Discrete Wavelet Transform and Ensemble Tree Algorithms for Air Pollutant Modeling: A Case Study

Snezhana Georgieva Gocheva-Ilieva[1], Atanas Valev Ivanov[2], Maya Plamenova Stoimenova-Minova[3], Stoyanka Krasimirova Koleva-Pavlova[4]

[1] University of Plovdiv Paisii Hilendarski, Bulgaria snow@uni-plovdiv.bg

[2] University of Plovdiv Paisii Hilendarski, Bulgaria aivanov@uni-plovdiv.bg

[3] University of Plovdiv Paisii Hilendarski, Bulgaria mstoimenova@uni-plovdiv.bg

[4] University of Plovdiv Paisii Hilendarski, Bulgaria stkoleva@uni-plovdiv.bg

Abstract: Air pollution is one of the greatest environmental problems of our time on a global and local scale. Elevated and even low but constant levels of harmful emissions in the air in urbanized areas pose serious risks to the health of the population. This paper develops a new approach for statistical modeling of time series of air pollutants, depending on meteorological factors. A new framework based on discrete wavelet transform (DWT) is proposed for decomposing pollutant's time series as a sum of components that represent trend, seasonality, and other specific characteristics. A key element in the applied DWT is an adaptive scheme for selecting the threshold value to control the reverse DWT's accuracy for achieving better prediction of the time series values. The resulting components are modeled with cutting-edge predictive ensemble tree algorithms, including bagging, boosting, and stacking techniques. This approach is tested with real data measured with a mobile automated station in the Plovdiv region, Bulgaria. All models are evaluated, analyzed, and cross-validated. The models are applied for short-term pollution forecasts.

Keywords: Wavelet analysis of time series, Random Forest, Arcing, Stacking

1. INTRODUCTION

The quality of the ambient air is of great importance for human health and all living organisms and is an important part of the complex ecological system of the earth. The protection of air from pollution raises the need of solution of many problems on a global, regional and local level to the entire society - governments, scientific organizations and centers, world associations and others. In particular, this includes scientific measurements and research on the impact on health and the environment of air pollutants such as particulate matter (PM), nitrogen dioxide (NO_2), nitrogen monoxide (NO), sulfur dioxide (SO_2), ground-level ozone (O_3), and other. A large number of medical papers study the trends between environment and public health. The state of the art on the effects of air pollution on the health of older adults during physical activities is presented in the recent evidence-based review [1].

Paper [2] is focused on the risks of polluted air affecting respiratory diseases. Other researchers are investigating the harm to patients with acute and chronic cardiovascular disease from exposure to particulate matter ($PM_{2.5}$) air pollution [3]. Even low concentrations of air pollutants are dangerous for human health, and in particular for children, the elderly and sick people [4], [5]. A systematic summary of the current state of ambient air quality and the harms of excessive levels of the main pollutants for public health, as well as prescriptions for corresponding control and management are presented in the reports of the World Health Organization (WHO) [6].

Along with medical data, the accumulation of huge amounts of measurement data in the field of air pollution provides great opportunities for their statistical processing and extraction of useful information. This gives a basis for applying various means and methods for modeling, analysis, establishing trends and causes, as well as for forecasting the levels of concentrations of harmful air emissions for the purpose of prevention and making adequate management decisions.

In recent decades, the number of scientific studies in this field has grown exponentially. Of these, we will note two main groups - multivariate statistical approaches and large-scale numerical methods, and the increasingly rapidly developing techniques and algorithms of machine learning (ML). Of the first group, popular statistical methods for time series analysis such as multiple linear regression (MLR), principal component analysis (PCA), Markov chain, autoregressive moving average (ARIMA), Gaussian state space analysis and others were used in [7]–[13]. Results from the large-scale complex numerical models are presented in [14], [15].

ML forecasting algorithms are widely used in the literature to increase the performance of forecasting various time series [16]–[19]. Recurrent neural networks (RNN) models using multi-layer perceptron (MLP) and Elman NN were built in [16] to forecast the daily maximum concentrations of SO_2, O_3, PM_{10}, NO_2, CO in the city of Palermo, Italy. To predict the levels of $PM_{2.5}$ Zhu et al. constructed ensemble model based on decision tree, random forest (RF), Adaboost, GBDT, K-nearest neighbor (KNN), XGBoost, LightgBM, CatBoost, support vector regression (SVR), Stacking, Blending, and MLR, respectively [17]. The goodness of fit of all ensemble learning models reached about 0.92-0.94 or 92% to 94% fitting with the initial $PM_{2.5}$ data. A multi-scale deep learning and optimal combination ensemble approach incorporating MLP and stacked sparse autoencoder methods for hourly air quality index (AQI) forecasting is considered in [18]. The authors of [19] model daily concentration of PM_{10} with a NN ensemble combining MLP, radial basis function, Elman NN and SVR.

In recent years, the application of new powerful hybrid methods such as continuous and discrete wavelet analysis, combined with ML, has gained popularity. Recent publications in this line in the field of environmental sciences are, for example [20]–[22]. A combination of function expansions in a wavelet series with ARIMA models is used in [20]. Correlation analysis and artificial neural networks (ANNs) including wavelet analysis to identify the linear and nonlinear associations, respectively, between the air pollution index and meteorological variables in two Chinese cities were applied in [21]. Twelve algorithms and large amount of meteorological predictors have been used to achieve correlation coefficients of the forecasts about 88-89%. Air pollution data (SO_2, NO_2, PM_{10}, $PM_{2.5}$, O_3, and CO) and four meteorological variables (temperature, barometric pressure, humidity, and wind speed) are modeled in [22] using wavelet de-noising, detrended correlation analysis, and long short-term memory (LSTM) methods.

In this study, a new methodology and framework for time series analysis is proposed, based on combinations of DWT and different ensemble tree ML methods. The goal is to remove noise from the data by adaptive selection and combination of wavelets, then construct predictive regression models with ensemble methods with increased performance and prediction ability. The developed approach is applied to the analysis and prediction of two separate air pollutants (NO_2 and SO_2), depending on four meteorological variables. Models are selected and evaluated using standard statistical metrics. In addition, a diagnosis of the residuals of the models is carried out to check their adequacy. Models are validated by applying them for short-term predictions to data not used in their construction.

Statistical computations are performed using Wolfram Mathematica, Salford Predictive Modeler (SPM), and SPSS software [23]–[25].

2. METHODOLOGY AND FRAMEWORK OF THE STUDY

For modeling a given time series *Y*, we apply the following methodology (see also Fig. 1):

1) We divide the data for *Y* into two parts - for constructing the models (learning sample) and for validation dataset.

