Environment International 175 (2023) 107931

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Full length article

Machine learning algorithms for high-resolution prediction of spatiotemporal distribution of air pollution from meteorological and soil parameters

Hai Tao[a,b,c], Ali H. Jawad[d], A.H. Shather[e], Zainab Al-Khafaji[f], Tarik A. Rashid[g], Mumtaz Ali[h], Nadhir Al-Ansari[i,*], Haydar Abdulameer Marhoon[j,k], Shamsuddin Shahid[l], Zaher Mundher Yaseen[m,n,*]

[a] School of Computer and Information, Qiannan Normal University for Nationalities, Duyun, Guizhou 558000, China
[b] State Key Laboratory of Public Big Data, Guizhou University, Guizhou, Guiyang 550025, China
[c] Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[d] Faculty of Applied Sciences, UniversitiTeknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[e] Dep of Computer Technology Engineering, Engineering Technical College, University of Alkitab, Iraq
[f] Department of Building and Construction Technologies Engineering, AL-Mustaqbal University College, Hillah 51001, Iraq
[g] Computer Science and Engineering Department, University of Kurdistan Hewler, Erbil, KR, Iraq
[h] UniSQ College, University of Southern Queensland, QLD 4350, Australia
[i] Dept. of Civil, Environmental and Natural Resources Engineering, Lulea Univ. of Technology, Lulea T3334, Sweden
[j] Information and Communication Technology Research Group, Scientific Research Center, Al-Ayen University, Thi-Qar, Iraq
[k] College of Computer Sciences and Information Technology, University of Kerbala, Karbala, Iraq
[l] Department of Hydraulics and Hydrology, School of Civil Engineering, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), 81310 Skudia, Johor, Malaysia
[m] Civil and Environmental Engineering Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
[n] Interdisciplinary Research Center for Membranes and Water Security, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia

ARTICLE INFO

Handling Editor: Xavier Querol

Keywords:
Air quality prediction
$PM_{2.5}$ concentration
Simulated annealing
Machine learning
Arid climate

ABSTRACT

This study uses machine learning (ML) models for a high-resolution prediction (0.1°×0.1°) of air fine particular matter ($PM_{2.5}$) concentration, the most harmful to human health, from meteorological and soil data. Iraq was considered the study area to implement the method. Different lags and the changing patterns of four European Reanalysis (ERA5) meteorological variables, rainfall, mean temperature, wind speed and relative humidity, and one soil parameter, the soil moisture, were used to select the suitable set of predictors using a non-greedy algorithm known as simulated annealing (SA). The selected predictors were used to simulate the temporal and spatial variability of air $PM_{2.5}$ concentration over Iraq during the early summer (May-July), the most polluted months, using three advanced ML models, extremely randomized trees (ERT), stochastic gradient descent backpropagation (SGD-BP) and long short-term memory (LSTM) integrated with Bayesian optimizer. The spatial distribution of the annual average $PM_{2.5}$ revealed the population of the whole of Iraq is exposed to a pollution level above the standard limit. The changes in temperature and soil moisture and the mean wind speed and humidity of the month before the early summer can predict the temporal and spatial variability of $PM_{2.5}$ over Iraq during May-July. Results revealed the higher performance of LSTM with normalized root-mean-square error and Kling-Gupta efficiency of 13.4% and 0.89, compared to 16.02% and 0.81 for SDG-BP and 17.9% and 0.74 for ERT. The LSTM could also reconstruct the observed spatial distribution of $PM_{2.5}$ with MapCurve and Cramer's V values of 0.95 and 0.91, compared to 0.9 and 0.86 for SGD-BP and 0.83 and 0.76 for ERT. The study provided a methodology for forecasting spatial variability of $PM_{2.5}$ concentration at high resolution during the peak pollution months from freely available data, which can be replicated in other regions for generating high-resolution $PM_{2.5}$ forecasting maps.

* Corresponding authors.
E-mail addresses: haitao@sgmtu.edu.cn (H. Tao), ali288@uitm.edu.my (A.H. Jawad), akhsh@uoalkitab.edu.iq (A.H. Shather), zainabal-khafaji@uomus.edu.iq (Z. Al-Khafaji), tarik.ahmed@ukh.edu.krd (T.A. Rashid), Mumtaz.Ali@usq.edu.au (M. Ali), nadhir.alansari@ltu.se (N. Al-Ansari), haydar@alayen.edu.iq (H.A. Marhoon), sshahid@utm.my (S. Shahid), z.yaseen@kfupm.edu.sa (Z.M. Yaseen).

https://doi.org/10.1016/j.envint.2023.107931
Received 13 December 2022; Received in revised form 18 March 2023; Accepted 11 April 2023
Available online 15 April 2023

1. Introduction

1.1. Research background

Air pollution substantially affects public health (Jamei et al., 2022; Li et al., 2020; Mokoena et al., 2020), crop yield (Burney and Ramanathan, 2014; Rollin et al., 2022), human productivity (Chen and Zhang, 2021), social activities (Liu et al., 2022; Yan et al., 2019) and economy (Jiang et al., 2020; Mujtaba and Shahzad, 2021). Globally, 9 out of 10 people breathe polluted air, which causes approximately 7 million premature death annually (Fowler et al., 2020). The global annual cost of disease burden due to higher fine particulate matter ($PM_{2.5}$), a major constituent of air pollution, is nearly U$21 billion (Lelieveld et al., 2015; World Bank Group and IHME, 2016). It is also responsible for 1.8 billion working days loss. Its cumulative effects on different sectors are equivalent to U$2.9 trillion economic loss, nearly 3.3% of global GDP (Kjellström et al., 2019). Despite significant economic losses, air pollution is gradually increasing in most parts of the globe, particularly in developing countries. A study showed that global mean population-weighted $PM_{2.5}$ concentrations increased by 38%, and the related excess deaths increased from 89% in 1960 to 124% in 2009 (Butt et al., 2017). A continuous increase in air pollution would cause an increase in related healthcare costs to USD 176 billion in 2060 and an annual working-day loss of 3.7 billion (Lanzi, 2016).

Though $PM_{2.5}$ concentration depends on various sources, their movement, transport and deposition depend on various meteorological and earth's surface physical factors (Elminir, 2005; Fu et al., 2021; Hashim et al., 2021; Mokoena et al., 2020; Qi et al., 2020; Wang et al., 2018). A study by He et al. (2017) showed that a 70% variation in air pollution in China depends on different meteorological variables. Precipitation improves air quality by forcing fine air particulate down to the ground (De Nevers, 2010). Air temperature and wind influence air movement and, thus, air pollution (Yang et al., 2020). High humidity is positively related to air $PM_{2.5}$ levels. Another study showed that temperature and wind speed are major meteorological factors defining air pollution in China (Li et al., 2019).

The above studies indicate the possibility of forecasting $PM_{2.5}$ from meteorological variables. Therefore, several attempts have been made to forecast $PM_{2.5}$ from meteorological variables (Bai et al., 2016). The early initiatives were based on conventional statistical regression models, including linear regression (Pérez et al., 2000), generalized linear regression (Wu et al., 2013), and autoregressive moving averages (Wang and Guo, 2009). The conventional regression-based models provide prediction by fitting linear relationships between meteorological variables and air $PM_{2.5}$ concentration. However, in practice, the relationships are often not linear. For example, the rainfall amount and the concentration of $PM_{2.5}$ in the air have an inverse but nonlinear relationship. Several nonlinear statistical regression models have been developed to overcome this challenge (Cobourn, 2010; Sorek-Hamer et al., 2013). However, the capability of nonlinear statistical models is very limited. The highly nonlinear and complex relationships are often encountered between meteorological variables and $PM_{2.5}$ which is not possible to map using those techniques.

1.2. Machine learning literature review

The excellent ability of machine learning (ML) models to map nonlinear relationships has opened a new avenue for better prediction of $PM_{2.5}$ with sufficient lag time. Several ML models have been employed in recent years for $PM_{2.5}$ concentration predictions. This includes artificial neural networks (Casallas et al., 2021; Ventura et al., 2019), support vector regression (Zhang et al., 2021), extreme learning machines (Yin et al., 2021), gradient boosting (He et al., 2022), and deep learning (Xiao et al., 2020). Several studies also compared the performance of different ML algorithms to find the best method. For example, Li et al. (2021) used a combination of support vector regression, random forest, and neural networks to predict $PM_{2.5}$ concentrations. Wu et al. (2020) used a combination of random forest and gradient boosting decision trees to predict $PM_{2.5}$ concentrations. A comprehensive review conducted for various ML methods used for $PM_{2.5}$ prediction, including support vector regression, random forest, artificial neural networks, and deep learning (Peng et al., 2022). The studies showed the excellent performance of ML models in forecasting $PM_{2.5}$ concentration (Danesh Yazdi et al., 2020). In recent years, ML algorithms are advanced further to handle more randomness in data, store memory to predict multiple lead time values, and create a more comprehensive parameter range to map higher data nonlinearity (Bagheri, 2022; Chang et al., 2020; Cui et al., 2022; Karimian et al., 2019; Xiao et al., 2019). The LSTM is one such algorithm which uses a broader range of hyperparameter values. It helps LSTM to work on a wide range of parameters to explore the pattern of a complex multidimensional series. Its capacity to store memory also helps to improve its prediction capability. Stochastic gradient descent-backpropagation (SGD-BP) is another advanced version of the ANN algorithm which uses optimized search to minimize the model error. Its stochastic nature to converge to higher accuracy has made it superior to many other ML algorithms (Katongtung et al., 2022). The decision tree-based models have also been improved by adding randomness in generating the trees. Extremely randomized tree (ERT) is one such decision tree algorithm that is computationally more efficient (Afshar et al., 2022; Sachdeva and Kumar, 2022). These three algorithms have been widely used in many environmental studies, including pollution modelling (Bacanin et al., 2022; Ibrahim et al., 2022; Pruthi and Liu, 2022; Q. He et al., 2022), but have not been explored yet in forecasting spatiotemporal distribution of $PM_{2.5}$.

Long-Short Memory Network (LSTM), a robust version of the deep learning model, was developed to forecast $PM_{2.5}$ concentration in Beijing, China (Niu et al., 2023). The authors incorporated dew point temperature and wind speed as additional parameters for air quality index, CO, NO_2 and PM_{10}. The research finding approved the potential of the LSTM model to forecast $PM_{2.5}$ multiple days ahead. Although ML models have shown a noticeable improvement, researchers are moving toward even more robust methodologies for ML process enhancement. Recently, Hu et al., (2023) predicted $PM_{2.5}$ and O_3 concentrations using hybridized convolutional neural network and bidirectional (CNN)-LSTM-gated recurrent unit (GRU)The authors developed their hybrid CNN-LSTM-GRU model based on six pollution indicators, which provided better accuracy than the standalone ML models. On a similar mechanism of hybridized ML models, several other researchers developed models for PM2.5 concentration prediction (Eren et al., 2023; Kim et al., 2022; Wood, 2022). Attempts have been made to forecast air $PM_{2.5}$ concentration on different time scales, including hourly (Kanabkaew, 2013), daily (Xiao et al., 2020), and monthly or seasonal (Wu et al., 2022) scales. The seasonal forecasting of $PM_{2.5}$ provides an early warning of air pollution conditions with sufficient lag time and, therefore, can be used for awareness development and mitigate its effect (Jiang et al., 2017; Wu et al., 2022). However, forecasting the seasonal variability of $PM_{2.5}$ needs a long-term prediction model. The selection of predictors that can indicate long-term changes in $PM_{2.5}$ can be employed to overcome this challenge.

2. Research gap and motivation

The previous studies mainly concentrated on meteorological variables for $PM_{2.5}$ predictions. Studies showed that soil moisture is important in forming suspended fine air particulates (Wang et al., 2018). The transportation and fate of these particulates depend on wind speed, rainfall, and other meteorological variables. Therefore, considering all these factors is important for the reliable prediction of $PM_{2.5}$. The seasonal prediction of $PM_{2.5}$ also needs predictors that define possible long-term changes in air $PM_{2.5}$ concentration. The trends in soil moisture or meteorological variables can be suitable indicators of possible long-term changes in $PM_{2.5}$ concentration. For example, the difference in soil moisture between two consecutive months indicates its increasing or decreasing nature. Such information can be used for the seasonal prediction of $PM_{2.5}$. However, the inclusion of the changing pattern of predictors for forecasting long-term change in $PM_{2.5}$ concentration has not been well studied.

The Middle Eastern region is one of the world's most polluted regions (Elbayoumi et al., 2013). Air pollution caused nearly 13,000 premature deaths in the region in 2017 (Myllyvirta, 2020). The $PM_{2.5}$ level in most cities in the Middle East is 5 to 10 times higher than the prescribed limit (Saad, 2021). The human and economic costs estimated for air pollution demonstrated higher than 3% of GDP in some middle east countries (Heger et al., 2022). Iraq is one of the most polluted countries by $PM_{2.5}$ in the region. The average $PM_{2.5}$ of 39.6 $\mu g/m^3$ has made Iraq among the ten most polluted countries globally (Al-Aseel, 2022). The crude oil-driven power generators, due to poor electrical infrastructure, fires from refineries, and war-induced pollution along with other conventional sources, have made the pollution level in the country much higher than in many other countries in the region (Zwijnenburg, 2015). The country needs mitigation measures to reduce the impacts of air pollution to improve people's living standards and economic development. Forecasting air pollution can help to cope with the forthcoming hazards.

