Indian Journal of Multidisciplinary Research and Studies
ISSN: 3108-1681 | Volume- 2, Issue- 4 | October - December 2026

Optimizing Rainfall Forecasting in Iraq: A Comparative Study of XGBoost, Nature-inspired Algorithms, and Tweedie Models

Taha Darwassh Hanawy Hussein[1], Wakkaa Ali Hadba[2], Ozlam Abdulhakeem Mahmood [3]
[1] College of Pharmacy and Clinical Laboratory Sciences, University of Kirkuk, Kirkuk, Iraq
[2] College of Management and Economics, University of Kirkuk, Kirkuk, Iraq
[3] College of Computer Science and Software Department, Kirkuk University, Kirkuk/Iraq

Abstract

In fields such as agriculture and disaster management, it is essential to obtain accurate rainfall predictions. However, due to discontinuous precipitation and a large number of zero-valued data, modeling becomes difficult. In these complex situations, traditional statistical methods often cannot achieve highly accurate predictions. Using meteorological data from the Abu Ghraib station in Iraq (2011–2025), this article examines the performance of machine learning (ML) models against statistical frameworks. Therefore, we utilized an XGBoost model, nature-inspired optimization techniques (Whale Particle Optimization (WPO) and Hybrid Bio-inspired Optimization Algorithm (HBOA)) and the Tweedie model. The nature-inspired optimization techniques were utilized to refine the parameter tuning of the XGBoost algorithm separately. Specifically, it provides a comparative analysis of four models: XGBoost, WPO-enhanced XGBoost, HBOA-enhanced XGBoost, and Tweedie. The findings showed that the XGBoost model is clearly better than other models. Briefly, the XGBoost model achieved the highest coefficient of determination (R2=0.1483) and the lowest Root Mean Square Error (RMSE=3.3367). Although the WPO-enhanced XGBoost reduced the Mean Absolute Error (MAE) to 0.8026, it failed to outperform basic XGBoost. Additionally, the Tweedie model performed worst, with a negative R2 (-0.1116). Accordingly, the results demonstrated the superiority of ML models over statistical frameworks in dealing with complex situations (discontinuous features and a high frequency of zero-valued data). Nevertheless, further research is needed to enhance XGBoost with advanced methods, such as Bayesian optimization.

Keywords: Regression Algorithms, Tweedie Model, Whale Optimization Algorithm, XGBoost, Hybrid Bio-inspired optimization Algorithms

1. Introduction

Predicting future weather patterns is very important in critical sectors such as disaster management, agriculture, and transportation. Previously, weather forecasting was based on numerical models that analyzed satellite data such as temperature, humidity, and wind. Although the models were efficient, they used high computing resources and had several limitations Attri et al. (2025). Recently, machine learning (ML) models have become key tools

Published: 12 November 2026
DOI: https://doi.org/10.70558/IJMRS.2026.v2.i4.301260
Copyright © 2026 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

for data analysis in different disciplines. Aiming to identify and detect complex, unconventional patterns and achieve optimum accuracy and the highest prediction speed.

The machine learning (ML) algorithms, like XGBoost, are popular in this regard because of their effectiveness and quick training process using advanced parameter optimization methods. However, improving the effectiveness of such models is a scientific issue that has led to algorithms such as WPO and HBOA, which optimize model parameters. Regardless of the advantages of ML, conventional statistical models like Tweedie are used in certain areas, and their effectiveness is compared to the effectiveness of new models (Kumar et al., 2024). In rain-fed agriculture, where the amount of precipitation is vital for successful work, precipitation forecasting becomes crucial. Precipitation modeling faces problems because of discontinuity and scarcity of data and the large share of zeroes, preventing conventional models from high-quality forecasting (Nelson et al., 2018).

Among the popular models used for data analysis in such cases is the Tweedy model, known for its ability to handle dispersed and complex distributions, such as those in precipitation data. The family of Tweedy distributions is more flexible in the handling of data that has zeros and continuous values. The Poisson-Gamma distribution is one of the Tweedy family distributions, which is well-suited for handling precipitation data efficiently. Many studies have shown the importance of this technique in improving our understanding of how climate affects key areas, as well as in providing better solutions to climate and environmental issues (Suhaila, 2023; Dunn, 2004; Hasan & Dunn, 2010; Dalal et al., 2025). This paper will examine modern techniques for the analysis of complex data and how to deal with the challenges involved in modeling.

The structure of the paper is as follows. In Section 2, the current literature is reviewed in great detail, with particular focus on the theory behind the Tweedie distribution, the XGBoost method and nature-inspired optimization algorithms. The research methodology, which includes data acquisition, data pre-processing methods, as well as the performance metrics used for evaluation, is presented in Section 3. Experimental results and preliminary analysis of the proposed predictive approaches are provided in Section 4. The obtained results are analyzed in depth in Section 5, focusing on the trade-offs in terms of performance between baseline machine learning models and the optimized ones. Finally, Section 6 concludes this research and outlines future recommendations.

2. Background and Literature Review

In spite of numerous studies performed on Tweedie models and the XGBoost algorithm, there is still a lack of research combining them to form new approaches. Besides, there is also a lack of research on applying biologically inspired optimizers, like the WPO algorithm, to optimize the XGBoost algorithm. In addition, the present comparative analyses do not offer an adequate study of the robustness of the XGBoost algorithm when dealing with data with zero inflation or extreme events. The Tweedie models belonging to the exponential dispersion models are widely used while working with positive data with a zero component – a phenomenon that occurs frequently in precipitation data having days with no rainfall. These models have great advantages in handling extreme dispersion or data with heavy tails, making them ideal for

modeling rainfall data. Nevertheless, if these models perform poorly, it may be necessary to resort to alternative algorithms, such as XGBoost, which is known to perform efficiently with this kind of data. This research shows a scarcity of studies and applications of biologically inspired optimizers, such as WPO, to improve XGBoost performance. This area is of great importance for studying applications that predict and simulate rainfall data with higher quality (Dalal et al., 2025; Museba, 2022; Xie et al., 2025).

