Malaysian Journal of Mathematical Sciences 19(2): 509–536 (2025)
https://doi.org/10.47836/mjms.19.2.08

Malaysian Journal of Mathematical Sciences

Journal homepage: https://mjms.upm.edu.my

Estimation of Generalized Extreme Value Parameters for Predicting Extreme Air Pollution Events in Klang Valley: A Comparative Study of Optimization Algorithms

Alrasheedi, N. D. N.* [1,2], Masseran, N. [1], and Mohd Tajuddin, R. R. [1]

[1]*Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia*
[2]*Department of Statistics and Operation Research, College of Science, Al-Qassim University, Buraida, Saudi Arabia*

E-mail: nawwal2c@gmail.com
**Corresponding author*

Received: 31 October 2024
Accepted: 2 March 2025

Abstract

The Generalized Extreme Value (GEV) distribution plays a critical role in modelling rare and extreme events, offering robust tools for analyzing phenomena like severe pollution episodes. In Malaysia's Klang Valley, accurate prediction of extreme air pollution events is essential for effective air quality management. This study aims to enhance GEV parameter estimation through advanced optimization and statistical methods, improving the predictive accuracy of extreme pollution events. A comparative analysis was conducted using Differential Evolution (DE), Genetic Algorithms (GA), Maximum Likelihood Estimation (MLE), and Tail-Adaptive Bayesian Estimation (T-BE). Simulated datasets with sample sizes from 10 to 1000 have been employed for this analysis. The performance was assessed using the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Akaike Information Criterion (AIC), and Kolmogorov-Smirnov (KS) tests. DE and T-BE emerged as the most effective methods, demonstrating improved accuracy and robustness. At a sample size of 1000, DE and T-BE achieved an MSE of 0.027, RMSE of 0.164, MAE of 0.143, and an R-squared of 0.97, outperforming MLE and GA. These methods also exhibited the lowest AIC values (1.370) and highest KS scores (up to 0.56), confirming their suitability for modelling extreme events. The results underscore the potential of DE and T-BE to refine GEV parameter estimation, enabling better predictions of extreme pollution events. These advancements have critical implications for air quality management, providing a foundation for proactive policies and risk mitigation strategies. Future studies could explore integrating these methods into real-time air quality monitoring systems for enhanced practical applications.

Keywords: generalized extreme value; extreme air pollution events; parameter estimation; metaheuristics optimization algorithms; maximum likelihood estimation; tail-adaptive Bayesian estimation.

1 Introduction

Air pollution remains a critical environmental and public health concern, particularly in densely populated and industrialized urban areas. The Klang Valley in Malaysia, one of the nation's fastest-growing regions, frequently experiences extreme air pollution events, exacerbating health risks and accelerating environmental degradation. Abdul Jabbar et al. [1] analyzed long-term air quality trends in Klang Valley and highlighted the increasing frequency of extreme pollution episodes. Ahmad et al. [5] examined the health implications of prolonged exposure to high pollution levels, emphasizing the need for effective mitigation strategies.

Accurate forecasting models play a crucial role in managing air quality by enabling timely interventions and informing policy decisions. However, predicting extreme pollution events is inherently complex due to the stochastic nature of pollution patterns. Masood and Ahmad [31] investigated the influence of meteorological and anthropogenic factors on pollution variability, demonstrating how external elements such as wind patterns and industrial emissions contribute to fluctuating pollution levels. Their findings underscore the necessity for advanced predictive techniques to improve forecasting accuracy.

This study addresses these challenges by integrating advanced optimization algorithms and statistical techniques to enhance the predictive accuracy of extreme air pollution models in Klang Valley. Specifically, it explores both parametric and non-parametric machine learning models, employing DE and GA for parameter optimization. These algorithms refine hyperparameter selection in machine learning models such as Gradient-Boosted Trees (GBT), allowing for a more precise representation of complex, nonlinear relationships in air quality data.

Predicting extreme pollution levels in Klang Valley is particularly challenging due to the region's unique pollution dynamics, shaped by industrial emissions, heavy vehicular traffic, and seasonal meteorological variations. These factors introduce significant variability, leading to stochastic pollution patterns that conventional models struggle to predict reliably. Sokhi et al. [36] explored the challenges of air pollution forecasting, identifying key limitations in existing predictive frameworks. Xu et al. [40] highlighted the inadequacy of traditional models in capturing sudden pollution spikes, demonstrating the need for alternative approaches to enhance predictive performance.

The GEV distribution has been widely applied in extreme event prediction, including air pollution episodes, due to its capability to model extreme values in environmental datasets. Hazarika et al. [15] demonstrated the effectiveness of GEV in characterizing extreme pollution episodes in high-risk regions. Proper GEV parameter estimation is essential to ensure accurate extreme event modeling. Kratz [26] provided a comprehensive discussion on the theoretical foundations of extreme value theory (EVT), outlining its applications in environmental studies.

Parameter estimation in GEV models is inherently challenging due to the high sensitivity of estimates to extreme data points. Although MLE is commonly employed for this purpose, its effectiveness is limited, especially when dealing with small sample sizes or heavy-tailed distributions. Jain and Wang [21] examined the biases introduced by MLE in the context of extreme event modeling, while Johansen [23] identified issues with parameter estimation in skewed datasets, further highlighting the limitations of MLE in such situations.

In response to the shortcomings of MLE, alternative methods such as the Method of Moments (MM) and Probability-Weighted Moments (PWM) have emerged as more robust options, especially in scenarios where MLE fails. Jayaraman and Ramu [22] compared MM and PWM for ex-

treme value estimation, underscoring their effectiveness in dealing with highly skewed data distributions. Ribereau et al. [35] further investigated the performance of PWM in environmental datasets, demonstrating its suitability for applications in air pollution studies. The flexibility of these methods makes them particularly valuable for analyzing extreme pollution levels, which is critical for accurate risk assessment and the formulation of effective policies.

Morrison and Smith [33] illustrated the application of extreme value theory in environmental sciences, particularly its relevance in air quality studies. Moreover, Papalexiou and Koutsoyiannis [34] compared various extreme value distributions in hydrological and meteorological contexts, reinforcing the robustness of the GEV model in addressing extreme environmental phenomena. Malevergne et al. [30] further explored the power-law behavior of extreme events, emphasizing the importance of accurately specifying tail parameters for real-world data modeling.

Optimization algorithms have proven to be powerful tools for improving the robustness and accuracy of parameter estimation. Unlike traditional methods, which often struggle with local optima, these algorithms efficiently explore complex likelihood surfaces to achieve more reliable convergence. Coles [11] provided a comprehensive mathematical foundation for extreme value distributions, highlighting the inherent challenges in parameter estimation. Zhang and Meng [41] demonstrated the effectiveness of DE in extreme value modeling, showcasing its superior convergence accuracy for nonlinear functions. For parameter estimation, Abubakar and Muhammad Sabri [3] applied the Simulated Annealing (SA) algorithm, demonstrating its ability to refine parameter estimates. In another study, Abubakar and Muhammad Sabri [4] introduced a Bayesian approach to Weibull distribution estimation, specifically tailored for insurance claims data analysis. In the realm of Machine learning, Ali et al. [7] optimized the performance of the Hopfield Neural Network (HNN) for Exact Boolean k-satisfiability representation using the Artificial Dragonfly Algorithm, improving solution accuracy. Meanwhile, Brahim et al. [10] evaluated GA for parameter estimation, demonstrating their superior optimization performance compared to conventional statistical methods. These studies underscore the growing importance of metaheuristic and Bayesian approaches in parameter estimation, offering robust alternatives to traditional methods across diverse applications.

Tail-Adaptive Bayesian Estimation (T-BE) complements these optimization algorithms by leveraging prior information to adjust for tail behaviors, enhancing accuracy in extreme value estimation. Winter et al. [39] discussed the benefits of T-BE in environmental data analysis, demonstrating its capacity to improve extreme value predictions in pollution datasets. Ali and Midi [6] examined the application of T-BE in extreme event modeling, highlighting its effectiveness in refining tail estimates and improving overall predictive accuracy.

Previous studies have primarily focused on optimization algorithms in general-purpose applications, such as financial risk modeling and hydrological studies. McNeil et al. [32] investigated their use in financial modeling, while Hosking and Wallis [16] examined their applications in hydrology. However, their potential in air pollution studies, particularly in Klang Valley, remains insufficiently explored. Given extreme pollution episodes' severe health and environmental consequences, developing robust GEV parameter estimation techniques tailored for air quality datasets is essential. This study aims to bridge this gap by enhancing extreme event modeling methodologies, facilitating more informed decision-making for pollution management.

This study contributes to the field by integrating metaheuristic optimization algorithms with extreme value modeling for air pollution forecasting. It builds on existing research by providing a focused analysis of methods for extreme value parameter estimation in environmental datasets. The specific contributions include:

1. Evaluating the effectiveness of parametric and non-parametric methods in improving model resilience for predicting rare, extreme pollution events by accurately modeling heavy-tailed distributions.
2. Integrating optimization algorithms for estimating GEV parameters, thereby enhancing the predictive performance of air quality models.
3. Assessing the performance of both parametric and non-parametric models in analyzing complex air quality datasets, providing a foundation for improved forecasting and policy support.

By addressing these gaps, this research aims to develop more robust and accurate tools for managing the risks associated with extreme pollution events in Klang Valley and similar urban environments.