2) We select several threshold values for adaptive calibration of wavelet models. In our case, parts of the standard deviation of *Y* are used.

3) To the training sample we apply various types of DWT and inverse transformations with different thresholds for de-noising the learning sample. We select the wavelet models for which the noise is smal (for example, in a selected confidence interval). The following DWTs are used in our study ($w_p,\ p=1,2,3$): biorthogonal spline wavelet, Daubechies and Haar, labeled B, D, and H, respectively.

4) We average the selected de-noised wavelet models by linear stacking and obtain the model *wY*.

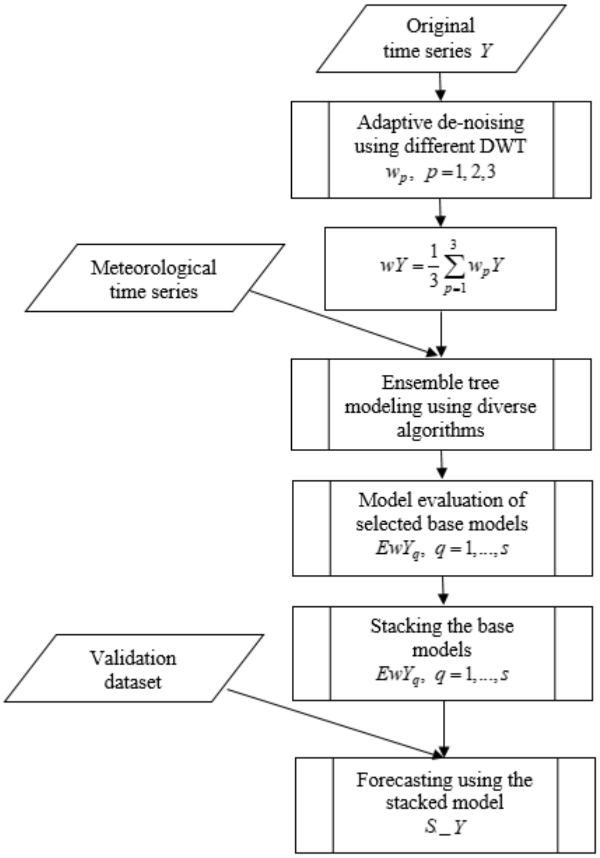

Original time series Y
Adaptive de-noising using different DWT w_p, $p = 1, 2, 3$
$wY = \frac{1}{3}\sum_{p=1}^{3} w_p Y$
Meteorological time series
Ensemble tree modeling using diverse algorithms
Model evaluation of selected base models EwY_q, $q = 1, \ldots, s$
Stacking the base models EwY_q, $q = 1, \ldots, s$
Validation dataset
Forecasting using the stacked model S_Y

Fig. 1. The framework of the study.

5) Using data from weather variables for the entire time period, we build *s* (*s*>=3) number of diverse ML models. In our case, we use three ensemble learning algorithms.

6) We evaluate the performance of ensemble models and select the best of each type as base models.

7) We stack the selected base models by averaging and denote the resulting model by *S_Y*.

8) We apply this model to forecast *Y* on the test sample.

3. DESCRIPTION OF USED METHODS

The specific methods used in this paper are: wavelet transform, RF, Arcing (ARC), CART Ensembles and Bagging (CEB), and simple averaging for stacking. The three ensemble methods are based on the Classification and regression trees (CART) method, presented in [26].

A. Wavelet Transform

The WT is an improved analogue of the Fourier transform. The theoretical basis and some practical issues of WT are presented in [27]–[30]. The WT is applied for modeling of nonstationary processes occurring over finite spatial and temporal domains. It uses generalized local base functions (wavelets) that can be stretched and translated with a flexible resolution in both frequency and time [29].

In the continuous case, the idea of WT is based on a decomposition of the signal $f(t) \in L^2(R)$ in terms of the so-called "mother wavelet" or simply “wavelet”.

$$\psi_{a,b}(t)=\frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right), \quad (1)$$

where a denotes the scale (dilation) and b denotes the position (translation) of the wavelet. Various types of wavelets are used in the literature [27]–[30].

The WT is defined as a convolution integral

$$W(a,b)=\frac{1}{\sqrt{|a|}}\int_t f(t)\psi^*\left(\frac{t-b}{a}\right)dt \quad (2)$$

where ψ^* is the complex conjugate of ψ defined on the open “time and scale" real (a,b) half plane. The original function $f(t)$ can be formally reconstructed using the wavelet coefficients by the inversion formula [29]:

$$f(t)=\frac{1}{c_\psi}\int\frac{\mathrm{d}a}{a^2}\int\frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right)W_{(a,b)}\,\mathrm{d}b \quad (3)$$

where

$$c_\psi=\int_0^{+\infty}\frac{|\hat{\psi}(\omega)|}{\omega}\mathrm{d}\omega<+\infty,$$

and $\hat{\psi}(\omega)$ is the Fourier transform of $\psi(\omega)$.

In the DWT when the function $f(t)$ is given in table form, the parameters are discretized and the coefficients and integrals are represented approximately by sums [30]. DTW also applies series decomposition as a linear sum of component series, by setting different parameters.

B. Random Forest

RF is among the most popular and effective ML methods. It is used for classification and regression. It was proposed by Lio Breimen in 2001 in [31]. The RF model includes tens or hundreds of independent CART trees. For each individual tree, the initial data sample is randomly divided into two parts - about 2/3 is used to train the model, and the rest, called out-of-bag (OOB), is used for testing. When building the tree, a random subsample of the entire set of predictors is selected to split the data at each node. RF trees are usually unpruned. The predictions of all trees are averaged and the resulting ensemble model is obtained.

The RF hyperparameters are used to calibrate and stop the algorithm. The most important of them are: number of trees, limits for the number of observations in parent and child nodes (m1:m2), number of randomly selected predictors from all predictors of each splitting of a node. Lagged variables can also be used as predictors.

C. Arcing

Arcing (acronym for “adaptive resampling and combining”) is a class of ensemble algorithms using the boosting paradigm. Arcing was introduced by Breiman in [32] in 1998, but its capabilities have not yet been sufficiently explored. The general idea of boosting algorithms is to generate a sequence of multiple single but dependent models of an ensemble. In Arcing, if we have already generated K single models, the next (K+1)th model is constructed with a new training dataset of initial size N from the original data by weighted random sampling. The weight of each selected observation is calculated by a formula that takes into account its assigned weights from all previous K models. The weight is greater for those observations that contributed to more error. For the case of regression, the final model is grown by taking the average of all trees at one vector. The algorithm can be applied to classification and regression, both for decision trees and ensembles of NN or other algorithms. Data can be of numeric or string type. The main effect of arcing and bagging is to reduce both bias, variance and test error [32], [33].

