Prediction of extreme events in climatic time series using: classical, bayesian, and machine learning approaches

Jefferson Conza Juan Riofrio Saba Infante Eusebio Ariza

January 6, 2026

Abstract

Extreme meteorological, hydrological, and climatic events have increased in both frequency and intensity over recent decades as a consequence of climate change, posing significant challenges to social stability, food security, ecosystems, financial systems, public health, and critical infrastructure. Extreme Value Theory provides a rigorous probabilistic framework for modeling rare events and quantifying uncertainty associated with unusually large or small observations.

In this work, we propose a methodological framework for the detection and prediction of extreme values in climatic time series, with particular emphasis on forecasting block maxima over future periods. We compare classical extreme value models based on the Generalized Extreme Value (GEV) distribution with Bayesian inference techniques and machine learning approaches for parameter estimation and predictive performance. The proposed methods are evaluated using both simulated data and real climatic datasets from Ecuador, allowing for a systematic assessment of their ability to predict extreme maxima under realistic conditions.

Predictive performance is assessed using quantitative metrics that measure the accuracy and reliability of maximum value forecasts. The results highlight the strengths and limitations of each modeling approach and provide insights into their suitability for extreme event prediction in climatic applications, with potential relevance for risk assessment and decision-making in the context of climate change.

Keywords: Extreme events; Time series analysis; Generalized Extreme Value distribution; Bayesian methods; Machine learning.

1 Introduction

Extreme events constitute a fundamental object of study in many scientific disciplines, particularly in environmental and climatic sciences, due to their potentially severe social, economic, and ecological consequences. Phenomena such as extreme precipitation, floods, droughts, heat waves, and unusually high or low temperatures, although infrequent, are often responsible for disproportionate levels of damage to infrastructure, agriculture, public health, and natural ecosystems. As a result, the statistical analysis and prediction of extreme values have become essential tools for risk assessment and decision-making in climate-sensitive contexts [5, 19, 28].

From a statistical viewpoint, extreme observations arise naturally as maxima or minima of stochastic processes and display characteristics that differ markedly from those of typical or average behavior. Climatic time series are often highly variable and exhibit skewed distributions, heavy tails, temporal dependence, seasonality, and clustering of extreme events. These properties complicate standard modeling approaches and necessitate specialized techniques capable of adequately describing tail behavior and quantifying uncertainty in rare-event regimes [29].

The classical framework for modeling extreme values is provided by Extreme Value Theory (EVT). The foundational results of Fisher and Tippett [9], later formalized by Gnedenko [11], establish that the distribution of properly normalized block maxima converges to one of three limiting forms: Gumbel, Fréchet, or Weibull. These distributions are unified under the Generalized Extreme Value (GEV) family [15], which has become a standard tool in applied extreme value analysis. EVT has been successfully applied in numerous areas, including climatology, hydrology, finance, and engineering, and provides a theoretically sound basis for extrapolating beyond the range of observed data.

Despite its strong theoretical foundations, the practical implementation of EVT presents several challenges. In many applications, inference relies on relatively small samples of extreme observations, leading to substantial estimation uncertainty. Moreover, climatic time series frequently exhibit non-stationary behavior due to seasonal patterns, long-term trends, and climate variability. Extensions of EVT, including threshold-based approaches and point process representations, have been developed to address some of these issues by allowing for greater modeling flexibility [2, 5, 26].

Bayesian methods provide an alternative inferential framework that is particularly well suited to extreme value analysis. By incorporating prior information through prior distributions and employing Markov chain Monte Carlo techniques, Bayesian approaches allow for a coherent treatment of parameter uncertainty, flexible hierarchical modeling, and the construction of full predictive distributions for future extremes [1, 3, 6, 12]. These features are especially valuable when extreme observations are scarce or when expert knowledge is available.

In recent years, machine learning and deep learning techniques have gained increasing attention in the analysis and prediction of time series data. These methods have demonstrated strong predictive capabilities in a variety of applications and offer the ability to model complex nonlinear relationships and temporal dependencies. Several studies have proposed machine learning approaches specifically designed to improve the prediction of extreme events, either by incorporating loss functions motivated by extreme value theory or by adapting recurrent neural network architectures to better capture extreme behavior [7, 14, 21, 25, 27, 31, 32, 33].

Motivated by these developments, the present work aims to analyze and compare classical EVT-based models, Bayesian inference methods, and machine learning approaches for the prediction of extreme values in climatic time series. The focus is placed on forecasting future block maxima and evaluating predictive performance under different modeling assumptions. The proposed framework is assessed using both simulated data and real climatic datasets from Ecuador, with the objective of providing insight into the advantages and limitations of each approach and their potential applicability in climate-related risk assessment.

2 Problem Statement

Let

$$X_T = (x_1, x_2, \ldots, x_T), \qquad x_t \in \mathbb{R}^d,\ T \in \mathbb{N},$$

denote a multivariate time series representing a climatic system of interest, where each vector x_t contains d observed variables at time t, such as temperature, precipitation, humidity, or wind speed. The aim of this work is to analyze and predict extreme values derived from this series, with particular emphasis on the prediction of maximum values over future time periods.

2.1 Block maxima construction

Following a block-based perspective, the observed series is divided into consecutive, non-overlapping segments associated with a fixed prediction horizon. Let $l \in \mathbb{N}$ denote the length of the prediction period and let $k \in \mathbb{N}$ be such that $kl < T$. The available information up to time kl is given by the partial time series

$$X_{kl} = (x_1, x_2, \ldots, x_{kl}).$$

The objective is to characterize extreme behavior over the subsequent period of length l by focusing on the maximum value attained by a specific component of the series during that interval.

Definition 1 (Future block maximum)**.** Given the input series X_{kl}, the future block maximum is defined as

$$M_{kl} = \max \left\{ x_{kl+1}^{(d)}, x_{kl+2}^{(d)}, \ldots, x_{(k+1)l}^{(d)} \right\},$$

where $x_t^{(d)}$ denotes the d-th component of the multivariate observation x_t.

This quantity represents the extreme outcome of interest over the next prediction block and constitutes the target variable in the forecasting problem.

Under suitable regularity conditions, the statistical behavior of block maxima admits an asymptotic characterization through the results of Extreme Value Theory, which provide a theoretical basis for modeling extremes and extrapolating beyond the range of observed values [5, 9, 11].