Some studies have found that Tweedie family distributions, such as the gamma and Poisson-gamma distributions, are suitable for modeling seasonal rainfall at Australian stations. The Tweedie family distributions are especially useful in making a simulation of zero and heavy rainfall events (Hasan & Dunn, 2015). The models developed using the Tweedie model in Australia have proven successful in explaining the influence of climate on monthly rainfall rates by incorporating cyclical and climatic components into the GLM framework (Hasan & Dunn, 2012). Some studies conducted in Malaysia have found that Tweedy models, along with seasonal and climatic components, can explain monthly rainfall effectively (Suhaila, 2023). Studies indicated that the Tweedy composite model with Lasso reduction is an effective statistical framework for predicting rainfall intensity. It also helps address the challenges of multicollinearity associated with large-scale climate data (Hayati & Permatasari, 2023). Studies in Sweden confirmed that the Tweedy distribution is a suitable model for analyzing monthly precipitation, with behavior similar to that found in Australian environments, particularly the effect of coastal location on the model's parameters (Rydén, 2025).

The Whale Optimization Algorithm (WOA), which is based on the hunting behavior of humpback whales, is an efficient algorithm that combines exploration and exploitation in scientific search spaces. It has been found to be successful in optimizing statistical models and machine learning algorithms with particular emphasis on climate modeling and rainfall prediction. Unlike WOA and PSO, the WPO has an extra feature of overcoming local optima. As a result of its bubble search technique, WOA has been tested on 29 mathematical optimization problems and 6 structural design problems, demonstrating better performance than any modern algorithm (Cai et al., 2024; Dalal et al., 2025; Mirjalili & Lewis, 2016; Pham et al., 2020; Rydén, 2025; Yang, 2014). XGBoost can be considered one of the advanced machine learning algorithms thanks to its high efficiency and ability to deal with complex data. This algorithm works based on a gradient-based loss function for tree models. This helps minimize errors and increase the overall stability of the model in environmental science. There have been studies on estimating temperature patterns in forests using meteorological data from open fields (Chen & Guestrin, 2016). Furthermore, XGBoost has predicted solar radiation and the levels of groundwater in the state of Selangor in Malaysia, highlighting the ability of the method to work with time series and climate data that have natural noise (Chen & Guestrin, 2016; Ghafarian et al., 2022; Trojovský & Dehghani, 2023). XGBoost is one of the methods used for predicting climate and environmental factors(So & Valdez, 2025). Besides XGBoost, there are other algorithms similar to tree boosting, namely CatBoost, which proved to be efficient for the analysis of insurance data through a Tweedie Boosting model designed to handle imbalanced or zero-inflated data. The above-mentioned models are used to improve the results produced by XGBoost and produce better results in the case of data with abnormal distributions (Zhou et al., 2022). To deal with the problem of unevenly distributed classes of

railway accidents and the insufficient cause analysis of the mentioned issue, a Whale Optimization Algorithm-based approach was suggested for improving the performance of the XGBoost algorithm. As a result, a new WOA-XGBoost model was developed.

Furthermore, a number of researchers (Dalal et al., 2025; Museba, 2022; Xie et al., 2025) investigated a machine learning algorithm designed to improve the analysis of people's credit risk compared with conventional approaches. An XGBoost algorithm combined with adaptive particle swarm optimization proved more flexible under different environmental conditions. The experiments have revealed that the suggested approach, based on decision trees and an XGBoost algorithm, is superior due to its new dynamic properties and is substantiated by the particle swarm algorithm. In the industrial domain, the WOA was used to enhance the effectiveness of the XGBoost algorithm for anomaly detection.

2.1. Tweedie Regression

The Tweedie regression model is an example of a type of Generalized Linear Model (GLM) whose response variable follows a Tweedie distribution. This regression model is ideal for modeling data that has a combination of zero and positive continuous values, and it is used to predict rainfall data, which could have zeroes or overdispersion (Cordeiro et al., 2021), (Dunn & Smyth, 2008), and (Dunn & Smyth, 2001). This regression model is characterized by a family of distributions specified by the following functional forms (Kurz, 2017; Hadba, 2025). As equation (1).

$$f(y|\theta,\varphi) = a(y,\varphi)\exp\left[\frac{y\theta - \kappa(\theta)}{\varphi}\right] \quad (1)$$

where θ is the natural parameter and φ > 0, the parameter φ is called the dispersion, and μ means the variation of random distributions is a special case of EDMS where Var (y) = κ (θ) φ, respectively. The Tweedie family is specified by the EDM variable as E(Y) = μ = κ(θ). Several familiar distributions characterize the medium relationship in ventilation. For example, natural (P = 0) ,Var (μ) = φμP for P ∈ (0, 1). The Tweedie family (p = 3) includes. Poisson (P = 1), Gamma (P = 2) and the reverse Gaussian. The P ∈ (1, 2) choice is the most supportive of the semi-continuous results. Tweedie distributions in this range, which belong to the so-called Poisson-Gamma distributions (Ghafarian et al., 2022), (Obiora et al., 2021), and (Osman et al., 2021), are characterized by the parameter P. Table 1 summaries Tweedie EDMs (Jørgensen, 1997). In the exponential dispersion model, the mean and variance of a variable X are given by equations (2) and (3), respectively, where $K(\theta)$ represents the cumulant function.The dependent variable is modeled via its conditional mean, as shown in equation (4), and its variance is related to the mean through the exponential variance function presented in equation (5), where phi is the dispersion parameter and p is the power parameter. A logarithmic link function, as described in equation (6), is employed, making the mean an exponential function of the independent variables, as shown in equation (7). The parameters beta, phi, and p are estimated by maximizing the log-likelihood function given in equation (8).

