Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Statistical Benchmark of Metaheuristic Algorithms for Nonlinear Soil Constitutive Model Calibration

Mohammad Alibabaei Shahraki

alibabaei_mohammad@yahoo.com

Technical and Vocational University

Research Article

Keywords: Constitutive model calibration, Geotechnical optimization, Hyperbolic model, Duncan–Chang model, Modified Cam-Clay model, Metaheuristic algorithms

Posted Date: August 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10576567/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Statistical Benchmark of Metaheuristic Algorithms for Nonlinear Soil Constitutive Model Calibration

Mohammad Alibabaei Shahraki[1*]

1- Department of Civil Engineering, Technical and Vocational University, Tehran, Iran

Abstract

Accurate calibration of nonlinear constitutive models is essential for reliable prediction of geomaterial behavior in computational geotechnical engineering. This study presents a benchmarking framework for evaluating eight population-based metaheuristic optimization algorithms for calibrating three widely used constitutive models: Hyperbolic, Duncan–Chang, and Modified Cam–Clay. Synthetic datasets with a signal-to-noise ratio of 20 dB were analyzed under identical computational settings using 30 independent runs for each algorithm. Performance was evaluated through descriptive statistics, convergence analysis, Friedman ranking, Wilcoxon signed-rank tests with Holm–Bonferroni correction, Dolan–Moré performance profiles, and parameter estimation accuracy analysis. The results indicate that optimization performance depends strongly on constitutive model complexity and parameter identifiability. While several algorithms show comparable behavior for simpler calibration problems, larger performance differences emerge as nonlinearity and parameter interactions increase. The findings also demonstrate that low objective-function values do not necessarily imply accurate recovery of constitutive parameters, highlighting the influence of equifinality in nonlinear inverse problems. The proposed framework provides a reproducible basis for comparing optimization algorithms and supports more comprehensive evaluation of constitutive model calibration strategies in computational geotechnical engineering.

Keywords: Constitutive model calibration; Geotechnical optimization; Hyperbolic model; Duncan–Chang model; Modified Cam-Clay model; Metaheuristic algorithms

1. Introduction

Advanced numerical modeling, particularly the Finite Element Method (FEM), has become an indispensable and standard tool in the design, analysis, and performance prediction of geotechnical structures [1, 2]. The accuracy and reliability of predictions derived from such modeling depend directly and nonlinearly on the quality of input parameters for soil constitutive

* Email: alibabaei_mohammad@yahoo.com

models. These parameters, which represent the physical and mechanical characteristics of soil, are typically extracted through laboratory tests (e.g., triaxial, oedometer) or in-situ field investigations (e.g., cone penetration, pressuremeter) [3, 4]. However, the precise determination of these parameters is consistently challenged by the inherent complexities of soil behavior. Soil, as a multiphase, heterogeneous, and anisotropic material, exhibits highly nonlinear responses, path-dependent behavior, strain-rate sensitivity, and susceptibility to measurement noise [5, 6]. These complexities ensure that the relationship between model parameters and observed soil response is not straightforward but constitutes a complex inverse problem requiring advanced identification methodologies.

The primary challenge in soil parameter calibration lies in the multi-modal nature of the error landscape. Due to high parameter correlation, different combinations of parameters may produce similar responses within specific strain ranges—a phenomenon known as solution non-uniqueness [7, 8]. Furthermore, laboratory data are invariably accompanied by varying levels of noise and uncertainty arising from instrument errors, sample heterogeneity, and boundary condition limitations [9, 10]. Traditional calibration approaches, often based on trial-and-error or gradient-based optimization, fail when confronted with such complex error surfaces. Gradient-based methods are highly sensitive to initial starting points and readily become trapped in local minima, particularly when the objective function is non-smooth or noisy [11, 12]. These limitations frequently yield parameter sets that, while mathematically optimal, are physically unstable and lack generalizability to different loading conditions.

In recent decades, metaheuristic optimization algorithms have emerged as powerful and robust alternatives to classical optimization methods in geotechnical engineering. Inspired by natural, evolutionary, physical, or social behaviors, these algorithms do not require gradient information and possess strong global search capabilities [13, 14]. Their core mechanism relies on balancing two fundamental phases: exploration (searching new regions of the solution space) and exploitation (refining promising solutions), enabling them to escape local minima and converge toward global optima [15, 16]. Algorithms such as Genetic Algorithms (GA) [17], [18], Differential Evolution (DE) [19], and Particle Swarm Optimization (PSO) [20] have been successfully applied to various engineering problems, including geotechnical and geological parameter estimation [5, 10, 21]. The principal advantage of these methods lies in their ability to handle complex, discontinuous, and noisy objective functions—characteristics commonly encountered in soil behavior modeling.

Nevertheless, the rapidly expanding diversity of new metaheuristic algorithms over the past two decades has introduced a new challenge for researchers and practitioners. Algorithms such as Grey Wolf Optimizer (GWO), Harris Hawks Optimization (HHO), Slime Mould Algorithm (SMA), Cloud Drift Optimization (CDO), and Crayfish Optimization Algorithm (COA) have been proposed for solving complex optimization problems. Their search strategies differ considerably, motivating comparative evaluation under a common benchmarking framework. Each algorithm employs a unique search mechanism: some operate based on social hierarchy, others on hunting behavior, and still others on physical phenomena such as cloud drift or crustacean behavior. The fundamental question arises: which of these algorithms is most suitable for calibrating complex, nonlinear soil constitutive models? Answering this question cannot rely on speculation but requires comprehensive empirical comparison grounded in statistical evidence. Unfortunately, the existing literature in this domain has not yet matured sufficiently to provide definitive guidance, as most studies conduct fragmented comparisons without standardized methodologies.

This study presents a unified benchmarking framework for comparing eight metaheuristic algorithms for the calibration of three nonlinear constitutive soil models. The framework combines identical computational settings, a physically-constrained objective function, repeated independent runs, and statistical analysis to enable a consistent comparison across algorithms. The benchmark includes eight metaheuristic algorithms—Cloud Drift Optimization (CDO) [22], Chimp Optimization [23], Grey Wolf Optimizer (GWO) [24], Harris Hawks Optimization (HHO) [25], Crayfish Optimization Algorithm (COA) [26], Slime Mould Algorithm (SMA) [27], Seagull Optimization Algorithm (SOA) [28], and Particle Swarm Optimization (PSO) [20]—in calibrating three constitutive models of varying complexity: Hyperbolic, Duncan–Chang, and Modified Cam-Clay. These models were chosen to represent a spectrum of complexity: the Hyperbolic Model (representing basic nonlinear elasticity), the Duncan–Chang Model (representing empirical nonlinear stress-strain behavior widely used in practice), and the Modified Cam-Clay (MCC) Model (representing critical state soil mechanics and elastoplasticity). Each model presents unique challenges regarding parameter interdependence, sensitivity, and nonlinearity, making them ideal candidates for rigorous calibration testing. To ensure result significance, this study employs a rigorous validation protocol comprising 30 independent runs, Friedman ranking tests, Wilcoxon signed-rank tests with Holm–Bonferroni correction, and Dolan–Moré performance profiles.

The objective function combines RMSE with penalties for boundary violations,

parameter consistency, and local regularization to discourage physically unrealistic parameter estimates. Overall, the proposed benchmark provides a consistent basis for comparing optimization strategies for constitutive soil model calibration.

Since Cloud Drift Optimization (CDO) was originally proposed by the author in previous work, special care was taken to reduce potential experimental bias. All competing algorithms were implemented using their original formulations and the parameter settings recommended in their respective original publications. No algorithm-specific parameter tuning or problem-dependent optimization was performed for any method, including CDO. Furthermore, all optimization algorithms were evaluated using the same objective function, which combines the root mean square error (RMSE) with boundary, parameter-consistency, and local regularization penalties. This unified formulation ensures that performance differences arise from the optimization strategies rather than from differences in objective function design. Accordingly, the reported results should be interpreted as a comparison of optimization strategies under a common benchmark rather than as evidence of the universal superiority of any particular algorithm.

1.1. Research Gap

Despite the growing application of metaheuristic algorithms in geotechnical engineering, several important research gaps remain that limit the comparability and generalizability of existing studies.

First, most studies focus on applying one or a small number of algorithms (commonly GA or PSO) to a single soil model, with limited consideration of more recent metaheuristic approaches [16]. As a result, systematic benchmarking across classical and modern algorithms under identical conditions remains limited, making it difficult to assess their relative performance in a consistent manner. In many cases, classical evolutionary algorithms are emphasized, while newer swarm-intelligence or physics-based methods are not fully explored [7, 8]. Second, many existing comparisons lack rigorous statistical validation. Due to the stochastic nature of metaheuristic algorithms, single-run or limited-run analyses can be misleading [29, 30]. However, a significant number of studies rely mainly on mean and standard deviation reporting without applying non-parametric statistical tests such as Friedman or Wilcoxon signed-rank tests. In addition, performance profile analysis is rarely adopted in geotechnical optimization studies [31].

Finally, reproducibility remains a major concern in computational geotechnical research. In many cases, detailed algorithm settings, datasets, or implementation codes are not fully reported [32], which restricts independent verification and cross-study comparison. These limitations highlight the need for a unified and transparent benchmarking framework with standardized experimental conditions and statistical evaluation procedures.

2. Soil Constitutive Models

Accurate mathematical representation of soil behavior is paramount in geotechnical engineering design and analysis. Soil exhibits highly nonlinear, stress-dependent, and path-dependent behavior, which necessitates the use of constitutive models capable of capturing these complexities. In this study, three distinct constitutive models were selected to benchmark the performance of metaheuristic optimization algorithms. These models were chosen to represent a spectrum of complexity: the Hyperbolic Model (representing basic nonlinear elasticity), the Duncan–Chang Model (representing empirical nonlinear stress-strain behavior widely used in practice), and the Modified Cam-Clay (MCC) Model (representing critical state soil mechanics and elastoplasticity). Each model presents unique challenges regarding parameter interdependence, sensitivity, and nonlinearity, making them ideal candidates for rigorous calibration testing.

2.1. Hyperbolic Model

The Hyperbolic Model, originally proposed by Kondner [33], is one of the most widely used constitutive relationships for representing the nonlinear stress–strain behavior of geomaterials. The model assumes that the stress response asymptotically approaches an ultimate value with increasing strain, thereby capturing the gradual stiffness degradation observed in laboratory testing. The constitutive relationship adopted in this study is expressed as: [33]:

$$\sigma = \frac{G_0 \gamma}{1 + \gamma/\gamma_{ref}} \tag{1}$$

Where σdenotes shear stress, γ is shear strain, G_0 represents the initial shear modulus, and γ_{ref} is the reference strain controlling the curvature of the stress–strain response.

Calibration Challenges: This model was selected because it represents a low-dimensional nonlinear calibration problem with moderate parameter interaction. Although only two parameters are involved, strong correlation may exist between G_0 and γ_{ref}, particularly in

the presence of measurement noise. Consequently, the Hyperbolic Model provides a suitable benchmark for evaluating the convergence characteristics and parameter-identification capability of optimization algorithms.

2.2. Duncan–Chang Model

The Duncan–Chang Model [34] model is one of the most extensively applied nonlinear constitutive models in geotechnical engineering and forms the basis of numerous finite-element analyses of earth structures. The model describes the stress–strain response using a hyperbolic relationship and accounts for the progressive mobilization of soil strength. The model can be expressed as:

$$\varepsilon_a = \frac{q}{E_i\left(1 - \frac{q}{q_f}\right)} \tag{2}$$

Where ε_a is the axial strain, $q = (\sigma_1 - \sigma_3)$ is the deviatoric stress, E_i is the initial tangent modulus, and q_f is the asymptotic failure stress. The initial tangent modulus is commonly represented as

$$E_i = KP_a\left(\frac{\sigma_3}{P_a}\right)^n \tag{3}$$

Where K is the modulus number, n is the modulus exponent, P_a is the atmospheric pressure, and σ_3 is the confining pressure.

Calibration Challenges: The Duncan–Chang model introduces a higher degree of nonlinearity than the Hyperbolic Model due to the interaction between stiffness and strength parameters. The resulting calibration problem exhibits stronger parameter coupling and increased sensitivity to experimental noise, making it an appropriate intermediate-complexity benchmark for optimization algorithms.

2.3. Modified Cam-Clay Model

The Modified Cam-Clay Model (MCC) [35] model is a fundamental constitutive framework within Critical State Soil Mechanics and is widely used to describe the elastoplastic behavior of clays. Unlike empirical constitutive relationships, MCC incorporates yield-surface evolution, volumetric hardening, and stress-path dependency within a unified theoretical

framework. The yield function of the MCC model is expressed as:

$$f(q, p') = q^2 + M^2 p'(p' - p_c) = 0 \quad (4)$$

Where q is the deviatoric stress, p' is the effective mean stress, M is the critical-state stress ratio, and p_c is the pre-consolidation pressure. The volumetric behavior is governed by the compression and swelling indices, λ and κ, which control plastic and elastic compressibility, respectively.

Calibration Challenges: In contrast to the Hyperbolic and Duncan–Chang models, the MCC model represents an elastoplastic constitutive formulation with coupled stress and volumetric responses. Consequently, the associated calibration problem involves stronger parameter interactions, nonlinearity, and identifiability challenges. For this reason, MCC serves as a demanding benchmark for assessing the robustness and reliability of optimization algorithms in constitutive parameter estimation.

Collectively, these three models provide a comprehensive testbed for optimization algorithms. The Hyperbolic model tests basic nonlinear fitting, the Duncan–Chang model tests power-law hardening calibration, and the Modified Cam-Clay model tests logarithmic elastoplastic parameter identification. The parameters defined herein constitute the search space vectors $\mathbf{x}$ for the metaheuristic algorithms described in Section 4, where the objective is to minimize the discrepancy between model predictions and experimental data.

