atmosphere

MDPI

Article

Modeling PM2.5 and PM10 Using a Robust Simplified Linear Regression Machine Learning Algorithm

João Gregório [1,†], Carla Gouveia-Caridade [1] and Pedro J. S. B. Caridade [2,*]

[1] SpaceLayer Technologies, Uburu-IQ, Av. Emídio Navarro, 33, 3000-151 Coimbra, Portugal
[2] CQC-ISM and Department of Chemistry, University of Coimbra Rua Larga, 3004-545 Coimbra, Portugal
* Correspondence: pedrojcaridade@uc.pt
† Current address: National Physical Laboratory, Technology & Innovation Centre, 99 George Street, Glasgow G1 1RD, UK.

Abstract: The machine learning algorithm based on multiple-input multiple-output linear regression models has been developed to describe PM2.5 and PM10 concentrations over time. The algorithm is fact-acting and allows for speedy forecasts without requiring demanding computational power. It is also simple enough that it can self-update by introducing a recursive step that utilizes newly measured values and forecasts to continue to improve itself. Starting from raw data, pre-processing methods have been used to verify the stationary data by employing the Dickey–Fuller test. For comparison, weekly and monthly decompositions have been achieved by using Savitzky–Golay polynomial filters. The presented algorithm is shown to have accuracies of 30% for PM2.5 and 26% for PM10 for a forecasting horizon of 24 h with a quarter-hourly data acquisition resolution, matching other results obtained using more computationally demanding approaches, such as neural networks. We show the feasibility of using multivariate linear regression (together with the small real-time computational costs for the training and testing procedures) to forecast particulate matter air pollutants and avoid environmental threats in real conditions.

Keywords: machine learning; multivariate linear regression; time series forecasting; forecasting; particulate-matter; environmental data analysis

Citation: Gregório, J.; Gouveia-Caridade, C.; Caridade, P.J.S.B. Modeling PM2.5 and PM10 Using a Robust Simplified Linear Regression Machine Learning Algorithm. *Atmosphere* **2022**, *13*, 1334. https://doi.org/10.3390/atmos13081334

Academic Editor: Paulo Artaxo

Received: 22 July 2022
Accepted: 19 August 2022
Published: 22 August 2022

1. Introduction

Air pollution is one of the largest environmental concerns presently affecting people across the world (in developed and developing countries alike). The World Health Organization (WHO) estimates that there were 3.7 million premature deaths globally in 2012 due to air pollution [1–3]. Ambient air pollutants are diverse and varied; however, of all the contaminants, a few stand out as the most dangerous to human health, such as particulate matter (PM), ozone, nitrogen dioxide, sulfur dioxide, and carbon monoxide. Particulate matter consists of fine particles from organic and inorganic sources with aerodynamic diameters smaller than 10 µm (PM10) and 2.5 µm (PM2.5); it is produced by on-road and off-road vehicles, with contributions from power plants, industrial boilers, incinerators, petrochemical plants, aircraft, ships, and so on, depending on the locations and prevailing winds [4]. Due to their small sizes, $\leq$10 µm (PM10) and $\leq$2.5 µm (PM2.5), these fine particles are able to penetrate deep into the lungs and be absorbed into the bloodstream, causing damage to the organism [1,3,5–7]. The level of damage they can cause varies depending on the concentration and type [1,8,9]. Given the nature of their absorption, they mostly cause damage to the respiratory system, although the cardiovascular and neurological systems can also be affected [10–13]. Some of the less severe short-term effects include irritation of the mucous areas, such as the eyes, nose, and throat; effects could also include headaches and nausea, which disappear with time. However, chronic exposures to high levels are also linked to more serious conditions and can cause upper respiratory infections, such as bronchitis and emphysema [1,8]. Regarding long-term effects, PM exposure is also

linked to chronic respiratory diseases (e.g., asthma and lung cancer) [10], cardiovascular ailments [11,14], and brain damage [12].

Considering the danger posed by particulate matter, it is necessary to survey and monitor the ambient levels in the atmosphere [3,15]. However, since particulate matter is an extremely heterogeneous mixture, it becomes very difficult to measure and quantify (see, e.g., References [16,17]). This task was made easier by standardizing fine particles with diameters equal to or smaller than 10 μm as the official measure of ambient particle pollution [18]. Moreover, the increase in emissions and recognition of the heightened dangers of fine particles with an average diameter equal to or smaller than 2.5 μm made it necessary to also pay attention to these particles [3,19,20]. Considering these facts, the ability to accurately predict and forecast PM10 and PM2.5 values is of great interest to public health because it would reveal ambient levels of pollutants and enable the general population to take preventive actions.

The traditional approach to forecasting environmental variables implies the construction of deterministic models, which require extensive knowledge of parameters, such as air current flows, particle diffusion, and relevant chemical reactions [21]. The drawbacks of these approaches lie in the data acquisition and model construction processes. Moreover, usable data for all necessary parameters are hard to collect, and even if the necessary data are available, the algorithm construction and refinement process are lengthy, demanding, and often produce inaccurate models given to the chaotic nature of the atmosphere [22].

The development and application of data-driven algorithms (e.g., machine learning (MLA)) applied to the forecasting of pollutants have increased substantially in recent years, becoming quantitative alternatives to time series [8,23–26]. The advantage of these types of algorithms is that they are capable of capturing and finding the underlying patterns hidden in data; thus, they could be used for forecasting without the need for any prior assumptions. This means that not only is the model building process less demanding than that of the traditional approach but also that the model applications to new data are more straightforward and extremely fast [25,26]. Cabaneros et al. [27] conducted an extensive review on air pollution forecasting using artificial neural networks, while Kalajdjieski et al. [28] used multi-modal data and deep neural networks to generative approaches enhanced with standard data augmentation methods for handling imbalanced datasets. Many other references can be found in the literature regarding MLA, e.g., References [29–38].