Equations (2) and (3) are expressed as follows:

$$E(X) = \lambda K'(\theta) \tag{2}$$

$$Var(X) = \lambda K''(\theta) \tag{3}$$

Conditional mean: $$E(Y_i \mid X_i) = \mu_i \tag{4}$$

Variance function: $$Var(Y_i) = \varphi \mu_i^p \tag{5}$$

Link function (log link): $$g(\mu_i) = \log(\mu_i) = X_i^T \beta \tag{6}$$

$$\mu_i = \exp(X_i^T \beta) \tag{7}$$

Log-likelihood function: $$\ell(\beta, \varphi, p) = \Sigma \log f(y_i \mid \mu_i, \varphi, p) \tag{8}$$

Table 1: A summary of Tweedie EDMs (Jørgensen, 1997)

Distribution	p	α	M
Extreme stable	$p < 0$	$1 < \alpha < 2$	$(0, \infty)$
Normal	$p = 0$	$\alpha = 2$	$\mathbb{R}$
[Do not exist]	$0 < p < 1$	$2 < \alpha < \infty$	
Poisson	$p = 1$	$\alpha \to -\infty$	$(0, \infty)$
Compound Poisson	$1 < p < 2$	$\alpha < 0$	$(0, \infty)$
Gamma	$p = 2$	$\alpha = 0$	$(0, \infty)$
Positive stable	$p > 2$	$0 < \alpha < 1$	$(0, \infty)$
Extreme stable	$p \to \infty$	$\alpha = 1$	$\mathbb{R}$

2.2. Whale Optimization Algorithm (WOA)

The Whale Optimization Algorithm is based on the social behaviors of humpback whales. This algorithm has been developed based on the bubble-net fishing approach (Mirjalili & Lewis, 2016). This type of whale is considered a part of the baleen whale family. The baleen whale family consists of seven species of whales and is distinguished by the special sleeping behavior where one of its two brain hemispheres is active, while the other is resting. Humpback whales live in herds and eat small fish and krill. The unique hunting behavior of the humpback whales is represented in making circular bubble patterns around fish to surround them and make them

easier to catch. Whale Particle Optimization (WPO) is a combined algorithm that combines the classic WOA with Particle Swarm Optimization (PSO) in order to escape from the local optima and keep the particles' diversity. The concepts of WPO are the following (Madlool, 2024):

First: Prey Encircling

Whales are identified as the largest mammals on our planet. The size of an adult whale ranges from 30 meters and can be as heavy as 180 tons. Whales prey on fish schools (Attri et al., 2025). Merghelli and Lewis (Khodadadi et al., 2024) proposed a novel nature-inspired optimization algorithm called the Whale Optimization Algorithm. The WOA is modeled on the Bubble-net hunting strategy used by humpback whales to prey on fish schools or krill found at or near the surface of the water. In order to adopt this technique, humpback whales swim in a circle around their prey in a decreasing circle formation while at the same time rotating to make bubbles in circles or in the shape of the number "9".

The WOA uses a 50% probability to choose between the shrinking encircling strategy and the spiral updating strategy, with the aim of improving the whale's position during the optimization process (Kaveh & Ghazaan, 2017). The mathematical representation of these strategies is shown in the next section. It is assumed that the optimal position at any time is the same position as the prey or the nearest whale at that time. Therefore, the whales position themselves towards this optimal position (Dalal et al., 2025; Hayati & Permatasari, 2023). The search and targeting process can be represented by a mathematical model as follows:

$$\vec{D} = \left|\vec{C} \cdot \overrightarrow{X_P}(t) - \vec{X}(t)\right| \tag{9}$$

$$\vec{X}(t+1) = \overrightarrow{X_P}(t) - \vec{A} \cdot \vec{D} \tag{10}$$

where t represents the current iteration, $\vec{A}$, $\vec{D}$ are coefficient $\vec{X}_p$ is the position vector of the prey, and $\vec{X}$ denotes the position vector of a whale. The vectors $\vec{A} \rightarrow$ and $\vec{A} \rightarrow$ are calculated as follows:

$$\vec{A} = \overrightarrow{2a} \cdot \overrightarrow{r_1} - \vec{a}$$
$$\vec{C} = 2 \cdot \overrightarrow{r_2} \tag{11}$$

where components of $\vec{a}$ are linearly decreased from 2 to 0 over the course of iterations, and r1,r2 are random vectors in [0,1] (Madlool, 2024).

Second: Bubble Net Attack Mechanism:

This strategy is simulated using two main methods (Madlool, 2024).

a) **Reducing the Encircling Radius**: The search agents gradually move towards the prey by shrinking the search area.

b) **Spiral Position Update**: The spiral path taken by the whale towards the prey is simulated using a logarithmic spiral mathematical function.

2.2.1 Spiral Bubble

WOA has been developed based on the behavior of humpback whales when they hunt their prey through bubble-netting. This method consists of two major stages. First, there is an The exploration phase involves searching for a variety of solutions, followed by an exploitation phase aimed at refining the best solutions discovered. At its core, the Whale Optimization Algorithm (WOA) is based on the concept of a spiral path, which describes how a whale approaches its prey (see Figure 1). This algorithm is characterized by its simplicity and reliance on a very limited number of parameters, making it extremely easy to implement. (Mahmood et al., 2023).

Network Attack Patterns: The bubble-net strategy, used by whales to trap their prey, is mathematically formulated within the Whale Optimization Algorithm (WOA) as follows:

$$X(\vec{t}+1) = \overrightarrow{D'} \cdot \exp^{ul} \cdot \cos(2\pi l) + X^*(t)$$
$$\overrightarrow{D'} = \left|\overrightarrow{X^*}(t) - X\overrightarrow{(t)}\right| \quad (12)$$

Employing u as a control parameter that dictates the shape of the logarithmic spiral, along with l, which is chosen randomly from [-1, 1], the model calculates the distance of the i-th whale from the target (best solution discovered so far). The humpback whales use two distinct approaches for hunting: one involves circling around prey with circles having a progressively decreasing radius, while the second approach uses a spiral path. In the algorithm for the optimization, the position of the whale is moved through any of the above two routes with equal probability. An attempt is made to mimic the hunting behavior of these whales. The mathematical model of the above-mentioned strategy is presented briefly below (Madlool, 2024) and (Mahmood et al., 2023):

$$X(t+1) = \begin{cases} \vec{X} * (t) - \vec{A}\vec{D}p < 0.5 \\ \overrightarrow{D^1} e^{bt} \cos(2\pi t) + \overrightarrow{X^*}(t) p < 0.5 \end{cases} \quad (13)$$

where p is a number between 0 and 1 that is selected at random.