3. Objective Functions

The accuracy of constitutive model calibration depends directly on the formulation of the objective function used to quantify the discrepancy between model predictions and observed data. In inverse analysis problems, the objective function serves as the criterion that guides the optimization process toward the parameter set that best reproduces the measured material response. To ensure a fair and unbiased comparison among optimization algorithms, this study adopts a unified objective function for all constitutive models and optimization methods. The objective function is based solely on the discrepancy between predicted and observed responses and does not incorporate algorithm-specific modifications or artificial landscape transformations. This formulation allows the benchmark results to reflect the intrinsic search capability, convergence behavior, and parameter-identification performance of the competing optimization algorithms. For all calibration problems considered in this study, the objective

function is defined using the Root Mean Square Error (RMSE) between the constitutive model predictions and the reference dataset.

3.1. Root Mean Square Error (RMSE)

cause it provides a direct measure of the discrepancy between predicted and observed responses while maintaining sensitivity to large deviations. RMSE is widely used in constitutive model calibration, system identification, and parameter estimation problems due to its simplicity, interpretability, and compatibility with both deterministic and stochastic optimization methods. The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i^{pred} - y_i^{obs}\right)^2} \tag{5}$$

Where N is the total number of observations, y_i^{obs} denotes the observed response, and y_i^{pred} represents the corresponding model prediction obtained using a candidate parameter vector. A lower RMSE value indicates a closer agreement between the constitutive model response and the reference data. Consequently, the calibration problem can be formulated as the following optimization task:

$$\min_{\boldsymbol{x}} RMSE(\boldsymbol{x}) \tag{6}$$

Where $\boldsymbol{x}$ denotes the vector of constitutive model parameters to be identified.

3.2. Optimization Formulation

The constitutive model calibration problem is formulated as a bounded nonlinear optimization problem. For each constitutive model, the objective is to identify the parameter vector that minimizes the RMSE while satisfying physically meaningful parameter bounds. The optimization problem can be expressed as:

$$\min_{\boldsymbol{x}} RMSE(\boldsymbol{x}) \tag{7}$$

subject to

$$\boldsymbol{xL} \leq \boldsymbol{x} \leq \boldsymbol{xU} \tag{8}$$

Parameter bounds are enforced directly within the optimization algorithms through boundary-control mechanisms. Therefore, no additional penalty functions, artificial regularization terms, or landscape modifications are introduced into the objective function. This approach ensures that the comparison remains focused on the optimization capability of the algorithms rather than on problem-specific objective-function engineering.

By employing a common RMSE-based objective function and identical parameter bounds, all competing algorithms are evaluated under the same calibration framework, allowing statistically meaningful comparisons of convergence performance, robustness, and parameter recovery accuracy.

4. Optimization Algorithms

o ensure a comprehensive and unbiased assessment of optimization performance, both deterministic and metaheuristic optimization approaches were considered in this study. Deterministic methods provide an important baseline for constitutive model calibration because they exploit gradient information and have long been established in parameter-estimation problems. Metaheuristic algorithms, on the other hand, offer greater flexibility for handling nonlinear, multimodal, and nonconvex optimization landscapes. Accordingly, the benchmark framework includes one deterministic optimization method and several state-of-the-art metaheuristic algorithms. All algorithms were evaluated under identical parameter bounds, objective-function definitions, stopping criteria, and statistical assessment procedures.

4.1. Cloud Drift Optimization

CDO simulates the atmospheric drift motion of cloud particles influenced by adaptive environmental forces. The algorithm employs a fitness-based weighting mechanism to guide particles toward promising regions while maintaining diversity through stochastic drift velocities. The core position update is governed by an adaptive weight $w_{i,j}$and drift vector vb:

$$X_{i(j)}^{(t+1)} = X^*(j) + 0.8 \cdot vb(j) \cdot (w_{i,j} \cdot X_A(j) - X_B(j)) \tag{9}$$

Where $X^*$is the best-known solution, $X_A, X_B$are random agents, and $vb \sim U(-0.2a, 0.2a)$represents a random adjustment controlled by factor a, mimicking cloud

formation refinement.

Although CDO was originally proposed by the author, all competing algorithms were implemented using identical objective functions, parameter bounds, population sizes, stopping criteria, and statistical evaluation procedures to ensure a fair and unbiased comparison.

4.2. Chimp Optimization Algorithm

Inspired by the intelligent hunting behavior of chimpanzees, this algorithm models four cooperative roles: Driver, Barrier, Chaser, and Attacker. The final position is determined by the collective influence of these four independent search agents:

$$\vec{X}_{chimp}^{t+1} = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3 + \vec{X}_4}{4} \tag{10}$$

Where $\vec{X}_1 \ldots \vec{X}_4$represent the positions updated by each hunting role, enabling adaptive switching between global and local search.

4.3. Grey Wolf Optimizer

GWO mimics the leadership hierarchy of grey wolves, where the top three solutions (alpha, beta, delta) guide the population. The search agents update their positions by encircling the prey based on the estimated location of the best three agents:

$$\vec{X}_{t+1} = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \tag{11}$$

Where $\vec{X}_1, \vec{X}_2, \vec{X}_3$are the position vectors calculated relative to the alpha, beta, and delta wolves, respectively.

4.4. Harris Hawks Optimization

HHO is based on the surprise pounce behavior of Harris hawks, utilizing an energy factor $E$to dynamically switch between exploration and exploitation phases. The phase transition is defined as:

$$E = 2E_0\left(1-\frac{t}{T}\right) \tag{12}$$

Where $E_0$is the initial energy and $T$is the maximum iterations. When $|E| < 1$, the algorithm performs local exploitation using various besiege strategies.

4.5. Crayfish Optimization Algorithm

Proposed by Jia et al. (2023), COA simulates the summer resort behavior and competition mechanisms of crayfish. The algorithm balances exploration (seeking suitable temperatures) and exploitation (competition for food) based on a temperature factor. The position update during the competition phase is:

$$X_i^{t+1} = X_i^t + C \cdot (X_{best}^t - X_i^t) + D \cdot (X_{random}^t - X_i^t) \tag{13}$$

Where C and D are competition coefficients influenced by the temperature factor, ensuring dynamic adaptation to the search landscape.

4.6. Slime Mould Algorithm

SMA simulates the oscillatory foraging behavior of slime mould using adaptive weights based on fitness quality. The weight update mechanism allows the algorithm to connect food sources with optimal paths:

$$W = \begin{cases} 1 + r \cdot \log\left(\dfrac{F_{best} - F_i}{F_{best} - F_{worst}}\right) & \text{if condition} \\ 1 - r \cdot \log\left(\dfrac{F_{best} - F_i}{F_{best} - F_{worst}}\right) & \text{otherwise} \end{cases} \tag{14}$$

Where F_i is the fitness of the current agent, enabling positive and negative feedback propagation.

4.7. Seagull Optimization Algorithm

SOA models the migration and attack behaviors of seagulls. The algorithm uses a linearly

decreasing control parameter F_c to guide agents toward the best solution using a spiral movement pattern. The position update is formulated as:

$$X_i^{t+1} = (D_s \cdot e^{b \cdot l} \cdot \cos(2\pi l)) + X_{best} \quad (15)$$

Where D_s is the distance vector influenced by F_c, b is a spiral constant, and l is a random parameter, facilitating precise local exploitation around the best solution.

4.8. Particle Swarm Optimization

PSO simulates the social behavior of bird flocks, where particles adjust their trajectories based on personal and global best experiences. The velocity update equation is defined as:

$$v_i^{t+1} = wv_i^t + c_1 r_1 (pbest_i - x_i^t) + c_2 r_2 (gbest - x_i^t) \quad (16)$$

Where w is the inertia weight, and c_1, c_2 are acceleration coefficients controlling the convergence behavior.

5. Experimental Setup

To ensure a fair, reproducible, and statistically robust comparison of the eight metaheuristic algorithms, a comprehensive experimental framework was designed within the MATLAB environment. This framework standardizes the computational budget, defines physically meaningful search spaces for soil parameters, and incorporates realistic data uncertainties. All algorithms were implemented using MATLAB R2023a, leveraging the Global Optimization Toolbox and Statistics and Machine Learning Toolbox for auxiliary functions, with consistent random stream initialization to ensure comparability across independent runs.

5.1. Algorithm Configuration and Computational Budget

To ensure a fair and unbiased comparison, all optimization algorithms were evaluated under a common computational framework. The benchmark included one deterministic optimization method, Nonlinear Least Squares (NLS), and eight metaheuristic algorithms: Cloud Drift Optimization (CDO), Chimp Optimization Algorithm, Grey Wolf Optimizer (GWO), Harris Hawks Optimization (HHO), Crayfish Optimization Algorithm (COA), Slime Mould Algorithm (SMA), Seagull Optimization Algorithm (SOA), and Particle Swarm Optimization (PSO). The following configuration was adopted for all population-based

algorithms:

- Population Size: 100 search agents. This value was selected to provide sufficient population diversity while maintaining reasonable computational cost.
- Maximum Iterations: 30 generations. This termination criterion was selected to provide a balanced trade-off between solution quality and computational efficiency.
- Independent Runs: 30 independent runs were performed for each stochastic algorithm and constitutive model. Multiple runs were necessary to account for random initialization effects and to support subsequent statistical analyses, including Friedman ranking and Wilcoxon signed-rank testing.
- Termination Condition: All metaheuristic algorithms were terminated after reaching the maximum number of iterations. Early stopping criteria were intentionally avoided to ensure identical computational budgets across competing methods.
- Random Number Generation: Reproducibility was ensured using MATLAB's Mersenne Twister random number generator. Different random seeds were assigned to the 30 independent runs through the *rng* function.

The deterministic NLS approach was implemented using MATLAB's *lsqcurvefit* function. Since NLS does not employ population-based search, convergence was governed by the default optimization tolerances and stopping criteria of the MATLAB Optimization Toolbox. The selected computational budget was intentionally kept identical for all metaheuristic algorithms to ensure equal computational effort. The objective of this study is not to identify the optimal parameterization of individual optimizers but rather to evaluate their relative performance under a common and controlled benchmark environment. To minimize subjective tuning bias, all algorithms were implemented using parameter settings recommended in their original publications or widely adopted reference implementations. No problem-specific hyperparameter tuning was performed for any optimization method. Furthermore, identical objective functions, parameter bounds, datasets, and evaluation procedures were used for all algorithms to ensure a transparent and unbiased comparison.

5.2. Parameter Search Space Definition

A critical aspect of inverse parameter identification is the definition of physically plausible bounds for each constitutive model parameter. Search spaces that are too wide may

lead to inefficient exploration and unphysical solutions, while overly restrictive bounds may exclude the true optimal parameters. The bounds adopted in this study (Table 1) were derived from extensive geotechnical literature, typical laboratory test ranges for sands and clays, and preliminary sensitivity analyses.

Table 1. Parameter Bounds for Constitutive Models

Model	Parameter	Physical Meaning	Lower Bound	Upper Bound	True Value
Hyperbolic	G_0	Initial shear modulus	10	5000	45
	γ_{ref}	Reference shear strain	0.001	0.5	0.008
Duncan–Chang	K	Modulus number	50	5000	950
	n	Modulus exponent	0.1	1.5	0.45
	Rf	Failure ratio	0.5	1.0	0.82
Modified Cam-Clay	λ	Compression index	0.05	0.50	0.20
	κ	Swelling index	0.01	0.15	0.05
	M	Critical state slope	0.8	1.5	1.20
	pc	Pre-consolidation pressure (kPa)	50	1000	400

Rationale for Bound Selection:

- **Hyperbolic Model**: The initial shear modulus G_0 spans a wide range representative of different soil stiffness conditions, from relatively soft geomaterials to dense granular soils. The reference strain γ_{ref} was bounded to capture the transition from near-linear small-strain behavior to the nonlinear response typically.

- **Duncan–Chang Model**: The modulus number K and modulus exponent n were selected based on ranges commonly reported for granular and cohesionless soils. The failure ratio R_f was restricted to values between 0.5 and 1.0 because this parameter represents the fraction of asymptotic failure stress mobilized by the hyperbolic constitutive relationship and is generally constrained within this interval.

- **Modified Cam-Clay Model**: The compression index λ and swelling index κ were bounded according to ranges typically reported for normally consolidated and lightly

overconsolidated clays. The critical-state parameter M was selected to cover commonly observed critical-state friction characteristics of fine-grained soils. The pre-consolidation pressure p_c was assigned a sufficiently broad range to represent different consolidation histories while remaining within realistic geotechnical conditions. These bounds collectively allow the calibration framework to capture a wide spectrum of clay behavior without introducing unrealistic parameter combinations.

5.3. Synthetic Data Generation and Noise Injection

To isolate algorithmic performance from uncertainties inherent in real laboratory data, synthetic experimental datasets were generated for each constitutive model using MATLAB's numerical computing capabilities. The procedure was as follows:

1. **Reference Parameter Selection:** A reference parameter vector $\mathbf{x}^*$ was specified for each constitutive model based on the values presented in Table 1.

2. **Data Point Generation:** For each model, $N = 50$ data points were generated by evaluating the corresponding constitutive equations over representative stress or strain ranges. The generated responses were treated as the noise-free reference behavior.

3. **Noise Injection:** To emulate experimental uncertainty, additive Gaussian noise was superimposed on the synthetic responses according to:

$$y_i^{obs} = y_i^{model}(\boldsymbol{x}^*) + \varepsilon_i \tag{17}$$

Where $y_i^{model}(\boldsymbol{x}^*)$ denotes the constitutive model prediction corresponding to the reference parameter vector and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is a normally distributed random variable with zero mean and variance σ^2.

