Iraqi National Journal of Earth Science

www.earth.uomosul.edu.iq

Forecasting PM2.5 Daily Concentration in Baghdad, Iraq Based on Improving Random Forest Algorithm

Zakariya Nafi Shehab [1*], Zakariya Yahya Algamal [2], Raid Mahmood Faisal [3]

1 Department of Environmental Systems and Information, Environmental Research Center, University of Mosul, 41002 Mosul, Nineveh, Iraq.
2 Department of Statistics and Informatics Science, College of Computer Science and Mathematics, University of Mosul, 41002 Mosul, Nineveh, Iraq.
3 Department of Environmental Technology, College of Environmental Sciences, University of Mosul, 41002 Mosul, Nineveh, Iraq.

Article information

Received: 15- Nov-2024

Revised: 16- Dec -2024

Accepted: 07- Jan -2025

Available online: 01- Jan -2026

Keywords:
Coati Optimization Algorithm,
Random Forest Algorithm,
PM2.5,
Air Pollution,
Forecasting,

Correspondence:
Name: Zakariya Nafi Shehab

Email:
zakariyashehab@uomosul.edu.iq

ABSTRACT

Forecasting air quality in urban areas is complex due to difficulties in accurately defining emission flux density and the meteorological fields. Combustion gases from human and social activities are the most significant sources of PM 2.5, which is a major air pollutant. Accurate and reliable prediction of PM 2.5 levels is crucial for assessing health risks. Forecasting PM2.5 daily concentration, in general, has been predicted by Random Forest (RF) as a machine learning algorithm. However, the RF performance is highly sensitive to the choice of its hyperparameters, which usually necessitates careful tuning. Consequently, searching for the optimal set of RF hyperparameters constitutes an essential step when attempting to improve model efficiency. Various techniques have come into view for effective hyperparameter tuning of RF. Meta-heuristic optimization methods, with their strong local search abilities, can prevent the training network from getting trapped in local optima and increase the likelihood of identifying the global optimum. This paper proposes employing the Coati Optimization Algorithm (COA), a meta-heuristic approach, to improve RF hyperparameter determination, and, consequently, forecasting PM 2.5 concentrations. Daily PM 2.5 concentrations in Baghdad, Iraq, from 2019 to 2023 are gathered to train RF models and assess the proposed COA performance. The effectiveness of COA is estimated using several metrics. Overall, our proposed COA approach demonstrates superior performance in terms of evaluation criteria compared to other methods in both training and testing daily PM 2.5 concentrations.

DOI: 10.33899/injes.v26i1.60200,

التنبؤ بالتركيز اليومي لـ PM2.5 في بغداد، العراق بناءً على خوارزمية الغابات العشوائية المحسنة

زكريا نافع محمود شهاب [1*]، زكريا يحيى الجمال [2]، رائد محمود فيصل [3]

[1] قسم النظم والمعلومات البيئية، مركز بحوث البيئة، جامعة الموصل، الموصل، نينوى، العراق.

[2] قسم الإحصاء وعلوم المعلومات، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، نينوى، العراق.

[3] قسم تقانات البيئة، كلية العلوم البيئية، جامعة الموصل، الموصل، نينوى، العراق.

معلومات الارشفة

تاريخ الاستلام: 15– نوفمبر – 2024

تاريخ المراجعة: 16– ديسمبر – 2024

تاريخ القبول: 07– يناير – 2025

تاريخ النشر الالكتروني: 01– يناير – 2026

الكلمات المفتاحية:

خوارزمية التحسين الكواتي،

PM2.5،

خوارزمية الغابة العشوائية،

تلوث الهواء،

التنبؤ،

المراسلة:

الاسم: زكريا نافع شهاب

Email: zakariyashehab@uomosul.edu.iq

الملخص

يعد التنبؤ بجودة الهواء في المناطق الحضرية أمرًا معقدًا بسبب الصعوبات في تحديد كثافة تدفق الانبعاثات بدقة وتأثير مجالات الأرصاد الجوية. وتُعد غازات الاحتراق الناتجة عن الأنشطة البشرية والاجتماعية أهم مصادر انبعاثات PM 2.5، وهي ملوث رئيس للهواء. يعد التنبؤ الدقيق والموثوق بمستويات PM 2.5 أمرًا بالغ الأهمية لتقييم المخاطر الصحية. وقد تم التنبؤ بالتركيز اليومي لـ PM2.5 بشكل عام بواسطة الغابة العشوائية (RF) كخوارزمية للتعلم الآلي. ومع ذلك، فإن أداء الترددات اللاسلكية حساس للغاية لاختيار المعلمات المفرطة التي تتطلب عادةً ضبطًا دقيقًا. وبالتالي، فإن البحث عن المجموعة المثلى من المعلمات التشعبية للترددات اللاسلكية يشكل خطوة أساسية عند محاولة تحسين كفاءة النموذج. وقد ظهرت تقنيات مختلفة لضبط المعلمات الفائقة الفعالة للترددات اللاسلكية. يمكن لطرائق التحسين الفوقية المجتهدة، بقدراتها القوية في البحث المحلي، أن تمنع شبكة التدريب من الوقوع في فخ التفاوتات المحلية وتزيد من احتمالية تحديد المستوى الأمثل العالمي. تقترح هذه الورقة البحثية استخدام خوارزمية كواتي للتحسين (COA)، وهو نهج مجتهد– متغاير، لتحسين تحديد المعلمات المفرطة للترددات اللاسلكية، وبالتالي التنبؤ بتركيزات PM 2.5. تم جمع التركيزات اليومية لـ PM 2.5 في بغداد بالعراق من عام 2019 إلى عام 2023 لتدريب نماذج الترددات اللاسلكية وتقييم أداء COA المقترح. تم تقدير فعالية COA باستخدام عدة مقاييس. بشكل عام، يُظهر نهجنا المقترح لتقييم أداء النماذج المقترحة أداءً متفوقًا من حيث معايير التقييم مقارنة بالطرق الأخرى في كل من التدريب واختبار تركيزات جسيمات 2.5 اليومية.

DOI: 10.33899/injes.v26i1.60200,

Introduction

Air pollution gives rise to environmental issues such as global warming, depletion of the ozone layer, and the occurrence of acid rain. Air quality degradation, particularly in urban regions, arises from swift industrialization, infrastructure expansion, and urban development. While these advancements aim to accommodate the rapid population growth in these areas, they concurrently generate detrimental effects on human health and the environment equally (Jamil and Shehab, 2021; Wood, 2022).