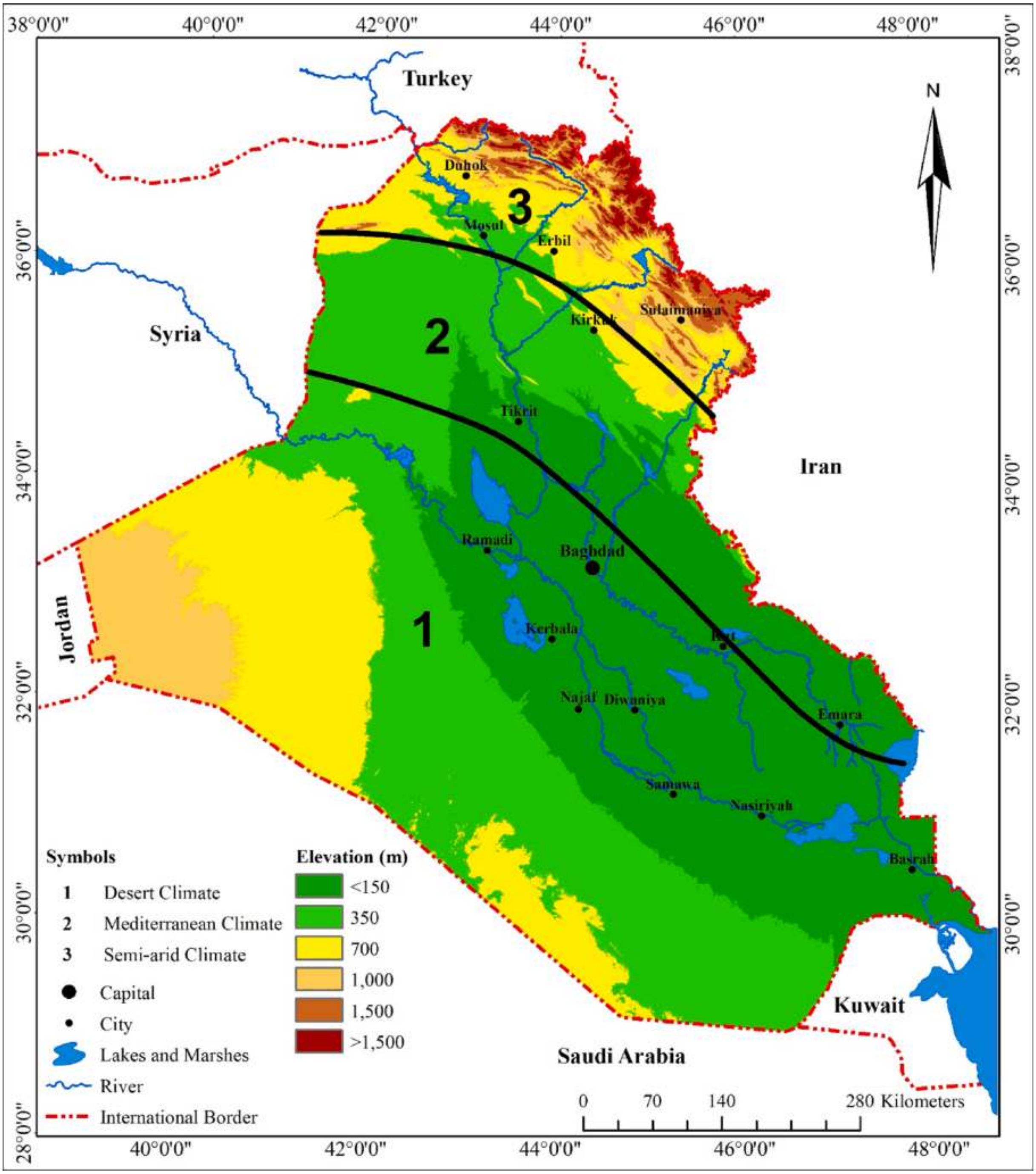

Fig. 1. Geography of Iraq. The location of Iraq in the middle east is also depicted. (Desert Climate Zone (1), Semi-Arid Climate Zone (2), Mediterranean Climate Zone (3)).

3. Research objectives

This study intends to develop a seasonal $PM_{2.5}$ forecasting model for Iraq for high-resolution mapping of possible spatial patterns in $PM_{2.5}$ in early summer, the country's most air-polluted period. Different lags of soil moisture, wind speed, rainfall, temperature, and relative humidity and their trends in the previous season were considered to select the predictors for forecasting $PM_{2.5}$ concentration in early summer. This allowed the prediction of the spatial pattern of seasonal $PM_{2.5}$ concentration before the beginning of the season. The spatial map of predicted $PM_{2.5}$ can be employed to increase awareness of possible pollution disasters and take necessary mitigation measures.

4. Investigated region and data

4.1. Geography of Iraq

Iraq lies between latitude: 29°15′N – 38°15′N, and longitude: 38°45′ – 48°45′E has a total area of 438,320 km^{2,} including 924 km^2 of inland waters (Fig. 1). Iran surrounds it to the east, Turkey to the north, Syria and Jordan to the west, Saudi Arabia and Kuwait to the south, and the Persian Gulf to the southeast. The country can be classified into three topographic units (Jaradat, 2003): mountainous region, which cover an area of 92,000 km^2, in the north and northeast, with elevation ranges from 1,700 to nearly 3,000 m; undulating lands covering 42,000 km^2 in the south and the west with an average altitude varies from 200 to 1000 m, and the plain covers 69.2% of the land extended from the central north to the south including the Arabian deserts in the west.

The climate of Iraq is mainly continental, subtropical semi-arid type, with the north and northeastern mountainous regions having a Mediterranean climate. Iraq has four climatic seasons, namely, (1) a hot and dry summer (May to September); (2) cool and wet winter (December to February); (3) spring (March to April); and (4) autumn (October to November) (Abd Alraheem et al., 2022). Iraq is one of the most vulnerable countries to climate-related hazards. Sand and dust storms are among the most devastating hazards, which have become more recurrent with the temperature rise in recent years (Al-Kasser, 2021). Air pollution due to sand and dust is every year problem, particularly during early summer (May-July). Therefore, the country is considered one of the countries globally with high $PM_{2.5}$ pollution. A global study of population exposure to $PM_{2.5}$ revealed 100% of the Iraqi population is exposed to a $PM_{2.5}$ level exceeding the WHO limit (Cohen et al., 2017). Iraq has a population of nearly 45 million, growing by 2.3% annually. Most of the population lives in the central-east and south alluvial lands, where pollution is high due to being surrounded by deserts. Exposure of a large population to high $PM_{2.5}$ would gradually increase the country's economic burden for public health care and negatively affect economic growth if measures are not.

4.2. Data description and sources

This study used fine particulate matter ($PM_{2.5}$) data developed by Washington University in St. Louis (V5.GL.02) (Castillo et al., 2021). It was developed by integrating Aerosol Optical Depth (AOD), estimated using different satellite sensors, with the GEOS-Chem chemical transport model. The model was trained and validated with in-situ $PM_{2.5}$ data using the Geographically Weighted Regression method to generate the final product. The data are available on a monthly time scale with a spatial resolution of 0.01° × 0.01° at https://wustl.app.box.com/v/ACAG-V5GL03-GWRPM25/folder/183614225548.

The present study used ERA5 meteorological and soil data (Hersbach et al., 2020) to predict $PM_{2.5}$. ERA5 is the fifth generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis of a large number of global atmospheric, land and ocean variables. ERA5 data is generated at a horizontal resolution of 0.1° × 0.1° by combining the ECMWF model output produced through the physical principle of atmospheric circulation and the in-situ data gathered from across the globe.

The list of ERA variables used to predict $PM_{2.5}$ concentration is provided in Appendix A**.** These datasets were downloaded from:

https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview. The resolution of $PM_{2.5}$ data is 0.01° × 0.01°. It was aggregated to the ERA5 resolution of 0.1° × 0.1°. The monthly pollution data are available for the period 1998–2021. Therefore, ERA5 and $PM_{2.5}$ data for the period 1998–2021 with a spatial resolution of 0.1° × 0.1° was used for model training and validation.

5. Methodology overview

The steps followed for the spatiotemporal prediction of $PM_{2.5}$ are shown using a flowchart in Fig. 2. The present study considered four meteorological and one soil variable to predict $PM_{2.5}$ concentration. A partial correlation analysis was conducted between each predictor and $PM_{2.5,}$ and the predictors found significantly correlate with $PM_{2.5}$ was used for further analysis. Different lags and differences in lag values were used in a non-greedy feature selection method known as Simulated Annealing (SA) to select the final set of predictors. The selected predictors were used in three ML algorithms, LSTM, SGD-BP, and ERT, for predicting $PM_{2.5}$. The hyperparameters of the ML models were optimized using Bayesian optimization to improve the model's prediction capability. Details of the methods are discussed in the following subsections.

5.1. Partial correlation

Partial correlation was initially used to assess the influence of different variables on $PM_{2.5}$ concentrations. It was used to assess the association of each variable with $PM_{2.5}$ by removing the influence of other variables. Partial correlation between two variables, X_1 and $X_{2,}$ by removing the influence of X_3 is measured (Bhagat et al., 2020) as,

$$r_{x_1x_2.x_3} = \frac{r_{x_1x_2} - r_{x_1x_3} r_{x_2x_3}}{\sqrt{1 - r^2_{x_1x_3}}\sqrt{1 - r^2_{x_2x_3}}} \tag{1}$$

where $r_{x_ax_b}$ is the correlation between X_a and X_b, where a and b can be 1, 2 or 3 in the present case.

5.2. Feature selection using non-greedy wrapper

The factors that showed higher partial correlation were selected. Their different lags and differences in lag values were used to select the final set of predictors using a non-greedy wrapper method. The non-greedy wrapper was selected considering its ability to escape the localize traps. Several non-greedy wrappers, including Naïve Bayes, genetic algorithm, and SA, are available for ML model feature selection. In this study, SA was selected due to its better capability in feature selection (Kirkpatrick et al., 1983). The SA method introduces randomness in the selection procedure, which allows SA to initiate new search spaces to obtain better optima (Jamei et al., 2021).

5.3. Machine learning models

5.3.1. Long-short term memory (LSTM)

LSTM (Hochreiter and Schmidhuber, 1997) is an advanced version of the recurrent neural network (RNN). The RNN is a version of the neural network which uses a feedback network to recognize the previous data for prediction. This ability to memorize the data sequence has made such a network better predict environmental events that follow a seasonal pattern. The major problem of RNN is short memory; therefore, it can't provide good predictions for longer data sequences (Jamei et al., 2023a). This is due to the drastic declination of the effect of input on the

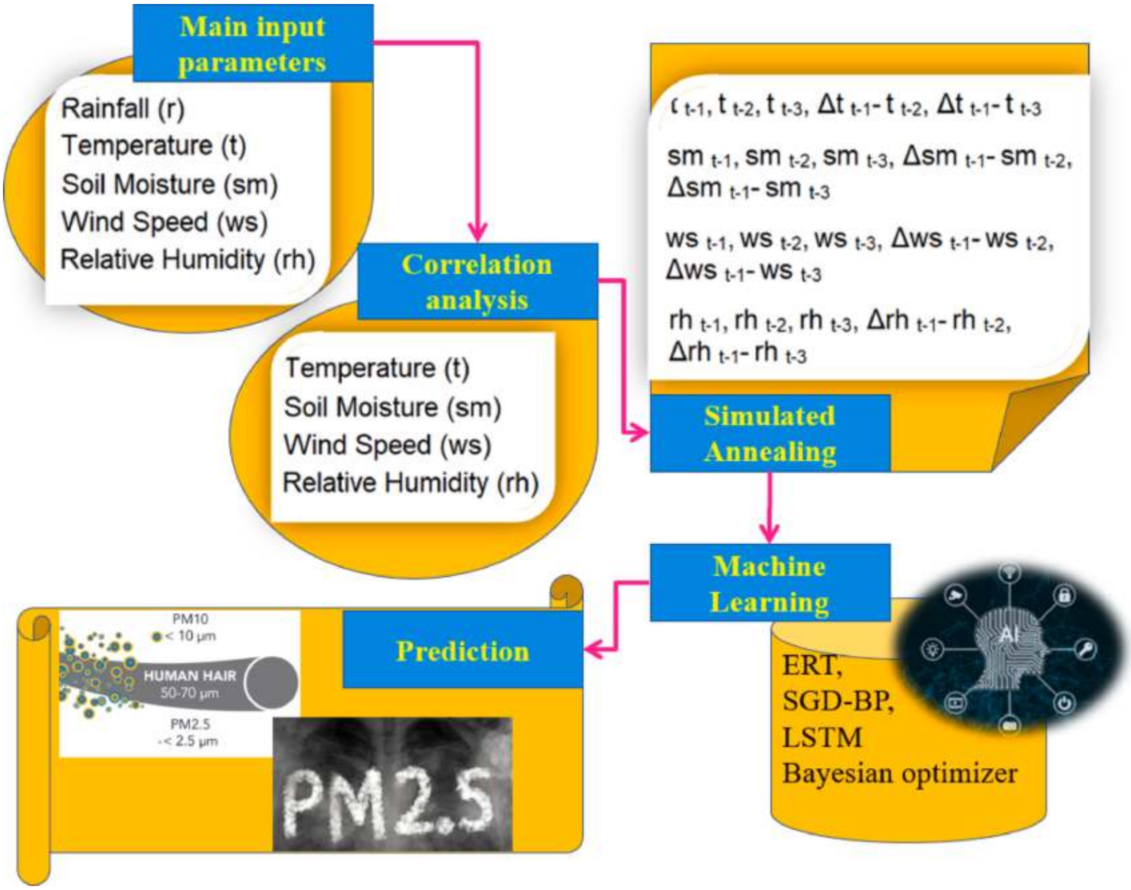
Main input parameters
Rainfall (r)
Temperature (t)
Soil Moisture (sm)
Wind Speed (ws)
Relative Humidity (rh)
Correlation analysis
Temperature (t)
Soil Moisture (sm)
Wind Speed (ws)
Relative Humidity (rh)
t_{t-1}, t_{t-2}, t_{t-3}, $\Delta t_{t-1}-t_{t-2}$, $\Delta t_{t-1}-t_{t-3}$
sm_{t-1}, sm_{t-2}, sm_{t-3}, $\Delta sm_{t-1}-sm_{t-2}$, $\Delta sm_{t-1}-sm_{t-3}$
ws_{t-1}, ws_{t-2}, ws_{t-3}, $\Delta ws_{t-1}-ws_{t-2}$, $\Delta ws_{t-1}-ws_{t-3}$
rh_{t-1}, rh_{t-2}, rh_{t-3}, $\Delta rh_{t-1}-rh_{t-2}$, $\Delta rh_{t-1}-rh_{t-3}$
Simulated Annealing
Machine Learning
ERT,
SGD-BP,
LSTM
Bayesian optimizer
AI
Prediction
PM10
< 10 μm
HUMAN HAIR
50-70 μm
PM2.5
< 2.5 μm
PM2.5

Fig. 2. The flowchart showing the steps followed for predicting $PM_{2.5}$ in Iraq.

hidden layer with time. LSTM can solve this problem. LSTM improves the capability of RNN using a memory unit (Hunt et al., 2022; Van Houdt et al., 2020).