The noise level was selected to achieve a Signal-to-Noise Ratio (SNR) of 20 dB, representing a moderate level of measurement variability commonly encountered in geotechnical laboratory testing. The same synthetic datasets were used for all optimization algorithms to ensure consistency and comparability of the benchmark results. Because the true parameter values are known in advance, this framework enables direct evaluation of objective-function convergence, parameter recovery accuracy, robustness, and statistical repeatability across competing optimization methods.

5.4. Implementation Details and Reproducibility

All experiments were conducted on identical computational hardware and within a unified MATLAB environment to eliminate variability arising from computational infrastructure. The source code, synthetic datasets, and optimization settings are provided to facilitate reproducibility and independent verification of the results. Key implementation choices include:

- Vectorization: All constitutive model equations and objective-function evaluations were fully vectorized to improve computational efficiency and ensure consistent execution across optimization algorithms.

- Constraint Handling: Parameter bounds were enforced directly through algorithm-specific boundary-control mechanisms. For population-based algorithms, a reflection-based boundary handling strategy was employed to preserve feasibility while maintaining population diversity.

- Objective Function Evaluation: The same RMSE-based objective function was used for all optimization algorithms. No additional penalty functions, parameter-centering mechanisms, artificial regularization terms, or landscape modifications were introduced. Consequently, all optimization methods were evaluated under an identical and unbiased objective-function framework.

- Software Environment: All simulations and optimization procedures were implemented in MATLAB R2024a. The deterministic Nonlinear Least Squares (NLS) method was implemented using the lsqcurvefit function available in the MATLAB Optimization Toolbox.

All algorithms were executed using their original formulations and the parameter settings recommended in their respective publications or standard reference implementations, without problem-specific hyperparameter tuning. Furthermore, identical datasets, parameter bounds, and termination criteria were employed for all optimization methods. This experimental design ensures reproducibility, consistency, and a fair comparison between deterministic and metaheuristic optimization approaches. The resulting benchmark framework provides a robust basis for the statistical analyses and comparative evaluations presented in the subsequent sections.

6. Statistical Analysis

In comparative optimization studies, reporting mean error values alone is insufficient to draw scientifically valid conclusions. Due to the stochastic nature of metaheuristic algorithms, performance metrics exhibit variability across independent runs. Therefore, rigorous non-parametric statistical tests are essential to determine whether observed performance differences are statistically significant or merely artifacts of random variation. This section details the comprehensive statistical framework employed to evaluate and rank the eight metaheuristic algorithms across the three soil constitutive models. The analysis proceeds hierarchically: (1) the Friedman test assesses overall ranking differences across all algorithms, (2) the Wilcoxon signed-rank test performs pairwise comparisons against the best-performing algorithm, (3) the Success Rate quantifies reliability in achieving acceptable solutions, and (4) Performance Profiles provide a holistic visualization of algorithmic robustness across all problem instances.

6.1. Friedman Test

The Friedman test is a non-parametric alternative to the repeated-measures ANOVA, designed to detect differences in treatments across multiple test attempts. In the context of algorithm benchmarking, it evaluates whether the performance rankings of k algorithms across Nindependent problem instances (here, combinations of soil models and independent runs) differ significantly. Unlike parametric tests, the Friedman test does not assume the normality of the underlying error distributions, making it ideal for optimization results that often exhibit skewness or heteroscedasticity.

Procedure: For each independent run $i(i = 1, \dots, N)$, algorithms are ranked based on their objective function value F, with rank $r_{ij} = 1$ assigned to the best-performing algorithm and $r_{ij} = k$ to the worst. Ties are assigned average ranks. The average rank for algorithm j is computed as:

$$\bar{R}_j = \frac{1}{N}\sum_{i=1}^{N} r_{ij} \tag{18}$$

The Friedman statistic χ_F^2 quantifies the deviation of observed average ranks from the expected uniform distribution:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} \bar{R}_j^2 - \frac{k(k+1)^2}{4}\right] \quad (19)$$

Where N denote the number of independent problem instances (30 runs × 3 models = 90 total observations). k denote the number of algorithms compared ($k = 8$). $\bar{R}_j$ denote the average rank of the j-th algorithm.

For $k > 4$ and $N > 10$, the χ_F^2 statistic follows a chi-squared distribution with $k - 1$ degrees of freedom. However, to improve accuracy for moderate sample sizes, the Iman-Davenport correction transforms the statistic into an F-distributed variable:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} \quad (20)$$

This corrected statistic F_F follows an F-distribution with $(k-1)$ and $(k-1)(N-1)$ degrees of freedom. A significant p-value ($p < 0.05$) rejects the null hypothesis that all algorithms perform equivalently, warranting post-hoc pairwise analysis.

Application in This Study: The Friedman test was applied separately for each soil constitutive model and collectively across all models to identify globally superior algorithms. Average ranks $\bar{R}_j$provide an intuitive metric: lower values indicate consistently better performance. Algorithms with $\bar{R}_j < (k+1)/2 = 4.5$ outperform the median expectation.

6.2. Wilcoxon Signed-Rank Test

While the Friedman test identifies global differences, the Wilcoxon signed-rank test provides focused pairwise comparisons to determine whether a specific algorithm (e.g., the top-ranked CDO) significantly outperforms each competitor. This non-parametric test assesses whether the median difference between paired observations is zero, making no assumptions about distribution normality.

Procedure: For a given soil model, let $x_i$and $y_i$denote the final objective function values obtained by two algorithms across the i-th independent run ($i = 1, \ldots, n$, where $n = 30$). The difference for each pair is:

$$d_i = x_i - y_i \tag{21}$$

Differences equal to zero are excluded. The absolute differences $| d_i |$ are ranked, and signs are restored to obtain signed ranks. The test statistic W is the smaller of the sum of positive ranks (W^+) and negative ranks (W^-):

$$W = \min(W^+, W^-) \tag{22}$$

For $n \geq 10$, the distribution of W approximates a normal distribution, enabling computation of a z-score:

$$Z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{23}$$

The p-value is derived from the standard normal distribution. A two-tailed test with significance level $\alpha = 0.05$ is employed; $p < 0.05$ indicates a statistically significant difference between the two algorithms.

Wilcoxon tests were conducted between the best-performing algorithm (identified via Friedman ranking) and each of the seven competitors, for each soil model separately. This yields $3 \times 7 = 21$ pairwise comparisons. To control for family-wise error rate due to multiple testing, the Holm-Bonferroni correction was applied to adjust p-values sequentially. Results are reported as: (1) raw p-values, (2) adjusted p-values, and (3) significance indicators (†: $p < 0.05$, ‡: $p < 0.01$, §: $p < 0.001$).

6.3. Success Rate

Beyond statistical significance, practical utility demands that an algorithm reliably converges to solutions of acceptable quality. The Success Rate metric quantifies this reliability by measuring the proportion of independent runs in which the algorithm achieves an objective function value below a predefined threshold $E_{\text{threshold}}$. The Success Rate is formally defined as:

$$\text{Success Rate} = \frac{\#(E < E_{\text{threshold}})}{N_{\text{runs}}} \times 100\% \quad (24)$$

Where: $\#(E < E_{\text{threshold}})$= number of runs where the final error E(RMSE or penalized F) is below the threshold. N_{runs}= total number of independent runs ($N_{\text{runs}} = 30$). $E_{\text{threshold}}$= problem-specific acceptance criterion.

Threshold Selection: $E_{\text{threshold}}$was determined empirically for each soil model based on the known "true" parameters used in synthetic data generation:

- **Hyperbolic Model:** $E_{\text{threshold}} = 0.05 \times F(\mathbf{x}^*)$, where $\mathbf{x}^*$is the true parameter vector.
- **Duncan–Chang Model:** $E_{\text{threshold}} = 0.07 \times F(\mathbf{x}^*)$(higher tolerance due to increased nonlinearity).
- **Modified Cam-Clay Model:** $E_{\text{threshold}} = 0.10 \times F(\mathbf{x}^*)$(most challenging calibration).

These thresholds represent engineering-acceptable deviations: solutions within 5–10% of the optimal error are deemed practically useful for geotechnical design.

Interpretation: A Success Rate ≥ 90% indicates high reliability; 70–90% denotes moderate reliability; < 70% suggests the algorithm is prone to premature convergence or stagnation for that problem class.

6.4. Performance Profile (Dolan–Moré)

Performance profiles, introduced by Dolan and Moré [36], provide a cumulative distribution function-based visualization of algorithmic performance across a suite of test problems. Unlike aggregate statistics that mask problem-specific behavior, performance profiles reveal how frequently an algorithm is within a factor $\tau$of the best solver for any given problem instance.

Formulation: Let Pbe the set of $N_p$problem instances (here, $N_p = 90$: 3 models × 30 runs) and S be the set of k algorithms. For problem $p \in P$ and solver $s \in S$, define the performance ratio:

$$r_{p,s} = \frac{t_{p,s}}{\min_{s' \in S} t_{p,s'}} \tag{25}$$

Where $t_{p,s}$is the performance metric (final objective value F) for solver $s$on problem p. By definition, $r_{p,s} \geq 1$, and $r_{p,s} = 1$indicates that solver $s$achieved the best result for problem p. The performance profile $\rho_s(\tau)$ for solver $s$is the probability that its performance ratio is within a factor $\tau$of the best:

$$\rho_s(\tau) = \frac{1}{N_p} \mid [p \in P \mid r_{p,s} \leq \tau] \mid \tag{26}$$

Where $|\cdot|$ denotes set cardinality. $\rho_s(\tau)$is a non-decreasing, piecewise-constant function plotted for $\tau \geq 1$.

Interpretation Guidelines:

- **Top-left corner ($\tau = 1$):** $\rho_s(1)$ represents the probability that solver $s$is the *absolute best* for a randomly selected problem. Higher values indicate superior peak performance.
- **Right tail ($\tau \to \infty$):** $\lim_{\tau \to \infty} \rho_s(\tau)$equals the fraction of problems for which solver $s$successfully terminated. Values < 1 indicate failures to converge.
- **Curve dominance:** A solver whose profile lies above and to the left of another is considered more robust and efficient across the problem suite.

Performance profiles were constructed using the final penalized objective value Fas the performance metric $t_{p,s}$. Profiles were generated for: (1) all 90 problem instances combined, and (2) separately for each soil constitutive model. This dual perspective reveals both overall algorithmic robustness and model-specific strengths/weaknesses. The Dolan–Moré framework complements the Friedman and Wilcoxon tests by providing an intuitive, visual summary of relative algorithmic performance across the entire experimental landscape.

This multi-layered statistical approach ensures that conclusions about algorithmic superiority are both mathematically rigorous and practically meaningful. The Friedman test establishes global ranking significance, the Wilcoxon tests validate pairwise advantages, Success Rate quantifies engineering reliability, and Performance Profiles offer a holistic

visualization of robustness. Together, these methods provide an unambiguous, reproducible basis for recommending optimal metaheuristic strategies for nonlinear soil constitutive model calibration.

7. Result

It should be emphasized that the objective of this benchmark is not to establish a universally superior optimization algorithm. According to the No-Free-Lunch theorem, optimizer performance is inherently problem dependent. Therefore, the reported rankings should be interpreted only within the scope of nonlinear soil constitutive model calibration under the adopted experimental conditions.

The experimental results of the comprehensive benchmarking study are presented and analyzed in this section. All findings are based on 30 independent runs per algorithm per soil constitutive model, ensuring statistical robustness and reproducibility. The performance of eight metaheuristic algorithms—CDO, Chimp, GWO, HHO, COA, SMA, SOA, and PSO—was evaluated across three constitutive models of increasing complexity: Hyperbolic, Duncan–Chang, and Modified Cam-Clay.

The analysis is structured into four focused subsections. First, Section 7.1 presents descriptive performance statistics, including mean error, standard deviation, and Success Rate, providing an initial overview of algorithmic accuracy and reliability. Second, Section 7.2 reports the results of non-parametric statistical tests (Friedman ranking and Wilcoxon signed-rank tests) to determine whether observed performance differences are statistically significant. Third, Section 7.3 examines convergence behavior and algorithmic stability through mean convergence curves and Dolan–Moré performance profiles. Finally, Section 7.4 evaluates parameter estimation accuracy by comparing calibrated parameters against the known "true" values used in synthetic data generation.

All numerical results, including raw error values and statistical metrics, are summarized in Table 2 (Descriptive Statistics), Figure 5 (Friedman Ranks), and Figure 4 (Wilcoxon Results). Visual comparisons are provided in Figures 1–12, which illustrate error distributions, convergence trajectories, significance heatmaps, and parameter recovery accuracy. The complete dataset and MATLAB source codes are available in the supplementary material to ensure full reproducibility.

7.1. Descriptive Performance Comparison

The descriptive statistical performance metrics of eight metaheuristic algorithms across three soil constitutive models, aggregated over 30 independent runs, are summarized in Table 2. This subsection presents an objective analysis of the raw performance data, focusing on central tendency, dispersion, and reliability metrics, without premature inferential claims.

