atmosphere

MDPI

Article

A New Method for the Evaluation and Visualization of Air Pollutant Level Predictions

Jana Faganeli Pucer

Faculty of Computer and Information Science, University of Ljubljana, 1000 Ljubljana, Slovenia; jana.faganeli@fri.uni-lj.si; Tel.: +386-1-479-8226

Abstract: Accurately predicting air pollutant levels is very important for mitigating their effects. Prediction models usually fail to predict sudden large increases or decreases in pollutant levels. Conventional measures for the assessment of the performance of air pollutant prediction models provide an overall assessment of model behavior, but do not explicitly address model behavior when large changes are observed. In our work, we propose a method to automatically label the observed large changes. We also propose two visualization methods and two measures that can help assess model performance when sudden large changes in pollutant levels occur. The developed measures enable the assessment of model performance only for large changes (MAE of large changes), or weigh the model residuals by the rate of change (WErr), making the evaluation measures "cost-sensitive". To show the value of the novel evaluation and visualization methods, we employ them in the evaluation of three empirical examples—different statistical models used in real-life settings and a popular atmospheric dispersion model. The proposed visualizations and measures can be a valuable complement to conventional model assessment measures when the prediction of large changes is as important as (even if they are rare) or more important than predictions of other levels.

Keywords: air pollutant levels; prediction model; large changes; performance measure; visualization

check for updates

Citation: Faganeli Pucer, J. A New Method for the Evaluation and Visualization of Air Pollutant Level Predictions. *Atmosphere* **2022**, *13*, 1456. https://doi.org/10.3390/atmos13091456

Academic Editor: Joaquim Arteta

Received: 14 August 2022
Accepted: 2 September 2022
Published: 8 September 2022

1. Introduction

Inhalation of polluted air is a major environmental health risk [1–4]. Its short- and long-term exposure can lead to severe adverse health effects such as reduced lung function, respiratory infections, asthma, heart disease, and diabetes. This is why most countries have put legislation in place to dictate how air pollutants should be measured, evaluated, and forecasted with regard to specific predefined limit values. In Europe, air pollutant level measurements and their limit values are defined in Directive 2008/50/EC [5]. The directive also dictates that member states have to forecast the exceedances of the limit values. Accurate forecasting of elevated levels of air pollutants can alleviate their impact on human health. EU member states have to predict the exceedances of the daily limit value for PM_{10}, the exceedances of the information threshold for ozone (O_3), and the exceedances of the alert threshold for sulfur dioxide (SO_2) and nitrogen dioxide (NO_2).

Different types of models are used to forecast air pollutant levels, most commonly photochemical atmospheric dispersion models [6–10] and different statistical models. Some statistical models used for the prediction of pollutant levels are described in Taheri S. and Sodoudi [11], Lu and Wang [12], Dutot et al. [13], Faganeli Pucer et al. [14], Sharma et al. [15], Kocijan et al. [16] and de Gennaro et al. [17]. Photochemical atmospheric dispersion models usually produce spatial three-dimensional rasters of pollutant levels as outputs while statistical models mostly produce point predictions. Models are usually validated by comparing their outputs to pollutant measurements from a particular location (point measurements) [18], so both types of models (chemical dispersion and statistical models) are validated as point predictions. There have been several criteria used for the assessment of the performance of air quality models, such as the measures that evaluate the distance between point prediction and measured values e.g., root mean square error (RMS) [13,19–26], mean absolute error (MAE) [13,14,25–27], and mean absolute percentage error [28]. Those measures are useful when models predict pollutant levels as continuous values (e.g., regression models). Different indexes that assess model performance as normalized measures have also been used [15] for example the correlation coefficient, Willmott Index [25,29], Nash–Sutcliffe Efficiency [30], and Legates and McCabe Index [31]. Multiple evaluation measures used in air pollutant level forecasting are described in Sharma et al. [15], Kocijan et al. [16], and Carslaw and Ropkins [32]. When the model predicts the exceedance of a predefined threshold (e.g., PM_{10} daily limit value of 50 $\mu g/m^3$) or predicts the range (class) of the pollutant level, the evaluation is performed in terms of accuracy, true positive rate, false positive rate [17,26,33,34], or the success index [13,17]. If the model predicts the probability of exceedance of a predefined threshold, the performance can be evaluated by using the logarithmic score [14].

Measures that assess the mean or squared mean (MSE, MAE) of the residuals estimated as the difference between the observed and modeled values provide an overall assessment of model performance, which is intuitive (probably MAE is the most intuitive measure). When using MSE instead of MAE for the assessment of model performance large residuals provide a greater contribution to the overall error. All the coefficients and indices listed above give a comparison of modeled and measured levels. If we model daily air pollutant levels with the simplest model possible, the "persistence model" (today's level is the same as yesterday's), and assess its performance using the previously mentioned measures, the persistence model does not seem that inefficient. For example, if we assess the performance of the persistence model in terms of MAE, in most instances the error would not be very high as sudden large changes in pollutant levels are not very common in Europe [35]. The residuals (of the persistence model) would be small when the change in pollutant level is small from one day (one modelled value) to the next. Still, persistence is not a good model as it is incapable of predicting changes in pollutant levels. All measures listed above cannot assess the performance of the models when we are particularly interested in their ability to predict large pollutant level changes. The conventional measures are not "cost-sensitive".

For model evaluation, different visualizations are also used. Still, the most popular visualization is the plot depicting the observed values against the modeled values [18,24,26,36]. Different scatter plots, residual plots, and quantile-quantile plots are also frequently used [18,36]. For the visual evaluation of air quality models, the Taylor diagram [37] and the Target plot [38], which enable the visualization of more than one metric in the same plot, or the polar coordinate diagram of the relative prediction error [26], which shows the distribution of the errors for different times of the day, are also used. Recently, some visualizations for feature importance have been proposed [21]. All these visualizations show interesting aspects of model performance but tell nothing about model performance when predicting large changes.

When we were developing models to predict PM_{10} and O_3 levels with the Slovenian Environment Agency (ARSO) [14], air quality experts emphasized that the critical task when predicting pollutant levels is to correctly predict the onset and offset of an episode of high PM_{10} or ozone levels. According to ARSO, correctly predicting large increases and decreases is more important than the average model performance. They also wanted measures that would enable them to evaluate model performance during these rare events. This was the main motivation for our work.

High PM_{10} levels in Slovenia occur in winter when the meteorological situation is stable with an unfavorable dispersion situation (temperature inversion and low wind conditions) [39]. This is usually known as air stagnation, which is characterized by stable weather, low wind in the lower atmosphere, and no precipitation [27,40,41]. Even emissions are affected by the meteorological situation, e.g., a sudden decrease in temperature affects the amount of indoor heating and discourages people from biking or walking to work, which increases the amount of traffic. Such a decrease in temperature accompanied by stable, low wind, low-temperature weather in Slovenia marks the beginning of an episode of high PM_{10} levels. When such an episode is forecasted, precautions can be taken to limit PM emissions (e.g., limit traffic or discourage people from burning wood). Similarly, high ozone levels are not only induced by the share of ozone precursors and solar radiation, but also by air stagnation [40]. High ozone levels are common in summer when solar radiation is high and hours of daylight are long [42]. If high ozone levels are forecasted for Slovenia, people are advised not to exercise outdoors and to avoid going to the mountains. As such measures affect people's lives it is also important to carefully predict the end of such episodes, so the implemented measures can cease. A similar situation occurs in all of central Europe. This is the main reason that air quality experts want their models to be able to predict large increases and decreases.

