sustainability

MDPI

Article

Prediction of Daily Mean PM_{10} Concentrations Using Random Forest, CART Ensemble and Bagging Stacked by MARS

Snezhana Gocheva-Ilieva *, Atanas Ivanov and Maya Stoimenova-Minova

Department of Mathematical Analysis, Faculty of Mathematics and Informatics, Paisii Hilendarski University of Plovdiv, 4000 Plovdiv, Bulgaria; aivanov@uni-plovdiv.bg (A.I.); mstoimenova@uni-plovdiv.bg (M.S.-M.)
* Correspondence: snow@uni-plovdiv.bg

Abstract: A novel framework for stacked regression based on machine learning was developed to predict the daily average concentrations of particulate matter (PM_{10}), one of Bulgaria's primary health concerns. The measurements of nine meteorological parameters were introduced as independent variables. The goal was to carefully study a limited number of initial predictors and extract stochastic information from them to build an extended set of data that allowed the creation of highly efficient predictive models. Four base models using random forest, CART ensemble and bagging, and their rotation variants, were built and evaluated. The heterogeneity of these base models was achieved by introducing five types of diversities, including a new simplified selective ensemble algorithm. The predictions from the four base models were then used as predictors in multivariate adaptive regression splines (MARS) models. All models were statistically tested using out-of-bag or with 5-fold and 10-fold cross-validation. In addition, a variable importance analysis was conducted. The proposed framework was used for short-term forecasting of out-of-sample data for seven days. It was shown that the stacked models outperformed all single base models. An index of agreement IA = 0.986 and a coefficient of determination of about 95% were achieved.

Keywords: air pollution; machine learning; stacking; rotation ensemble; bagging; selective ensemble; diversity strategy

Citation: Gocheva-Ilieva, S.; Ivanov, A.; Stoimenova-Minova, M. Prediction of Daily Mean PM_{10} Concentrations Using Random Forest, CART Ensemble and Bagging Stacked by MARS. *Sustainability* **2022**, *14*, 798. https://doi.org/10.3390/su14020798

Academic Editors: Baojie He, Ayyoob Sharifi, Chi Feng and Jun Yang

Received: 8 December 2021
Accepted: 7 January 2022
Published: 11 January 2022

1. Introduction

Nowadays, environmental protection is one of the main problems that require complex solutions. Air pollution by harmful aerosols and particulates has a strong negative impact on human health, causing various diseases, according to a number of medical studies [1–3]. In particular, the concentrations of particulate matter with a diameter of 10 microns or less (PM_{10}) can harm the lung tissues and throat, aggravate asthma, and increase respiratory illness. To reduce pollution in the European Union, regulations and measures were adopted in 2008 to control the concentrations of the most harmful air pollutants [4,5]. As a member of the EU, Bulgaria implements legal measures for the preservation of atmospheric clean air. Even though the last 3–4 years have seen a good reduction in harmful pollution, it is still above the prescribed limits in many cities. Further improvement requires continued monitoring and careful study of the air quality state in order to achieve the sustainable development of a healthy urban atmosphere on a global and local scale.

The availability of a large volume of data regarding pollutants over time enables the performance of analyses to extract essential information about the interactions between various factors that determine the level of pollution. Meteorological changes are among the important conditions for the formation and transport of pollution and cannot be ignored in this type of analysis [6,7]. Hence, mathematical and statistical modeling of the respective time series data has an important function in providing reliable tools for analysis and forecasting.

Two types of regression models for forecasting time series data can be systematized from the available literature: parametric and nonparametric, with the latter covering

mainly computer-based methods with machine learning (ML). Parametric methods include multiple linear regression (MLR), the linear mixed model, the Box–Jenkins autoregressive integrated and moving average (ARIMA), numerical sensitivity analysis, and more. These methods provide global models, simultaneously describing all data with a common preset and fixed dependence type. The publications in this field for univariate and multivariate time series, including those for Bulgaria, can be found in [8–15].

The ML methods are based on a data-driven approach. The algorithm of a given ML method extracts a model closely related to the specific empirical data and can vary significantly from one dataset to another. Widely popular among ML-based methods for time series of air pollutants are as follows: artificial neural network (ANN), autoregressive neural network, deep learning, random forest (RF), support vector machine (SVM), decision trees, and more. The applications of this group of methods for time series of air pollutants and different hybrid applications are numerous. Of these, we will specify [7,15–20] and review papers [21,22]. Sophisticated models of time-series of air pollutants are built using deep learning, including the long short-term memory (LSTM) approaches, which involve deep layers and massive neurons to achieve superior descriptive ability on the training set (see for instance [23,24]).

Over the years, in the field of ML, growing attention has been directed towards ensemble methods for regression. These are characterized by the construction of numerous ensemble component learners using the same algorithm and combination of the resulting outcomes. The most frequently used techniques include resampling, bagging, and boosting. When bagging, the trees in the ensemble are parallel and independent, and when boosting, they are sequential. The main advantage of the ensemble methods is that they lead to a dramatic decrease in test set errors and a significant reduction in variance [25]. In particular, ensemble tree methods, such as RF and bagged classification and regression trees, were used in [16] to study the spatial distribution of PM_{10} concentrations, and bagged trees and RF models for $PM_{2.5}$ forecasting were built in [26]. More ensemble methods, including the tree-based, are considered in [26–31].

Despite the good qualities of ensemble methods based on weak learners, random subspaces, and randomness in the algorithms, the created models can be unstable and lacking in predictive power [27,32,33]. Selective ensemble algorithms are suggested to improve the generalization ability of the ensemble models. Overall, the goal is to identify and remove the weak learners, which reduces the performance of the ensemble. This allows the creation of a new ensemble that is not only smaller in size, but also more accurate than ensembles generated by non-selective algorithms. Under development are different approaches for the construction of selective ensembles in the case of decision trees, for ANN, or for other types of learners that use multi-objective techniques, statistical measures, pruning techniques, etc. [27,32–34].

Recently, a special class of ML approaches applying the stacked generalization paradigm has gained popularity. Stacked generalization (stacking) is a type of meta-modeling used to combine the predictions of several heterogeneous ML models using another learning algorithm to learn and predict which combinations of the models generally give better performance. The idea to combine the predictions of the regression models of time series data dates back to the 1970s in papers [35,36]. In 1992 Wolpert [37] introduced the term "stacking" for classification problems, which was also adopted for regression by Breiman in [38].

In environmental science, stacking methods are still applied relatively rarely. For example, time series data of air pollutants are modeled with weighted support vector regression (SVR), boosting, and chance theory in [39], with Lasso, AdaBoost, XgBoost, and multi-layer perceptron (MLP) stacked by SVR in [40], and with bagged trees, random subspaces, and selective techniques in [26]. The current state of ensemble methods, including stacking approaches, is presented in review papers [27,31].

The main goal of this study was to develop a new approach to the stacked generalization paradigm for the prediction of the time series of an air pollutant based on a limited

number of predictors. A novel stacking framework was designed using the ensemble methods RF and CART ensemble and bagging (CART-EB) and their rotation variants as base learners, and multivariate adaptive regression splines (MARS) as a meta-learner. The proposed approach was applied to model and forecast the average daily concentrations of PM_{10}, the main ambient air pollutant and concern of health in Bulgarian cities and also in other European cities and worldwide. The stacking framework was developed as a part of the cloud Internet of Things (IoT) platform EMULSION [41].

2. Materials and Methods

2.1. Initial Datasets

In the empirical study, time series were used for PM_{10} and accompanying time series for the region of the town of Burgas, a typical Bulgarian urban area. The data were taken from official measurements of the certified automated monitoring station in Dolno Ezerovo (42.518892, 27.375144) [42] and meteorological time series from [43]. The study area is located in the large Burgas Bay on the Black Sea, eastern Bulgaria, and includes several islandless lakes. Lukoil, the largest oil refinery in the Balkans, the Southern industrial zone, and the Pan-European corridor VIII are in close proximity. The Burgas region has a humid subtropical climate (Köppen climate classification Cfa) with continental influences. The average air temperature during the high season is 24 °C and the average winter temperature is about 4–5 °C.

An analysis of the status of air quality in the studied area of Burgas shows that the only problematic pollutant is PM_{10}, reporting exceedances of the established limits. Its corresponding mean values over the last several years are (in $\mu g/m^3$): 49.1 (2015), 45.5 (2016), 45.8 (2017), 41.8 (2018), 33.0 (2019), and 36.8 (2020). The mean levels of other harmful aerosols in the last five years have been relatively low (O_3: 48.0 $\mu g/m^3$; NO_x: 9.7 $\mu g/m^3$ NO: 3.8 $\mu g/m^3$; NO_2: 13.1 $\mu g/m^3$; CO: 0.4 mg/m^3; SO_2: 10.6 $\mu g/m^3$). The $PM_{2.5}$ values were measured with a mobile station, and no exceedances were reported. For Bulgaria, this indicator was relatively low—less than 19 $\mu g/m^3$ per year [42,44].

The initial data consisted of daily mean concentrations of PM_{10} over six years and four months, from 1 January 2015 to 28 April 2021, or a total of $N = 2310$ days. In addition to variable *PM*10, nine meteorological time series were taken into consideration as follows: *MaxT*, (°C)—maximum daily air temperature, *MinT*, (°C)—minimum daily air temperature, *WindSpeed*, (m/s)—wind speed, *Humidity*, (%)—relative air humidity, *Pressure*, (mbar)—atmospheric pressure, *Cloud*, (%)—cloud conditions, *Rain* (cm), *Weather*$ (categorical), and *WindDir*$, (categorical)—wind direction (The categorical variable *WinDir*$ takes 16 different values, corresponding to the geographical wind direction. *Weather*$ assumes the standard string notions, such as sunny, light rain shower, overcast, blizzard, etc.

The analyses considered the permissible official European and national upper limits for PM_{10}. They are as follows: average daily mean up to 50 $\mu g/m^3$, which must not be exceeded more than 35 times within one calendar year, and 40 $\mu g/m^3$ on average per year [4,5]. The prescribed upper limits by the World Health Organization (WHO) for PM_{10} are 50 $\mu g/m^3$ daily and 20 $\mu g/m^3$ annually.

Table 1 shows the descriptive statistics of the initial continuous type data.

Table 1. Descriptive statistics of the initial data for PM_{10} and meteorological variables.

Statistics \ Variable	*PM10* ($\mu g/m^3$)	*MaxT* (°C)	*MinT* (°C)	*WindSpeed* (m/s)	*Humidity* (%)	*Pressure* (mbar)	*Cloud* (%)	*Rain* (cm)
N valid	2175	2310	2310	2310	2310	2307	2310	2310
N missing	135	0	0	0	0	3	0	0
Mean	41.91	17.33	10.79	13.93	70.61	1016.73	34.67	2.02
Median	37.01	17.00	11.00	13.00	71.00	1016.00	26.00	0.00
Std. Deviation	19.223	8.503	7.129	6.464	9.938	6.993	28.73	5.6851
Variance	369.595	72.293	50.827	41.781	98.774	48.901	825.55	32.320
Skewness	1.980	−0.088	−0.169	1.431	−0.166	0.258	0.778	5.155
Kurtosis	8.361	−0.901	−0.794	2.848	−0.415	0.247	−0.538	35.395
Minimum	4.77	−8	−10	2	40.00	990	0.0	0.0
Maximum	248.80	37	25	51	96.00	1040	100.0	69.5

Some problematic values of PM_{10} were observed. For example, the maximum value of PM_{10} in the Burgas region is 248.8 $\mu g/m^3$, and the mean of PM_{10} for the six years is 41.91 $\mu g/m^3$. The extreme value of 248.8 $\mu g/m^3$ in the period of 26–30 March 2020 was due to unfavorable weather conditions caused by strong winds and transboundary dust particles from the Aral Sea region [42]. Since the maximum of PM_{10} is the only outlier, this value was substituted with the second largest one. The number of missing values for PM_{10} was 135, or about 6.2% of the dataset. All missing data were replaced using linear interpolation in the modeling process and denoted by the same variable names. In addition, Table 1 shows some large values of the coefficients of skewness (1.980) and kurtosis (8.361) for *PM*10. This could affect the direct application of classical regression methods.

Figure 1 shows the sequence plot of *PM*10, where the horizontal red line indicates the permissible upper daily limit of 50 $\mu g/m^3$ in the European union. Large peaks and exceedances are observed during the winter months.