2.2 Prediction of future maxima

Given the observed input sequence X_{kl}, the central problem considered in this article is the prediction of the future maximum M_{kl} over the next block of length l. This prediction task involves learning the relationship between historical multivariate information and future extreme outcomes.

In practice, the accurate prediction of M_{kl} is challenging due to the limited number of extreme events, the presence of temporal dependence, interactions among multiple variables, and potential non-stationarities commonly observed in climatic time series.

2.3 Modeling perspectives

Different methodological approaches can be employed to address the prediction of extreme values. Classical approaches rely on Extreme Value Theory and parametric models to describe the tail behavior of block maxima. Bayesian methods extend this framework by incorporating prior information and providing a coherent quantification of uncertainty. More recently, machine learning techniques have been proposed as alternative tools for extreme value prediction, aiming to capture complex temporal and multivariate patterns directly from the data without relying on explicit distributional assumptions.

2.4 Purpose of the study

The purpose of this study is to compare these different modeling perspectives—classical EVT-based models, Bayesian approaches, and machine learning techniques—with respect to their ability to predict future block maxima in multivariate climatic time series. The comparison is conducted using both simulated data and real climatic datasets from Ecuador, with a focus on predictive performance and practical applicability in the context of extreme event analysis.

3 Theoretical background of models

This section provides the theoretical foundations of the modeling approaches considered in this work. We begin by reviewing the classical Extreme Value Theory framework for modeling block maxima. Although the prediction problem introduced in Section 2 is formulated in a supervised and multivariate setting, the classical block maxima approach provides the asymptotic justification for modeling extreme outcomes and serves as a theoretical benchmark for the methods considered in subsequent sections.

3.1 Block maxima approach

The block maxima approach constitutes a fundamental methodology in Extreme Value Theory for modeling the behavior of extremes arising from a sequence of random variables. While the variable of interest in Section 2 is the future maximum M_{kl} over a fixed prediction horizon, its statistical behavior can be studied through the classical framework of block maxima, which describes the asymptotic distribution of maxima computed over historical blocks of observations.

Let $\{X_1, X_2, \ldots\}$ be a sequence of independent and identically distributed random variables with common distribution function F. Consider a partition of the data into m non-overlapping blocks of equal size n, given by

$$(X_1, \ldots, X_n),\ (X_{n+1}, \ldots, X_{2n}),\ \ldots,\ (X_{(m-1)n+1}, \ldots, X_{mn}).$$

For each block $j = 1, \ldots, m$, define the block maximum

$$M_n^{(j)} = \max\left\{X_{(j-1)n+1}, X_{(j-1)n+2}, \ldots, X_{jn}\right\}.$$

The sequence $\{M_n^{(1)}, M_n^{(2)}, \ldots, M_n^{(m)}\}$ represents the observed block maxima and constitutes the primary object of interest in the block maxima framework.

Distribution of block maxima

The distribution function of the block maximum $M_n^{(j)}$ can be expressed as

$$\Pr\left(M_n^{(j)} \le x\right) = \Pr\left(X_{(j-1)n+1} \le x, \ldots, X_{jn} \le x\right) = \{F(x)\}^n,$$

assuming independence within each block. As $n \to \infty$, this distribution degenerates at the upper endpoint of F, making direct modeling impractical.

To overcome this degeneracy, it is necessary to consider normalized block maxima of the form

$$\frac{M_n^{(j)} - b_n}{a_n},$$

where $a_n > 0$ and $b_n \in \mathbb{R}$ are suitable normalizing sequences.

Extremal types theorem

The Fisher–Tippett–Gnedenko theorem states that if there exist sequences $\{a_n\}$ and $\{b_n\}$ such that the normalized block maxima converge in distribution, that is,

$$\Pr\left(\frac{M_n^{(j)} - b_n}{a_n} \le x\right) \longrightarrow G(x), \qquad n \to \infty,$$

for a non-degenerate distribution function G, then G must belong to the Generalized Extreme Value (GEV) family [5, 9, 11].

This theorem provides a complete characterization of all possible non-degenerate limiting distributions for normalized maxima and justifies the use of the GEV distribution as a universal model for block maxima.

Generalized Extreme Value distribution

The cumulative distribution function of the GEV distribution is given by

$$G(x;\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \quad 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0,$$

where $\mu \in \mathbb{R}$ is a location parameter, $\sigma > 0$ is a scale parameter, and $\xi \in \mathbb{R}$ is a shape parameter controlling the tail behavior.

Types of extreme value distributions

Depending on the value of the shape parameter ξ, the GEV distribution reduces to one of three classical types:

- **Gumbel type** ($\xi = 0$): In the limit as $\xi \to 0$, the GEV distribution converges to the Gumbel distribution,
$$G(x) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}, \qquad x \in \mathbb{R}.$$
- **Fréchet type** ($\xi > 0$): This case corresponds to heavy-tailed distributions with unbounded support,
$$G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \qquad x > \mu - \frac{\sigma}{\xi}.$$
- **Weibull type** ($\xi < 0$): This case corresponds to distributions with a finite upper endpoint at $\mu - \sigma/\xi$,
$$G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \qquad x < \mu - \frac{\sigma}{\xi}.$$

These three types exhaust all possible non-degenerate limits for normalized block maxima.

Likelihood-based inference

Assume that the block maxima

$$\{M_n^{(1)}, M_n^{(2)}, \ldots, M_n^{(m)}\}$$

are independent and identically distributed realizations from a Generalized Extreme Value distribution with parameters (μ,σ,ξ), with $\sigma > 0$ and $\xi \neq 0$. The likelihood function based on the observed block maxima is given by

$$\mathcal{L}(\mu,\sigma,\xi) = \prod_{j=1}^{m} f\left(M_n^{(j)} \mid \mu,\sigma,\xi\right),$$

where the GEV density function is

$$f(x \mid \mu, \sigma, \xi) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi-1}\exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\},$$

defined for

$$1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0.$$

Substituting this expression into the likelihood yields

$$\mathcal{L}(\mu, \sigma, \xi) = \frac{1}{\sigma^m}\exp\left\{-\sum_{j=1}^{m}\left[1+\xi\left(\frac{M_n^{(j)}-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}\prod_{j=1}^{m}\left[1+\xi\left(\frac{M_n^{(j)}-\mu}{\sigma}\right)\right]^{-1/\xi-1}.$$

Taking the natural logarithm, the log-likelihood function $L = \log\mathcal{L}(\mu, \sigma, \xi)$ can be written as

$$L(\mu, \sigma, \xi) = -m\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{j=1}^{m}\log\left[1+\xi\left(\frac{M_n^{(j)}-\mu}{\sigma}\right)\right] - \sum_{j=1}^{m}\left[1+\xi\left(\frac{M_n^{(j)}-\mu}{\sigma}\right)\right]^{-1/\xi},$$

subject to the constraints

$$1+\xi\left(\frac{M_n^{(j)}-\mu}{\sigma}\right) > 0, \qquad j = 1, \ldots, m.$$

The maximum likelihood estimates $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ are obtained by numerically maximizing $L(\mu, \sigma, \xi)$ under the above constraints. In general, no closed-form solution exists for this optimization problem. In practice, numerical maximization is commonly performed using the `fgev` function from the `evd` package in R [30].