Table 2. Descriptive statistical performance metrics of eight metaheuristic algorithms across three soil constitutive models

Model	Algorithm	Mean Error	Std Error	Median Error	Min Error	Max Error	Success Rate
Hyperbolic	CDO	84739.70	22022.13	87587.70	41085.59	117650.02	10.00
	Chimp	73380.23	22281.21	64470.37	42138.30	132309.25	6.67
	GWO	78374.86	29264.27	68533.66	47792.55	173333.62	10.00
	HHO	299626.40	502662.57	123639.81	44832.69	2618893.23	3.33
	COA	85634.33	48445.56	78330.90	45829.14	322712.60	10.00
	SMA	73003.84	23331.17	76701.23	39381.61	115564.28	**16.67**
	SOA	113969.97	56938.12	98505.73	42674.35	289275.49	3.33
	PSO	342797.36	194626.74	325877.46	64903.10	936581.73	0.00
Duncan-Chang	CDO	331807.98	295793.91	209562.88	72028.24	1124939.75	**33.33**
	Chimp	848943.72	522812.82	676266.61	297380.26	2704909.47	0.00
	GWO	834370.16	677886.56	550589.17	96682.25	3088859.50	6.67
	HHO	599563.13	415814.38	480743.21	125193.87	1603813.16	6.67
	COA	612601.04	464671.18	425582.31	126330.61	1820351.03	10.00
	SMA	797908.62	659409.73	495182.41	121638.74	1984569.93	3.33
	SOA	1139503.54	639710.55	1004262.16	266070.08	2424501.99	0.00
	PSO	808409.61	339449.08	749077.74	351159.11	1716098.58	0.00
Cam-Clay	CDO	0.35	0.01	0.35	0.34	0.38	**83.33**
	Chimp	0.43	0.06	0.42	0.34	0.60	10.00
	GWO	0.43	0.06	0.40	0.36	0.56	6.67
	HHO	0.39	0.06	0.36	0.34	0.55	43.33
	COA	0.49	0.17	0.43	0.35	1.23	6.67
	SMA	0.37	0.02	0.36	0.34	0.42	56.67
	SOA	0.50	0.07	0.52	0.35	0.64	3.33
	PSO	0.46	0.07	0.45	0.36	0.66	0.00

The results presented in Table 2 demonstrate substantial variation in optimization performance across both constitutive models and optimization algorithms. Three general observations can be drawn from the descriptive statistics.

First, algorithm rankings vary considerably among the Hyperbolic, Duncan–Chang, and Modified Cam-Clay models, indicating that optimization performance is strongly influenced by the characteristics of the calibration problem. This observation suggests that conclusions drawn from a single constitutive model may not necessarily generalize to other calibration scenarios.

Second, several algorithms exhibit relatively large standard deviations compared with their corresponding mean errors, reflecting noticeable run-to-run variability. Such behavior may be associated with sensitivity to stochastic initialization, differences in exploration–exploitation balance, or the presence of multiple local optima within the search landscape.

Third, mean error alone does not provide a complete representation of algorithm behavior. The Success Rate metric complements the error statistics by quantifying the frequency with which an algorithm achieves solutions that satisfy predefined engineering acceptance criteria. Consequently, both accuracy and consistency should be considered when evaluating optimization performance.

Hyperbolic Model

For the Hyperbolic model, mean calibration errors range from 73,003.84 to 342,797.36. The lowest mean errors are observed for SMA, Chimp, GWO, CDO, and COA, all of which produce results within a relatively narrow range compared with the overall spread of the dataset. Moreover, the standard deviations associated with these methods indicate a substantial degree of overlap in their performance distributions, suggesting that differences among them should be interpreted cautiously.

In contrast, HHO, SOA, and PSO exhibit higher mean errors and greater variability, indicating less consistent convergence behavior for this particular calibration problem. Success Rates vary between 0% and 16.67%, with relatively low values observed for all algorithms. Considering the acceptance threshold of ($E_{\text{threshold}}$=50,000), these results suggest that obtaining engineering-acceptable solutions remains challenging for most optimization strategies under the selected acceptance threshold. Consequently, occasional successful runs may occur even for algorithms with comparatively large average errors, highlighting the importance of evaluating both central tendency and reliability metrics.

Duncan–Chang Model

The Duncan–Chang model appears to represent a more demanding calibration problem, as reflected by the generally higher error levels reported across all algorithms. A noticeable separation is observed between the lowest reported mean error and the remainder of the algorithm set; however, the large standard deviations indicate that this distinction should be interpreted alongside the observed run-to-run variability.

For some algorithms, the large gap between minimum and maximum errors implies that high-quality solutions can be obtained in certain runs, whereas other runs converge to less favorable regions of the search space. Such behavior is indicative of a challenging optimization problem characterized by multiple competing local optima.

Success Rates remain generally low across the Duncan–Chang model, with several algorithms failing to satisfy the acceptance criterion in any independent run. Given the threshold value of ($E_{\text{threshold}}$=160,000), these findings highlight the difficulty of consistently obtaining satisfactory calibration results for the Duncan–Chang model and demonstrate the sensitivity of performance assessments to the selected acceptance criterion.

Among the investigated constitutive models, the Duncan–Chang formulation exhibited the most challenging optimization landscape, as evidenced by the combination of large error magnitudes, high variability, and generally limited success in satisfying the acceptance criterion.

Modified Cam-Clay Model

For the Modified Cam-Clay model, calibration errors are reported on a different scale, ranging approximately from 0.35 to 0.50. Although the absolute differences among algorithms are smaller than those observed in the previous models, relative differences in performance remain evident.

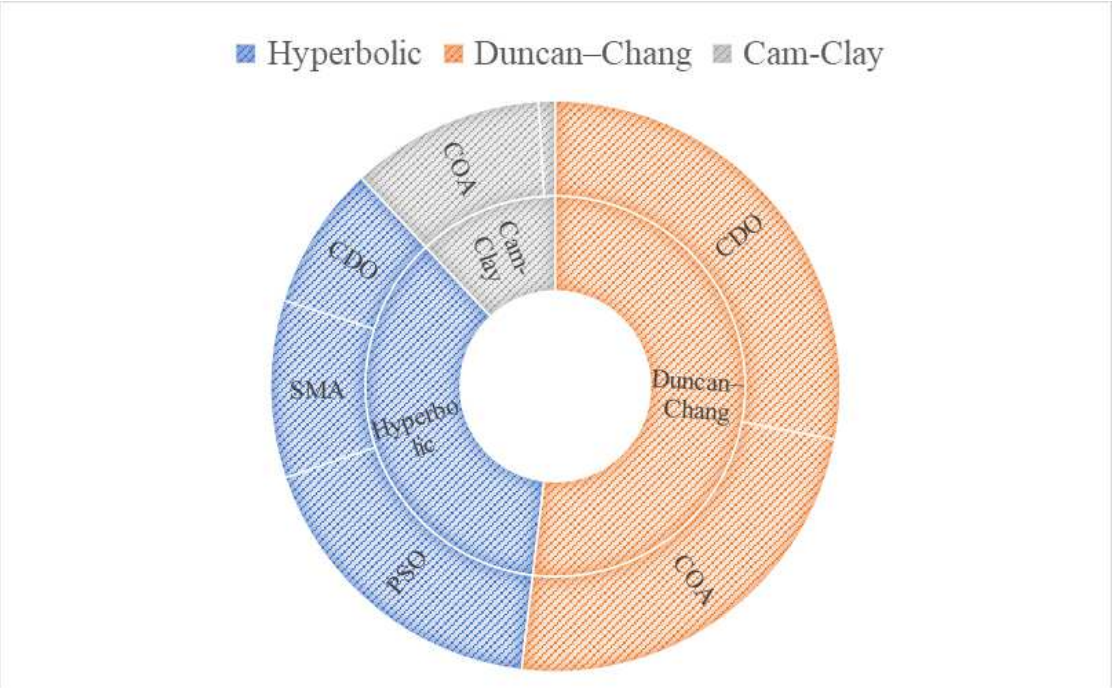
While several algorithms produced comparable mean errors, substantially different Success Rates were observed, indicating that similar average accuracy does not necessarily imply comparable reliability across repeated optimization runs.

These observations emphasize that evaluation based solely on mean error may obscure important information regarding robustness and repeatability. Algorithms with similar average performance can exhibit substantially different levels of reliability when variability measures are taken into account.

Variability and Reliability Considerations

To facilitate comparison of variability across models with different error scales, the coefficient of variation (CV = Std/Mean) is examined in Figure 1. Unlike standard deviation, the CV provides a normalized measure of dispersion and enables a more consistent assessment of algorithm stability across distinct calibration problems. Consequently, the combined interpretation of mean error, standard deviation, Success Rate, and coefficient of variation offers

644 a more comprehensive characterization of optimization performance than any individual metric
645 considered in isolation.



646 Figure 1. The coefficient of variation (CV = Std/Mean) of metaheuristic algorithms across three soil
647 constitutive models

648 CV values in Figure 1 indicate that algorithmic stability is problem-dependent. CDO
649 demonstrates high consistency for the Cam-Clay model but exhibits greater variability for the
650 Duncan–Chang model. Conversely, PSO shows consistently high variability across all models,
651 suggesting potential limitations in handling the multi-modal error landscapes characteristic of
652 soil parameter calibration.

653 Figure 2 presents a heatmap of normalized mean calibration errors for the eight
654 optimization algorithms across the three constitutive models. By normalizing the error values
655 within each model, the figure facilitates comparison of relative algorithm behavior while
656 preserving the actual mean error values reported in Table 2. Darker colors correspond to lower
657 normalized errors and therefore indicate more favorable calibration outcomes within each
658 model.

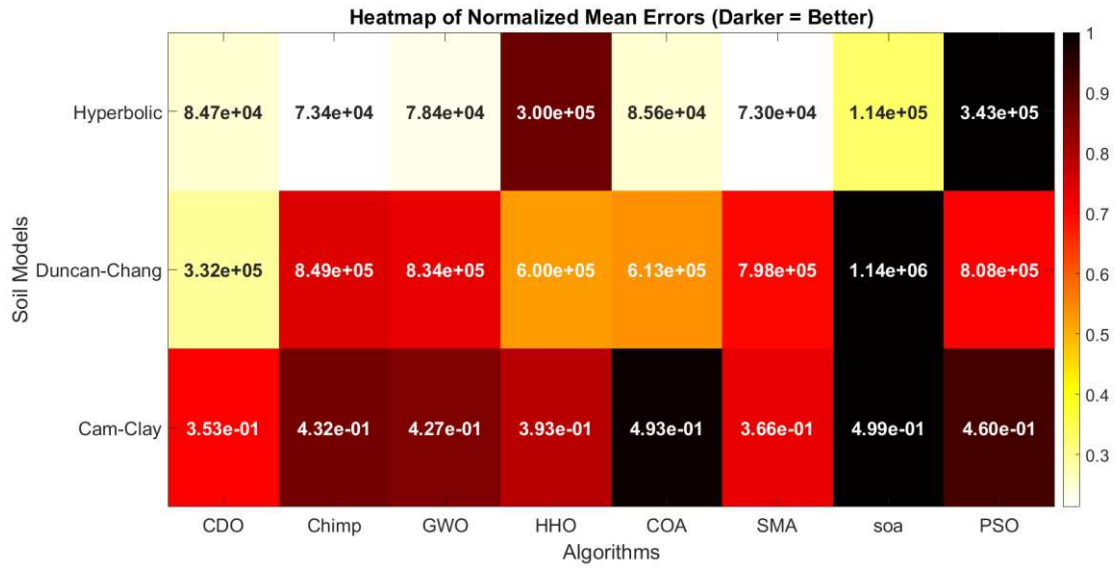


Figure 2. Heatmap of normalized mean calibration errors for eight metaheuristic algorithms across three soil constitutive models

The heatmap highlights the strong dependence of optimization performance on the underlying constitutive model. For the Hyperbolic model, several algorithms—including SMA, Chimp, GWO, CDO, and COA—produce relatively similar mean errors, resulting in a comparatively uniform color distribution among the best-performing methods. This observation suggests that multiple optimization strategies are capable of achieving comparable calibration quality for this problem.

For the Duncan–Chang model, a wider spread of normalized errors is observed across the algorithm set. The increased contrast in color intensity indicates that algorithm behavior becomes more differentiated as the complexity of the calibration problem increases. At the same time, the large standard deviations reported in Table 2 suggest that these differences should be interpreted together with the observed run-to-run variability.

In the Modified Cam–Clay model, the range of mean errors is considerably narrower than in the previous models. The resulting heatmap pattern is therefore more homogeneous, indicating that several algorithms achieve comparable average calibration accuracy. Nevertheless, differences in Success Rate and variability metrics reveal that similar mean errors do not necessarily imply equivalent reliability.

A comparison across rows indicates that the Duncan–Chang model is associated with the largest error magnitudes and variability levels, whereas the Modified Cam–Clay model exhibits the smallest overall error range. These observations suggest that the relative difficulty of the

calibration problem varies substantially among constitutive models and that algorithm rankings are influenced by the characteristics of the underlying search landscape.

Overall, Figure 2 reinforces the conclusion that optimization performance is problem-dependent. Rather than identifying a universally superior algorithm, the results demonstrate that different optimization strategies exhibit varying levels of effectiveness, robustness, and consistency depending on the constitutive model being calibrated.

7.2. Statistical Ranking and Significance

Figure 3 presents a comparative visualization of the calibration error distributions obtained by eight metaheuristic algorithms across the three soil constitutive models over 30 independent runs. The boxplots illustrate the statistical spread of the objective function values, including the median (central line), interquartile range (box boundaries), whiskers (data range), and outliers (red crosses). The red dashed horizontal line in each subplot indicates the success threshold ($E_{\text{threshold}}$), below which solutions are considered engineering-acceptable. Success Rates, displayed in the upper portion of each subplot, quantify the percentage of runs achieving this threshold, providing a complementary metric to the error distributions for assessing algorithmic reliability.