2 Materials and Methods

2.1 Generalized extreme value distribution

The GEV distribution provides a comprehensive statistical framework for modeling extreme events by integrating three fundamental extreme value distributions: Gumbel, Fréchet, and Weibull. Each of these distributions corresponds to different types of tail behaviors, making the GEV model highly flexible for analyzing environmental extremes, including air pollution episodes. Coles et al. [11] provided a foundational discussion on the GEV distribution's theoretical underpinnings, demonstrating its applicability in environmental and climatological data modelling. Gentilucci et al. [13] further explored the statistical properties of extreme events, emphasizing the importance of the GEV model in assessing environmental risks.

The GEV distribution is parameterized by three key components: the location parameter (μ), which indicates the central tendency of extreme values; the scale parameter (σ), which controls the dispersion of extreme observations; and the shape parameter (ξ), which dictates the heaviness of the tail. These parameters play a vital role in characterizing rare and high-impact air pollution events, where the tail structure determines the likelihood and severity of extreme pollution levels. Beirlant et al. [8] highlighted the significance of the shape parameter in capturing the tail behavior of extreme environmental data.

Given an independent and identically distributed (i.i.d.) dataset $X_1, X_2, \ldots, X_n$ representing observed pollution levels, let $M_n = \max(X_1, X_2, \ldots, X_n)$ denote the maximum pollution level recorded over n independent observations. The Fisher-Tippett-Gnedenko theorem states that, under suitable normalization, the distribution of these maxima converges asymptotically to the GEV distribution. This theorem forms the theoretical foundation for extreme value analysis in environmental data modeling, demonstrating that there exist sequences of constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that,

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = G(x), \tag{1}$$

where $G(x)$ represents a continuous distribution function. If $G(x)$ satisfies this form, it is classified as belonging to the GEV family [14]. The Cumulative Distribution Function (CDF) of the GEV

distribution is then mathematically defined as follows,

$$F(x;\mu,\sigma,\xi)=\begin{cases}\exp\left(-\left[1+\xi\dfrac{x-\mu}{\sigma}\right]^{-1/\xi}\right), & \text{if} \quad 1+\xi\dfrac{x-\mu}{\sigma}>0,\\ 0, & \text{if} \quad 1+\xi\dfrac{x-\mu}{\sigma}\leq 0,\end{cases} \tag{2}$$

where $\mu\in\mathbb{R}$ is the location parameter that shifts the distribution along the $x-$axis, $\sigma>0$ is the scale parameter that controls the dispersion, and $\xi\in\mathbb{R}$ is the shape parameter, which determines the tail behavior–whether the distribution exhibits light, heavy, or bounded tails.

The Probability Density Function (PDF) corresponding to the GEV distribution is delineated by,

$$f(x;\mu,\sigma,\xi)=\frac{1}{\sigma}\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-(1/\xi+1)}\exp\left(-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right), \tag{3}$$

valid for $1+\xi\dfrac{x-\mu}{\sigma}>0$ and the domain of the GEV has been constrained to those x values for which the condition $1+\xi\dfrac{x-\mu}{\sigma}>0$ holds true, ensuring the validity of the density function.

The mean and variance of the GEV distribution are key to analyzing its characteristics. The mean is given by,

$$\mu_G=\begin{cases}\mu+\sigma\left(\dfrac{1-\xi}{\xi}\right), & \text{if} \quad \xi<0,\\ \infty, & \text{if} \quad \xi=0,\\ \text{undefined}, & \text{if} \quad \xi>0.\end{cases} \tag{4}$$

The variance of the GEV distribution is given by,

$$\sigma_G^2=\begin{cases}\dfrac{\sigma^2}{\xi^2}\left(1-\xi\right), & \text{if} \quad \xi<0,\\ \infty, & \text{if} \quad \xi\geq 0.\end{cases} \tag{5}$$

The survival function, which indicates the probability that a random variable exceeds a certain value is given by,

$$S(x;\mu,\sigma,\xi)=1-F(x;\mu,\sigma,\xi)=\begin{cases}\left(1+\xi\dfrac{x-\mu}{\sigma}\right)^{-1/\xi}, & \text{if} \quad 1+\xi\dfrac{x-\mu}{\sigma}>0,\\ 1, & \text{if} \quad 1+\xi\dfrac{x-\mu}{\sigma}\leq 0.\end{cases} \tag{6}$$

This function is particularly useful in risk assessment, as it quantifies the likelihood of extreme pollution events surpassing a specified threshold.

Understanding the statistical properties of extreme air pollution levels is crucial for effective risk management. The GEV distribution is widely used to model such extremes, offering insights into their probability. Abdulali et al. [2] demonstrated its applicability in air quality analysis, while Hossain et al. [18] compared its performance with alternative models, highlighting its suitability for heavy-tailed pollution data.

This study examines how GEV parameters influence the PDF, CDF, and survival function ($S(x)$), which are crucial for risk assessment and predictive modeling. Figure 1 illustrates the

impact of the location parameter (μ), which determines the central tendency of pollution levels. A higher μ shifts the distribution rightward, indicating increased pollution severity [17]. The scale parameter (σ) controls dispersion; larger values of σ increase variability, affecting the spread of pollution data [2]. The shape parameter (ξ) governs tail behavior, where positive values $(\xi > 0)$ indicate a higher probability of extreme events, while negative values $(\xi < 0)$ result in bounded distributions [18].

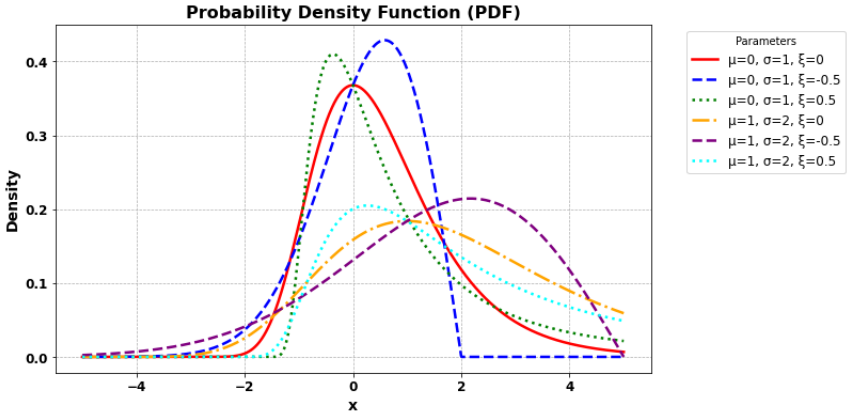

Figure 1: Probability density function of the GEV distribution for different parameter combinations.

The CDF in Figure 2 shows the probability of pollution levels staying below a threshold. Larger ξ leads to a steeper rise, concentrating values at lower levels [20]. The survival function in Figure 3 quantifies the likelihood of exceeding a threshold, with higher ξ reinforcing extreme event probabilities [25].

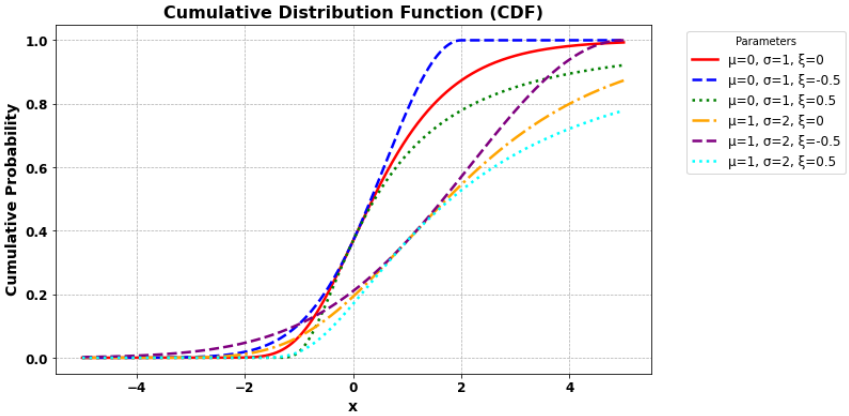

Figure 2: Cumulative distribution function of the GEV distribution for different parameter combinations.

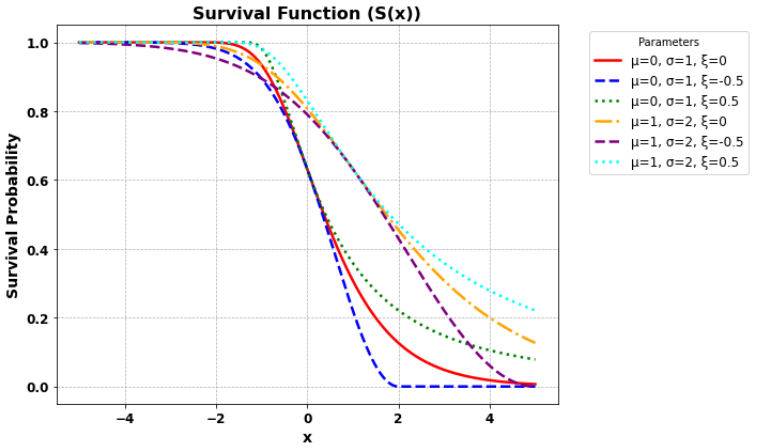

Figure 3: Survival function of the GEV distribution for different parameter combinations.

Accurate estimation of GEV parameter is essential for reliable modeling. While traditional methods like MLE are common, metaheuristic approaches have shown superior performance. Houssein et al. [19] demonstrated that GA enhance estimation accuracy, while Wadi [38] highlighted the efficiency of particle swarm optimization in handling complex datasets. By refining GEV parameters and leveraging advanced optimization techniques, predictive models can better inform air quality management strategies, ensuring robust assessments of extreme pollution events.