Arcing hyperparameters set at the beginning of the analysis are: number of trees, limits on the minimum number of observations in parent and child nodes, method of validation for individual models (with k-fold cross-validation, external sampling, and others).

In the current study, we apply the original variant of Breiman's algorithm, also known as Arc-x4 [32], implemented in [24].

D. CART Ensembles and Bagging

The third ML method used for the generation of base models is a typical representative of the bagging class. It was developed by Breiman in [34]. Unlike arced trees in the ensemble, the bagged trees are independent of each other. In CART Ensembles and bagging (CEB), each single tree is constructed by resampling the dataset from the original data, but without using weights. In other characteristics, the arcing and CEB algorithms have the same assumptions and hyperparameters.

In this study, we will build the ensemble bagging models with CART trees, implemented in [24].

E. Stacking

The stacking paradigm implies combining at least three diverse models created with different algorithms. These models are called base models or level 1 models [33]. At the next level 2, the base models are used as predictors to build a new model, which is done with a different algorithm compared to the level 1 algorithms. It is possible to apply several algorithms at level 2, after which, in turn, the resulting models can stack to level 3, and so on. The procedure is carried out in order to obtain a new, more efficient model.

In this study, we will run averaging stacking twice: for wavelet models, and then for selected base models created using RF, Arcing and CEB algorithms.

Model Evaluation

To evaluate the constructed models of time series $Y=\{Y_1,Y_2,...,Y_N\}$, we use the statistical measures root mean squared error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination R^2, calculated by the expressions:

$$RMSE=\sqrt{\frac{1}{N}\sum\nolimits_{t=1}^{N}(Y_t-P_t)^2}, \qquad (4)$$

$$MAPE=\frac{100}{N}\sum\nolimits_{t=1}^{N}\left|\frac{Y_t-P_t}{Y_t}\right|, \qquad (5)$$

$$R^2=\frac{\left\{\sum\nolimits_{t=1}^{N}(Y_t-\overline{Y})(P_t-\overline{P})\right\}^2}{\sum\nolimits_{t=1}^{N}(Y_t-\overline{Y})\sum\nolimits_{t=1}^{N}(P_t-\overline{P})} \qquad (6)$$

where $P=\{P_1,P_2,...,P_N\}$ is the predicted series, $\overline{Y}$, $\overline{P}$ are the mean values of the two time series, N is the sample size.

We assess the adequacy of the models by diagnosing their residuals with analysis of the plot of their autocorrelation function (ACF). In addition, the Durbin-Watson (DW) test for the absence of serial correlation in linear regression will be used. The DW statistic takes values in the interval [0, 4], where a value close to 2.0 indicates a lack of autocorrelation.

4. DESCRIPTION OF DATA

We will demonstrate the proposed methodology for data modeling for two air pollutants – NO_2 and SO_2 in the city of Plovdiv, Bulgaria. Plovdiv is the second largest city after the capital Sofia, with about 360 thousand inhabitants. It is built around 7 hills in the south-central region of the country. The terrain is low flat, with an altitude of 164 m. The climate is transitional-continental. The range of mean annual temperatures varies from -3°C in winter to 31°C in summer. The average annual relative humidity is 73% and the average annual rainfall is 540 mm. In Plovdiv, weak winds prevail (0 - 5 m/s), with winds of up to 1 m/s being up to 95% of the year. This is a prerequisite for the retention of harmful concentrations of air pollutants, especially in the winter period.

The studied data for NO_2 and SO_2 are on an average daily scale from 11 April 2021 to 24 March 2023 or in total for N1=720 days. They were measured by Alphasense automatic monitoring station with optical sensors, located on the roof of the new building of Plovdiv University. Four meteorological time series were also measured for the same period: Temperature, Humidity, Pressure and Luminosity. We denote the relevant variables by NO2, SO2, Temp, Hum, Press, and Lum.

The initial raw data had less than 10% missing values. They were replaced by the average known values for the same dates measured in subsequent years.

To conduct the analyses, we divide all available N1=720 data into two parts. We will use the first part for N=713 days for a learning set, and the last 7 measurements - for validating the models and evaluating their forecasts. Descriptive statistics of the data for N=713 are shown in Table I.

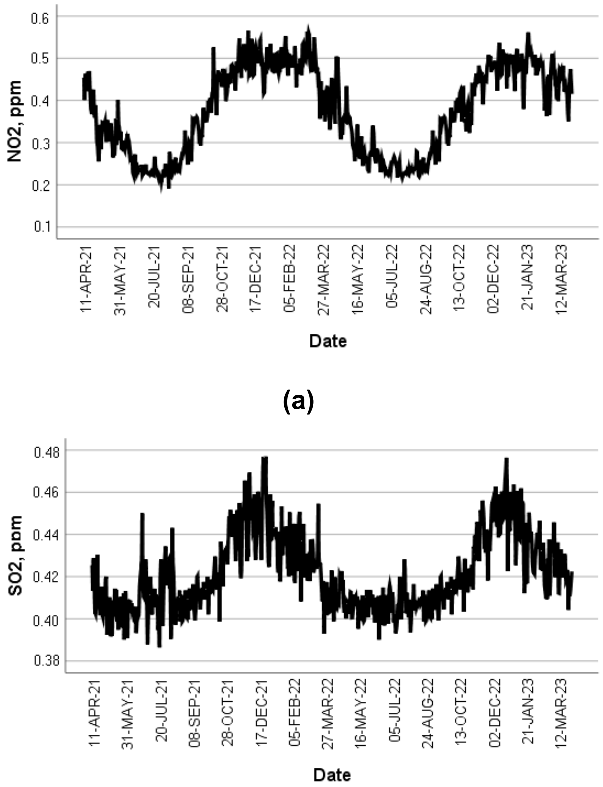
Table I shows that pollutant levels both for NO_2 and SO_2 are problematic. The permissible average daily concentrations for the European Union, Bulgaria, as well as for the WHO [6] for NO_2 and SO_2 are 120 $\mu g/m^3$=0.062 ppm and 125 $\mu g/m^3$=0.047 ppm, respectively. An excess of about 6 times was observed for NO_2 and 9 times for SO_2.

Sequence plots of the pollutant's variables are shown in Fig. 2. An annual cyclicity is observed.

Table I. Descritive statistics of the used variables.