Implications for extreme value modeling

The block maxima approach provides a theoretically justified framework for modeling extremes and constitutes the classical foundation of extreme value analysis. Within the context of this work, it serves as a benchmark for comparing Bayesian inference methods and machine learning approaches for the prediction of extreme values.

3.2 Bayesian approach

Bayesian inference constitutes an alternative and powerful framework for parameter estimation in extreme value models, particularly well suited to situations where available data are limited and uncertainty plays a central role. In extreme value analysis, the number of observed block maxima is often small, which may lead to unstable estimates under classical inference methods. Bayesian methodology addresses this limitation by allowing the incorporation of prior information through a prior distribution on the model parameters.

Let θ denote the vector of unknown parameters associated with the extreme value model under consideration. In the Bayesian paradigm, prior knowledge or expert information about θ is encoded in a prior distribution $\pi(\theta)$. This prior distribution is then combined with the information provided

by the observed data through the likelihood function $L(\theta \mid \mathbf{x})$ to obtain the posterior distribution of the parameters. According to Bayes' theorem, the posterior distribution is given by

$$p(\theta \mid \mathbf{x}) = \frac{L(\theta \mid \mathbf{x})\,\pi(\theta)}{\int L(\theta \mid \mathbf{x})\,\pi(\theta)\,d\theta},$$

where $\mathbf{x}$ denotes the observed data and the denominator serves as a normalizing constant.

Within the block maxima framework described in Section 3.1, Bayesian inference is applied to the parameters of the Generalized Extreme Value (GEV) distribution. This approach provides a coherent probabilistic description of parameter uncertainty and forms the basis for predictive inference on future extreme events. In particular, the Bayesian framework naturally leads to predictive distributions for future block maxima, which are of primary interest in the prediction problem formulated in Section 2.

The flexibility of Bayesian inference allows for the use of informative or weakly informative prior distributions, depending on the availability of prior knowledge, and facilitates uncertainty quantification through posterior summaries. These features make Bayesian methods especially attractive for extreme value modeling in climatic applications, where data scarcity and high-impact events are common.

3.2.1 Bayesian formulation for the Gumbel model

In order to illustrate the Bayesian inference procedure in a simpler setting, we first consider the special case of the Generalized Extreme Value distribution corresponding to the Gumbel type ($\xi = 0$). This case is frequently used in practice and serves as a convenient starting point for Bayesian modeling of extremes.

Let

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

denote a sample of observed block maxima, assumed to be independent and identically distributed realizations from a Gumbel distribution with location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma > 0$, denoted by

$$X_i \sim \text{Gumbel}(\mu, \sigma), \qquad i = 1, \ldots, n.$$

The cumulative distribution function of the Gumbel distribution is given by

$$F_X(x; \mu, \sigma) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}, \qquad x \in \mathbb{R},$$

and the corresponding probability density function is

$$f_X(x; \mu, \sigma) = \frac{1}{\sigma}\exp\left\{-\frac{x-\mu}{\sigma} - \exp\left(-\frac{x-\mu}{\sigma}\right)\right\}.$$

Likelihood function. Assuming independence, the likelihood function for the parameter vector $\theta = (\mu, \sigma)$ is

$$L(\theta \mid \mathbf{x}) = \prod_{i=1}^{n} f_X(x_i; \mu, \sigma) = \frac{1}{\sigma^n}\exp\left\{-\sum_{i=1}^{n}\exp\left(-\frac{x_i-\mu}{\sigma}\right) - \sum_{i=1}^{n}\frac{x_i-\mu}{\sigma}\right\}. \tag{1}$$

Prior dis tributions. Following Rostami and Adam (2015), we assume independent prior distributions for the parameters μ and σ, of the form

$$p(\mu, \sigma) = p(\mu)\, p(\sigma).$$

The prior distribution for the location parameter μ is taken as

$$p(\mu) \propto \exp\left\{ -\exp\left(-\frac{\mu - \mu_0}{\sigma_0} \right) - \frac{\mu - \mu_0}{\sigma_0} \right\},$$

which corresponds to a Gumbel-type prior with hyperparameters $\mu_0 \in \mathbb{R}$ and $\sigma_0 > 0$. For the scale parameter σ, we assume the prior

$$p(\sigma) \propto \sigma \exp\left(-\frac{\sigma^2}{2\lambda_0^2} \right), \qquad \sigma > 0,$$

where $\lambda_0 > 0$ is a scale hyperparameter. These choices ensure positivity of σ and yield a proper prior distribution.

Posterior distribution. Combining the likelihood and the prior distributions, the joint posterior density of (μ, σ) is given, up to a normalizing constant, by

$$p(\mu, \sigma \mid \mathbf{x}) \propto L(\theta \mid \mathbf{x})\, p(\mu)\, p(\sigma).$$

Substituting the expressions above, the posterior distribution can be written as

$$\begin{aligned} p(\mu, \sigma \mid \mathbf{x}) \;&\propto\; \frac{1}{\sigma^n} \exp\left\{ -\sum_{i=1}^{n} \exp\left(-\frac{x_i - \mu}{\sigma} \right) - \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} \right\} \\ &\quad \times \exp\left\{ -\exp\left(-\frac{\mu - \mu_0}{\sigma_0} \right) - \frac{\mu - \mu_0}{\sigma_0} \right\} \sigma \exp\left(-\frac{\sigma^2}{2\lambda_0^2} \right) \\ &= \frac{1}{\sigma^{n-1}} \exp\left\{ -\sum_{i=1}^{n} \exp\left(-\frac{x_i - \mu}{\sigma} \right) - \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} - \exp\left(\frac{\mu - \mu_0}{\sigma} \right) - \frac{\mu - \mu_0}{\sigma} - \frac{\sigma^2}{2\lambda_0^2} \right\} \end{aligned}$$

Closed-form expressions for posterior summaries are not available, and inference must therefore rely on numerical methods. In practice, posterior sampling is commonly performed using Markov chain Monte Carlo techniques, which allow for estimation of posterior moments, credible intervals, and predictive quantities.