The schematic diagram of an LSTM is presented in Appendix A. A traditional LSTM consists of three layers: (i) forget, (ii) input, and (iii) output. The Forget layer deletes the memory of the precedent state (C_{t-1}) based on the present input and antecedent hidden output. It uses a sigmoid function (f_t) which yields an output of either 0 or 1 (Jamei et al., 2023b). It filters out the memory when the output is zero. Otherwise, it passes to the next layer. This can be presented below (Hochreiter and Schmidhuber, 1997),

$$C_t^f = C_{t-1} \times f_t \tag{2}$$

The input layer generates a state (C_t) from the output of the forget layer (C_t^f) and the present input (C_t^i). It also uses a forget logic. It filters out the present input if it is unexpected. It uses a *tanh* function to scale the values in the range of −1 to 1.

$$C_t = C_t^f + C_t^i \tag{3}$$

The output layer considers the present input and present state of the cell to generate a hidden state (H_t) and cell output (O_t). This layer also uses a *tanh* function to scale the cell state. Besides, different biases are incorporated at different layers (Hochreiter and Schmidhuber, 1997),

$$H_t = O_t \times tanh(C_t) \tag{4}$$

5.3.2. *Stochastic gradient descent with back propagation neural network (SGD-BP)*

Conventional ANN uses a backpropagation algorithm for model tuning. SGD-BP (Amari, 1993) combines SGD and BP error minimization techniques for model training. The schematic diagram of an SGD-BP is shown in Appendix B. Gradient descent is an optimization method employed to search the variables' values that minimize a function. It does this by estimating the target function gradient. Generally, a 1st order derivative of the function for input variables is estimated to locate the optimum values. A negative gradient indicates new values that provide the evaluation function's lower estimate. A learning rate is used to guide the changes in the input variables. The process is iterated until the threshold minimum of the function output is reached. It often fails to find the evaluation function minimum in case of noisy input data. SGD can overcome this. The SGD minimizes a loss function of the prediction model for the calibration data (Ernst, 2014; Ye, 2022). The SGD-BPANN estimate the net input function (net) from the inputs and their weights (Amari, 1993),

$$net = \sum_{i=0}^{n} w_i x_i \tag{5}$$

The output is estimated from the net input function using the following equation:

$$o = \frac{1}{1 + e^{-net}} \tag{6}$$

SGD-BP uses a function gradient (-∇C) at a tangent vector location that changes rapidly. This concept is employed to estimate the new weight (W^+) from the present weight (W) using a learning rate of η (Amari, 1993):

$$W^+ = W - \eta \nabla C \tag{7}$$

In SGD-BP, the evaluation function slope of model parameters is a probabilistic approximation or noisy. It helps avoid the statistical noise from noisy data affecting the gradient signal.

5.3.3. Extremely randomized trees (ERT)

ERT (Geurts et al., 2006) is an ensemble decision algorithm like a random forest (RF). However, it generates a large number of unpruned trees using the calibration data and an aggregated average of the output of all trees as the final output in case of regression. The major difference between ERT with RF is that it randomly samples the features at each split location of the trees using a greedy algorithm. This allows it to select the optimum split point.

ERT calibrates each tree in the ensemble tree t ∈ {1, ⋯, T} with the complete set of calibration data. Each sample is a d-dimensional feature vector f_j. The samples are employed to generate the tree root, and then the split function is used until leaf nodes are generated from the sample. Two children subsets are generated uniformly without replacement of complete feature set f. ERT selects split points randomly from all samples in calibration rather than bootstrap. It makes the decision trees less related and reduces the variance, which is eventually realized by increasing the ensemble tree number. This also allows ERT to add randomness in growing trees and improve the prediction capability compared to simple RF and other tree-based algorithms (Geurts et al., 2006). Three hyperparameters are required to optimize for better performance of ERT, the number of trees, the number of inputs to select randomly, and the minimum sample size to create a new split point. Like RF, ERT is efficient in computation and modelling high-dimension features. However, extra efficiency in ERT comes from the increased randomness (Padmaja et al., 2020).

5.3.4. Bayesian optimization

The ML models used in this study have several hyperparameters. The hyperparameters of LSTM are the number of LSTM cells, number of hidden layers, learning rate and batch size. The ERT hyperparameters are the number of estimators, max features, min sample split, min sample leaf and max depth, while the SGD-BP hyperparameters include learning rate, momentum, batch size, and regularization parameters. These hyperparameter values must be selected properly to improve the learning processes. However, its optimization is a challenge in ML modelling (Weiqi et al., 2022). The use of optimization algorithms with ML models can help in the selection of hyperparameters and enhance model prediction. Generally, random or grid search algorithms are used for parameter tuning. This study used Bayesian optimization for faster and better optimization of model parameters (Dewancker et al., 2016). The Bayesian method is a global optimization technique which builds a probability function to tune the parameters based on a fitness function using calibration data. Like the grid or random search method, it iteratively fits the evaluation function with different parameter sets to choose the best set. However, the Bayesian method uses a probabilistic approach to define hyperparameters as the probability of fitness value, which is known as a surrogate of the fitness function (Snoek et al., 2012),

$$P(fitness|hyperparameters) \tag{8}$$

This surrogate drives the method to find the next hyperparameter set that performs best based on the surrogate. The selected set is then evaluated based on the actual fitness function, and the surrogate model is modified based on the evaluation result. The process is repeated until a threshold is reached.

The continuous modification of the Bayesian model helps to select the next set of hyperparameters in an informed manner. It helps the method to be more accurate in finding global optima. Unlike random search, the Bayesian algorithm is driven by its past status; therefore, it is much faster to find the optimum. Several studies reported better performance of the Bayesian method for ML hyperparameter tuning (Gao et al., 2021; Yin and Li, 2022).

5.3.5. Performance metrics

The prediction performance of the models in replicating $PM_{2.5}$ time series at each location was validated using normalized root mean square error (NRMSE) in %, root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), Willmott's modified coefficient of agreement (MD), and Kling-Gupta efficiency (KGE) (Faskari et al., 2022; Yaseen, 2021). The equation of the performance metrics and their range and optimum values are provided as follow:

$$NRMSE = 100*\frac{\sqrt{\frac{1}{N}*\sum_{i=1}^{N}(S_i - O_i)^2}}{sd(O_i)}\ 0 - \infty \quad Optimal \quad Value = 0 \tag{9}$$

$$RMSE = \sqrt{\frac{1}{N}*\sum_{i=1}^{N}(S_i - O_i)^2}\ 0 - \infty \quad Optimal \quad Value = 0 \tag{10}$$

$$MAE = \frac{1}{N}*\sum_{i=1}^{N}|S_i - O_i|\ 0 - \infty \quad Optimal \quad Value = 0 \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}(O_i - \mu_o)^2}\ 0 - 1 \quad Optimal \quad Value = 1 \tag{12}$$

$$md = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)}{\sum_{i=1}^{n}(|S_i - \mu_o| + |O_i - \mu_o|)}\ 0 - 1 \quad Optimal \quad Value = 1 \tag{13}$$

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\mu_s}{\mu_o} - 1\right)^2 + \left(\frac{\sigma_s/\mu_s}{\sigma_o/\mu_o} - 1\right)^2}\ -1 - \infty \quad Optimal \quad Value = -1 \tag{14}$$

The performance of the models in reconstructing the spatial variability of the forecasted $PM_{2.5}$ map was evaluated using two spatial indices, Mapcurves and Cramer's V. Mapcurves (Hargrove et al., 2006) measures the similarity between two maps using the following equation:

$$Mapcurves = \sum\left[\left(\frac{C}{B+C}\right)\left(\frac{C}{A+C}\right)\right] \tag{9}$$

where C is the degree of intersection between the two maps, A and B are the total area of observed and modelled $PM_{2.5}$ distribution maps.

Cramer's V assesses the spatial agreement between two maps as (Cramér, 1946),

$$Cramer'sV = \sqrt{\frac{x^2}{N(\min(m,n) - 1}} \tag{10}$$

where x^2 is Chi-Square, N is the number of grid points, m and n are the rows and columns in the map. Mapcurves and Cramer's V values range between 0 and 1, where 1 represents the best match.

6. Application results and analysis

6.1. Spatiotemporal distribution of $PM_{2.5}$ in Iraq

Fig. 3 shows the spatial distribution of annual mean $PM_{2.5}$ over Iraq. The annual mean $PM_{2.5}$ in the country varies between 10.8 and 98.4 μg/m^3. The $PM_{2.5}$ concentration is higher in the south, more than 75 μg/m^3 at most locations. It is also high in some places in the central-west and northeast regions. The lower average $PM_{2.5}$ is mostly in the elevated northern region and the western desert plains. However, the average $PM_{2.5}$ is higher than the limit of World Health Organization (WHO) defined in 2022 (5 μg/m^3) and the WHO limit of 10 μg/m^3 defined in 2005 (Pai et al., 2022) at all locations in Iraq. It means the population of the whole of Iraq is exposed to polluted air and, thus, the severity of air pollution in Iraq.

The monthly distribution of air pollution levels is shown in Fig. 4. The $PM_{2.5}$ values at all grid points over Iraq for a month were used to prepare the boxplot of the month. The boxplot of each month of the year revealed the seasonal variation of $PM_{2.5}$ levels in Iraq. Fig. 4 shows the

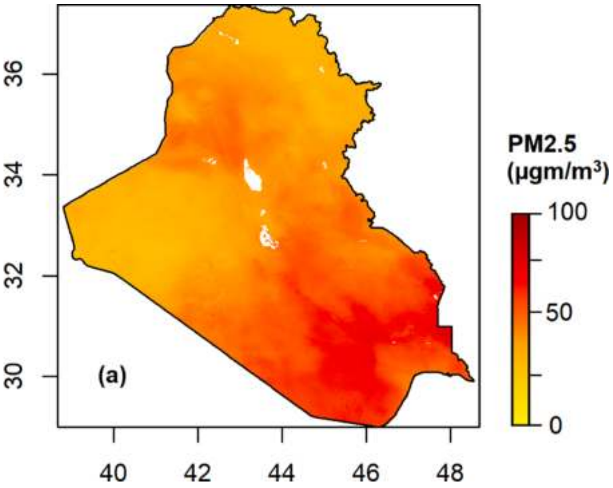

Fig. 3. Spatial distribution of annual mean $PM_{2.5}$ over Iraq for the period 1998–2021.

higher pollution in the early summer months (May to July), while the lowest is from November to January. The low $PM_{2.5}$ was also noticed in August. The country experiences a medium pollution level in spring and autumn. The lower limit of the whiskers indicates the lowest pollution level in Iraq. The figure shows that the lowest level of the whisker of the boxes was always more than 10 μg/m^3. This indicates the pollution level is higher than the WHO (2022) prescribed threshold at all locations. Outliers in the $PM_{2.5}$ level were noticed in most months, particularly in June and July. The high values of $PM_{2.5}$ in these two months indicate the extreme pollution level at some locations.

The mean $PM_{2.5}$ values in the early summer months were higher than 55 μg/m^3. The lower limit of the whiskers was more than 21.3 μg/m^3 in all these months. This indicates high pollution levels in these months all over Iraq. Therefore, this study focused on $PM_{2.5}$ concentration prediction from May to July to facilitate early measures and planning before the beginning of the high pollution period. The spatial distribution of mean $PM_{2.5}$ concentration from May to July (MJJ) is shown in Fig. 5. It shows a similar to the annual spatial pattern of $PM_{2.5}$ in MJJ. However, the pollution level was higher than the annual mean. The $PM_{2.5}$ over the whole south was more than 80 μg/m^3 during these three months. The $PM_{2.5}$ in the less polluted northern and western regions were also more than 40 μg/m^3. The high pollution level over the country during MJJ indicates their severe implications for public health and the economy and expresses the need for its prediction.

6.2. Selection of predictors of $PM_{2.5}$

The spatial distribution of the predictors used for forecasting air $PM_{2.5}$ concentration over Iraq is shown in Fig. 6. The rainfall in the country varies from above 900 mm in the north to less than 100 mm in the south and west. Most parts of Iraq receive rainfall of less than 200 mm. The mean temperature is low in the mountainous region in the north and high in the rest of the country, particularly in the south and central west. The annual mean temperature is more than 25 °C in most of the country. The soil moisture follows a similar pattern to rainfall. It is high in the northern high rainfall region and less in the rest of the country. It is less than 0.1 m^3/m^3 over most of the country, indicating dry soil dominates the study area. The annual mean relative humidity in the country varies from 18 to 59%, indicating a desert climate. It is high in the country's northern half, including the far north and central west, while low in the south. The wind speed follows a different pattern than the other factors considered in this study. It is high along a throng extended from the southeast to the northwest, passing through the country's central region. The analysis revealed low rainfall, high temperature, and low soil moisture dominating the country. None of the factors follows the $PM_{2.5}$ concentration pattern over Iraq, as shown in

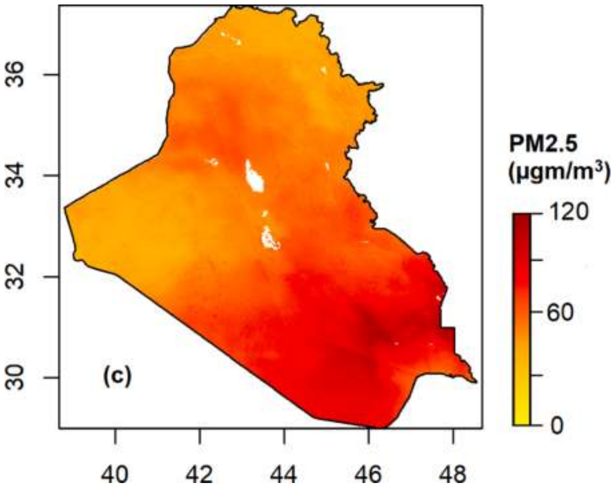

Fig. 5. Spatial distribution of mean $PM_{2.5}$ during May-June over Iraq for the period 1998–2021.