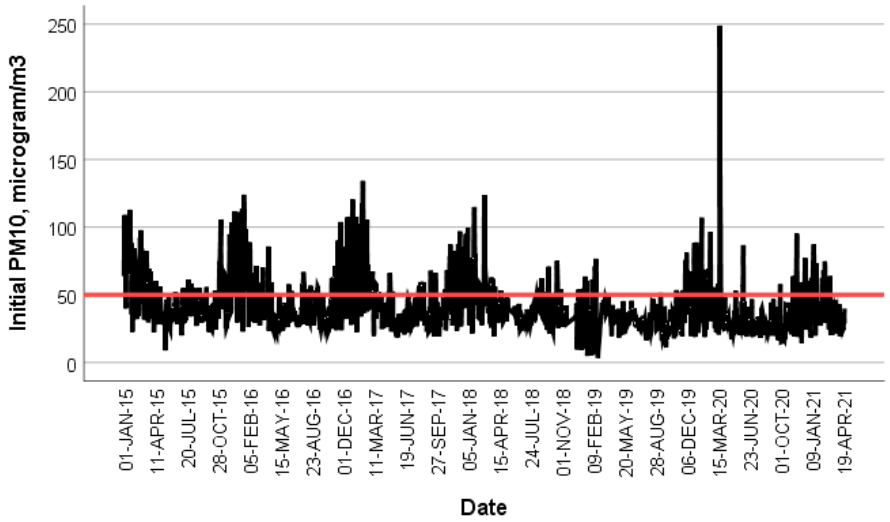


Figure 1. Sequence plot of the measured daily concentrations of PM_{10} in the Burgas region, Bulgaria. The horizontal red line indicates the permissible upper daily limit value of 50 $\mu g/m^3$ for PM_{10} according to European and national standards.

2.2. Model Assumptions

Every model has its assumptions and limitations. The current study aims to demonstrate the construction of models for PM_{10}. According to reports from the European agency, this is Bulgaria's most problematic pollutant over recent years [44]. Measurements of meteorological factors have been used as predictors; however, this does not limit the suggested approach. If easy retrieval of data regarding other pollutants and factors is possible, their in-

fluence can be modeled directly. As compensation, the use of lagged variables of *PM*10 and meteorological time series includes the stochastic influence of all other factors, including critical inventory factors, emission intensity, and other measurements or yet-undetermined factors. We should note that the suggested framework is intended for short-term forecasts of PM_{10} levels as a component of mobile applications within a developed cloud IoT platform. For this reason, the study is concentrated on algorithms with a preferably small initial number of variables.

2.3. Stacking Generalization

On the one hand, stacking can be viewed as a generalization of many ensemble methods. On the other hand, it can be viewed as a specific combination method that combines by learning [27]. According to Wolpert's terminology [37], the original data and the models constructed for them in the first stage are referred to as level-0 data and level-0 models (or base-learners), respectively. Then, the obtained set of cross-validated predictions and the second stage learning algorithm are considered as the level-1 data and the level-1 generalizer (or meta-learner), as in [37,38]. The target variable is the same for the two levels. For this approach, it is believed that a combination of several models outperforms a single "best" model [45]. It should be noted that the concept of "stacking" assumes a combination of the results from various algorithms, unlike the concept of the "ensemble method", where one algorithm is applied to various subsets. A simplified outline of the idea of stacking is shown in Figure 2.

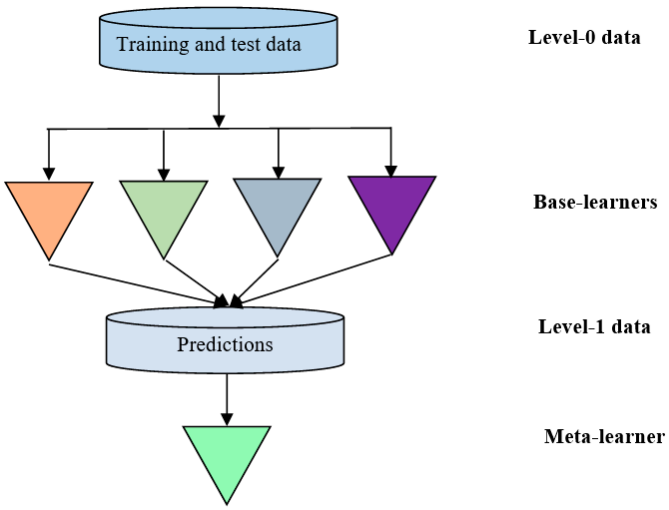


Figure 2. General stacking framework.

2.4. Proposed Stacking Framework

Following the general guidelines from Section 2.3, we propose a stacking framework in five steps, as illustrated in Figure 3.

Step 1: Level-0 data

The target variable is *PM*10—measured PM_{10} values. However, we use the variable *PM*10_7, derived from *PM*10, where the last seven values are deleted, to validate the predictive power of the models. In terms of analysis, the considered influencing variables on the PM_{10} concentrations are categorized into four groups of predictors—environmental parameters (meteorological variables), rotated principal component variables (PCs), lagged variables (of *PM*10, meteorological time series, and PCs), and temporal variables (*Day*, *Year*$, and *Season*$).

Step 2: Construction of base models

In stacking, it is considered that the best results can be obtained by combining three to eight different base learners at level 0 [37,38,45]. Therefore, in our study, we chose to select four base models. With different combinations of level-0 data we will build several level-0 models using RF, rotation RF (RotRF), selective CART-EB (Sel-EB), and selective rotation CART-EB (SelRot-EB) methods. The corresponding *RotRF* and *SelRotEB* models are constructed using PCs and lagged PCs instead of meteorological factors. Sel-EB and RotSel-EB type models are built by the proposed simplified selective ensemble algorithm.

Step 3: Training and evaluation of the base models

In this step, we perform the training and cross-validation of the generated set of the base models. Models based on RF are trained using the standard out-of-bag (OOB) subsamples and the models built on CART-EB, using 10-fold CV. A Wilcoxon signed rank test (WSRT) is applied to select heterogeneous models and check the required diversity [46]. In addition, diagnostics is performed on the residuals of the models using partial autocorrelation functions (PACFs). The statistics and performance measures of successful base models are calculated.

Step 4: Level 1 data (Predictions)

Four end models with the best statistical indicators are selected from the many base models. Their predicted values for *PM*10_7, including the last 7 days of the time series, are considered level-1 data.

Step 5: Stacked models

The selected four base models are regressed with the MARS method. Several MARS models were built, linear and up to the second degree of interaction between predictors. The training and evaluation are conducted using a 5- or 10-fold CV. The resulting stacked models with the best statistics are denoted by *S-MARS*. Finally, the model's predicted and forecasted values are compared with those of the base models and with the measured values of PM_{10}.

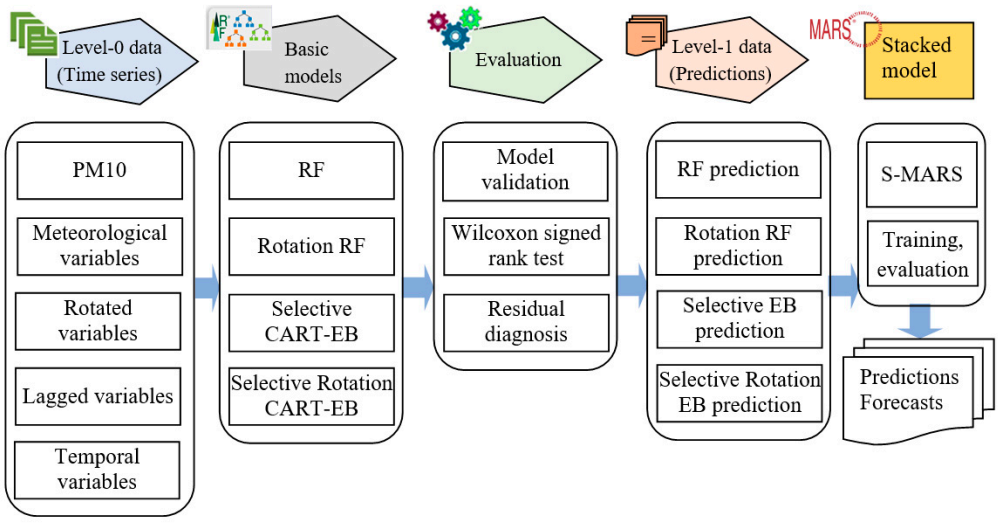


Figure 3. Scheme of the proposed stacking framework.

2.5. Methods

To build the models, we used the following methods, described briefly below: (1) factor analysis (FA) with principal component analysis (PCA) to obtain rotated space and relevant principal components (PCs) from the multicollinear independent variables; (2) RF; (3) CART-EB and selective CART-EB; and (4) MARS.

2.5.1. Principal Component Analysis and Factor Analysis to Obtain Rotation Space of PCs

Principal component analysis was introduced by Hotteling [47] as a technique for deriving linear combinations of collinear variables into uncorrelated new variables (PCs) that capture the most possible variance. For this purpose, the overall correlation matrix and its eigenvalues are initially calculated, and the corresponding PCs are extracted [48]. In the next step, one can use FA to retain all or some PCs (also called factors) to reduce the data dimensionality. The procedure ends with the rotation of the factor variables and thus a rotated space or solution to the problem is formed. A factor model is adequate when each original variable is grouped to only one factor, and the factors themselves are well separated from each other. Studies that apply PCA and FA abound in environmental science (see for instance [8,15]).

2.5.2. Random Forest and Rotation RF

Random forest is a powerful statistical ML and ensemble technique developed by Leo Breiman, based on his bagging algorithm and the random subspace method of Ho [49–51]. RF is suitable for regression and classification, and can handle all types of data—numeric, categorical, and nominal. Moreover, there are no special restrictive assumptions about the distribution of the data. The RF algorithm builds hundreds of independent binary trees using identically distributed randomized learning samples. The initial set of data with volume N is randomly divided into two, of which about one third is used for a test set called out-of-bag (OOB). The other part is used to form the learning sets with volume N through the bootstrap algorithm for each tree in the ensemble. To avoid overfitting and reduce variance, each model is trained with OOB. In this way, a large ensemble of independent decision trees is generated, which form the forest, each of which builds its predictive model of the target. The final RF model is obtained after averaging each predicted value from the ensemble component trees. The main input hyperparameters of RF are the number of trees T in the ensemble, the minimum cases in the parent node of the tree, and Q—the number of randomly selected predictors from all given M features at each node of each tree. Usually, $Q = 3$ [49].

Rotation RF is a variant of RF that uses rotated PCs in the set of features, where PCs replace some initial variables from which they are generated.

2.5.3. Proposed Selective CART Ensemble and Bagging Algorithm and its Rotation Variant

The general CART-EB is a learning method that aims to reduce variance within a noisy dataset. It is applicable for regression and classification problems. To improve the stability of the CART method, bagging (abbreviated from bootstrap aggregation) is applied in an ensemble of several trees [50]. In bagging, each tree is built independently with a random sample of data in a training set, by randomly extracting cases from the initial dataset with replacement. A random subset of Q features is selected at each decision, split as in the RF algorithm. Additionally, an initial CART tree is built on the full sample. After calculating predictions from all trees, the values for each case are averaged and the final target prediction is obtained [27,50]. A general scheme of the CART-EB algorithm is shown in Figure 4.

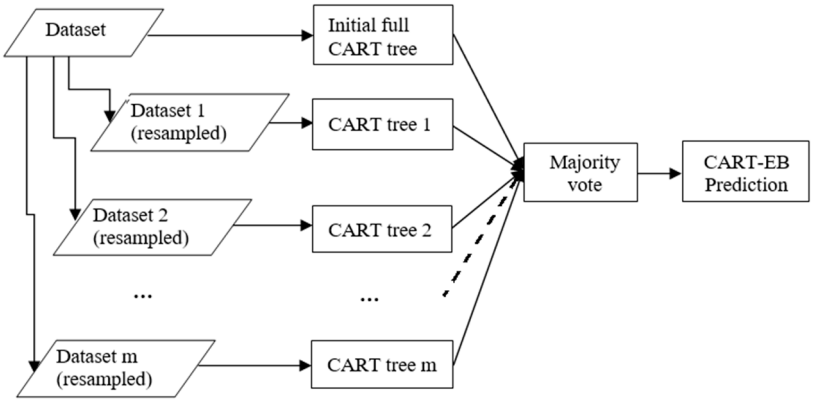


Figure 4. Flowchart of the CART-EB algorithm.