3.2.2 Metropolis Hastings algorithm

In order to summarize and explore the posterior distribution $p(\mu, \sigma \mid \mathbf{x})$, closed-form expressions are generally unavailable, and numerical methods are required. A standard approach consists of using Monte Carlo sampling techniques, in particular the Metropolis–Hastings (MH) algorithm [17, 22], which belongs to the class of Markov chain Monte Carlo (MCMC) methods.

The MH algorithm generates a Markov chain whose stationary distribution coincides with the target posterior distribution. This makes it possible to approximate posterior summaries of the parameters and, consequently, to perform inference and prediction of extreme values using both simulated and real data.

Let $\mathbf{x} = (x_1, \ldots, x_n)$ denote the observed block maxima, assumed to follow a Gumbel(μ, σ) distribution as described in Section 3.2.1. The marginal posterior distributions for the parameters

μ and σ are obtained by ignoring terms in the joint posterior density that do not involve the parameter of interest [13]. Up to proportionality, these marginals are given by

$$p(\mu \mid \mathbf{x}, \sigma) \propto \exp\left\{ -\sum_{i=1}^{n} \exp\left(-\frac{x_i - \mu}{\sigma} \right) - \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} - \exp\left(-\frac{\mu - \mu_0}{\sigma_0} \right) - \frac{\mu - \mu_0}{\sigma_0} \right\},$$

and

$$p(\sigma \mid \mathbf{x}, \mu) \propto \sigma^{1-n} \exp\left\{ -\sum_{i=1}^{n} \exp\left(-\frac{x_i - \mu}{\sigma} \right) - \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} - \frac{\sigma^2}{2\lambda_0^2} \right\}, \qquad \sigma > 0.$$

The Metropolis–Hastings algorithm proceeds as follows.

Step 1. Initialization. Choose initial values $\mu^{(0)}$ and $\sigma^{(0)}$.

Step 2. Iterative updates. Given that the Markov chain is currently at $(\mu^{(j)}, \sigma^{(j)})$, proceed as follows.

- Draw a candidate value for the location parameter from a normal proposal distribution,

  $$\mu^{\text{can}} \sim \mathcal{N}\left(\mu^{(j)},\, \tau_\mu^2 \right),$$

  where τ_μ^2 is a suitably chosen tuning variance.

  The candidate μ^{can} is accepted with probability

  $$\alpha_\mu = \min\left\{ 1,\ \frac{p(\mu^{\text{can}} \mid \mathbf{x}, \sigma^{(j)})}{p(\mu^{(j)} \mid \mathbf{x}, \sigma^{(j)})} \right\},$$

  and the update is given by

  $$\mu^{(j+1)} = \begin{cases} \mu^{\text{can}}, & \text{if } u_1 < \alpha_\mu, \\ \mu^{(j)}, & \text{otherwise,} \end{cases}$$

  where $u_1 \sim \mathcal{U}(0, 1)$.

  For numerical stability, the acceptance probability is evaluated using the logarithm of the ratio,

  $$\begin{aligned} \log\left(\frac{p(\mu^{\text{can}} \mid \mathbf{x}, \sigma^{(j)})}{p(\mu^{(j)} \mid \mathbf{x}, \sigma^{(j)})} \right) \;=\; & \frac{n}{\sigma^{(j)}} \left(\mu^{\text{can}} - \mu^{(j)} \right) + \frac{\mu^{(j)} - \mu^{\text{can}}}{\sigma_0} \\ & + \exp\left(-\frac{\mu^{(j)} - \mu_0}{\sigma_0} \right) - \exp\left(-\frac{\mu^{\text{can}} - \mu_0}{\sigma_0} \right) \\ & + \sum_{i=1}^{n} \left[\exp\left(-\frac{x_i - \mu^{(j)}}{\sigma^{(j)}} \right) - \exp\left(-\frac{x_i - \mu^{\text{can}}}{\sigma^{(j)}} \right) \right]. \end{aligned}$$

- Given the updated value $\mu^{(j+1)}$, draw a candidate value for the scale parameter,

  $$\sigma^{\text{can}} \sim \mathcal{N}\left(\sigma^{(j)},\, \tau_\sigma^2 \right),$$

restricted to $\sigma^{\text{can}} > 0$.

The candidate σ^{can} is accepted with probability

$$\alpha_\sigma = \min\left\{1,\ \frac{p(\sigma^{\text{can}} \mid \mathbf{x}, \mu^{(j+1)})}{p(\sigma^{(j)} \mid \mathbf{x}, \mu^{(j+1)})}\right\},$$

with the update

$$\sigma^{(j+1)} = \begin{cases} \sigma^{\text{can}}, & \text{if } u_2 < \alpha_\sigma, \\ \sigma^{(j)}, & \text{otherwise,} \end{cases}$$

where $u_2 \sim \mathcal{U}(0,1)$.

Again, the acceptance probability is computed via the log-ratio

$$\begin{aligned} \log \frac{p(\sigma^{\text{can}} \mid \mathbf{x}, \mu^{(j+1)})}{p(\sigma^{(j)} \mid \mathbf{x}, \mu^{(j+1)})} &= (1-n)\left[\log(\sigma^{\text{can}}) - \log(\sigma^{(j)})\right] + \frac{(\sigma^{(j)})^2 - (\sigma^{\text{can}})^2}{2\lambda_0^2} \\ &\quad + \sum_{i=1}^{n} (x_i - \mu^{(j+1)}) \left(\frac{1}{\sigma^{(j)}} - \frac{1}{\sigma^{\text{can}}}\right) \\ &\quad + \sum_{i=1}^{n} \left[\exp\left(-\frac{x_i - \mu^{(j+1)}}{\sigma^{(j)}}\right) - \exp\left(-\frac{x_i - \mu^{(j+1)}}{\sigma^{\text{can}}}\right)\right]. \end{aligned}$$

Step 3. Iteration. Repeat Step 2 for a sufficiently large number of iterations, discarding an initial burn-in period and retaining the remaining samples to approximate the posterior distribution.