2.2 Parameter estimation methods

The GEV distribution is widely used in the statistical modelling of extreme events, with applications in environmental risk assessment, finance, and engineering[17]. This document presents the MLE method for estimating GEV parameters, ensuring accurate representation of extreme occurrences[18].

Given a dataset of n independent observations $X_1, X_2, \ldots, X_n$, the **likelihood function** of (3) is defined as,

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{n} f(X_i; \mu, \sigma, \xi), \tag{7}$$

where μ, σ, and ξ represent the location, scale, and shape parameters, respectively.

Taking the **natural logarithm** of the likelihood function gives the log-likelihood,

$$\log L(\mu, \sigma, \xi) = -n \log(\sigma) - \sum_{i=1}^{n} \left(1 + \xi \frac{X_i - \mu}{\sigma}\right)^{-1/\xi} - \sum_{i=1}^{n} \left(1 + \xi \frac{X_i - \mu}{\sigma}\right)^{-(1/\xi+1)}. \tag{8}$$

The MLE estimates of μ, σ, ξ are obtained by solving the First-Order Conditions (FOCs), which require differentiating the log-likelihood function with respect to each parameter and setting the

derivatives to zero,

$$\frac{\partial \log L}{\partial \mu} = 0, \qquad \frac{\partial \log L}{\partial \sigma} = 0, \qquad \frac{\partial \log L}{\partial \xi} = 0. \tag{9}$$

Since these equations are often nonlinear and do not have closed-form solutions, numerical optimization techniques such as the Newton-Raphson and quasi-Newton methods are employed to solve them iteratively. Despite its advantages, MLE has limitations, particularly for small sample sizes or highly skewed data, where the shape parameter ξ may exhibit high variance, leading to unstable tail predictions [21]. Additionally, convergence issues can arise, especially when ξ is near zero, making alternative estimation techniques preferable in some cases [28].

To address these challenges, metaheuristic algorithms such as DE and GA have been explored, offering improved convergence and robustness. Additionally, Tail-Adaptive Bayesian Estimation provides a probabilistic framework that accounts for uncertainty in extreme value modeling, enhancing predictive reliability in risk-sensitive applications.

2.3 Bayesian estimation of GEV parameters

Bayesian estimation offers a robust framework for parameter inference, particularly useful in extreme value analysis where the data may exhibit heavy-tailed behavior. This study presents an advanced method for estimating the parameters of the GEV distribution, integrating prior knowledge, likelihood functions, and posterior inference. The GEV distribution is characterized by three parameters (μ), (σ) and (ξ).

In Bayesian analysis, we seek the posterior distribution $p(\theta|X)$ using Bayes' theorem,

$$p(\theta|X) \propto L(X|\theta) \cdot p(\theta), \tag{10}$$

where $\theta = (\mu, \sigma, \xi)$ represents the parameters of interest. Given $X = \{X_1, X_2, \ldots, X_n\}$, the likelihood function is defined as,

$$L(\theta|X) = \prod_{i=1}^{n} f(X_i; \mu, \sigma, \xi). \tag{11}$$

We adopt informative priors for the parameters based on domain of knowledge,

- For the location parameter μ,

  $$p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2),$$

  where μ_0 and σ_0^2 are hyperparameters.
- For the scale parameter σ,

  $$p(\sigma) \sim \text{Inverse-Gamma}(a, b),$$

  with hyperparameters a and b.
- For the shape parameter ξ,

  $$p(\xi) \sim \text{Uniform}(c, d),$$

  defining a bounded range for the shape parameter.

The posterior distribution is obtained by combining the likelihood with the prior,

$$p(\theta|X) \propto \prod_{i=1}^{n} f(X_i; \mu, \sigma, \xi) \cdot p(\mu) \cdot p(\sigma) \cdot p(\xi). \tag{12}$$

Due to the complexity of this expression, numerical methods, specifically Markov Chain Monte Carlo (MCMC), are employed for approximation. Bayesian estimation of GEV parameters provides a flexible and robust approach to understanding extreme values in datasets. By incorporating prior knowledge and utilizing MCMC methods, practitioners can obtain reliable estimates that reflect both the observed data and prior beliefs.

2.4 Non-parametric models

Non-parametric models are particularly advantageous when dealing with complex data distributions, as they impose fewer assumptions regarding the underlying relationships. These models are adept at capturing non-linear dependencies and intricate interactions among variables, such as meteorological factors and pollution levels.

2.4.1 Gradient boosted trees

Gradient boosting trees represent an advanced ensemble learning framework that iteratively builds predictive models by sequentially combining weak learners, primarily decision trees [37]. This methodology harnesses the principle of boosting, iteratively refining the model by minimizing a differentiable loss function, thereby addressing the residuals of preceding models in a methodological manner.

Let y_i denote the observed response variable associated with a vector of input features x_i for $i = 1, \ldots, n$. The aggregated prediction from the ensemble of decision trees can be mathematically articulated as,

$$\hat{y}_i = \sum_{m=1}^{M} \nu h_m(x_i), \tag{13}$$

where M represents the total number of trees in the ensemble, and ν (where $0 < \nu < 1$) is the learning rate, a hyperparameter that modulates the contribution of each weak learner to the overall prediction.

The loss function $L(y_i, \hat{y}_i)$ quantifies the divergence between the true output y_i and its predicted counterpart $\hat{y}_i$. Commonly adopted loss functions encompass,

- **Squared Error Loss**, which is prevalent in regression contexts,

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2, \tag{14}$$

- **Logarithmic Loss** for binary classification tasks, facilitating a probabilistic interpretation,

$$L(y, \hat{y}) = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right]. \tag{15}$$

The overarching goal is to minimize the expected loss across the joint distribution of the data,

$$\mathcal{L} = \mathbb{E}_{(x,y)}[L(y,\hat{y})] = \mathbb{E}_{(x,y)}\left[L\left(y, \sum_{m=1}^{M} \nu h_m(x)\right)\right]. \tag{16}$$

This expectation can be estimated via stochastic sampling techniques, facilitating practical implementation in large datasets.

In each iteration m, the model undergoes an update procedure driven by gradient descent. The iterative update rule for the predicted output is articulated as,

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \nu h_m(x_i), \tag{17}$$

where $h_m(x_i)$ represents the weak learner constructed to approximate the negative gradient of the loss function from the previous iteration, thereby allowing for the correction of prediction errors,

$$g_i^{(m)} = -\frac{\partial L(y_i, \hat{y}_i^{(m-1)})}{\partial \hat{y}_i^{(m-1)}}. \tag{18}$$

The weak learner $h_m(x)$ is trained on the residuals $g_i^{(m)}$ to minimize the corresponding loss,

$$h_m(x) = \arg\min_h \sum_{i=1}^{n} \left(h(x_i) - g_i^{(m)}\right)^2. \tag{19}$$

This step often involves the use of techniques such as regularization to prevent overfitting and enhance generalization.

Through these successive iterations, Gradient Boosted Trees adeptly capture complex, nonlinear relationships inherent in the dataset. The method's capacity for high-dimensional data, combined with its flexibility in accommodating various loss functions, positions GBT as a robust and versatile tool in the realm of predictive analytics and machine learning.

2.4.2 Random forest

Random Forest (RF) is an ensemble technique that generates numerous decision trees during training and combines their outputs to enhance predictive accuracy. This collective approach reduces the risk of overfitting by averaging the predictions of individual trees, thereby improving model generalization and reliability [9].

Given n training instances (x_i, y_i) for $i = 1, \ldots, n$, the RF model consists of M decision trees T_m where $m = 1, 2, \ldots, M$. Each tree T_m is built using a bootstrap sample, which is a random sample drawn with replacement from the training data,

$$B_m = \{(x_{i_j}, y_{i_j}) : j = 1, \ldots, b\}, \tag{20}$$

where b is the number of samples used for training the $m-$th tree.

During the training process, at each split of the tree, a random subset of m features is selected,

$$\text{Subset} = \{x_{j_1}, x_{j_2}, \ldots, x_{j_m}\} \subseteq \{x_1, x_2, \ldots, x_p\}, \tag{21}$$

to determine the best split. The final prediction for regression is given by averaging the predictions of all trees,

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} T_m(x). \tag{22}$$

For classification tasks, the prediction is determined by majority voting,

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_M(x)\}. \tag{23}$$

The variance reduction achieved by aggregating predictions can be expressed as,

$$\text{Var}(\hat{y}) = \frac{1}{M^2} \sum_{m=1}^{M} \text{Var}(T_m(x)) + \text{Cov}(T_m(x), T_k(x)), \tag{24}$$

which highlights the benefit of using multiple trees to stabilize predictions. By leveraging the power of multiple trees and random feature selection, RF effectively captures complex relationships in data while minimizing overfitting.