In this study, we propose a new simplified selective CART-EB algorithm. In our scheme, the statistical measure of the quality of the model is the index of agreement (*IA* or *d*), calculated by the expression [52]:

$$IA = d = 1 - \frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} \left(|P_i - \overline{O}| + |O_i - \overline{O}| \right)^2} \quad (1)$$

where P_i is predicted values, O_i and $\overline{O}$ are observed values and its mean, respectively, and N is the sample volume. IA is a dimensionless and bounded measure in $[0,\ 1]$, with values closer to 1 indicating better agreement between the model and the target variable.

The suggested selective algorithm is as follows: Let ensemble E with n_s component trees is a model of the observed target variable O. We use d_{E} to denote its IA. The trees are removed one by one and the IAs of the corresponding reduced ensembles are calculated as $d_j,\ \ j = 1, 2, \ldots,\ \ n_s$. Each component tree T_j, such that $d_j > d_E$, is a candidate for removal from the ensemble. T_j is also known as "negative" tree. It is easy to understand that the maximum increase in d_{E} is obtained through the simultaneous removal of all the negative trees. At the same time, the algorithm presents a new type of diversity.

The selective rotation CART-EB algorithm is a variant of selective CART-EB, where the PCs are used as predictors with their lagged variables to replace the original continuous-type meteorological samples.

2.5.4. Multivariate Adaptive Regression Splines

MARS is a non-parametric data mining method developed in [53]. It is designed to predict numeric outcomes and provides models similar to traditional regression that may include partial terms of the nonlinear type. The dependent variable $y = y(X)$ can be predicted using p independent variables $X = \left(X_1, X_2, \ldots, X_p\right)$ defined in $\mathbb{R}^N$. The MARS model $\hat{y} = \hat{y}_{[L]}$ is presented in the form:

$$\hat{y}_{[M]} = b_0 + \sum_{j=1}^{L} b_j BF_j(X) \quad (2)$$

where $b_0, b_j,\ \ j = 1, 2, \ldots, L$ are the coefficients, BF_j are its basis functions (BF), and L is their number. The one-dimensional BF is written as:

$$BF_j(X) = \max_{X_k}(0, X_k - c_{k,j}) \ \text{ or } BF_j(X) = \max_{X_k}(c_{k,j} - X_k, 0)$$

where the nodes (points of slopes) are denoted by $c_{k,j} \in X_k$. The non-linear interactions are presented as the products of other BFs. The usual MARS hyperparameters are the

maximum number of BFs and the maximum degree of interactions. For each model, the MARS algorithm defines variables and nodes so as to minimize a predefined loss function, such as the mean square error. More about MARS methodology and implementation can be found in [53].

2.5.5. Performance Measures

In this work, we apply four typical evaluation measures to assess and compare the accuracy and predictive ability of the constructed models. In addition to the index of agreement (*IA*) in Equation (1), the other three evaluation indices are root mean squared error (*RMSE*), fractional mean bias (FB), and the coefficient of determination R^2, calculated by the expressions:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - P_i)^2}, \qquad FB = 2\frac{\overline{O} - \overline{P}}{\overline{O} + \overline{P}}, \qquad R^2 = \frac{\left\{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\left(O_i - \overline{O}\right)\right\}^2}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2 . \sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2} \tag{3}$$

where P_i and O_i stand for model predicted and target values, respectively, and $\overline{P}$, $\overline{O}$ are their mean values. *RMSE* is used to assess the model accuracy. The FB index measures the tendency of a model to over-predict with values close to 2 and under-predict with values close to -2. The coefficient of determination R^2 is a measure of the proportion of the total variation of target values explained by the model and expresses linear relationships. A good predictive model should have a value close to 0 for *RMSE* and FB, and a value close to 1 for *IA* and R^2. Statistical analyses were conducted using the Salford Predictive Modeler (SPM) from Minitab, IBM SPSS, and the author's code in Wolfram Mathematica software.

3. Results

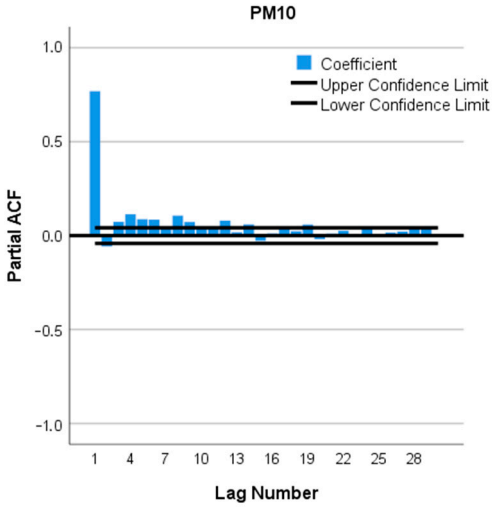
The analysis will be carried out according to the framework in Section 2.4.

3.1. Pre-Processing Level-0 Data

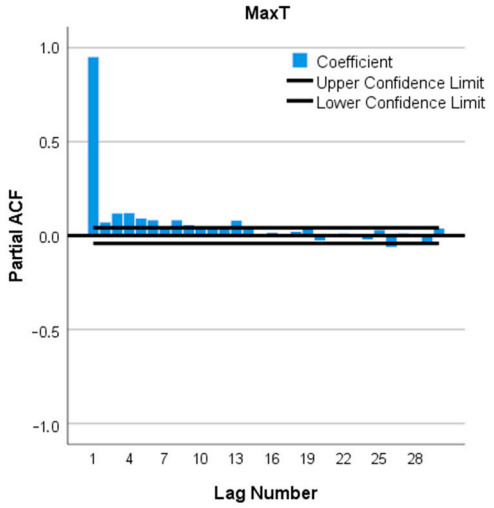
The continuous time series in the initial data was *PM*10 and the seven meteorological variables *MaxT*, *MinT*, *WindSpeed*, *Humidity*, *Pressure*, *Cloud*, and *Rain*. We will carefully investigate their characteristics for further use.

3.1.1. Construction of the Additional Samples

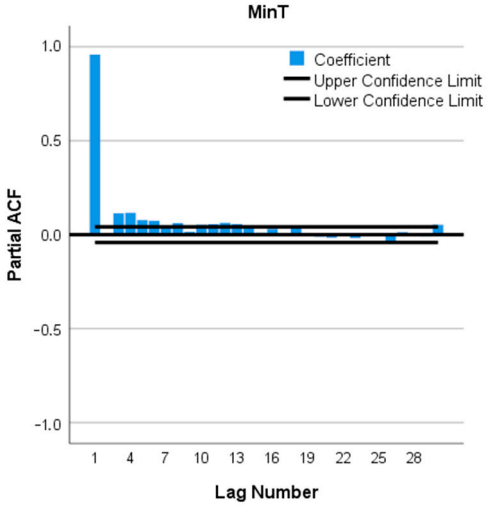
Along with the initial samples from Section 2.1, we involved additional variables that affect PM_{10}. With the aid of the PACF, we found that *PM*10 and the other seven variables in Table 1 had large PACF coefficients in lag 1. Figure 5 shows the PACF of *PM*10 and the seven continuous meteorological factors. These results indicate that lagged variables can be used as additional data samples. These variables will be denoted by "*name*"<1>. Additionally, using PCA and FA, some rotated variables were extracted from the independent continuous meteorological variables. In addition, we introduced three temporal variables, namely *Day* with values 1, 2, . . . , *N* (ordinal), the nominal *Year*$, and *Season*$ with four values (winter, spring, summer, and autumn).



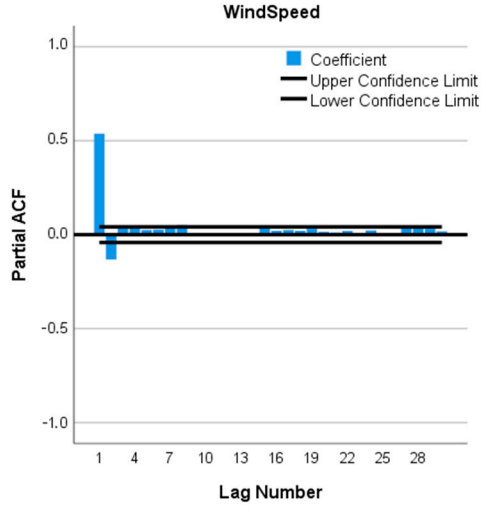
(**a**)



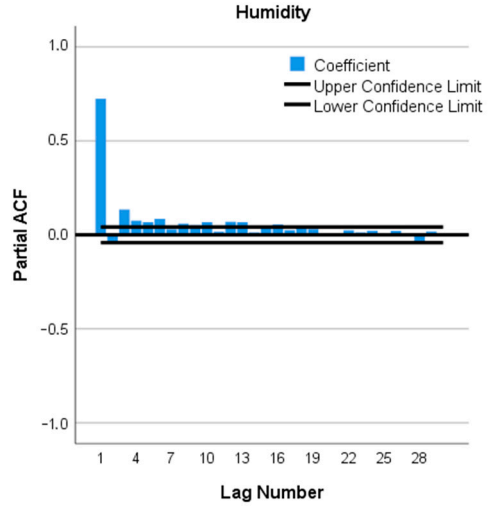
(**b**)



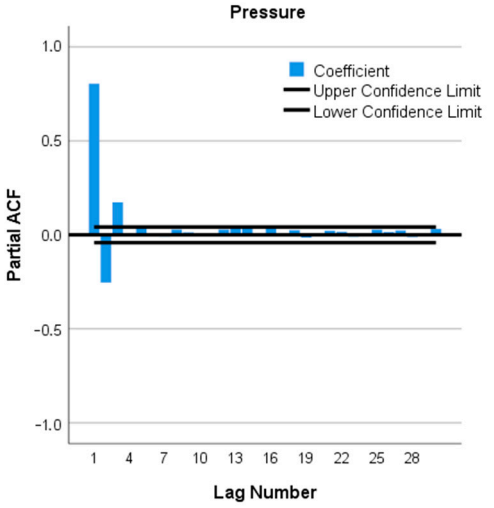
(**c**)



(**d**)

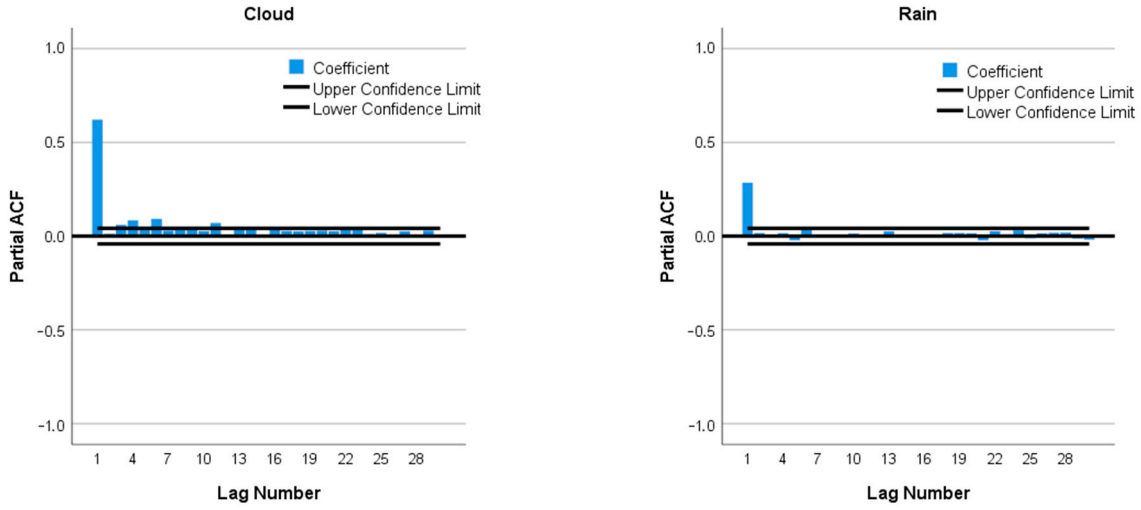


(**e**)



(**f**)

Figure 5. *Cont.*



(**g**) (**h**)

Figure 5. The PACF of the studied time series.

3.1.2. Determination of Principal Components

We transformed the seven continuous meteorological variables using PCA and FA. Their correlation matrix had a small determinant, det = 0.013. In addition, the Kaiser–Meyer–Olkin Measure of Sampling Adequacy (KMO) was equal to 0.636 > 0.5, and Bartlett's test of sphericity was significant with a p-value equal to zero. From these statistics, we concluded that the considered seven variables were multicollinear and suitable for PCA and FA. Six PCs (factors) were extracted with a total variance of 99.555%. Thus, the loss of information was negligible, within 0.5%. After factor rotation using the Promax method, we obtained the pattern matrix shown in Table 2. Of the six factors ($PC1$, $PC2$, . . . , $PC6$), the $PC1$ groups *MaxT* and *MinT* and the other variables are single factors. All the variables are very well separated in PCs, as they have large loadings with only one factor (the corresponding term from the main diagonal), and relative to other factors, the loadings are negligible, with absolute values less than 0.1. Therefore, we can assume that the factor analysis is statistically valid [48].