In our work, we present two new measures for the evaluation of prediction model performance adapted for the evaluation of large pollutant level changes. We also propose two new visualizations for the visual evaluation of model performance in these extreme situations.

The remainder of the paper is organized as follows. In Section 2 we define the pollutant level changes and discuss what constitutes large changes, we propose new visualizations and measures that evaluate large changes. In Section 3 we present three examples of the applicability of the presented visualizations and measures. We conclude the paper with Section 4.

2. The Visualization and Evaluation of Large Changes

In this section, we define changes in pollutant levels (see Section 2.1), large pollutant level changes (see Section 2.2) and propose two new visualizations (see Section 2.3) and two new measures (see Section 2.4) for the evaluation of air quality models. Figure 1 represents the steps taken to get to the final model evaluation.

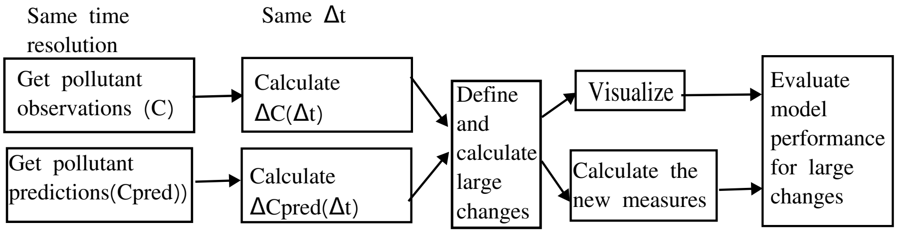

Same time resolution
Same Δt
Get pollutant observations (C)
Get pollutant predictions(Cpred))
Calculate ΔC(Δt)
Calculate ΔCpred(Δt)
Define and calculate large changes
Visualize
Calculate the new measures
Evaluate model performance for large changes

Figure 1. The flow of the proposed methods for air pollutant model predictions visualization and evaluation.

2.1. *Defining Pollutant Level Changes*

We denote the change in pollutant level as ΔC(x,Δ T), where $C(x, T)$ denotes the observed value of an investigated pollutant at a certain location (x) at a certain time and $C(x, T + t)$ the same pollutant t time in advance at the same location (x). $Cpred(x, T)$ denotes the level predicted by a certain model for the same location as $C(x, T)$. As all observations and predictions are made for the same location x, we will discard x from the equations and write the changes in pollutant level as ΔC(Δt), which is defined in Equation (1). As an example, we can evaluate a statistical model that predicts today's PM_{10} levels in Ljubljana and we want to assess if the model correctly predicted how much the daily average PM_{10} level will increase today from the level that was observed yesterday, so we define:

$$\Delta C(\Delta t) = C(T + t) - C(T). \tag{1}$$

$C(T)$ is the last available observation before we run the model, e.g., if we are predicting the PM_{10} level for tomorrow and the last available PM level for today is at midnight. ΔC(Δt) denotes the difference in observed pollutant level (measured values) at time instant T and time instant $T + t$.

We also define the predicted change in pollutant level ΔCpred(Δt):

$$\Delta Cpred(\Delta t) = Cpred(T + t) - C(T). \tag{2}$$

Cpred$(T + t)$ is the value predicted by the model that is located at t time instants in the future. ΔC(Δt) describes how much the pollutant level changed from one observation to the next (the true increase) while ΔCpred(Δt) describes the change assessed by the model.

2.1.1. Defining the ΔCpred(Δt) for Photochemical Atmospheric Dispersion Models

When assessing the performance of statistical models, we define ΔCpred(Δt) (described in Section 2.1) as the difference between the predicted value and the last observed value. If we are dealing with photochemical atmospheric dispersion models (e.g., model EMEP MSC-W [6] model, see Section 3.3), the observed value is not available or at least is not taken into account by the model. Photochemical dispersion models are usually run without past measurements as inputs.

If we want to evaluate such models using the new measures, we must define ΔCpred(Δt) differently than with statistical models. We propose defining ΔCpred(Δt) as the difference between the predicted level at time T and the predicted level at time $T + t$. This shows how the model is capable of predicting large changes in pollutant levels disregarding the actual predicted levels. We speculate that sometimes photochemical dispersion models are capable of predicting large changes correctly, but the produced value (C(T)) is not correct due to some inherent biases.

2.2. Defining Large Pollutant Level Changes

Intuitively, a large change is a change in pollutant level between time T and $T + t$ that is much larger than the average change in pollutant levels in this time interval. Defining what a large change is can be very ambiguous. It could be defined by a field expert or could be defined statistically. For example, an expert on air pollution can quickly identify sudden large increases or decreases in pollutant levels at a certain location, while people who do not deal with these issues cannot spot these changes. This is why we propose a method based on the statistical properties of pollutant level changes in a certain time interval.

We propose modeling the differences ΔC(Δt) with the t-distribution [43], which is very similar to the normal distribution, but it can accommodate heavier tails. The t-distribution is a symmetric and bell-shaped distribution. It is a class of distributions, not one distribution. When fitting a t-distribution, we have to specify the degrees of freedom. The larger the number of degrees of freedom, the more similar to the normal distribution it becomes (when the t-distribution has infinite degrees of freedom it becomes the normal distribution). At lower degrees of freedom, it can model heavier tails than the normal distribution [44]. The t-distribution is usually assumed to be centered at 0 and is not scaled. In our case, we use a t-distribution with a location and a scale parameter (*metRology R* package). The location parameter enables the distribution not to be centered at 0 and the scale parameter enables the adaptation of the width of the distribution. To fit the distributions, we use the *fitdistr* function of the *'fitdistrplus'* [45] *R* [46] package.

Figure 2 shows the fit of a normal and t-distribution to ΔC(Δt) values from Ljubljana, Slovenia. The shape of the fitted t-distribution gives a better fit than the normal distribution. Using the fitted distribution we can evaluate the limits of the largest ΔC(Δt) values (e.g., the largest 5%, 10% level changes).

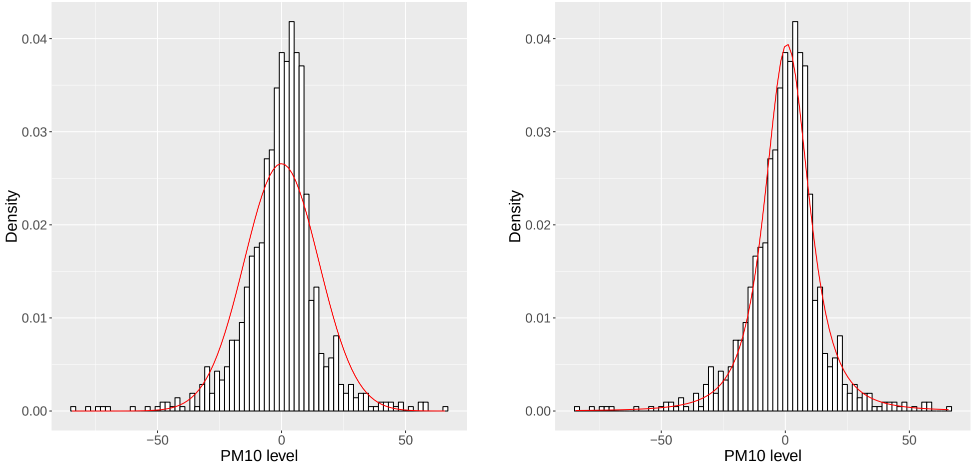


Figure 2. Fitting the normal (**left**) and t-distribution (**right**) on ΔC(Δt) PM_{10} values from Ljubljana, Slovenia for the period 2014–2016.

