate extreme value statistics. To ensure independence among concatenated segments, we first estimate the longitudional decorrelation scales: approximately 1500 km for temperature and around 50 km for the more heterogeneous precipitation fields in midlatitudes. Using this, we construct extended stationary time series (of length 12,896 years) by concatenating temperature and precipitation data from every 5th longitude (approximately 1500 km apart at T42 resolution, which has a 64×128 latitude/longitude transform grid) and corresponding latitudes from opposite hemispheres. For example, precipitation data from 40.46°N 0.00°E, 40.46°N 14.06°E, 40.46°N 28.12°E, 40.46°N 42.18°E, 40.46°N 56.25°E, ..., 40.46°S 0.00°E, 40.46°S 14.06°E, 40.46°S 28.12°E,..., etc. are combined into a single extended record that leverages the model's zonal and hemispheric symmetry. This design allows us to treat the simulation as a statistically controlled experiment for testing how GEV parameter estimates converge with increasing sample length.

2.1.2 GFDL ESM-4 Model While the aquaplanet simulation provides a long, statistically stationary dataset for exploring EVT, it is inherently idealized—it lacks regional variability, a representation of land, and low-frequency interannual variability. To assess extreme event statistics in a more realistic climate context, we also analyze the *esm-piControl* simulation from the GFDL-ESM4 Earth System Model (19), part of the Coupled Model Intercomparison Project Phase 6 (CMIP6). As part of the CMIP6 design, preindustrial control experiments are run to study naturally occurring, unforced variability of the preindustrial climate system (21). The esm-piControl experiment sets atmospheric CO_2 concentrations to 1850 levels and then it evolves naturally via the coupled carbon cycle, without any anthropogenic emissions, for 500 years (21). The result is 500 years of stationary climate data free of long-term warming trends but fully representing internally generated variability, produced by a fully coupled model that includes interactive ocean, sea ice, land, and atmosphere components. This enables us to apply EVT methods to a more realistic climate system with internal variability.

2.2 Extreme Value Theory Methods

Extreme Value Theory (EVT) provides a statistical framework for studying extreme events across a wide range of disciplines, including climate science. EVT provides a class of models that allow us to estimate the probability of extreme events beyond the observational record (11; 22). We use the classical block maxima formulation of EVT to describe the statistical behavior of annual extremes.

Suppose we have a sequence of daily precipitation observations $X = \{X_1,\ X_2,\ X_3\ , ...\}$ grouped into blocks of length m. Within each block, we extract the maximum single-day precipitation amount. For blocks of length one year, this results in a dataset of annual maxima. According to the Fisher–Tippett–Gnedenko theorem (23; 24), as the block size m increases, the distribution of block maxima can only converge to one of three possible limiting extreme value distributions, no matter the distribution of the underlying process (25; 26): Gumbel (Type I), Frechet (Type II), and reversed-Weibull (Type III). These three distributions can be combined into a single Generalized Extreme Value (GEV) distribution, with probability density function (pdf)

$$g(x) = \frac{1}{\sigma}\left[1 + \frac{\xi(x-\mu)}{\sigma}\right]^{-1-\frac{1}{\xi}} \exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right) \quad \text{for} \quad 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0 \tag{1}$$

and cumulative distribution function (cdf)

$$G(x) = \exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right) \tag{2}$$

Here, x are the block maxima, μ is the location parameter, σ is the scale parameter, and ξ is the shape parameter. Because our goal is to assess finite-sample behavior rather than propose alternative estimators, we use maximum likelihood estimation (MLE) throughout.

The location parameter is a measure of the center of the distribution (e.g., Figure 1a). The scale parameter is proportional to the standard deviation of the distribution (e.g., Figure 1b). The shape parameter describes the tail behavior of the GEV distribution and of the underlying process (e.g., Figure 1c) and is critical for risk assessment. For $\xi \approx 0$, we have a Type I distribution with

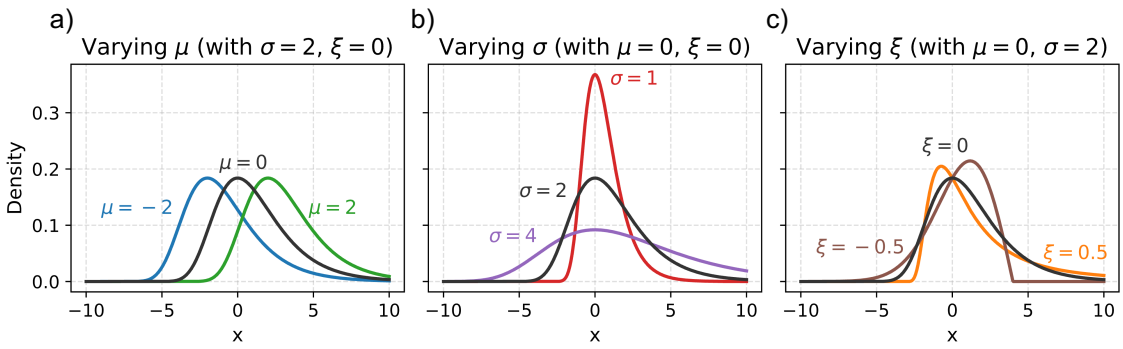

Figure 1. (a) A GEV distribution where the scale and shape parameters are fixed and the location parameter varies. (b) Same as (a) but the scale parameter varies. (c). Same as (a) but the shape parameter varies. The black distribution is a GEV distribution with $\mu = 0$, $\sigma = 2$, and $\xi = 0$, and it is the same for all three subplots.

exponentially decaying tails. A Type II distribution with $\xi > 0$ has polynomially decaying ("fat") tails, whereas a Type III distribution with $\xi < 0$ is associated with bounded observations X_i—and therefore bounded block maxima. For bounded distributions, like those often observed for temperature (e.g., 27), the upper bound is equal to $\mu - \sigma/\xi$. When $\xi \geq 0.5$, the variance becomes undefined, the implications of which will be described later.

A key application of EVT is estimating return periods, which represent the expected time interval between extreme events of a given magnitude or greater (26). We calculate the return period in years using the cdf as follows:

$$R(x) = \frac{1}{\omega\left[1 - G(x)\right]}, \tag{3}$$

where ω is the sampling frequency, or the number of blocks per year.

In addition to computing the return period, we can compute the return level, which represents the average event magnitude exceeded on average once every T years:

$$z_T = \mu + \frac{\sigma}{\xi}\left[\left(-\ln\left(1 - \frac{1}{T}\right)\right)^{-\xi} - 1\right] \quad \text{for} \quad \xi \neq 0. \tag{4}$$

Evaluating Eq. (3) at a set of precipitation exceedance levels provides the return period for each, and it can be used to construct a return level curve. Alternatively, one can directly evaluate Eq. (4) across a range of return periods T to obtain the same return level curve. By fitting a GEV distribution, we can now explicitly calculate the probabilities of events even more extreme than those observed in the time series.

In addition to the block maxima approach, EVT can also be applied using Peaks-Over-Threshold (POT) methods. Using only block maxima loses data, as there may be extreme events in a one-year block that were not the annual maximum, but are still considered "extreme." If we choose a large enough threshold, u, the distribution of values $(X - u)$, called "excesses," conditional on $X > u$ is approximately distributed as a Generalized Pareto Distribution (GPD) with cdf

$$H(x) = 1 - \left(1 + \frac{\xi(x - u)}{\sigma^*}\right)^{-1/\xi} \tag{5}$$

Similar to the GEV distribution, one can also construct a return level curve using the cdf of the GPD distribution (11), which has the form

$$z_{Tu} = u + \frac{\sigma^*}{\xi}\left((T\lambda)^{\xi} - 1\right) \quad \text{for} \quad \xi \neq 0, \tag{6}$$

where $\sigma^* = \sigma + \xi(u - \mu)$, u is the chosen threshold, and λ is the average annual exceedance rate. The parameter ξ in Eqs. (5) and (6) is equal to that of the corresponding GEV distribution. Similarly, the scale parameter of the GPD distribution σ^* is related to the parameters of the GEV distribution (11).

While POT methods might initially seem more relevant for risk assessment than block maxima, there are a few properties that make this difference less important when looking at return levels for

rare events such as those with return periods of 100 or 1000 years. When modeling peaks over thresholds, we now have to consider the clustering of extreme values, and it would require the estimation of an extremal index to statistically decluster the data (28). This provides another source of uncertainty. Importantly, for rare events, the return level curves provided by GEV and GPD in Eqs. (4) and (6), respectively, are asymptotically equivalent (18). This is the case when:

- The GEV block sizes m are large
- The return period is long: $T \to \infty$
- The threshold is chosen so that the same class of events is studied: $u \approx \mu$

2.3 Comparing Block Maxima and Threshold-Based EVT Methods

To assess the extreme precipitation statistics in an idealized aquaplanet simulation, we apply two classical approaches from extreme value theory: the Generalized Extreme Value (GEV) distribution, which models block maxima, and the Generalized Pareto distribution (GPD), which models exceedances above a high threshold.