Table 2. Pattern matrix of the rotated seven meteorological variables [1].

Variable	Component (PC)					
	*PC*1	*PC*2	*PC*3	*PC*4	*PC*5	*PC*6
MinT	**1.046**	0.024	−0.009	0.033	0.034	0.053
MaxT	**0.921**	−0.028	0.011	−0.040	−0.043	−0.064
WindSpeed	0.001	**1.001**	0.001	−0.002	−0.002	−0.002
Rain	−0.001	0.001	**0.998**	0.002	0.001	0.003
Humidity	−0.001	−0.002	0.002	**0.998**	−0.003	0.001
Pressure	−0.005	−0.002	0.001	−0.003	**0.997**	−0.004
Cloud	−0.004	−0.001	0.005	0.005	−0.004	**0.993**

[1] Extraction method: principal component analysis. Rotation method: Promax with Kaiser normalization. The highest loadings are in bold.

The resulting six rotated factors form a rotated subspace, which we used to build some of the base models. We should especially note that the established multicollinearity between the meteorological variables was not very strong when KMO = 0.636 << 1. This allowed us to use both the initial meteorological variables and their PCs as predictors in the ML models, but not simultaneously. The PACFs of the extracted PCs have the same character as given in Figure 5.

To build the heterogeneous base models, we used the following four datasets of predictors:

- **Dataset A:** *PM*10_7<1>, *MinT*, *MaxT*, *WindSpeed*, *Rain*, *Humidity*, *Pressure*, *Cloud*, *Weather*$, *WinDir*$, *Day*, *Year*$, *Season*$;
- **Dataset B:** *PM*10_7<1>, *PC*1, *PC*2, *PC*3, *PC*4, *PC*5, *PC*6, *Weather*$, *WinDir*$, *Day*, *Year*$, *Season*$;
- **Dataset C:** *PM*10_7<1>, *PM*10_7<2>, *MinT*, *MaxT*, *WindSpeed*, *Humidity*, *Pressure*, *Cloud*, *Rain*, *MaxT*<1>, *MinT*<1>, *WindSpeed*<1>, *Humidity*<1>, *Pressure*<1>, *Cloud*<1>, *Rain*<1>;
- **Dataset D:** *PM*10_7<1>, *PC*1, *PC*2, *PC*3, *PC*4, *PC*5, *PC*6, *PC*1<1>, *PC*2<1>, *PC*3<1>, *PC*4<1>, *PC*5<1>, *PC*6<1>, *Day*, *Year*$, *Season*$.

3.2. Construction and Comparison of the Base Models

With different combinations of level-0 data, we built several level-0 models using RF, rotation RF (RotRF), selective CART-EB (Sel-EB), and selective rotation CART-EB (SelRot-EB) algorithms. The corresponding RotRF and SelRot-EB models were constructed using PCs and lagged PCs instead of meteorological factors. The Sel-EB and SelRot-EB type models were built by the proposed simplified selective ensemble algorithm.

3.2.1. Diversity Strategies

In order to build heterogeneous base models, we explored the following five strategies:

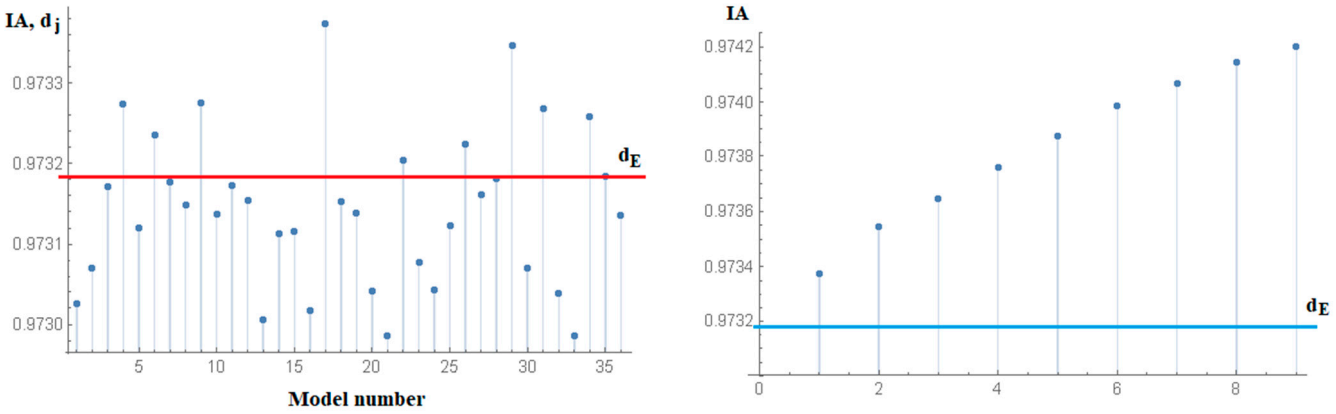
1. Use of test and training subsets of varying size and composition, based on initial and additional data samples;
2. Selection of different algorithms as base-learners;
3. Building selective ensembles to improve some base-learners;
4. Variation of the hyperparameters of the algorithms;
5. Using different model validation techniques.

To test the diversity of base models, the use of WSRT is recommended since it is less restrictive to the sample distribution, nature of the error, and less susceptible to outliers [46]. The null hypothesis of WSRT is that the differential series $d(t) = A(t) - B(t)$ has zero median, where $A(t)$, $B(t)$ are two related samples.

Several models with each of the four methods (RF, RotRF, Sel-EB, and SelRot-EB) and datasets A, B, C, and D were built. The RF-based models were trained with the OOB test sample and the rest with 10-fold cross-validation. The hyperparameters were the number of trees in the ensemble (T), minimum cases in parent node 10, minimum cases in terminal node equal to 5 (recommended in [48]), and the number of randomly selected features at each node of each tree was taken as Q = 3 or Q = 4.

3.2.2. Construction of Selective CART-EB Ensembles

We will take a closer look at the results of our suggested simplified algorithm for constructing a selective CART ensemble with the following example. Using dataset D, we generate a standard RotEB ensemble model *RotEB*36 with $n_s = 36$ trees. Its initial CART tree is represented below by number 36. The value of IA for the model is $d_E = 0.97318$. We remove the trees one by one and calculate d_j, $j = 1, 2, \ldots, n_s$ of all reduced ensembles with 35 trees. Figure 6a shows the graph of the derived updated values of IA, together with d_E. We identify nine negative trees (above the line d_E). The removal of each of the nine trees should lead to an improvement of the ensemble's IA by $d_j - d_E$. In our case, the negative trees' numbers in descending order of d_j are: 17, 29, 4, 9, 31, 34, 6, 26, and 22.



(**a**) (**b**)

Figure 6. Example of selective ensemble algorithm, where the horizontal lines represent the d_E value of the initial ensemble model *RotEB*36. (**a**) Updated values of IA following the removal of each of the ensemble's trees from *RotEB*36; (**b**) IA values of the selective ensembles (4).

Of the many combinations, we choose to remove groups of negative trees in the following manner. We designate the sequence of negative trees as:

$$Mt = \{m17,\ m29,\ m4,\ m9,\ m31,\ m34,\ m6,\ m26,\ m22\}$$

and construct the sums $\sum_{i=1}^{k} Mt(i), \quad k = 1, 2, \ldots, 9$. The resulting nine new selective ensemble models are calculated using the following expression:

$$sel\_k = \frac{n_s.RotEB36 - \sum_{i=1}^{k} Mt(i)}{n_s - k}, \quad k = 1, 2, \ldots, 9; \quad n_s = 36 \tag{4}$$

Figure 6b illustrates the increase in the IA of ensembles *sel_k*. Table 3 contains the evaluation measures for the derived selective ensembles (4). We can see that all the statistical indicators improve as k increases. From the following models, we will choose the last (for $k = 9$) as a base model, denoted by *SelRotEB*1. It contains 27 trees.

Table 3. Evaluation statistics of the *RotEB*36 and the 9 selective ensemble models from (4).

Ensemble Model	Number of Trees	RMSE	FB	AI	R^2
RotEB36	36	5.65492	0.00680	0.973181	0.9168
Sel_1	35	5.63835	0.00643	0.973373	0.9169
Sel_2	34	5.62213	0.00654	0.973545	0.9169
Sel_3	33	5.61483	0.00635	0.973646	0.9170
Sel_4	32	5.60487	0.00614	0.97376	0.9173
Sel_5	31	5.59607	0.00611	0.973873	0.9174
Sel_6	30	5.58497	0.00613	0.973985	0.9176
Sel_7	29	5.57812	0.00596	0.974066	0.9180
Sel_8	28	5.57221	0.00587	0.974144	0.9182
Sel_9	27	5.56676	0.00575	0.974201	0.9186

We applied the same procedure to the CART-EB model built with 31 trees using dataset C. Eight negative trees were obtained. The constructed maximal selective ensemble model, according to formula of the type (4) for $k = 1, 2, \ldots, 8, \quad n_s = 31$ was chosen as the base model and is denoted by *SelEB*1.

3.2.3. Base Models

Several candidates for base models were built for the target variable *PM*10_7. Table 4 presents the construction parameters and statistics of the four selected base models *RF*1, *RotRF*1, *SelEB*1, and *SelRotEB*1. Specific sets of the proposed five types of diversities were applied. Different datasets were used. There were some common elements, such as *PM*10_7 <1>, included in all models. The other main predictors in the first and third models were the initial meteorological variables, while the second and fourth used their respective PCs (rotated subspace). The RF models contained about 10 times more trees in the ensembles *RF*1 and *SelEB*1, and *RotRF*1 and *SelRotEB*1.

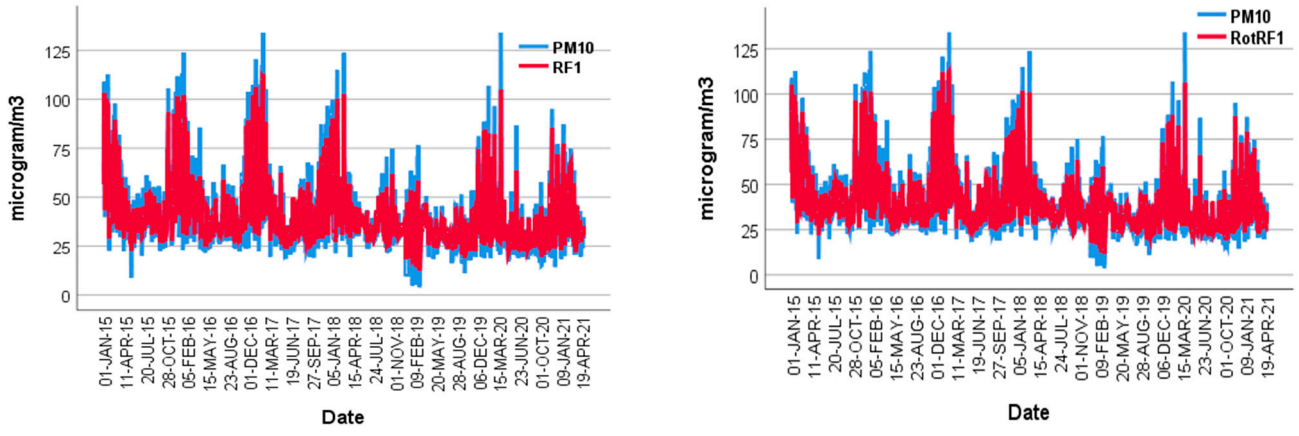
Table 4. Parameters and statistics of the four selected base models [a].

Model	Method	Dataset 0	Number of Trees T	Q	Trained Sample R^2	RMSE	FB	IA, d	R^2
*RF*1	RF	A	200	3	0.7126	5.3020	0.0043	0.9758	0.9358
*RotRF*1	RotRF	B	400	4	0.7114	5.0588	0.0045	0.9784	0.9393
*SelEB*1	Sel-EB	C	23 out of 31	3	-	5.8095	−0.0011	0.9721	0.9082
*SelRotEB*1	SelRot-EB	D	27 out of 36	3	-	5.5668	0.0058	0.9742	0.9186

[a] The first two models were trained with the OOB test sample and the rest with the 10-fold cross-validation technique.