2.3. Visualizations of ΔC(Δt) for the Assessment of Model Performance

To assess regression model performance, we usually examine different plots. The preferred plots are those showing the predicted values against the actual values or different residual plots. To better understand how the models perform with large changes, we propose two additional visualizations. The visualizations do not focus on the predicted values but on the predicted change.

Figure 3 shows the residuals of the model against the observed values (ΔC(Δt)). The figure shows the predictions against the observed change in PM_{10} levels in Ljubljana in the period from 2014 to 2016. In this example, a random forest model [47] was trained as described in [14] (this is a re-analysis of those models) for predicting the daily mean PM_{10} levels in the morning of the current day. The time series of observed and predicted values of different models for Ljubljana are shown in the Appendix B. The vertical lines in the plot show the limits of the 95% and 90% intervals (the intervals containing those percentages of data as described in Section 2.2). The large changes are represented by points lying outside the selected interval (on the right of the upper limit and the left of the lower limit of the interval). The width of the interval is arbitrary, defined by the individual performing the analysis (whether 95% or 90% or even 85%, 80% of the largest changes are observed). In these examples, the large change was defined statistically. For smaller changes, the residuals are distributed evenly around the *x*-axis, but for large changes, they are not. The plot shows that the observed model is prone to underestimate large increases (on the right values lie above the *x*-axis) and large decreases (on the left values lie under the *x*-axis), especially where the largest changes are observed (95% interval). We cannot spot such model behavior by examining an ordinary residual plot.

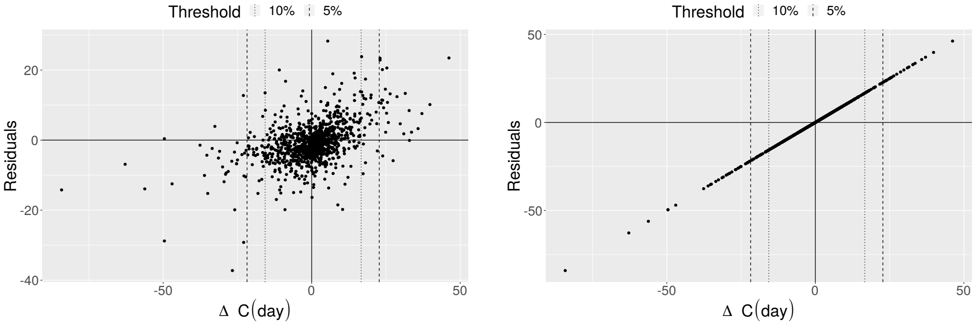


Figure 3. Residuals of the random forest model (**left**) and the persistence model (**right**) for PM_{10} levels forecasted for Ljubljana against ΔC(1 day) actual levels. The persistence model (on the right) is not good as it does not predict any change so the residuals are equal to the observed changes.

Figure 4 shows the actual change ΔC(Δt) against the predicted change ΔCpred(Δt). The figure shows the same predictions for Ljubljana as the first visualization. The dashed vertical lines represent the limits of the 95% interval. The diagonal ($y = x$) shows the location of the "perfect" predictions. The closer the points lie to the line, the more accurate the predicted ΔC(Δt) are. The greatest mistake a model can make is to wrongly predict the direction of the change when large changes occurred in reality; that is, predicting an increase instead of a large decrease or a decrease instead of a large increase. So all points on the left of the y-axis (most importantly on the left of the left dashed line) should be lying below the x-axis and the points on the right of the y-axis (most importantly on the right of the right dashed line) should be lying below the x-axis. The further the points lie from the y-axis the more important it is that they lie on the correct side of the x-axis. In this regard, the example model from Figure 4 performs well; there is only one point lying on the left side of the left dashed line and above the x-axis. From Figure 4 we can also see that the model mostly underestimates large changes.

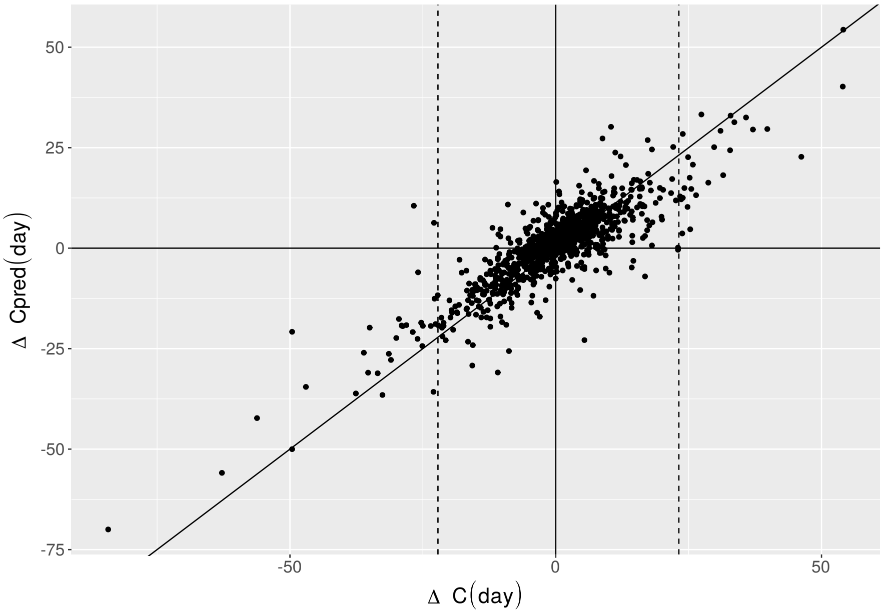


Figure 4. ΔC(1 day) against ΔCpred(1 day) for Ljubljana as predicted by the random forest model.

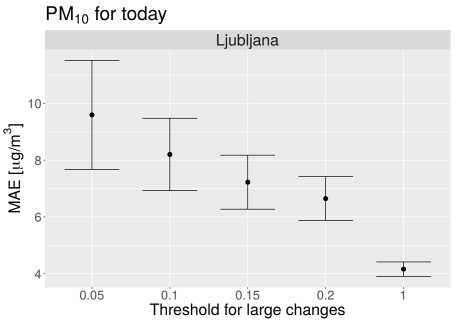
The visualization shown in Figure 3 helps evaluate how the residuals relate to observed changes in pollutant levels if there is some kind of expected behavior associated with the amount of the observed change, e.g., in our case the model underestimates almost all large changes. The visualization from Figure 4 helps evaluate the relationship between the observed changes and the predicted changes. The greatest mistake air quality models can make is to predict a decrease instead of a large increase or vice versa and, by looking at Figure 4, it is easy to spot such instances.

2.4. Performance Measures

As discussed in the Introduction section, numeric measures for the evaluation of model performance usually evaluate the overall (unweighted) performance of the model or the classification performance (accuracy, sensitivity, specificity, etc.) when a threshold is predefined. To evaluate the performance of our models when the prediction of large changes is crucial, we propose two new performance measures.

The first proposed evaluation measure based on the proposed visualization shown in Figure 3 is to calculate the MAE or the MSE only for the data lying outside of the predefined intervals. Figure 5 shows MAE with two standard errors for different shares of largest changes taken into account (from the upper 5% of the largest to the 20% of the largest ΔC(Δt)) we can compare it to the same measure calculated for all values (see Figure 5).

In our example (see Figure 5), MAE increases when we narrow the width of the observed intervals. MAE is the largest when only the most extreme changes are observed.