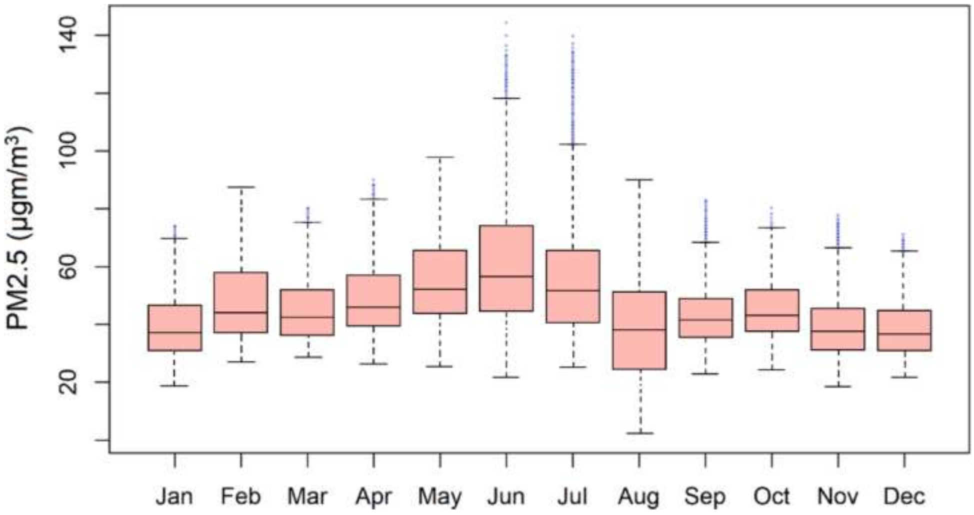

Fig. 4. The boxplot of monthly mean $PM_{2.5}$ over Iraq for the period 1998–2021.

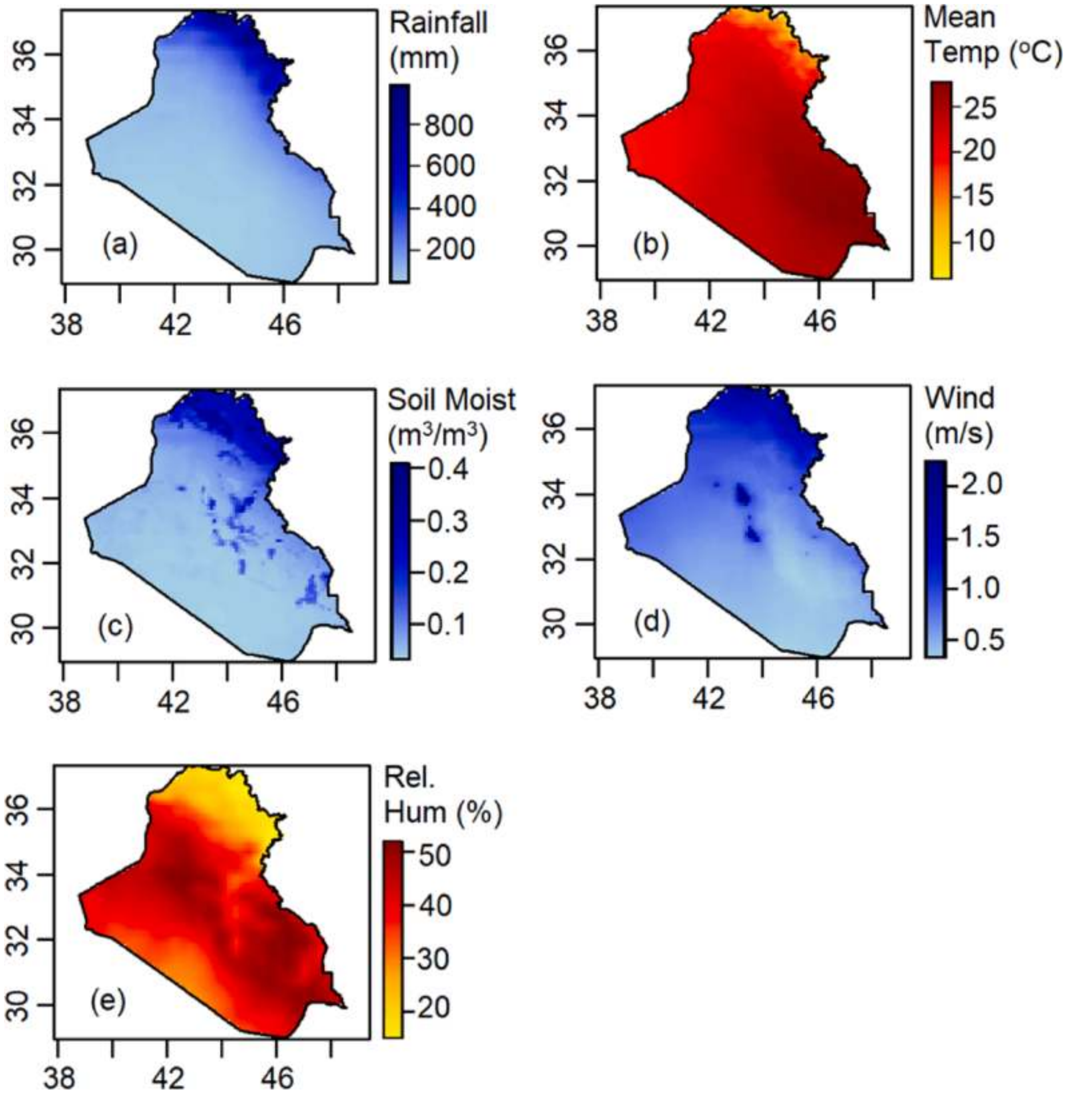

Fig. 6. Spatial distribution of the predictors used in the present study: (a) rainfall, (b) mean temperature, (c) soil moisture, (d) wind speed, and (e) relative humidity.

Fig. 5. The $PM_{2.5}$ distribution also does not follow the population distribution, which is mostly concentrated in the west, particularly the central west. It indicates the complexity of $PM_{2.5}$ distribution over the country and the difficulty in prediction.

The partial correlation of the predictors with $PM_{2.5}$ is presented in Fig. 7. The data at all the grid points were merged and then used for estimating partial correlation. The partial correlation estimated the association of each factor with $PM_{2.5}$ by removing the influence of other factors. The figure shows the higher partial correlation of wind speed and temperature with $PM_{2.5}$ (0.42), followed by relative humidity (-0.38), soil moisture (-0.31) and rainfall (-0.22). The positive correlation of wind speed and temperature with $PM_{2.5}$ indicates higher wind speed and temperature accelerate $PM_{2.5}$ concentration. The relationship of relative humidity, soil moisture and rainfall with $PM_{2.5}$ was negative, indicating their increase reduces $PM_{2.5}$ concentration. The relationships are physically justifiable. The analysis showed a significant correlation of $PM_{2.5}$ with wind speed, temperature, relative humidity and soil moisture at a 99% confidence interval. In contrast, the relationship of $PM_{2.5}$ with rainfall was significant at a 95% confidence interval. Therefore, only wind speed, temperature, relative humidity, and soil moisture were used to derive the predictors for better prediction accuracy. It has been mentioned earlier that rainfall in Iraq is less than 200 mm in most parts of the country. The low rainfall is incapable of wet deposition of suspended $PM_{2.5}$ in the air. Therefore, the present study found less influence of rainfall on the $PM_{2.5}$ concentration in Iraq.

The present study used the preliminarily selected four variables' lags and time difference values to select the final set of predictors using SA. The predictors used for selecting the final set are shown in the first column of Table 1. The three time-lags of each factor (t-1, t-2 and t-3) and their difference between two periods, $\Delta_{t-1\ -\ t-2}$ and $\Delta_{t-1\ -\ t-3}$ were used. The difference in a variable between two periods indicates the change of the variable with time, which helps to indicate how it may change in the near future. Therefore, those values and the time lags were used in the present study. The finally selected predictors using SA from the set of preliminarily selected predictors are given in the second column of Table 1. The SA selected four variables most suitable for predicting $PM_{2.5}$ concentration, tm_{t-1}, $tm_{(t-1\ -\ t-2)}$, ws_{t-1}, $sm_{(t-1\ -\ t-3)}$, rh_{t-1}. It means the first lag of temperature, wind speed and relative humidity or the values of those parameters in April, and the changes in temperature and soil moisture in recent two consecutive months or the differences between April and March as the most important estimating $PM_{2.5}$ concentration in early summer (MJJ) in Iraq.

6.3. *Prediction of $PM_{2.5}$ concentration*

The selected predictors were used in three ML models to forecast $PM_{2.5}$ concentration in MJJ. Data at all grid points over Iraq for 1998–2021 were merged, and then a 70:30 ratio was used for model

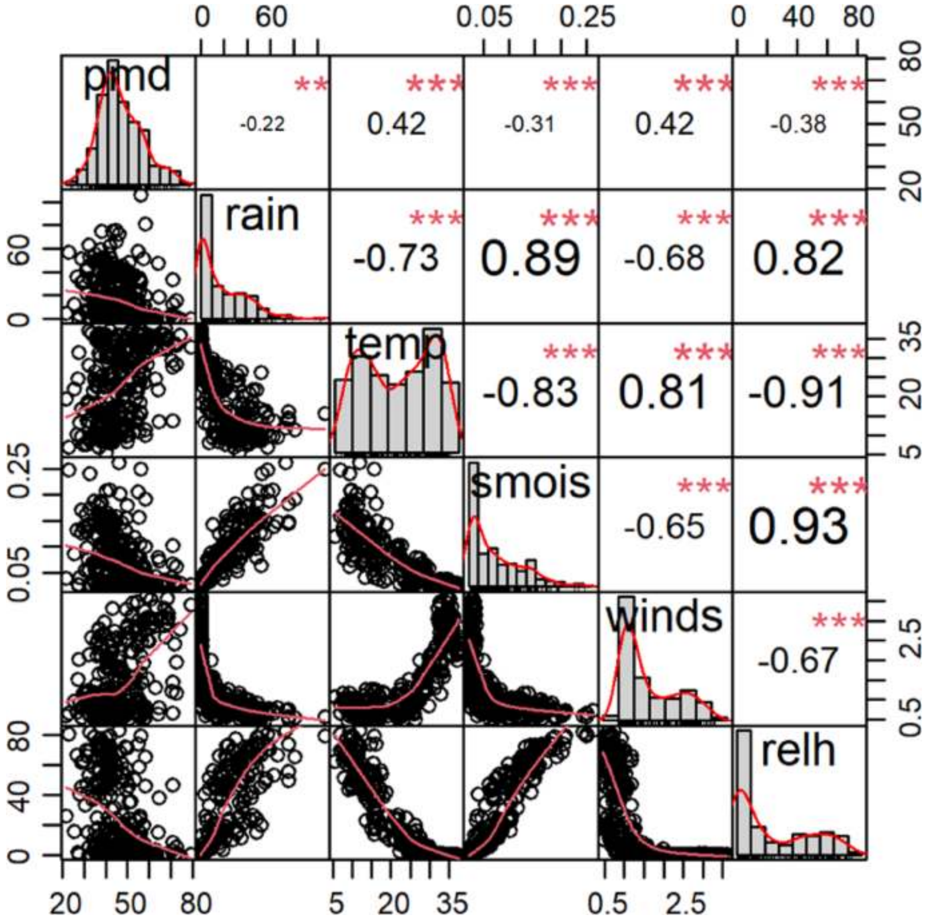

Fig. 7. Partial correlation of $PM_{2.5}$ with different predictors and their significance [pmd = PM2.5 concentration in μgm/m^3; rain = precipitation (mm); temp = mean temperature in °C; smois = ratio of water volume to total volume; winds = wind speed in m/s; relh = relative humidity in %].

Table 1
The set of preliminarily selected predictors and the final set of chosen predictors using SA.

Preliminary predictors	Final predictors
tm_{t-1}, tm_{t-2}, tm_{t-3}, $tm_{(t-1\ -\ t-2)}$, $tm_{(t-1\ -\ t-3)}$ ws_{t-1}, ws_{t-2}, ws_{t-3}, $ws_{(t-1\ -\ t-2)}$, $ws_{(t-1\ -\ t-3)}$ sm_{t-1}, sm_{t-2}, sm_{t-3}, $sm_{(t-1\ -\ t-2)}$, $sm_{(t-1\ -\ t-3)}$ rh_{t-1}, rh_{t-2}, rh_{t-3}, $rh_{(t-1\ -\ t-2)}$, $rh_{(t-1\ -\ t-3)}$	tm_{t-1}, $tm_{(t-1\ -\ t-2)}$, ws_{t-1}, $sm_{(t-1\ -\ t-3)}$, rh_{t-1}

*tm: mean temperature; ws: windspeed; sm: soil moisture; rh: relative humidity.

training and validation. The PM2.5 was predicted at 8384 grids covering entire Iraq. Five selected predictors for each of the grids were used to generate the input matrix of 5 × 8384 to predict the PM2.5 at 8384 grids. The Bayesian algorithm optimized the ML hyperparameters to get the best performance. The models' relative performance during the validation period was evaluated using statistical indices and comparison plots, as discussed in subsequent paragraphs.

The performance of the models with validation data based on six statistical metrics is shown in Fig. 8. The metrics were used to estimate the performance of the models at each grid point, and their values at all grid points were used to prepare the boxplots. The results showed better performance of LSTM in terms of all six metrics used. The mean RMSE, MAE, NRMSE (%), R2, MD and KGE of LSTM were 8.2 μgm/m^3, 5.8 μgm/m^3, 13.4%, 0.92, 0.9 and 0.89, compared to 10.3 μgm/m^3, 6.5 μgm/m^3, 16.02%, 0.86, 0.85 and 0.81 for SGD-BP and 12.1 μgm/m^3, 7.8 μgm/m^3, 17.9%, 0.85, 0.79 and 0.74 for ERT. The interquartile range (IQR) and complete range of the metrics for LSTM were much narrower compared to the other two models, indicating the better performance of LSTM at all locations. For example, the highest NRMSE of LSTM was 17.3% which was lower than the mean NRMSE of ERT. A similar result was noticed for KGE. It indicates a much better performance of LSTM than the other two models. The SGD-BP showed a better performance than ERT but less than LSTM. Therefore, the performance of the models can be ranked in order: LSTM, SGD-BP and ERT.

The observed and predicted $PM_{2.5}$ in different early summer months using three ML models are shown in Fig. 9. The predicted data at all locations were used to prepare the boxplots. The figure shows good accuracy in predicted $PM_{2.5}$ by all models in all months. However, ERT and SGD-BP over-predicted the $PM_{2.5}$ at some locations where they are less and under-predicted at some locations where they are high. In contrast, LSTM was more reliable in replicating the $PM_{2.5}$ values at all locations. The other two models also showed relatively less accuracy in different months. For example, ERT performed relatively poorly in July and SGD-BP in May. However, LSTM could reliably predict the mean, IQR, and complete data range $PM_{2.5}$ concentration for all months. The model also replicated the extreme values observed in June and July.