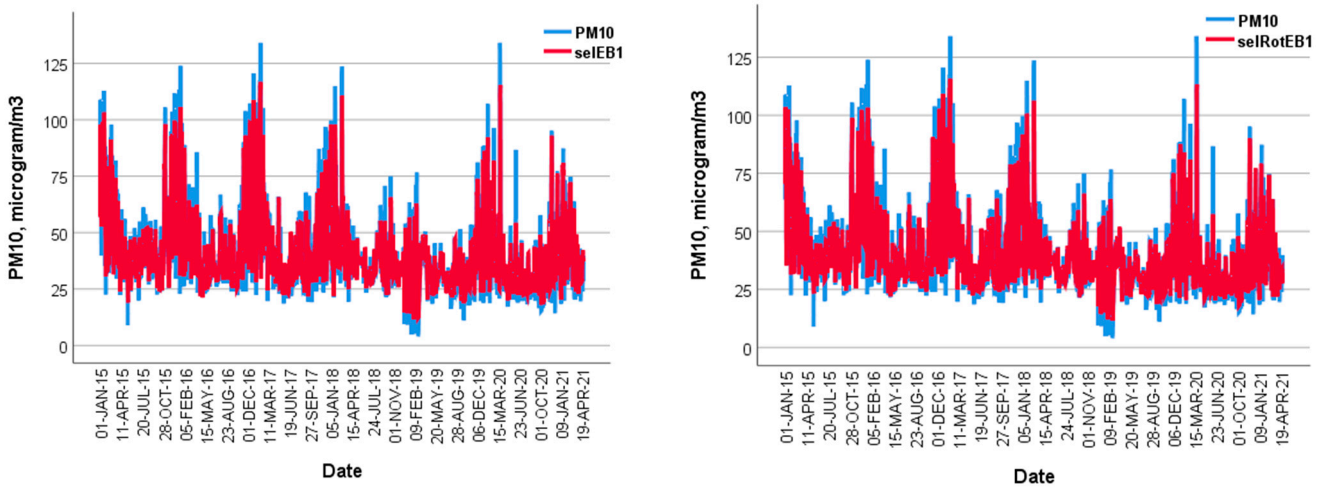
From the calculated evaluation measures (1) and (3) in Table 4, it can be seen that the statistics are close. The most favorable statistics for RMSE, AI, and R^2 were for the model *RotRF*1. On the other hand, *SelEB*1 showed the lowest indicators (except for FB).

A comparison of the prediction quality of the four base models for the initial *PM*10 data is shown in Figure 7. Due to the large amount of data, no visible difference was observed. The comparison with the sorted *PM*10 data using scatter plots is illustrated in Figure 8. From both Figures 7 and 8, it can be seen that the highest and somewhat lowest values of the PM_{10} concentrations were not very well predicted. As already mentioned, this is a common shortcoming of ensemble methods that average their predictions.



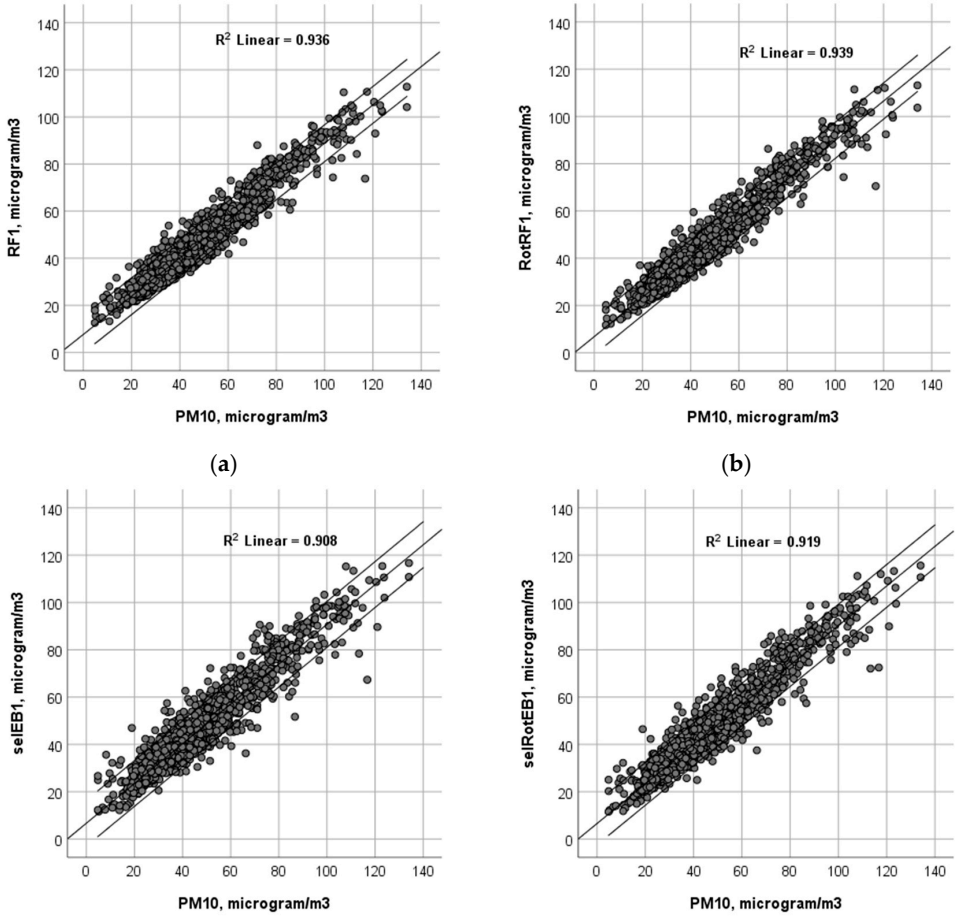
(**a**) (**b**)

Figure 7. *Cont.*



(**c**) (**d**)

Figure 7. Quality of the coincidence of the measured values of PM_{10} and the values predicted by the base models with 5% confidence intervals for: (**a**) model *RF*1; (**b**) model *RotRF*1; (**c**) model *selEB*1; and (**d**) model *SelRotEB*1.



(**a**) (**b**)

(**c**) (**d**)

Figure 8. Quality of the coincidence of the measured values of PM_{10} and the ones predicted by the base models with 5% confidence intervals: (**a**) model *RF*1; (**b**) model *RotRF*1; (**c**) model *SelEB*1; and (**d**) model *SelRotEB*1.

3.3. Evaluation of the Base Models and Variable Importance

Following our proposed stacking framework in Sections 2.4 and 3.2.1, we must also analyze the diversity and residuals of the base models. Another important task is to determine the degree of influence of predictors in the models. The most important of these are meteorological variables.

3.3.1. Checking for Diversity

The results of the WSRT are shown in Table 5. All the significance values (p-values) were less than 0.01, leading us to reject the null hypothesis of the test and conclude that there were significant differences between the four base models at the confidence level of 95%.

Table 5. Wilcoxon signed rank test results to check the difference between the four selected base models.

	Test Statistics [a]					
	***RF*1-*RotRF*1**	***RF*1-*SelEB*1**	***RF*1-*SelRotEB*1**	***RotRF*1-*SelEB*1**	***RotRF*1-*SelRotEB*1**	***SelEB*1-*SelRotEB*1**
Z (Standardized test statistics)	−2.666 [b]	−2.721 [c]	−2.925 [b]	−6.297 [c]	−2.270 [c]	−5.313 [b]
Asymptotic Sig. (2-tailed)	0.008	0.007	0.003	0.000	0.023	0.000

[a] Wilcoxon signed ranks test; [b] based on negative ranks; [c] based on positive ranks.

3.3.2. Analysis of Residuals

To assess the reliability of the models, we need to study their residuals in more detail. First, we must check whether the residuals are merely white noise with no significant serial correlation. However, the author of [54] states that "in fact, there is currently no general diagnostic statistics for nonlinear autocorrelation relationships." In particular, this also applies to ML models, including models that use lagged time series of the dependent variable and/or of predictors. In these cases, standard statistical tests, such as Durbin–Watson, Ljung–Box, Breusch–Godfrey, etc., which are valid for the assumption of linearity, cannot be used [54]. For nonlinear time series models, it is recommended to study the autocorrelation functions (ACF) plot of the model residual [55,56]. If the data constitute a large sample from an independent white noise sequence, approximately 95% of the sample autocorrelations should lie between the bounds $\pm 1.96/\sqrt{N}$ [57]. For our four models, the results are shown in Figure 9. We can conclude that models' residuals are white noise and do not contain significant autocorrelation.

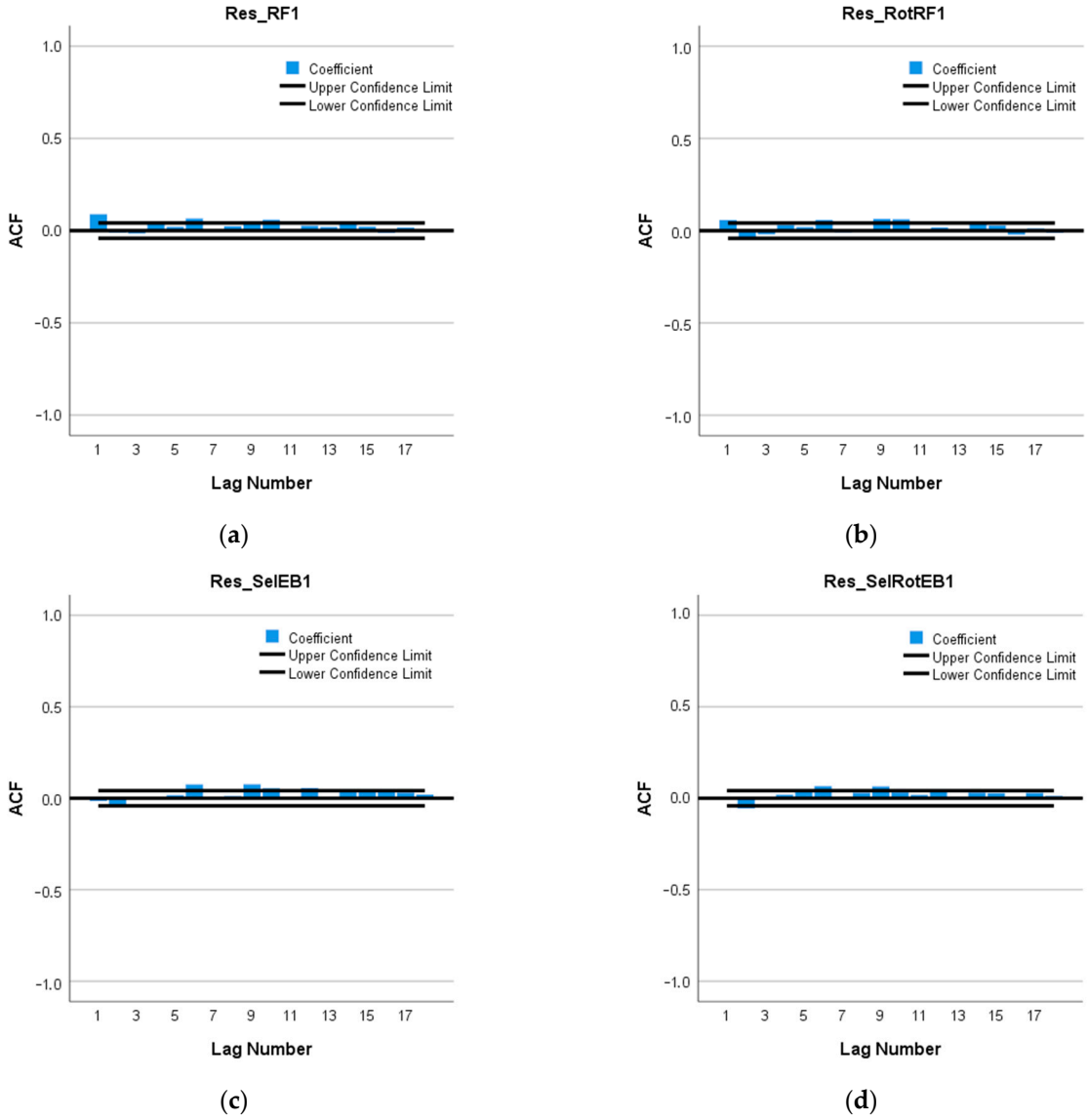


Figure 9. Residual ACF of the selected four base models of PM_{10}: (**a**) model *RF*1; (**b**) model *RotRF*1; (**c**) model *SelEB*1; and (**d**) model *SelRotEB*1.