Posterior samples obtained via the Metropolis–Hastings algorithm can be used to compute point estimates, credible intervals, and predictive distributions for extreme values. This approach enables inference on maxima and minima under both simulated scenarios and real climatic datasets, in line with the objectives of this study.

3.2.3 Hamiltonian Monte Carlo algorithm

The Hamiltonian Monte Carlo (HMC) method was originally introduced by Duane et al. [8] in the context of molecular dynamics; see Neal [23] for a comprehensive theoretical and practical treatment. HMC improves upon the Metropolis–Hastings (MH) algorithm by employing a directed proposal mechanism that exploits gradient information of the log-posterior distribution to guide the Markov chain toward regions of higher posterior density. As a consequence, the HMC algorithm typically exhibits substantially higher acceptance rates and reduced random-walk behavior compared to traditional random-walk MH schemes, particularly in complex or correlated parameter spaces [10].

Let $\theta \in \mathbb{R}^d$ denote the vector of model parameters and let $p(\theta \mid \mathbf{x})$ be the posterior distribution of interest. In Hamiltonian Monte Carlo, an auxiliary momentum variable $\omega \in \mathbb{R}^d$ is introduced, independent of θ, with distribution

$$\omega \sim \mathcal{N}(0, \Sigma),$$

where Σ is a symmetric positive definite mass matrix, typically chosen as a diagonal matrix with constant elements, for instance $\Sigma = I_d$ [16].

Under this construction, the joint distribution of (θ, ω) is defined up to a normalizing constant and factorizes as

$$p(\theta, \omega \mid \mathbf{x}) \propto p(\theta \mid \mathbf{x})\, p(\omega).$$

The parameter vector θ is interpreted as the position of a particle, while ω represents its momentum. The potential energy of the system is defined as

$$U(\theta) = -\log p(\theta \mid \mathbf{x}),$$

and the kinetic energy associated with the momentum variable is given by

$$K(\omega) = \frac{1}{2}\, \omega^\top \Sigma^{-1} \omega.$$

The total energy of the system, known as the Hamiltonian, is therefore defined as

$$H(\theta, \omega) = U(\theta) + K(\omega).$$

Consequently, the (unnormalized) joint density of (θ, ω) can be written as

$$p(\theta, \omega \mid \mathbf{x}) \propto p(\theta \mid \mathbf{x})\, p(\omega) \propto p(\theta \mid \mathbf{x}) \exp\left(-\frac{1}{2}\omega^\top \Sigma^{-1} \omega\right) \propto \exp[-H(\theta, \omega)]\,,$$

where proportionality holds up to a constant independent of (θ, ω).

Hamiltonian dynamics. For continuous time t, the evolution of the system is governed by the Hamiltonian dynamics equations

$$\frac{d\theta}{dt} = \frac{\partial H(\theta, \omega)}{\partial \omega} = \Sigma^{-1}\omega, \qquad \frac{d\omega}{dt} = -\frac{\partial H(\theta, \omega)}{\partial \theta} = \nabla_\theta \log p(\theta \mid \mathbf{x}),$$

where $\nabla_\theta \log p(\theta \mid \mathbf{x})$ denotes the gradient of the log-posterior distribution with respect to θ.

Leapfrog integration. In practice, the Hamiltonian dynamics cannot be solved analytically and must be approximated numerically. A solution of Hamilton's equations defines a deterministic trajectory, or path, in the augmented space (θ, ω) along which the total energy is approximately conserved. Within an MCMC iteration, this trajectory is used to generate candidate values of the parameter vector θ. One standard approach to simulate such dynamics is to discretize time using the Leapfrog integrator [20], which preserves volume and reversibility.

Given a step size $\varepsilon > 0$, the Leapfrog updates are given by

$$\begin{aligned}
\omega^{(t+\varepsilon/2)} &= \omega^{(t)} + \frac{\varepsilon}{2}\, \nabla_\theta \log p(\theta^{(t)} \mid \mathbf{x}), \\
\theta^{(t+\varepsilon)} &= \theta^{(t)} + \varepsilon\, \Sigma^{-1} \omega^{(t+\varepsilon/2)}, \\
\omega^{(t+\varepsilon)} &= \omega^{(t+\varepsilon/2)} + \frac{\varepsilon}{2}\, \nabla_\theta \log p(\theta^{(t+\varepsilon)} \mid \mathbf{x}).
\end{aligned}$$

Acceptance step. Since the target distribution is the joint density of (θ, ω), the transition to a new candidate state $(\theta^{\text{can}}, \omega^{\text{can}})$ obtained after L Leapfrog steps is accepted with probability

$$\alpha\big\{(\theta, \omega), (\theta^{\text{can}}, \omega^{\text{can}})\big\} = \min\left\{1,\ \frac{p(\theta^{\text{can}}, \omega^{\text{can}} \mid \mathbf{x})}{p(\theta, \omega \mid \mathbf{x})}\right\}.$$

Using the Hamiltonian formulation, this acceptance probability can be written equivalently as

$$\alpha\big\{(\theta, \omega), (\theta^{\text{can}}, \omega^{\text{can}})\big\} = \min\left\{1,\ \exp[H(\theta, \omega) - H(\theta^{\text{can}}, \omega^{\text{can}})]\right\}.$$

Algorithm: Hamiltonian Monte Carlo. The Hamiltonian Monte Carlo (HMC) algorithm can be summarized as follows:

1. Give an initial position $\theta^{(0)}$ and set $i = 1$.
2. Draw a candidate momentum $\omega^{\text{can}} \sim \mathcal{N}_d(0, I_d)$ and an auxiliary variable $u \sim \mathcal{U}(0, 1)$.
3. Set
$$(\theta^{(I)}, \omega^{(I)}) = (\theta^{(i-1)}, \omega^{\text{can}}) \quad \text{and} \quad H_0 = H\left[\theta^{(I)}, \omega^{(I)}\right].$$
4. Repeat the Leapfrog solution L times:
   - Update the momentum:
   $$\omega^{\text{can}} = \omega^{\text{can}} + \frac{\varepsilon}{2} \nabla_\theta \mathcal{L}\left(\theta^{(i-1)}\right),$$
   where the gradient $\nabla_\theta \mathcal{L}(\theta^{(i-1)})$ is evaluated before the movement.
   - Update the position:
   $$\theta^{(i-1)} = \theta^{(i-1)} + \varepsilon\, \omega^{\text{can}}.$$
   - Update the momentum again:
   $$\omega^{(\text{can})} = \omega^{(\text{can})} + \frac{\varepsilon}{2} \nabla_\theta \mathcal{L}\left(\theta^{(i-1)}\right),$$
   where the gradient $\nabla_\theta \mathcal{L}(\theta^{(i-1)})$ is evaluated after the movement.
5. Set
$$(\theta^{(L)}, \omega^{(L)}) = (\theta^{(i-1)}, \omega^{\text{can}}), \quad \text{and} \quad H_1 = H\left[\theta^{(L)}, \omega^{(L)}\right].$$
6. Compute the acceptance probability:
$$\alpha_{\text{HMC}} = \alpha\left[(\theta^{(I)}, \omega^{(I)}), (\theta^{(L)}, \omega^{(L)})\right] = \min\{1, \exp(H_0 - H_1)\}.$$
7. Set $\theta^{(i)} = \theta^{(L)}$ if $\alpha_{\text{HMC}} > u$, and $\theta^{(i)} = \theta^{(I)}$ otherwise.
8. Set $i = i + 1$ and return to Step 2 until convergence.

Gradients for the Gumbel model. For the Gumbel model introduced in Section 3.2.1, let
$$Z_i = \frac{x_i - \mu}{\sigma}, \qquad i = 1, \ldots, n.$$
By taking log in (1), the log-likelihood function is obtained
$$\mathcal{L}(\mu, \sigma) = -n \log \sigma - \sum_{i=1}^{n} Z_i - \sum_{i=1}^{n} \exp(-Z_i).$$
The gradients with respect to the parameters are given by
$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{1}{\sigma}\left(n - \sum_{i=1}^{n} \exp(-Z_i)\right),$$
and
$$\frac{\partial \mathcal{L}}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)\left[1 - \exp(-Z_i)\right].$$

Reparametrization. To ensure positivity of the scale parameter, we apply the reparametrization

$$\eta = \log \sigma, \qquad \sigma = \exp(\eta).$$

Under this transformation, the gradient with respect to η becomes

$$\frac{\partial \mathcal{L}}{\partial \eta} = -n + \sum_{i=1}^{n} Z_i \left[1 - \exp(-Z_i)\right].$$

The gradient vector used in HMC is therefore

$$\nabla_\theta \mathcal{L}(\mu, \eta) = \left(\frac{\partial \mathcal{L}}{\partial \mu}, \frac{\partial \mathcal{L}}{\partial \eta} \right).$$

The HMC framework provides an efficient alternative to Metropolis–Hastings for Bayesian inference in extreme value models and is particularly well suited for high-dimensional parameter spaces and strongly correlated posterior distributions.

3.3 Machine Learning

While classical and Bayesian extreme value models provide a solid theoretical framework for inference and prediction, they rely on explicit distributional assumptions and often require careful model specification. As an alternative and complementary approach, machine learning methods offer flexible, data-driven tools capable of capturing complex temporal and multivariate dependencies without the need for predefined parametric forms.

In the context of extreme event prediction, machine learning models can be trained to learn latent representations of the underlying dynamics directly from the data, potentially improving predictive performance when the data-generating mechanisms are highly nonlinear or non-stationary. In this subsection, we focus on recurrent neural network architectures, which are particularly well suited for time series analysis. We begin by describing the Long Short-Term Memory (LSTM) model and its integration with extreme value modeling through a probabilistic formulation.

3.3.1 Long Short-Term Memory

Long Short-Term Memory (LSTM) networks are a class of recurrent neural networks (RNNs) specifically designed to process sequential data and to overcome the vanishing gradient problem that affects standard RNNs. By incorporating a gated memory mechanism, LSTM networks are able to capture and retain long-term temporal dependencies, which makes them particularly suitable for time series modeling and prediction.

In this work, LSTM models are employed as a data-driven alternative to classical extreme value models. The objective is to learn latent temporal representations that are informative for the prediction of extreme events. The formulation presented below follows the standard LSTM architecture, as described for example in [31]

LSTM cell formulation. Let $x_t \in \mathbb{R}^p$ denote the input vector at time t, and let $h_{t-1} \in \mathbb{R}^d$ and $C_{t-1} \in \mathbb{R}^d$ denote the hidden state and the cell state at time $t-1$, respectively. The LSTM cell is defined through the following gating mechanisms.

- **Forget gate.** The forget gate controls which information from the previous cell state is retained:

$$f_t = \sigma\Big(W^{(f)}(h_{t-1}, x_t) + b^{(f)} \Big),$$

where $W^{(f)}$ and $b^{(f)}$ are the weight matrix and bias vector, respectively, and $\sigma(\cdot)$ denotes the logistic sigmoid function.

- **Input gate.** The input gate regulates the incorporation of new information into the cell state:
$$i_t = \sigma\Big(W^{(i)}(h_{t-1}, x_t) + b^{(i)}\Big).$$
- **Cell state update.** Candidate values for the cell state are computed as
$$C_t^{\text{cad}} = \tanh\Big(W^{(C)}(h_{t-1}, x_t) + b^{(C)}\Big),$$
and the updated cell state is given by
$$C_t = f_t \odot C_{t-1} + i_t \odot C_t^{\text{cad}},$$
where $\odot$ denotes element-wise multiplication.
- **Output gate.** The output gate determines the output of the LSTM unit as a function of the updated cell state. It is given by
$$o_t = \sigma\Big(W^{(o)}(h_t, x_t) + b^{(o)}\Big),$$
and the hidden state is computed as
$$h_t = o_t \tanh(C_t),$$
where o_t denotes the output gate activation, $W^{(o)}$ is the corresponding weight matrix, $b^{(o)}$ is the bias term, and h_t represents the hidden state output of the LSTM unit.

The collection of LSTM parameters is denoted by
$$\theta = \Big(W^{(f)}, b^{(f)}, W^{(i)}, b^{(i)}, W^{(C)}, b^{(C)}, W^{(o)}, b^{(o)}\Big).$$

LSTM as a latent Markov representation for extremes. To model the temporal evolution of extreme events, we adopt a Markovian perspective in which the distribution of future extremes depends on a state variable summarizing past information. In a classical Markov setting, the sequence of extreme values is described by the conditional distribution
$$p(X_{n+1} = x \mid X_n = y) = p(X_n = x \mid X_{n-1} = y),$$
which assumes that the distribution of the extreme value depends only on the state attained at the previous time step [14]. While this assumption provides a parsimonious description, it may be too restrictive when complex temporal dependencies are present.