In this regard, high population density is normally linked to high air quality degradation resulting from increased energy consumption, which can be attributed to both residential and industrial activities. Some major sources of this problem are the use of fossil fuels for power generation and heating purposes, which leads to the release of greenhouse gases (such as ozone, methane, carbon dioxide, nitrogen oxides), sulphur oxides, and particulate matter into the atmosphere (Wang et al., 2019). Moreover, increasing numbers of vehicles across cities, which has become the cause of congestion, worsened air quality and have strong effects on locals

(Mohd Shafie et al., 2022). Exposure to outdoor air pollution, largely due to PM2.5, has led to an estimated 3.3 million premature deaths annually across the globe, with a possible doubling of that number by 2050 (Lelieveld et al., 2015).

PM 2.5 is among the major environmental air contaminants, characterized by small size, high toxicity, easy diffusivity, and long suspension time in the air (Xu et al., 2022), which seriously threatens the environment, human health, and socio-economic contexts. PM2.5 not only carries harmful compounds but also travels extensively across distances, carrying an appreciable effect on air quality and visibility (Tsurumi and Managi, 2020). Exposure to PM2.5 raises the risk of cancer, respiratory, and cardiovascular diseases, which are major threats to human health.

Precise and efficient prediction of future PM2.5 levels holds considerable importance in environmental management (Wang et al., 2024). Human activities associated with fossil fuel consumption in transportation, power generation, heating/cooling, and various industrial processes are primary sources of anthropogenic PM2.5. Additionally, events like crop-residue burning, human-induced forest fires, and deforestation-related dust storms intermittently contribute significantly to PM2.5 levels in specific regions (Tian et al., 2021; Wood, 2022). The complexity of atmospheric conditions, including temperature inversion layers and thermal updrafts, significantly influences the accumulation or dispersion of air pollutants. However, these intricacies often elude customary meteorological measurements (Murthy et al., 2020).

PM2.5 concentration forecasting models have gained increasing interest and have been continuously updated and redeveloped in recent years. For environmental protection supervision, it is critically required that PM2.5 concentrations are continuously monitored in real time, which requires reliable methods for forecasting to determine whether air quality standards are met (Wu et al., 2022). The accurate forecast of spatiotemporal variations of air pollution requires intricate algorithms in addressing the complexities of tracking this type of pollution (Muthukumar et al., 2022). Researchers cited the need for strong estimation models that would analyse levels of air pollution to give meaningful insights and aid in informed policy formulation (Southerland et al., 2022). Therefore, it is very critical to adopt new data-driven methodologies so as to achieve high accuracy in air quality predictions.

Of late, deep learning as well as machine learning have been placed among the most important tools in earth and environmental sciences (Faisal and Shehab, 2025; Shehab and Faisal, 2025; Zhong et al., 2021)**.** They were applied to many complex tasks, like urban flood prediction (Kao et al., 2021), estimation of water quality (Algamal et al., 2025; Jamil and Shehab, 2021; Najah Ahmed et al., 2019; Shehab et al., 2024), and regional air pollution forecasting (Wong et al., 2021). These technologies have greatly improved the ability to forecast PM 2.5 levels and issue warnings on air pollution. Some algorithms of machine learning, namely SVM as well as RF, have already shown a huge potential in PM 2.5 forecasting because of their excellent ability to learn complex variable patterns and relationships. That capacity easily permits the perfect capturing of nonlinear dynamics within the PM 2.5 concentration series and, therefore, better accuracy in its forecasting (Wang et al., 2024). Such models can clearly be a great alternative to the traditional atmospheric models, which are inventory-based because they use statistics and are not limited by the presence of extensive inventories (Feng et al., 2020). An ensemble learning method, the RF algorithm is employed in this research due to its great flexibility in PM 2.5 forecasting. On that note, its strength in adaptability to handle complex relationships within data averts overfitting, a common problem in most machine learning models.

This research thus aims to go beyond the limitations associated with short-term and coarse temporal resolution in prediction and provide a method that will enable the building of a robust model for the prediction of daily variations reliably and accurately. This paper focuses on the development and extension of precise models for the forecasting and prediction of PM 2.5

concentrations using an RF model based on environmental and meteorological variables. For this purpose, daily data are used from a ground-level sensor at Baghdad U.S. Embassy, located in one of the most highly congested areas within Baghdad, Iraq. This dataset will form the basis for training, testing, and validation of the model proposed herein. This dataset contains daily measurements of PM 2.5 for four years.

Data and methods

Study area

The daily PM2.5 average concentration dataset was compiled from the Baghdad U.S. Embassy air quality monitoring station from 1st March 2019 to 1st June 2023. Undoubtedly, the data obtained from the singular monitoring post are solely indicative of the immediate conditions of the city centre, providing a comprehensive urban outlook but struggling to identify specific pollution hotspots within the city. Baghdad stands as the largest urban centre in Iraq (Fig. 1), accounting for about 22% of the national population, and therefore encounters substantial air pollution threats stemming from different sources such as vehicle emissions, industrial activities, power generation, open burning, and waste disposal. This has led to an upward trend in average yearly PM 2.5 concentration spanning over the previous ten years. Baghdad was ranked among the most polluted cities, depending on PM2.5 concentrations globally in 2022.