3.3.3. Variable Importance

An important part of modeling air pollutants is assessing the impact of the predictors used in the models. Table 6 shows the relative weights of the predictors in the base models. All the models determined a 100% primary effect of the lagged variable *PM*10_7 <1>, which in this case was shown to some extent as a stochastic factor, including measured and immeasurable complex effects from the previous day's PM_{10} concentrations.

Of the meteorological predictors (including their lagged variables), the more significant, weighing more than 9 to 27 relative units, were *MinT* and *MaxT* (or *PC*1), followed by *WindSpeed* (*PC*2). The variables *WinDir*$ and *Weather*$ had less importance. These factors determine the main meteorological conditions for the formation and retention of PM_{10} in ambient air [58]. Therefore, we can assume that our models correctly detect the contribution of the conditions under consideration, including the additional data samples. The temporal variable *Day* also had a large relative weight, with 12, 10, and 24 relative units in the first, second, and fourth models, respectively.

Table 6. Relative variable importance in the base models [a].

Base Model	Variable Importance
*RF*1	*PM10_7*<1> (100), *MinT* (12), *Day* (12), *WindSpeed* (9), *MaxT* (9), *WinDir*$ (6), *Rain* (4), *Cloud* (3), *Humidity* (2), *Pressure* (1)
*RotRF*1	*PM*10_7<1> (100), *PC*2 (13), *Day* (10), *PC*1 (9), *Season*$ (7), *WinDir*$ (6), *PC*3 (4), *PC*5 (1)
*SelEB*1	*PM*10_7<1> (100), *PM*10_7<2> (50), *MinT*<1> (17), *Pressure*<1> (16), *MaxT*<1> (15), *WindSpeed* (14), *MaxT* (12), *Cloud* (11), *Pressure* (10), *Rain* (9), *Humidity* (8), *Cloud*<1> (8), *WindSpeed*<1> (8), *Humidity*<1> (7), *Rain*<1> (5)
*SelRotEB*1	*PM*10_7<1> (100), *PC*1<1> (27), *Day* (24), *Weather*$ (20), *PC*5<1> (19), *WinDir*$ (18), *PC*1 (17), *PC*2 (16), *PC*5 (13), *PC*2 (12), *PC*6 (11)

[a] Variables with zero importance are not included.

From the results so far, we can conclude that the selected base models are statistically tested and reliable and can be used for forecasting. As shown in Table 4, the RF-based models had higher coefficients of determination, reaching R^2 = 94%, compared to those with selective CART-EB, for which R^2 was about 91–92%. Similarly, the RMSE of the RF and RotRF models was smaller. As the number of trees increased, the statistics stabilized, which is especially characteristic of the CART-EB algorithm. When the statistical indicators are close, the simpler model is chosen; however, it must be one that retains good diversity in the final selection of base models.

3.4. Construction and Analysis of Stacked MARS Models

We used the predicted values from the base models as predictors in constructing the MARS models for target *PM*10_7.

3.4.1. Construction and Evaluation Statistics of the Stacked Models

Different values of the hyperparameters of the MARS algorithm were set. In this case, the main hyperparameter was the maximum number of basis functions (BFs) and the order of interactions between them. We only used linear and up to the second degree of BFs. Table 7 shows the statistics of the selected stacked models *S-MARS*1 and *S-MARS*2. The linear model, *S-MARS*1, was obtained with a set of 80 BFs and, after pruning in the backward step of the algorithm, L = 29 BFs remained. The second-order model, *S-MARS*2, was built using a maximum of 60 BFs, resulting in L = 30 BFs remaining. The *S-MARS*1 was tested with 5-fold CV and the *S-MARS*2 with 10-fold cross-validation.

Table 7. Statistics of the two staked S-MARS models and the reference model *AvrStacked*.

Model	Number of BFs; L	R^2 Test	RMSE	FB	IA, d	R^2
*S-MARS*1	80; 29	0.9351	4.3341	−0.00014	0.9860	0.9462
*S-MARS*2	60; 30	0.9349	4.2522	−0.00012	0.9865	0.9482
AvrStacked	-	-	5.2325	0.00347	0.9769	0.9329

A comparison of the results of the performance measures with those of the base models shows a clear superiority of *S-MARS*1 and *S-MARS*2 on all four measures (1) and (3). Table 8 gives the relative importance of the base models in the stacked *S-MARS*1 and *S-MARS*2. For *S-MARS*1, the model *RotRF*1 (100 points) has the highest contribution, followed by *RF*1 (53), *SelEB*1 (49), and *SelRotEB*1 (50). For *S-MARS*2, the relative importance is somewhat different—100, 43, 21, and 35, respectively. From Table 4 and Figure 8, we obtain that the best base model is *RotRF*1, which in Table 8 has the highest relative weight of 100. A comparison of the *S-MARS*1 and *S-MARS*2 statistics with those of *RotRF*1 shows the superiority of the stacked models in all statistics.

Table 8. Relative importance of base models in the staked models.

Model	*RF*1	*RotRF*1	*SelEB*1	*SelRotEB*1
*S-MARS*1	52.87	100	49.44	50.30
*S-MARS*2	42.88	100	21.27	34.81
AvrStacked	25	25	25	25

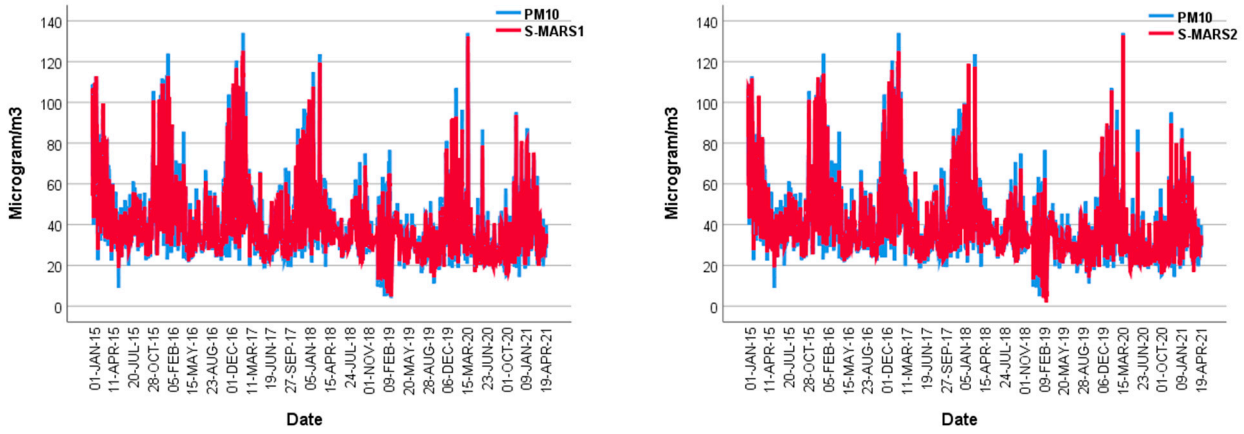
Having the predictions of the four base models, let us also calculate the standard stacked average model as a reference model. We denote it by *AvrStacked*. Its values are calculated using the expression:

$$ArvStaked = \frac{RF1 + RotRF1 + SelEB1 + SelRotEB1}{4} \quad (5)$$

From Tables 7 and 8, it is seen that this reference model has weaker statistical indicators compared to the S-MARS models.

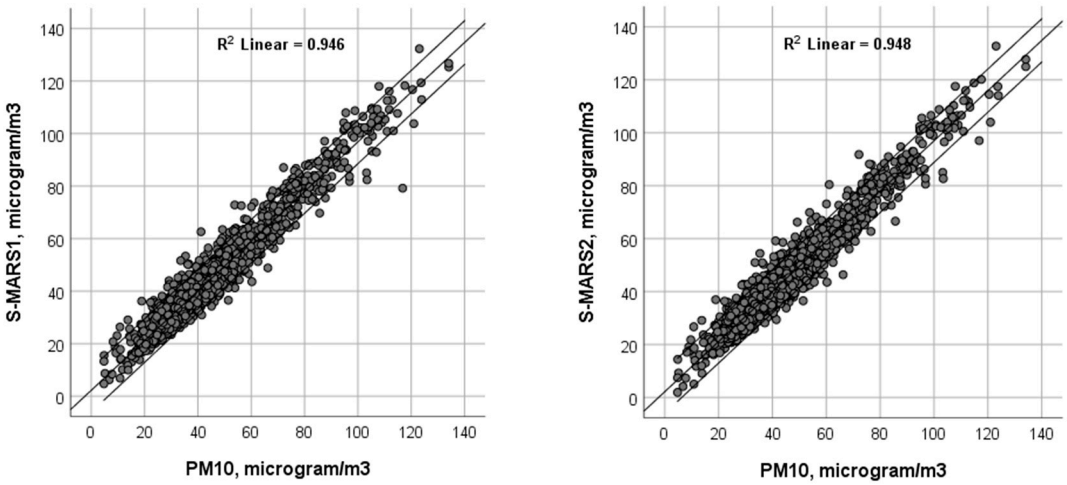
The *S-MARS*1 model has the Equation (A1) and BFs (A2) in Appendix A. The Equation (A1) and BFs (A2) serve to calculate each approximate value of the time series *PM*10 if the corresponding values of the four models at time t are known.

The prediction quality of the stacked models is visualized in Figure 10. Figure 11a,b show the scatter plots comparing the predicted levels versus PM_{10} measured with 5% confidence intervals. Again, much better prediction was observed, and in particular of the largest and smallest values, compared to the corresponding results of the base models from Figures 7 and 8, respectively.



(**a**) (**b**)

Figure 10. Quality of the coincidence of the measured values of PM_{10} and the ones predicted by the stacked models with 5% confidence intervals for (**a**) model *S-MARS*1 and (**b**) model *S-MARS*2.



(**a**) (**b**)

Figure 11. Scatter plots of predicted values with stacked S-MARS models versus measured PM_{10} values for (**a**) linear model *S-MARS*1 and (**b**) second-degree model *S-MARS2*.

3.4.2. Reliability Evaluation of S-MARS Models

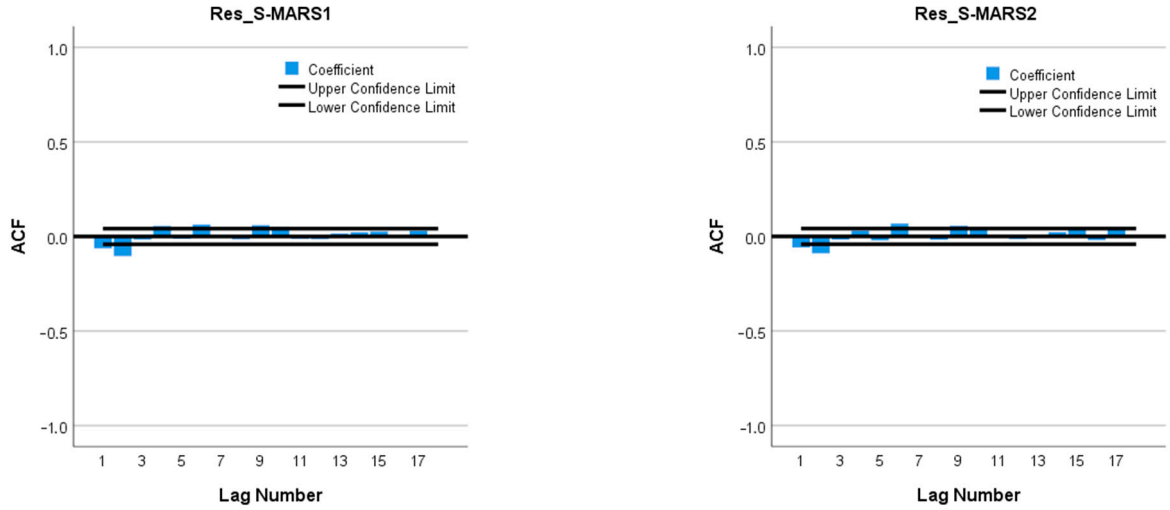
To provide statistical evidence and compare the stacked models, we used two statistical tools: WSRT and an ACF plot of their residuals. Using the WSRT in Table 9, *S-MARS*1, *S-MARS*2, and *PM*10 were compared. Since all the p-values were insignificant (Sig. > 0.05), the null hypothesis was retained; thus, there was no statistically significant difference between the compared variables. In particular, the WSRT for *S-MARS*1 and *S-MARS*2 was insignificant (Sig. = 0.150 > 0.05). Therefore, the two stacked models can be considered as relatively equal. According to the parsimonious principle, we chose the linear model as simpler.

Table 9. Statistics of Wilcoxon signed rank test to check the closeness between the stacked models and the target variable.