Variable	Statistics				
	Mean	Std. Dev.	Variance	Min	Max
NO2, ppm	0.383	0.102	0.01	0.191	0.565
SO2, ppm	0.420	0.018	0.0003	0.386	0.476
Temp, °C	16.24	9.31	86.65	-3.03	35.29
Hum, %	59.40	19.26	370.89	22.16	100.00
Press,	995.83	6.91	47.73	975.46	1018.10

Pa					
Lum, Luxe	3873	1719	2955626	176	7652



(a)

(b)

Fig. 2. Sequence plots of the pollutant's variables: (a) NO2, (b) SO2.

5. RESULTS

We conduct the modeling and analyzes according to the proposed methodology and framework in section II.

A. Wavelet De-noising

First, we will use the training sample with N=713 observations for de-noising with three different families of wavelets – biorthogonal splines (B), Daubechies (D), and Haar wavelets (H).

The decomposition of time series *Y* in DWT is carried out by calculating its coefficients in stages or refinements of the type illustrated in Fig. 3. The initial main signal *Y* is denoted by a0. First, it is decomposed into a sum of new signal a1 and noise signal d1, then a1 is decomposed into sum of signal a2 and noise signal d2, etc. The number of refinements depends on the choice of wavelet and the hyperparameters of the algorithm used. The final decomposition after the *r* stage is:

$$Y = a_r + d_1 + d_2 + ...d_r \qquad (7)$$

Fig. 4 shows the decomposition of DWT coefficients using biorthogonal spline of the NO2 using 4 refinements.

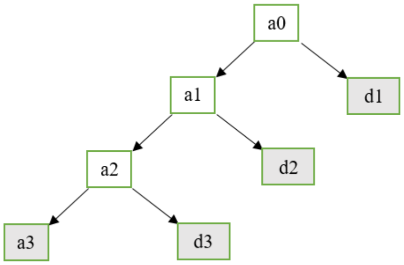


Fig. 3. A tree view of decomposed DWT coefficients for $r = 3$.

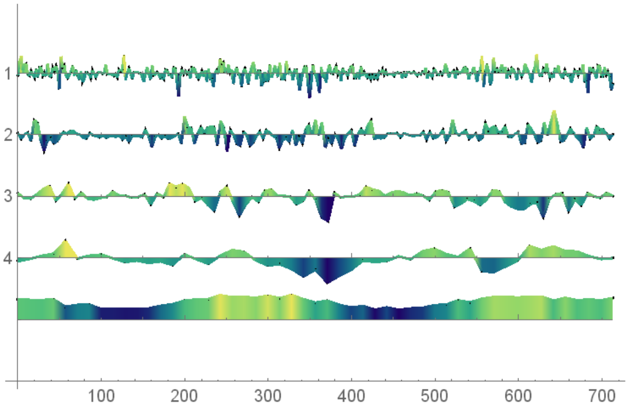


Fig. 4. Wavelet list plot of decomposed DWT coefficients using biorthogonal spline (B).

After performing a DWT, an inverse transform is applied to recover the original data. In this procedure, the main parameter is the threshold value for determining the accuracy of the extracted wavelet model. We denote this threshold by *th*. In this research, we have applied an adaptive determination of the threshold value, comparing the result with the ACF of the residuals of the wavelet model. Good results were obtained at values $th = \mathrm{std},\ \mathrm{std}/2,\ \mathrm{std}/3$, and std4, where the std is the standard deviation of the original time series variable. We denote the selected wavelet models (w-models) for NO2 at th=std/3 by B_NO2, D_NO2, and H_NO2, respectively for B, D and H DWT. Their plots are given in the upper part of Fig. 5. Similarly, the lower part of Fig. 5 shows the graphs for SO2 at th=std/4 as well.

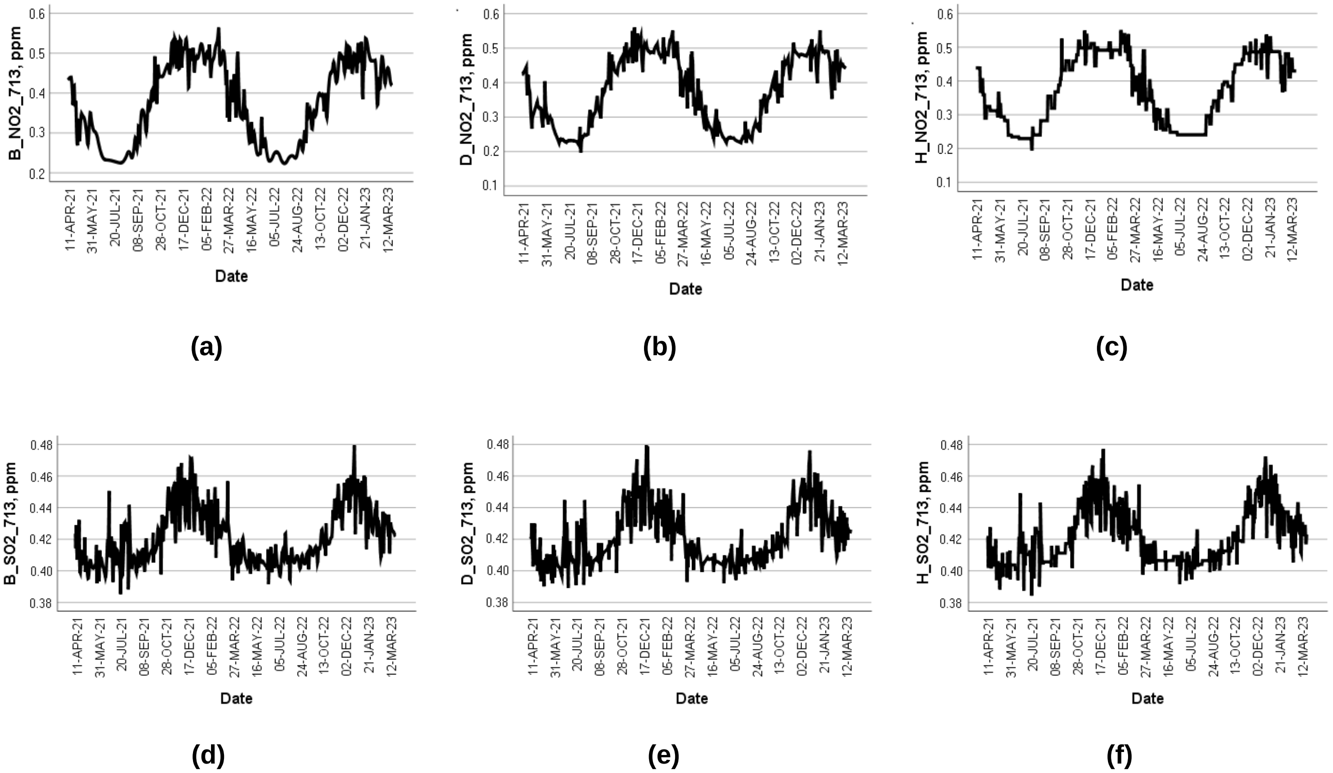
The approximations obtained after the inverse DWT for both pollutants are very good and within the limits of the ACF of their residuals. The corresponding statistical measures (4)-(6) and DW statistics of these models are given in the upper rows of Table II. DW is calculated by regression of a given model with the target *Y*. It is seen that the DW statistics are close to 2, so there is no autocorrelation in the models.