In the proposed approach, the Long Short-Term Memory (LSTM) network learns a latent representation that acts as a Markov state. Specifically, the observed past is replaced by the hidden state of the LSTM,
$$X_{t-1} \;\longrightarrow\; h_t = f_\theta(x_{t-\Delta}, \ldots, x_t),$$
where $f_\theta(\cdot)$ denotes the nonlinear mapping induced by the LSTM with parameters θ.

Under this formulation, the Markov property is expressed in terms of the latent state, leading to the approximation
$$p(X_t \mid X_{t-1}) \;\approx\; p(X_t \mid h_t),$$
which allows the conditional distribution of extreme values to depend on a rich, data-driven summary of past observations rather than on a single lagged value.

Gumbel likelihood parameterized by LSTM features. To connect the latent representations learned by the LSTM network with extreme value modeling, we assume that the target variable X_t, conditional on the hidden state h_t, follows a Gumbel distribution. Specifically, we consider

$$X_t \mid h_t \sim \text{Gumbel}(\mu_t, \sigma), \qquad \mu_t = \mu(h_t) = \omega^\top h_t,$$

where $\omega \in \mathbb{R}^d$ is a parameter vector that links the latent features to the location parameter of the distribution, and $\sigma > 0$ is a scale parameter.

This formulation allows the distribution of extreme values to vary dynamically over time through the hidden state h_t, which summarizes relevant information from past observations. As discussed in the previous subsection, the hidden state is obtained as

$$h_t = f_\theta(x_{t-\Delta}, \ldots, x_t),$$

where $f_\theta(\cdot)$ denotes the nonlinear mapping implemented by the LSTM with parameters θ.

For simplicity and numerical stability, we fix the scale parameter to $\sigma = 1$. Under this assumption, the negative log-likelihood of a single observation X_t given the latent state h_t and parameter vector ω is given by

$$-\log p(X_t \mid h_t, \omega) = -X_t + \omega^\top h_t + \exp\Big(-\big(X_t - \omega^\top h_t\big)\Big).$$

Optimization objective. Let

$$x_i = (x_{i(t-\Delta)}, \ldots, x_{it}) \in \mathbb{R}^{\Delta+1}$$

denote the i-th input window extracted from the time series, and let $h_{it} \in \mathbb{R}^d$ be the corresponding hidden representation produced by the LSTM. The parameters of the model consist of the LSTM parameters θ and the Gumbel location parameter vector ω.

The global objective function is defined as the minimization of the cumulative negative log-likelihood over all windows and time indices:

$$\min_{\theta,\omega} \sum_i \sum_t -\log p(x_{it} \mid h_{it}, \omega) = \min_{\theta,\omega} \sum_i \sum_t \Big[-x_{it} + \omega^\top h_{it} + \exp\Big(-\big(x_{it} - \omega^\top h_{it}\big)\Big)\Big].$$

This optimization problem corresponds to a probabilistic regression task for the location parameter of the Gumbel distribution, where the LSTM network learns latent temporal features that dynamically parameterize extreme value behavior. As a result, the proposed framework provides a flexible, data-driven approach for extreme value prediction in complex and potentially non-stationary time series.

3.3.2 Gated Recurrent Unit (GRU)

Gated Recurrent Units (GRUs) are a class of recurrent neural networks designed to process sequential data, similar to LSTM networks but with a simpler internal structure. GRUs regulate the flow of information through update and reset gates, enabling the network to capture long-term dependencies while reducing the number of parameters and computational cost.

In this work, we adopt the deep learning approach proposed by [25] for the prediction of extreme values based on block maxima. The method consists of embedding the Generalized Extreme Value (GEV) distribution within a Gated Recurrent Unit (GRU) architecture, which is a type of recurrent neural network (RNN) designed to process sequential data such as time series [**?**].

GRU networks employ gating mechanisms to control the flow of information within the network, allowing it to decide which information should be retained and which should be discarded at each time step. Using the GRU architecture, the hidden representation extracted from the input time series X_{kl} is used to estimate the parameters μ_{kl}, σ_{kl}, and ξ_{kl} of the GEV distribution associated with the block maximum M_{kl}.

Step 1: GRU encoding of the input time series. Let X_{kl} denote the input multivariate time series up to time kl, as defined in Section 2. The sequence X_{kl} is fed into a GRU network in order to obtain a latent representation:

$$h_{kl} = \mathrm{GRU}(X_{kl}),$$

where $h_{kl} \in \mathbb{R}^d$ represents the hidden state of the GRU at time kl.

Step 2: Mapping the GRU state to GEV parameters. The hidden state h_{kl} is then used as input to a two-layer fully connected neural network, which produces the parameters of the GEV distribution associated with the block maximum at time kl. Specifically, we define

$$(\mu_{kl}, \sigma_{kl}, \xi_{kl})^\top = W_2\big(W_1 h_{kl} + b_1\big) + b_2,$$

where W_1, b_1, W_2, and b_2 are trainable parameters of the neural network. We denote by

$$\theta = (W_1, b_1, W_2, b_2)$$

the collection of parameters associated with this mapping.

Step 3: Predictive distribution of block maxima. Given the estimated parameters $(\mu_{kl}, \sigma_{kl}, \xi_{kl})$, the predicted block maximum $\widehat{M}_{kl}$ is assumed to follow a GEV distribution:

$$\widehat{M}_{kl} \mid \mu_{kl}, \sigma_{kl}, \xi_{kl} \sim \mathrm{GEV}(\mu_{kl}, \sigma_{kl}, \xi_{kl}).$$

Learning procedure and loss function. The learning process consists of minimizing the negative log-likelihood of the observed block maxima $\{M_{kl}\}$ under the predicted GEV distribution. The primary loss function is defined as

$$L_1 = -\sum_{k=1}^{n} \log \mathrm{GEV}(M_{kl}; \mu_{kl}, \sigma_{kl}, \xi_{kl}).$$

To enforce the support constraint of the GEV distribution,

$$1 + \xi_{kl}\frac{M_{kl} - \mu_{kl}}{\sigma_{kl}} > 0,$$

a penalty term is introduced, as proposed in [25]:

$$L_2 = \sum_{k=1}^{n} \max\left(-1 - \xi_{kl}\frac{M_{kl} - \mu_{kl}}{\sigma_{kl}},\, 0\right).$$

The final objective function to be minimized is given by

$$L = L_1 + \lambda L_2,$$

where $\lambda > 0$ is a hyperparameter controlling the influence of the penalty term.