To establish a reference, we first confirm that the two canonical EVT approaches produce consistent estimates of rare-event behavior in the idealized aquaplanet. Figure 2 compares these methods in the idealized aquaplanet simulation using the nearly 13,000-year concatenated time series of daily precipitation at 40.46° North and South. A GEV distribution is fit to the annual maxima (Figure 2a), and the probability density function, cumulative density function, and return level curves are generated (Figure 2c–e). Alternatively, by selecting a threshold of 28.5 mm—approximately the 99.8th percentile of the time series—a GPD is fit to the daily precipitation values exceeding this level (Figure 2b) and the other curves are also generated (Figure 2c–e).

The return level curves generated by these two methods look quite similar, which is expected, as both methods characterize the upper tail, or extremes, of the same underlying distribution. Figure 2c–e overlays the GEV and GPD fits. The probability density functions (Figure 2c) and cumulative density functions (Figure 2d) confirm that the annual maximum precipitation distribution (blue) closely resembles the tail of the full daily precipitation distribution (red). The return level curves (Figure 2e) show that for return periods greater than approximately 6 years, the two methods yield almost identical return levels. The return periods for which these two methods converge is slightly longer if a different threshold is used. We see this convergence on similar timescales across all latitudes in this aquaplanet simulation. While both approaches are theoretically justified, because the block maxima approach requires fewer assumptions (e.g., no declustering or threshold selection), we use the GEV framework in all subsequent analyses.

3 Results

3.1 How Limited Data Distort Perceived Extreme Event Probability

While the aquaplanet simulation offers nearly 13,000 years of data, real-world records are far shorter, often limited to just a few decades. To understand how the choice of segment length affects estimated risk, we emulate realistic data constraints by subsampling shorter blocks (e.g., 10, 30, and 100 years) from the 13,000-year aquaplanet simulation and fit a GEV distribution to each block's annual maxima. These subsamples represent either single short records or individual ensemble members of a multi-realization experiment. For simplicity, these slices are non-overlapping and continuous, but this method can be expanded to looking at time series generated from a moving block bootstrap method (29).

Using each of these slices of data, we can construct a GEV distribution of annual maxima, and then use that to fit a return level curve. The sample of return level curves empirically estimates the uncertainty in return levels for a given return period. Even though short records may technically allow for the application of EVT, the results do not guarantee reliability for the probability of rare events, especially when extrapolating an order of magnitude beyond the record length.

Figure 3 illustrates how block length affects estimated return level curves. With just 10 years of data in a block (Fig. 3a), the median return level curve, defined as the pointwise median return level at each return period, does not follow that of the full 13,000-year dataset (in red). It tends to underestimate the return level for all return periods greater than approximately 10 years. There is also substantial spread for return periods longer than about 10 years, as reflected by the wide confidence intervals. For instance, the interquartile range (IQR) for the 100-year return level is 13.0 mm, and it nearly doubles to 28.6 mm for the 1000-year return level—including an upper bound that is unlikely to be physically realizable.

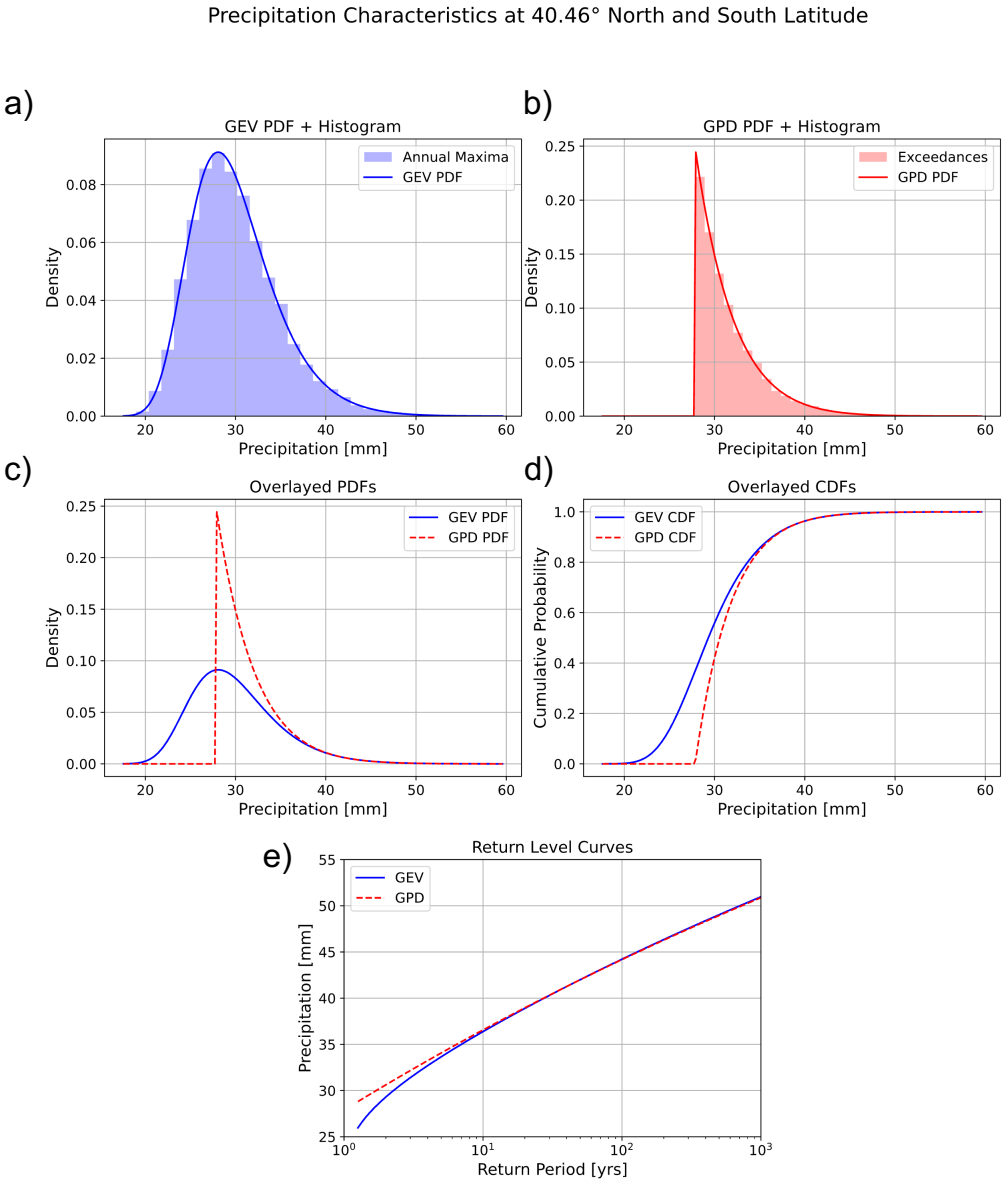

Figure 2. Distributions based upon 12,896 years of data from 40.46° North and South latitude in the aquaplanet. a) Histogram of annual maximum daily precipitation and the fitted Generalized Extreme Value (GEV) distribution. (b) Histogram of daily precipitation greater than 28.4mm and a fitted Generalized Pareto Distribution (GPD). (c) The probability density functions (PDFs) from panels (a) and (b) overlaid. (d) The cumulative density functions (CDFs) from (a) and (b) overlaid. (e) The return level curves constructed for the GEV distribution and the GPD overlaid.

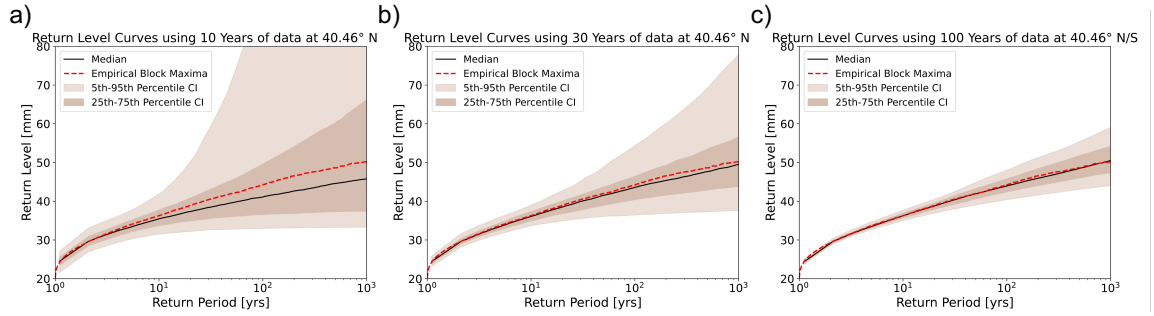

Figure 3. Return-level curves for annual maximum daily precipitation at 40.46 degrees using (a) 10 year (b) 30 year and (c) 100 year slices of the entire 12,896 year precipitation time series. The black line indicates the pointwise median of the return-level curves. The dashed red line indicates the return-level curve generated from the entire time series. Dark brown shading indicates the 25th to 75th percentile. Light brown shading indicates the 5th to 95th percentiles.