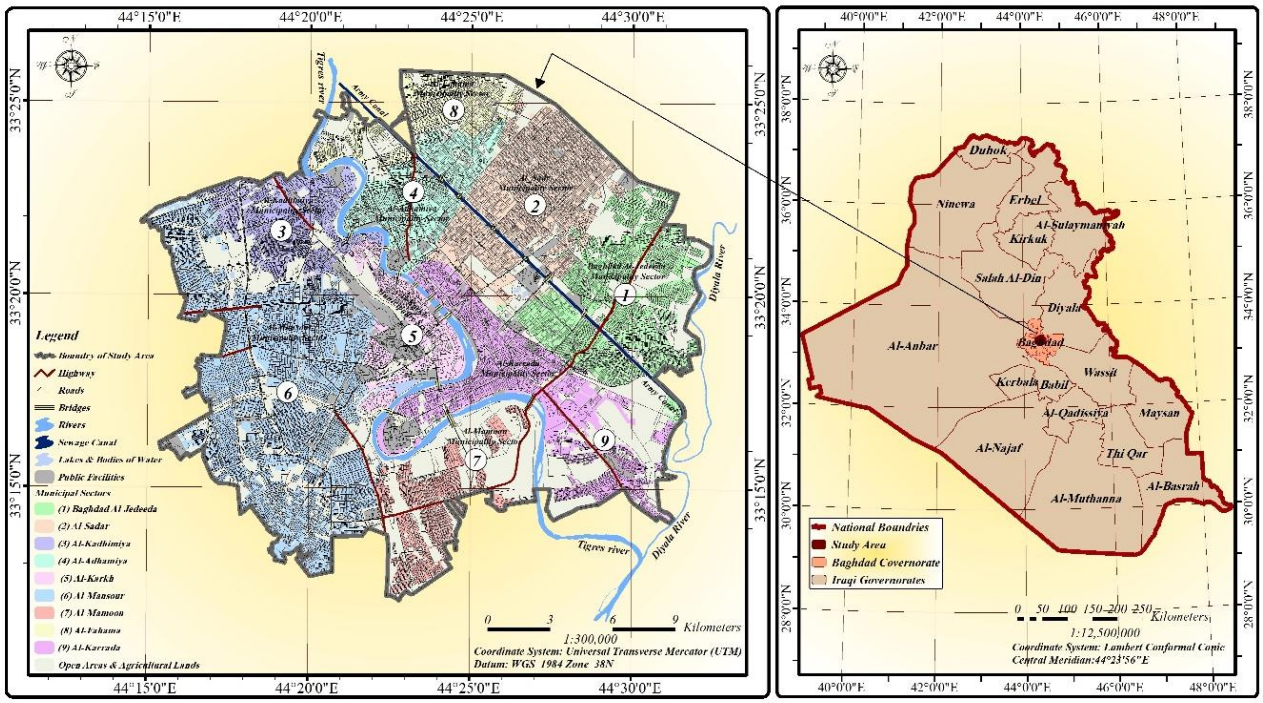

Fig. 1. Map of Baghdad highlighting metropolitan areas and municipalities.

Random Forest algorithm

RF has developed into a robust as well as versatile tool in the machine learning landscape, demonstrating remarkable performance in diverse applications and employed in regression and classification tasks equally (Hasnain et al., 2023). This technique gains from the power of ensemble learning by creating an ensemble of decision trees, each of them trained on a dissimilar random subset of training data. This will then merge the predictions of multiple decision trees, reducing overfitting, generally providing an increase in accuracy and a boost in the robustness of the model. The algorithm creates multiple bootstrap samples around the training data by randomly drawing subsets of it with replacement. A decision tree is built for each bootstrap sample. Inherent in the tree induction process: at each split, instead of

considering all features, the algorithm randomly selects a subset of features to consider at each node. This randomness in features goes a step further in reducing overfitting and increasing the variety of the forest. In prediction, each tree in the forest casts a vote (Josso et al., 2023). For classification, it chooses the most common class among the trees as a prediction. It uses the average of all trees' predictions for regression. RF is found to be at par with other state-of-the-art performance measures across a wide array of tasks all the time. Its ensemble nature reduces overfitting and improves generalizability, hence more reliable predictions. Besides, it is resistant to noise and outliers in data, retaining its accuracy even on challenging datasets (Yang et al., 2020). The RF also gives the significance of distinct features so that one can see the logic of the model's decision process. Then, the power of RF is that, with its adaptability and flexibility, it can be employed for both regression and classification tasks, which very much extends the domains of problems it is useful for.

The proposed improvement

Hyperparameters are those parameters that require defining before implementing an algorithm or before running an algorithm (Khalid and Javaid, 2020). In the case of Data Science, tuning of hyperparameters is one of the significant stages in the workflow. Done correctly, hyperparameter tuning can indeed take a worthless model and turn it into a model ready for production to make real decisions (Probst et al., 2019; Singh et al., 2021).

The identification of the appropriate hyperparameters plays a crucial part in the RF prediction accuracy and learning time (Daviran et al., 2021). However, the decision-making of the right combination of hyperparameters is very critical and consumes a lot of time (Zhu et al., 2022).

Tuning hyperparameters manually consumes a lot of time, and at the same time, a manual process needs a deeper understanding of the RF algorithm as well as its hyperparameters. As for the issues regarding the manual specification of hyperparameters, it is necessary to use hyperparameter optimization to search for the best configuration automatically (Ge et al., 2023). The methods to apply to find the values of the hyperparameters include randomized search (RS), Bayesian optimization (BO), cross-validation (CV), and grid search (GS). This process was previously conducted for all the possible combinations of hyperparameters, and the set of hyperparameters that yielded the best results in the chosen criterion was returned. However, these approaches are computationally expensive, and they do not exhaust all the possible combinations of hyperparameters (Abdulsaed et al., 2023).

Hence, there is a need for better and more effective methods for hyperparameter tuning for RF. Over the last few years, metaheuristic optimization algorithms have been extensively applied for solving the problem of hyperparameter tuning (Abdulsaed et al., 2023; Algamal et al., 2021). Recently, various new nature-inspired algorithms were proposed by researchers to extend and upgrade the exploration and exploitation of the existing algorithms. The Coati Optimization Algorithm (COA) has emerged as one of the most popular because of its high efficiency (Dehghani et al., 2023; Jia et al., 2023).