Share	n	MAE	2 St.err.
0.05	61	9.60	1.92
0.10	115	8.20	1.28
0.15	171	7.22	0.95
0.20	228	6.65	0.78
1.00	1065	4.16	0.25

Figure 5. Model performance on the entire dataset (figure left *x*-axis 1) and on different shares of highest ΔC(Δt). The figure on the left shows MAE with two standard errors for the example discussed above (see Figures 3 and 4). There is a significant increase in MAE for the largest ΔC(Δt) compared to MAE for all instances. We can also display the numerical results in a table (see left; Share—the share of ΔC(Δt) labeled as large, n—number of instances, MAE—mean absolute errors, 2 St. err—two standard errors).

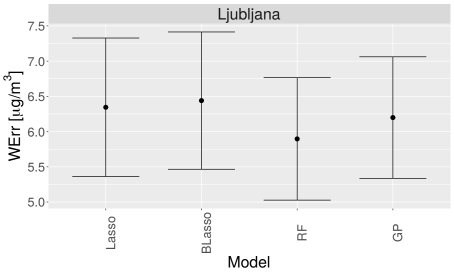
To assess the behavior of all predictions at once, not splitting them as we did with the first proposed measure, we propose to weight the large changes ΔC(Δt) as more important than smaller changes. We propose to weight each absolute error (residual) with the associated absolute value of ΔC(Δt) and divide it by the average absolute value of ΔC(Δt). We will denote the weighted error as WErr and calculate it as:

$$WErr = \frac{|\Delta C(\Delta t)||Cpred - C|}{\overline{|\Delta C(\Delta t)|}}. \tag{3}$$

This way, the residuals of the models predicting large changes are weighed more than residuals of smaller changes.

Both presented measures focus on large changes. The first one only evaluates the residuals for the chosen percentage of largest changes and ignores all others. The second measure tries to evaluate all model residuals at the same time but weighs the residuals for large changes more, and the residuals for smaller changes less.

Figure 6 shows a comparison of different models used for PM_{10} prediction in Ljubljana in the period from 2014 to 2016 based on $WErr$. The random forest model shows a superior performance over other models in terms of $WErr$. Still, the confidence intervals of the RF and GP models show a great overlap.



Model	WErr	2 st.err.
Lasso	6.35	0.98
BLasso	6.44	0.97
RF	5.90	0.87
GP	6.20	0.86

Figure 6. Figure on the left shows WErr with two standard errors for different models trained and tested on the same data (PM_{10} in Ljubljana) the used models are Lasso–L1-regularized regression, BLasso–Bayesian lasso, RF—random forest, GP—Gaussian process. The table on the right shows the numerical values (WErr—calculated as described in Equation (3), 2 st. err—two standard errors).

3. How Can the New Visualizations and Measures Improve the Evaluation of Our Models?

In this section, we show three examples of model evaluation using the methodology described above. We do not train and test new models but render a new analysis of existing models. The first two examples comprise the comparison of different statistical models for PM_{10} and ozone forecasting for the current and following day (a re-analysis of the results presented in [14]); the third example shows a case of a comparison of the performance of a statistical model with a photochemical dispersion model.

3.1. Evaluation of Different Models for PM_{10} Prediction in Nova Gorica, Slovenia

Nova Gorica is a town in western Slovenia with around thirty thousand inhabitants. It is located in the Mediterranean part of Slovenia close to the Italian border. PM_{10} levels are higher in winter than in summer, but they are usually quite low. In winter, PM_{10} levels are affected by increased emissions from indoor heating and an unfavorable dispersion situation. Another contributing factor is the long-range pollution from heavily industrialized and densely populated northern Italy, which has a notable effect on this part of Slovenia [39]. Meteorological conditions (especially the changes in the dispersion situation) can induce sudden increases or decreases in pollutant levels. The following example shows the comparison in performance for different models predicting PM_{10} levels (see [14]) for the current day (morning predictions for the current day).

Figure 7 shows the comparison of the performances of the lasso (Lasso) [48], Bayesian lasso (BLasso) [49], random forest (RF) [47] and Gaussian process (GP) [50] models in Nova Gorica tested on data from 2014 to 2016. The models are briefly described in Appendix A. Figure 7 top left shows that in terms of MAE all four models perform equivalently, while, in terms of WErr, the GP model outperforms all other models. When observing MAE for different percentages of highest Δ C(1 day) values (as described in Section 2.4) the GP model outperforms other models in the prediction of the largest changes (see Figure 7 top right). If we only judge the performance of the models with conventional measures we can conclude that all models perform equivalently. When we explore model performance for large changes, we can see the clear advantage of using the GP model. If we want our model to predict large changes well, the GP model is the best model and it clearly outperforms the RF model.

3.2. Evaluation of Different Models for Ozone Prediction in Koper, Slovenia

Koper is a town on the Slovenian coast. It is home to about thirty thousand inhabitants and lies across the bay from Trieste (Italy), which is a much larger city. The climate in Koper is Mediterranean with hot, dry summers and mild winters. Long sunny, warm days accompanied by local emissions and emissions from across the bay of Trieste and neighboring Italian regions contribute to high ozone levels [39,51]; one of the highest in Slovenia (apart from high altitude measurement sites).

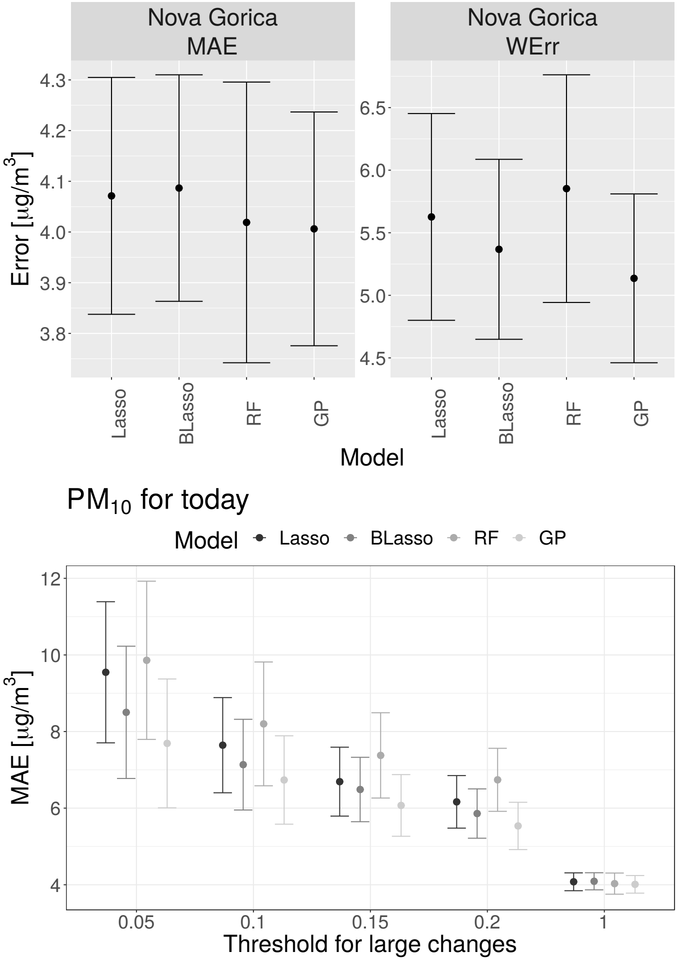


Figure 7. The top left figure shows MAE for different models for PM$_{10}$ prediction for today in Nova Gorica, while figure top right shows WErr for the same models. The lower figure shows MEA for different Δ C(1 day).