While it may be clear that one should not attempt to make statements about 100- or 1000-year events using a single 10 year block of data, we show that even an ensemble of 10 year blocks still systematically underestimates extreme event probabilities. These errors are greatly improved when using 30 years blocks (Fig. 3b), where individual blocks have less error and the aggregate of many 30 year blocks is also less biased. The median return level curves still slightly underestimate the magnitude of 100- and 1000-year events, but as expected, increasing the data length reduces this uncertainty – both the IQR and 5th–95th percentile ranges narrow significantly. Subsampling the time series into 100 year blocks further reduces the bias and uncertainty (Fig. 3c). The IQR for the 1000-year return level is 12.8 mm with 30-year blocks and 6.8 mm with 100-year blocks.

In addition to Figure 3, which subdivides the same time series into different numbers of 10, 30, and 100 year blocks, we can systematically test the effects of sample length versus the number of samples. Figure 4 demonstrates this explicitly by fixing the number of samples and varying block size (left to right) and fixing block size and varying the sample count (top to bottom). The total length of data which is then divided into samples is noted in the top left corner of each subplot. We see that as long a given time series is divided into at least 20-30 samples, the resulting samples approximately span the full distribution and approximately match the "true" uncertainty shown by Figure 3. This matches expectations from large sample theory (e.g. central limit theorem) arguments. Furthermore, we see that block sizes greater than approximately 30 years (at this specific latitude) reduce bias in the median return level estimates (Figure 3). As long as one has more than 20-30 ensemble members, it is always more beneficial to have as large of a block size as possible from which return levels are calculated.

The results above demonstrate that given the same length of data in aggregate, shorter block lengths distort the apparent rarity of extremes, even when multiple blocks are available. The statistical instability arises not only from limited data length, but from how sampling variability affects the parameters of the fitted GEV distribution. In particular, the shape parameter ξ, which governs the heaviness or boundedness of the distribution tail, plays an outsized role in determining return level accuracy. To better understand this mechanism, we next examine how biases in ξ directly propagate into biases in return level estimates across many subsampled realizations of the aquaplanet data.

3.2 Shape Parameter as Driver of Return Level Uncertainty

Among GEV parameters, the shape parameter (ξ) exerts the dominant control on rare-event probability and thus on perceived risk. Even small deviations in ξ can lead to large errors in return levels because of the nonlinear form of the GEV return level equation (Eq. 4). To investigate this sensitivity, we use the full 13,000-year precipitation time series as the reference ("true") distribution and compare it to results from shorter, subsampled segments of 10, 30, and 50 years.

Using the full 13,000-year time series, we compute both the true shape parameter and the true 1000-year return level at 40.46° North and South latitude. We then compare those reference values with the values estimated from every 10-, 30-, and 50-year block. Figure 5 shows how the bias in the shape parameter relates to the relative bias in the 1000- and 100-year return levels for each subsampled time series. Relative bias can be thought of as a percent error and is defined as the difference between the estimated and the true return levels divided by the true return level. The median bias is shown by a magenta point, and dashed contours show fixed kernel density estimates.

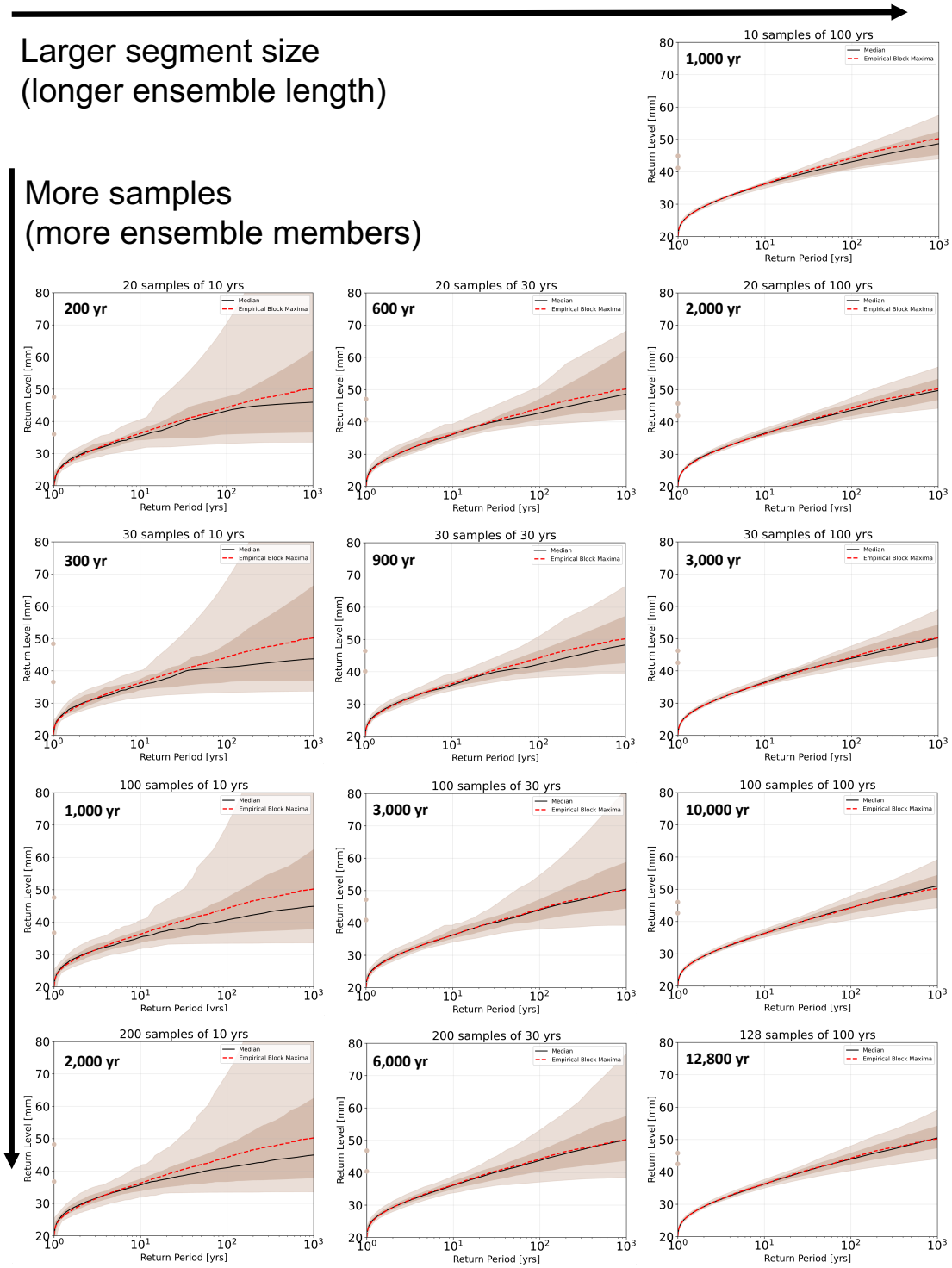

Figure 4. Resampled return-level curves for annual maximum daily precipitation at 40.46 degrees, similar to Fig. 3. From top to bottom, subplanels show an increasing number of samples of the 12,896 year time series used to generate return level curves. From left to right, longer segments of data are used to generate the return level curve. The black line indicates the pointwise median of the return-level curves. The dashed red line indicates the return-level curve generated from the entire time series. Dark brown shading indicates the 25th to 75th percentile. Light brown shading indicates the 5th to 95th percentiles. Brown dots on the y-axis indicate the 25th and 75th percentiles of return level corresponding to a 100 year return period. Bold text annotations indicate the total data length used to generate each subpanel.

In general, underestimating the shape parameter is associated with underestimating the return level, while overestimation of ξ inflates return levels.