COA is a population-based metaheuristic approach, where each coati represents an individual within the population. The location of each coati within the search space affects the decision variables' values. Therefore, in COA, coatis serve as potential solutions to the problem at hand. Initially, the coatis are randomly placed within the search space as defined thereafter (Dehghani et al., 2023):

$$X_i : X_{i,j} = Lb_j + r * (Ub_j - Lb_j), \quad i = 1, 2, ..., N ; j = 1, 2, ..., m \tag{1}$$

One interesting behavior in this exploration phase is that one group of coatis climbs the tree to scare the iguana down to ground level, where others are waiting to capture it. This behavior shows the ability of the algorithm to explore or examine certain aspects in a particular context. The algorithm mimics the behavior based on the notion that half the population climbs

the trees to force the iguana down, and the remaining half wait at ground level to catch the iguana as it falls, and it looks like the following Eq. (2).

$$X_i^{new} : X_{i,j}^{new} = X_{i,j} + r*(I_j - \vartheta * X_{i,j}), i = 1,2,.....,[N/2]; j = 1,2.....,m \quad (2)$$

Once the iguana drops to the ground, it is assigned a random position (I^G) within the search space, and the coatis under the tree adjust their position according to Eqs. (3) and (4). If the objective function value at the new position calculated for each coati is an improvement over the current value, the new position is accepted. Conversely, if the new position yields a worse objective function value, the coati remains in its original position (Eq. 5) (Dehghani et al., 2023).

$$I^G : I_j^G = Lb_j + r*(Ub_j - Lb_j), j = 1,2,.....,m \quad (3)$$

$$X_i^{new} : X_{i,j}^{new} = \begin{cases} X_{i,j} + r*(I_j^G - \vartheta * x_{i,j}), F(I^G) < F(X_i) \\ X_{i,j} + r*(X_{i,j} - I_j^G), else \end{cases} \quad (4)$$

with $i = \left[\frac{N}{2}\right]+1, \left[\frac{N}{2}\right]+2,.....,N$, $j = 1,2,.....,m$

$$X_i = \begin{cases} X_i^{new}, F(X_i^{new}) < F(X_i) \\ X_i, else \end{cases} \quad (5)$$

where: X_i is the current position of the ith coati, X_i^{new} is the newly calculated position for the ith coati, I^G is the randomly generated position of the iguana, F is the objective function value.

In the next phase, the exploitation phase, the behavior of coatis employing a predator-escape strategy is described. This strategy keeps the coati close to its current position and in a safe stance, enhancing the algorithm's exploitation capability. To accomplish this, a random position is created near each coati using Eqs. (6) and (7). If this newly generated position results in an improved objective function value, the coati assumes the new position; otherwise, it retains its original position (Eq. 5).

$$Lb_j^{local} = \frac{Lb_j}{t}, Ub_j^{local} = \frac{Ub_j}{t}, t = 1,2,.....,T \quad (6)$$

$$X_i^{new} : X_{i,j}^{new} = X_{i,j} + (1-2*r)*(Lb_j^{local} + r*(Ub_j^{local} - Lb_j^{local})), i = 1,2,...,N, j = 1,2,..,m \quad (7)$$

To optimize the hyperparameters of RF using COA as an improvement proposition, the position vector *X* of each coati is defined as a vector with dimension *D*, representing the coati's position in the COA. Consequently, each vector *X* corresponds to a specific configuration of the RF, with each dimension of *X* representing a distinct hyperparameter of the RF. Therefore, we have five positions that each coati in the swarm will search for them. Consequently, our proposed improvement is as follows:

Step 1: The number of coatis, N_{coati} , is set to 30, and the maximum number of iterations is T =500.

Step 2: The positions of each coati are randomly specified. For the number of trees in the forest (N_{trees}), the position is randomly generated from a uniform distribution between 10 and 100. While for the minimum number of samples at leaf node (MN_{leaf}) and for the minimum

number of samples considered to split an internal node ($MN_{internal}$), the position for each one is randomly generated from a uniform distribution within the range [1,25] and range [2,20], respectively. Further, the position of the maximum depth of the tree (M_{tree}) is randomly generated from a uniform distribution between 1 and 200. Finally, the number of features ($N_{features}$) taken for the best split is considered depending on the number of lags of the series. A representation of the coati position (Fig. 2).

Step 3: The fitness function is defined as

$$\text{fitness} = \min\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i\,(N_{trees}, MN_{leaf}, MN_{internal}, M_{tree}, N_{features})})^2\right]. \quad (8)$$

Step 4: The positions of the coati are updated using Eq. (4) and Eq. (7), respectively.

Step 5: Steps 3 and 4 are repeated until a T is reached.

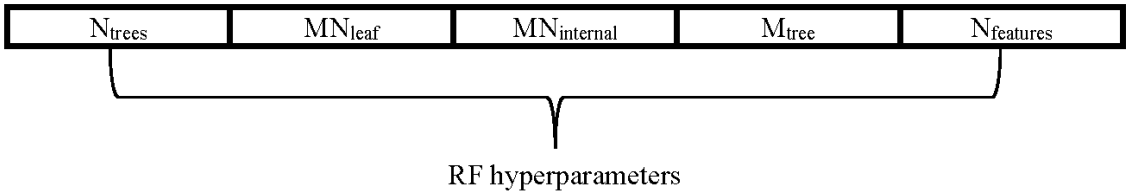

Fig. 2. Position of each coati in the COA algorithm.

Forecasting evaluation criteria

Four assessment measures, RMSE, MAE, direction accuracy (DA), and R^2 are adopted in this research to estimate the effectiveness of the projected forecasting method. Their simple mathematical expressions (Table 1).

Table 1: Forecasting evaluation criteria.

Evaluation criterion	Mathematical formula
MAE	$\frac{1}{n}\sum_{t=1}^{n}\lvert y_t - \hat{y}_t\rvert$
RMSE	$\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$
DA	$\frac{1}{n}\sum_{t=1}^{n} z_t, \quad z_t = \begin{cases} 1, & \text{if } (y_{t+1} - y_t)(\hat{y}_{t+1} - \hat{y}_t) \geq 0 \\ 0, & \text{otherwise} \end{cases}$
R^2	$1 - \left[\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \Big/ \sum_{t=1}^{n}(\hat{y}_t - \bar{y}_t)^2\right]$

Results analysis

Our goal in using our proposed approach, COA, is to demonstrate that an adequate choice of hyperparameters can produce better PM2.5 daily concentration forecasting. Comprehensive comparison tests using RS, BO, CV, and GS are used to examine the forecasting performance of our proposed algorithm, COA.

The data are divided into two sets: the training data set, including 1270 PM2.5 daily concentrations (from 1 March 2019 to 31 December 2022), and the testing data set, including 151 PM2.5 daily concentrations (from 1 January 2023 to 1 June 2023). Figures (3 and 4) depict the daily PM2.5 concentrations' pattern over time for the training and testing data sets. Both figures proved that the daily PM2.5 concentration pattern is nonlinear and non-stationary over

time. The descriptive statistics for both training and testing data sets (Table 2) were used to investigate the behavior of daily PM2.5 concentrations. The daily PM2.5 concentrations' pattern has trend, fluctuation, asymmetry, and intermittency.