Figure 8 shows the comparison of the Lasso, BLasso, RF, and GP models for the prediction of the daily maximal values of ozone in Koper (tested on 2014–2016 data). On average, according to MAE, the best performing model is GP (see Figure 8 up left), but when observing WErr the GP model does not stand out much (see Figure 8 upper figure on the right). According to WErr, all observed models perform similarly although the GP model exhibits the best performance. Overall RF and GP models show a similar performance when predicting large changes; they both outperform the two linear models, which is not evident when only observing the MAE for all instances. By using our evaluation measures, we can conclude that when it is crucial to carefully predict large changes the non-linear models (RF or GP models) are the best models for predicting ozone in Koper.

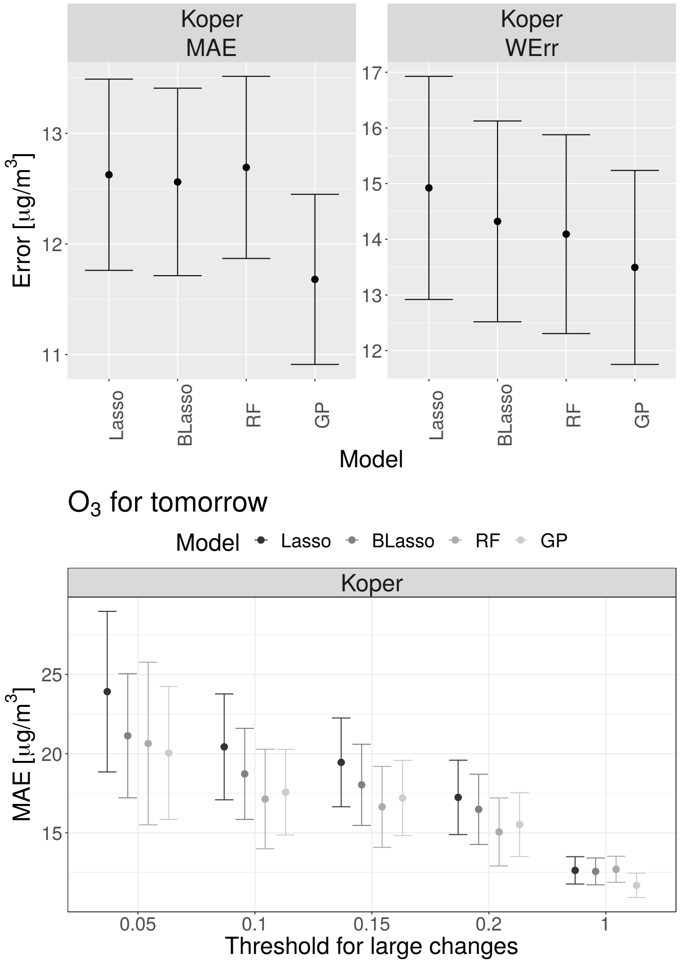


Figure 8. The top left figure shows MAE for different models for O_3 prediction for the following day in Koper, while the top right Figure shows WErr for the same models. The lower figure shows MEA for different Δ C(2 day).

3.3. Comparison of the EMEP MSC-W Model with Statistical Models for the Prediction of PM_{10} Levels

The same evaluation methodology can also be applied to photochemical dispersion models (they are briefly described in Appendix A). As an example, we compare the results of the best performing statistical model for the prediction of PM_{10} levels for the next day (Δt = 2 days) in Ljubljana (GP model) to the predictions of the EMEP MSC-W model [6]. We compare both predictions as point predictions for the same location (modeling point) where the measurements were performed. We selected the EMEP MSC-W model because the results are available online and it is a popular photochemical dispersion model used for the assessment of long-range transport pollution in Europe. The presented example is not intended as a strict comparison of statistical and photochemical dispersion models, nor as a critique of a model type. This is simply a case study on how we can compare the performance of two different types of models as point predictions using the presented methodology. The ΔC(2 days) are calculated as described in Section 2.1.1.

Figure 9 shows similar visualizations to the ones we observed when comparing different statistical models. If we examine the top plots (MAE and WErr) we can conclude that the GP model is a much better performing model than the EMEP model. When analyzing only large changes, we see that as we restrict the observed percentage of the largest changes, the difference between the GP and EMEP models becomes smaller. For only extreme changes, the EMEP model outperforms the GP model. Still, this outperformance of the EMEP model for large changes is not evident in the WErr plot (Figure 9) in the top left. This leads us to question, how is this possible? To further analyze this inconsistency, we visualize our results as proposed in Section 2.3. The top of Figure 10 shows the plots of the residuals against ΔC(2 days) for the two observed models. The changes predicted by the EMEP model sometimes exhibit very high deviation from the observed changes (even when the changes in pollutant levels are small), but in some instances the difference between the actual and predicted changes for very large changes is small. The EMEP model does not underestimate large decreases as consistently as the GP model, but similar to the GP model, generally underestimates large increases. Figure 10 shows ΔC(2 days) against ΔCpred(2 days) for the observed models. As in the previous figure, the predictions of the GP model show smaller divergence (in this figure this means less scatter around the $y = x$ line). The EMEP MSC-W model contrary to the GP model often predicts an increase instead of a large decrease (points lying on the left of the left dashed line and above the x-axis). It also predicts more increases instead of large decreases as compared to the GP model. By looking at those plots, we can better compare the performance of both models and understand why the MAE for extreme changes (the largest 5%) is smaller for the EMEP model, but when observing WErr the GP model clearly outperforms the EMEP model. Some EMEP model residuals are large when small changes in pollutant levels occur and even if they are weighed with much smaller weights than large changes, they increase WErr substantially. By observing the plots in Figure 9 the performance of the GP model looks much more coherent than that of the EMEP model, especially when observing the bottom plots where the points lie much closer to the $y = x$ line for the GP model.

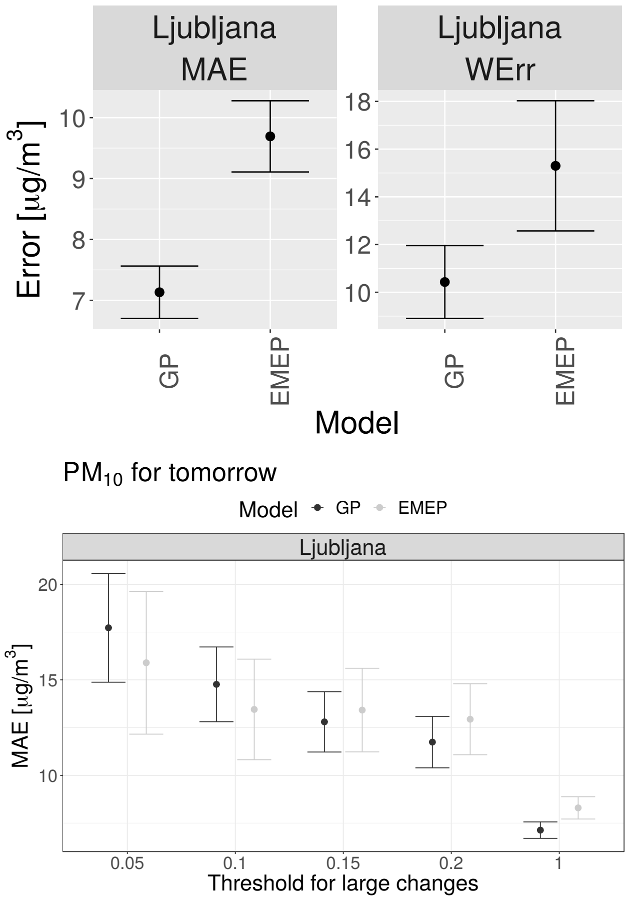
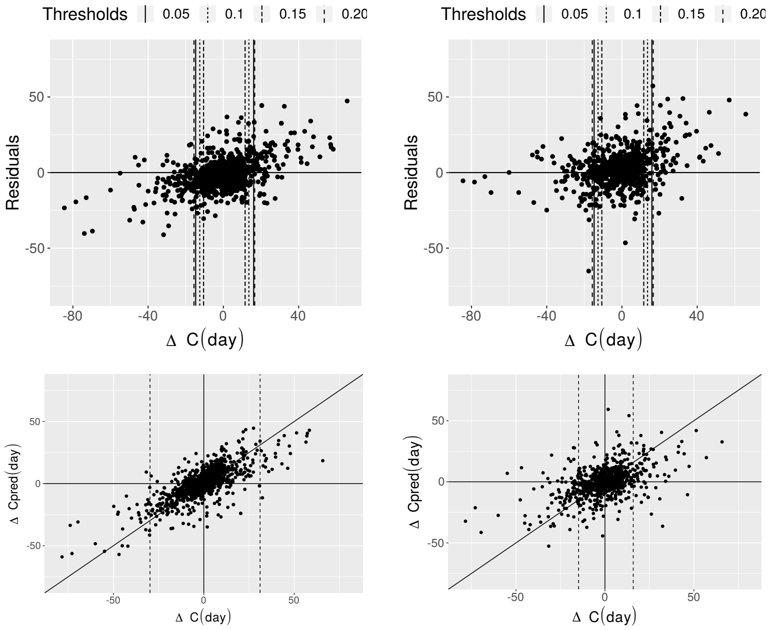