Despite the MLA's popularity, the results achieved are not conclusive and are usually computationally costly even with graphical unit processing acceleration [39]. Therefore, the air pollution forecast remains an important and active area of research, where high-quality outcomes and fast schemes are key to limiting the exposure to pollution [40] and having the ability to handle imbalanced datasets and exploit multimodal data [41]. This work proposes a robust and simple algorithm based on linear regressions methods with enough viability to accurately forecast near-future PM10 and PM2.5 concentrations and is capable of being taught in the framework of machine learning with minimal computational costs. The main novelty is that the algorithm, being regression-based, is fast-acting and allows for faster forecasts without the requirement of demanding computational power. It is also simple enough that it can self-update, by introducing a recursive step that utilizes newly measured values and forecasts to continue to improve itself. Linear regression methods are labeled as simple machine learning algorithms [42]. Due to their simplicity, many publications focus their applications in many fields. Juneng et al. [43], used 10 predictor variables in the logarithmic scale to study the PM10 concentrations over Malaysia based on multi-station and multi-properties. They reported that 35–50% of the PM10 variations can be explained by these variables. Ng et al. [44] also studied the Malaysian PM10 concentrations with multiple linear regression and regression with time series error models obtaining 17% of MAPE. Shams et al. [45] employed artificial neural networks and multiple linear regression models for predicting SO_2 concentrations using seven input variables. The low value of $R^2 = 0.708$ shows the limitation of a simple application of the linear regression method. In turn, Okkaoglu et al. [46] reported a detailed study on daily PM10, periodicity,

and the harmonic regression model, attempting to capture the hidden periodicity of the time series. They rely on the use of periodograms having the requirement of a stationary series achieving a high degree of precision on annual concentrations. For the periodograms, they employed a trigonometric series and Fourier analysis. Bai et al. [47] carried out a thorough review (see also Reference [46]) on air pollution forecasts based on statistical forecasting methods, artificial intelligence methods, and numerical forecasting methods. They have also reported hybrid models that may improve the forecast accuracy.

2. Data and Methods

The PM10 and PM2.5 datasets employed in this work were obtained from the Copernicus Atmosphere Monitoring Service [48] (CAMS), reported on an hourly basis from 1 October 2016 to 30 September 2017 for 40.192169 N –8.414162 W location. The geographical coordinates correspond to a location in the central part of Portugal, although any location could have been employed in the present study. The dataset from CAMS is of the ENSEMBLE type [49,50]; it gathered information from eight European models and calculated via a median value approach [51]. These products generally yield better performances than the individual model products both for the forecast and analysis. For the particular case, the 0 km of altitude (surface level) and the analysis batches were chosen since they were validated, while no other sources of data were employed.

Before training the MLA with the dataset, a preliminary analysis of the data showed the need for pretreatment, see the flowchart in Figure 1. First, the dataset was not balanced in terms of size consistency with gaps, which made it impossible to properly use the time series forecasting model based on machine learning principles. Since the data gap number and sizes were short and scarce, a piecewise linear interpolation model [52] was employed to overcome the dataset limitation. This simple model connects the two known adjacent points, (x_1, y_1) and (x_2, y_2), with a straight line and allows determining a set of i values in-between them:

$$y_i = y_1 + (x_i - x_1)\frac{y_2 - y_1}{x_2 - x_1} \quad (1)$$

A major issue in the case of data inference is that balancing the dataset cannot lead to changes in the time series properties, compromising the analytics of the training, testing, and forecast. To confirm that the data set properties have been maintained after the balancing procedure, the Dickey–Fuller [53] test for stationary was applied to the original data set and on the continuous data set

$$y_t - y_{t-1} = (\rho - 1)y_{t-1} + u_t \quad (2)$$

where y_t is the testing variable, t is the time index, u_t is the error term, and ρ is the test coefficient. In the case of $\rho > 0.05$, the series is non-stationary, with a limiting value of 1 being completely non-stationary. The greater the difference between zero and $\rho - 1$ is, the greater the certainty with which the series can be called stationary. This also allowed to check for the time-series stationarity since any regression MLA can only be applied to stationary data [54]. If it is not stationarity, a detrending process has to be conducted before being able to train the MLA. Only after these steps can the training process of the MLA take place. Note that it is a common assumption in most time series methods that the data must be stationary [46]. Even if regression methods do not have assumptions regarding data stationarity, ensuring that data used in this study are stationary, they will facilitate future comparisons with works that utilize non-regression methods.

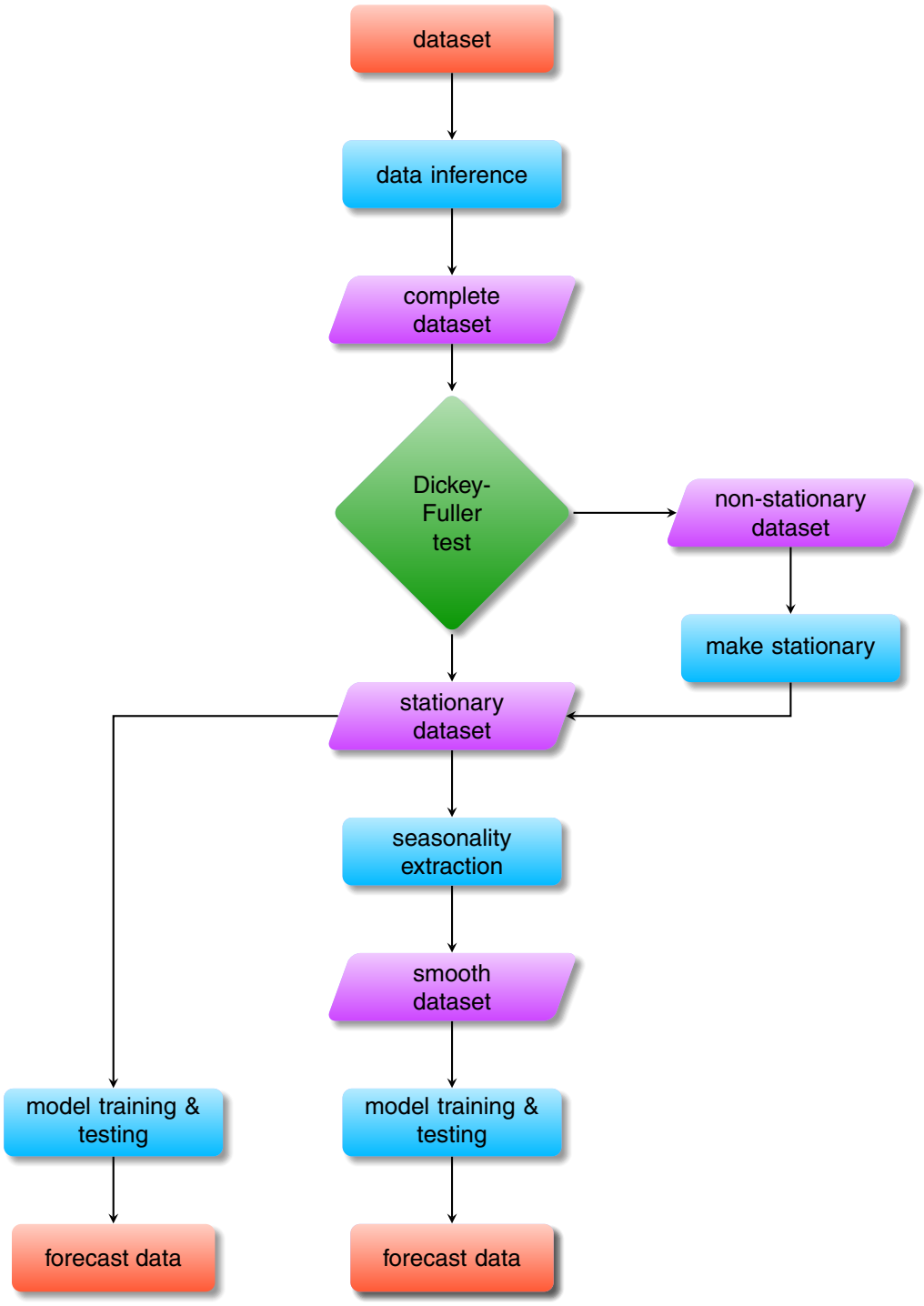

Figure 1. Data processing flowchart and algorithm application.