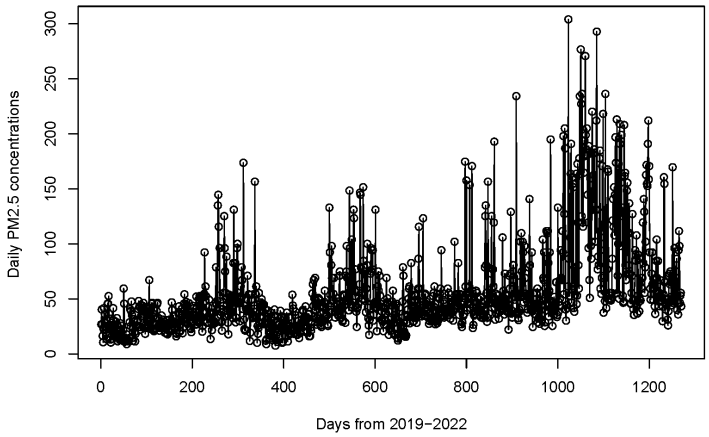

Fig. 3. Time series of the daily PM2.5 concentrations for the training data set.

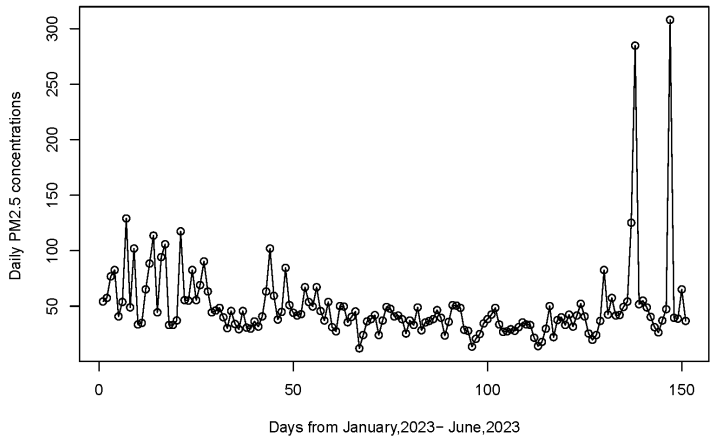

Fig. 4. Time series of the daily PM2.5 concentrations for the testing data set.

Table 2: Statistical descriptive daily PM2.5 concentrations.

	Training data set	Testing data set
Mean	55.51	49.19
Standard deviation	42.47	35.83
Skewness	2.27	4.74
Kurtosis	5.77	29.51
Minimum	7.44	12.00
Maximum	303.98	308.02

The forecasting performances of the COA, RS, BO, CV, and GS models in terms of evaluation criteria (Tables 3 and 4) for training and testing data sets, respectively. According to the results of Tables 3 and 4, it can be shown from the forecasting results of daily PM2.5 concentrations that the suggested approach, COA, may significantly improve forecasting accuracy and generalization capacity because it has lower MAE, RMSE, greater DA, and greater R^2 values. COA performs better than the RS, BO, CV, and GS methods. From Table (3), compared to RS, BO, CV, and GS, the reduction in terms of MAE and RMSE of COA is 87.26%, 85.53%, 86.01%, 86.55% and 87.22%, 86.75%, 86.99%, and 87.16% respectively.

Similarly, for the testing data set (Table 4) compared to RS, BO, CV, and GS, the corresponding criteria, in terms of MAE, decreased by 87.61%, 86.67%, 86.91%, and 87.21% respectively. While in terms of RMSE, it decreased by 83.61%, 83.05%, 83.34% and 83.54% respectively.

Meta-heuristic algorithms perform well in predicting, even in the absence of data processing, except for the traditional techniques; RS, BO, CV, and GS performance could be caused by the arbitrary hyperparameter settings. Moreover, in the comparison between CV, RS, and GS in estimating the best hyperparameters of RF, the results of the evaluation indicators based on the training set show that the CV method is superior to the RS and GS methods. In addition, the application of the RS method revealed that RS obtained non-satisfactory results.

Table 3: Forecasting results of the COA and the benchmark approaches RS, BO, CV, and GS for the training set.

	COA	RS	BO	CV	GS
MAE	0.124	0.974	0.857	0.886	0.922
RMSE	1.308	10.237	9.877	10.057	10.189
DA	0.984	0.651	0.694	0.681	0.663
R^2	0.988	0.614	0.711	0.704	0.625

Table 4: Forecasting results of the COA and the benchmark approaches RS, BO, CV and GS for the testing set.

	COA	RS	BO	CV	GS
MAE	0.205	1.655	1.538	1.567	1.603
RMSE	1.789	10.918	10.558	10.738	10.87
DA	0.964	0.635	0.678	0.665	0.647
R^2	0.972	0.598	0.695	0.688	0.609