2.2.2 Search for Prey

This exploration scheme involves a random search for prey, which is achieved through varying the vector $\vec{A}$ (Madlool, 2024). For the foraging of the sperm whales, the search for prey is stochastic with respect to their present locations. During this exploration scheme, the value of vectors $\vec{A}$ is varied in a random fashion within the range -1 to 1. This is done to encourage search agents to move away from the whale's reference location, thereby increasing the diversification of the search process. In this stage, a search agent is selected at random, and its location is altered using the scheme above. This method will be useful for improving the algorithm's ability to conduct an overall exploration of the search space, even when $|\vec{A}|$ is greater than 1. This scheme can be mathematically modeled using the following equations (Madlool, 2024):

$$\vec{D} = \left|\vec{C}\vec{X}_{rand} - \vec{X}\right| \quad (14)$$

$$\vec{X}(t+1) = \vec{X}_{rand} - \vec{AD} \quad (15)$$

Let $\vec{X}_{rand}$ represents the location of the random vector, indicating the randomly selected whale. There are many similarities between the WOA and other evolutionary computation algorithms, such as the Genetic Algorithm. Both these algorithms attempt to discover the optimal solution using the foraging pattern of animals. Algorithms in this category begin with a randomly selected initial population of solutions and refine the solutions generation after generation (Madlool, 2024). The WOA follows a similar approach, wherein the best current solutions are used to direct the search towards the optimal solution; thus, demonstrating a natural phenomenon seen in other species. Compared with the Genetic Algorithm, the Whale Optimization Algorithm is easier to use. Nevertheless, there are certain differences between the two algorithms, including the fact that the former does not have the features of evolution, namely, crossover and mutation. Currently, researchers try to improve the Whale Optimization Algorithm and make it incorporate a crossover procedure (Madlool, 2024). The algorithm called the x1 Method is as follows:

$$Y = w1x1 + w2x2 + \cdots + wixi \quad (16)$$

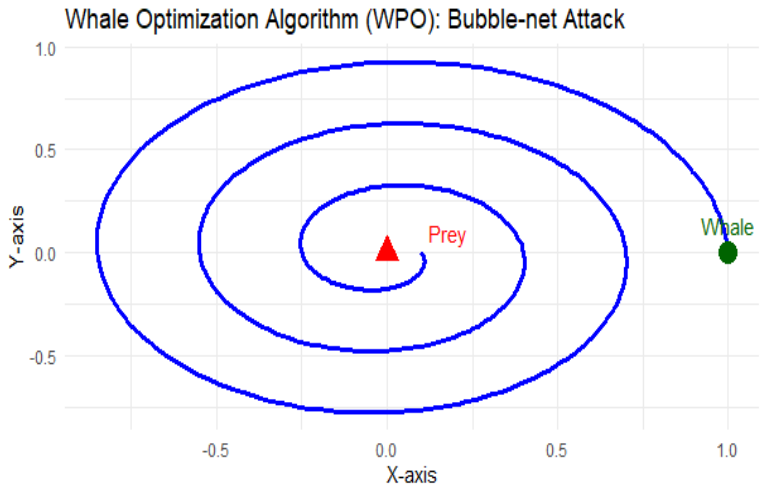


Figure 1**:** The search and exploitation process within the WPO using a spiral bubble model

2.3 Hybrid Bio-inspired Optimization Algorithms (HBOA)

Hybrid nature-inspired optimization algorithms (HBOA) model natural processes and combine various types of optimization techniques in order to strike an appropriate balance between exploration and exploitation. Hybrid optimization algorithms aim to address the weaknesses of individual algorithms by incorporating characteristics associated with biological evolution and the collective behaviors of insects, such as PSO-GA, PSO-GWO, and GWO. Combining these two characteristics improves accuracy, accelerates convergence, and increases robustness in solving difficult problems. It is evident from the scientific literature that various hybrid techniques have been developed, such as PSO-GA, PSO-GWO, HHO-WOA, and others (Rahman & Mohamad-Saleh, 2018) :

1. **Nature integrated with Nature**: PSO and GA or WOA and GWO are examples of that.

2. **Nature integrated with Traditional Mathematical Methods:** For example, combining PSO and Gradient Descent.

3. **Nature integrated with Machine Learning:** Using evolutionary algorithms to determine appropriate weights for neural networks or select appropriate features for them.

In the current study, the simplified HBOA technique is applied to optimize the parameter p in the Tweedie regression model. The process involves the application of three approaches, which include selection, crossover, and mutation techniques. The process of using these approaches is done in order to minimize the root mean squared error. The use of the Hybrid Binary Obstacle Avoidance Algorithm (HBOA) in Tweedie regression models illustrates how hybrid algorithms enhance the predictive accuracy of statistical models (Cai et al., 2024; Hadba, 2025; Kurz, 2017). It depends on the following factors (Cai et al., 2025). First, to refine the solutions, a genetic algorithm-inspired hybridization technique is used. Second, to enhance data diversity, a mutation mechanism that produces random Gaussian noise is used. Third, to ensure optimal results, performance-based selection and replacement strategies are implemented.

Despite being simple, this approach is very good in dynamically balancing exploration and exploitation, which makes it a good approach for the effective optimization of parameters in regression models (Cai et al., 2025) and (Nguyen & Le, 2021). HBOA is a hybrid optimization approach modeled on nature and biology. It uses various approaches to achieve a balance between exploring the solution space and exploiting the best solution, with the aim of obtaining the optimal solution to a problem. The main use of HBOA is parameter optimization.

Fitness Function (Zhang & Li, 2020) and (Hanif, 2020):

$$F(x) = \Sigma\,(i = 1\ to\ n)\ wi * fi(x) \tag{17}$$

where F(x) is the fitness value of solution x, fi(x) is the evaluation of criterion i, and wi denotes the weight assigned to each criterion.

$$\mathrm{X\{t+1\}} = \mathrm{Xt} + \alpha * (\mathrm{X_best} - \mathrm{Xt}) + \beta * \mathrm{Rt} \tag{18}$$

where X_t is the current position, X_{best} refers to the best solution found so far, α, β are learning and exploration parameters, and R_t is a random variable used to promote exploration.

Hybrid Strategy:

$$\mathrm{X_\{t+1\}} = \gamma * \mathrm{X_local} + (1-\gamma) * \mathrm{X_global} \tag{19}$$

where X_{local} is the local solution,X_{global} is the global solution, and γ is a parameter that balances exploitation and exploration.

2.4 Extreme Gradient Boosting algorithm (XGBoost)

Gradient boosting has become one of the most significant approaches in machine learning, owing to its ability to aggregate various simple models like decision trees within a single system in order to increase prediction accuracy and decrease both bias and variance at the same time. One of the developments in gradient boosting includes that made by Chen & Guestrin

(2016) in (Chen & Guestrin, 2016), which involves developing a new distributed weighted quantile sketch approach, which can manage weighted data with a theoretical guarantee, leading to a scalable and efficient approach termed XGBoost in (Chen & Guestrin, 2016).