The core of the proposed algorithm is the machine-learning component. For a simple and robust forecast, it is necessary that the model be trained effectively in a short computational time. In this particular case, the multiple-input multiple-output linear regression model was selected. It is based on the fact that it does not have the drawbacks of multiple linear regression models, such as the accumulation of errors along the forecasting horizon in recursive models and the conditional independence assumption in direct models. In contrast to traditional single-output learning, multi-output learning can concurrently predict

multiple outputs. The outputs can be of various types and structures, and the problems that can be solved are diverse. These models are based on vector-valued functions, such as

$$\mathbf{Y}^{(i)}[y_{t+1}, ..., y_{t+H}] = \mathbf{X}^{(i)}(y_t, ..., y_{t-d+1}) \cdot \mathbf{C}_{(d+1)\times i} + \mathbf{w}_{H\times i} \quad (3)$$

where **Y** and **X** are the output and input vectors, respectively, for a forecast of H steps ahead using d input values where i represents the size of the training set, or, in other words, the total number of equations for which the coefficients matrix, **C**, will be minimized via a least-squares approach. Finally, **w** is the noise matrix [24,55–57].

The optimization procedure was carried out by minimizing the sum of the square elements on the diagonal of the residual sum of squares and cross-product matrices [58])

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \quad (4)$$

which gives the least possible trace for the coefficients matrix as well as minimizes the generalized variance of the system. The final coefficients can be stored in a simplified vector equation similar to Equation (3) alongside the input vector in order to produce the forecasts

$$\mathbf{Y}[y_{t+1}, \ldots, y_{t+H}] = \mathbf{X}(y_t, \ldots, y_{t-d+1}) \cdot \mathbf{C} \quad (5)$$

In this work, the methods and models were implemented in the Python programming language aimed at forecasting a total of 24 steps ahead, $H = 24$, using the same number of input values, $d = 24$. The structure of the MLA is reported in Algorithm 1. The full set of data consisted of a total of 8784 data points, from which 80 % have been employed for training and cross-validation and the remaining 20 % for testing, as suggested by Pareto's principle [59]. From the 6980 data points used for training, a ten-fold-winner-takes-all cross-validation process was employed [60,61]. As seen in the Algorithm 1 construction, this was conducted by splitting that portion of the data into ten equal random parts, using nine to train the model and the remaining one to validate it. This process was repeated ten times, always changing the validation segment. Afterward, the iteration that yielded the best results was replicated and tested on the remaining 1804 data points that comprised the test set. The testing process was carried out three times on the same set but with different start and end points to account for variations of the model and to add statistical meaning to the results [24].

The performance of the model was evaluated using the mean average percentage error (MAPE), which is the standardized approach to demonstrate the prediction accuracy of a forecasting method in the statistics and machine learning theory [26,62]. It is given by:

$$\text{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{X_i - X_{\text{forecast},i}}{X_i}\right| \quad (6)$$

where X_i is the actual value and $X_{\text{forecast},i}$ is the corresponding forecast value for a data set of size N, while the multiplication by 100 is converted to a percentage. Beyond the MAPE criteria, a qualitative analysis of the results was conducted by checking the coherence between the real and forecasted values in predicting the correct hazard level. This was conducted by counting the number of measured values that crossed the threshold between hazard levels and converting the count into percentages and afterward doing the same with the forecasted values. The disparity between the values allowed for a comparative analysis. Another aspect that was explored was the seasonality extraction or seasonal decomposition and the effects on model performance [24]. This was carried out by smoothing the original data set, considering a weekly and monthly signal repetition that generated two extra data sets. These new data sets disregarded any noise by considering only the overall trend of the data and the seasonal effects.

Algorithm 1: Multiple-input multiple-output linear regression algorithm.

Input : Full balanced dataset $\mathbf{X}$ size N
begin
  $N_{\text{train}} = \text{Int}(0.8 \times N)$
  $\mathbf{X}^{(1)}_{\text{train}} \longleftarrow \text{Split.RAND}(\mathbf{X}, size = N_{\text{train}})$
  $\mathbf{X}^{(1)}_{\text{test}} \longleftarrow \mathbf{X}_{\text{train}} \cap \mathbf{X}$
  $\mathbf{X}^{(2)}_{\text{train}} = \mathbf{X}^{(1)}_{\text{train}}$
  $\mathbf{X}^{(2)}_{\text{test}} = \mathbf{X}^{(1)}_{\text{test}}$
  $p \longleftarrow \text{Dickey–Fuller}(\mathbf{X}^{(2)}_{\text{train}})$
  if $(p > 0.05)$ **then**
    $\mathbf{X}^{(2)}_{\text{train}}$ non-stationary
    $\mathbf{X}^{(3)}_{\text{train}} \longleftarrow \text{Savitzky–Golay.Filter}(\mathbf{X}^{(2)}_{\text{train}})$
    $\mathbf{X}^{(3)}_{\text{train}}$ stationary
  else
    $\mathbf{X}^{(2)}_{\text{train}}$ stationary
  end
end
for $k = 0$ **to** $n = 10$ **do**
  for $i = 0$ **to** $n = 10$ **do**
    $\mathbf{X}^{(i,1)}_{\text{train}} \longleftarrow \text{Split.RAND}(\mathbf{X}^{(1)}_{\text{train}}, size = N/10)$
    $\mathbf{X}^{(i,2)}_{\text{train}} \longleftarrow \text{Split.RAND}(\mathbf{X}^{(2)}_{\text{train}}, size = N/10)$
    $\mathbf{X}^{(i,3)}_{\text{train}} \longleftarrow \text{Split.RAND}(\mathbf{X}^{(3)}_{\text{train}}, size = N/10)$
    $\text{LinearRegression}(\mathbf{X}^{(i,1)}_{\text{train}})$
    $\text{LinearRegression}(\mathbf{X}^{(i,2}_{\text{train}})$
    $\text{LinearRegression}(\mathbf{X}^{(i,3)}_{\text{train}})$
    $\text{MAEP}^{(i,1)} \longleftarrow \text{LinearRegression.Predict}(\mathbf{X}^{(2)}_{\text{train}})$
    $\text{MAEP}^{(i,2)} \longleftarrow \text{LinearRegression.Predict}(\mathbf{X}^{(i,2)}_{\text{train}})$
    $\text{MAEP}^{(i,3)} \longleftarrow \text{LinearRegressionPredict}(\mathbf{X}^{(i,3)}_{\text{train}})$
  end
  $\text{bestModel}^{(k)} \longleftarrow \min(\text{MAEP}^{(i,j)})$
end
$\text{bestModel} \longleftarrow \text{bestModel}^{(k)}$
Output: $H = 24; d = 24$