The performance evaluation of the model using density scatter plots is shown in Fig. 10. The density of the points is presented using color in the figure. The red color represents the higher density, while the blue the less density. The alignment of the points along the diagonal of the plot indicates a better prediction of the observed values. Fig. 10 shows a better prediction of observed $PM_{2.5}$ using LSTM. The LSMT predicted $PM_{2.5}$ data are densely concentrated along the diagonal of the plot, while those were little spread vertically for both ERT and SGD-BPANN. The vertically spreading pattern was more for ERT than SGD-BP. The results indicate the best performance using LSTM, followed by SGD-BP and ERT. The performance the models were also evaluated using Taylor

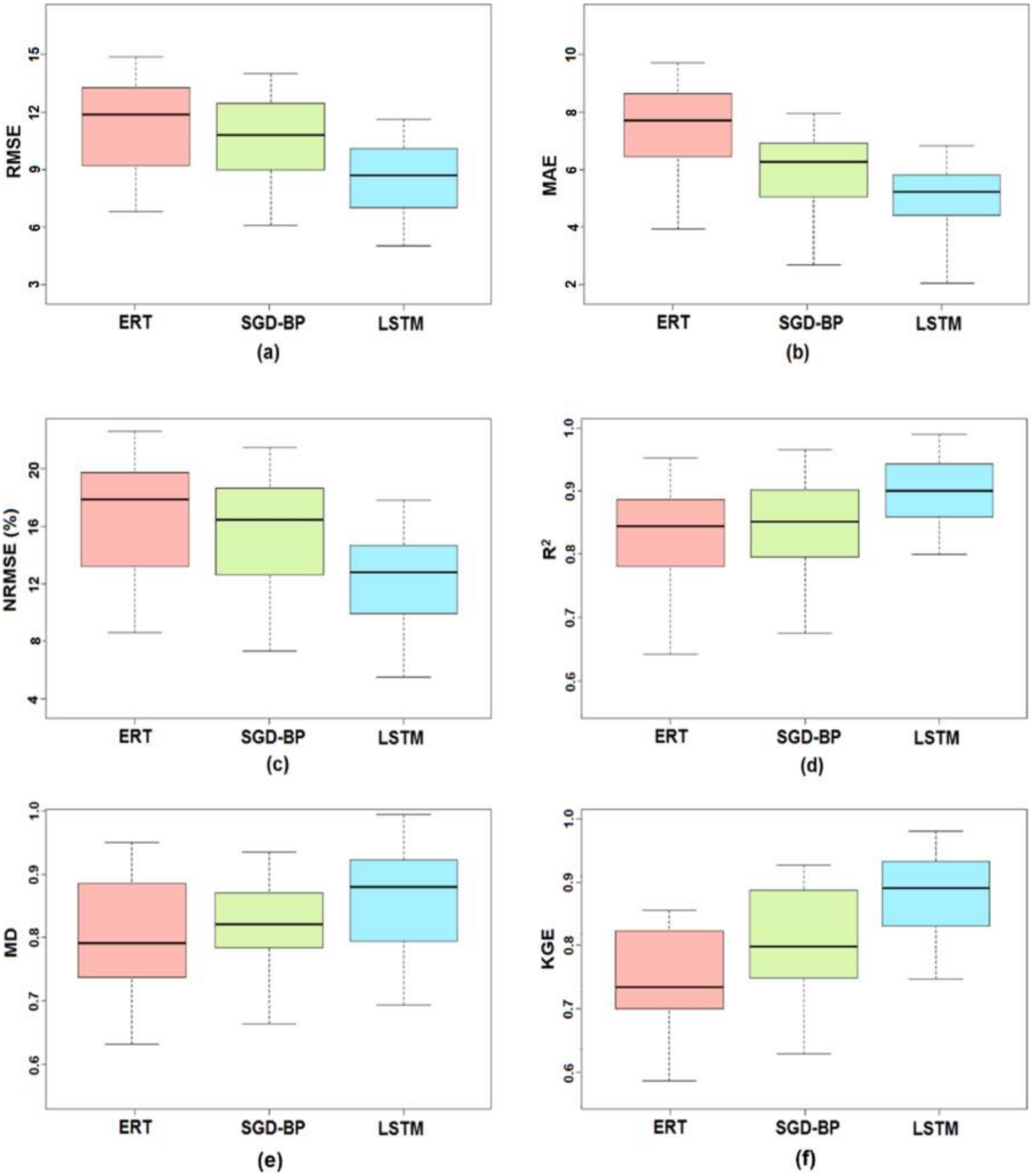

Fig. 8. Performance of the ML models in predicting $PM_{2.5}$ at different grid locations over Iraq during the validation period.

diagram, violin plot and radar chart. Obtained results are presented in supplementary Figures S3 to S5. Those figures also showed the best performance using LSTM, followed by SGD-BP and ERT.

The predicted $PM_{2.5}$ at different locations was used to prepare the spatial distribution of the $PM_{2.5}$ over Iraq for the validation period. The predicted maps using different ML models are shown in Fig. 11. The predicted maps were compared with the observed map presented in Fig. 3 to show the capability of the model to reconstruct the spatial pattern of $PM_{2.5}$. The figure shows that all models could reconstruct the observed spatial distribution of $PM_{2.5}$. All models could estimate the high $PM_{2.5}$ in the south and low values in the north and west. The patches of extremely high $PM_{2.5}$ values in the south and moderately high $PM_{2.5}$ patches in low-concentration regions in the central west were reliably estimated by all the models. However, the spatial extents of the patches estimated by different models differed. The LSTM replicated all low and high patterns accurately. The ERT and SGD-BP could not properly reconstruct the high $PM_{2.5}$ region in the south. The ERT also failed to replicate the moderate PM2.5 region in the central west reliably.

The capability of the models to estimate the spatial distribution of $PM_{2.5}$ was statistically evaluated using MapCurve, Cramer's V, spatial correlation, and mean spatial bias. Obtained results are presented in Table 2. The results showed the better capability of LSTM in reconstructing the spatial distribution of $PM_{2.5}$ during MJJ. The MapCurve, Cramer's V, and spatial correlation between the observed and the LSTM predicted $PM_{2.5}$ maps were 0.95, 0.91, and 0.97, indicating the nearly

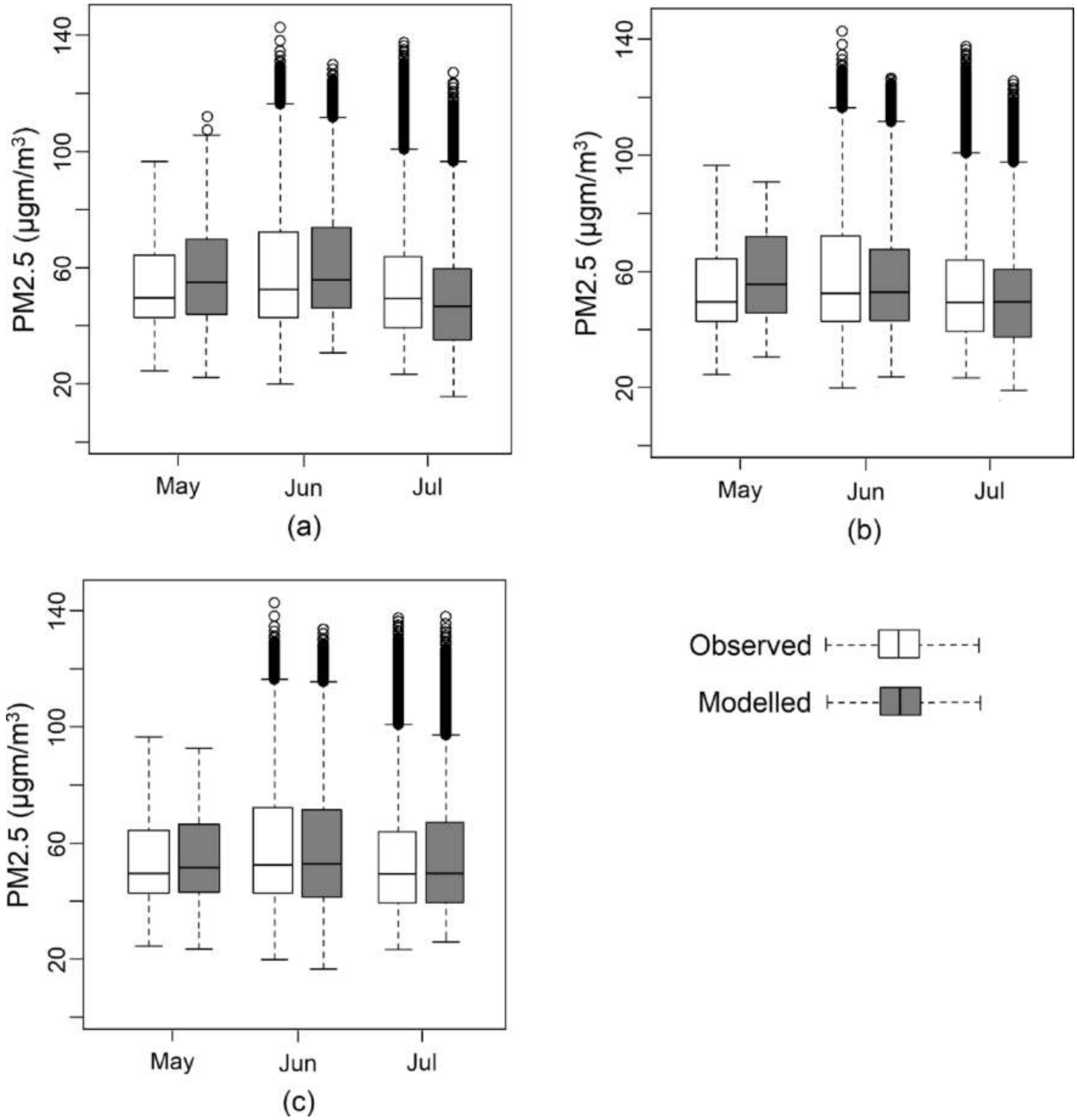

Fig. 9. The boxplots show the capability of (a) ERT, (b) SGD-BP, and (c) LSTM in reproducing observed $PM_{2.5}$ concentrations during May-July.

ideal performance of the model. The mean bias (%) in the LSTM $PM_{2.5}$ map was only 1.3% compared to 6.5% for SGD-BP and 8.2% for ERT.

The spatial distribution of the bias in estimated $PM_{2.5}$ by different models is shown in Fig. 12. The bias maps were generated to show the spatial distribution of the performance of the models. The red color in the map indicates positive bias or overestimation, the blue represents underestimation, and the white represents nearly zero bias. All the maps show the randomness in the spatial distribution of the bais. The bais (%) range over Iraq was only −3.2–6.7% for LSTM, compared to −12.5–14.1 for SGD-BP and −19.8–20.0% for ERT. The bias in LSTM predicted $PM_{2.5}$ was nearly zero at most locations in southern high $PM_{2.5}$ regions. However, it slightly overpredicted the $PM_{2.5}$ values in the low $PM_{2.5}$ regions in the far north and the west. However, over- and underprediction was much less than SGD-BP and ERT.

7. Discussion

Air pollution, specifically the concentration of $PM_{2.5}$, is a major public health concern worldwide. Accurate prediction of $PM_{2.5}$ levels is essential for effectively managing and mitigating its adverse health effects. In recent years, ML techniques have gained popularity in air quality prediction due to their ability to handle large datasets with high-dimensional features and capture complex nonlinear relationships between predictors and responses. Several studies have compared the performance of ML models with simple statistical models, like the linear regression model (Kim et al., 2022) and the generalized additive model (Li et al., 2017), in predicting $PM_{2.5}$ levels. However, it is worth noting that the performance of ML models may depend on the specific dataset and modeling techniques used and that proper validation and interpretation of the results for the practical application of these models. Therefore, the present study compared the performance of three advanced ML algorithms to find the most efficient one for predicting the spatiotemporal variability of summer $PM_{2.5}$ in Iraq. Several studies showed an increasing trend in $PM_{2.5}$ in Iraq (Boys et al., 2012; Coskuner et al., 2018; Shihab, 2021). All the studies also found that the $PM_{2.5}$ concentrations and increasing rates were higher during the summer months. Therefore, this study only attempted to predict the spatiotemporal distribution of $PM_{2.5}$ concentration over Iraq during the summer

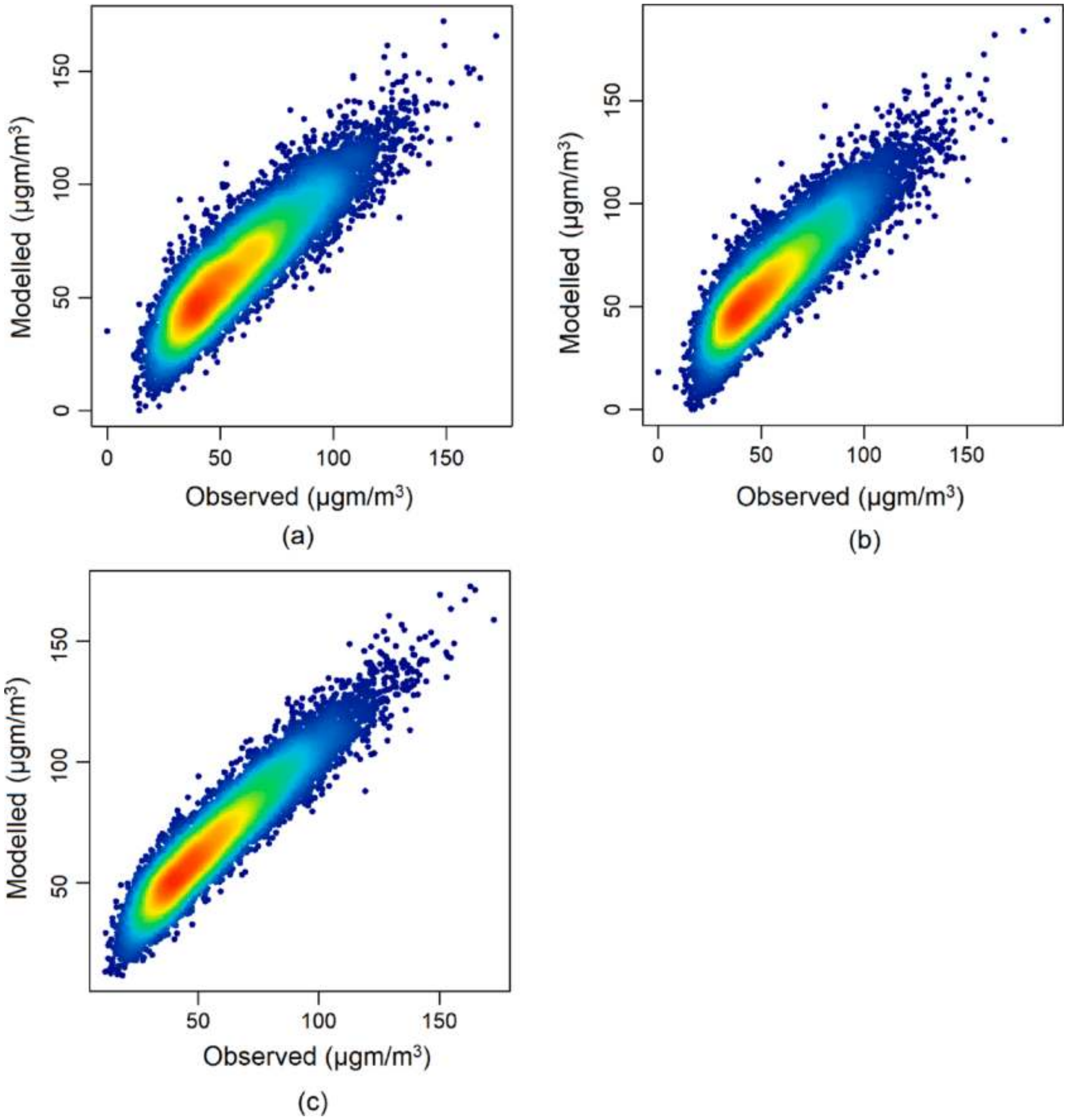

Fig. 10. The scatter plots of the observed and simulated $PM_{2.5}$ using (a) ERT, (b) SGD-BP, and (c) LSTM during the prediction period (May-July).

months.