Following the proposed methodology, we average the w-models with the expression

$$wY = (B\_Y + D\_Y + H\_Y)/3 \quad (8)$$

where *Y*=NO2 or *Y*=SO2. With this linear combination, the statistics of the averaged models were found to improve over their composite models, as seen in Table II.

Fig. 6 shows the quality of fitting of the w-models to the measured data. Excellent correspondence is observed. In particular, we should note that the combined w-models very well approximate the high and low values of the two time series, as can be seen from Fig. 6.



(a) (b) (c)

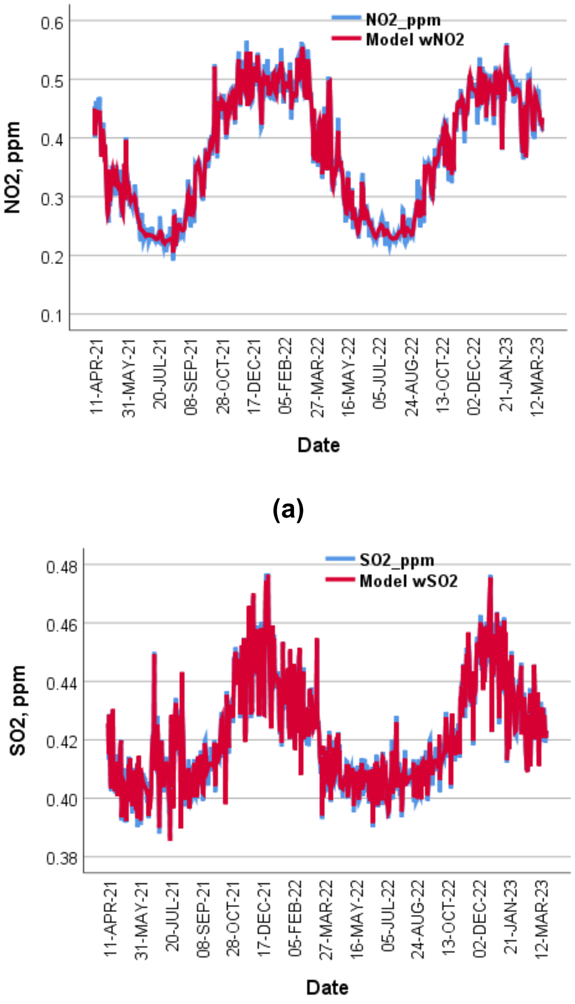
(d) (e) (f)

Fig. 5. Sequence plots of selected wavelet models for NO2 using the DTW: (a) biorthogonal spline (B); (b) Daubechies (D); (c) Haar (H); and for SO2: (d) B wavelet; (e) D wavelet; (f) H wavelet.

Table II. Performance statistics of the built models.

Model	Statistics				
	Cases	RMSE	MAPE	*R*^{2}	DW
B_NO2	713	0.0139	3.08	0.982	1.875
D_NO2	713	0.0144	3.17	0.981	1.760
H_NO2	713	0.0132	2.95	0.984	1.914
B_SO2	713	0.0022	0.40	0.985	2.329
D_SO2	713	0.0017	0.30	0.991	2.321
H_SO2	713	0.0018	0.33	0.990	2.176
wNO2	**713**	**0.0106**	**2.36**	**0.990**	**2.055**
wSO2	**713**	**0.0013**	**0.24**	**0.995**	**2.384**
NA30	720	0.0106	2.26	0.990	2.033
NRF300	720	0.0112	2.42	0.990	1.935
NC30	720	0.0106	2.30	0.990	1.970
SA15	720	0.0042	0.74	0.945	1.889
SRF200	720	0.0041	0.73	0.943	1.791

SC40	720	0.0040	0.70	0.952	1.854
S_NO2	**720**	**0.0106**	**2.30**	**0.990**	**2.027**
S_SO2	**720**	**0.0039**	**0.70**	**0.957**	**1.867**

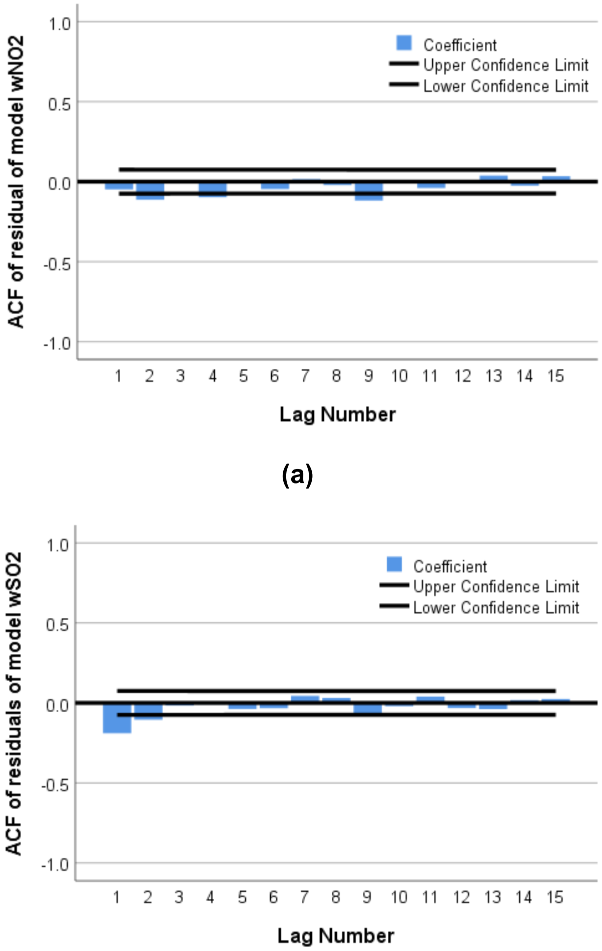


(a)

(b)

Fig. 6. Comparison between measured data and w-models for: (a) NO2 and wNO2, (b) SO2 and wSO2.

The ACF of the residuals of the averaged w-models are shown in Fig. 7. Also from Table II the DW statistics are close to 2, so there is no autocorrelation. Also, as the residuals are small and almost within the confidence intervals, we can conclude that the models are adequate.