$\mathbf{X}_{\text{forecast}}[y_{t+1}, \ldots, y_{t+H}] \longleftarrow \mathbf{X}_{\text{bestModel}}[y_t, \ldots, y_{t-d+1}]$

3. Results and Discussion

In this section, we present the results of the forecasted data. Note that by the construction of the algorithm, both original and deseasonalized data sets are empirically considered as stationary since the Dickey–Fuller test was performed and the Savitzky–Golay [63] filter applied in case of non-stationarity of the subset. The outcomes from the applications on the undecomposed time series are the primary focus and serve as the comparison for results involving the decomposed time series. In all cases, the computational cost was ∼10 min of real-time for the complete process in a non-GPU-accelerated one-core process, showing the simplicity of the method in the training and validation steps.

3.1. Original Data

The results of the MLA application on the original data without any kind of seasonality extraction are shown in Figure 2 for the third test iteration. The results of PM10 forecasting, shown by time evolution and correlation diagrams, are in very good agreement between input (real) and predicted values, suggesting that the model is able to cope with the complexity of the pollutant over time and fast concentration variation. The correlation diagram shows that the majority of the points follow the bisector line, while the linear fit is indistinguishable from the bisection. This assessment seems to be further supported by the comparison of the forecast values alongside the real values over the cumulative time scale. The two lines are indistinguishable within the scale, showing that the global trend of the MLA also follows the input data. For a more quantitative analysis, the error is also reported with the cumulative timeline. The mean MAPE value for all three testing iterations is of 26.6% and individual errors are mostly small except in a few situations, which correspond to inflection points of the original signal.

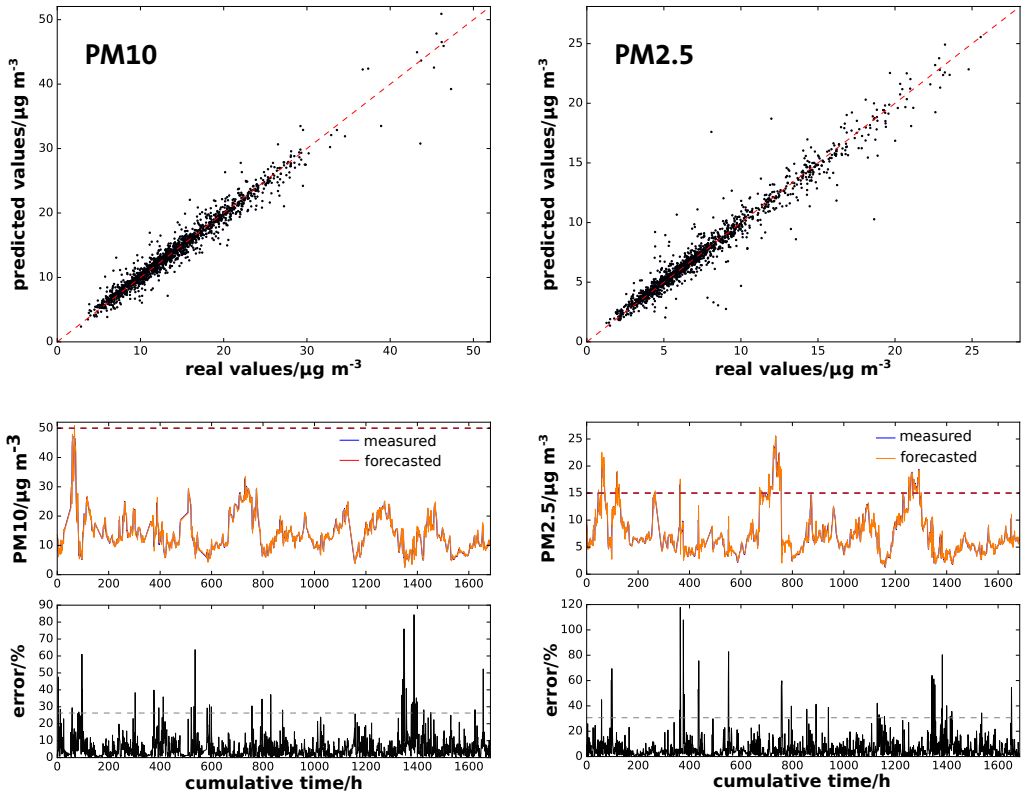

Figure 2. Dispersion, concentration, and error over time dependencies for the third test iteration of forecasting without seasonality decomposition. Dispersion plots: $R^2 = 0.933$ for PM10 and $R^2 = 0.920$ for PM2.5.

An additional parameter for the present analysis is related to the health hazard ranges of PM25 and PM10 pollutants. The limiting values reported by the European Environmental Agency [64] for the moderate air quality index is 25 µg m^{-3} for PM2.5 and 50 µg m^{-3} for PM10. Although the reported value is for 24 h exposure, it was used instantaneously, cast now as a typical value. Since the hazard level is closely related to the impact of the pollutant, overestimation of the forecast may lead to false positives or false negatives if underestimated. It is quite surprising that, from a strictly qualitative perspective, only a single point was forecast outside the first hazard level (false positive) of 50 µg m^{-3} (see Figure 2, the dashed color line in the PM10 over time).

For the PM2.5 forecast, a similar trend to PM10 was also observed in Figure 2, an expected behavior since it is well known that the PM2.5 and PM10 are highly correlated

variables [65] and share the same physical–chemical sources [66]. The spread of the data along the correlation diagram is more visible but still shows a high degree of forecast consistency. Quantitatively, the calculated MAPE is 31.5 %, a value somewhat larger than the one of PM10, while only 0.7 % are false positives above the hazard level. The general trend shows that the forecast values are in good agreement with the real values along the same cumulative time and the largest individual error measurements, such as in PM10, due to the inflection points of the time series. It is interesting to see that the two largest values (above 100%) are sudden variations in the PM2.5 concentration, which cannot be accounted by our model. Note that no value was removed from the complete data set and no attempt was made to check statistical or experimental outliers. If these points were actually outliers and could have been removed from the dataset, MAPE values ∼20% would have been obtained.