The magnitude of the median biases in both the shape parameter and return level decrease when going from blocks of 10 to 30 years of data (cf. Figs. 5a–b and 5c–d) as the tails of the distribution become better sampled. The difference is less extreme but still noticeable when going from samples of 30 to 50 years (c.f. Figs. 5c–d and 5e–f). We also see that the joint distribution of shape parameter bias vs. return level bias becomes more constrained with more data. Unsurprisingly, there is also a larger relative error when estimating 1000-year return levels (Figure 5b, d, f) compared to 100-year return levels (Figure 5a, c, e). It is important to note that while the median biases decrease in magnitude, any individual 30- or even 50-year segment of the time series is still likely to be biased in the shape parameter, potentially leading to a bias in the return level by ±20%. However, an ensemble median of multiple 30- and 50-year samples is likely to produce an unbiased estimate of the shape parameter and thus return level, as seen by the magenta dot being located near (0,0) (Fig. 5).

3.3 Variation in Bias and Return-Level Uncertainty Across Latitudes

While the previous sections focused on a single latitude, we can use the aquaplanet simulation to systematically assess how sample size limitations impact extreme value estimates across latitudes for both precipitation and temperature. We compare the "true" GEV parameters and return levels with estimates derived from 10-, 30-, and 100-year time series segments to assess how shape parameter and return level biases vary.

Figure 6 shows the relationship between the true shape parameter and the median bias in estimated shape parameter for precipitation (panel a) and temperature (panel b) across all latitudes in the aquaplanet simulation. For both variables, negative shape parameters are underestimated, with shorter samples amplifying this bias. In the few locations where the true shape parameter $\xi > 0$, we see a slight positive bias in the shape parameter estimates, such that short samples overestimate the heaviness of the tail. As expected, the bias magnitude decreases sharply with longer time series (Figure 6). However, a notable distinction emerges between temperature and precipitation: for temperature, even 30-year samples often yield substantially underestimated shape parameters, highlighting the challenges of estimating bounded distributions from limited data. Conceptually, this is because, for distributions with a hard upper bound (i.e., negative ξ), shorter samples are less likely to capture values near that bound, making the distribution appear more tightly constrained than its asymptotic limit. In contrast, for unbounded or heavy-tailed distributions, occasional large extremes are more likely to appear even in small samples, helping mitigate the shape parameter bias.

The sign and magnitude of the finite-sample bias are consistent with both second-order asymptotic theory and prior numerical studies (e.g., 16; 17; 18; 14). In EVT, the convergence of block maxima to the limiting GEV form is governed by a second-order term $A(t)$, whose sign determines whether the block-maxima CDF approaches the limit from above or below (16; 17; 18). In practice, this means that distributions with bounded upper tails (reversed-Weibull; $\xi < 0$), such as temperature, tend to yield negative ξ biases in finite samples, while heavy-tailed (Fréchet; $\xi > 0$) distributions such as precipitation tend to yield positive biases. This theoretical behavior aligns closely with the observed bias patterns across latitudes and provides a more theoretical explanation for why temperature extremes are systematically underestimated and precipitation extremes are occasionally inflated when using short time series. Crucially, we see that bounded extremes (e.g., temperature) are systematically underpredicted from short data, while heavy-tailed extremes (e.g., precipitation) have a potential to be exaggerated.

When looking at Eq. (4), it is clear that small changes in the shape parameter (on the order of ±0.25 as seen in Figure 5) greatly influence the return level, more so than small errors in estimating the location or scale parameters (not shown), due to nonlinearities in the dependence of return levels on ξ. Based on this fact and the results above, we argue that the shape parameter drives the return level bias across latitudes. To test this, we compute the relative bias in the 1000-year return level at each latitude and plot it against the corresponding shape-parameter bias (Figure 7). Each point represents the average of all 10-, 30-, or 100-year segment estimates, with shading denoting the length of segments.

These results confirm that shape parameter bias is the dominant source of error in return level estimation, both in sign and magnitude, across a range of latitudes and climate variables. Underestimating the shape parameter leads to systematically lower return levels, while overestimating it inflates risk. Although averages across multiple longer samples (>30 years) reduce these biases, individual short records can continue to produce substantial variability even

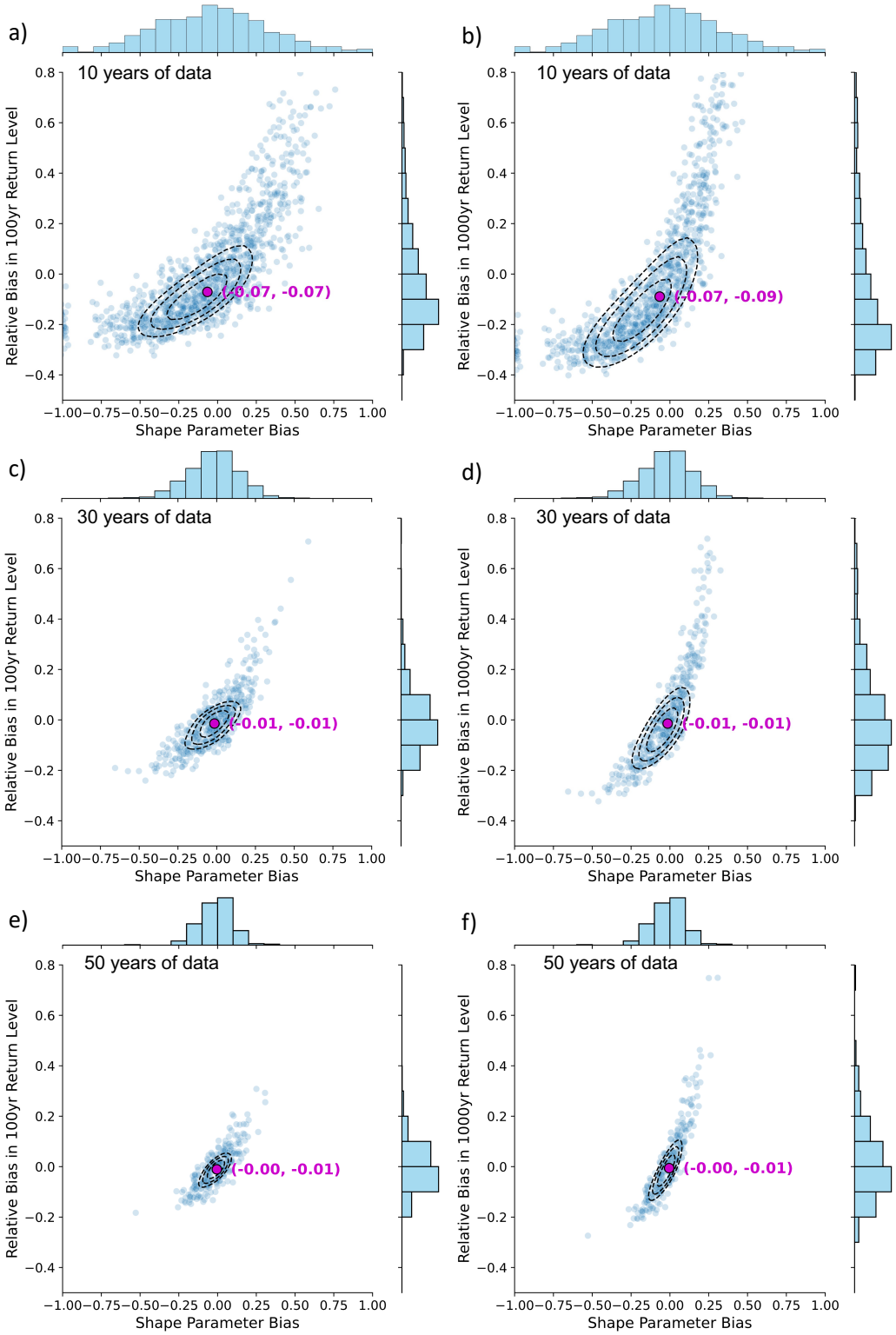

Figure 5. Joint distributions of shape parameter bias and bias in 100 and 1000-year return levels. (a) Scatter plot showing the joint distribution of bias in the shape parameter compared to the relative bias in the 100 year return level using 10-year segments of the precipitation time series. Relative bias is defined as the difference between the estimated and true return levels divided by the true return level. The magenta point is the location of the median bias in the shape parameter and the median relative return level bias. Marginal histograms are shown on each axis for illustrative purposes. Kernel density contours are plotted at values of 0.7, 0.8, and 0.9. (b) Same as (a) but illustrating the relative bias in the 1000-year return level. (c) Same as (a) but using 30-year segments of data. (d) Same as (a) but using 30-year segments of data and for the 1000-year return level. (e) Same as (a) but using 50-year segments of data. (f) Same as (a) but using 50-year segments of data and for the 1000-year return level.