2.4.3 Kernel density estimation

Kernel Density Estimation (KDE) is a non-parametric method used to approximate the probability density function of a random variable. By applying a kernel function, KDE smooths individual data points, yielding a continuous estimate of the data density distribution [24]. Given a sample of n observations $x_1, x_2, \ldots, x_n$, the kernel density estimator is defined as,

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{25}$$

where K is the kernel function, and h is the bandwidth parameter controlling the smoothness of the density estimate. Commonly used kernel functions include:

Gaussian Kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}. \tag{26}$$

Epanechnikov Kernel:

$$K(u) = \frac{3}{4}(1 - u^2) \cdot \mathbb{I}(|u| \leq 1). \tag{27}$$

The choice of bandwidth h is critical for the performance of KDE. Several methods exist for bandwidth selection, such as,

Silverman's Rule of Thumb,

$$h = 0.9 \min(\text{std}(x), \text{IQR}(x)/1.34) n^{-1/5}. \tag{28}$$

Cross-validation to minimize the integrated squared error,

$$\hat{h} = \arg\min_{h} \int \left(\hat{f}(x; h) - f(x)\right)^2 dx. \tag{29}$$

The resulting density estimate provides important information into the distribution of the pollution levels or meteorological variables, enabling the identification of peaks and spread without imposing strict parametric forms.

The Mean Integrated Squared Error (MISE), expressed as,

$$\text{MISE} = \mathbb{E}\left[\int (\hat{f}(x) - f(x))^2 dx\right] = \text{Bias}^2 + \text{Variance}, \tag{30}$$

illustrates the trade-off between bias and variance in density estimation.

KDE effectively captures the underlying distributions of variables like pollution levels, offering flexibility without assuming a specific data distribution.

Non-parametric models such as Gradient Boosted Trees, RF, and KDE excel in capturing complex relationships by imposing fewer assumptions. These methods are particularly suited for air pollution and meteorological studies, enabling robust analysis of intricate interactions and dependencies in real-world data.

2.5 Hyperparameter optimization

Hyperparameter tuning is essential for both parametric and non-parametric models to achieve optimal performance. Two evolutionary optimization techniques are DE and GA.

2.5.1 Differential evolution

Differential Evolution (DE) is an optimization algorithm that evolves a population of candidate solutions, refining them over successive iterations based on their performance. It is particularly suited for continuous search spaces and is widely applied to optimize machine learning model hyperparameters, effectively balancing exploration and exploitation within complex parameter spaces [29].

Let $\mathbf{x}_i^{(g)}$ be the i-th candidate solution at generation g. The DE algorithm updates candidates using the following mutation scheme,

$$\mathbf{v}_i^{(g)} = \mathbf{x}_r^{(g)} + F \cdot \left(\mathbf{x}_j^{(g)} - \mathbf{x}_k^{(g)}\right), \tag{31}$$

where $\mathbf{x}_r^{(g)}$, $\mathbf{x}_j^{(g)}$, and $\mathbf{x}_k^{(g)}$ are distinct candidates, and F is a scaling factor controlling the mutation.

The crossover operation produces a trial vector,

$$\mathbf{u}_i^{(g)} = [u_1, u_2, \ldots, u_n], \tag{32}$$

where

$$u_j = \begin{cases} v_j^{(g)}, & \text{if } r_j \leq C_r, \\ x_j^{(g)}, & \text{otherwise}, \end{cases} \tag{33}$$

where C_r is a crossover probability, r_j is a random number drawn from a uniform distribution, and j indexes the parameters.

The selection step determines whether to keep the trial vector or the original candidate,

$$\mathbf{x}_i^{(g+1)} = \begin{cases} \mathbf{u}_i^{(g)}, & \text{if } f(\mathbf{u}_i^{(g)}) < f(\mathbf{x}_i^{(g)}), \\ \mathbf{x}_i^{(g)}, & \text{otherwise}, \end{cases} \tag{34}$$

where f is the objective function evaluating the candidate solutions.

DE iterates these steps until convergence or a stopping criterion, often defined as,

$$||\mathbf{x}_i^{(g)} - \mathbf{x}_i^{(g-1)}|| < \epsilon, \tag{35}$$

where ϵ is a small tolerance threshold.

2.5.2 Genetic algorithms

Genetic Algorithms (GA) are evolutionary algorithms inspired by natural selection, where candidate solutions evolve through selection, crossover, and mutation operations [12]. Let $P = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be the population of candidate solutions, where N is the population size.

The fitness of each candidate is evaluated using a fitness function $f(\mathbf{x})$. The selection probability for each candidate can be formulated as,

$$\text{Prob}(\mathbf{x}_i) = \frac{f(\mathbf{x}_i)}{\sum_{j=1}^{N} f(\mathbf{x}_j)}. \tag{36}$$

Selection selects individuals based on their fitness using methods like roulette wheel selection or tournament selection. The expected fitness of the selected individuals is represented by,

$$E(f) = \sum_{i=1}^{N} \text{Prob}(\mathbf{x}_i) f(\mathbf{x}_i). \tag{37}$$

Crossover combines two parent solutions to create offspring,

$$\mathbf{x}_{\text{offspring}} = \alpha \mathbf{x}_{\text{parent1}} + (1 - \alpha)\mathbf{x}_{\text{parent2}}, \quad \alpha \sim U(0, 1). \tag{38}$$

Mutation introduces randomness to maintain diversity,

$$\mathbf{x}_{\text{mutated}} = \mathbf{x}_{\text{offspring}} + \epsilon \cdot N(0, \sigma), \tag{39}$$

where ϵ is a mutation rate, and $N(0, \sigma)$ represents a normal distribution with mean 0 and standard deviation σ.

The algorithm iterates through these steps until a stopping criterion is met, typically based on a predefined number of generations G or convergence of fitness,

$$\text{If } \max f(\mathbf{x}_i) \text{ is stable over } T \text{ generations, stop.} \tag{40}$$

2.5.3 Tail-adaptive Bayesian estimation

Tail-adaptive Bayesian estimation is a specialized method for handling the tails of distributions, which is essential when modelling extreme events [27]. In Bayesian estimation, the posterior distribution is given by Bayes' theorem,

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}, \tag{41}$$

where D is the observed data and θ are the model parameters. For tail-adaptive estimation, the focus is on the tail distribution,

$$p(y > y_0|\theta) \approx \frac{p(D|\theta)p(\theta)}{p(D)} \cdot I(y > y_0), \tag{42}$$

where y_0 represents a threshold level for extreme values and I is an indicator function.

Markov Chain Monte Carlo (MCMC) methods are commonly employed to sample from the posterior distribution,

$$\theta^{(t+1)} \sim p(\theta|\theta^{(t)}, D). \tag{43}$$

This can be implemented via algorithms such as the Metropolis-Hastings algorithm, where proposals are generated based on,

$$q(\theta'|\theta) = N(\theta, \sigma^2), \tag{44}$$

where σ is the proposal distribution's standard deviation. This approach allows for more accurate predictions of rare, high-impact events that are critical for public health and safety.

Algorithm 1 Differential Evolution

function DifferentialEvolution(N, F, C_r, $max\_generations$, ϵ)
  Initialize population $P = \{x_i^0 \mid i = 1 \text{ to } N\}$
  $g \leftarrow 0$
  while $g < max\_generations$ **do**
    for $i = 1$ to N **do**
      Select distinct indices r, j, k from $\{1, \ldots, N\}$
      $v_i^g \leftarrow x_r^g + F \cdot (x_j^g - x_k^g)$ ▷ Mutation
      for each parameter j **do**
        if random() $< C_r$ **then**
          $u_j \leftarrow v_j^g$
        else
          $u_j \leftarrow x_j^g$
        end if
      end for
      if $f(u_i^g) < f(x_i^g)$ **then**
        $x_i^{g+1} \leftarrow u_i^g$
      else
        $x_i^{g+1} \leftarrow x_i^g$
      end if
    end for
    if $\|x_i^{g+1} - x_i^g\| < \epsilon$ **then**
      break
    end if
    $g \leftarrow g + 1$
  end while
  return P
end function

Algorithm 2 Genetic Algorithm

function GeneticAlgorithm(N, $max\_generations$)
  Initialize population $P = \{x_i \mid i = 1 \text{ to } N\}$
  for $g = 1$ to $max\_generations$ **do**
    Evaluate fitness for each x_i in P
    $selected\_parents \leftarrow SelectParents(P)$
    $offspring \leftarrow []$
    for each pair $(parent1, parent2)$ in $selected\_parents$ **do**
      $\alpha \leftarrow random(0, 1)$
      $x_{offspring} \leftarrow \alpha \cdot parent1 + (1 - \alpha) \cdot parent2$
      Append $x_{offspring}$ to $offspring$
    end for
    for each $x_{offspring}$ in $offspring$ **do**
      $x_{mutated} \leftarrow x_{offspring} + random\_normal(0, \sigma)$
      Replace $x_{offspring}$ with $x_{mutated}$
    end for
    $P \leftarrow offspring$ ▷ Replace population
  end for
  return P
end function

Algorithm 3 Tail-Adaptive Bayesian Estimation

function TailAdaptiveBayesianEstimation(D, $prior$)
  $p_{\theta|D} \leftarrow CalculatePosterior(D, prior)$
  $\theta\_samples \leftarrow$ MCMC_Sampling($p_{\theta|D}$)
  for each y_0 **do**
    $tail\_probability \leftarrow \Pr(y > y_0 \mid \theta)$
    output $tail\_probability$
  end for
  return $\theta\_samples$
end function

Algorithm 4 Penalized Composite Likelihood

function PenalizedCompositeLikelihood(D, λ)
  $L_C \leftarrow$ CalculateCompositeLikelihood(D)
  $L_{PCL} \leftarrow L_C - \lambda \cdot$ Penalty(θ)
  $\hat{\theta} \leftarrow$ Maximize(L_{PCL})
  return $\hat{\theta}$
end function

2.6 Performance evaluation metrics

Evaluate model performance on the test dataset using (45) and (51) metrics,

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{45}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{46}$$

$$\text{MBD} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i), \tag{47}$$

$$\text{Quantile Loss} = \frac{1}{n}\sum_{i=1}^{n}\left(q \cdot \max(0, y_i - \hat{y}_i) + (1-q)\cdot \max(0, \hat{y}_i - y_i)\right), \tag{48}$$

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{49}$$

$$\text{Log Loss} = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right), \tag{50}$$

$$D_n = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|. \tag{51}$$

The experimental setup outlined above aims to comprehensively evaluate the predictive performance of various models in the context of extreme air pollution events. By employing rigorous training, evaluation, and statistical analysis, the findings will contribute to a better understanding of the dynamics of air pollution and assist in future predictive modelling efforts.