3.2. Seasonal Decomposed Data

The performance of the MLA algorithm can also be quantified for the seasonal decomposed data, in a similar analysis to the one carried out for the original data. Figure 3 shows the results for the MLA using the weekly decomposed data while the analogous plots in Figure 4 correspond to the results for the application on monthly averaged data. Both sets have been obtained, as stated previously, by applying a third degree polynomial Savitzky–Golay filter [63]. It is interesting to note that in the case of weekly decomposition, a higher value of the mean MAPE at the third iteration was found when compared to the original data: 39.7% for PM10 and 33.4% for the case of PM2.5. In contrast, the monthly seasonality treatment gave a mean MAPE at the same level of the original data: 27.0% and 31.8% for PM10 and PM2.5. Such behaviors in the case of the weekly seasonality data were due to the averaging of the full 7 days that diminished the structure of the small data inflections, leading to a more difficult accommodation of the trend by the linear model (see the errors associated with the real values in Figure 3). The correlation diagram found in Figure 2 corroborates our findings, with the major differences occurring for higher concentration values, in which the moving window average filters were smoothed out in the MLA model. For the case of the monthly decomposition, the data after the moving average filters had smaller fluctuations of lower concentrations allowing more accurate forecasts of the pollutants. The weekly averaged data for PM10 showed a perfect health hazard coherence prediction, showing no false-positives in opposition to the monthly averaged.

The PM2.5 forecasting yielded statistical results in-line with those of PM10 with weekly and monthly decompositions. Both correlation plots indicate that the model achieves a good description of the data, while the error is approximately below 30%. In this case, the season decomposition did not bring major discrepancies when compared to the original data due to the fact that the PM2.5 series had negligible seasonal variation. From a qualitative perspective, the health hazard level prediction reached 98.1 and 98.4 for the weekly and monthly decomposed signals, respectively, with the number of false-positives being less than 1%. Prior to the comparative statistical analysis, the results using the time series data set with different levels of seasonal decompositions were in good agreement with each other, yielding similarly good results with MAPE values near the 30 % margin.

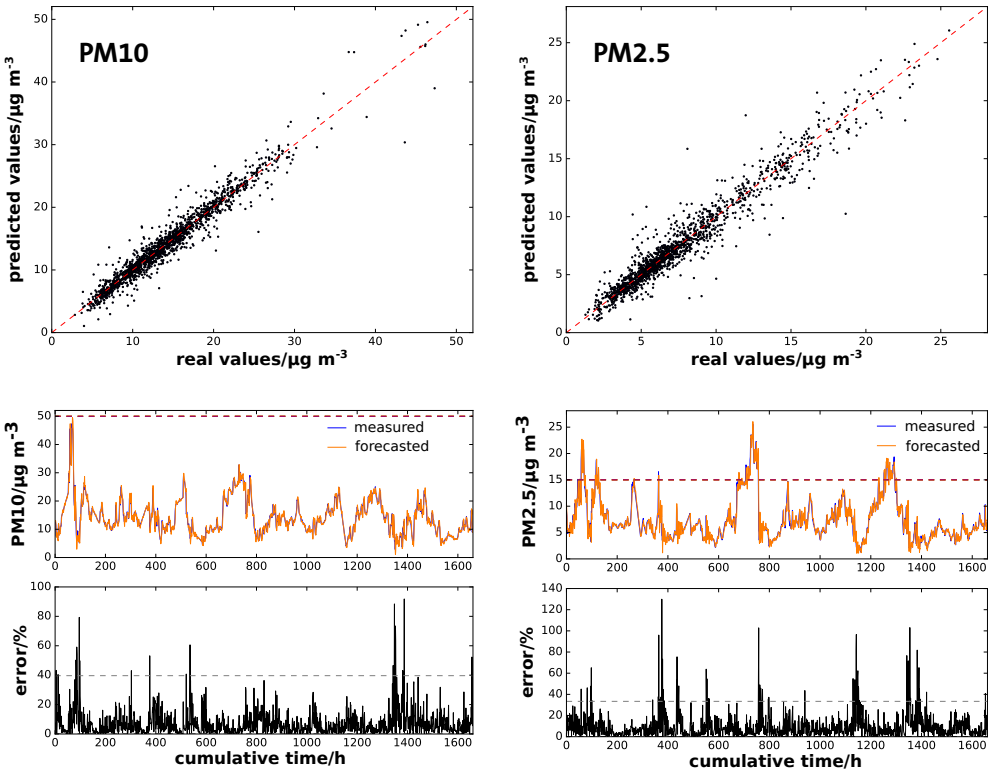

Figure 3. Dispersion, concentration, and errors over time dependencies for the third test iteration of forecasting without seasonality decomposition with weekly seasonality decomposition. Dispersion plots: $R^2 = 0.917$ for PM10 and $R^2 = 0.909$ PM2.5.

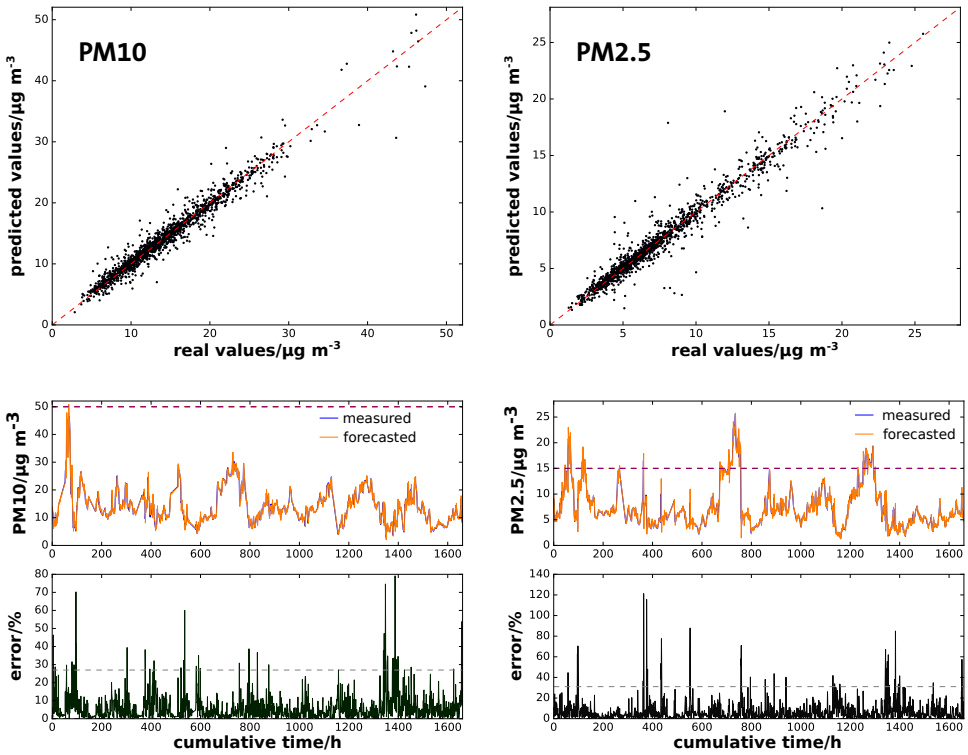

Figure 4. Dispersion, concentration, and errors over time dependencies for the third test iteration of forecasting without seasonality decomposition with monthly seasonality decomposition. Dispersion plots: $R^2 = 0.932$ for PM10 and $R^2 = 0.917$ PM2.5.