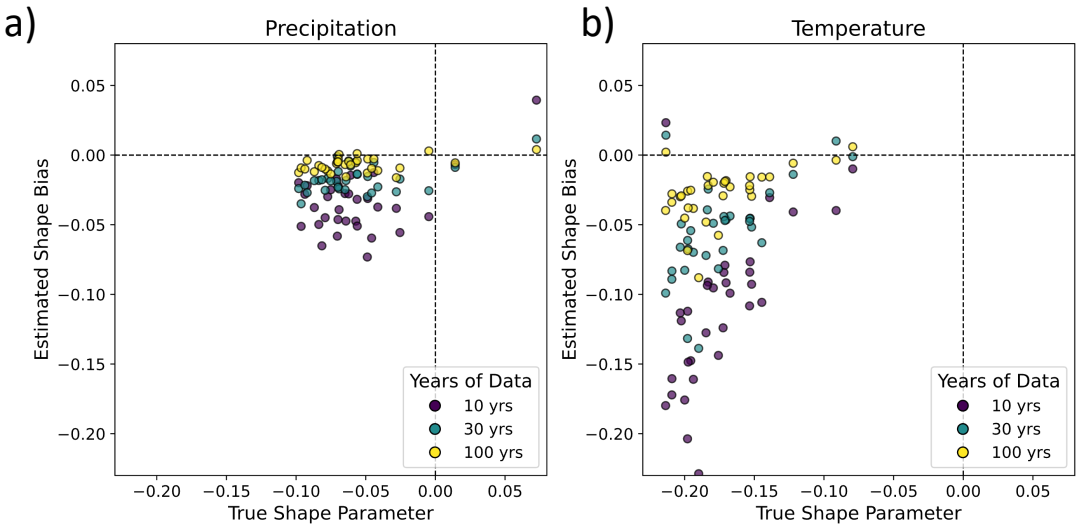

Figure 6. (a) Scatter plot of the true shape parameter of the precipitation distribution for all latitudes in the idealized aquaplanet simulation plotted against the median estimated bias in the shape parameter when using 10, 30, and 100 years of data. Points are shaded by the years of data used in the time series that yields these estimates. (b) Same as (a) but for temperature in the idealized aquaplanet simulation.

when distributions are relatively well-behaved (e.g., Fig 5).

3.4 Application to GFDL ESM4

The real world is quite far from the idealized aquaplanet model. Precipitation and temperature extremes vary spatially, are subject to low-frequency variability, and data are limited to much less than 10,000 years. To evaluate how our insights generalize to a more realistic setting, we apply the same framework to the 500-year preindustrial control simulation from the GFDL-ESM4 Earth System Model, a part of the CMIP6 archive. While the simulation is shorter and includes greater internal variability, it offers a more realistic setting to test the robustness of our methods and assess the practical implications for uncertainty quantification in return level estimates under observationally-relevant constraints.

Figure 8 shows the GEV parameters derived from the full 500-year dataset for both precipitation and temperature. These serve as our reference or "true" parameters for this section. While even 500 years may not fully constrain the tail behavior in arid regions, these parameter maps provide a practical baseline against which to evaluate finite-sample biases. The location parameter for precipitation is larger in typically wetter regions (Figure 8a) and the location parameter for temperature is larger in typically hotter regions (Figure 8b). The scale parameter, associated with the width of the GEV distribution, is larger in locations with more variability from one year to another in precipitation/temperature (Figure 8c–d). Consistent with expectations from the aquaplanet analysis, the shape parameter for precipitation is near zero, but slightly positive, for much of the planet, whereas all regions exhibit negative shape parameters for temperature, implying a bounded distribution.

To assess finite-sample effects, we subdivide the 500-year simulation into 10-, 30-, and 50-year segments and re-estimate the 100-year return levels for both variables. Figure 9 shows the resulting biases in both the shape parameter and return level for precipitation. We see large positive biases in the shape parameter ξ in parts of the tropics for 10-year samples, which translate into substantial overestimates of return levels (Figure 9a, b). In contrast, mid- and high-latitude regions, where ξ is typically negative or near zero, tend to show negative shape biases, resulting in slightly underestimated return levels. The static-like noise in Figure 9 is largely due to random sampling variability. When dividing the time series into 30 and 50 year blocks of data, this bias is greatly reduced, but some subtropical regions with large positive shape parameters ($\xi \geq 0.5$) continue to exhibit large return-level biases even when the median shape bias is near zero (Figure 9d).

This apparent contradiction arises because the return level equation (Eq. 4) is highly nonlinear in ξ, particularly when $\xi \gtrsim 0.5$ (as is the case in climatologically arid regions). Additionally, maximum likelihood estimation of the shape parameter breaks down for shape parameters $\xi > 0.5$. For these reasons, evaluation should be focused on regions with shape parameters $\xi < 0.5$ and in

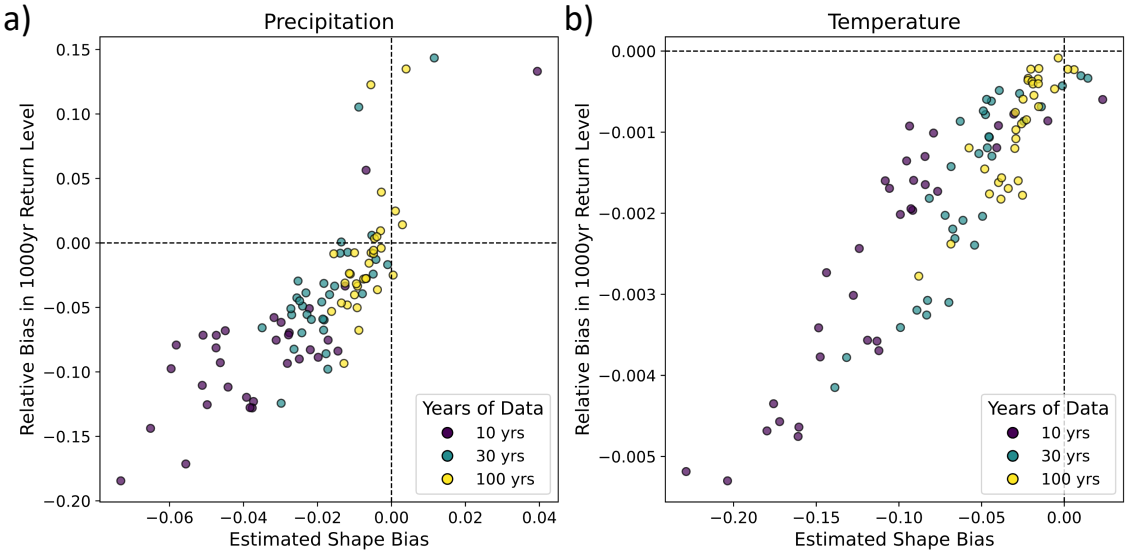

Figure 7. (a) A scatter plot of the estimated bias in the shape parameter of the precipitation distribution for at all latitudes in the idealized aquaplanet simulation plotted against the relative bias in the 1000 year return level when using 100, 30, and 10 years of data. Points are shaded by the years of data used in the time series that yields these estimates. (b) Same as (a) but for temperature in the idealized aquaplanet simulation.