The present study showed a higher capacity of LSTM in predicting $PM_{2.5}$ compared to the other two ML models considered in this study. This may be attributed to the ability of LSTM to capture temporal dependencies, handle highly nonlinear relationships, and scalability (Gers et al., 2000; Hochreiter and Schmidhuber, 1997). LSTM models are designed to capture long-term dependencies and patterns in sequential data, making them well-suited for analyzing time series data, such as $PM_{2.5}$ pollution levels. Unlike other ML models, LSTM models have a memory component that enables them to capture information from previous time steps and incorporate it into the current prediction, improving their accuracy (Yu et al., 2019). Various factors influence $PM_{2.5}$ pollution levels, and the relationships between these factors and $PM_{2.5}$ pollution levels can be complex and nonlinear. LSTM models can learn these complex relationships and capture nonlinear dependencies between the input features and the target variable. LSTM models can be trained on large datasets with many input features and handle data with varying sampling frequencies (Li et al., 2021). This makes them well-suited for analyzing $PM_{2.5}$ pollution data, which can have several input features. Several studies have shown that LSTM models outperform other ML models (Niu et al., 2016). Several studies have applied LSTM to predict $PM_{2.5}$ concentrations in different regions. The studies showed varying performances of the model in different regions (Karimian et al., 2019; Yu et al., 2022). The variation in the performance of these models is due to various factors such as data quality, feature selection, and modeling parameter optimization. This study showed a better performance of the LSTM compared to its application in other studies. The present study showed that LSTM could forecast the monthly mean $PM_{2.5}$ concentration during summer months with an RMSE of 8.2 µgm/m^3, MAE of 5.8 µgm/m^3 and R^2 of 0.92. The results showed better performance of the model in predicting $PM_{2.5}$ in this study. This may be mainly due to the selection of input features and optimization of model parameters.

The present study's novelty is using the changing pattern of different meteorological and soil parameters and their time lags for the prediction. The selected predictors using SA revealed the changes in temperature and soil moisture, along with wind speed and relative humidity before the beginning of summer, can predict the monthly variability and spatial distribution of $PM_{2.5}$ during the early summer months. The rising temperature in the spring causes a gradual decrease in soil moisture. The

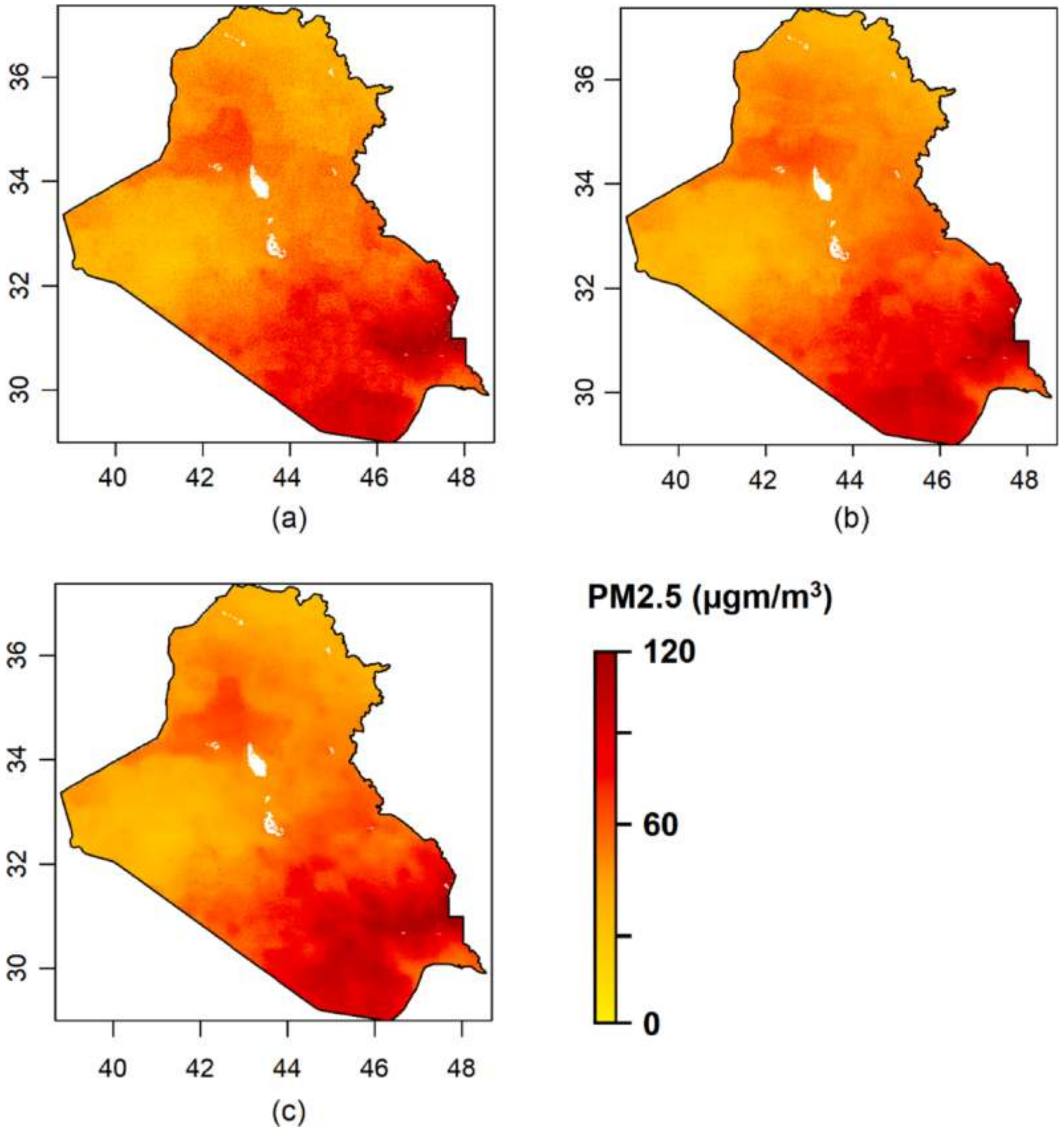

Fig. 11. The spatial distribution of the simulated $PM_{2.5}$ (May-July)using (a) ERT, (b) SGD-BP, and (c) LSTM during the validation period.

Table 2
Performance of the models in reconstructing the spatial pattern of observed $PM_{2.5}$ during early summer.

Metrics	ERT	SGD-BP	LSTM
MapCurve	0.83	0.90	0.95
Cramer's V	0.76	0.86	0.91
Spatial Correlation	0.88	0.93	0.97
Bias (%)	8.2	6.5	1.3

pollution level during the summer is high when a high temperature rises during spring, causing rapid soil drying. The higher wind and less humidity help in transmitting the fine particulates. Less soil moisture indicates dry soil and, thus, more potential as a source of dust pollution. Therefore, selected predictors by SA are physically justifiable. The difference in predictor values between two-time lags defines the direction and magnitude of the change in the predictor, indicating a possible direction and intensity of change in pollution in the future. For example, if the difference in soil moisture between two consecutive months is negative, it indicates a possible drying of soil with time. A higher magnitude of the difference indicates a rapid soil drying, possibly implying more air pollution from dust. Including this new feature as input considerably improves the performance of the models.

This study used meteorological parameters of ERA5 of the preceding months to predict the spatial distribution of $PM_{2.5}$ pollution in the summer months. The ERA5 data is updated daily with a latency of about 5 days. Therefore, it is possible to use those data to predict seasonal $PM_{2.5}$ levels five days after the beginning of the season. Climate reanalysis datasets of ERA5 can provide valuable information on weather patterns and atmospheric conditions, which can be useful for predicting air quality levels in a region. To generate a comprehensive picture of the climate system, these datasets use advanced modeling techniques to assimilate data from various sources, including satellite observations, ground-based measurements, and weather models. Recent studies (Al-Hasani and Shahid, 2022; Karami et al., 2022), showed the reliability of ERA5 data in estimating the meteorological variables of Iraq. Therefore, it can be expected that the model developed in this study using ERA5 data can be used for reliable forecasting of the spatial distribution of seasonal $PM_{2.5}$ concentration over Iraq.

Several studies in China predicted daily or hourly $PM_{2.5}$ concentrations at a resolution of 0.01° using meteorological variables as predictors (Dong et al., 2022; Li et al., 2021). However, such a high-resolution (0.01° resolution) prediction of $PM_{2.5}$ is not possible in many other regions due to the unavailability of high-resolution meteorological

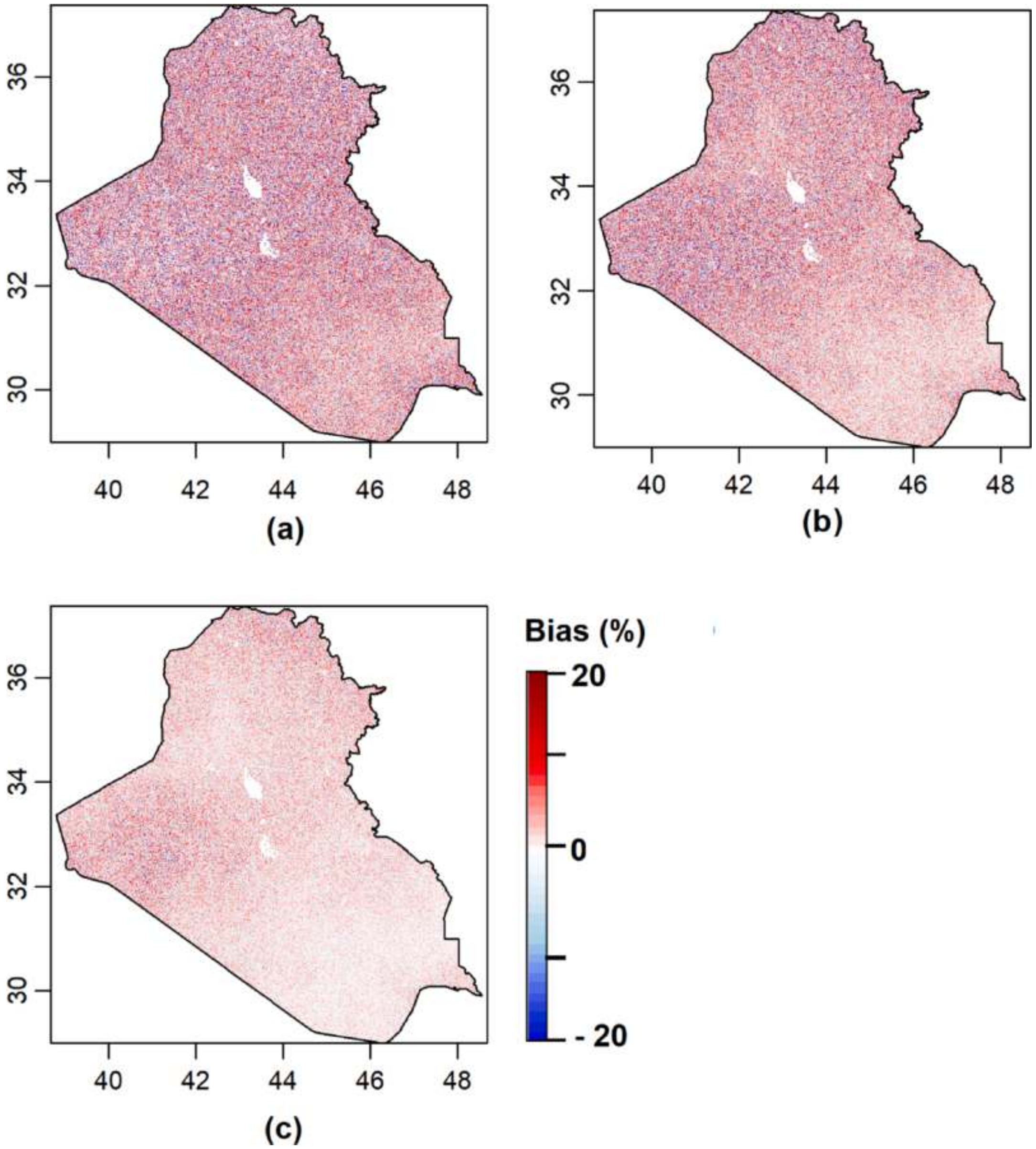

Fig. 12. The spatial distribution of the bias (%) in simulated $PM_{2.5}$ using (a) ERT, (b) SGD-BP, and (c) LSTM during the prediction period (May-July).

variables. Only MODIS high-resolution (0.01°) air pollution and temperature data are available daily at higher resolution. However, the only air temperature was insufficient for predicting $PM_{2.5}$ in Iraq as other variables like wind, relative humidity and soil moisture significantly affect $PM_{2.5}$ concentration in Iraq. Therefore, the prediction of $PM_{2.5}$ concentration in the study area was limited to 0.1 resolution considering the availability of quality meteorological and soil data at this resolution. For this purpose, 0.01° resolution $PM_{2.5}$ data were aggregated to generate 0.1° resolution $PM_{2.5}$ concentration. Here, it should be noted that uncertainty in estimation is added during downscaling but not upscaling. Therefore, for comparison or model development, lower-resolution data are always upscaled to coarser resolution (Salman et al., 2022). Similarly, it was not possible to develop a model for forecasting $PM_{2.5}$ at daily or hourly scale due to the unavailability of high-resolution daily and hourly $PM_{2.5}$ concentration data for Iraq. Such studies can be conducted in the future when high spatial and temporal resolution pollution data are available.