(a)

(b)

Fig. 7. ACF of the residuals of the selected w-models for: (a) NO2, (b) SO2.

B. Building and Evaluating of Base Ensemble Tree Models

The use of DWT was intended to remove the influence of noise in the raw data. When we aim to predict future values of a time series, we need regression type models as well as predictors whose values are known in the future. To evaluate the forecasts, we introduce the four meteorological variables described in Table I, with values up to N1=720 observations. Model variables wNO2 and wSO2 were selected as target variables.

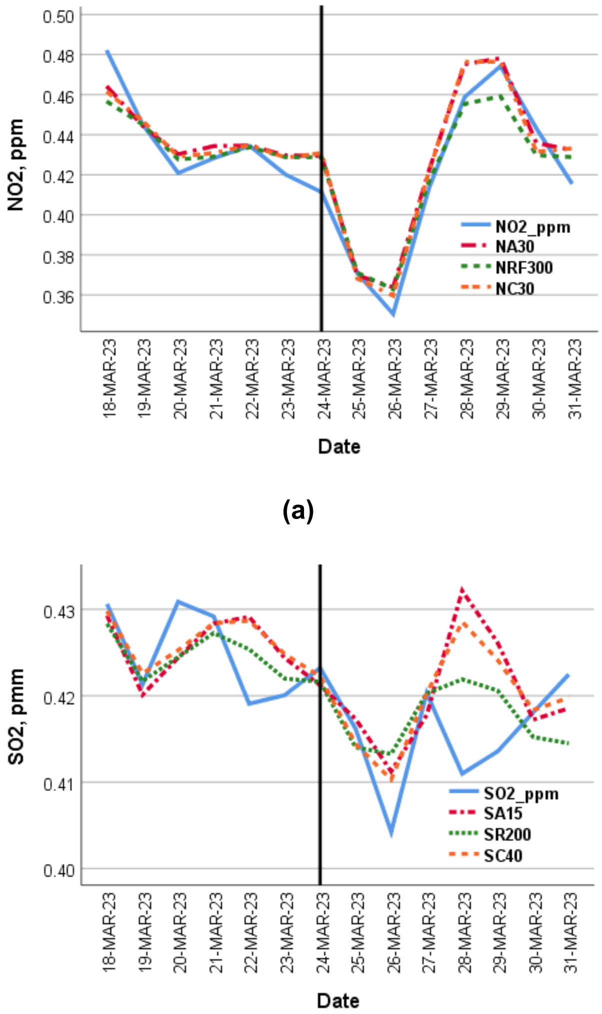
Multiple ensemble models were constructed with their hyperparameters varied as follows:

1. For RF (R): the number of trees in the ensemble were set to 100, 200 and 300; procedure for testing the models – the OOB by default; number of randomly selected predictors – 2, 3, and 4; minimum cases in terminal node – 5.
2. For Arcing (A) models: 20, 30, 40, and 50 trees in each ensemble model; m1:m2 = 10:5, 5:5; 10-fold cross-validation.
3. For CEB (C) models - as for Arcing.

All models were evaluated according to criteria (4)-(6) for all N1=720 observations. The corresponding selected best models are indicated by adding the number of trees to their name, for e.g. NR200, NA30, etc. Their remaining parameters are further described in Table II.

From the results shown in Table II, it can be seen that the models are improved in stages, following the developed modeling framework (see Fig. 1). Finally, after the last step, the S_NO2 and S_SO2 models showed the best statistics. The only exception was obtained for model NA30, whose MAPE = 0.26 is minimal. However, when examining the ACF of the residuals, this model gives way to the stacked model S_NO2.

The forecasts for the last 7 test days for the three best ensemble models are illustrated in Fig. 8 along with their confidence intervals. A relatively very good quality of the predicted values is observed. The quality of the NO2 forecasts is significantly better, most likely due to better statistics compared to SO2.



(a)

(b)

Fig. 8. Forecasts of the selected base models for the last 14 days: to the left of the vertical line – for the last 7 days of the training sample, to the right - for the test sample: (a) for NO2, (b) for SO2.

C. Building and Evaluation of the Stacked ML Models

The last step of the proposed methodology is the stacking of the selected best base models from the previous section. For this purpose, we averaged the resulting base models separately for NO2 and SO2. For NO2 the sacked model S_NO2 is found by

$$S\_NO2 = (NA30 + NR300 + NC30)/3 . \quad (9)$$

Analogically, we calculate the values of the stacked model for SO2, denoted by S_SO2 according to the formula

$$S\_SO2 = (SA15 + SR200 + SC40)/3 . \qquad (10)$$

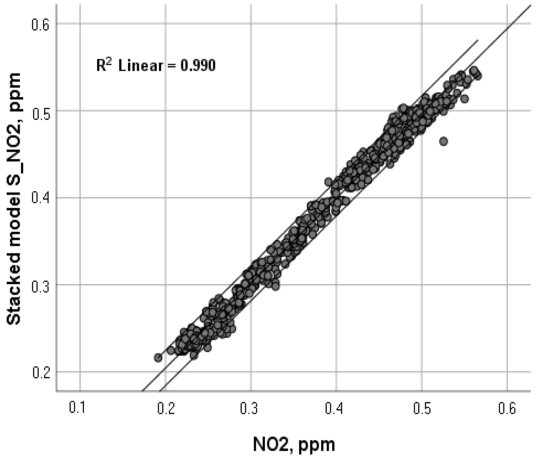
The statistical indices of the two stacked models S_NO2 and S_SO2 are shown at the bottom of Table II. The comparison with the others shows that they improve the performance of all other models. For prediction of NO2 with model S_NO2 we obtained coefficient of determination R^2.=0.990 and MAPE=2.3%. With model S_SO2 we have R^2=0.957, RMSE=0.0039 and MAPE-0.70, respectively.

Fig. 9 presents a scatter-plot for comparing the values of the stacked models with the initial variables NO2 and SO2, supplemented up to N1=720 observations. Very good agreement with the measured data and small deviations from the confidence interval are observed.