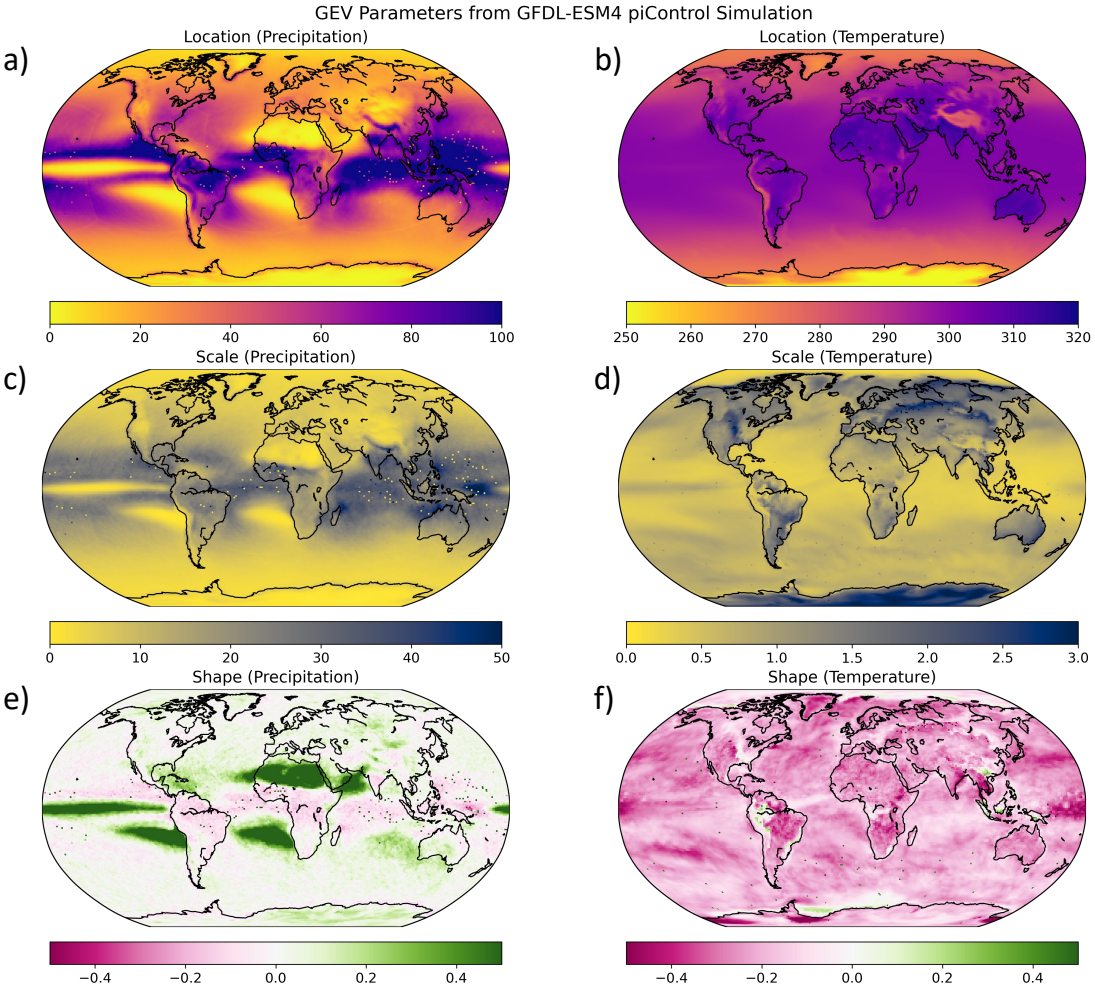

Figure 8. Plots of the GEV (a, b) location parameter, (c, d) scale parameter, and (e, f) shape parameters for (a,c,e) precipitation and (b,d,f) temperature in the 500 year GFDL ESM4 preindustrial control simulation. Location and scale parameters have units of mm and K for precipitation and temperature respectively, and the shape parameter is unitless.

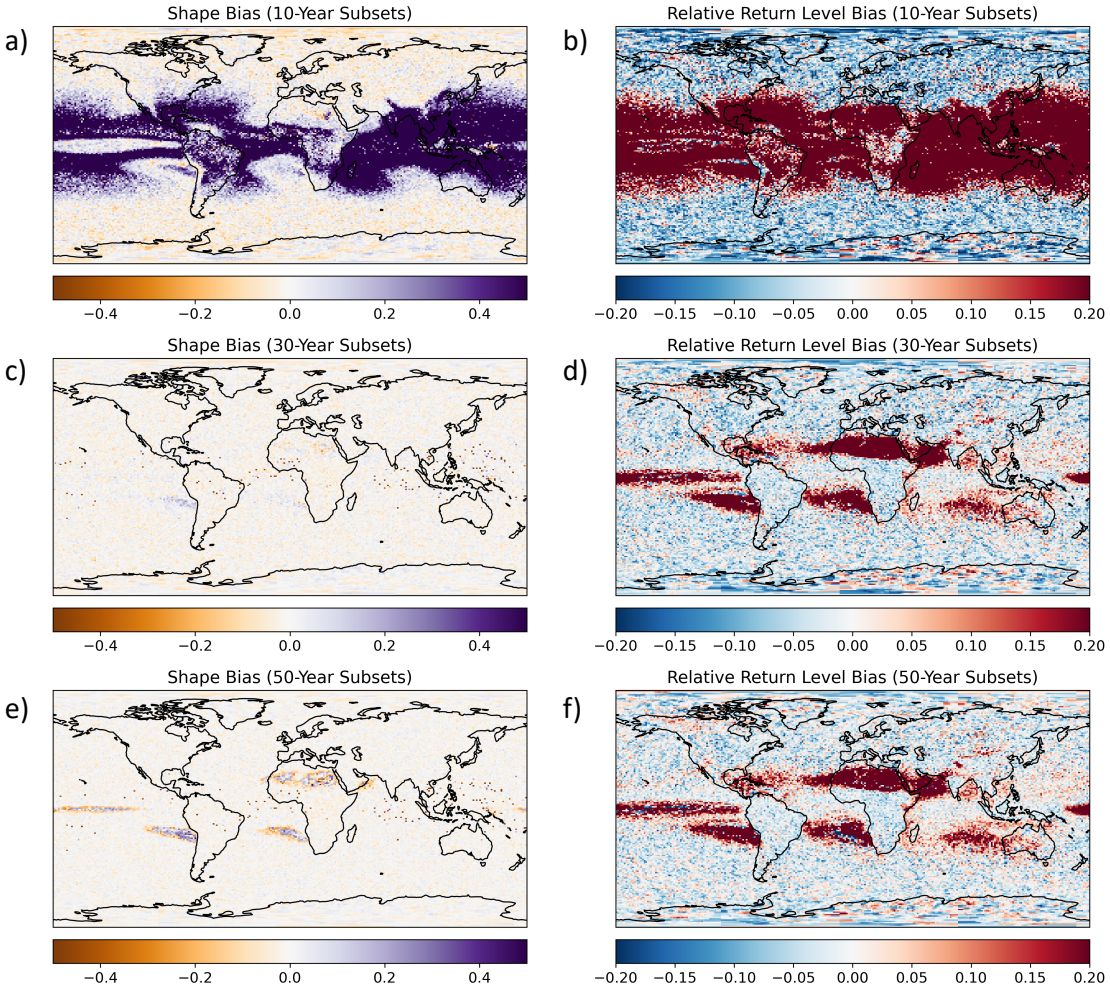

Figure 9. (a) Map of the median shape bias obtained by subsampling the precipitation time series into 10-year subsets. (b) Map of the median relative return level bias in the 100-year return level for precipitation estimated from 10-year segments of data. (c) Same as (a) but for 30 years of data. (d) Same as (b), but for 30 years of data. (e) Same as (a) but for 50 years of data. (f) Same as (b), but for 50 years of data.

those regions, 30 years of data produces $< 1\%$ bias in estimating the 100-year return level for precipitation.

In contrast to the idealized aquaplanet simulation, we find that for precipitation, the bias in the shape parameter is not strongly correlated with the true value of ξ. Instead, we see a correspondence between the **scale** parameter and the bias in the shape parameter. Figure 9a shows that regions with very large scale parameters ($\sigma > 20$ mm day^{-1}, as seen in Fig. 8c), particularly in the tropics, are also regions where 10-year segments produce the largest shape and return level biases. Conceptually, this is because when the precipitation maxima are highly variable from year to year, even 10 or 20 years of data will have dramatically different maxima, giving the false impression of a fat-tailed distribution with a large, positive shape parameter ($\xi > 4$) if the distribution is not properly sampled. This means we are more likely to obtain large overestimates of the shape parameter and associated return levels. With 30 years of data, this effect is significantly reduced, and in most locations, the estimates fall much closer to those derived from the full dataset.