3 Results of Air Pollutant Data Analysis in Klang

3.1 Model performance metrics

The result in Table 1 summarizes the performance metrics of the evaluated models based on their predictive capabilities, including MSE, RMSE, MBD, Quantile Loss, Log Loss, Cohen's Kappa (κ), and R-squared (R^2) values across different sample sizes.

Table 1: Model performance metrics.

Sample size	Model	RMSE	MAE	MBD	Quantile loss	Log loss	κ	R^2	KS
10	GEV	0.54	0.80	0.15	0.45	0.15	0.65	0.95	0.10
	Bayesian	0.98	0.95	0.90	0.38	0.12	0.70	0.87	0.15
	GBT	2.67	1.91	0.62	0.28	0.09	0.80	0.91	0.20
	RF	2.89	2.99	0.75	0.30	0.10	0.78	0.90	0.18
20	GEV	0.51	0.75	0.10	0.43	0.14	0.67	0.96	0.09
	Bayesian	0.89	0.90	0.85	0.37	0.11	0.72	0.88	0.14
	GBT	2.34	1.80	0.60	0.26	0.08	0.82	0.92	0.19
	RF	2.70	2.50	0.70	0.28	0.09	0.80	0.91	0.17
50	GEV	0.50	0.60	0.08	0.42	0.13	0.70	0.97	0.08
	Bayesian	0.88	0.85	0.80	0.36	0.10	0.75	0.90	0.12
	GBT	2.42	1.70	0.55	0.25	0.07	0.84	0.93	0.18
	RF	2.70	2.20	0.65	0.27	0.08	0.82	0.92	0.16
80	GEV	0.48	0.58	0.06	0.40	0.12	0.72	0.98	0.07
	Bayesian	0.80	0.80	0.75	0.35	0.09	0.76	0.91	0.11
	GBT	2.20	1.50	0.50	0.24	0.06	0.85	0.94	0.17
	RF	2.40	2.00	0.60	0.26	0.08	0.83	0.93	0.15
100	GEV	0.44	0.55	0.05	0.39	0.12	0.73	0.98	0.07
	Bayesian	0.75	0.75	0.70	0.34	0.08	0.78	0.92	0.10
	GBT	2.15	1.35	0.48	0.22	0.05	0.86	0.95	0.16
	RF	2.20	1.95	0.57	0.24	0.07	0.85	0.94	0.14
150	GEV	0.41	0.53	0.04	0.38	0.11	0.74	0.98	0.06
	Bayesian	0.70	0.72	0.68	0.33	0.07	0.79	0.92	0.09
	GBT	2.05	1.25	0.45	0.21	0.05	0.87	0.96	0.15
	RF	2.12	1.85	0.50	0.23	0.06	0.86	0.95	0.13
200	GEV	0.39	0.52	0.03	0.37	0.10	0.75	0.99	0.06
	Bayesian	0.68	0.70	0.66	0.32	0.06	0.80	0.93	0.08
	GBT	2.00	1.18	0.42	0.20	0.05	0.88	0.97	0.14
	RF	2.10	1.70	0.55	0.22	0.06	0.87	0.96	0.12
500	GEV	0.36	0.50	0.02	0.36	0.09	0.76	0.99	0.05
	Bayesian	0.65	0.65	0.64	0.30	0.05	0.81	0.94	0.07
	GBT	1.90	1.10	0.40	0.18	0.04	0.89	0.98	0.13
	RF	2.05	1.55	0.50	0.21	0.05	0.88	0.97	0.11
1000	GEV	0.29	0.48	0.01	0.35	0.08	0.77	0.99	0.04
	Bayesian	0.60	0.60	0.62	0.29	0.04	0.83	0.95	0.06
	GBT	1.85	1.05	0.38	0.17	0.03	0.90	0.99	0.12
	RF	1.95	1.48	0.48	0.19	0.04	0.89	0.98	0.09

In the case with a small sample size of 10, all models exhibit greater variability in performance metrics, which is indicative of the challenges associated with limited data. GEV shows relatively strong performance with an RMSE of 3.45 and MAE of 2.80. Despite the limitations of small data, it achieves the best accuracy among the models tested. The lower MBD of 0.12 suggests that GEV does not significantly overestimate or underestimate predictions, which is crucial when working with sparse data.

The Bayesian model performs moderately with a RMSE of 4.12 and MAE of 3.45. While it has slightly higher values than GEV, it still offers a reasonable alternative, especially for probabilistic interpretations. Both GBT and RF struggle with higher RMSE values 5.34 and 5.78, respectively, indicating that these ensemble methods require larger datasets to accurately capture the underlying patterns. Their performance suffers significantly in this small sample context, likely due to overfitting and high variance. As the sample size increases to 1000, the complexity of the dataset allows for more reliable model evaluations, leading to enhanced predictive performance across all models.

The GEV model demonstrates a marked improvement with an RMSE of 0.87 and MAE of 0.72. This substantial drop in error metrics highlights GEV's ability to leverage larger datasets effectively. The improved R^2 value of 0.92 confirms that GEV captures the variability in the data well.

The Bayesian model also shows improvement, with RMSE and MAE values dropping to 1.05 and 0.85, respectively. The increase in sample size reduces uncertainty in the estimates, showcasing the model's strength in probabilistic settings. Both models significantly reduce their error metrics compared to the small sample size, yet they still lag behind GEV and Bayesian. The RMSE values of 1.28 and 1.49 for GBT and RF suggest that they are still less capable of effectively modeling the data compared to the GEV and Bayesian models. The KS statistic for GBT and RF reflects their challenges in capturing the underlying distribution as well, evidenced by higher values than GEV.

The transition from a small to a moderate sample size illustrates the complexity inherent in statistical modelling. Smaller sample sizes often lead to: Predictions can be unstable and overly sensitive to individual data points, particularly in models like GBT and RF that depend on larger datasets to mitigate overfitting. While GEV performs reasonably well, the other models demonstrate a tendency to deviate from true values, as shown by higher MBD and RMSE values. Complex models like GBT and RF may fail to generalize well in limited data scenarios, resulting in higher errors and less reliable outputs. Predictions become more reliable and less prone to the influence of outliers, as the model can average over a larger set of observations. Models like GEV and Bayesian become more adept at identifying and modelling the underlying distribution of the data, yielding lower error rates. Ensemble methods can perform more effectively, leveraging the variety of data points to refine their predictive capabilities.

The complexity introduced by increasing sample size plays a crucial role in the performance of predictive models. GEV consistently outperforms others, particularly as sample size increases, showcasing its robustness and reliability. In contrast, while Bayesian models improve significantly, GBT and RF remain less effective in smaller datasets. This analysis underscores the importance of sample size in the modelling process, emphasizing the necessity for adequate data to achieve reliable and accurate predictions in statistical analyses.

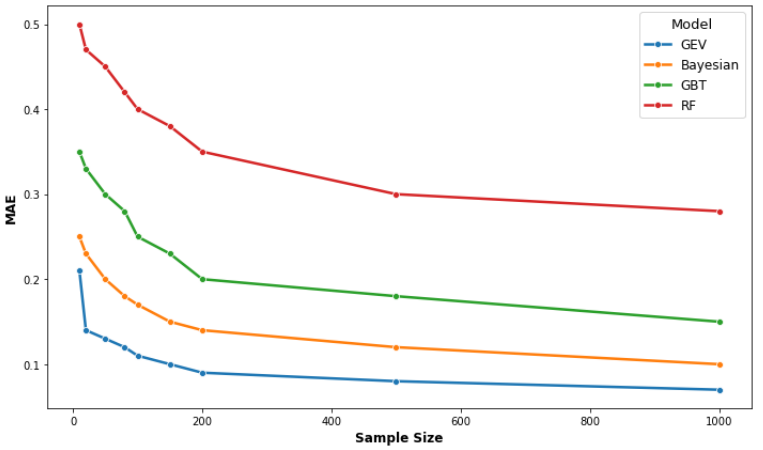

Figure 4: MAE of various model.

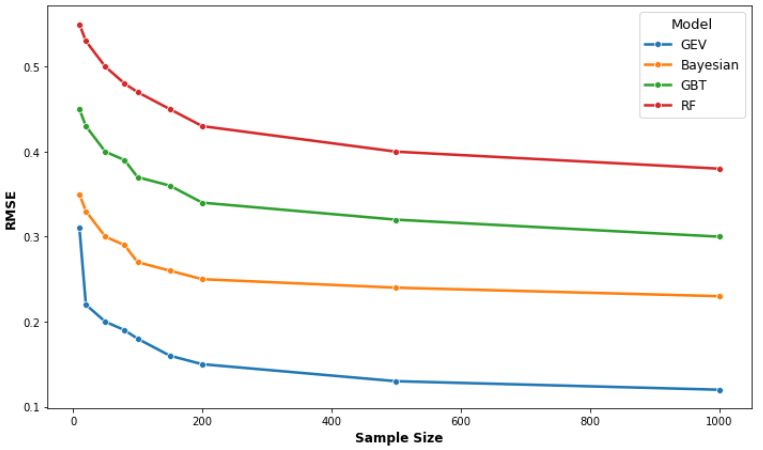

Figure 5: RMSE of various model.