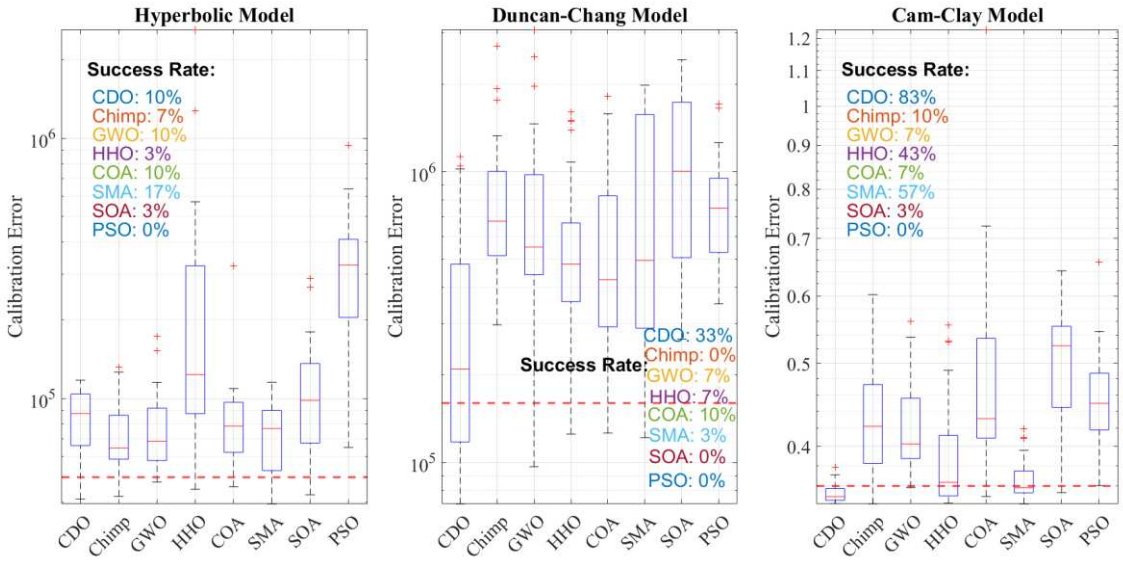


Figure 3. Comparison of calibration error distributions and Success Rates for eight metaheuristic algorithms across three soil constitutive models: (a) Hyperbolic Model, (b) Duncan–Chang Model, and (c) Modified Cam-Clay Model

Figure 3 presents a comprehensive visualization of the Wilcoxon signed-rank test results, comparing the performance of CDO against each of the seven competing algorithms across the

three soil constitutive models. The heatmap employs a color-coded scheme to indicate statistical significance: green cells denote cases where CDO significantly outperforms the competitor, red cells indicate superior performance by the competitor algorithm, and yellow cells represent cases where no statistically significant difference was detected ($p \geq 0.05$). The p-values displayed in each cell have been adjusted using the Holm–Bonferroni correction to control for family-wise error rate across multiple comparisons, ensuring the robustness of statistical inferences.

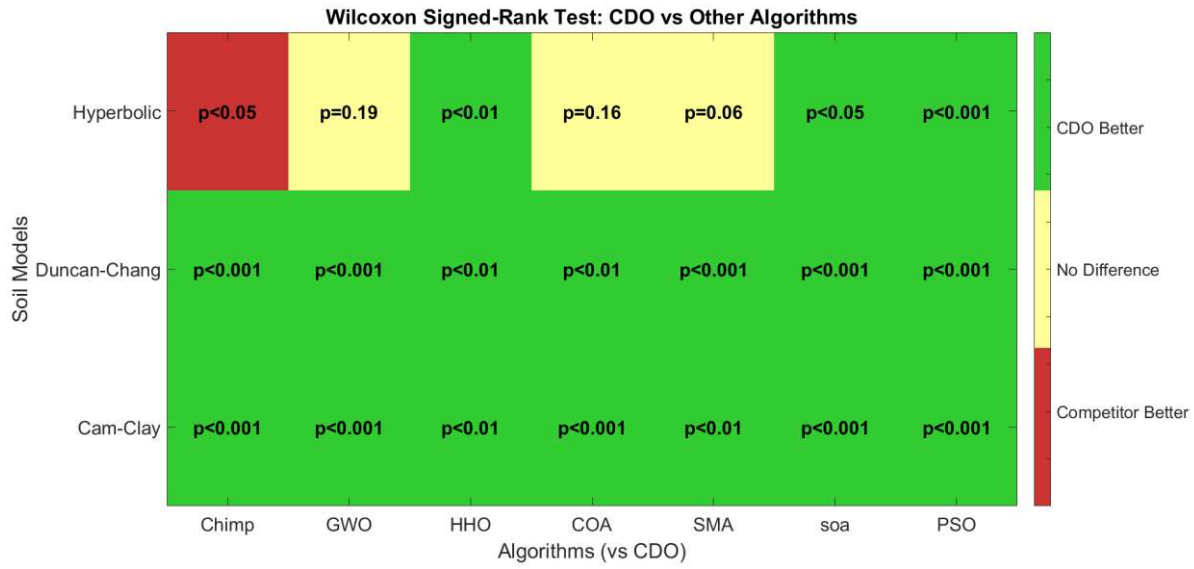


Figure 4. Heatmap visualization of Wilcoxon signed-rank test results comparing CDO against seven competing algorithms across three soil constitutive models

The Wilcoxon signed-rank test results shown in Figure 4 provide a pairwise statistical comparison of the optimization algorithms across the three constitutive models. The results indicate that the statistical relationships among algorithms are strongly dependent on the characteristics of the calibration problem.

For the Hyperbolic model, a mixed pattern of statistical significance is observed. Several algorithm pairs exhibit no statistically significant differences after Holm–Bonferroni correction, suggesting that multiple optimization strategies achieve comparable performance for this calibration task. At the same time, statistically significant differences are identified for certain comparisons, indicating that performance variations are not entirely attributable to stochastic variability.

For the Duncan–Chang model, a larger number of statistically significant pairwise differences are observed. This finding is consistent with the greater spread of error values and

higher variability reported in Table 2, suggesting that algorithm behavior becomes more differentiated as the complexity of the calibration problem increases.

A similar pattern is observed for the Modified Cam–Clay model, where statistically significant differences remain evident despite the relatively narrow range of mean errors. This observation highlights that algorithms exhibiting similar average performance may nevertheless differ in terms of reliability and consistency across repeated runs.

Overall, the Wilcoxon analysis confirms that algorithm rankings are not uniform across constitutive models and that performance differences should be interpreted within the context of the specific calibration problem under consideration..

Figure 5 presents the Friedman test results, providing a non-parametric statistical ranking of the eight metaheuristic algorithms across the three soil constitutive models. The left panel displays a heatmap of average ranks for each algorithm-model combination, where lower values indicate superior performance. The right panel shows the overall algorithm ranking based on average ranks aggregated across all models. The Friedman test evaluates whether the observed performance differences are statistically significant, with the Iman-Davenport correction applied to ensure accuracy for the given sample size (N = 90 observations: 30 runs × 3 models).

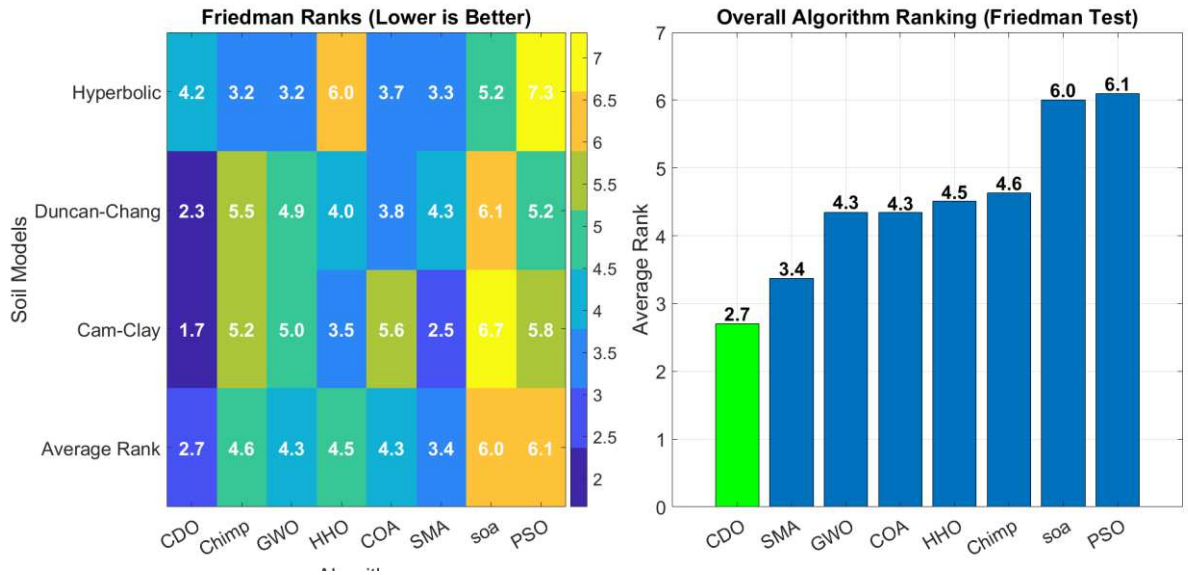


Figure 5. Friedman test results for algorithm ranking: (a) Heatmap of Friedman ranks across soil models, and (b) Overall algorithm ranking based on average ranks

Figure 5 summarizes the Friedman ranking analysis across the three constitutive models. The results indicate that the relative ranking of algorithms varies among calibration problems,

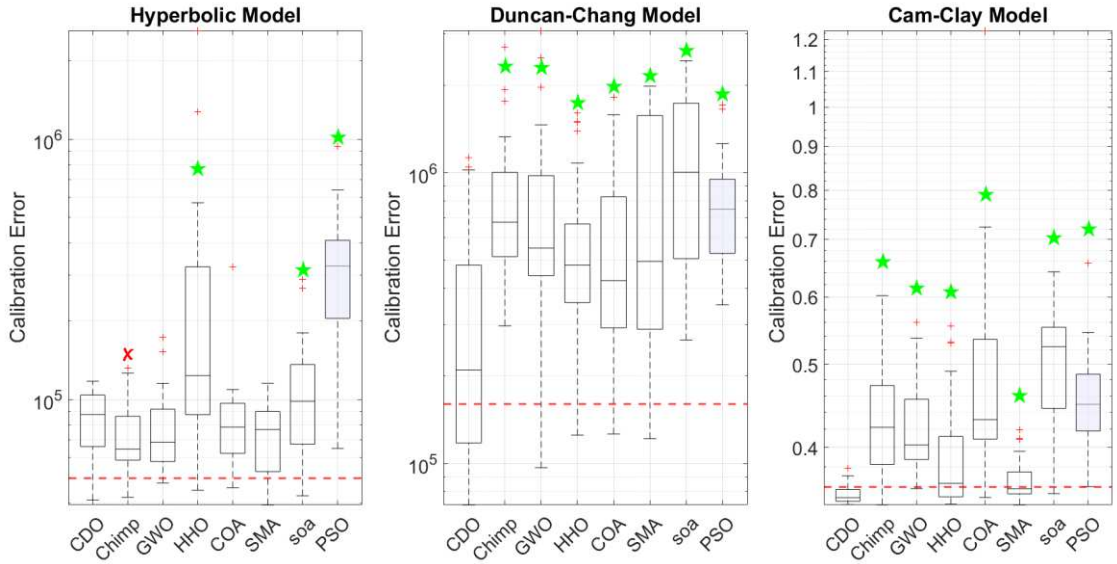
742 reinforcing the observation that optimization performance is problem-dependent.

743 Several algorithms consistently occupy favorable ranking positions across multiple
744 models, whereas others exhibit stronger dependence on the characteristics of individual
745 constitutive formulations. In particular, noticeable differences are observed between rankings
746 obtained for the Hyperbolic, Duncan–Chang, and Modified Cam–Clay models, indicating that
747 no single ranking structure is preserved across all calibration scenarios.

748 The overall Friedman ranking provides a useful summary of average performance across
749 the benchmark set; however, it should be interpreted together with the model-specific rankings
750 and variability measures. Algorithms with similar average ranks may exhibit substantially
751 different behavior in terms of robustness, convergence consistency, and sensitivity to problem
752 complexity.

753 The Friedman test, together with the Iman–Davenport correction, rejects the null
754 hypothesis of equivalent algorithm performance ($p < 0.05$), indicating that statistically
755 significant differences exist among the investigated optimization methods. Nevertheless, the
756 ranking results should be viewed as reflecting performance within the selected benchmark
757 problems rather than evidence of universal superiority for any particular algorithm.

758 Figure 6 illustrates the distribution of calibration errors obtained by eight metaheuristic
759 algorithms across the three soil constitutive models over 30 independent runs. The box plots
760 display the median, interquartile range, whiskers, and outliers (marked with red crosses and
761 green stars) of the objective function values. The red dashed horizontal line in each subplot
762 indicates the success threshold ($E_{\text{threshold}}$), below which solutions are considered engineering-
763 acceptable. This visualization provides insight into both the central tendency and variability of
764 algorithmic performance, highlighting differences in reliability and convergence stability.



765 Figure 6. Comparison of calibration error distributions for eight metaheuristic algorithms across three
766 soil constitutive models: (a) Hyperbolic Model, (b) Duncan–Chang Model, and (c) Modified Cam-
767 Clay Model

768 Figure 6 provides additional insight into the distribution of calibration errors by
769 illustrating the median, dispersion, and occurrence of outliers for each algorithm.

770 For the Hyperbolic model, several algorithms exhibit comparable median error levels,
771 although differences in dispersion are evident. Some methods display wider interquartile ranges
772 and a larger number of extreme observations, indicating greater sensitivity to initialization and
773 search-path variability.

774 For the Duncan–Chang model, the distributions are generally wider and more
775 asymmetric, reflecting the increased difficulty of the calibration problem. The presence of large
776 outliers for several algorithms suggests that convergence to suboptimal regions of the search
777 space remains a significant challenge.

778 For the Modified Cam–Clay model, the distributions are considerably more compact,
779 indicating improved consistency across repeated runs. Nevertheless, noticeable differences in
780 spread and outlier occurrence remain visible among algorithms, demonstrating that similar mean
781 errors do not necessarily correspond to equivalent levels of reliability.

782 Taken together, the boxplot analysis complements the descriptive statistics and statistical
783 tests by highlighting the importance of considering variability, robustness, and distributional
784 characteristics in addition to average calibration accuracy..