The XGBoost works based on the gradient boosting technique, where gradient descent is combined with boosting techniques through a particular algorithm referred to as Gradient Boosting Machines (GBM). In GBM, several models are combined, and each model is assigned weights depending on the errors and the performance (Suhaila, 2023). Misclassification points have more weight than those that have been correctly classified. Through the use of gradient descent, GBM uses second-order statistics in order to improve the model. Using this particular procedure, XGBoost first develops a decision tree in order to get the data pattern and correct any errors. The decision trees are added incrementally through boosting techniques in order to increase the accuracy (Khodadadi et al., 2024). Finally, the whole of the decision trees is combined into one model through specific weights. Due to its advanced technical features and high performance, this algorithm is efficient and can be used to solve complex problems. With its advanced technical features and high performance, this algorithm is an ideal choice for tackling complex classification and prediction tasks with high reliability and efficiency (Kaveh & Ghazaan, 2017; Mahmood et al., 2023).

The function of the prediction stage can be computed as follows (Singh et al., 2022).

$$P_i^{(s)} = \sum_{k=1}^{s} P_k(y_i) = P_i^{(s-1)} + P_s(y_i) \tag{20}$$

where $P_s(y_i)$ refers to the step t in the learner. The original function can be modified to prevent overfitting.

$$O^{(s)} = \sum_{k=1}^{n} l(y - y_i) + \sum_{k=1}^{s} \theta(f_i)$$
$$\theta(f_i) = \varphi T + 0.5\lambda \|w\|^2 \tag{21}$$

The boosting technique uses a series of small decision trees to gradually improve the prediction ability of the model, hence improving its accuracy, especially in areas where it is poor. The shrinkage parameter λ is important for controlling how quickly improvement occurs because it reduces the weight of each new tree, thereby allowing a larger number of trees to be used for correction. Boosting differs from bagging in that it builds models progressively in iterations rather than simply averaging outcomes in every iteration as in bagging (Gao, 2024).

Boosting, unlike bagging and random forest, will overfit slowly as the number of trees (B) grows too large. However, such overfitting will normally occur quite slowly, if at all. In order to find the appropriate value of B, cross-validation is used. Furthermore, λ (the shrinkage parameter) is always a positive number used to control the learning process in boosting, thereby preventing sudden jumps. The parameter d (the number of splits in each tree) controls the complexity of the boosting model. The good choice of d is d = 1, meaning that the tree is a stump (Gao, 2024).

Boosting algorithm in regression:

1. Set

$f\hat{}(x) = 0$ and ri = yi for all i in the training set.

2. For $b = 1, 2, \ldots, B$, repeat: (a) Fit a tree

$f\hat{}b$ with d splits (d + 1 terminal nodes) to the training data (X, r).

(b) Update

f^ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x) \tag{22}$$

3. Research methodology

3.1. Data Collection and Preparation:

The data used to derive the mean precipitation values were obtained from the Ministry of Agriculture, Agricultural Meteorological Center, Baghdad Governorate - Abu Ghraib Station, situated at a longitude of 44.23° E and a latitude of 33.32° N. The data set contains the values of rainfall on a daily basis from April 15, 2011, to November 30, 2025, and represents the dependent variable in this case. Additionally, some influential independent variables were selected, such as Maximum Temperature, Minimum Temperature, Average Temperature, Maximum Relative Humidity, Minimum Relative Humidity, Average Relative Humidity, Total Solar Radiation, Average Wind Speed, Maximum Wind Speed, Wind Direction, and Evaporation Rate. These values were standardized to avoid any bias due to differences in units of measurement. The complete dataset was then divided into two separate sets: 80% for training the predictive models and 20% for testing the performance and accuracy of these models.

3.2. Evaluation Metrics

The prediction accuracy of the models was measured using a set of metrics. These metrics provide a comprehensive view of each model's explanatory power and prediction error (Mauricio Palacio, 2020).

3.2.1 Coefficient of Determination (R^2)

R^2 is used to measure how well a model explains the variability in the data (Mauricio Palacio, 2020). If R^2 equals 1, this means the model accounts for all variability, providing a perfect prediction. When $R^2 = 0$, the model provides no improvement over the explanatory power of the mean. Additionally, when $R^2 < 0$, the model performs worse than relying solely on the mean. Its mathematical formula is:

$$R^2 = 1 - (SSR/SST) \tag{23}$$

where SSR and SST are the residual and total sum of squares, respectively. While SSR is calculated from $\sum(y_i - \hat{y}_i)^2$, SST is determined as $\sum(y_i - \bar{y})^2$. Here, y_i refers to the actual values, $\hat{y}_i$ is the predicted value, and $\bar{y}$ denotes the mean of the actual values.

3.2.2 Mean absolute error (MAE)

This metric is commonly used to evaluate the performance of predictive models; it represents the average of the absolute differences between actual values and predicted values, providing a clear indicator of prediction accuracy. Its formula is expressed as follows:.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{N} |yi - \hat{}yi| \tag{24}$$

3.2.3 Root Mean Squared Error (RMSE)

RMSE is considered one of the key measures of determining the effectiveness of predictions in a model because it indicates how far off the predictions have been from the actuals on average. A smaller RMSE means better performance of the model and its predictive power. RMSE is calculated using the formula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{25}$$

4. Results and Discussion

The probability distribution analysis for the Tweedy coefficient p is depicted in Figure 2. The estimated value of the Tweedy coefficient p = 1.7 is included in the range $1 < p < 2$, which means the distribution belongs to the compound family of distributions Poisson-Gamma. Thus, the variable under consideration has a probability distribution concentrated close to zero but taking positive non-zero continuous values.