3.3. Statistical Analysis

Table 1 shows a summarizes the statistical results obtained by the application of the MLA developed in this work. It shows the standard deviations of the iteration MAPE values. These values are of great importance because they can be used as tie-breaking criteria to choose the best model outcomes. Comparing PM10 forecasts, the MLA using weekly decomposed data yielded the highest MAPE value, 32.3 %, and with a very high standard deviation between measurements, 6.53%. Such behavior indicates that the model is unstable between iterations. Between the applications of the original and monthly decomposed data, MAPE values were similar, and despite the original being slightly lower, 26.6 %, its standard deviation was also higher, 0.24%. Weighting these two aspects, both these applications were viable, the choice of the model was based on accuracy and coherence. PM2.5 forecasting had a similar analysis. Weekly decomposition led to standard deviations of very high (2.23) testing iterations, which made it doubtful regarding coherence. Between the original and the monthly decomposed data set, results were similar and the MAPE was lower for the original, 31.5 %, while the standard deviation was lower, 0.9, for the monthly decomposed data. Moreover, for the weekly decomposition results between iterations, the MAPE in this case increased nearly 30% between the three testing iterations. Since the dataset between iterations was randomly chosen to force a strong validation of the MLA, the moving window average may have led a more complex concentration variation overtime than could not be accounted by the linear-regression.

Table 1. Summary of the MLA results for PM10 and PM2.5 for the three iteration validations. The probability for the false positives and respective coherence for the health hazard levels defined by EEA [64] are also shown. All values (except for standard deviation (SD)) are in percentages.

			Original Data	Weekly Decomp.	Monthly Decomp.
PM10	**MAPE**	iter. 1	26.8	27.2	27.3
		iter. 2	26.6	30.0	27.0
		iter. 3	26.3	39.7	27.0
		mean	26.6	32.3	27.1
		SD	0.2	6.5	0.2
	hazard level	coherence	99.9	100	99.9
		false positives	0.06	-	0.06
PM2.5	**MAPE**	iter. 1	32.5	29.7	32.8
		iter. 2	31.3	33.8	31.6
		iter 3	30.7	33.4	31.1
		mean	31.5	32.3	31.8
		SD	0.9	2.2	0.9
	hazard level	coherence	98.5	98.1	98.4
		false positives	0.8	1.0	0.8

4. Conclusions

In this work, it has been presented a simple but robust machine learning algorithm for the forecast of PM10 and PM2.5 concentrations. The developed MLA was able to predict PM10 and PM2.5 variations coherently; a replication of the results could easily be achieved to account for proper statistical meaning. The algorithm may also be employed in other variables since the training can be carried out in a simple and fast way with minimal computational resources. The mean MAPE calculated had a typical value of 30% for both air pollutants, with the best results obtained for training with the original data. The application of the weekly decomposition method to the PM10 original data used to remove seasonality worsened the results by approximately 6%. Since the goal was to forecast near-future events, the weekly seasonality removed important trends, i.e., hourly increasing traffic,

by smoothing out the data. For PM2.5, all of the testing iterations produced accurate results. In both cases, the health hazard coherence was around 98% with false positives less than 1% of the overall data. Multi-step time-series forecasting is an ambitious approach due to the uncertainties of predictions for large forecasting horizons [67,68]. Linear regression machine learning algorithms are among the most widely used to address these problems given that they are versatile and easy to implement. Substantiated by the knowledge that approaches based on multivariate regression are shown to regularly be better than those based on multiple regression, an attempt at forecasting PM10 and PM2.5 variations in the atmosphere using a multi-output MLA can be used for a fast warning system.

Author Contributions: The authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: The CQC-IMS is financially supported by the Portuguese Science Foundation ("Fundação para a Ciência e a Tecnologia"—FCT)—Projects UIDB/00313/2020 and UIDP/00313/2020 (national funds). This work was also supported by the Copernicus Academy–European Union's Earth Observation Programme.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Data can be obtained via Copernicus Services—CAMS as in Ref. [48] or by requesting to the authors.

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

References

1. *World Health Statistics 2022: Monitoring Health for the SDGs, Sustainable Development Goals*; World Health Organization: Geneve, Switzerland, 2022.
2. Wang, L.; Zhong, B.; Vardoulakis, S.; Zhang, F.; Pilot, E.; Li, Y.; Yang, L.; Wang, W.; Krafft, T. Air Quality Strategies on Public Health and Health Equity in Europe—A Systematic Review. *Int. J. Environ. Res. Public Health* **2016**, *13*, 1196. [CrossRef] [PubMed]
3. Chu, Y.; Liu, Y.; Li, X.; Liu, Z.; Lu, H.; Lu, Y.; Mao, Z.; Chen, X.; Li, N.; Ren, M.; et al. A Review on Predicting Ground PM2.5 Concentration Using Satellite Aerosol Optical Depth. *Atmosphere* **2016**, *7*, 129. [CrossRef]
4. Leung, D.Y. Outdoor-indoor air pollution in urban environment: Challenges and opportunity. *Front. Environ. Sci.* **2015**, *2*, 69. [CrossRef]
5. Katsouyanni, K. Ambient air pollution and health. *Br. Med. Bull.* **2003**, *68*, 143–156. [CrossRef]
6. Macintyre, H.L.; Heaviside, C.; Neal, L.S.; Agnew, P.; Thornes, J.; Vardoulakis, S. Mortality and emergency hospitalizations associated with atmospheric particulate matter episodes across the UK in spring 2014. *Environ. Int.* **2016**, *97*, 108–116. [CrossRef]
7. Xie, R.; Sabel, C.E.; Lu, X.; Zhu, W.; Kan, H.; Nielsen, C.P.; Wang, H. Long-term trend and spatial pattern of PM 2.5 induced premature mortality in China. *Environ. Int.* **2016**, *97*, 180–186. [CrossRef]
8. De Mattos Neto, P.S.; Cavalcanti, G.D.; Madeiro, F.; Ferreira, T.A. An Approach to Improve the Performance of PM Forecasters. *PLoS ONE* **2015**, *10*, e0138507. [CrossRef]
9. Gautam, S.; Yadav, A.; Tsai, C.J.; Kumar, P. A review on recent progress in observations, sources, classification and regulations of PM2.5 in Asian environments. *Environ. Sci. Pollut. Res.* **2016**, *23*, 21165–21175. [CrossRef]
10. Fajersztajn, L.; Veras, M.; Barrozo, L.V.; Saldiva, P. Air pollution: A potentially modifiable risk factor for lung cancer. *Nat. Rev. Cancer* **2013**, *13*, 674–678. [CrossRef]
11. Feng, J.; Yang, W. Effects of Particulate Air Pollution on Cardiovascular Health: A Population Health Risk Assessment. *PLoS ONE* **2012**, *7*, e33385. [CrossRef]
12. Kim, H.; Kim, W.H.; Kim, Y.Y.; Park, H.Y. Air Pollution and Central Nervous System Disease: A Review of the Impact of Fine Particulate Matter on Neurological Disorders. *Front. Public Health* **2020**, *8*, 921. [CrossRef]
13. Sîrbu, C.A.; Stefan, I.; Dumitru, R.; Mitrica, M.; Manole, A.M.; Vasile, T.M.; Stefani, C.; Ranetti, A.E. Air Pollution and Its Devastating Effects on the Central Nervous System. *Healthcare* **2022**, *10*, 1170. [CrossRef]
14. Breitner, S.; Schneider, A.; Devlin, R.B.; Ward-Caviness, C.K.; Diaz-Sanchez, D.; Neas, L.M.; Cascio, W.E.; Peters, A.; Hauser, E.R.; Shah, S.H.; et al. Associations among plasma metabolite levels and short-term exposure to PM2.5 and ozone in a cardiac catheterization cohort. *Environ. Int.* **2016**, *97*, 76–84. [CrossRef]
15. Gozzi, F.; Della Ventura, G.; Marcelli, A. Mobile monitoring of particulate matter: State of art and perspectives. *Atmos. Poll. Res.* **2016**, *7*, 228–234. [CrossRef]
16. Marcazzan, G.M.; Vaccaro, S.; Valli, G.; Vecchi, R. Characterisation of PM10 and PM2.5 particulate matter in the ambient air of Milan (Italy). *Atmos. Environ.* **2001**, *35*, 4639–4650. [CrossRef]