The optimization of L is performed using stochastic gradient-based methods, such as the Adam optimizer [18]. This training strategy enables the GRU-based model to jointly learn temporal representations and the parameters of the GEV distribution, providing a flexible framework for predicting extreme values from time series data.

3.4 Evaluation Metrics

To assess the predictive performance of the proposed models, we consider a set of standard objective error measures commonly used in forecasting and extreme value prediction. These metrics quantify the discrepancy between the observed extreme values

$$M_1, \ldots, M_T,$$

and their corresponding predicted values

$$\widehat{M}_1, \ldots, \widehat{M}_T.$$

1. **Symmetric Mean Absolute Percentage Error (SMAPE).**

   The SMAPE measures the relative percentage difference between observed and predicted values in a symmetric manner, reducing the effect of scale. It is defined as

   $$\text{SMAPE} = \frac{100\%}{T} \sum_{t=1}^{T} \frac{|M_t - \widehat{M}_t|}{|M_t| + |\widehat{M}_t|}.$$

2. **Root Mean Square Error (RMSE).**

   The RMSE penalizes large prediction errors more heavily due to the quadratic term and is given by

   $$\text{RMSE} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \left(M_t - \widehat{M}_t\right)^2}.$$

3. **Mean Absolute Error (MAE).**

   The MAE measures the average magnitude of the prediction errors without emphasizing large deviations. It is defined as

   $$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} |M_t - \widehat{M}_t|.$$

4. **Mean Square Error (MSE).**

   The MSE corresponds to the squared version of the RMSE and is given by

   $$\text{MSE} = \frac{1}{T} \sum_{t=1}^{T} \left(M_t - \widehat{M}_t\right)^2.$$

In all cases, M_t denotes the observed extreme value at time t, while $\widehat{M}_t$ represents the corresponding predicted value obtained from the fitted extreme value model (e.g., GEV-, Gumbel-, or machine learning–based approaches). These metrics provide complementary perspectives on predictive accuracy and are used jointly to compare the performance of the competing models.

3.5 Data set

This section describes the datasets used for the implementation, training, and evaluation of the proposed algorithms for extreme value prediction. All modeling approaches considered in this work—classical EVT-based models, Bayesian inference methods, and machine learning architectures—are explicitly implemented and tested using these datasets.

4 Discussion

This section discusses the results obtained from the full implementation of the proposed algorithms for extreme value prediction. The comparison encompasses classical block maxima models, Bayesian inference methods based on MCMC and HMC, and machine learning approaches using LSTM and GRU architectures.

5 Conclusion

In this work, we developed, implemented, and evaluated a comprehensive set of algorithms for the prediction of extreme values in climatic time series. The proposed framework integrates classical Extreme Value Theory, Bayesian inference techniques, and modern machine learning models within a unified and coherent methodology.

Acknowledgments

References

[1] Amin, M., et al. (2015). Bayesian inference for extreme value models. *Journal of Applied Statistics*.

[2] Balkema, A. A., and de Haan, L. (1974). Residual life time at great age. *Annals of Probability*, 2, 792–804.

[3] Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2006). *Statistics of Extremes: Theory and Applications*. Wiley.

[4] Chavez-Demoulin, V., and Davison, A. C. (2012). Modelling time series extremes. *REVSTAT*, 10, 109–133.

[5] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer.

[6] Coles, S., and Tawn, J. (1996). A Bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society, Series C*, 45, 463–478.

[7] Ding, Y., et al. (2019). Deep learning for time series forecasting with extreme events. *Proceedings of the AAAI Conference on Artificial Intelligence*.

[8] Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2), 216–222.

[9] Fisher, R. A., and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24, 180–190.

[10] Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1), 110–120.

[11] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44, 423–453.

[12] Gómez, H. W., et al. (2019). Bayesian modeling of extreme events. *Computational Statistics*.

[13] González, R. G., Parra, M. I., Acero, F. J., and Martín, J. (2019). An improved method for the estimation of the Gumbel distribution parameters. *arXiv preprint* arXiv:1902.07963.

[14] Goytom, I., and Sankaran, K. (2019). Forecasting maxima in climate time series. In *Proceedings of the 9th International Workshop on Climate Informatics*.

[15] Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press.

[16] Hartmann, M., and Ehlers, R. S. (2017). Bayesian inference for generalized extreme value distributions via Hamiltonian Monte Carlo. *Communications in Statistics–Simulation and Computation*, 46(7), 5285–5302.

[17] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109.

[18] Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980.

[19] Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer.

[20] Leimkuhler, B., and Reich, S. (2004). *Simulating Hamiltonian Dynamics*. Cambridge University Press.

[21] Li, Y., et al. (2023). Deep learning approaches for extreme event forecasting. *Neural Computing and Applications*.

[22] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087–1092.

[23] Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC.

[24] Neves, C., and Cordeiro, G. M. (2020). Extreme value regression models. *Statistical Modelling*, 20, 487–514.

[25] Nishino, Y., et al. (2022). Time series forecasting with extreme value awareness. *IEEE Transactions on Neural Networks*.

[26] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3, 119–131.

[27] Riofrio, J., et al. (2023). Machine learning methods for climatic extreme prediction. *Environmental Modelling & Software*.

[28] Rydell, J. (2013). Applications of extreme value theory in environmental sciences. *Environmental Statistics*.

[29] Sakthivel, R., and Nandhini, S. (2024). Climate extremes and their socio-economic impacts. *Climate Dynamics*.

[30] Stephenson, A. G. (2002). *evd: Extreme value distributions*. R News, 2(2), 31–32.

[31] Xiao, J., Deng, T., and Bi, S. (2024). Comparative analysis of LSTM, GRU, and transformer models for stock price prediction. In *Proceedings of the International Conference on Digital Economy, Blockchain and Artificial Intelligence*.

[32] Zangana, H., and Obeyd, A. (2024). Time series prediction using HMM and recurrent neural networks. *Applied Artificial Intelligence*.

[33] Zhang, Q., et al. (2024). Extreme event adaptive GRU for multivariate time series forecasting. *Pattern Recognition*.