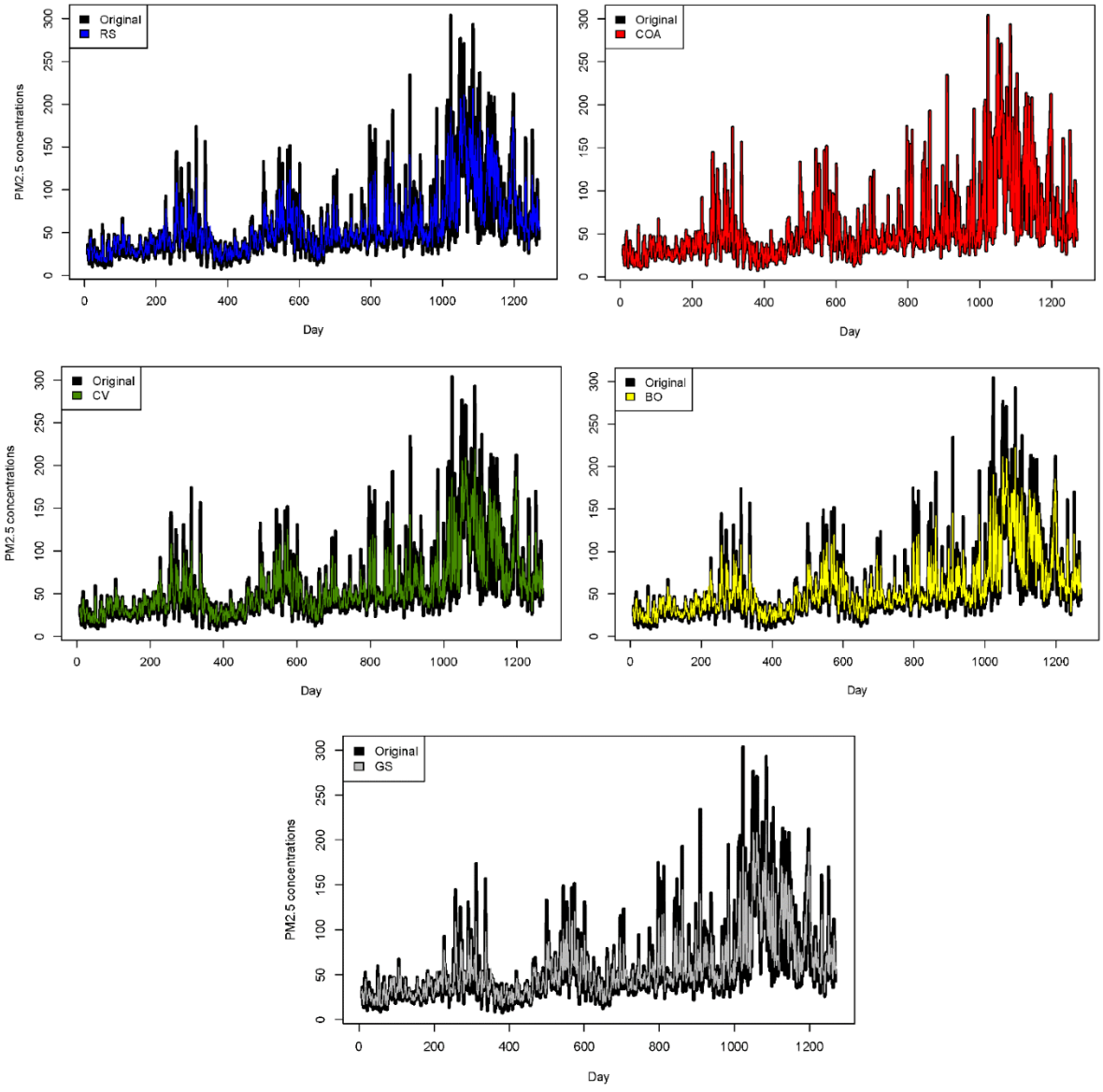

Fig. 5. Forecasting results in the training data set based on methods used.

From time-series plots of forecasting, the model accuracy of RF based on the COA algorithm is stable and superior to others for both the training and testing datasets. As demonstrated in Figures 5 and 6, the COA model's predicted daily PM2.5 levels are essentially in line with the actual values, demonstrating the high quality of the model's forecast. In addition, training and testing PM2.5 concentrations greatly increase the prediction capacity of COA. This is combined with the rather smooth time series of the COA. As a result, the COA model predicts daily PM2.5 concentrations with remarkable accuracy. Moreover, as shown in Figures 5 and 6, the PM2.5 concentrations forecast depending on CV are slightly closer to the actual value in comparison to the BO method. On the other hand, the PM2.5 concentrations forecast of RS method performs poorly in their forecasts over time.