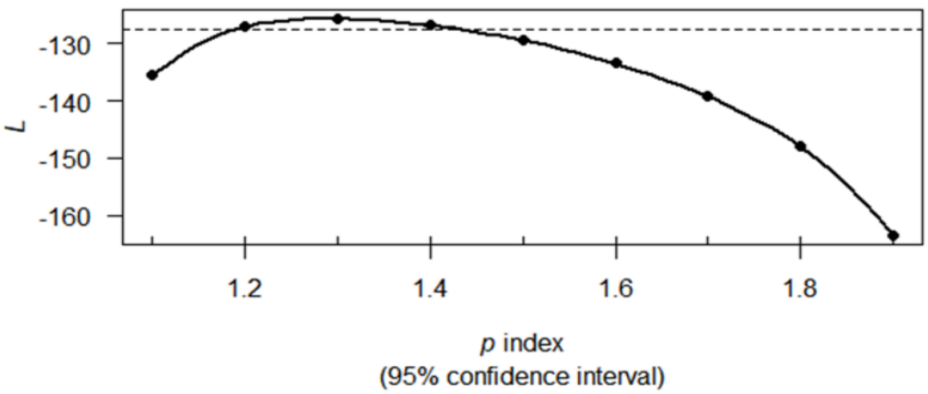


Figure 2: Likelihood profile for the Tweedie p Index with 95% Confidence Interval,p=1.7

The data were analyzed using the R program, with 70% used for training and 30% for testing; the results are listed in Table 2.

Table 2: Performance comparison table of prediction algorithms using R^2, MAE, and RMSE metrics

Model	AIC	p	phi	RMSE	MAE	R2
tweedie	5826.252	1.7		3.817732	0.8251875	**-0.111628**
XGBoost + WPO				3.435915	0.8025986	**0.096991775**
XGBoost		-	-	3.336765	0.8363192	**0.148356035**
XGBoost + HBOA				3.627336	0.8646015	**-0.006427434**



Figure 3: Performance of models using the RMSE metric

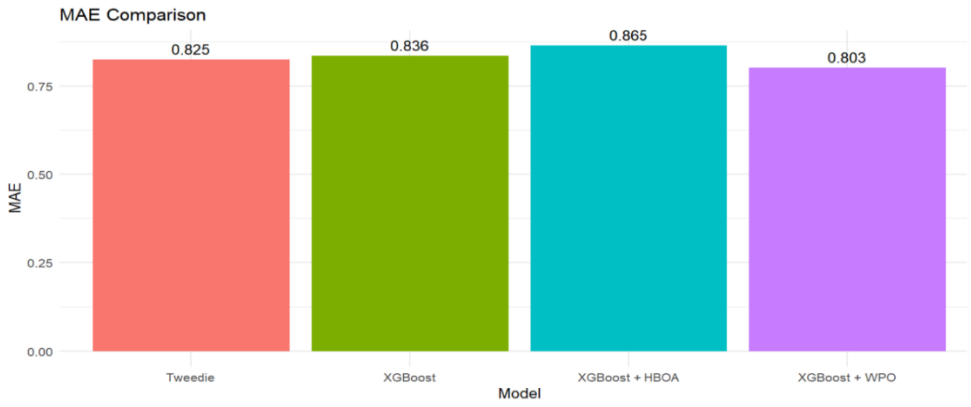


Figure 4: Performance of the algorithms using the MAE metric

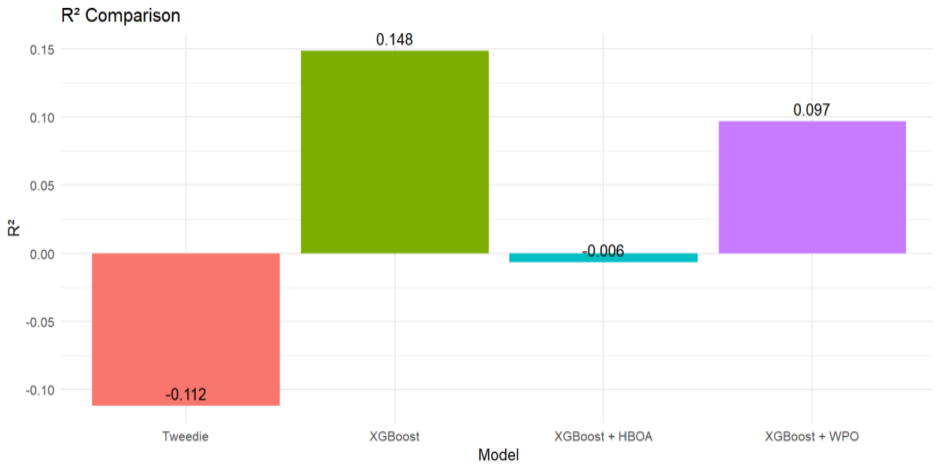


Figure 5: Performance of the algorithms using the R^2 metric

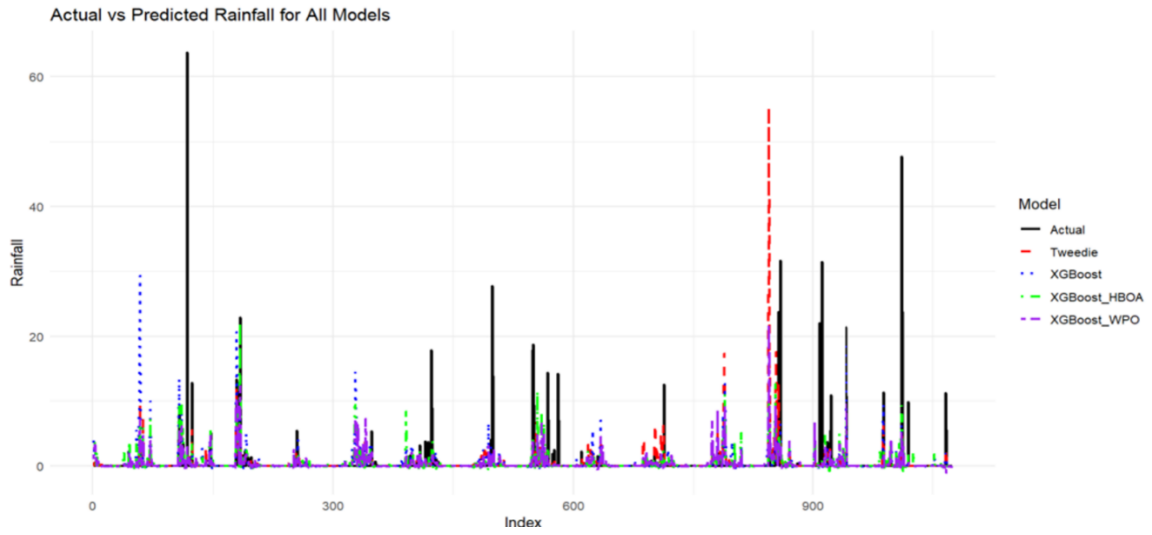


Figure 6: Model's prediction algorithms for rainfall amounts

5. Discussion

As shown in Figure 3 below, the XGBoost model managed to get the smallest RMSE value, around 3.337, implying better accuracy than the other models studied. The XGBoost model, when applied with WPO and HBOA, came after that and showed little or no significant improvement over the base model. The Tweedie model had the largest error value, around 3.818, indicating poor performance compared with the others in this experimental setting. In general, applying the optimization techniques does not provide much improvement in performance. Better outcomes could be obtained by using more refined parameter settings or a bigger dataset.