Figure 9. The top left figure shows MAE for GP and EMEP MSC-W models for PM_{10} prediction for the following day in Ljubljana, top right figure shows WErr for the same models. The lower figure shows MAE for different GP and EMEP MSC-W models for different percentages of highest values Δ C(2 day).

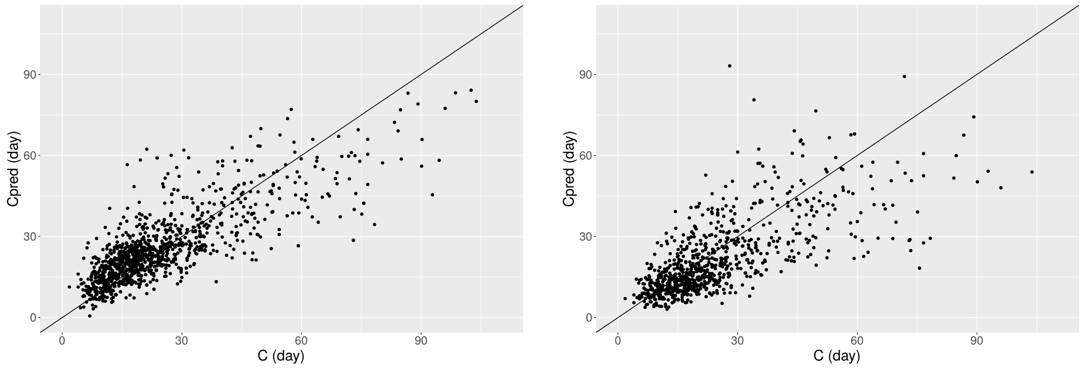


GP model EMEP MSC-W model

Figure 10. Plots on top show the residuals (GP and EMEP MSC-W models) against the observed change in PM$_{10}$ level in Ljubljana. Plots of ΔC(2 days) against ΔCpred(2 days) (GP and EMEP MSC-W models) for PM$_{10}$ level in the lower of the plots for Ljubljana.

Usual Assessment of Model Performance

We usually assess the performance of the models with the help of plots showing the measured PM$_{10}$ levels (results of different models) against the levels predicted by our models (see Figure 11). These plots can help us with the usual model assessment. In Figure 11 on the right we can observe the performance of the EMEP model. The predicted levels are further from the true levels as PM$_{10}$ levels get higher. The spread around the $y = x$ line is quite large. In Figure 11 left, we can observe the performance of the GP model for the same location (Ljubljana). Here the spread around the $y = x$ line is smaller as with the EMEP model, but it still increases with increasing PM$_{10}$ levels. The GP model generally underestimates large PM$_{10}$ levels. From these two plots, it is impossible to say anything about model predictions of different level changes.



GP model EMEP MSC-W model

Figure 11. Observed PM$_{10}$ levels from Ljubljana against predicted PM$_{10}$ levels (GP and EMEP MSC-W models) for tomorrow. Results from the same models as discussed above.

4. Conclusions

In our article, we have presented new and "cost-sensitive" measures for the performance evaluation of prediction models used for the prediction of air pollutant levels (point predictions). We have also presented two new visualizations. Our first method is based on the estimation of large pollutant level changes that occur in the time interval Δt. We use the identified large changes to evaluate the performance of prediction models only in the event large changes were observed (MAE for different percentages of largest changes). The second evaluation metric weights the model residuals with the absolute value of the observed change (WErr). We have also proposed two new visualizations, where the predicted values or model residuals are plotted against the change in pollutant level, which allows for a graphic evaluation of model performance related to the rate of the observed change in Δt. As shown in Section 3, the proposed visualizations and measures can better provide an understanding of model performance. The reanalysis of existing models shows that models usually underestimate large changes.

When examining model performance with our measures and visualizations, we showed that, when only large changes are important, the evaluation should be performed with the first measure—MAE for the selected percentage of largest changes. When large changes are as important as any other prediction, conventional measures should be used, such as MAE or MSE. When all results are important but predictions of large changes are proportionally more important, the WErr measure demonstrates a good trade-off between our first measure and conventional measures. The same measures could be used for the evaluation of meteorological models or other environmental models.

The presented measures and visualizations are useful only when dealing with models that predict some temporal variable (that changes with time) and we have a strictly defined time step; e.g., one day predictions are made for the next day. The measures are useful when large changes are more important than usual changes and are not frequent.

As properly predicting large changes is very important for relevant institutions to fulfill the requirements dictated by the European Directive 2008/50/EC [5] and to implement adequate measures, the proposed evaluation measures and visualizations of model predictions should provide further insight into model performance. In future work, we will employ the new measures and visualizations alongside the ones generally used and use them to better evaluate our air quality models.

Funding: This work was supported by the Slovenian Research Agency (ARRS Project J1-1716 and ARRS Program P2-0209).

Data Availability Statement: The code is freely available at https://gitlab.com/janafp/model_evaluation.git, accessed on 1 September 2022.

Acknowledgments: The author would like to thank the department of Air quality of the Slovenian Environment Agency (ARSO) for their data and help with the formulation of the problem.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. The Evaluated Models

L1-reguarised regression (lasso) [48], Bayesian lasso [49], random forest [47] and Gaussian processes [50] are machine learning models. The examples we analyzed above (see Section 3) were all used as regression models, but they can also be used as classification models.

L1-regularized regression also referred to by the shortened name **lasso** (least absolute shrinkage and selection operator) is a type of regularized model. In our case, we use the regularized linear regression. Regularization is a technique implemented to avoid model overfitting by adding a penalty parameter on the coefficients of the linear model. In the case of lasso the coefficients are shrunk towards zero which reduces the number of coefficients and reduces the complexity of the model and mutli-collinearity.

Bayesian lasso is a Bayesian version of the lasso regression where constraints are imposed through prior distributions. The regularization parameter is treated like a model parameter and is fit simultaneously with the coefficients. The Laplace prior provides constraints that have the same analytic form as the L1 penalty used in lasso.

The **random forest** model is an ensemble non-linear model composed of multiple decision trees. It trains different decision trees on bootstrapped samples of the training set (bagging) and a random subset of features. The final result for the regression is the average of the results of the decision trees.