	Test Statistics [a]		
	***S-MARS*1, *PM*10**	***S-MARS*2, *PM*10**	***S-MARS*1, *S-MARS*2**
Z (Standardized test statistics)	−0.973 [c]	−1.025 [c]	−1.441 [c]
Asymp. Sig. (2-tailed)	0.331	0.305	0.150

[a] Wilcoxon signed ranks test; [c] based on positive ranks.

The errors ACF are shown in Figure 12. They are within the respective bounds. Therefore, we can assume that the residuals of *S-MARS*1 and *S-MARS*2 are white noise, and do not contain serial autocorrelation.



(**a**) (**b**)

Figure 12. Residual ACF of the stacked models *S-MARS*1 and *S-MARS*2 (**a**) model *S-MARS*1 and (**b**) model *S-MARS*2.

3.4.3. Forecasting of Holdout Sample for 7 Days Using Staked S-MARS Models

We will demonstrate the results of our *S-MARS*1 and *S-MARS*2 models for short-term forecasting of PM_{10} concentrations. So far, all models were built and analyzed for the target *PM*10_7, in which the last seven values of PM_{10} were unknown as holdout sample data. The forecasts of these 7 days are shown in Figure 13, together with the values of the measured PM_{10} for the last 21 days of the initial sample. On the left side of the vertical line are the stacked predictions for the penultimate 14 days, and on the right are seven forecasts for the holdout sample.

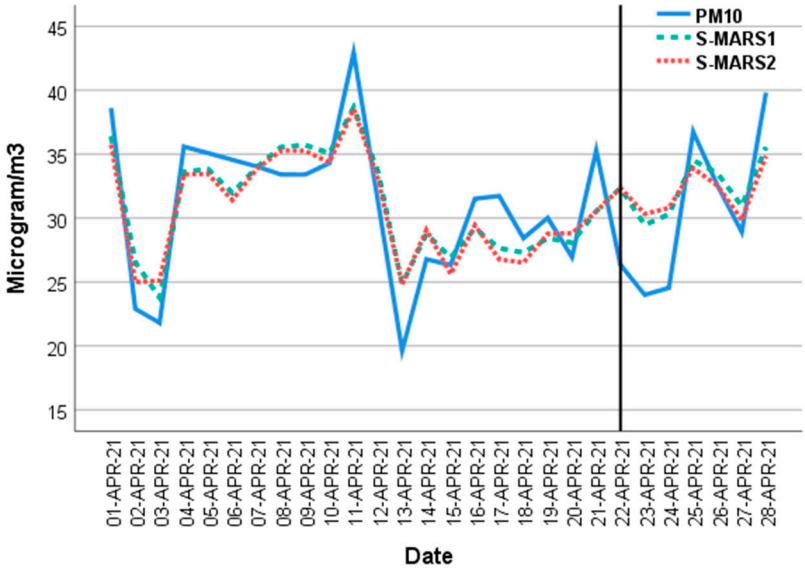


Figure 13. Forecasting of the measured PM_{10} levels for the last 28 days using the stacked S-MARS models. The predicted values are plotted to the left of the vertical line, and data forecasts for the 7-day holdout sample are provided to the right.

From the performed reliability assessment, we can conclude that all constructed models are statistically valuable, and the proposed framework has great potential for forecasting time series of air pollution concentrations under the model assumptions.

4. Discussion with Conclusions

In this paper, we studied the dependence of time series for the concentrations of the air pollutant PM_{10} on meteorological time series, including maximum and minimum average daily air temperature, wind speed, relative air humidity, atmospheric pressure, cloud conditions, rain, and weather and wind direction as categorical variables. The data are for a seaside region of Bulgaria over a period of more than 6 years. To find a more parsimonious approach to improving the forecasting quality of time series for PM_{10}, we developed a novel ensemble-based stacking framework.

With a limited number of initial predictors, we conducted a careful study to find existing relationships between them. In this way, we constructed additional data samples. The proposed framework combines ML techniques at two levels, utilizing the paradigm of stacked generalization. For the first time, CART ensemble and bagging (CART-EB) and rotation CART-EB with principal components were used to build base models. In addition, a simplified selective ensemble algorithm was proposed and implemented. A reduction in the number of trees in the CART-EB ensembles was achieved and the number of models was expanded through greater diversity. Although with slightly lower performance than RF1 and RotRF1, the obtained selective CART-EB and RotEB models at level 0 had relatively good statistical indicators.

Perhaps the most challenging part in the application of the framework is the selection of the four models at level 0 so that they are diverse among each other and that each approximates the PM_{10} data sufficiently well. We achieved the required diversity property by suggesting five strategies. After that, with the help of the MARS method, the forecasts made using these four base models were combined into two stacked models—linear *S-MARS*1 and second-degree *S-MARS*2. We showed no significant difference between these two models and chose the parsimonious *S-MARS*1 as the final one. With this linear stacked MARS model, the following performance measures were achieved: RMSE = 4.3341, fractional bias FB = -0.00014, index of agreement IA = 0.9860, and coefficient of determination $R^2 = 95\%$. All constructed models were cross-validated, and their residuals were examined to establish their reliability and the lack of serial autocorrelation.

Another notable aspect is the interesting result that stacking with MARS overcomes some of the disadvantages of the ensemble models, such as poor prediction of the highest and lowest values due to averaging.

Meteorological factors are a standard component of the data samples used to model the concentrations of air pollutants. In a recent publication [7], meteorological variables are the only predictors. An ANN model has been developed for forecasting hourly data for PM_{10} for winter periods in three settlements in Poland. The evaluation measures R^2 from 0.452 to 0.848, RMSE values from 8.80 to 23.56, and index of agreement (IA) from 0.693 to 0.957 were achieved. They are comparable to the corresponding statistics of our base models. Additionally, daily mean PM_{10} concentrations are forecasted in [17] using ANN, with hourly PM_{10} measurements one-day in advance, and local meteorological and some deterministic data, such as Sahara dust alert in Montseny and Barcelona, Spain. The performance indices of the model were R^2 = 0.86 and R^2 = 0.73 for the two cities, respectively. These results are inferior to ours, which is probably partly due to the greater variability of the hourly data used.

Our model is comparable to the stacked model presented in [40] for forecasting average daily pollution with fine particulate matter ($PM_{2.5}$) in Beijing, China. First, base models were developed using LASSO, AdaBoost, XgBoost, and MLP optimized by the genetic algorithm (GA-MLP). Then, these base models were combined by SVR. The stacked model reached the coefficient of determination R^2 = 90%. Another paper, similar to our study, is [39]. The authors analyzed and predicted the measured hourly concentrations of PM_{10} in four automatic stations in Cairo, Egypt, depending on seven meteorological and three temporal variables. Three machine learning methods were studied and compared—support vector regression, boosting, and stacking ensemble, applying the chance weight of its target variables. Stacking models showed the best statistics, with R^2 values up to 64%.

In [16], three ML spatial models using RF, bagged classification, regression trees (bagged CART), and mixture discriminate analysis (MDA) for the hazard prediction of PM_{10}, measured in 75 stations in the Barcelona Province, Spain, were developed and analyzed. Thirteen important variables were used as predictors: minimum temperature, maximum temperature, precipitation, wind speed, wind direction, elevation, road density, normalized difference vegetation index (NDVI), topographic wetness index (TWI), land use, terrain roughness index (TRI), and distance from the water body. The bagged CART and RF models achieved an accuracy of 92% to 93%, respectively, with a precision of about 86%. Although in different contexts, these results are in good agreement with those obtained from our base models (see Table 4).

Some other similar studies in which stacked selective ensembles are implemented can be mentioned. In [26], such models were proposed to forecast hourly $PM_{2.5}$ levels based on meteorological and pollutant predictors and a set of lagged variables. The authors explored random subspace, inclusive subspace, bagging, selective ensemble, and SVR-based stacking techniques. A large number of models for different data samples were obtained. Their statistical measures are comparable to ours for both base and stacked models.

The presented stacking framework and the results of this study show that the proposed approach allows one to obtain high-quality and reliable ensemble models for predicting the levels of air pollutants. Its implementation is useful for the protection of the population by providing timely information based on weather forecasts and other easily accessible data. In future studies, we plan to expand the application of the developed time series framework to forecast air and water pollution levels, financial and foreign exchange markets, etc., characterized by greater volatility.

Author Contributions: Conceptualization and methodology, S.G.-I.; software, S.G.-I. and A.I.; validation, all authors; investigation, all authors; resources, A.I. and M.S.-M.; data curation, M.S.-M.; writing—original draft preparation, S.G.-I.; writing—review and editing, S.G.-I.; funding acquisition, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This study emanated from research conducted with the financial support of the Bulgarian National Science Fund (BNSF), grant number KP-06-IP-CHINA/1 (КП-06-ИП-КИТАЙ/1).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used in this study are freely available on the official websites provided in references [42,43].

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Model S-MARS1

The stacked linear *S-MARS*1 model has the equation

$$\begin{aligned}\hat{y} = \hat{P}M10\_7 = {} & 75.1403 - 0.696284\,BF2 - 0.423695\,BF3 + 0.674278\,BF5 - 0.69262\,BF6 + 1.15976\,BF7 \\ & -1.46369\,BF9 + 8.49668\,BF15 - 6.91818\,BF17 - 4.42493\,BF19 + 4.07217\,BF21 - 0.918303\,BF23 \\ & -2.75369\,BF25 + 2.87188\,BF27 - 0.46759\,BF29 + 0.228861\,BF31 - 0.285907\,BF39 + 0.569981\,BF41 \\ & +3.02017\,BF47 + 3.8429\,BF49 - 1.59911\,BF51 - 2.59165\,BF53 - 7.48239\,BF57 + 7.56856\,BF59 \\ & +9.48026\,BF61 + 1.06406\,BF63 - 1.59165\,BF65 - 2.13434\,BF71 - 5.5955\,BF75 - 3.44223\,BF77\end{aligned} \quad \text{(A1)}$$

where BFs are calculated by the expressions:

$$
\begin{array}{lll}
BF2 = \max(0, 81.8456 - RotRF1); & BF3 = \max(0, SelEB1 - 45.6374); & BF5 = \max(0, RF1 - 43.7164); \\
BF6 = \max(0, 43.7164 - RF1); & BF7 = \max(0, SelEB1 - 83.8242); & BF9 = \max(0, RotRF1 - 90.2285); \\
BF15 = \max(0, RF1 - 87.9986); & BF17 = \max(0, RF1 - 84.2722); & BF19 = \max(0, RF1 - 89.9127); \\
BF21 = \max(0, RF1 - 82.2843); & BF23 = \max(0, SelRotEB1 - 78.9109); & BF25 = \max(0, RF1 - 102.162); \\
BF27 = \max(0, SelRotEB1 - 103.255); & BF29 = \max(0, RF1 - 25.6409); & BF31 = \max(0, SelEB1 - 24.8643); \\
BF39 = \max(0, SelRotEB1 - 38.9905); & BF41 = \max(0, RotRF1 - 40.7856); & BF47 = \max(0, SelEB1 - 72.0619); \\
BF49 = \max(0, SelRotEB1 - 52.7836); & BF51 = \max(0, SelRotEB1 - 57.5217); & BF53 = \max(0, SelRotEB1 - 51.8321); \\
BF57 = \max(0, RotRF1 - 74.2716); & BF59 = \max(0, RotRF1 - 75.2887); & BF61 = \max(0, RotRF1 - 61.7364); \\
BF63 = \max(0, SelRotEB1 - 69.5625); & BF65 = \max(0, SelEB1 - 67.475); & BF71 = \max(0, SelEB1 - 75.6904); \\
BF75 = \max(0, RotRF1 - 62.9263); & BF77 = \max(0, RotRF1 - 60.1102) &
\end{array} \quad \text{(A2)}
$$