Figure 4 below illustrates a comparative analysis of four predictive models based on MAE, where lower values indicate better models. From the data, it is clear that the XGBoost model with WPO recorded the least MAE of approximately 0.803, meaning that it was the most accurate at minimizing absolute error. The second most accurate model was the Tweedie model, with an MAE of approximately 0.825, and the third was the XGBoost model, with an MAE of 0.836. On the other hand, the model that recorded the highest MAE (approximately 0.865) was the XGBoost model with HBOA, which performed the worst in this case. The Ordinary XGBoost (Baseline) achieved the minimum RMSE of 3.337, thus showing the highest predictive accuracy overall. It had an R^2 of 0.148, suggesting that about 15% of the variance of rainfall data could be explained, which is the best result compared to other tested models. Figure 5 shows a comparison between the predictive performances of various models according to the coefficient of determination (R^2). An R^2 closer to 1 indicates a better capability of explanation, whereas a negative value suggests poor performance of a model. The XGBoost model outperformed the other models, achieving the highest R^2 (about 0.148), demonstrating its explanatory power. The XGBoost + WPO model came next with a positive R^2 value (about 0.097), but it was not enough to surpass the baseline XGBoost model. Meanwhile, the XGBoost + HBOA model got a value close to zero but negative (about −0.006), which shows low explanatory capability. The Tweedie model showed the worst performance in comparison with other models, with a negative value of R^2 (about −0.112).

In general, the graph shows that the XGBoost algorithm is the most efficient algorithm for representing the data. Meanwhile, there were optimization algorithms like HBOA that did not show any significant improvement in this analysis.

Among other findings is that XGBoost integrated with WPO yielded significant improvements in RMSE and MAE compared to the traditional Tweedie model. Nevertheless, its R^2 was 0.097, which is lower than the R^2 of the regular XGBoost algorithm. It means that the optimization with WPO slightly reduced the explained variance while decreasing the absolute errors. In addition, the XGBoost-integrated HBOA performed worse than both XGBoost and WPO, with an RMSE of 3.254, an MAE of 0.786, and an R^2 close to zero. Moreover, Traditional Tweedie yielded the maximum RMSE (3.818) and the minimum R^2 (-0.112), indicating poor performance in the analyzed case. The value of AIC of Traditional Tweedie is equal to 5826.25.

In Figure 6, actual values are compared to predictions (Tweedie, WPO, XGBoost, and HBOA). The results indicate that XGBoost struggles with the characteristics of this dataset and requires significant improvements, such as parameter tuning or increasing the training data size, to improve performance.

6. Conclusions and Recommendations

The results clearly show that the XGBoost model outperformed the other models tested in predictive performance, achieving the highest R^2 and the lowest RMSE. This shows its greater ability to comprehend data and handle complex patterns compared to statistical models. While the use of the WPO algorithm improved the MAE score, no similar improvement was observed in other scores, such as R^2 and RMSE, indicating that optimizing a single measure does not necessarily lead to improvements in model performance. On the other hand, when used with the HBOA algorithm, the model did not achieve good performance compared to the base model, as the R^2 value decreased and errors increased.

The Tweedie model was the worst-performing out of all models considered here because of negative R^2 values. These findings point to deficiencies in the Tweedie model when working with this kind of data, compared to more sophisticated machine learning models. Thus, the conclusions drawn from the analysis confirm the superiority of reinforcement learning models, especially the XGBoost model, in terms of prediction accuracy compared to traditional statistical models. The current research highlights the need to analyze the influence of optimization algorithms, as partial optimization may not yield improvements in overall performance. Use the basic XGBoost model as your primary model for similar data, given its balanced and efficient performance. Avoid applying optimization algorithms haphazardly; analyze the influence of each on model performance before doing so. Think of new ways to combine the XGBoost model with various optimization algorithms like Bayesian Optimization or Grid/Random Search to allow you to make comparisons between the algorithms used here.

References

Attri, S., Gupta, S., & Singh, S. (2025). Risk forecasting using artificial intelligence and machine learning: Evidence from stock and cryptocurrency markets. In Artificial

intelligence for financial risk management and analysis (pp. 187–196). IGI Global Scientific Publishing.

Cai, T., Ma, F., Ye, Z., Zhou, W., Wang, M., He, Q., … Shen, J. (2025). Dynamic niche technology-based hybrid breeding optimization algorithm for multimodal feature selection. Scientific Reports, 15(1), Article 8001.

Cai, T., Qiao, Z., Ye, Z., Pan, H., Wang, M., Zhou, W., … Bai, W. (2024). Dynamic niches-based hybrid breeding optimization algorithm for solving multi-modal optimization problem. Mathematics, 12(17), Article 2779.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Association for Computing Machinery.

Cordeiro, G. M., Labouriau, R., & Botter, D. A. (2021). An introduction to Bent Jørgensen's ideas. Brazilian Journal of Probability and Statistics, 35(1), 2–20. https://doi.org/10.1214/19-BJPS458

Dalal, S., Rani, U., Lilhore, U. K., Dahiya, N., Batra, R., Nuristani, N., & Le, D. N. (2025). Optimized XGBoost model with whale optimization algorithm for detecting anomalies in manufacturing. Journal of Computational and Cognitive Engineering, 4(4), 413–423.

Gao, J. (2024). R-squared (R2)—How much variation is explained? Research Methods in Medicine & Health Sciences, 5(4), 104–109.

Ghafarian, F., Wieland, R., Lüttschwager, D., & Nendel, C. (2022). Application of extreme gradient boosting and Shapley additive explanations to predict temperature regimes inside forests from standard open-field meteorological data. Environmental Modelling & Software, 156, Article 105466.

Hadba, W. A. (2025). Tweedie regularization model and application. Communications in Statistics—Theory and Methods, 54(6), 1729–1746.