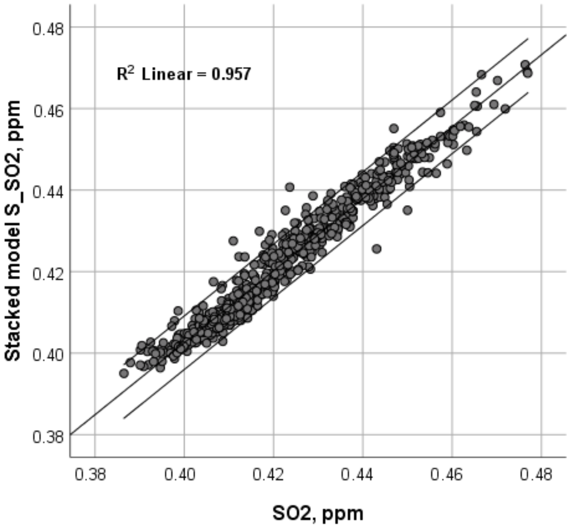
D. Forecasting the Test Sample Using the Stacked Models

The forecast values only for the last 7 days, saved for model validation, are shown in Fig. 10.

We can conclude that the proposed methodology and research framework are effective means of increasing the performance of ML methods for modeling and forecasting air pollution.

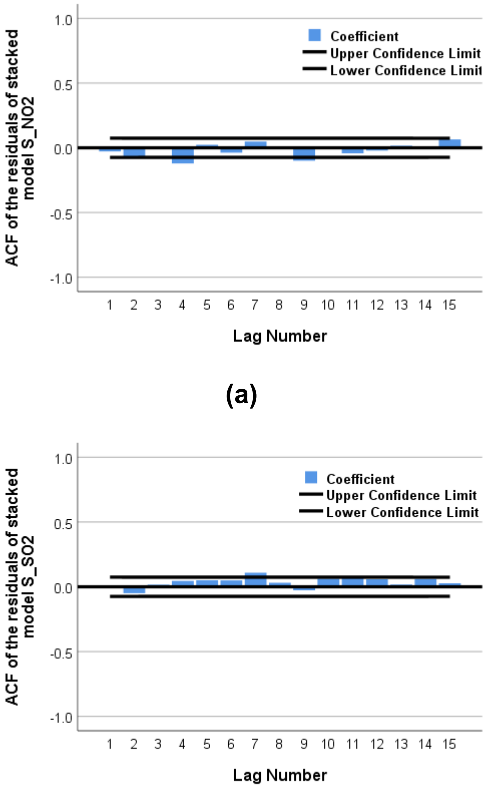


(a)



(b)

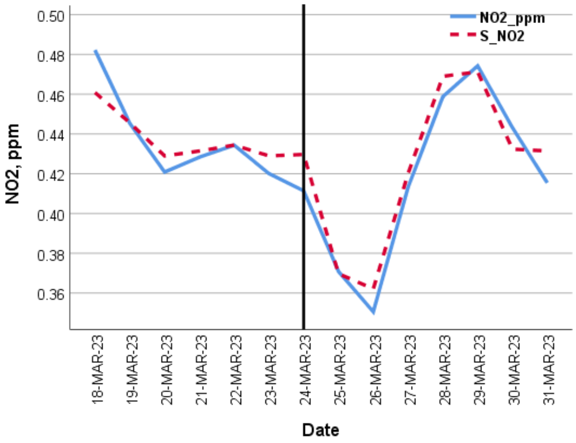
Fig. 9. Scatter plots of stacked: models versus measured pollutants' variables: (a) for NO2, (b) for SO2.



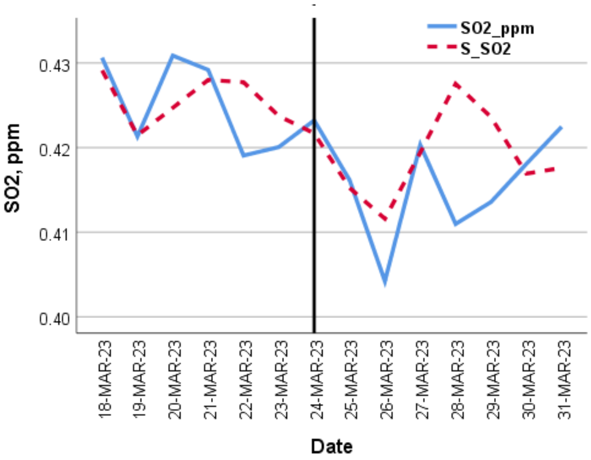
(a)

(b)

Fig. 10. ACF plot of residuals of the stacked models: (a) for S_NO2, (b) for S_SO2.



(a)



(b)

Fig. 11. Forecasts of the stacked models for the last 14 days: to the left of the vertical line – for the last 7 days of the training sample, to the right - for the test sample: (a) for NO2, (b) for SO2.

6. CONCLUSION

In this research, we developed a methodology and framework for time series modeling and forecasting based on wavelet analysis and machine learning algorithms. We applied three DWTs for de-noising the initial time series. We averaged the obtained w-models, which resulted in an increase in performance.

Prediction on a test sample was carried out with three ensemble tree methods – arcing, random forest and ensemble bagging. One best ML model from each of the three approaches was selected. The results were again stacked to a new improvement. All models are validated and tested for autocorrelation. The methodology was applied for modeling and short-term forecasts for two air pollutants (NO2 and SO2).

The proposed and demonstrated approach showed great ability for statistical research of complex time series from the field of environmental science.

7. ACKNOWLEDGEMENT

This study has emanated from a research conducted with the financial support of the Bulgarian National Science Fund (BNSF), Grant number KP-06-IP-CHINA/1.

REFERENCES

[1] A. Andrade, A. D'Oliveira, L. C. De Souza, L. Stabile, and G. Buonanno, "Effects of air pollution on the health of older adults during physical activities: Mapping review," International Journal of Environmental Research and Public Health, vol. 20, no. 4, art. 3506, February 2023. DOI 10.3390/ijerph20043506

[2] G.-P. Bălă, R.-M. Râjnoveanu, E. Tudorache, R. Motișan, and C. Oancea, "Air pollution exposure—the (in)visible risk factor for respiratory diseases," Environmental Science and Pollution Research, vol. 28, no. 16, pp. 19615-19628, March 2021. doi: 10.1007/s11356-021-13208-x

[3] J. Finch, D. W. Riggs, T. E. O'Toole, C. A. Pope III, A. Bhatnagar, and D. J. Conklin, "Acute exposure to air pollution is associated with novel changes in blood levels of endothelin-1 and circulating angiogenic cells in young, healthy adults," AIMS Environmental Science, vol. 6, no. 4, pp. 265-276, 2019. doi: 10.3934/environsci.2019.4.265

[4] Z. Chen, B. Wang, Y. Hu, X. Cui, and T. Shi, "Short-term effects of low-level ambient air NO2 on the risk of incident stroke in Enshi city, China," International Journal of Environmental Research and Public Health, vol. 19, no. 11, art. 6683, May 2022.