Gaussian process is a Bayesian non-linear model where a Gaussian process is used as a prior probability distribution over functions. Gaussian processes are a set of random variables that have a multivariate normal distribution. Gaussian process models represent a Bayesian equivalent to artificial neural networks.

Atmospheric dispersion models are mathematical formulations of the physics and chemistry of the atmosphere [52]. They combine meteorology, pollutant emissions, atmospheric chemistry, and transport of pollutants, as well as removal processes, to predict air pollution concentrations at different locations and at different temporal resolutions. They also consider the transformation and chemical reactions of the pollutants.

Appendix B. Time Plots of Observed and Predicted Levels

Figure A1 shows time series of observed and modeled values by different models for Ljubljana.

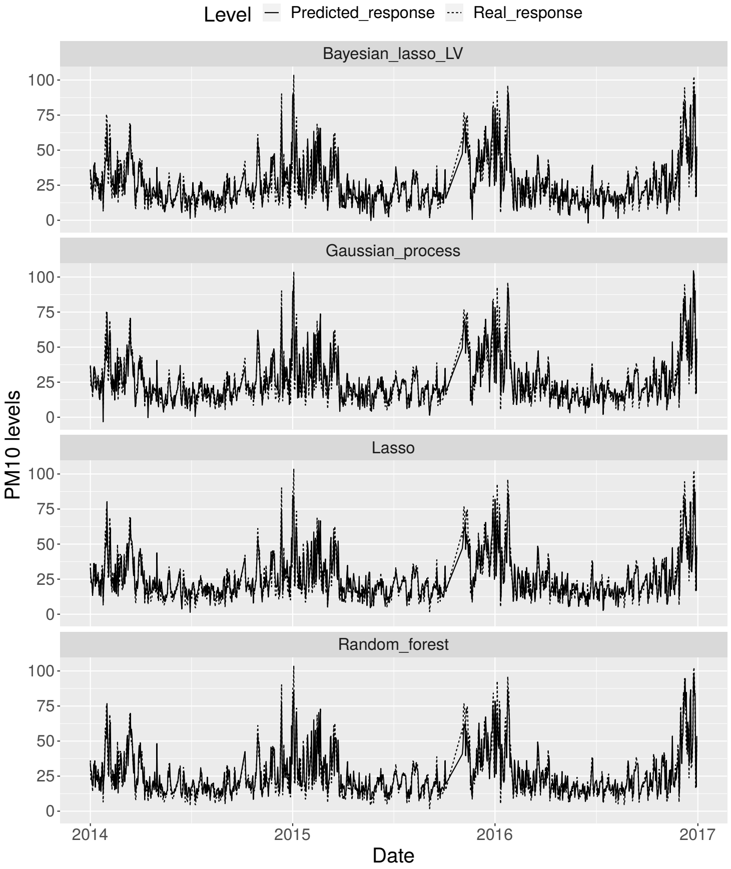


Figure A1. Observed and predicted PM_{10} levels by different models for today for Ljubljana. The dashed line represents the observed levels.

References

1. World Health Organization. *Health Aspects of Air Pollution with Particulate Matter, Ozone and Nitrogen Dioxide: Report on a WHO Working Group, Bonn, Germany, 13–15 January 2003*; WHO Regional Office for Europe: Copenhagen, Denmark, 2003.
2. Khaefi, M.; Geravandi, S.; Hassani, G.; Yari, A.; Soltani, F.; Dobaradaran, S.; Moogahi, S.; Mohammadi, M.; Mahboubi, M.; Alavi, N.; et al. Association of particulate matter impact on prevalence of chronic obstructive pulmonary disease in Ahvaz, southwest Iran during 2009–2013. *Aerosol Air Qual. Res.* **2017**, *17*, 230–237. [CrossRef]
3. Momtazan, M.; Geravandi, S.; Rastegarimehr, B.; Valipour, A.; Ranjbarzadeh, A.; Yari, A.R.; Dobaradaran, S.; Bostan, H.; Farhadi, M.; Darabi, F.; et al. An investigation of particulate matter and relevant cardiovascular risks in Abadan and Khorramshahr in 2014–2016. *Toxin Rev.* **2018**, *38*, 290–297. [CrossRef]
4. Shah, A.S.; Lee, K.K.; McAllister, D.A.; Hunter, A.; Nair, H.; Whiteley, W.; Langrish, J.; Newby, D.; Mills, N. Short term exposure to air pollution and stroke: Systematic review and meta-analysis. *BMJ* **2015**, *350*, h1295. [CrossRef] [PubMed]
5. European Council. Directive 2008/50/EC of the European Parliament and of the Council. *Decis. Counc.* **2008**, *29*, 169–212.
6. Simpson, D.; Benedictow, A.; Berge, H.; Bergström, R.; Emberson, L.D.; Fagerli, H.; Flechard, C.R.; Hayman, G.D.; Gauss, M.; Jonson, J.E.; et al. The EMEP MSC-W chemical transport model–technical description. *Atmos. Chem. Phys.* **2012**, *12*, 7825–7865. [CrossRef]
7. Baker, K.; Scheff, P. Photochemical model performance for $PM_{2.5}$ sulfate, nitrate, ammonium, and precursor species SO_2, HNO_3, and NH_3 at background monitor locations in the central and eastern United States. *Atmos. Environ.* **2007**, *41*, 6185–6195. [CrossRef]
8. Mailler, S.; Menut, L.; Khvorostyanov, D.; Valari, M.; Couvidat, F.; Siour, G.; Turquety, S.; Briant, R.; Tuccella, P.; Bessagnet, B.; et al. CHIMERE-2017: From urban to hemispheric chemistry-transport modeling. *Geosci. Model Dev.* **2017**, *10*, 2397–2423. [CrossRef]
9. Grell, G.; Peckham, S.; Schmitz, R.; McKeen, S.A.; Frost, G.; Skamarock, W.; Eder, B. Fully coupled "online" chemistry within the WRF model. *Atmos. Environ.* **2005**, *39*, 6957–6975. [CrossRef]
10. Horowitz, L.; Walters, S.; Mauzerall, D.; Emmons, L.; Rasch, P.; Granier, C.; Tie, X.; Lamarque, J.F.; Schultz, M.; Tyndall, G. A global simulation of tropospheric ozone and related tracers: Description and evaluation of MOZART, version 2. *J. Geophys. Res. Atmos.* **2003**, *108*, 4784. [CrossRef]
11. Taheri Shahraiyni, H.; Sodoudi, S. Statistical modeling approaches for PM_{10} prediction in urban areas; A review of 21st-century studies. *Atmosphere* **2016**, *7*, 15. [CrossRef]
12. Lu, W.Z.; Wang, D. Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme. *Sci. Total Environ.* **2008**, *395*, 109–116. [CrossRef] [PubMed]
13. Dutot, A.L.; Rynkiewicz, J.; Steiner, F.E.; Rude, J. A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions. *Environ. Model. Softw.* **2007**, *22*, 1261–1269. [CrossRef]
14. Faganeli Pucer, J.; Pirš, G.; Štrumbelj, E. A Bayesian approach to forecasting daily air-pollutant levels. *Knowl. Inf. Syst.* **2018**, *57*, 635–654. [CrossRef]
15. Sharma, E.; Deo, R.C.; Prasad, R.; Parisi, A. A hybrid air quality early-warning framework: An hourly forecasting model with online sequential extreme learning machines and empirical mode decomposition algorithms. *Sci. Total Environ.* **2020**, *709*, 135934. [CrossRef]
16. Kocijan, J.; Hančič, M.; Petelin, D.; Božnar, M.Z.; Mlakar, P. Regressor selection for ozone prediction. *Simul. Model. Pract. Theory* **2015**, *54*, 101–115. [CrossRef]
17. de Gennaro, G.; Trizio, L.; Di Gilio, A.; Pey, J.; Pérez, 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*, 875–883. [CrossRef]
18. Thunis, P.; Georgieva, E.; Pederzoli, A. A tool to evaluate air quality model performances in regulatory applications. *Environ. Model. Softw.* **2012**, *38*, 220–230. [CrossRef]
19. Chaloulakou, A.; Grivas, G.; Spyrellis, N. Neural network and multiple regression models for PM_{10} prediction in Athens: A comparative assessment. *J. Air Waste Manag. Assoc.* **2003**, *53*, 1183–1190. [CrossRef]
20. Lu, W.Z.; He, H.D. Comparison of three prediction strategies within $PM_{2.5}$ and PM_{10} monitoring networks. *Atmos. Pollut. Res.* **2019**, *11*, 590597.
21. Gu, J.; Yang, B.; Brauer, M.; Zhang, K. Enhancing the evaluation and interpretability of data-driven air quality models. *Atmos. Environ.* **2021**, *246*, 118125. [CrossRef]
22. Mao, W.; Wang, W.; Jiao, L.; Zhao, S.; Liu, A. Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. *Sustain. Cities Soc.* **2021**, *65*, 102567. [CrossRef]
23. Vazquez Santiago, J.; Inoue, K.; Tonokura, K. Modeling Ground Ozone Concentration Changes after Variations in Precursor Emissions and Assessing Their Benefits in the Kanto Region of Japan. *Atmosphere* **2022**, *13*, 1187. [CrossRef]
24. Gregório, J.; Gouveia-Caridade, C.; Caridade, P. Modeling PM2.5 and PM10 Using a Robust Simplified Linear Regression Machine Learning Algorithm. *Atmosphere* **2022**, *13*, 1334. [CrossRef]
25. Huang, L.; Zhu, Y.; Zhai, H.; Xue, S.; Zhu, T.; Shao, Y.; Liu, Z.; Emery, C.; Yarwood, G.; Wang, Y.; et al. Recommendations on benchmarks for numerical air quality model applications in China–Part 1: PM 2.5 and chemical species. *Atmos. Chem. Phys.* **2021**, *21*, 2725–2743. [CrossRef]