Hanif, I. (2020). Implementing extreme gradient boosting (XGBoost) classifier to improve customer churn prediction. In Proceedings of the 1st International Conference on Statistics and Analytics (ICSA 2019). European Alliance for Innovation. https://doi.org/10.4108/eai.2-8-2019.2290338

Hasan, M. M., & Dunn, P. K. (2010). Two Tweedie distributions that are near-optimal for modelling monthly rainfall in Australia. International Journal of Climatology, 31(9), 1389–1397.

Hasan, M. M., & Dunn, P. K. (2012). Understanding the effect of climatology on monthly rainfall amounts in Australia using Tweedie GLMs. International Journal of Climatology, 32(7), 1006–[end page needed].

Hasan, M. M., & Dunn, P. K. (2015). Seasonal rainfall totals of Australian stations can be modelled with distributions from the Tweedie family. International Journal of Climatology, 35, 3093–3101.

Hayati, M., & Permatasari, R. (2023). Rainfall prediction in statistical downscaling using Tweedie compound response and LASSO penalty. International Journal of Scientific Research in Science, Engineering and Technology, 10(3), 537–546.

Jørgensen, B. (1997). The theory of dispersion models. CRC Press. https://books.google.com/books?id=0gO7bgs_eSYC

Kaveh, A., & Ghazaan, M. I. (2017). Enhanced whale optimization algorithm for sizing optimization of skeletal structures. Mechanics Based Design of Structures and Machines, 45(3), 345–362.

Khodadadi, N., El-Kenawy, E. S. M., Faridmarandi, S., Ghahfarokhi, M. S., Ibrahim, A., & Mirjalili, S. (2024). A novel version of whale optimization algorithm for solving optimization problems. In Handbook of whale optimization algorithm (pp. 591–600). Academic Press.

Kumar, R., Goel, R., Sidana, N., Sharma, A. P., Ghai, S., Singh, T., … Ahmad, V. (2024). Enhancing climate forecasting with AI: Current state and future prospect. F1000Research, 13, Article 1094.

Kurz, C. F. (2017). Tweedie distributions for fitting semicontinuous health care utilization cost data. BMC Medical Research Methodology, 17(1), Article 171.

Madlool, Z. A. (2024). Develop whale optimization algorithm (WOA) in genetic method to predict the optimal treatment for diseases. Journal of Al-Qadisiyah for Computer Science and Mathematics, 16(4), 300–310.

Mahmood, S., Bawany, N. Z., & Tanweer, M. R. (2023). A comprehensive survey of whale optimization algorithm: Modifications and classification. Indonesian Journal of Electrical Engineering and Computer Science, 29(2), 899–910.

Palacio, S. M. (2020). Machine-learning applied methods [Doctoral dissertation, Universitat de Barcelona]. Universitat de Barcelona Digital Repository. https://hdl.handle.net/2445/169629

Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. Advances in Engineering Software, 95, 51–67.

Museba, T. (2022). Adaptive particle swarm optimized XGBoost ensemble algorithm for online credit scoring [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-1660274/v1

Nelson, C. D., Ngare, P., & Odongo, L. (2018). A Poisson-Gamma model for zero inflated rainfall data. Journal of Probability and Statistics, 2018.

Nguyen, D. T., & Le-Hoai, L. (2011). Tradeoff time-cost-quality optimization using hybrid bio-inspired optimization algorithms. Applied Soft Computing, 11(4), 3846–3858. https://doi.org/10.1016/j.asoc.2011.02.001

Obiora, C. N., Ali, A., & Hasan, A. N. (2021). Implementing extreme gradient boosting (XGBoost) algorithm in predicting solar irradiance. In 2021 IEEE PES/IAS PowerAfrica (pp. 1–5). IEEE.

Osman, A. I. A., Ahmed, A. N., Chow, M. F., Huang, Y. F., & El-Shafie, A. (2021). Extreme gradient boosting (XGBoost) model to predict the groundwater levels in Selangor, Malaysia. Ain Shams Engineering Journal, 12(2), 1545–1556.

Pham, Q. V., Mirjalili, S., Kumar, N., Alazab, M., & Hwang, W. J. (2020). Whale optimization algorithm with applications to resource allocation in wireless networks. IEEE Transactions on Vehicular Technology, 69(4), 4285–4297.

Rahman, I., & Mohamad-Saleh, J. (2018). Hybrid bio-inspired computational intelligence techniques for solving power system optimization problems: A comprehensive survey. Applied Soft Computing, 69, 72–130.

Rydén, J. (2025). Statistical analysis of monthly precipitation in Sweden using the Tweedie distribution. Theoretical and Applied Climatology, 156(10), Article 512. https://doi.org/10.1007/s00704-025-05792-6

Singh, A. K., Kumar, P., Ali, R., Al-Ansari, N., Vishwakarma, D. K., Kushwaha, K. S., … Heddam, S. (2022). An integrated statistical-machine learning approach for runoff prediction. Sustainability, 14(13), Article 8209.

So, B., & Valdez, E. A. (2025). Zero-inflated Tweedie boosted trees with CatBoost for insurance loss analytics. Applied Soft Computing, 177, Article 113226.

Suhaila, J. (2023). Tweedie models for Malaysia rainfall simulations with seasonal variabilities. Journal of Water and Climate Change, 14(10), 3648–3670.

Trojovský, P., & Dehghani, M. (2023). A new bio-inspired metaheuristic algorithm for solving optimization problems based on walruses behavior. Scientific Reports, 13(1), Article 8775.

Xie, Y., Mi, G., Tan, D., & Liu, C. (2025). WOA-XGBoost based railway accident type prediction and cause analysis. Engineering Letters, 33(3), 732–742.

Yang, X. S. (2014). Nature-inspired optimization algorithms (Vol. 20). Elsevier.

Yang, X. S. (2014). Nature-inspired optimization algorithms. Elsevier. https://books.google.com/books?id=BbHVAQAAQBAJ

Zhang, X., & Li, Y. (2020). HBOA for parameter optimization in predictive models. Expert Systems with Applications, 159, Article 11359.

Zhou, H., Qian, W., & Yang, Y. (2022). Tweedie gradient boosting for extremely unbalanced zero-inflated data. Communications in Statistics—Simulation and Computation, 51(9), 5507–5529.