8. Conclusions

ML models have been used to predict $PM_{2.5}$ concentration in early summer over Iraq to provide an early warning of the possible spatial and temporal pattern of air $PM_{2.5}$ during this high pollution period in the country. The performance of ML algorithms significantly depends on the inputs used. The present study's novelty is using the changing pattern of different meteorological and soil parameters and their time lags for the prediction. The selected predictors using SA revealed the changes in temperature and soil moisture, along with wind speed and relative humidity before the beginning of summer, can predict the monthly variability and spatial distribution of $PM_{2.5}$ during the early summer months.

The LSTM model developed in the study efficiently simulates the $PM_{2.5}$ distribution in Iraq with high accuracy. The LSTM reconstructed the high-resolution $PM_{2.5}$ distribution map of Iraq during the validation period with high accuracy. It could replicate the high $PM_{2.5}$ zones in the south and low $PM_{2.5}$ zones in the north with negligible bias. It also predicted the observed high pollution in July compared to other models. The results indicate the suitability of the LSTM model for air $PM_{2.5}$

concentration forecasting in Iraq. The capability of LSMT to store the prediction in the preceding step helped it to provide reliable predictions for the long term. Therefore, it was more capable of predicting $PM_{2.5}$ in the early summer months with the previous season's meteorological and soil moisture data.

The study used freely available ERA5 meteorological and soil parameters to predict high-resolution $PM_{2.5}$ distribution data, which are also available in the public domain. The data can be used for any other region for developing a model following the procedure discussed in this paper for pollution prediction and early warning. In the future, other advanced optimization algorithms can be integrated with ML models to improve the predictability of the ML models. Other earth's surface features, such as vegetation index and normalized water index, can be considered for selecting a wide range of predictors for better prediction of $PM_{2.5}$.

CRediT authorship contribution statement

Hai Tao: Conceptualization, Formal analysis, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. **Ali H. Jawad:** Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **A.H. Shather:** Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Zainab Al-Khafaji:** Data curation, Formal analysis, Validation, Visualization, Writing – original draft, Writing – review & editing. **Tarik A. Rashid:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Mumtaz Ali:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Nadhir Al-Ansari:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Haydar Abdulameer Marhoon:** Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Shamsuddin Shahid:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Zaher Mundher Yaseen:** Project administration, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors appreciate the respected reviewers and handling editors for their constructive comments. In addition, the fiirst author acknowledge the support received by the National Natural Science Foundation of China (No.61862051), the Science and Technology Foundation of Guizhou Province (No.[2019]1299, No.ZK[2022]449), the Top-notch Talent Program of Guizhou province (No.KY[2018]080), the Natural Science Foundation of Education of Guizhou province (No.[2019]203) and the Funds of Qiannan Normal University for Nationalities (No. qnsy2019rc09). Finally, an appreciation is declared to Al-Mustaqbal University College (Grant number: MUC-E-0122).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2023.107931.

References

Abd Alraheem, E., Jaber, N.A., Jamei, M., Tangang, F., 2022. Assessment of future meteorological drought under representative concentration pathways (RCP8. 5) scenario: case study of Iraq. Knowledge-Based Eng. Sci. 3, 64–82.

Afshar, F., Seyedabrishami, S., Moridpour, S., 2022. Application of Extremely Randomised Trees for exploring influential factors on variant crash severity data. Sci. Rep. 12, 11476. https://doi.org/10.1038/s41598-022-15693-7.

Al-Aseel, Z., 2022. Will Iraq overcome its pollution crisis? [WWW Document]. URL https://amwaj.media/article/pollution-a-danger-haunting-the-lives-of-iraqis.

Al-Hasani, A.A.J., Shahid, S., 2022. Spatial distribution of the trends in potential evapotranspiration and its influencing climatic factors in Iraq. Theor. Appl. Climatol. 150, 677–696.

Al-Kasser, M.K., 2021. Air Pollution in Iraq Sources and Effects, in: IOP Conference Series: Earth and Environmental Science. IOP Publishing, p. 12014.

Amari, S., 1993. Backpropagation and stochastic gradient descent method. Neurocomputing 5, 185–196. https://doi.org/10.1016/0925-2312(93)90006-o.

Bacanin, N., Sarac, M., Budimirovic, N., Zivkovic, M., AlZubi, A.A., Bashir, A.K., 2022. Smart wireless health care system using graph LSTM pollution prediction and dragonfly node localization. Sustain. Comput. Informatics Syst. 35, 100711 https://doi.org/10.1016/j.suscom.2022.100711.

Bagheri, H., 2022. A machine learning-based framework for high resolution mapping of PM2.5 in Tehran, Iran, using MAIAC AOD data. Adv. Sp. Res. 69, 3333–3349. https://doi.org/10.1016/j.asr.2022.02.032.

Bai, Y., Li, Y., Wang, X., Xie, J., Li, C., 2016. Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. Atmos. Pollut. Res. 7, 557–566. https://doi.org/10.1016/j.apr.2016.01.004.

Bhagat, S.K., Tiyasha, T., Tung, T.M., Mostafa, R.R., Yaseen, Z.M., 2020. Manganese (Mn) removal prediction using extreme gradient model. Ecotoxicol. Environ. Saf. 204, 111059 https://doi.org/10.1016/j.ecoenv.2020.111059.

Boys, B., Martin, R. V, van Donkelaar, A., MacDonell, R., Hsu, N.C., 2012. Time series analysis of global surface PM2. 5 from remote-sensed aerosol optical depth, in: AGU Fall Meeting Abstracts. pp. A24C-08.

Burney, J., Ramanathan, V., 2014. Recent climate and air pollution impacts on Indian agriculture. PNAS 111, 16319–16324. https://doi.org/10.1073/pnas.1317275111.

Butt, E.W., Turnock, S.T., Rigby, R., Reddington, C.L., Yoshioka, M., Johnson, J.S., Regayre, L.A., Pringle, K.J., Mann, G.W., Spracklen, D.V., 2017. Global and regional trends in particulate air pollution and attributable health burden over the past 50 years. Environ. Res. Lett. 12, 104017 https://doi.org/10.1088/1748-9326/aa87be.

Casallas, A., Ferro, C., Celis, N., Guevara-Luna, M.A., Mogollón-Sotelo, C., Guevara-Luna, F.A., Merchán, M., 2021. Long short-term memory artificial neural network approach to forecast meteorology and PM2.5 local variables in Bogotá, Colombia. Model. Earth Syst. Environ. https://doi.org/10.1007/s40808-021-01274-6.

Castillo, M.D., Kinney, P.L., Southerland, V., Arno, C.A., Crawford, K., van Donkelaar, A., Hammer, M., Martin, R. V, Anenberg, S.C., 2021. Estimating Intra-Urban Inequities in PM2. 5-Attributable Health Impacts: A Case Study for Washington, DC. GeoHealth 5, e2021GH000431.

Chang, F.-J., Chang, L.-C., Kang, C.-C., Wang, Y.-S., Huang, A., 2020. Explore spatio-temporal PM2.5 features in northern Taiwan using machine learning techniques. Sci. Total Environ. 736, 139656 https://doi.org/10.1016/j.scitotenv.2020.139656.

Chen, S., Zhang, D., 2021. Impact of air pollution on labor productivity: Evidence from prison factory data. China Econ. Q. Int. 1, 148–159. https://doi.org/10.1016/j.ceqi.2021.04.004.

Cobourn, W.G., 2010. An enhanced PM2.5 air quality forecast model based on nonlinear regression and back-trajectory concentrations. Atmos. Environ. 44, 3015–3023. https://doi.org/10.1016/j.atmosenv.2010.05.009.

Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., Pope 3rd, C.A., Shin, H., Straif, K., Shaddick, G., Thomas, M., van Dingenen, R., van Donkelaar, A., Vos, T., Murray, C.J.L., Forouzanfar, M.H., 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. Lancet (London, England) 389, 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6.

Coskuner, G., Jassim, M.S., Munir, S., 2018. Characterizing temporal variability of PM2. 5/PM10 ratio and its relationship with meteorological parameters in Bahrain. Environ. Forensic 19, 315–326.

Cramér, H., 1946. Mathematical Methods of Statistics (PMS-9) Princeton University Press. Princeton, NJ, USA.

Cui, F., Al-Sudani, Z.A., Hassan, G.S., Afan, H.A., Ahammed, S.J., Yaseen, Z.M., 2022. Boosted artificial intelligence model using improved alpha-guided grey wolf optimizer for groundwater level prediction: Comparative study and insight for federated learning technology. J. Hydrol. 606, 127384 https://doi.org/10.1016/j.jhydrol.2021.127384.

Danesh Yazdi, M., Kuang, Z., Dimakopoulou, K., Barratt, B., Suel, E., Amini, H., Lyapustin, A., Katsouyanni, K., Schwartz, J., 2020. Predicting fine particulate matter (PM2. 5) in the greater london area: An ensemble approach using machine learning methods. Remote Sens. 12, 914.

De Nevers, N., 2010. Air pollution control engineering. Waveland press.

Dewancker, I., McCourt, M., Clark, S., 2016. Bayesian optimization for machine learning: A practical guidebook. arXiv Prepr. arXiv1612.04858.

Dong, L., Hua, P., Gui, D., Zhang, J., 2022. Extraction of multi-scale features enhances the deep learning-based daily PM2. 5 forecasting in cities. Chemosphere 308, 136252.

Elbayoumi, M., Ramli, N.A., Yusof, N.F.F.M., Al Madhoun, W., 2013. Spatial and seasonal variation of particulate matter (PM10 and PM2. 5) in Middle Eastern classrooms. Atmos. Environ. 80, 389–397.

Elminir, H.K., 2005. Dependence of urban air pollutants on meteorology. Sci. Total Environ. 350, 225–237. https://doi.org/10.1016/j.scitotenv.2005.01.043.

Eren, B., Aksangür, İ., Erden, C., 2023. Predicting next hour fine particulate matter (PM2. 5) in the Istanbul Metropolitan City using deep learning algorithms with time windowing strategy. Urban Clim. 48, 101418.

Ernst, O.K., 2014. Stochastic gradient descent learning and the backpropagation algorithm. Univ. California, San Diego, La Jolla, CA, Tech. Rep.

Faskari, S.A., Ojim, G., Falope, T., Abdullahi, Y.B., Abba, S.I., 2022. A Novel Machine Learning based Computing Algorithm in Modeling of Soiled Photovoltaic Module. Knowledge-Based Eng. Sci. 3, 28–36.

Fowler, D., Pyle, J.A., Sutton, M.A., Williams, M.L., 2020. Global Air Quality, past present and future: an introduction. Philos. Trans. A. Math. Phys. Eng. Sci. 378, 20190323. https://doi.org/10.1098/rsta.2019.0323.

Fu, M., Le, C., Fan, T., Prakapovich, R., Manko, D., Dmytrenko, O., Lande, D., Shahid, S., Yaseen, Z.M., 2021. Integration of complete ensemble empirical mode decomposition with deep long short-term memory model for particulate matter concentration prediction. Environ. Sci. Pollut. Res. 1–12.

Gao, H., Zhong, S., Zhang, W., Igou, T., Berger, E., Reid, E., Zhao, Y., Lambeth, D., Gan, L., Afolabi, M.A., Tong, Z., Lan, G., Chen, Y., 2021. Revolutionizing membrane design using machine learning-Bayesian optimization. Environ. Sci. & Technol. 56, 2572–2581. https://doi.org/10.1021/acs.est.1c04373.

Gers, F.A., Schmidhuber, J., Cummins, F., 2000. Learning to Forget: Continual Prediction with LSTM. Neural Comput. 12, 2451–2471. https://doi.org/10.1162/089976600300015015.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42.

Hargrove, W.W., Hoffman, F.M., Hessburg, P.F., 2006. Mapcurves: a quantitative method for comparing categorical maps. J. Geogr. Syst. 8, 187–208. https://doi.org/10.1007/s10109-006-0025-x.

Hashim, B.M., Al-Naseri, S.K., Al Maliki, A., Sa'adi, Z., Malik, A., Yaseen, Z.M., 2021. On the investigation of COVID-19 lockdown influence on air pollution concentration: regional investigation over eighteen provinces in Iraq. Environ. Sci. Pollut. Res. 1–19.

He, J., Gong, S., Yu, Y., Yu, L., Wu, L., Mao, H., Song, C., Zhao, S., Liu, H., Li, X., Li, R., 2017. Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major Chinese cities. Environ. Pollut. 223, 484–496. https://doi.org/10.1016/j.envpol.2017.01.050.

He, W., Meng, H., Han, J., Zhou, G., Zheng, H., Zhang, S., 2022. Spatiotemporal PM2. 5 estimations in China from 2015 to 2020 using an improved gradient boosting decision tree. Chemosphere 296, 134003.

Heger, M., Vashold, L., Palacios, A., Alahmadi, M., Acerbi, M., 2022. Blue Skies, Blue Seas: Air Pollution, Marine Plastics, and Coastal Erosion in the Middle East and North Africa. World Bank Publications.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049.

Hochreiter, S., Schmidhuber, J.J., 1997. Long short-term memory. Neural Comput. 9, 1–32. https://doi.org/10.1162/neco.1997.9.8.1735.

Hu, J., Chen, Y., Wang, W., Zhang, S., Cui, C., Ding, W., Fang, Y., 2023. An optimized hybrid deep learning model for PM2. 5 and O3 concentration prediction. Air Qual. Atmos. Heal, 1–15.