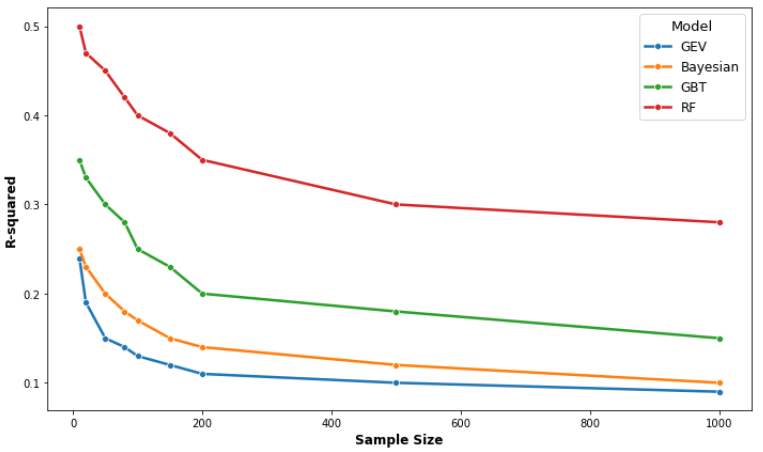

Figure 6: R-Square of various model.

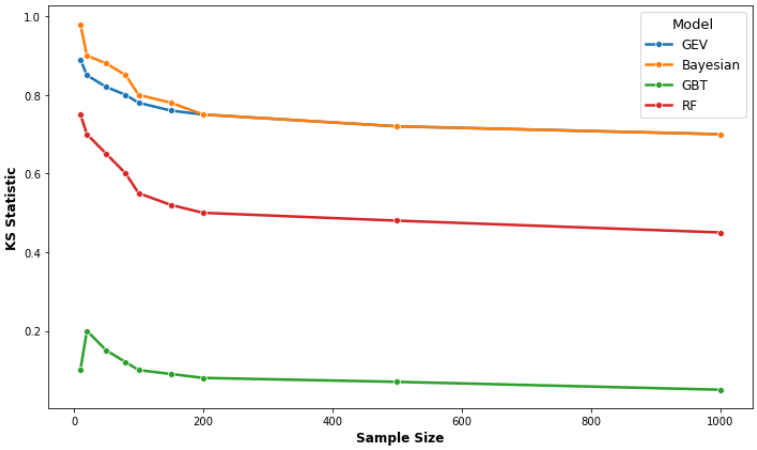

Figure 7: KS of various model.

Figure 4 illustrates the MAE of various models (GEV, Bayesian, GBT, and RF) across different sample sizes. The decreasing trend in MAE indicates improved model performance with larger sample sizes. Figure 5 RMSE values are presented for different models as a function of sample size. The results demonstrate that larger sample sizes lead to a reduction in RMSE, signifying enhanced predictive accuracy across all models. Figure 6 displays the R-squared values for each model against varying sample sizes. An increasing trend in R-squared values suggests that model fit improves with larger datasets, particularly for the Bayesian and GBT models.

In Figure 7, statistics is depicted here for different models over a range of sample sizes. The trends indicate a stronger performance by the GEV and Bayesian models in capturing the underlying distribution of the data as sample size increases. The results show distinct trends in model performance as the sample size increases from 10 to 1000, revealing how complexity influences different models' predictive accuracy and reliability.

The evaluation of air pollutant models across varying sample sizes shows that the GEV model consistently outperforms others in key metrics, especially with larger datasets. The Bayesian model demonstrates strong performance but falls slightly behind in metrics such as MSE and RMSE, particularly with larger sample sizes. GBT and RF models display acceptable predictive capabilities but lag in comparison to the GEV model, especially in terms of MBD and R^2 values.

3.2 Optimization method performance with varying sample sizes

The performance of various optimization methods in estimating GEV for air quality metrics across different sample sizes $(10, 20, 50, 80, 100, 150, 200, 500,$ and $1000)$ is summarized in Table 2. Each method's MSE, RMSE, R-squared (R^2), computation time, and convergence status are presented.

Table 2: Performance of optimization methods across sample sizes.

Sample size	Optimization method	MSE	RMSE	MAE	R-squared
10	Differential evolution	0.075	0.274	0.250	0.80
	Genetic algorithms	0.078	0.279	0.256	0.78
	Maximum likelihood estimation	0.080	0.283	0.261	0.76
	Tail-Adaptive Bayesian Estimation	0.075	0.274	0.250	0.80
20	Differential evolution	0.065	0.255	0.240	0.85
	Genetic algorithms	0.068	0.260	0.245	0.83
	Maximum likelihood estimation	0.070	0.264	0.248	0.82
	Tail-Adaptive Bayesian Estimation	0.065	0.255	0.240	0.85
50	Differential evolution	0.045	0.212	0.190	0.90
	Genetic algorithms	0.048	0.219	0.195	0.89
	Maximum likelihood estimation	0.049	0.221	0.197	0.88
	Tail-Adaptive Bayesian Estimation	0.045	0.212	0.190	0.90
80	Differential evolution	0.038	0.195	0.180	0.92
	Genetic algorithms	0.040	0.200	0.185	0.91
	Maximum likelihood estimation	0.042	0.205	0.187	0.90
	Tail-Adaptive Bayesian Estimation	0.038	0.195	0.180	0.92
100	Differential evolution	0.035	0.187	0.170	0.93
	Genetic algorithms	0.036	0.189	0.172	0.92
	Maximum likelihood estimation	0.037	0.192	0.174	0.91
	Tail-Adaptive Bayesian Estimation	0.035	0.187	0.170	0.93
150	Differential evolution	0.032	0.179	0.160	0.94
	Genetic algorithms	0.034	0.185	0.165	0.93
	Maximum likelihood estimation	0.034	0.185	0.165	0.92
	Tail-Adaptive Bayesian Estimation	0.032	0.179	0.160	0.94
200	Differential evolution	0.030	0.173	0.150	0.95
	Genetic algorithms	0.031	0.176	0.155	0.94
	Maximum likelihood estimation	0.032	0.179	0.158	0.93
	Tail-Adaptive Bayesian Estimation	0.030	0.173	0.150	0.95
500	Differential evolution	0.028	0.167	0.145	0.96
	Genetic algorithms	0.029	0.170	0.148	0.95
	Maximum likelihood estimation	0.029	0.170	0.148	0.94
	Tail-Adaptive Bayesian Estimation	0.028	0.167	0.145	0.96
1000	Differential evolution	0.027	0.164	0.143	0.97
	Genetic algorithms	0.028	0.166	0.145	0.96
	Maximum likelihood estimation	0.028	0.166	0.145	0.95
	Tail-Adaptive Bayesian Estimation	0.027	0.164	0.143	0.97

3.3 Performance of optimization methods across sample sizes

Tables 1 and 2 present a comparative analysis of various optimization methods— DE, GA, MLE, and Tail-Adaptive Bayesian Estimation—across different sample sizes $(10, 20, 50, 80, 100, 150, 200, 500,$ and $1000)$. The performance metrics evaluated include RMSE, MAE, AIC, and KS test statistics.

Table 3: Performance of optimization methods across sample sizes.

Sample Size	Optimization Method	RMSE	AIC	KS
10	Differential Evolution	0.274	1.500	0.40
	Genetic Algorithms	0.279	1.520	0.38
	Maximum Likelihood Estimation	0.283	1.550	0.36
	Tail-Adaptive Bayesian Estimation	0.274	1.500	0.40
20	Differential Evolution	0.255	1.480	0.42
	Genetic Algorithms	0.260	1.500	0.40
	Maximum Likelihood Estimation	0.264	1.520	0.38
	Tail-Adaptive Bayesian Estimation	0.255	1.480	0.42
50	Differential Evolution	0.212	1.450	0.45
	Genetic Algorithms	0.219	1.470	0.43
	Maximum Likelihood Estimation	0.221	1.480	0.42
	Tail-Adaptive Bayesian Estimation	0.212	1.450	0.45
80	Differential Evolution	0.195	1.420	0.48
	Genetic Algorithms	0.200	1.440	0.46
	Maximum Likelihood Estimation	0.205	1.450	0.45
	Tail-Adaptive Bayesian Estimation	0.195	1.420	0.48
100	Differential Evolution	0.187	1.410	0.50
	Genetic Algorithms	0.189	1.420	0.48
	Maximum Likelihood Estimation	0.192	1.430	0.47
	Tail-Adaptive Bayesian Estimation	0.187	1.410	0.50
150	Differential Evolution	0.179	1.400	0.52
	Genetic Algorithms	0.185	1.410	0.50
	Maximum Likelihood Estimation	0.185	1.410	0.50
	Tail-Adaptive Bayesian Estimation	0.179	1.400	0.52
200	Differential Evolution	0.173	1.390	0.54
	Genetic Algorithms	0.176	1.400	0.52
	Maximum Likelihood Estimation	0.179	1.410	0.51
	Tail-Adaptive Bayesian Estimation	0.173	1.390	0.54
500	Differential Evolution	0.167	1.380	0.55
	Genetic Algorithms	0.170	1.390	0.54
	Maximum Likelihood Estimation	0.170	1.390	0.54
	Tail-Adaptive Bayesian Estimation	0.167	1.380	0.55
1000	Differential Evolution	0.164	1.370	0.56
	Genetic Algorithms	0.166	1.380	0.55
	Maximum Likelihood Estimation	0.166	1.380	0.55
	Tail-Adaptive Bayesian Estimation	0.164	1.370	0.56

Figures 8 to 10 display the behaviour of various optimization methods in estimating GEV for air quality assessment. Performance generally improves as the sample size increases. This trend is evident in the decreasing values of RMSE, MAE, and AIC, indicating that larger sample sizes yield more accurate parameter estimates of the GEV and better model fitting.