For temperature (Figure 10), the patterns are more consistent with theoretical expectations, largely because the scale parameter is relatively small. Since the shape parameter is negative nearly everywhere (Figure 8f), short samples exhibit negative bias in ξ, which in turn leads to underestimates of the return level (Figure 10). These return level biases are most pronounced in the midlatitudes and, though modest in percentage terms (up to 0.5%), they correspond to meaningful differences of up to 1.5°C in the 100-year temperature return level. Even with 30-year samples, temperature estimates remain biased across much of the world, suggesting that longer records may be needed to accurately characterize extreme temperature risk and for other bounded climate variables. While there are only ten 50-year segments in the 500-year time series, we see a much lower bias in shape parameter and relative return level using 50-year segments of

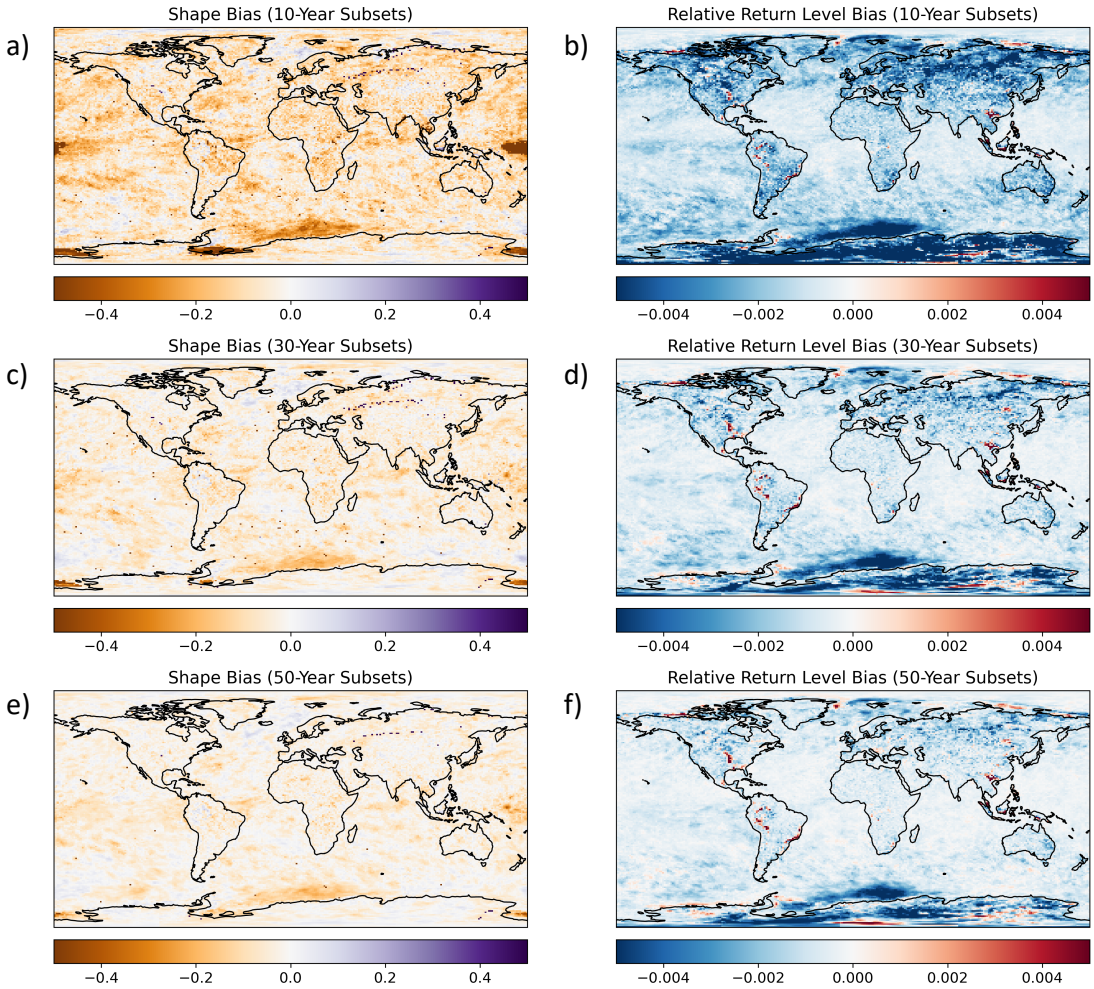

Figure 10. Same as Figure 9 but for temperature.

data (Figure 10e–f).

As shown earlier (Fig. 5), the ensemble median of many 50-year samples is approximately unbiased, but any single 50-year sample still frequently produces an incorrect estimate of return levels for both 100- and 1000-year events. Thus, achieving a 50-year record does not guarantee that a particular realization is sufficiently long for robust extrapolation to rare events. Instead, the total length of data n required to reliably estimate a T-year return level depends on the variability in the underlying GEV parameters, particularly the scale and shape parameters, and details of this relationship will be left to future work.

Taken together, these results show that finite-sample biases are not random but systematic and physically interpretable. They align with the statistical nature of the underlying process and the amplitude of variability. For practical applications, this means that return levels derived from lone 30 to 50-year observational records can significantly misrepresent rare-event probability, even though an ensemble of many such records would perform well, particularly for bounded variables. Recognizing these biases is crucial for interpreting design standards, insurance loss models, and adaptation thresholds derived from limited and/or synthetically generated data.

4 Discussion and Summary

This study systematically evaluates how finite sample size limitations affect the estimation of extreme event probabilities using generalized extreme value (GEV) distributions. Using long, stationary simulations from an idealized aquaplanet model and a 500-year GFDL-ESM4 control run, we quantify how biases in GEV parameters, especially the shape parameter ξ, propagate into estimates of 100- and 1000-year return levels. We show that the accuracy of rare-event statistics depends not only on the total data available, but critically on the length of the individual time series segments, whether they originate from a single realization or from an individual member of

an ensemble of simulations.

Across latitudes and variables, we find a consistent and physically interpretable pattern. Bounded, short-tailed distributions (negative ξ), typical of temperature extremes and many variables in the climate system (30), tend to produce underestimated ξ values in short samples, yielding overly conservative tails and underpredicting return levels. In contrast, heavy-tailed distributions (positive ξ) typical of precipitation in some locations (e.g., 31) exhibit the opposite tendency, inflating return levels from short records. These patterns are consistent with the second-order condition in extreme value theory, which governs how empirical block maxima converge to the asymptotic GEV form (16; 17). Practically, this means that records shorter than ∼30–50 years systematically bias the shape parameter, though this threshold varies by variable and latitude, with midlatitude temperature extremes and tropical precipitation requiring longer windows for reliable inference.

Our results also highlight an important distinction: while any individual 30-year or 50-year realization is likely to yield an inaccurate estimate of the return level, ensemble averages across many 30-year (or longer) samples tend to produce small median errors and are unbiased. Once the systematic bias in the shape parameter has been removed (after about 30-50 years) ensemble methods can be used to unbiasedly and more precisely estimate the probability of extreme events. Thus, reliable ensemble-based risk assessment requires both many members and sufficiently long members.

To address some of these challenges, bias-aware parameter estimation methods, such as the Generalized Maximum Likelihood Estimator (GMLE; (13)), L-moments methods (32), or Bayesian and machine-learning methods (e.g., 14; 33), can provide a pragmatic solution by constraining ξ toward physically plausible ranges and reducing finite-sample distortions in short records. Such prior-informed bias corrections, informed by the second order condition, would effectively pull short-record estimates of the shape parameter toward more realistic values (i.e. towards zero). In real-world applications where the shape parameter remains the most difficult to constrain, and observational noise and dataset inhomogeneities only exacerbate this issue, strategies that sample a plausible range of ξ values rather than relying on a single best-fit estimate may also help improve the robustness of risk assessments. For example, a 30-year estimate of $\xi = -0.6$ for temperature extremes is likely a biased underestimate. In such cases, decision-makers could consider a plausible range of less negative ξ values (e.g., -0.6 to -0.3) to better capture the uncertainty in rare event risk. By explicitly accounting for the biased estimate of the shape parameter, various techniques can improve the robustness of return level estimates from short observational records.

These findings have important implications for interpreting climate extremes in both observational datasets and model simulations, designing simulations, and analyzing observational or synthetic data records for climate extremes. Ensembles consisting of at least 30-50 years of data (based on variable of interest and location) are needed in order for the shape parameter to be unbiased and provide accurate estimates of extreme event probabilities. Moreover, the analysis of long synthetic datasets may inadvertently bias shape parameter estimates unless they are aggregated using using sufficiently long blocks.Moreover, our results support recent arguments (e.g., 34) that, for robust climate hazard characterization, ensembles of moderate-resolution simulations with long time series ($\geq$ 30 years) are likely more effective than relying on a small number of high-resolution runs.

While this study focuses on stationary settings, future work should extend these insights to nonstationary extremes and additional hazards (e.g., hail; (9)). By bridging theoretical insights from extreme value theory with large-sample climate modeling, this work clarifies how and why GEV-based estimates fail using finite data and offers practical guidance for bias-aware analysis of climate extremes that can improve the reliability of extreme-event metrics under real-world data constraints.

Acknowledgments

Model simulations for the aquaplanet simulation were performed on Caltech's Resnick High Performance Computing Cluster. We thank Sandra Yuter and an anonymous reviewer for helpful comments and feedback on drafts of this work.

Funding

This research was supported by Schmidt Sciences, LLC. R.N.P was also supported by an American Meteorological Society (AMS) Graduate Fellowship.

Author contributions

T.S. and R.N.P conceptualized the study, T.S. and R.N.P acquired funding, R.N.P led the investigation, formal analysis, visualization, and preparation of the original draft. T.S. supported reviewing and editing.