26. Zhang, Z.; Zeng, Y.; Yan, K. A hybrid deep learning technology for PM2.5 air quality forecasting. *Environ. Sci. Pollut. Res.* **2021**, *28*, 39409–39422. [CrossRef]
27. Faganeli Pucer, J.F.; Štrumbelj, E. Impact of changes in climate on air pollution in Slovenia between 2002 and 2017. *Environ. Pollut.* **2018**, *242*, 398–406. [CrossRef]
28. Cheng, C.H.; Huang, S.F.; Teoh, H.J. Predicting daily ozone concentration maxima using fuzzy time series based on a two-stage linguistic partition method. *Comput. Math. Appl.* **2011**, *62*, 2016–2028. [CrossRef]
29. Willmott, C.J.; Robeson, S.M.; Matsuura, K. A refined index of model performance. *Int. J. Clim.* **2012**, *32*, 2088–2094. [CrossRef]
30. McCuen, R.H.; Knight, Z.; Cutter, A.G. Evaluation of the Nash–Sutcliffe efficiency index. *J. Hydrol. Eng.* **2006**, *11*, 597–602. [CrossRef]
31. Legates, D.R.; McCabe, G.J. A refined index of model performance: A rejoinder. *Int. J. Clim.* **2013**, *33*, 1053–1056. [CrossRef]
32. Carslaw, D.; Ropkins, K. Openair—An R package for air quality data analysis. *Environ. Model. Softw.* **2012**, *27*, 52–61. [CrossRef]
33. Muñoz, E.; Martin, M.; Turias, I.; Jimenez-Come, M.; Trujillo, F. Prediction of PM_{10} and SO_2 exceedances to control air pollution in the Bay of Algeciras, Spain. *Stoch Environ. Res. Risk Assess.* **2014**, *28*, 1409–1420. [CrossRef]
34. Zhang, H.; Zhang, W.; Palazoglu, A.; Sun, W. Prediction of ozone levels using a Hidden Markov Model (HMM) with Gamma distribution. *Atmos. Environ.* **2012**, *62*, 64–73. [CrossRef]
35. Garrido-Perez, J.M.; García-Herrera, R.; Ordóñez, C. Assessing the value of air stagnation indices to reproduce PM10 variability in Europe. *Atmosphere* **2021**, *248*, 105258. [CrossRef]
36. Chang, J.C.; Hanna, S.R. Air quality model performance evaluation. *Meteorol. Atmos. Phys.* **2004**, *87*, 167–196. [CrossRef]
37. Taylor, K. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res. Atmos.* **2001**, *106*, 7183–7192. [CrossRef]
38. Stow, C.A.; Jolliff, J.; McGillicuddy, D.J., Jr.; Doney, S.C.; Allen, J.I.; Friedrichs, M.A.; Rose, K.A.; Wallhead, P. Skill assessment for coupled biological/physical models of marine systems. *J. Mar. Syst.* **2009**, *76*, 4–15. [CrossRef]
39. Gjerek, M.; Koleša, T.; Logar, M.; Matavž, L.; Murovec, M.; Rus, M.; Žabkar, R. *Kakovost zraka v Sloveniji v letu 2019*; Technical Report; ARSO: Ljubljana, Slovenija, 2019.
40. Horton, D.E.; Skinner, C.B.; Singh, D.; Diffenbaugh, N. Occurrence and persistence of future atmospheric stagnation events. *Nat. Clim. Chang.* **2014**, *4*, 698–703. [CrossRef]
41. Wang, X.; Dickinson, R.E.; Su, L.; Zhou, C.; Wang, K. PM2.5 pollution in China and how it has been exacerbated by terrain and meteorological conditions. *Bull. Am. Meteorol. Soc.* **2018**, *99*, 105–119. [CrossRef]
42. Krupa, S.; Manning, W.J. Atmospheric ozone: Formation and effects on vegetation. *Environ. Pollut.* **1988**, *50*, 101–137. [CrossRef]
43. Lange, K.; Little, R.; Taylor, J. Robust statistical modeling using the t distribution. *J. Am. Stat. Assoc.* **1989**, *84*, 881–896. [CrossRef]
44. Ahsanullah, M.; Golam Kibria, B.; Shakil, M. *Normal and Student's T Distributions and Their Applications*; Atlantis Press: Paris, France, 2014.
45. Delignette-Muller, M.L.; Dutang, C. fitdistrplus: An R Package for Fitting Distributions. *J. Stat. Softw.* **2015**, *64*, 1–34. [CrossRef]
46. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
47. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
48. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. [CrossRef]
49. Park, T.; Casella, G. The bayesian lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686. [CrossRef]
50. Schulz, E.; Speekenbrink, M.; Krause, A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *J. Math. Psychol.* **2018**, *85*, 1–16. [CrossRef]
51. Žabkar, R.; Rakovec, J.; Koračin, D. The roles of regional accumulation and advection of ozone during high ozone episodes in Slovenia: A WRF/Chem modelling study. *Atmos. Environ.* **2011**, *45*, 1192–1202. [CrossRef]
52. Holmes, N.; Morawska, L. A review of dispersion modelling and its application to the dispersion of particles: An overview of different dispersion models available. *Atmos. Environ.* **2006**, *40*, 5902–5928. [CrossRef]