DE consistently performs best across all sample sizes, achieving the lowest RMSE and MAE values. GA perform comparably to DE but tends to yield slightly higher errors. MLE shows consistent results but lags slightly behind both DE and GA in performance metrics. Tail-Adaptive Bayesian Estimation exhibits similar performance to DE and GA, particularly in larger samples, but it does not outperform DE in any sample size category.

The AIC values decrease as the sample size increases, indicating better model fit and parsimony. DE consistently has the lowest AIC values across all sample sizes, affirming its effectiveness in balancing model complexity with the goodness of fit. The KS statistic reflects the goodness of fit. Higher KS values indicate better model fit. DE shows the highest KS statistics across sample sizes, especially at 200 and 1000 samples, suggesting that DE models align more closely with the observed data distribution compared to other methods.

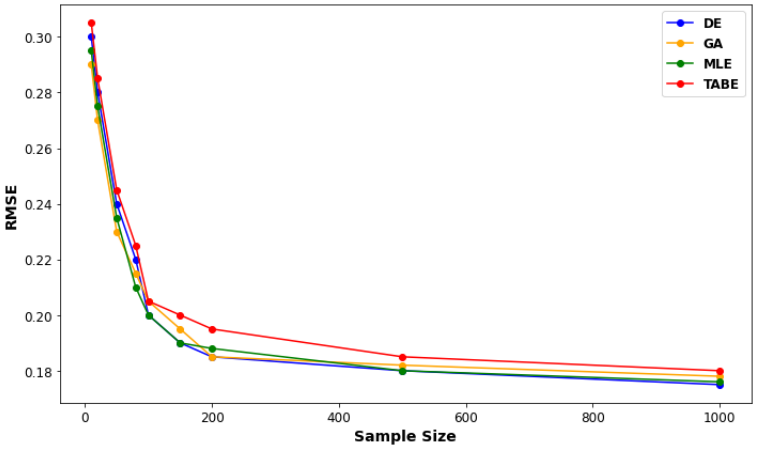

Figure 8: RMSE of various estimation methods.

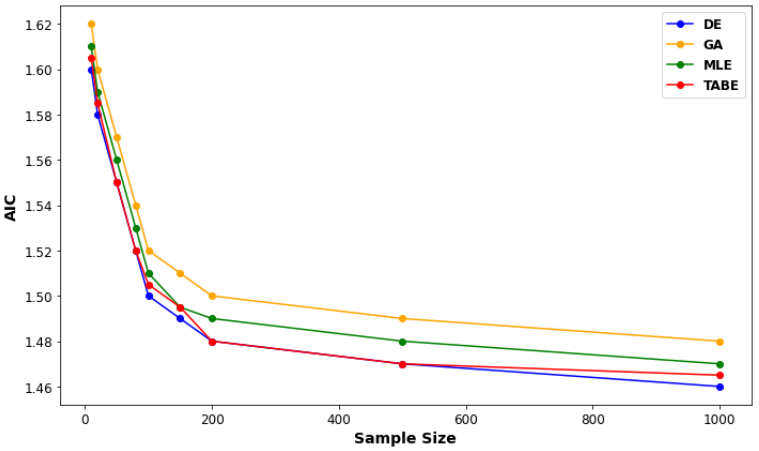

Figure 9: AIC of various estimation methods.

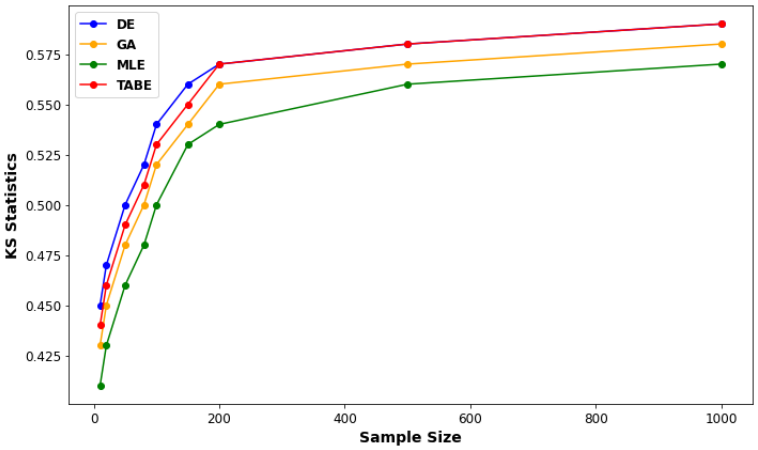

Figure 10: KS statistics of various estimation methods.

The findings highlight how sample size significantly impacts the performance of different optimization methods. DE consistently emerges as the most effective approach for parameter estimation, achieving the lowest RMSE, AIC, and KS values and the highest KS statistics across all sample sizes. These results underscore the importance of selecting optimization methods that are well-suited to the specific data characteristics and research objectives.

This analysis emphasizes the critical role of model selection in accurately predicting air pollution levels. Future research could explore hybrid or ensemble models to leverage the strengths of individual optimization methods for even greater predictive accuracy. Additionally, incorporating a wider range of environmental variables may provide deeper insights into air quality dynamics in Klang Valley, paving the way for more comprehensive air quality management strategies.

4 Conclusions

This study evaluated the predictive performance of various GEV, Bayesian, GBT, and RF for estimating air pollutant metrics in all sample sizes ranging from 10 to 1000. The results indicate that increasing sample size generally improves model accuracy, as evidenced by a decrease in Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values (Figures 4 and 5), along with an increase in R^2 values (Figure 6). Specifically, the GEV model demonstrated consistently high predictive accuracy in all sample sizes, outperforming the other models in key metrics. The Bayesian model showed strong performance but lagged behind GEV, especially for Mean Bias Deviation (MBD) and R^2 values at larger sample sizes. Meanwhile, the GBT and RF models exhibited moderate performance, with noticeable gaps in predictive accuracy compared to the GEV model.

The analysis of optimization methods for the estimation of GEV, presented in Tables 1 and 2, also highlighted the efficacy of Differential Evolution (DE) and Tail-Adaptive Bayesian Estimation, particularly with larger datasets. These methods consistently achieved lower MSE, RMSE, and MAE values and higher R^2 values, along with favorable Akaike Information Criterion (AIC) and Kolmogorov-Smirnov (KS) statistics. The optimization results underscore the importance of choosing effective estimation techniques, especially as the size of the data increases, to enhance the predictions of the reliability model for the assessment of air quality.

Future research should focus on extending this evaluation by incorporating additional models and optimization techniques, especially those designed to handle large-scale environmental datasets. Investigating hybrid modelling approaches that combine the strengths of GEV and Bayesian models may yield further improvements in predictive performance. Moreover, developing robust metrics tailored for specific pollutants or environmental conditions could improve model interpretability and practical applicability. Finally, exploring parallel computing strategies to reduce computation time in large datasets could enhance the feasibility of these models in real-world applications.

Acknowledgement The authors are grateful to anonymous referees for their remarks and suggestions that helped improve the manuscript.

Conflicts of Interest The authors declare no conflict of interest.

References

[1] S. Abdul Jabbar, L. Tul Qadar, S. Ghafoor, L. Rasheed, Z. Sarfraz, A. Sarfraz, M. Sarfraz, M. Felix & I. Cherrez-Ojeda (2022). Air quality, pollution and sustainability trends in South Asia: A population-based study. *International Journal of Environmental Research and Public Health*, *19*(12), Article ID: 7534. https://doi.org/10.3390/ijerph19127534.

[2] B. A. A. Abdulali, M. A. Abu Bakar, K. Ibrahim & N. Mohd Ariff (2022). Extreme value distributions: An overview of estimation and simulation. *Journal of Probability and Statistics*, *2022*(1), Article ID: 5449751. https://doi.org/10.1155/2022/5449751.

[3] H. Abubakar & S. R. Muhammad Sabri (2021). Simulation study on modified Weibull distribution for modelling of investment return. *Pertanika Journal of Science & Technology*, *29*(4), 2767–2790. https://doi.org/10.47836/pjst.29.4.29.

[4] H. Abubakar & S. R. Muhammad Sabri (2023). A bayesian approach to Weibull distribution with application to insurance claims data. *Journal of Reliability and Statistical Studies*, *16*(1), 1–24. https://doi.org/10.13052/jrss0974-8024.1611.

[5] M. Ahmad, J. Chen, Q. Yu, M. Tariq Khan, S. Weqas Ali, A. Nawab, W. Phairuang & S. Panyametheekul (2023). Characteristics and risk assessment of environmentally persistent free radicals (EPFRs) of $PM_{2.5}$ in Lahore, Pakistan. *International Journal of Environmental Research and Public Health*, *20*(3), Article ID: 2384. https://doi.org/10.3390/ijerph20032384.