Data availability

The GFDL-ESM4 esm-piControl run is part of the CMIP6 archive and can be found at https://esgf-node.llnl.gov/search/cmip6/. GFDL-FMS code needed to run the aquaplanet can be found on https://github.com/NOAA-GFDL/FMS. The specific run configuration is available from the authors on request.

References

[1] Arfanuzzaman M, Richard A Betts A G, Hirabayashi Y, Lissner T K, Gunn E L, Liu J, Morgan R, Mwanga S and Supratid S 2022 Water *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* ed Pörtner H O, Roberts D, Tignor M, Poloczanska E, Mintenbeck K, Alegría A, Craig M, Langsdorf S, Löschke S, Möller V, Okem A and Rama B (Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press) chapter 4 URL https://www.ipcc.ch/report/ar6/wg2/downloads/report/IPCC_AR6_WGII_Chapter04.pdf

[2] Jenkinson A F 1955 *Quarterly Journal of the Royal Meteorological Society* **81** 158–171 ISSN 1477-870X URL http://dx.doi.org/10.1002/qj.49708134804

[3] Katz R W, Parlange M B and Naveau P 2002 *Advances in Water Resources* **25** 1287–1304 ISSN 0309-1708 URL http://dx.doi.org/10.1016/S0309-1708(02)00056-8

[4] Villarini G, Smith J A, Ntelekos A A and Schwarz U 2011 *Journal of Geophysical Research* **116** ISSN 0148-0227 URL http://dx.doi.org/10.1029/2010JD015038

[5] Westra S, Alexander L V and Zwiers F W 2013 *Journal of Climate* **26** 3904–3918 URL https://doi.org/10.1175/jcli-d-12-00502.1

[6] Wehner M F, Duffy M L, Risser M, Paciorek C J, Stone D A and Pall P 2024 *Frontiers in Climate* **6** ISSN 2624-9553 URL http://dx.doi.org/10.3389/fclim.2024.1343072

[7] Zahid M, Blender R, Lucarini V and Bramati M C 2017 *Earth System Dynamics* **8** 1263–1278 ISSN 2190-4987 URL http://dx.doi.org/10.5194/esd-8-1263-2017

[8] Gessner C, Fischer E M, Beyerle U and Knutti R 2021 *Journal of Climate* 1–46 ISSN 1520-0442 URL http://dx.doi.org/10.1175/JCLI-D-20-0916.1

[9] Allen J T, Tippett M K, Kaheil Y, Sobel A H, Lepore C, Nong S and Muehlbauer A 2017 *Monthly Weather Review* **145** 4501–4519 ISSN 1520-0493 URL http://dx.doi.org/10.1175/MWR-D-17-0119.1

[10] Rocco M 2013 *Journal of Economic Surveys* **28** 82–108 ISSN 1467-6419 URL http://dx.doi.org/10.1111/j.1467-6419.2012.00744.x

[11] Coles S 2001 *An Introduction to Statistical Modeling of Extreme Values* (Springer London) ISBN 9781447136750 URL http://dx.doi.org/10.1007/978-1-4471-3675-0

[12] Hosking J R M, Wallis J R and Wood E F 1985 *Technometrics* **27** 251–261 ISSN 1537-2723 URL http://dx.doi.org/10.1080/00401706.1985.10488049

[13] Martins E S and Stedinger J R 2000 *Water Resources Research* **36** 737–744 ISSN 1944-7973 URL http://dx.doi.org/10.1029/1999WR900330

[14] Zeder J, Sippel S, Pasche O C, Engelke S and Fischer E M 2023 *Geophysical Research Letters* **50** ISSN 1944-8007 URL http://dx.doi.org/10.1029/2023GL104090

[15] Huang W K, Stein M L, McInerney D J, Sun S and Moyer E J 2016 *Advances in Statistical Climatology, Meteorology and Oceanography* **2** 79–103 ISSN 2364-3587 URL http://dx.doi.org/10.5194/ascmo-2-79-2016

[16] de Haan L and Ferreira A 2006 *Extreme Value Theory* (Springer New York) ISBN 9780387344713 URL http://dx.doi.org/10.1007/0-387-34471-3

[17] Alves M I F, Gomes M I, De Haan L and Neves C 2007 *REVSTAT-Statistical Journal* Vol. 5 No. 3 (2007): REVSTAT–Statistical Journal URL https://revstat.ine.pt/index.php/REVSTAT/article/view/53

[18] Bücher A and Zhou C 2021 *Statistical Science* **36** ISSN 0883-4237 URL http://dx.doi.org/10.1214/20-STS795

[19] Dunne J P, Horowitz L W, Adcroft A J, Ginoux P, Held I M, John J G, Krasting J P, Malyshev S, Naik V, Paulot F, Shevliakova E, Stock C A, Zadeh N, Balaji V, Blanton C, Dunne K A, Dupuis C, Durachta J, Dussin R, Gauthier P P G, Griffies S M, Guo H, Hallberg R W, Harrison M, He J, Hurlin W, McHugh C, Menzel R, Milly P C D, Nikonov S, Paynter D J, Ploshay J, Radhakrishnan A, Rand K, Reichl B G, Robinson T, Schwarzkopf D M, Sentman L T, Underwood S, Vahlenkamp H, Winton M, Wittenberg A T, Wyman B, Zeng Y and Zhao M 2020 *Journal of Advances in Modeling Earth Systems* **12** ISSN 1942-2466 URL http://dx.doi.org/10.1029/2019MS002015

[20] O'Gorman P A and Schneider T 2008 *Journal of Climate* **21** 3815–3832 URL https://doi.org/10.1175/2007jcli2065.1

[21] Eyring V, Bony S, Meehl G A, Senior C A, Stevens B, Stouffer R J and Taylor K E 2016 *Geoscientific Model Development* **9** 1937–1958 ISSN 1991-9603 URL http://dx.doi.org/10.5194/gmd-9-1937-2016

[22] Lucarini V, Faranda D, Freitas A C M, Freitas J M, Kuna T, Holland M, Nicol M, Todd M and Vaienti S 2016 Extremes and recurrence in dynamical systems URL https://arxiv.org/abs/1605.07006

[23] Fisher R A and Tippett L H C 1928 *Mathematical Proceedings of the Cambridge Philosophical Society* **24** 180–190 ISSN 1469-8064 URL http://dx.doi.org/10.1017/S0305004100015681

[24] Gnedenko B 1943 *The Annals of Mathematics* **44** 423 ISSN 0003-486X URL http://dx.doi.org/10.2307/1968974

[25] Carney M, Kantz H and Nicol M 2020 Analysis and simulation of extremes and rare events in complex systems *Advances in Dynamics, Optimization and Computation* Studies in systems, decision and control (Cham: Springer International Publishing) pp 151–182

[26] Wilks D S 2019 *Statistical methods in the atmospheric sciences* 4th ed (Philadelphia, PA: Elsevier Science Publishing)

[27] Zhang Y and Boos W R 2023 *Proceedings of the National Academy of Sciences* **120** ISSN 1091-6490 URL http://dx.doi.org/10.1073/pnas.2215278120

[28] Ferro C A T and Segers J 2003 *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **65** 545–556 ISSN 13697412, 14679868 URL http://www.jstor.org/stable/3647520

[29] Efron B and Tibshirani R 1994 *An Introduction to the Bootstrap* (Chapman and Hall/CRC) ISBN 9780429246593 URL http://dx.doi.org/10.1201/9780429246593

[30] Lucarini V, Faranda D, Wouters J and Kuna T 2014 *Journal of Statistical Physics* **154** 723–750 ISSN 1572-9613 URL http://dx.doi.org/10.1007/s10955-013-0914-6

[31] Brooks C E P and Carruthers N 1953 *Handbook of statistical methods in meteorology.* (HM Stationery Office)

[32] Hosking J R M 1990 *Journal of the Royal Statistical Society Series B: Statistical Methodology* **52** 105–124 ISSN 1467-9868 URL http://dx.doi.org/10.1111/j.2517-6161.1990.tb01775.x

[33] Rai S, Hoffman A, Lahiri S, Nychka D W, Sain S R and Bandyopadhyay S 2024 *Environmetrics* **35** ISSN 1099-095X URL http://dx.doi.org/10.1002/env.2845

[34] Schneider T, Behera S, Boccaletti G, Deser C, Emanuel K, Ferrari R, Leung L R, Lin N, Müller T, Navarra A, Ndiaye O, Stuart A, Tribbia J and Yamagata T 2023 *Nature Climate Change* **13** 887–889 ISSN 1758-6798 URL http://dx.doi.org/10.1038/s41558-023-01769-3