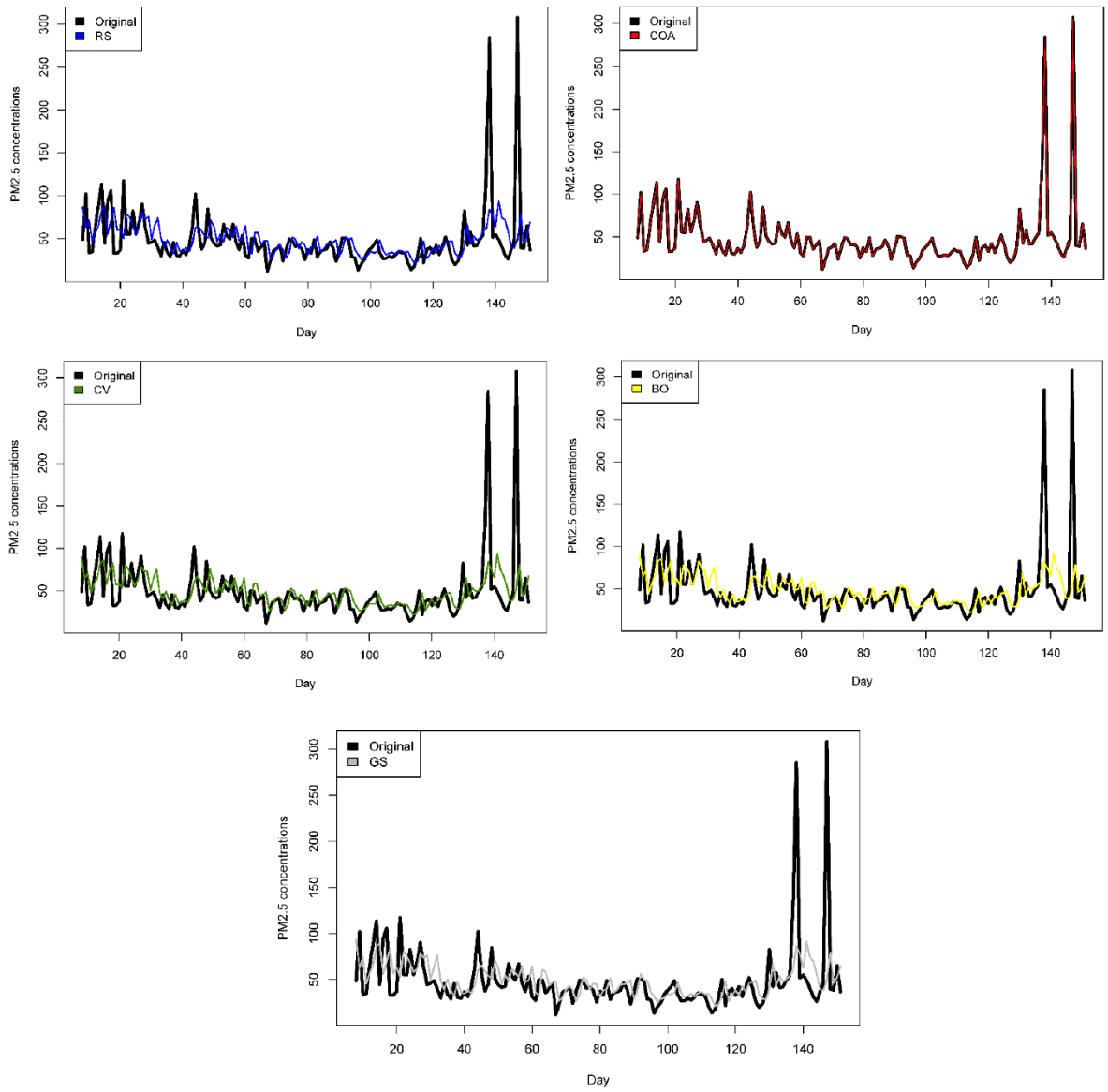

Fig. 6. Forecasting results in the testing data set based on the methods used.

Conclusion

The novelty of this approach is, in fact, the accurate estimation of PM2.5 concentrations. This study combines the RF model with a meta-heuristic optimization algorithm to estimate daily variations in PM2.5 concentrations within Baghdad, Iraq. This further enhances the exploration and exploitation capabilities of the COA to carry out hyperparameter optimization within RF. The results indicate that the COA-based methodology is superior compared to other RS, BO, CV, and GS methods, improving the predictive accuracy of the RF model significantly.

In particular, COA shows a better performance for all the metrics used to evaluate its performance, including MAE, RMSE, DA, and coefficient of determination (R^2). In the future, hybrid models and other meta-heuristic optimization algorithms may be developed to study the increased accuracy of PM2.5 forecasting to devise highly useful environmental policies.

The accurate prediction of PM2.5 concentration has a very critical role in the protection of public health and the management of the environment. This enhanced COA, as employed in this study for improving the prediction accuracy of daily PM2.5 levels, provides a potent tool for policymakers and health officials. With increased reliability in the forecasts, the authorities can initiate timely warnings and measures to decrease exposure, such as traffic restrictions, control over emissions by industries, and public advisories.

References

Abdulsaed, E.H., Alabbas, M. and Khudeyer, R.S., 2023. Hyperparameter Optimization for Convolutional Neural Networks using the Salp Swarm Algorithm. Informatica (Slovenia), 47(9), pp. 133–144. https://doi.org/10.31449/inf.v47i9.5148

Algamal, Z.Y., Qasim, M.K., Lee, M.H. and Mohammad Ali, H.T., 2021. Improving grasshopper optimization algorithm for hyperparameters estimation and feature selection in support vector regression. Chemometrics and Intelligent Laboratory Systems, 208. https://doi.org/10.1016/j.chemolab.2020.104196

Algamal, Z.Y., Shehab, Z.N. and Faisal, R.M., 2025. Spatiotemporal variation analysis of Tigris River water quality in Mosul, Iraq during 2020–2023 based on environmetric techniques. Environmental Earth Sciences, 84(2), 69. https://doi.org/10.1007/s12665-024-12086-z

Daviran, M., Maghsoudi, A., Ghezelbash, R. and Pradhan, B., 2021. A new strategy for spatial predictive mapping of mineral prospectivity: Automated hyperparameter tuning of random forest approach. Computers and Geosciences, 148. https://doi.org/10.1016/j.cageo.2021.104688

Dehghani, M., Montazeri, Z., Trojovská, E. and Trojovský, P., 2023. Coati Optimization Algorithm: A new bio-inspired metaheuristic algorithm for solving optimization problems. Knowledge-Based Systems, 259. https://doi.org/10.1016/j.knosys.2022.110011

Faisal, R.M. and Shehab, Z.N., 2025. Integrating GIS and comparative multi-criteria decision-making techniques for landslide susceptibility assessment. Progress in Physical Geography. https://doi.org/10.1177/03091333251380439

Feng, R., Gao, H., Luo, K. and Fan, J. ren., 2020. Analysis and accurate prediction of ambient PM2.5 in China using Multi-layer Perceptron. Atmospheric Environment, 232. https://doi.org/10.1016/j.atmosenv.2020.117534

Ge, D.M., Zhao, L.C., and Esmaeili-Falak, M., 2023. Estimation of rapid chloride permeability of SCC using hyperparameters optimized random forest models. Journal of Sustainable Cement-Based Materials, 12(5), pp. 542–560. https://doi.org/10.1080/21650373.2022.2093291

Hasnain, A., Sheng, Y., Hashmi, M.Z., Bhatti, U.A., Ahmed, Z. and Zha, Y., 2023. Assessing the ambient air quality patterns associated to the COVID-19 outbreak in the Yangtze River Delta: A random forest approach. Chemosphere, 314. https://doi.org/10.1016/j.chemosphere.2022.137638

Jamil, N.R. and Shehab, Z.N., 2021. Landscape Perspective to River Pollution: A Study of Bentong River, Malaysia. In Water Pollution and Management Practices. https://doi.org/10.1007/978-981-15-8358-2_2

Jia, H., Shi, S., Wu, D., Rao, H., Zhang, J. and Abualigah, L., 2023. Improve coati optimization algorithm for solving constrained engineering optimization problems. Journal of Computational Design and Engineering, 10(6), pp. 2223–2250. https://doi.org/10.1093/jcde/qwad095

Josso, P., Hall, A., Williams, C., Le Bas, T., Lusty, P., and Murton, B., 2023. Application of random-forest machine learning algorithm for mineral predictive mapping of Fe-Mn crusts in the World Ocean. Ore Geology Reviews, 162. https://doi.org/10.1016/j.oregeorev.2023.105671

Kao, I.F., Liou, J.Y., Lee, M.H. and Chang, F.J., 2021. Fusing stacked autoencoder and long short-term memory for regional multistep-ahead flood inundation forecasts. Journal of Hydrology, 598. https://doi.org/10.1016/j.jhydrol.2021.126371

Khalid, R. and Javaid, N., 2020. A survey on hyperparameters optimization algorithms of forecasting models in smart grid. In Sustainable Cities and Society (Vol. 61). Elsevier Ltd. https://doi.org/10.1016/j.scs.2020.102275