References

1. Janssen, N.; Fischer, P.; Marra, M.; Ameling, C.; Cassee, F. Short-term effects of $PM_{2.5}$, PM_{10} and $PM_{2.5–10}$ on daily mortality in the Netherlands. *Sci. Total. Environ.* **2013**, *463–464*, 20–26. [CrossRef]
2. Kappos, A.D.; Bruckmann, P.; Eikmann, T.; Englert, N.; Heinrich, U.; Höppe, P.; Koch, E.; Krause, G.H.; Kreyling, W.G.; Rauchfuss, K.; et al. Health effects of particles in ambient air. *Int. J. Hyg. Environ. Health* **2004**, *207*, 399–407. [CrossRef]
3. Kettunen, J.; Lanki, T.; Tiittanen, P.; Aalto, P.P.; Koskentalo, T.; Kulmala, M.; Salomaa, V.; Pekkanen, J. Associations of Fine and Ultrafine Particulate Air Pollution with Stroke Mortality in an Area of Low Air Pollution Levels. *Stroke* **2007**, *38*, 918–922. [CrossRef]
4. European Commission. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe. *Off. J. Eur. Union* **2008**, *152*, 1–44.
5. European Commission. Air Quality Standards. 2014. Available online: http://ec.europa.eu/environment/air/quality/standards.htm (accessed on 7 December 2021).
6. Seinfeld, J.H.; Pandis, S.N. Chapter 20. Wet deposition. In *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, 3rd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2006; pp. 856–888.
7. Nidzgorska-Lencewicz, J. Application of Artificial Neural Networks in the Prediction of PM_{10} Levels in the Winter Months: A Case Study in the Tricity Agglomeration, Poland. *Atmosphere* **2018**, *9*, 203. [CrossRef]
8. Ul-Saufie, A.Z.; Yahaya, A.S.; Ramli, N.A.; Rosaida, N.; Hamid, H.A. Future daily PM_{10} concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). *Atmos. Environ.* **2013**, *77*, 621–630. [CrossRef]
9. Tzanis, C.G.; Alimissis, A.; Philippopoulos, K.; Deligiorgi, D. Applying linear and nonlinear models for the estimation of particulate matter variability. *Environ. Pollut.* **2018**, *246*, 89–98. [CrossRef]
10. Dimov, I.; Georgiev, K.; Ostromsky, T.; Zlatev, Z. Computational challenges in the numerical treatment of large air pollution models. *Ecol. Model.* **2004**, *179*, 187–203. [CrossRef]
11. Dimov, I.; Georgieva, R.; Ostromsky, T.; Zlatev, Z.; Georgieva, R. Sensitivity studies of pollutant concentrations calculated by the UNI-DEM with respect to the input emissions. *Open Math.* **2013**, *11*, 1531–1545. [CrossRef]
12. Tsvetanova, I.; Zheleva, I.; Filipova, M.; Stefanova, A. Statistical analysis of ambient air PM_{10} contami-nation during winter periods for Ruse region, Bulgaria. In *Proceedings of the 13th National Congress on Theoretical and Applied Mechanics (NCTAM 2017), Sofia, Bulgaria, 6–10 September 2017*; Vassilev, V.M., Datcheva, M.D., Nikolov, S.G., Ivanova, Y.P., Eds.; MATEC Web of Conferences: New York, NY, USA, 2018; Volume 145, p. 1007. [CrossRef]
13. Veleva, E.; Zheleva, I. Statistical modeling of particle mater air pollutants in the city of Ruse, Bulgaria. In *Proceedings of the 13th National Congress on Theoretical and Applied Mechanics (NCTAM 2017), Sofia, Bulgaria, 6–10 September 2017*; Vassilev, V.M., Datcheva, M.D., Nikolov, S.G., Ivanova, Y.P., Eds.; MATEC Web of Conferences: New York, NY, USA, 2018; Volume 145, p. 1010. [CrossRef]
14. Zheleva, I.; Veleva, E.; Filipova, M. Analysis and modeling of daily air pollutants in the city of Ruse, Bulgaria. In *Proceedings of the 9th International Conference for Promoting the Application of Mathematics in Technical and Natural Sciences—AMiTaNS'17, Albena, Bulgaria, 21–26 June 2017*; Todorov, M., Ed.; AIP Conference Proceedings, American Institute of Physics: Melville, NY, USA, 2017; Volume 1895, p. 30007. [CrossRef]
15. Voukantsis, D.; Karatzas, K.; Kukkonen, J.; Räsänen, T.; Karppinen, A.; Kolehmainen, M. Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki. *Sci. Total Environ.* **2011**, *409*, 1266–1276. [CrossRef]
16. Choubin, B.; Abdolshahnejad, M.; Moradi, E.; Querol, X.; Mosavi, A.; Shamshirband, S.; Ghamisi, P. Spatial hazard assessment of the PM_{10} using machine learning models in Barcelona, Spain. *Sci. Total Environ.* **2019**, *701*, 134474. [CrossRef]
17. de Gennaro, G.; Trizio, L.; Di Gilio, A.; Pey, J.; Perez, N.; Cusack, M.; Alastuey, A.; Querol, X. Neural network model for the prediction of PM_{10} daily concentrations in two sites in the Western Mediterranean. *Sci. Total Environ.* **2013**, *463–464*, 875–883. [CrossRef]
18. Lasheras, F.S.; Nieto, P.J.G.; Gonzalo, E.G.; Bonavera, L.; Juez, F.J.D.C. Evolution and forecasting of PM_{10} concentration at the Port of Gijon (Spain). *Sci. Rep.* **2020**, *10*, art.11716. [CrossRef]

19. Tadano, Y.D.S.; Siqueira, H.V.; Alves, T.A. Unorganized machines to predict hospital admissions for respiratory diseases. In Proceedings of the 2016 IEEE Latin American Conference on Computational Intelligence (LA-CCI), Cartagena, Colombia, 2–4 November 2016. [CrossRef]
20. Belotti, J.T.; Castanho, D.S.; Araujo, L.N.; da Silva, L.V.; Alves, T.A.; Tadano, Y.S.; Stevan, S.L.; Corrêa, F.C.; Siqueira, H.V. Air pollution epidemiology: A simplified Generalized Linear Model approach optimized by bio-inspired metaheuristics. *Environ. Res.* **2020**, *191*, 110106. [CrossRef]
21. Ye, Z.; Yang, J.; Zhong, N.; Tu, X.; Jia, J.; Wang, J. Tackling environmental challenges in pollution controls using artificial intelligence: A review. *Sci. Total Environ.* **2019**, *699*, 134279. [CrossRef]
22. Cabaneros, S.M.; Calautit, J.K.; Hughes, B.R. A review of artificial neural network models for ambient air pollution prediction. *Environ. Model. Softw.* **2019**, *119*, 285–304. [CrossRef]
23. Xayasouk, T.; Lee, H.; Lee, G. Air Pollution Prediction Using Long Short-Term Memory (LSTM) and Deep Autoencoder (DAE) Models. *Sustainability* **2020**, *12*, 2570. [CrossRef]
24. Wang, J.; Li, J.; Wang, X.; Wang, J.; Huang, M. Air quality prediction using CT-LSTM. *Neural Comput. Appl.* **2020**, *33*, 4779–4792. [CrossRef]
25. Breiman, L. Arcing classifiers. *Ann. Stat.* **1998**, *26*, 801–824. Available online: https://www.jstor.org/stable/120055 (accessed on 7 December 2021).
26. Gu, K.; Xia, Z.; Qiao, J. Stacked Selective Ensemble for $PM_{2.5}$ Forecast. *IEEE Trans. Instrum. Meas.* **2019**, *69*, 660–671. [CrossRef]
27. Zhou, Z.-H. *Ensemble Methods: Foundations and Algorithms*; Chapman & Hall, CRC: Boca Raton, FL, USA, 2012.
28. Wang, H.; Jiang, Y.; Wang, H. Stock return prediction based on Bagging-decision tree. In Proceedings of the 2009 IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2009), Nanjing, China, 10–12 November 2009; pp. 1575–1580. [CrossRef]
29. Aydoğmuş, H.Y.; Ekinci, A.H.; Erdal, H.I.; Erdal, H. Optimizing the monthly crude oil price forecasting accuracy via bagging ensemble models. *J. Econ. Int. Financ.* **2015**, *7*, 127–136. [CrossRef]
30. Mohammed, A.; Asteris, P.; Koopialipoor, M.; Alexakis, D.; Lemonis, M.; Armaghani, D. Stacking Ensemble Tree Models to Predict Energy Performance in Residential Buildings. *Sustainability* **2021**, *13*, 8298. [CrossRef]
31. Mendes-Moreira, J.; Soares, C.; Jorge, A.; de Sousa, J.F. Ensemble approaches for regression. *ACM Comput. Surv.* **2012**, *45*, 1–40. [CrossRef]
32. Zhou, Z.-H.; Wu, J.; Tang, W. Ensembling neural networks: Many could be better than all. *Artif. Intell.* **2002**, *137*, 239–263. [CrossRef]
33. Zhou, Z.-H.; Tang, W. Selective Ensemble of Decision Trees. In Proceedings of the International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing RSFDGrC, Chongqing, China, 26–29 May 2003; Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 476–483. [CrossRef]
34. Zhu, X.; Ni, Z.; Cheng, M.; Jin, F.; Li, J.; Weckman, G. Selective ensemble based on extreme learning machine and improved discrete artificial fish swarm algorithm for haze forecast. *Appl. Intell.* **2017**, *48*, 1757–1775. [CrossRef]
35. Bates, J.M.; Granger, C.W.J. The Combination of Forecasts. *J. Oper. Res. Soc.* **1969**, *20*, 451–468. [CrossRef]
36. Newbold, P.; Granger, C.W.J. Experience with Forecasting Univariate Time Series and the Combination of Forecasts. *J. R. Stat. Soc. Ser. A* **1974**, *137*, 131. [CrossRef]
37. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
38. Breiman, L. Stacked regressions. *Mach. Learn.* **1996**, *24*, 49–64. [CrossRef]
39. Eldakhly, N.M.; Aboul-Ela, M.; Abdalla, A. A Novel Approach of Weighted Support Vector Machine with Applied Chance Theory for Forecasting Air Pollution Phenomenon in Egypt. *Int. J. Comput. Intell. Appl.* **2018**, *17*, 1850001. [CrossRef]
40. Zhai, B.; Chen, J. Development of a stacked ensemble model for forecasting and analyzing daily average $PM_{2.5}$ concentrations in Beijing, China. *Sci. Total Environ.* **2018**, *635*, 644–658. [CrossRef] [PubMed]
41. Ganchev, I.; Ji, Z.; O'Droma, M. Designing a cloud tier for the IoT platform EMULSION. *WSEAS T. Syst. Control* **2019**, *14*, 375–383, art.46.
42. Regional Inspectorate for Environment and Water, Burgas (in Bulgarian). Available online: http://riosvbs.com/home/menu/1296 or http://riosvbs.com/Files/%D0%A4%D0%9F%D0%A710%20%D0%94.%D0%95%D0%B7%D0%B5%D1%80%D0%BE%D0%B2%D0%BE%202021.xlsx (accessed on 7 December 2021).
43. World Weather Online, Burgas Historical Weather. Available online: https://www.worldweatheronline.com/burgas-weather-history/burgas/bg.aspx (accessed on 7 December 2021).
44. Air Quality in Europe—2020 Report. European Environment Agency. EEA Report 09/ 2020. Available online: https://www.eea.europa.eu//publications/air-quality-in-europe-2020-report. (accessed on 7 December 2021).
45. Džeroski, S.; Ženko, B. Is Combining Classifiers with Stacking Better than Selecting the Best One? *Mach. Learn.* **2004**, *54*, 255–273. [CrossRef]
46. Flores, B.E. The utilization of the Wilcoxon test to compare forecasting methods: A note. *Int. J. Forecast.* **1989**, *5*, 529–535. [CrossRef]
47. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417–441, 498–520. [CrossRef]

48. Izenman, A.J. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*; Springer: New York, NY, USA, 2008.
49. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
50. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
51. Ho, T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844. [CrossRef]
52. Willmott, C.J. On the validation of models. *Phys. Geogr.* **1981**, *2*, 184–194. [CrossRef]
53. Friedman, J.H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 1–141. [CrossRef]
54. De Gooijer, J.G.; Kumar, K. Some recent developments in non-linear time series modelling, testing, and forecasting. *Int. J. Forecast.* **1992**, *8*, 135–156. [CrossRef]
55. Gocheva-Ilieva, S.G.; Voynikova, D.S.; Stoimenova, M.P.; Ivanov, A.V.; Iliev, I.P. Regression trees modeling of time series for air pollution analysis and forecasting. *Neural Comput. Appl.* **2019**, *31*, 9023–9039. [CrossRef]
56. Livieris, I.E.; Stavroyiannis, S.; Pintelas, E.; Pintelas, P. A novel validation framework to enhance deep learning models in time-series forecasting. *Neural Comput. Appl.* **2020**, *32*, 17149–17167. [CrossRef]
57. Brockwell, P.J.; Davis, R.A. *Introduction to Time Series and Forecasting*, 3rd ed.; Springer: Berlin, Germany, 2016.
58. Wilks, D.S. *Statistical Methods in the Atmospheric Sciences*, 3rd ed.; Elsevier: Amsterdam, The Netherlands, 2011.