7.3. Convergence Behavior and Stability

Figure 7 illustrates the convergence behavior of the eight metaheuristic algorithms across the three soil constitutive models over 30 iterations. The plots display the mean calibration error (on a logarithmic scale) as a function of iteration number, with shaded regions representing the standard deviation across 30 independent runs. This visualization provides insights into both the convergence speed and stability of each algorithm, revealing how quickly they approach optimal solutions and how consistently they perform across different stochastic initializations. The convergence curves enable identification of premature convergence, stagnation behavior, and the ability of algorithms to escape local minima during the optimization process.

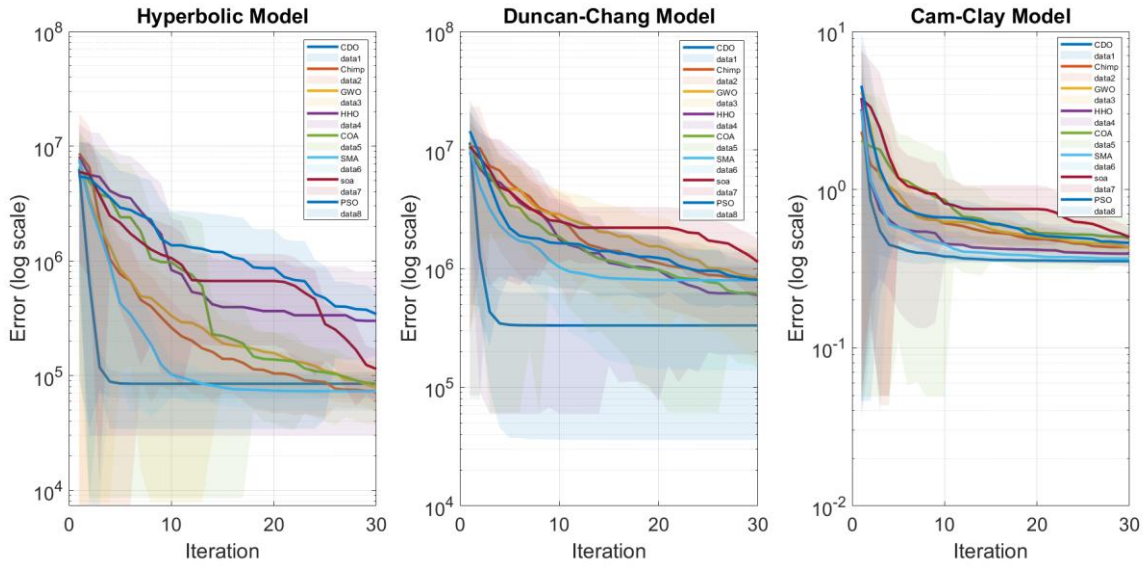


Figure 7. Convergence curves showing mean calibration error (± standard deviation) versus iteration for eight metaheuristic algorithms across three soil constitutive models: (a) Hyperbolic Model, (b) Duncan–Chang Model, and (c) Modified Cam-Clay Model

Figure 7 illustrates the convergence histories of the investigated optimization algorithms across the three constitutive models. In all cases, the largest reduction in objective function value occurs during the early iterations, followed by a gradual transition toward a stable region.

For the Hyperbolic model, several algorithms exhibit rapid initial error reduction, although differences in convergence rate and variability remain visible. Some methods converge to relatively stable solutions within the first iterations, whereas others continue to display fluctuations throughout the optimization process.

For the Duncan–Chang model, the convergence trajectories are generally more dispersed, indicating a more challenging optimization landscape. Wider confidence bands observed for

several algorithms suggest increased sensitivity to initialization and stronger run-to-run variability. In addition, some convergence curves exhibit plateau behavior, indicating reduced improvement during later iterations.

For the Modified Cam–Clay model, convergence is generally faster and more stable across all algorithms. The narrower confidence intervals suggest a lower level of variability compared with the Duncan–Chang model, which is consistent with the smaller error ranges reported in Table 2.

Overall, the convergence curves indicate that both convergence speed and convergence stability are influenced by the complexity of the constitutive model. The results further suggest that algorithm performance should be evaluated using both final solution quality and convergence characteristics rather than either criterion alone.

Figure 8 presents the Dolan–Moré performance profiles for the eight metaheuristic algorithms across the three soil constitutive models. Performance profiles provide a cumulative distribution function-based visualization that captures the probability of an algorithm achieving a solution within a factor $\tau$of the best solution found by any algorithm. The horizontal axis represents the performance ratio τon a logarithmic scale, while the vertical axis shows the cumulative probability $P(r_p \leq \tau)$. This analysis complements the descriptive statistics and pairwise significance tests by offering a holistic view of algorithmic robustness and efficiency across the entire problem suite, revealing how frequently each algorithm performs competitively relative to the best performer.

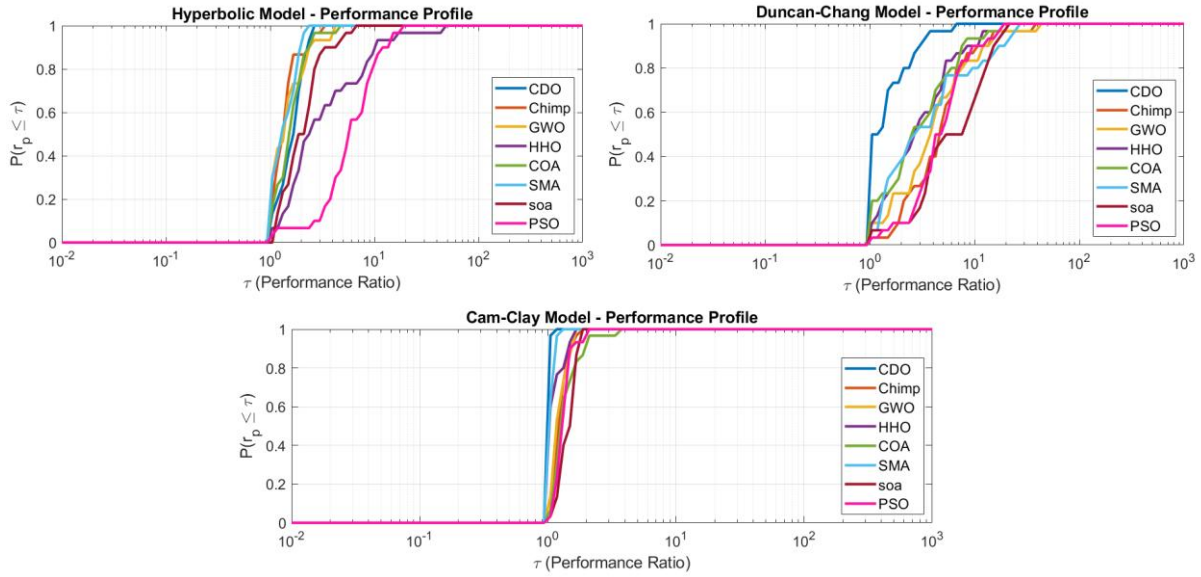


Figure 8. Dolan–Moré performance profiles for eight metaheuristic algorithms across three soil constitutive models: (a) Hyperbolic Model, (b) Duncan–Chang Model, and (c) Modified Cam-Clay Model

Figure 8 presents the Dolan–Moré performance profiles, which provide a probabilistic assessment of algorithm competitiveness across the benchmark problems.

For the Hyperbolic model, several algorithms exhibit similar performance profiles, indicating comparable probabilities of achieving near-best solutions. The substantial overlap among curves suggests that multiple optimization strategies remain competitive for this calibration problem.

For the Duncan–Chang model, greater separation among the performance profiles is observed. This finding is consistent with the descriptive statistics and statistical tests, which indicate stronger differentiation among algorithms as the complexity of the calibration problem increases.

For the Modified Cam–Clay model, the performance profiles are generally closer together, reflecting the relatively small differences in mean error among algorithms. Nevertheless, noticeable differences remain in the probability of achieving near-optimal solutions.

Taken together, the performance profiles demonstrate that algorithm competitiveness varies across constitutive models and that conclusions regarding optimization effectiveness are strongly dependent on the characteristics of the underlying calibration problem.

Figure 9 presents a multi-criteria performance comparison using radar charts to visualize the trade-offs between different performance metrics for each algorithm across the three soil constitutive models. The radar charts simultaneously display four key performance indicators: Mean Error (lower is better), Standard Deviation (lower is better), Minimum Error achieved (lower is better), and Success Rate (higher is better). This visualization technique enables immediate identification of algorithmic strengths and weaknesses, revealing which algorithms achieve balanced performance across all criteria versus those that excel in specific metrics while underperforming in others. The normalized scale (0 to 1) facilitates direct comparison across metrics with different units and magnitudes, where values closer to the center represent better performance for error metrics, and values closer to the periphery indicate superior Success Rate.

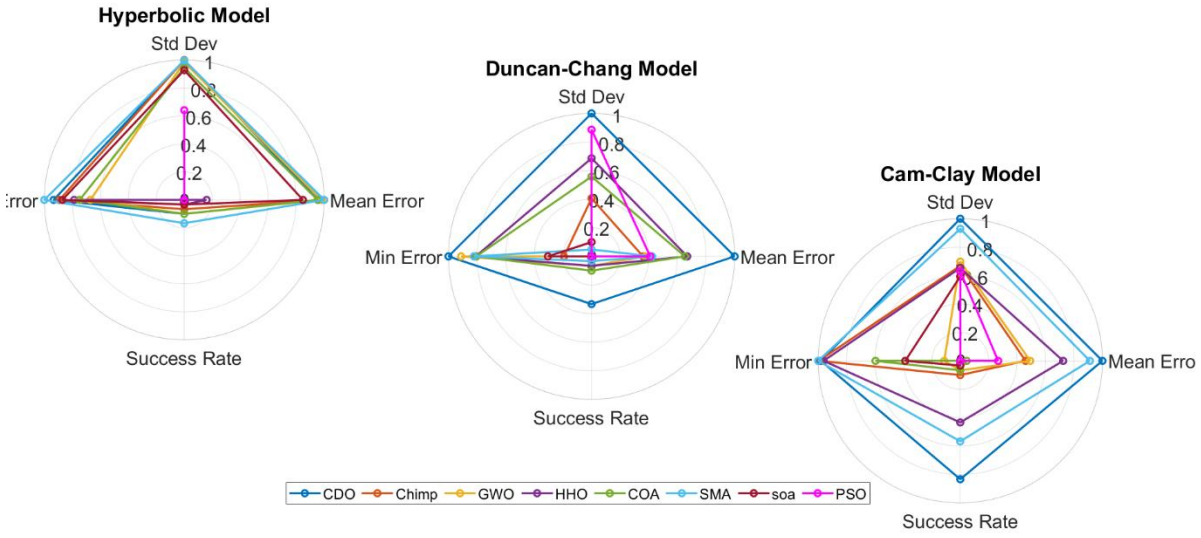


Figure 9. Multi-criteria performance comparison using radar charts for eight metaheuristic algorithms across three soil constitutive models: (a) Hyperbolic Model, (b) Duncan–Chang Model, and (c) Modified Cam-Clay Model

The radar chart visualization in Figure 9 provides a multi-criteria perspective on algorithm performance by simultaneously considering mean error, variability, best-case performance, and Success Rate. Unlike individual statistical indicators, the radar charts facilitate a holistic assessment of the trade-offs among different performance metrics. Three distinct patterns can be observed across the investigated constitutive models.

Hyperbolic Model: For the Hyperbolic model, the radar polygons exhibit substantial overlap, indicating relatively similar performance among several algorithms. The compact clustering suggests that the two-parameter calibration problem can be effectively addressed by

a range of optimization strategies. Although differences in mean error and Success Rate are present, they remain relatively modest compared with the variability observed in more complex models. Some algorithms exhibit slightly lower dispersion and higher Success Rates, whereas others display larger error ranges and greater variability. Overall, the radar chart indicates that the Hyperbolic model provides a comparatively balanced optimization landscape in which multiple algorithms are capable of producing competitive calibration results.

Duncan–Chang Model: The Duncan–Chang model exhibits noticeably greater separation among the radar polygons, reflecting increased differentiation in algorithm behavior. Differences are apparent across all performance metrics, including mean error, standard deviation, minimum error, and Success Rate. Several algorithms achieve lower average errors and higher frequencies of acceptable solutions, while others exhibit larger variability and reduced reliability. These observations suggest that optimization outcomes become more sensitive to the characteristics of the search strategy as the complexity of the constitutive model increases. The broader spread of performance indicators is consistent with the higher error magnitudes and variability reported in the descriptive statistics and statistical analyses.

Modified Cam–Clay Model: For the Modified Cam–Clay model, the radar charts reveal comparatively small differences in average error metrics among several algorithms, while more pronounced distinctions emerge in terms of variability and Success Rate. Certain algorithms achieve both low dispersion and high frequencies of acceptable solutions, indicating stable convergence behavior across repeated runs. In contrast, other methods exhibit larger variability and lower Success Rates despite achieving comparable average errors. This observation highlights the importance of considering reliability metrics in addition to central tendency measures when evaluating optimization performance. The results further suggest that similar mean errors do not necessarily correspond to equivalent levels of robustness or repeatability.

Overall, the radar chart analysis reinforces the conclusion that optimization performance is strongly dependent on the characteristics of the constitutive model. While some algorithms exhibit relatively stable behavior across multiple calibration problems, others show greater sensitivity to model complexity and search-space characteristics. Consequently, a comprehensive evaluation of optimization methods should consider multiple performance criteria simultaneously rather than relying exclusively on a single metric or ranking measure.

Cross-Model Synthesis: The radar charts provide a multi-criteria perspective on optimization performance by simultaneously considering accuracy, variability, best-case

899 performance, and Success Rate.

900 For the Hyperbolic model, the radar polygons exhibit substantial overlap, indicating
901 relatively similar performance among several algorithms. This observation is consistent with
902 the descriptive statistics, which showed limited separation among the best-performing methods.