Lelieveld, J., Evans, J.S., Fnais, M., Giannadaki, D. and Pozzer, A., 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature, 525(7569). https://doi.org/10.1038/nature15371

Mohd Shafie, S.H., Mahmud, M., Mohamad, S., Rameli, N.L.F., Abdullah, R. and Mohamed, A.F., 2022. Influence of urban air pollution on the population in the Klang Valley, Malaysia: a spatial approach. Ecological Processes, 11(1). https://doi.org/10.1186/s13717-021-00342-0

Murthy, B.S., Latha, R., Tiwari, A., Rathod, A., Singh, S. and Beig, G., 2020. Impact of mixing layer height on air quality in winter. Journal of Atmospheric and Solar-Terrestrial Physics, 197. https://doi.org/10.1016/j.jastp.2019.105157

Muthukumar, P., Cocom, E., Nagrecha, K., Comer, D., Burga, I., Taub, J., Calvert, C.F., Holm, J. and Pourhomayoun, M., 2022. Predicting PM2.5 atmospheric air pollution using deep learning with meteorological data and ground-based observations and remote-sensing satellite big data. Air Quality, Atmosphere and Health, 15(7). https://doi.org/10.1007/s11869-021-01126-3

Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M. and Elshafie, A., 2019. Machine learning methods for better water quality prediction. Journal of Hydrology, 578. https://doi.org/10.1016/j.jhydrol.2019.124084

Probst, P., Wright, M.N. and Boulesteix, A.L., 2019. Hyperparameters and tuning strategies for random forest. In Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 9, Issue 3, Wiley-Blackwell. https://doi.org/10.1002/widm.1301

Shehab, Z.N. and Faisal, R.M., 2025. Harnessing wind for hydrogen: comparative MCDM-GIS assessment of optimal plant locations. Journal of King Saud University – Engineering Sciences, 37(6), 30. https://doi.org/10.1007/s44444-025-00025-7

Shehab, Z.N., Faisal, R.M. and Ahmed, S.W., 2024. Multi-criteria decision making (MCDM) approach for identifying optimal solar farm locations: A multi-technique comparative analysis. Renewable Energy, 237. https://doi.org/10.1016/j.renene.2024.121787

Shehab, Z.N., Farhhan, A.F. and Faisal, R.M., 2024. Spatial variation influence of landscape patterns on surface water quality across an urbanized watershed in Mosul city, Iraq. Sustainable Water Resources Management, 10(5), 181. https://doi.org/10.1007/s40899-024-01162-8

Singh, P., Chaudhury, S. and Panigrahi, B.K., 2021. Hybrid MPSO-CNN: Multi-level Particle Swarm optimized hyperparameters of Convolutional Neural Network. Swarm and Evolutionary Computation, 63. https://doi.org/10.1016/j.swevo.2021.100863

Southerland, V.A., Brauer, M., Mohegh, A., Hammer, M.S., van Donkelaar, A., Martin, R.V., Apte, J.S., and Anenberg, S.C., 2022. Global urban temporal trends in fine particulate matter (PM2·5) and attributable health burdens: estimates from global datasets. The Lancet Planetary Health, 6(2). https://doi.org/10.1016/S2542-5196(21)00350-8

Tian, M., Gao, J., Zhang, L., Zhang, H., Feng, C. and Jia, X., 2021. Effects of dust emissions from wind erosion of soil on ambient air quality. Atmospheric Pollution Research, 12(7). https://doi.org/10.1016/j.apr.2021.101108

Tsurumi, T. and Managi, S., 2020. Health-related and non-health-related effects of PM2.5 on life satisfaction: Evidence from India, China and Japan. Economic Analysis and Policy, 67. https://doi.org/10.1016/j.eap.2020.06.002

Wang, J., Wu, T., Mao, J. and Chen, H., 2024. A forecasting framework on fusion of spatiotemporal features for multi-station PM2.5. Expert Systems with Applications, 238, 121951. https://doi.org/10.1016/j.eswa.2023.121951

Wang, W., Zhao, S., Jiao, L., Taylor, M., Zhang, B., Xu, G. and Hou, H., 2019. Estimation of PM2.5 Concentrations in China Using a Spatial Back Propagation Neural Network. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-019-50177-1

Wong, P.Y., Lee, H.Y., Zeng, Y.T., Chern, Y.R., Chen, N.T., Candice Lung, S.C., Su, H.J. and Wu, C.Da., 2021. Using a land use regression model with machine learning to estimate ground level PM2.5. Environmental Pollution, 277. https://doi.org/10.1016/j.envpol.2021.116846

Wood, D.A., 2022. Trend decomposition aids forecasts of air particulate matter (PM2.5) assisted by machine and deep learning without recourse to exogenous data. Atmospheric Pollution Research, 13(3). https://doi.org/10.1016/j.apr.2022.101352

Wu, A., Harrou, F., Dairi, A. and Sun, Y., 2022. Machine learning and deep learning-driven methods for predicting ambient particulate matters levels: A case study. Concurrency and Computation: Practice and Experience, 34(19). https://doi.org/10.1002/cpe.7035

Xu, Z., Niu, L., Zhang, Z., Hu, Q., Zhang, D., Huang, J. and Li, C. 2022. The impacts of land supply on PM2.5 concentration: Evidence from 292 cities in China from 2009 to 2017. Journal of Cleaner Production, 347. https://doi.org/10.1016/j.jclepro.2022.131251

Yang, L., Xu, H. and Yu, S., 2020. Estimating PM2.5 concentrations in Yangtze River Delta region of China using random forest model and the Top-of-Atmosphere reflectance. Journal of Environmental Management, 272. https://doi.org/10.1016/j.jenvman.2020.111061

Zhong, S., Zhang, K., Bagheri, M., Burken, J.G., Gu, A., Li, B., Ma, X., Marrone, B.L., Ren, Z.J., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B.M., Xiao, X., Yu, X., Zhu, J.J. and Zhang, H., 2021. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. Environmental Science and Technology, 55(19), pp. 12741–12754. https://doi.org/10.1021/acs.est.1c01339

Zhu, N., Zhu, C., Zhou, L., Zhu, Y. and Zhang, X., 2022. Optimization of the Random Forest Hyperparameters for Power Industrial Control Systems Intrusion Detection Using an Improved Grid Search Algorithm. Applied Sciences (Switzerland), 12(20). https://doi.org/10.3390/app122010456