[5] M. F. Ibrahim, R. Hod, M. A. B. A. Tajudin, W. R. W. Mahiyuddin, A. M. Nawi, and M. Sahani, "Children's exposure to air pollution in a natural gas industrial area and their risk of hospital admission for respiratory diseases," Environmental Research, vol. 210, art. 112966, July 2022.

[6] World Health Organization 2021 WHO global air quality guidelines: particulate matter (PM2.5 and PM10),

ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Available online at: https://apps.who.int/iris/handle/10665/345329

[7] S. Abdullah, N. N. L. M. Napi, A. N. Ahmed, A. M. Abdullah, and Z. T. A. Ramly, “Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia,” Atmosphere, vol. 11, no. 3, art. 289, March 2020.

[8] V. A. Reisen, E. Z. Monte, G. da Conceição Franco, A. M. Sgrancio, F. A. F. Molinares, P. Bondon, F. A. Ziegelmann, and B. Abraham, “Robust estimation of fractional seasonal processes: Modeling and forecasting daily average SO2 concentrations,” Mathematics and Computers in Simulation, vol. 146, pp. 27-43, April 2018. doi: 10.1016/j.matcom.2017.10.004

[9] I. Minkova, M. Filipova, and I. Zheleva, “Statistic study of particulate matter (PM10) air contamination in the city of Svishtov, Bulgaria,” AIP Conference Proceedings, AMITANS 2022 Conference Proceedings (to be published).

[10] M. Nikolova, M. Filipova, and I. Zheleva, “Statistic study of gaseous air contamination in the city of Ruse, Bulgaria,” AIP Conference Proceedings, AMITANS 2022 Conference Proceedings (to be published).

[11] E. Veleva, and I. R. Georgiev, “Seasonality of the levels of particulate matter PM10 air pollutant in the city of Ruse, Bulgaria,” AIP Conference Proceedings, vol. 2302, art. 030006, December 2020.

[12] Y. Alyousifi, K. Ibrahim, M. Othamn, W. Z. W. Zin, N. Vergne, and A. Al-Yaari, “Bayesian information criterion for fitting the optimum order of Markov chain codels: Methodology and application to air pollution data.” Mathematics, vol. 10, no. 13, art. 2280, June 2022.

[13] S. Kanchanasuta, S. Sooktawee, N. Bunplod, A. Patpai, N. Piemyai, and R. Ketwang, “Analysis of short-term air quality monitoring data in a coastal area,“ AIMS Environmental Science, 2021, vol. 8, no. 6, pp. 517-531, 2021. doi: 10.3934/environsci

[14] A. Luna, H. Navarro, and A. Moya, “SO2 and NO2 simulation and validation in metropolitan lima using WRF-chem model,” International Journal of Computational Methods and Experimental Measurements, vol. 8, no. 2, pp. 135-147, November 2020.

[15] V. Todorov, I. Dimov, S. Fidanova, T. Ostromsky, and R. Georgieva, “Optimized Monte Carlo methods for sensitivity analysis for large-scale air pollution model,” Studies in Computational Intelligence, vol. 1044, pp. 277-288, September 2022.

[16] U. Brunelli, V. Piazza, L. Pignato, F. Sorbello, and S. Vitabile, “Two-days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy,” Atmospheric Environment, vol. 41, no. 14, pp. 2967-2995, May 2007. doi: 10.1016/j.atmosenv.2006.12.013

[17] Y. Zhu, C. Liu, and J. Ma, “Prediction of PM2.5 concentration in Changchun based on ensemble learning model,” IEEE Proceedings - 18th International Conference on Computational Intelligence and Security (CIS 2022), pp. 79-83, December 2022.

[18] Z. Wang, H. Chen, J. Zhu, and Z. Ding, “Multi-scale deep learning and optimal combination ensemble approach for AQI forecasting using big data with meteorological conditions,” Journal of Intelligent and Fuzzy Systems, vol. 40, no. 3, pp. 5483-5500, March 2021.

[19] K. Siwek, and S. Osowski, “Improving the accuracy of prediction of PM10 pollution by the wavelet transformation and an ensemble of neural predictors.” Engineering Applications of Artificial Intelligence, vol. 25, no. 6, pp. 1246-1258, Sept. 2012.

[20] O. Mandrikova, N. Fetisova, and Y. Polozov, “Hybrid model for time series of complex structure with ARIMA components,” Mathematics, vol. 9, art. 1122, May 2021.

[21] Q. Guo, Z. He, S. Li, X. Li, J. Meng, Z. Hou, J. Liu, and Y. Chen, “Air pollution forecasting using artificial and wavelet neural networks with meteorological conditions,” Aerosol and Air Quality Research, vol. 20, no. 6, pp. 1429-1439, June 2020.

[22] Z. Zhang, H. Chen, and X. Huang, “Prediction of air quality combining wavelet transform, DCCA correlation analysis and LSTM model,” Applied Sciences, vol. 13, no. 5, art. 2796, February 2023. . https://doi.org/10.3390/app13052796

[23] Wolfram Mathematica, https://www.wolfram.com/mathematica/. Last accessed 15 August 2023

[24] Salford Predictive Modeler, https://www.minitab.com/en-us/products/spm/. Last accessed 15 August 2023

[25] IBM SPSS Statistics, https://www.ibm.com/products/spss-statistics Last accessed 15 August 2023

[26] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification and Regression Trees,”

Wadsworth Advanced Books and Software, Belmont, Canada, 1984.

[27] S. G. Mallat, “A theory for multiresolution signal decomposition: The wavelet representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pp. 674-693, July 1989.

[28] I. Daubechies, “Orthonormal bases of compactly supported wavelets,” Communications on Pure and Applied Mathematics, vol. 41, no. 7, pp. 909-996, Oct. 1988.

[29] K.-M. Lau, and H.-Y. Weng, ”Climate signal detection using wavelet transform: How to make a time series sing,” Bull. Amer. Meteor. Soc., vol. 76, pp. 2391-2402, 1995.

[30] T. V. Burrus, C. Burrus, K. Narasimhan, Y. Guo, and C. Li, “Introduction to Wavelets and Wavelet Transforms: A Primer,” Prentice Hall, Inc., New Jersey, 1998.

[31] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. doi:10.1023/A:1010933404324.

[32] L. Breiman, “Arcing classifiers,” Annals of Statistics, vol. 26, pp. 801-849, June 1998.

[33] L. I. Kuncheva, “Combining Pattern Classifiers: Methods and Algorithms,” John Wiley and Sons, Inc., Hoboken, 2004.

[34] L. Breiman, “Bagging predictors”, Machine Learning, vol. 24, no. 2, pp. 123-140, 1996. doi:10.1007/bf00058655

DOI: https://doi.org/10.15379/ijmst.v10i4.2251