Hunt, K.M.R., Matthews, G.R., Pappenberger, F., Prudhomme, C., 2022. Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States doi: 10.5194/hess-2022-53.

Ibrahim, S., Landa, M., Pešek, O., Brodský, L., Halounová, L., 2022. Machine Learning-Based Approach Using Open Data to Estimate PM2.5 over Europe. Remote Sens. 14, 3392. https://doi.org/10.3390/rs14143392.

Jamei, M., Ahmadianfar, I., Karbasi, M., Jawad, A.H., Farooque, A.A., Yaseen, Z.M., 2021. The assessment of emerging data-intelligence technologies for modeling Mg+ 2 and SO4− 2 surface water quality. J. Environ. Manage. 300, 113774.

Jamei, M., Ali, M., Malik, A., Karbasi, M., Sharma, E., Yaseen, Z.M., 2022. Air quality monitoring based on chemical and meteorological drivers: Application of a novel data filtering-based hybridized deep learning model. J. Clean. Prod. 374, 134011.

Jamei, M., Ali, M., Malik, A., Karbasi, M., Rai, P., Yaseen, Z.M., 2023a. Development of a TVF-EMD-based multi-decomposition technique integrated with encoder-decoder-bidirectional-LSTM for monthly rainfall forecasting. J. Hydrol. 129105.

Jamei, M., Karbasi, M., Ali, M., Malik, A., Chu, X., Yaseen, Z.M., 2023b. A novel global solar exposure forecast model based on air temperature: Designing a new multi-processing ensemble deep learning paradigm. Expert Syst, Appl, p. 119811.

Jaradat, A., 2003. Agriculture in Iraq: Resources, potentials, constraints, research needs and priorities. Agriculture.

Jiang, M., Sun, W., Yang, G., Zhang, D., 2017. Modelling seasonal GWR of daily PM2.5 with proper auxiliary variables for the Yangtze River Delta. Remote Sens. 9, 346. https://doi.org/10.3390/rs9040346.

Jiang, M., Kim, E., Woo, Y., 2020. The relationship between economic growth and air pollution-a regional comparison between China and South Korea. Int. J. Environ. Res. Public Health 17, 2761. https://doi.org/10.3390/ijerph17082761.

Kanabkaew, T., 2013. Prediction of Hourly Particulate Matter Concentrations in Chiangmai, Thailand Using MODIS Aerosol Optical Depth and Ground-Based Meteorological Data. EnvironmentAsia 6.

Karami, S., Ghassabi, Z., Rezazadeh, P., 2022. Investigating the mechanism of dust transferring from Iraq to the north of Alborz mountains in Iran. J. Air Pollut. Heal. 7, 375–398.

Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., Zhang, X., Sachdeva, S., 2019. Evaluation of different machine learning approaches to forecasting PM2.5 mass concentrations. Aerosol Air Qual. Res. 19, 1400–1410. https://doi.org/10.4209/aaqr.2018.12.0450.

Katongtung, T., Onsree, T., Tippayawong, N., 2022. Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes. Bioresour. Technol. 344, 126278 https://doi.org/10.1016/j.biortech.2021.126278.

Kim, B.-Y., Lim, Y.-K., Cha, J.W., 2022. Short-term prediction of particulate matter (PM10 and PM2. 5) in Seoul, South Korea using tree-based machine learning algorithms. Atmos. Pollut. Res. 13, 101547.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 80. https://doi.org/10.1126/science.220.4598.671.

Kjellström, T., Maître, N., Saget, C., Otto, M., Karimova, T., 2019. Working on a warmer planet: The effect of heat stress on productivity and decent work. International Labour Organization.

Lanzi, E., 2016. The economic consequences of outdoor air pollution. Organization for Economic Cooperation and Development.

Lelieveld, J., Evans, J.S., Fnais, M., Giannadaki, D., Pozzer, A., 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature 525, 367–371. https://doi.org/10.1038/nature15371.

Li, J., Garshick, E., Hart, J.E., Li, L., Shi, L., Al-Hemoud, A., Huang, S., Koutrakis, P., 2021. Estimation of ambient PM2.5 in Iraq and Kuwait from 2001 to 2018 using machine learning and remote sensing. Environ. Int. https://doi.org/10.1016/j.envint.2021.106445.

Li, Y., Lin, T.-Y., Chiu, Y.-H., 2020. Dynamic linkages among economic development, environmental pollution and human health in Chinese. Cost Eff. Resour. Alloc. 18, 32. https://doi.org/10.1186/s12962-020-00228-6.

Li, R., Wang, Z., Cui, L., Fu, H., Zhang, L., Kong, L., Chen, W., Chen, J., 2019. Air pollution characteristics in China during 2015–2016: Spatiotemporal variations and key meteorological factors. Sci. Total Environ. 648, 902–915. https://doi.org/10.1016/j.scitotenv.2018.08.181.

Li, S., Zhai, L., Zou, B., Sang, H., Fang, X., 2017. A generalized additive model combining principal component analysis for PM2. 5 concentration estimation. ISPRS Int. J. Geo-Information 6, 248.

Liu, H., Cui, W., Zhang, M., 2022. Exploring the causal relationship between urbanization and air pollution: evidence from China. Sustain. Cities Soc. 80, 103783 https://doi.org/10.1016/j.scs.2022.103783.

Mokoena, K.K., Ethan, C.J., Yu, Y., Quachie, A.T., 2020. Interaction effects of air pollution and climatic factors on circulatory and respiratory mortality in Xi'an, China between 2014 and 2016. Int. J. Environ. Res. Public Health 17, 9027. https://doi.org/10.3390/ijerph17239027.

Mujtaba, G., Shahzad, S.J.H., 2021. Air pollutants, economic growth and public health: implications for sustainable development in OECD countries. Environ. Sci. Pollut. Res. Int. 28, 12686–12698. https://doi.org/10.1007/s11356-020-11212-1.

Myllyvirta, L., 2020. Quantifying the economic costs of air pollution from fossil fuels key messages.

Niu, M., Wang, Y., Sun, S., Li, Y., 2016. A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2.5 concentration forecasting. Atmos. Environ. https://doi.org/10.1016/j.atmosenv.2016.03.056.

Niu, M., Zhang, Y., Ren, Z., 2023. Deep learning-based PM2. 5 long time-series prediction by fusing multisource data—a case study of Beijing. Atmosphere (Basel) 14, 340.

Padmaja, B., Prasa, V., Sunitha, K., 2020. A novel random split point procedure using extremely randomized (Extra) trees ensemble method for human activity recognition. EAI Endorsed Trans. Pervasive Heal. Technol. 6, 164824 https://doi.org/10.4108/eai.28-5-2020.164824.

Pai, S.J., Carter, T.S., Heald, C.L., Kroll, J.H., 2022. Updated world health organization air quality guidelines highlight the importance of non-anthropogenic PM2. Sci. Technol. Lett.

Peng, J., Han, H., Yi, Y., Huang, H., Xie, L., 2022. Machine learning and deep learning modeling and simulation for predicting PM2. 5 concentrations. Chemosphere 136353.

Pérez, P., Trier, A., Reyes, J., 2000. Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago. Chile. Atmos. Environ. 34, 1189–1196. https://doi.org/10.1016/s1352-2310(99)00316-7.

Pruthi, D., Liu, Y., 2022. Low-cost nature-inspired deep learning system for PM2.5 forecast over Delhi, India. Environ. Int. 166, 107373 https://doi.org/10.1016/j.envint.2022.107373.

Qi, C., Zhou, W., Lu, X., Luo, H., Pham, B.T., Yaseen, Z.M., 2020. Particulate matter concentration from open-cut coal mines: A hybrid machine learning estimation. Environ. Pollut. https://doi.org/10.1016/j.envpol.2020.114517.

Rollin, O., Aguirre-Gutiérrez, J., Yasrebi-de Kom, I.A.R., Garratt, M.P.D., de Groot, G.A., Kleijn, D., Potts, S.G., Scheper, J., Carvalheiro, L.G., 2022. Effects of ozone air pollution on crop pollinators and pollination. Glob. Environ. Chang. 75, 102529 https://doi.org/10.1016/j.gloenvcha.2022.102529.

Saad, N., 2021. Air Quality in Arab Countries: An Overview. Environment, Middle East, Pollution [WWW Document]. EcoMENA. URL https://www.ecomena.org/air-quality-arab/ (accessed 2.21.22).

Sachdeva, S., Kumar, B., 2022. Flood susceptibility mapping using extremely randomized trees for Assam 2020 floods. Ecol. Inform. 67, 101498 https://doi.org/10.1016/j.ecoinf.2021.101498.

Salman, S.A., Hamed, M.M., Shahid, S., Ahmed, K., Sharafati, A., Asaduzzaman, M., Ziarh, G.F., Ismail, T., Chung, E., Wang, X., 2022. Projecting spatiotemporal changes of precipitation and temperature in Iraq for different shared socioeconomic pathways with selected Coupled Model Intercomparison Project Phase 6. Int. J. Climatol..

Shihab, A.S., 2021. Assessment of ambient air quality of Mosul city/Iraq via Air Quality Index. J. Ecol. Eng. 22.

Snoek, J., Larochelle, H., Adams, R.P., 2012. Practical bayesian optimization of machine learning algorithms. Adv. Neural Inf. Process. Syst. 25.

Sorek-Hamer, M., Strawa, A.W., Chatfield, R.B., Esswein, R., Cohen, A., Broday, D.M., 2013. Improved retrieval of PM2.5 from satellite data products using non-linear methods. Environ. Pollut. 182, 417–423. https://doi.org/10.1016/j.envpol.2013.08.002.

Van Houdt, G., Mosquera, C., Nápoles, G., 2020. A review on the long short-term memory model. Artif. Intell. Rev. 53, 5929–5955. https://doi.org/10.1007/s10462-020-09838-1.

Ventura, L.M.B., de Oliveira Pinto, F., Soares, L.M., Luna, A.S., Gioda, A., 2019. Forecast of daily PM2.5 concentrations applying artificial neural networks and Holt-Winters models. Air Qual. Atmos. & Heal. 12, 317–325. https://doi.org/10.1007/s11869-018-00660-x.

Wang, W., Guo, Y., 2009. Air Pollution PM2.5 Data Analysis in Los Angeles Long Beach with Seasonal ARIMA Model. 2009 Int. Conf. Energy Environ. Technol doi: 10.1109/iceet.2009.468.

Wang, X., Dickinson, R.E., Su, L., Zhou, C., Wang, K., 2018. PM2.5 pollution in China and how it has been exacerbated by terrain and meteorological conditions. Bull. Am. Meteorol. Soc. 99, 105–119. https://doi.org/10.1175/bams-d-16-0301.1.

Weiqi, K., Weisong, W., Maoxing, Z., 2022. Integrated learning algorithms with Bayesian optimization for mild steel mechanical properties prediction. Knowledge-Based Eng. Sci. 3, 101–112.

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. Atmos. Pollut. Res. 13, 101352.

World Bank Group, IHME, 2016. The cost of air pollution: Strengthening the Economic Case for Action, The World Bank and Institute for Health Metrics and Evaluation University of Washington, Seattle doi: 10.1080/000368497326688.

Wu, D.-L., Lin, M., Chan, C.-Y., Li, W.-Z., Tao, J., Li, Y.-P., Sang, X.-F., Bu, C.-W., 2013. Influences of Commuting Mode, Air Conditioning Mode and Meteorological Parameters on Fine Particle (PM2.5) Exposure Levels in Traffic Microenvironments. Aerosol Air Qual. Res. 13, 709–720. https://doi.org/10.4209/aaqr.2012.08.0212.

Wu, Y., Lin, S., Shi, K., Ye, Z., Fang, Y., 2022. Seasonal prediction of daily PM2.5 concentrations with interpretable machine learning: a case study of Beijing. China. Environ. Sci. Pollut. Res. 29, 45821–45836. https://doi.org/10.1007/s11356-022-18913-9.

Xiao, C., Chen, N., Hu, C., Wang, K., Gong, J., Chen, Z., 2019. Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach. Remote Sens. Environ. 233 https://doi.org/10.1016/j.rse.2019.111358.

Xiao, F., Yang, M., Fan, H., Fan, G., Al-Qaness, M.A.A., 2020. An improved deep learning model for predicting daily PM2.5 concentration. Sci. Rep. 10, 20988. https://doi.org/10.1038/s41598-020-77757-w.

Yan, L., Duarte, F., Wang, D., Zheng, S., Ratti, C., 2019. Exploring the effect of air pollution on social activity in China using geotagged social media check-in data. Cities 91, 116–125. https://doi.org/10.1016/j.cities.2018.11.011.

Yang, J., Shi, B., Shi, Y., Marvin, S., Zheng, Y., Xia, G., 2020. Air pollution dispersal in high density urban areas: research on the triadic relation of wind, air pollution, and urban form. Sustain. Cities Soc. 54, 101941.

Yaseen, Z.M., 2021. An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: review, challenges and solutions. Chemosphere 277, 130126. https://doi.org/10.1016/j.chemosphere.2021.130126.

Ye, J.C., 2022. Artificial neural networks and backpropagation. Geomet. Deep Learn. Springer 91–112.

Yin, J., Li, N., 2022. Ensemble learning models with a Bayesian optimization algorithm for mineral prospectivity mapping. Ore Geol. Rev. 145, 104916 https://doi.org/10.1016/j.oregeorev.2022.104916.

Yin, S., Liu, H., Duan, Z., 2021. Hourly PM2.5 concentration multi-step forecasting method based on extreme learning machine, boosting algorithm and error correction model. Digit. Signal Process. 118, 103221 https://doi.org/10.1016/j.dsp.2021.103221.

Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput. 31, 1235–1270.

Yu, X., Wong, M.S., Liu, C.-H., Zhu, R., 2022. Synergistic data fusion of satellite observations and in-situ measurements for hourly PM2. 5 estimation based on hierarchical geospatial long short-term memory. Atmos. Environ. 286, 119257.

Zhang, P., Ma, W., Wen, F., Liu, L., Yang, L., Song, J., Wang, N., Liu, Q., 2021. Estimating PM2. 5 concentration using the machine learning GA-SVM method to improve the land use regression model in Shaanxi, China. Ecotoxicol. Environ. Saf. 225, 112772.

Zwijnenburg, W., 2015. Iraq's continuing struggle with conflict pollution. Peace Insight.