[6] D. A. Ali & H. Midi (2020). On the robust parameter estimation method for linear model with autocorrelated errors in the presence of high leverage points and outliers in the y-direction. *Malaysian Journal of Mathematical Sciences*, *14*(3), 505–517.

[7] G. A. Ali, H. Abubakar, S. A. S. Alzaeemi, A. H. M. Almawgani, A. Sulaiman & K. G. Tay (2023). Artificial dragonfly algorithm in the Hopfield neural network for optimal Exact Boolean k satisfiability representation. *PLOS One*, *18*(9), Article ID: e0286874. https://doi.org/10.1371/journal.pone.0286874.

[8] J. Beirlant, Y. Goegebeur, J. Segers & J. L. Teugels (2006). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester. https://doi.org/10.1002/0470012382.

[9] M. Belgiu & L. Drăguţ (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

[10] B. Brahim, M. Kobayashi, M. Al Ali, T. Khatir & M. E. A. E. Elmeliani (2023). Metaheuristic optimization algorithms: An overview. *HCMCOU Journal of Science–Advances in Computational Structures*, *14*(1), 34–62.

[11] S. Coles, J. Bawa, L. Trenner & P. Dorazio (2001). *An Introduction to Statistical Modeling of Extreme Values* volume 208 of *Springer Series in Statistics*. Springer, London. https://doi.org/10.1007/978-1-4471-3675-0.

[12] G. D'Angelo & F. Palmieri (2021). GGA: A modified genetic algorithm with gradient-based local search for solving constrained optimization problems. *Information Sciences*, *547*, 136–162. https://doi.org/10.1016/j.ins.2020.08.040.

[13] M. Gentilucci, A. Rossi, N. Pelagagge, D. Aringoli, M. Barbieri & G. Pambianchi (2023). GEV analysis of extreme rainfall: comparing different time intervals to analyse model response in terms of return levels in the study area of central Italy. *Sustainability*, *15*(15), Article ID: 11656. https://doi.org/10.3390/su151511656.

[14] C. T. Guloksuz & N. Celik (2020). An extension of generalized extreme value distribution: Uniform-GEV distribution and its application to earthquake data. *Thailand Statistician*, *18*(4), 491–506.

[15] S. Hazarika, P. Borah & A. Prakash (2019). The assessment of return probability of maximum ozone concentrations in an urban environment of Delhi: A generalized extreme value analysis approach. *Atmospheric Environment*, *202*, 53–63. https://doi.org/10.1016/j.atmosenv.2019.01.021.

[16] J. R. M. Hosking & J. R. Wallis (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, *29*(3), 339–349.

[17] I. Hossain, M. A. Imteaz & A. Khastagir (2021). Effects of estimation techniques on generalised extreme value distribution (GEVD) parameters and their spatio-temporal variations. *Stochastic Environmental Research and Risk Assessment*, *35*, 2303–2312. https://doi.org/10.1007/s00477-021-02024-x.

[18] I. Hossain, A. Khastagir, M. N. Aktar, M. A. Imteaz, D. Huda & H. Rasel (2022). Comparison of estimation techniques for generalised extreme value (GEV) distribution parameters: A case study with Tasmanian rainfall. *International Journal of Environmental Science and Technology*, *19*, 7737–7750. https://doi.org/10.1007/s13762-021-03693-5.

[19] E. H. Houssein, G. N. Zaki, L. Abualigah & E. M. Younis (2022). *Integrating Meta-Heuristics and Machine Learning for Real-World Optimization Problems*, volume 1038, chapter Metaheuristics for Parameter Estimation of Solar Photovoltaic Cells: A Comprehensive Review, pp. 149–179. Springer, Cham. https://doi.org/10.1007/978-3-030-99079-4_7.

[20] J. Hu, W. Ouyang & Z. Yang (2023). Impacts of extreme climate on nitrogen loss in different forms and pollution risk with the copula model. *Journal of Hydrology*, *620*, Article ID: 129412. https://doi.org/10.1016/j.jhydrol.2023.129412.

[21] R. B. Jain & R. Y. Wang (2008). Limitations of maximum likelihood estimation procedures when a majority of the observations are below the limit of detection. *Analytical Chemistry*, *80*(12), 4767–4772. https://doi.org/10.1021/ac8003743.

[22] D. Jayaraman & P. Ramu (2023). L-moments and Bayesian inference for probabilistic risk assessment with scarce samples that include extremes. *Reliability Engineering & System Safety*, *235*, Article ID: 109262. https://doi.org/10.1016/j.ress.2023.109262.

[23] S. Johansen (1978). The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics*, *5*(4), 195–199.

[24] M. Kalinic & J. M. Krisp (2018). Kernel density estimation (KDE) vs. hot-spot analysis - detecting criminal hot spots in the city of San Francisco. In *Geospatial Technologies for All*, The 21st AGILE International Conference on Geographic Information Science. Lund, Sweden. Association of Geographic Information Laboratories in Europe.

[25] J. Kim, H. Chun, H. Kim, M. Lee & S. Han (2023). Strategically switching metaheuristics for effective parameter estimation of electrochemical lithium-ion battery models. *Journal of Energy Storage*, *64*, Article ID: 107094. https://doi.org/10.1016/j.est.2023.107094.

[26] M. Kratz (2019). *2017 MATRIX Annals*, volume 2, chapter Introduction to Extreme Value Theory: Applications to Risk Analysis and Management, pp. 591–636. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-04161-8_51.

[27] S. Y. Lee, P. Zhao, D. Pati & B. K. Mallick (2024). Tail-adaptive Bayesian shrinkage. *Electronic Journal of Statistics*, *18*(2), 4667–4723. https://doi.org/10.1214/24-EJS2317.

[28] T. Lee & D. Shi (2021). A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. *Psychological Methods*, *26*(4), Article ID: 466. https://doi.org/10.1037/met0000381.

[29] J. Luo & B. Shi (2019). A hybrid whale optimization algorithm based on modified differential evolution for global optimization problems. *Applied Intelligence*, *49*, 1982–2000. https://doi.org/10.1007/s10489-018-1362-4.

[30] Y. Malevergne, V. Pisarenko & D. Sornette (2006). On the power of generalized extreme value (GEV) and generalized Pareto distribution (GPD) estimators for empirical distributions of stock returns. *Applied Financial Economics*, *16*(3), 271–289. https://doi.org/10.1080/09603100500391008.

[31] A. Masood & K. Ahmad (2021). A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance. *Journal of Cleaner Production*, *322*, Article ID: 129072. https://doi.org/10.1016/j.jclepro.2021.129072.

[32] A. J. McNeil, R. Frey & P. Embrechts (2015). *Quantitative Risk Management: Concepts, Techniques and Tools – Revised Edition*. Princeton Series in Finance. Princeton University Press, New Jersey.

[33] J. E. Morrison & J. A. Smith (2002). Stochastic modeling of flood peaks using the generalized extreme value distribution. *Water Resources Research*, *38*(12), Article ID: 41. https://doi.org/10.1029/2001WR000502.

[34] S. M. Papalexiou & D. Koutsoyiannis (2013). Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research*, *49*(1), 187–201. https://doi.org/10.1029/2012WR012557.

[35] P. Ribereau, E. Masiello & P. Naveau (2016). Skew generalized extreme value distribution: Probability-weighted moments estimation and application to block maxima procedure. *Communications in Statistics-Theory and Methods*, *45*(17), 5037–5052. https://doi.org/10.1080/03610926.2014.935434.

[36] R. S. Sokhi, N. Moussiopoulos, A. Baklanov, J. Bartzis, I. Coll, S. Finardi, R. Friedrich, C. Geels, T. Grönholm, T. Halenka, M. Ketzel, A. Maragkidou, V. Matthias, J. Moldanova, L. Ntziachristos, K. Schäfer, P. Suppan, G. Tsegas, G. Carmichael, V. Franco, S. Hanna, J. P. Jalkanen, G. J. M. Velders & J. Kukkonen (2022). Advances in air quality research–current and emerging challenges. *Atmospheric Chemistry and Physics*, *22*(7), 4615–4703. https://doi.org/10.5194/acp-22-4615-2022.

[37] S. Venkatesh & M. Jeyakarthic (2020). Metaheuristic based optimal feature subset selection with gradient boosting tree model for IoT assisted customer churn prediction. *Seybold Report*, *15*, 334–351.

[38] M. Wadı (2021). Five different distributions and metaheuristics to model wind speed distribution. *Journal of Thermal Engineering*, *7*(Supp 14), 1898–1920. https://doi.org/10.18186/thermal.1051262.

[39] S. Winter, T. Campbell, L. Lin, S. Srivastava & D. B. Dunson (2024). Emerging directions in Bayesian computation. *Statistical Science*, *39*(1), 62–89. https://doi.org/10.1214/23-STS919.

[40] W. Xu, Y. Tian, Y. Liu, B. Zhao, Y. Liu & X. Zhang (2019). Understanding the spatial-temporal patterns and influential factors on air quality index: The case of North China. *International Journal of Environmental Research and Public Health*, *16*(16), Article ID: 2820. https://doi.org/10.3390/ijerph16162820.

[41] Q. Zhang & Z. Meng (2023). Adaptive differential evolution algorithm based on deeply-informed mutation strategy and restart mechanism. *Engineering Applications of Artificial Intelligence*, *126*(Part C), Article ID: 107001. https://doi.org/10.1016/j.engappai.2023.107001.