17. Wallace, L.; Bi, J.; Ott, W.R.; Sarnat, J.; Liu, Y. Calibration of low-cost PurpleAir outdoor monitors using an improved method of calculating PM2.5. *Atmos. Environ.* **2021**, *256*, 118432. [CrossRef]
18. Spurny, K.R. Aerosol Filstration and Sampling. In *Advances in Aerosol Filtration*; Spurny, K.R., Ed.; CRC Press: Boca Raton, FL, USA, 1998; p. 415.
19. Harrison, R.M.; Deacon, A.R.; Jones, M.R.; Appleby, R.S. Sources and processes affecting concentrations of PM10 and PM2.5 particulate matter in Birmingham (U.K.). *Atmos. Environ.* **1997**, *31*, 4103–4117. [CrossRef]
20. Querol, X.; Alastuey, A.; Ruiz, C.R.; Artiñano, B.; Hansson, H.C.; Harrison, R.M.; Buringh, E.; Ten Brink, H.M.; Lutz, M.; Bruckmann, P.; et al. Speciation and origin of PM10 and PM2.5 in selected European cities. *Atmos. Environ.* **2004**, *38*, 6547–6555. [CrossRef]
21. Wilks, D.S. Statistics. *Stat. Methods Atmos. Sci.* **2011**, *100*, 100.
22. Lynch, P. The origins of computer weather prediction and climate modeling. *J. Comp. Phys.* **2008**, *227*, 3431–3444. [CrossRef]
23. Nazif, A.; Mohammed, N.I.; Malakahmad, A.; Abualqumboz, M.S. Application of Step Wise Regression Analysis in Predicting Future Particulate Matter Concentration Episode. *Water Air Soil Pollut.* **2016**, *227*, 117. [CrossRef]
24. Ben Taieb, S.; Bontempi, G.; Atiya, A.F.; Sorjamaa, A. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Syst. Appl.* **2012**, *39*, 7067–7083. [CrossRef]
25. Pérez, P.; Trier, A.; Reyes, J. Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. *Atmos. Environ.* **2000**, *34*, 1189–1196. [CrossRef]
26. Sfetsos, A.; Vlachogiannis, D. Time Series Forecasting of Hourly PM10 Using Localized Linear Models. *J. Soft. Eng. App.* **2010**, *3*, 374–383. [CrossRef]
27. Cabaneros, S.M.; Calautit, J.K.; Hughes, B.R. A review of artificial neural network models for ambient air pollution prediction. *Environ. Mod. Soft.* **2019**, *119*, 285–304. [CrossRef]
28. Kalajdjieski, J.; Zdravevski, E.; Corizzo, R.; Lameski, P.; Kalajdziski, S.; Pires, I.M.; Garcia, N.M.; Trajkovik, V. Air Pollution Prediction with Multi-Modal Data and Deep Neural Networks. *Remote Sens.* **2020**, *12*, 4142. [CrossRef]
29. Fan, J.; Li, Q.; Hou, J.; Feng, X.; Karimian, H.; Lin, S. A spatiotemporal prediction framework for air pollution based on deep RNN. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *4*, 15. [CrossRef]
30. Yi, X.; Zhang, J.; Wang, Z.; Li, T.; Zheng, Y. Deep distributed fusion network for air quality prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018 ; pp. 965–973.
31. Liu, B.; Yan, S.; Li, J.; Qu, G.; Li, Y.; Lang, J.; Gu, R. A sequence-to-sequence air quality predictor based on the n-step recurrent prediction. *IEEE Access* **2019**, *7*, 43331–43345. [CrossRef]
32. Ceci, M.; Corizzo, R.; Japkowicz, N.; Mignone, P.; Pio, G. Echad: Embedding-based change detection from multivariate time series in smart grids. *IEEE Access* **2020**, *8*, 156053–156066. [CrossRef]
33. Li, X.; Peng, L.; Hu, Y.; Shao, J.; Chi, T. Deep learning architecture for air quality predictions. *Environ. Sci. Poll Res.* **2016**, *23*, 22408–22417. [CrossRef] [PubMed]
34. Kök, I.; Şimşek, M.U.; Özdemir, S. A deep learning model for air quality prediction in smart cities. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017 ; IEEE: Piscataway, NJ, USA , 2017; pp. 1983–1990. [CrossRef]
35. Qi, Z.; Wang, T.; Song, G.; Hu, W.; Li, X.; Zhang, Z. Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 2285–2297. [CrossRef]
36. Li, T.; Shen, H.; Yuan, Q.; Zhang, X.; Zhang, L. Estimating ground-level PM2.5 by fusing satellite and station observations: A geo-intelligent deep learning approach. *Geophys. Res. Lett.* **2017**, *44*, 11–985. [CrossRef]
37. Yin, L.; Wang, L.; Huang, W.; Tian, J.; Liu, S.; Yang, B.; Zheng, W. Haze Grading Using the Convolutional Neural Networks. *Atmosphere* **2022**, *13*, 522. [CrossRef]
38. Kow, P.Y.; Chang, L.C.; Lin, C.Y.; Chou, C.C.; Chang, F.J. Deep neural networks for spatiotemporal PM2.5 forecasts based on atmospheric chemical transport model output and monitoring data. *Environ. Pollut.* **2022**, *306*, 119348. [CrossRef]
39. Justus, D.; Brennan, J.; Bonner, S.; McGough, A.S. Predicting the Computational Cost of Deep Learning Models. In Proceedings of the 2018 IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, 10–13 December 2018 ; pp. 3873–3882. [CrossRef]
40. Gardner-Frolick, R.; Boyd, D.; Giang, A. Selecting Data Analytic and Modeling Methods to Support Air Pollution and Environmental Justice Investigations: A Critical Review and Guidance Framework, 2022. *Environ. Sci. Technol.* **2022**, *56*, 2843-–2860. [CrossRef]
41. Kaur, H.; Pannu, H.S.; Malhi, A.K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions, 2019. *ACM Comput. Surv.* **2020**, *52*, 79. [CrossRef]
42. Ramsundar, B.; Zadeh, R.B. *TensorFlow for Deep Learning: From Linear Regression to Reinforcement Learning*; O'Reilly Media: Beijing, China, 2018.
43. Juneng, L.; Latif, M.T.; Tangang, F. Factors influencing the variations of PM10 aerosol dust in Klang Valley, Malaysia during the summer. *Atmos. Environ.* **2011**, *45*, 4370–4378. [CrossRef]
44. Ng, K.Y.; Awang, N. Multiple linear regression and regression with time series error models in forecasting PM10 concentrations in Peninsular Malaysia. *Environ. Monit. Assess.* **2018**, *190*, 63. [CrossRef]