903 For the Duncan–Chang model, greater differentiation among algorithms becomes
904 apparent. Differences in both error-related metrics and Success Rate suggest that algorithm
905 behavior is more strongly influenced by the characteristics of the optimization landscape for this
906 constitutive model.

907 For the Modified Cam–Clay model, several algorithms achieve comparable error levels,
908 while more pronounced differences emerge in terms of variability and Success Rate. These
909 results indicate that average calibration accuracy alone may not fully reflect algorithm
910 reliability.

911 A comparison across all three models reveals that algorithm rankings are not preserved
912 across constitutive formulations. Some algorithms demonstrate relatively stable behavior across
913 multiple models, whereas others exhibit stronger sensitivity to problem characteristics. This
914 observation reinforces the importance of evaluating optimization methods using a diverse set of
915 benchmark problems rather than relying on a single constitutive model.

916 From a practical perspective, the results suggest that optimization method selection
917 should consider not only solution accuracy but also convergence stability, robustness, and
918 consistency under repeated runs. Consequently, no single performance metric is sufficient to
919 characterize algorithm behavior across the full spectrum of soil constitutive model calibration
920 problems.

921 **Practical Implications for Geotechnical Practice:** The radar chart analysis provides
922 several observations that may assist researchers and practitioners when selecting optimization
923 methods for constitutive model calibration.

924 1. **For relatively simple calibration problems (e.g., Hyperbolic Model):** The
925 performance differences among algorithms are generally limited, and multiple
926 optimization strategies are capable of producing comparable calibration results. Under
927 such conditions, considerations such as computational cost, implementation simplicity,
928 and ease of parameter tuning may play a more important role in method selection.

2. **For moderately complex calibration problems (e.g., Duncan–Chang Model):** Performance differences become more pronounced, with noticeable variations in error levels, variability, and Success Rates among algorithms. These findings suggest that algorithm selection may have a greater influence on calibration outcomes as model complexity increases.

3. **For more demanding constitutive calibration problems (e.g., Modified Cam–Clay Model):** Reliability-related metrics, including variability and Success Rate, become increasingly important in addition to average calibration accuracy. The results indicate that algorithms with similar mean errors may nevertheless exhibit substantially different levels of consistency across repeated runs.

The multi-criteria visualization further highlights that optimization performance is strongly dependent on the characteristics of the constitutive model and the associated search landscape. Consequently, algorithm selection should be guided by the specific requirements of the calibration problem, including desired accuracy, robustness, and computational efficiency. Rather than identifying a universally optimal optimizer, the present benchmark provides a comparative framework for evaluating algorithm behavior across a range of representative soil constitutive model calibration scenarios.

7.4. Parameter Estimation Accuracy

The parameter estimation accuracy plots presented in Figure 10-12 provide additional insight into algorithm behavior by examining not only the objective function values but also the distributions of the calibrated constitutive parameters. The box plots enable assessment of both estimation bias, reflected by the proximity of the median to the reference value, and estimation stability, reflected by the spread of the distributions.

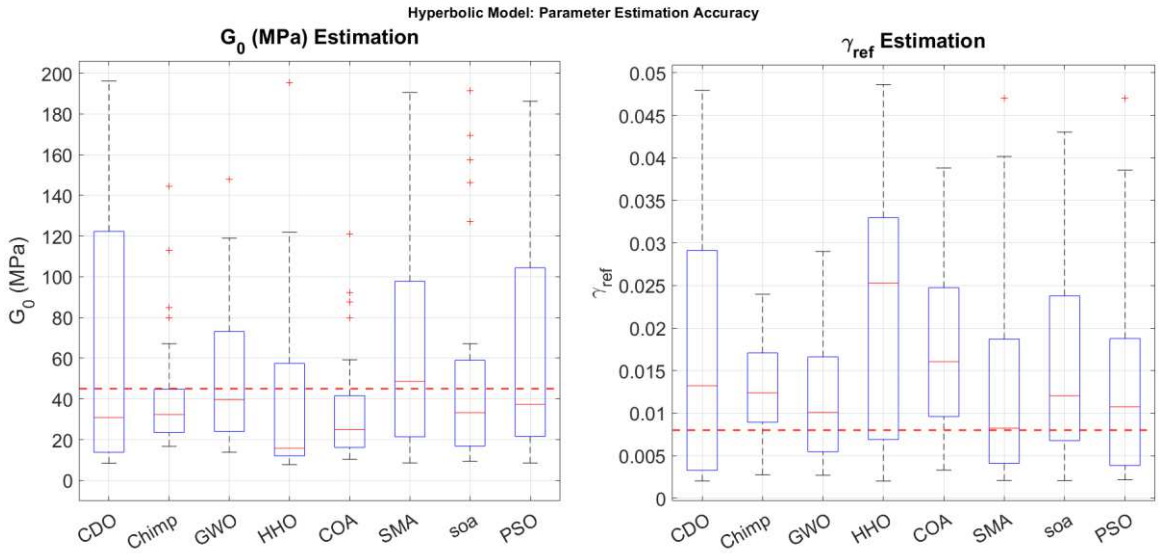


Figure 10. Parameter estimation accuracy for Hyperbolic Model: (a) Initial shear modulus (G_0), and (b) Reference shear strain (γ_{ref})

Hyperbolic Model

For the Hyperbolic model, the estimated values of both the initial shear modulus G_0 and the reference shear strain γ_{ref} generally cluster around the corresponding reference values. However, noticeable differences in dispersion are observed among algorithms. Some methods exhibit relatively compact interquartile ranges and limited outlier occurrence, whereas others display broader distributions and greater variability across independent runs.

The variability observed for $\gamma_{ref}$is generally larger than that observed for G_0, suggesting that parameters controlling the nonlinear curvature of the stress–strain response may be more sensitive to optimization uncertainty than stiffness-related parameters. Despite these differences, most algorithms are capable of producing parameter estimates within a reasonable range around the reference values.

Duncan–Chang Model

For the Duncan–Chang model, the distributions of the estimated parameters exhibit greater variability compared with the Hyperbolic model. The stiffness modulus number K shows noticeable differences in both dispersion and outlier occurrence among algorithms, indicating varying levels of estimation consistency.

The hardening exponent n generally exhibits wider distributions than the other

parameters, suggesting a greater degree of uncertainty in parameter identification. This observation is consistent with the strong influence of nonlinear stress dependency on the calibration process.

For the failure ratio R_f, the overall spread of the estimated values is comparatively smaller, although isolated extreme values remain visible for some algorithms. These results indicate that different model parameters may exhibit substantially different levels of identifiability even within the same constitutive formulation.

Modified Cam–Clay Model

For the Modified Cam–Clay model, differences among algorithms become more evident for certain parameters. The initial void ratio e_0 is generally estimated with relatively limited dispersion across most algorithms, while larger differences emerge for the compression index λ and reference pressure p_0.

Among the investigated parameters, p_0 exhibits the largest spread and the greatest occurrence of outliers, indicating a higher degree of estimation uncertainty. This behavior suggests that parameters associated with yielding and plastic response may be more difficult to identify consistently than parameters governing the initial state of the material.

Overall, the Modified Cam–Clay model exhibits greater variability in parameter estimation than the Hyperbolic model, reflecting the increased complexity of elastoplastic calibration.

Stability and Reliability Considerations

The parameter estimation results demonstrate that calibration quality cannot be assessed solely through objective function values. Algorithms producing similar mean errors may nevertheless exhibit different levels of parameter variability and estimation consistency.

Across all three constitutive models, differences in dispersion, outlier frequency, and median location indicate that both optimization performance and parameter identifiability influence the final calibration results. Consequently, reliable constitutive model calibration requires consideration of parameter distributions in addition to aggregate error metrics.

.

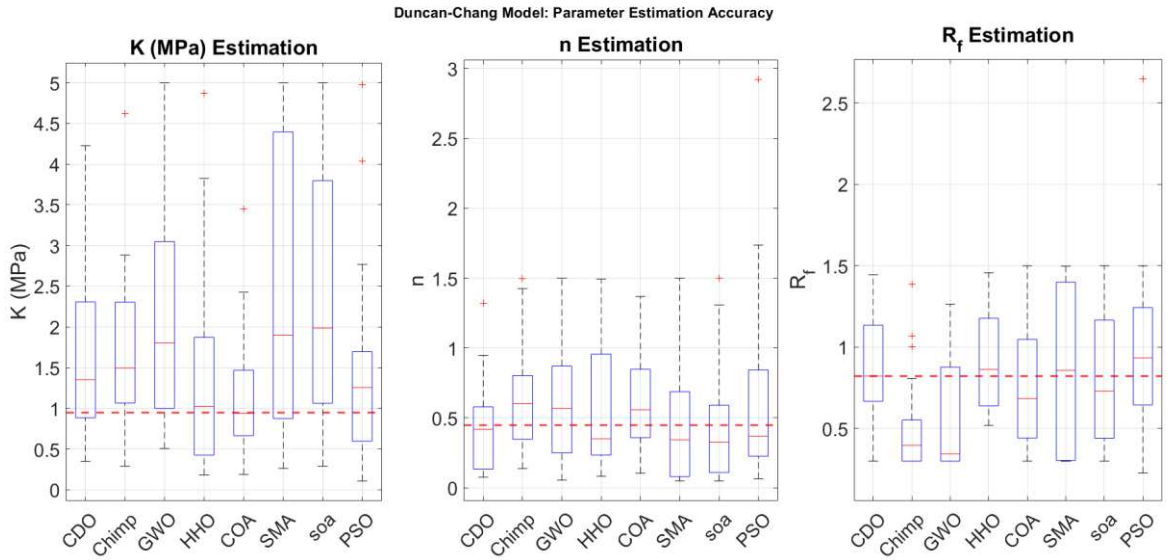


Figure 11. Parameter estimation accuracy for Duncan–Chang Model: (a) Stiffness modulus number (K), (b) Hardening exponent (n), and (c) Failure ratio (R_f)

Practical Implications for Geotechnical Engineering

The observed estimation patterns highlight several considerations relevant to practical geotechnical applications. First, parameter variability may influence the reliability of constitutive model predictions, particularly when calibrated parameters are subsequently used in numerical analyses. Second, certain parameters exhibit greater estimation uncertainty than others, indicating that parameter identifiability should be considered during model calibration and interpretation. Finally, the results emphasize the importance of multi-run statistical evaluation when comparing optimization algorithms, as single-run results may not adequately represent calibration robustness or repeatability.

Overall, the findings confirm that inverse parameter identification for nonlinear constitutive models remains a challenging optimization problem, particularly for formulations involving multiple interacting parameters and nonlinear constitutive relationships.

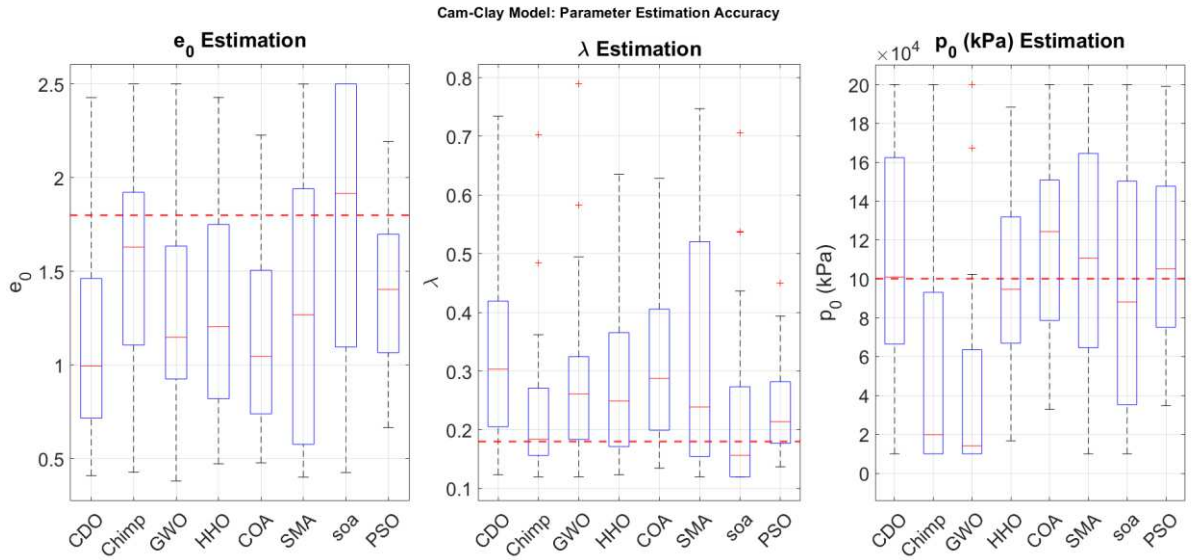


Figure 12. Parameter estimation accuracy for Modified Cam-Clay Model: (a) Initial void ratio (e_0), (b) Compression index (λ), and (c) Reference pressure (p_0)

To facilitate the reproducibility of the benchmarking process, the pseudocode of the calibration framework, including the objective function evaluation and the optimization loop, is provided as Figure A in Appendix A.

8. Discussion and limitations

The benchmarking results presented in this study indicate that optimization performance is strongly influenced by both the structural complexity of the constitutive model and the identifiability characteristics of the calibrated parameters. The observed differences among algorithms suggest that optimization effectiveness is problem-dependent and may vary considerably across constitutive formulations. While several algorithms achieve comparable performance for relatively simple calibration tasks, larger differences emerge as model complexity and parameter interactions increase.

A notable observation is the increasing variability associated with the calibration of more complex constitutive models. In particular, parameters governing nonlinear and elastoplastic behavior generally exhibit greater uncertainty than parameters associated with the initial stiffness or state variables. This finding highlights the importance of evaluating parameter identifiability alongside conventional goodness-of-fit measures when assessing calibration quality.

Another important result is the apparent decoupling between objective-function

minimization and parameter recovery accuracy. As demonstrated by the parameter estimation analyses, low calibration errors do not necessarily correspond to the most accurate reconstruction of the reference parameter values. Multiple parameter combinations may produce similar stress–strain responses while differing substantially in their individual parameter estimates. This phenomenon, commonly referred to as *equifinality*, suggests that calibration quality should not be evaluated solely on the basis of objective-function values. Incorporating parameter uncertainty analysis, physical constraints, or independent validation datasets may provide a more comprehensive assessment of calibration reliability.

Several limitations of the present study should be acknowledged. First, the benchmark was conducted using synthetic datasets with controlled noise levels (SNR = 20 dB). Although this approach facilitates a consistent comparison of optimization algorithms, it does not fully represent the measurement errors, boundary effects, sample disturbance, and other uncertainties commonly encountered in laboratory testing programs.

Second, the Modified Cam–Clay investigation was restricted to isotropic compression behavior. A more comprehensive calibration framework incorporating additional loading paths and constitutive parameters, such as swelling and critical-state parameters, would likely increase both parameter interaction and optimization complexity. Consequently, the findings reported here should be interpreted within the scope of the selected calibration scenarios.

Third, the benchmark considered a fixed computational budget and a predefined set of algorithmic parameters. Alternative population sizes, stopping criteria, parameter settings, or computational budgets may influence algorithm rankings and convergence behavior. Therefore, the reported results should not be interpreted as universally applicable to all optimization settings.

Finally, the penalized objective function introduced in this study represents a useful mechanism for incorporating engineering constraints into the calibration process. However, its effectiveness was evaluated only for the constitutive models and datasets considered herein. Further investigation is required to assess its applicability to other soil models, experimental datasets, and calibration frameworks. Future studies may also compare algorithm performance using conventional RMSE-based formulations and alternative objective functions to better understand the influence of penalty terms on optimization outcomes and algorithm rankings.

Overall, the results demonstrate that constitutive model calibration remains a challenging

inverse problem characterized by parameter interaction, uncertainty, and potential non-uniqueness of solutions. Consequently, robust evaluation of optimization algorithms should consider not only convergence performance but also parameter identifiability, solution stability, and repeatability across independent runs.

9. Conclusions

This study developed a statistically grounded benchmarking framework for evaluating metaheuristic optimization algorithms in the calibration of nonlinear soil constitutive models. Eight optimization algorithms were assessed across three constitutive formulations of varying complexity, and their performance was examined using descriptive statistics, convergence analysis, parameter estimation accuracy, and non-parametric statistical tests. Based on the results obtained under the conditions investigated, the following conclusions can be drawn:

Model-Dependent Performance: Optimization performance is strongly influenced by the complexity of the constitutive model and the characteristics of the calibration problem. While several algorithms exhibit comparable performance for the Hyperbolic model, greater differentiation emerges for the Duncan–Chang and Modified Cam–Clay models, indicating that algorithm behavior is highly problem-dependent.

Performance Variability Across Algorithms: Differences among algorithms become more pronounced as parameter interactions and nonlinearities increase. The results demonstrate that algorithm rankings are not preserved across constitutive formulations, suggesting that conclusions derived from a single benchmark problem may not be readily generalized to other calibration scenarios.

Statistical Evidence of Performance Differences: The non-parametric statistical analyses indicate that several observed differences among algorithms are statistically significant ($p < 0.05$). However, the magnitude and consistency of these differences vary across constitutive models, highlighting the importance of considering both statistical significance and practical relevance when interpreting optimization results.

Parameter Identifiability and Equifinality: A key finding of this study is the distinction between calibration accuracy and parameter recovery accuracy. Low objective-function values do not necessarily correspond to accurate identification of the reference constitutive parameters. The presence of equifinality demonstrates that multiple parameter combinations may reproduce similar stress–strain responses while differing substantially in their estimated

parameter values. Consequently, constitutive model calibration should consider parameter identifiability in addition to goodness-of-fit measures.

Importance of Multi-Criteria Evaluation: The combined analysis of error statistics, convergence behavior, Success Rate, parameter distributions, and statistical tests demonstrates that no single performance metric is sufficient to characterize optimization effectiveness. Comprehensive evaluation frameworks are therefore necessary to capture both accuracy and robustness aspects of algorithm behavior.

Implications for Geotechnical Calibration Problems: The results suggest that optimization method selection should consider the specific characteristics of the constitutive model, including parameter sensitivity, model complexity, and calibration objectives. For more complex constitutive formulations, robustness and repeatability become increasingly important alongside convergence quality.

Future Research Directions: Future studies should extend the proposed benchmarking framework to experimental laboratory datasets, investigate additional constitutive formulations and loading conditions, evaluate alternative objective functions, and explore multi-objective calibration strategies. Further research may also examine algorithm performance in higher-dimensional calibration problems involving coupled hydro-mechanical and elastoplastic processes.

Overall, the findings confirm that constitutive model calibration remains a challenging inverse problem characterized by parameter interaction, uncertainty, and non-uniqueness of solutions. The proposed benchmarking framework provides a structured basis for evaluating optimization algorithms and may support future studies aimed at improving the reliability and transparency of geotechnical parameter calibration.

10. Declaration

Conflict of Interest: The Cloud Drift Optimization (CDO) algorithm, which is among the algorithms evaluated in this study, was recently proposed by the author (Alibabaei Shahraki, 2025). To ensure objectivity, all algorithms were implemented with their default parameter settings as suggested in their original publications, and no additional tuning was performed that could favor CDO. The experimental design, including model selection, objective function formulation, and statistical analysis, was established prior to conducting the benchmarking study.

Funding: This research did not receive a specific grant from public, commercial, or not-for-profit funding agencies.

Data Availability: The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Artificial Intelligence Tools (AI): Artificial intelligence tools were used in this study solely for language polishing, grammar checking, and improving textual clarity. All authors are fully responsible for the intellectual content, research design, data analysis, interpretation of results, and formulation of scientific conclusions. AI tools were not involved in the generation of hypotheses, data, original ideas, or scientific reasoning.

Clinical Trial Number: Not applicable.

Consent to Participate: Not applicable.

Consent to Publish: Not applicable.

Ethics Declaration: Not applicable.

11. Reference

[1] Zeng, Z., L. Udpa, and S.S. Udpa, *Finite-element model for simulation of ferrite-core eddy-current probe.* IEEE Transactions on Magnetics, 2009. **46**(3): p. 905–909.

[2] Zona, A. and G. Ranzi, *Finite element models for nonlinear analysis of steel–concrete composite beams with partial interaction in combined bending and shear.* Finite Elements in Analysis and Design, 2011. **47**(2): p. 98–118.

[3] Chik, Z. and T. Islam. *Near surface soil characterizations through soil apparent resistivity: A case study*. in *2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*. 2013. IEEE.

[4] Southey, R.D., et al., *Using fall-of-potential measurements to improve deep soil resistivity estimates.* IEEE Transactions on Industry Applications, 2015. **51**(6): p. 5023–5029.

[5] Gonos, I.F. and I.A. Stathopulos, *Estimation of multilayer soil parameters using genetic algorithms.* IEEE Transactions on Power delivery, 2005. **20**(1): p. 100–106.

[6] Del Alamo, J., *A second order gradient technique for an improved estimation of soil parameters in a two-layer earth.* IEEE Transactions on Power Delivery, 2002. **6**(3): p. 1166–1170.

[7] Kang, M.-J., et al., *A nonlinear regression based approach for multilayer soil parameter estimation.* International Journal of Control and Automation, 2014. **7**(2): p. 65–74.

[8] Mohamed, A.W., H.Z. Sabry, and T. Abd-Elaziz, *Real parameter optimization by an effective differential evolution algorithm.* Egyptian Informatics Journal, 2013. **14**(1): p. 37–53.

[9] Dawalibi, F., J. Ma, and R. Southey, *Behaviour of grounding systems in multilayer soils: a parametric analysis.* IEEE Transactions on Power Delivery, 1994. **9**(1): p. 334–342.

[10] Calixto, W.P., et al., *Parameters estimation of a horizontal multilayer soil using genetic algorithm.* IEEE Transactions on Power Delivery, 2010. **25**(3): p. 1250–1257.

[11] Zhiqiang, H. and Z. Bin. *Soil model's inversion calculation based on genetic algorithm*. in *2011 7th Asia-Pacific International Conference on Lightning*. 2011. IEEE.

[12] Zou, J., et al., *Two-stage algorithm for inverting structure parameters of the horizontal multilayer soil.* IEEE transactions on magnetics, 2004. **40**(2): p. 1136–1139.

[13] Bäck, T., D.B. Fogel, and Z. Michalewicz, *Handbook of evolutionary computation.* Release, 1997. **97**(1): p. B1.

[14] Eiben, A.E. and J.E. Smith, *Introduction to evolutionary computing*. 2015: Springer.

[15] Črepinšek, M., S.-H. Liu, and M. Mernik, *Exploration and exploitation in evolutionary algorithms: A survey.* ACM computing surveys (CSUR), 2013. **45**(3): p. 1–33.
[16] Jesenik, M., et al., *Analytical modelling of a magnetization curve obtained by the measurements of magnetic materials' properties using evolutionary algorithms.* Applied Soft Computing, 2017. **52**: p. 387–408.
[17] Zames, G., *Genetic algorithms in search, optimization and machine learning.* Inf Tech J, 1981. **3**(1): p. 301.
[18] Holland, J.H., *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. 1992: MIT press.
[19] Das, S. and P.N. Suganthan, *Differential evolution: A survey of the state-of-the-art.* IEEE transactions on evolutionary computation, 2010. **15**(1): p. 4–31.
[20] Kennedy, J. and R. Eberhart. *Particle swarm optimization*. in *Proceedings of ICNN'95-international conference on neural networks*. 1995. ieee.
[21] Pereira, W.R., M.G. Soares, and L.M. Neto, *Horizontal multilayer soil parameter estimation through differential evolution.* IEEE Transactions on Power Delivery, 2015. **31**(2): p. 622–629.
[22] Alibabaei Shahraki, M., *Cloud drift optimization algorithm as a nature-inspired metaheuristic.* Discover Computing, 2025. **28**(1): p. 173.
[23] Khishe, M. and M.R. Mosavi, *Chimp optimization algorithm.* Expert systems with applications, 2020. **149**: p. 113338.
[24] Mirjalili, S., S.M. Mirjalili, and A. Lewis, *Grey Wolf Optimizer.* Advances in Engineering Software, 2014. **69**: p. 46–61.
[25] Heidari, A.A., et al., *Harris hawks optimization: Algorithm and applications.* Future generation computer systems, 2019. **97**: p. 849–872.
[26] Jia, H., et al., *Crayfish optimization algorithm.* Artificial Intelligence Review, 2023. **56**(Suppl 2): p. 1919–1979.
[27] Li, S., et al., *Slime mould algorithm: A new method for stochastic optimization.* Future Generation Computer Systems, 2020. **111**: p. 300–323.
[28] Dhiman, G. and V. Kumar, *Seagull optimization algorithm: Theory and its applications for large-scale industrial engineering problems.* Knowledge-based systems, 2019. **165**: p. 169–196.
[29] Veček, N., M. Črepinšek, and M. Mernik, *On the influence of the number of algorithms, problems, and independent runs in the comparison of evolutionary algorithms.* Applied Soft Computing, 2017. **54**: p. 23–45.
[30] Črepinšek, M., S.-H. Liu, and M. Mernik, *Replication and comparison of computational experiments in applied evolutionary computing: common pitfalls and guidelines to avoid them.* Applied Soft Computing, 2014. **19**: p. 161–170.
[31] Veček, N., et al., *Parameter tuning with Chess Rating System (CRS-Tuning) for meta-heuristic algorithms.* Information Sciences, 2016. **372**: p. 446–469.
[32] Črepinšek, M., et al., *Is a comparison of results meaningful from the inexact replications of computational experiments?* Soft Computing, 2016. **20**(1): p. 223–235.
[33] Kondner, R.L., *Hyperbolic stress-strain response: cohesive soils.* Journal of the Soil Mechanics and Foundations Division, 1963. **89**(1): p. 115–143.
[34] Duncan, J.M. and C.-Y. Chang, *Nonlinear analysis of stress and strain in soils.* Journal of the soil mechanics and foundations division, 1970. **96**(5): p. 1629–1653.
[35] Roscoe, K.H. and J.B. Burland, *On the generalized stress-strain behaviour of wet clay.* 1968.
[36] Dolan, E.D. and J.J. Moré, *Benchmarking optimization software with performance profiles.* Mathematical programming, 2002. **91**(2): p. 201–213.

12. Appendix A

The MATLAB source code used for all optimization runs, including the implementation of the eight metaheuristic algorithms, the constitutive models, and the objective function, is provided in the supplementary material and in the appendix for transparency and reproducibility. All algorithms were implemented using their standard formulations without problem-specific tuning.

```
Algorithm 1
Calibration Framework for Soil Constitutive Models

Initialize
    Experimental datasets
    Optimization algorithms
    Population size
    Maximum iterations
    Number of runs

FOR each Soil Model

    Define objective function

    FOR each Optimization Algorithm

        FOR each Independent Run

            Initialize population

            WHILE stopping criterion is not satisfied

                Evaluate objective function

                Update search agents

                Save best solution

            END WHILE

            Store
                Best parameters
                Objective value
                Convergence curve

        END FOR

    END FOR

    Compute
        Mean error
        Standard deviation
        Success rate

    Perform
        Friedman ranking test
        Wilcoxon signed-rank test

END FOR

Generate statistical figures

End
```

Figure A. Workflow of the proposed multi-algorithm benchmark framework for calibration of soil constitutive models.