45. Shams, S.R.; Jahani, A.; Kalantary, S.; Moeinaddini, M.; Khorasani, N. The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models for predicting SO_2 concentration. *Urban Clim.* **2021**, *37*, 100837. [CrossRef]
46. Okkaoğlu, Y.; Akdi, Y.; Ünlü, K.D. Daily PM10, periodicity and harmonic regression model: The case of London. *Atmos. Environ.* **2020**, *238*, 117755. [CrossRef]
47. Bai, L.; Wang, J.; Ma, X.; Lu, H. Air Pollution Forecasts: An Overview. *Int. J. Environ. Res. Public Health* **2018**, *15*, 780. [CrossRef] [PubMed]
48. Generated Using Copernicus Atmosphere Monitoring Service Information 2020. Available online: https://atmosphere.copernicus.eu/data (accessed on 20 January 2020).
49. Galmarini, S.; Bianconi, R.; Klug, W.; Mikkelsen, T.; Addis, R.; Andronopoulos, S.; Astrup, P.; Baklanov, A.; Bartniki, J.; Bartzis, J.C.; et al. Ensemble dispersion forecasting—Part I: Concept, approach and indicators. *Atmos. Environ.* **2004**, *38*, 4607–4617. [CrossRef]
50. Galmarini, S.; Bianconi, R.; Addis, R.; Andronopoulos, S.; Astrup, P.; Bartzis, J.C.; Bellasio, R.; Buckley, R.; Champion, H.; Chino, M.; et al. Ensemble dispersion forecasting—Part II: Application and evaluation. *Atmos. Environ.* **2004**, *38*, 4619–4632. [CrossRef]
51. Marécal, V.; Peuch, V.H.; Andersson, C.; Andersson, S.; Arteta, J.; Beekmann, M.; Benedictow, A.; Bergström, R.; Bessagnet, B.; Cansado, A.; et al. A regional air quality forecasting system over Europe: The MACC-II daily ensemble production. *Geosci. Model Dev.* **2015**, *8*, 2777–2813. [CrossRef]
52. Terry, W.R.; Lee, J.B.; Kumar, A. Time series analysis in acid rain modeling: Evaluation of filling missing values by linear interpolation. *Atmos. Environ.* **1986**, *20*, 1941–1943. [CrossRef]
53. Dickey, D.A.; Fuller, W.A. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *J. Amer. Stat. Ass.* **1979**, *74*, 427–431. [CrossRef]
54. Spiegel, M.R.; Stephens, L.J. *Schaum's Outline of Theory and Problems of Probability and Statistics*; McGraw-Hill: New York, NY, USA , 2008. [CrossRef]
55. Bontempi, G. Long term time series prediction with multi-input multi-output local learninge. In Proceedings of the 2nd ESTSP 2008 , Porvoo, Finland, 17–19 September 2008; pp. 145–154.
56. Ben Taieb, S.; Sorjamaa, A.; Bontempi, G. Multiple-output modeling for multi-step-ahead time series forecasting. *Neurocomputing* **2010**, *73*, 1950–1957. [CrossRef]
57. Bontempi, G.; Ben Taieb, S.; Le Borgne, Y.A. Machine learning strategies for time series forecasting. *Lect. Notes Bus. Infor. Proc.* **2013**, *138*, 62–77. [CrossRef]
58. Qin, J.; Guo, J.; Xu, X.; Kong, T.; Wang, X.; Ma, L.; Wurm, M. A universal and fast method to solve linear systems with correlated coefficients using weighted total least squares. *Meas. Sci. Technol.* **2021**, *33*, 015017. [CrossRef]
59. Sanders, R. The pareto principle: Its use and abuse. *J. Serv. Mark.* **1987**, *1*, 37–40. [CrossRef]
60. Shao, J. Linear model selection by cross-validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486–494. [CrossRef]
61. Fushiki, T. Estimation of prediction error by using K-fold cross-validation. *Stat. Comput.* **2009**, *21*, 137–146. [CrossRef]
62. Makridakis, S. Accuracy measures: Theoretical and practical concerns. *Int. J. Forecast.* **1993**, *9*, 527–529. [CrossRef]
63. Savitzky, A.; Golay, M.J. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36*, 1627–1639. [CrossRef]
64. *Air Quality in Europe 2021*; Technical Report; European Environment Agency: Copenhagen, Denmark , 2022.[CrossRef]
65. Gehrig, R.; Buchmann, B. Characterising seasonal variations and spatial distribution of ambient PM10 and PM2.5 concentrations based on long-term Swiss monitoring data, 2003. *Atmos. Environ.* **2003**, *37*, 2571–2580. [CrossRef]
66. Chow, J.C.; Watson, J.G. Review of PM2.5 and PM10 apportionment for fossil fuel combustion and other sources by the Chemical Mass Balance receptor model. *Energy Fuels* **2002**, *16*, 222–260. [CrossRef]
67. Khashei, M.; Bijari, M. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Appl. Soft. Comput.* **2011**, *11*, 2664–2675. [CrossRef]
68. Kumar, R.; Kumar, P.; Kumar, Y. Multi-step time series analysis and forecasting strategy using ARIMA and evolutionary algorithms. *Int. J. Inf. Technol.* **2022**, *14*, 359–373. [CrossRef]