IOŚ-PIB Environmental Protection and Natural Resources
Ochrona Środowiska i Zasobów Naturalnych
Vol. 36 No 3(105): 15-29
DOI 10.2478/oszn-2025-0018

Lech Łobocki*,**

Improving Deterministic Air Quality Forecasts Using Supervised Machine Learning: A Feasibility Study

* Warsaw University of Technology, Warsaw, Poland
** Institute of Environmental Protection – National Research Institute, Warsaw, Poland
e-mail: lech.lobocki@pw.edu.pl

Keywords:
air quality, air quality prediction, supervised machine learning, deterministic-statistical forecast

Abstract
The aim of this study is to investigate the potential, methods, and benefits of applying supervised machine-learning techniques to enhance deterministic air quality forecasts. These forecasts are produced at the Institute of Environmental Protection–National Research Institute using a numerical grid-based model that solves a system of conservation equations describing atmospheric dynamics as well as pollutant transport and transformation. Four alternative machine-learning models were tested, yielding similar results. The outcomes indicate the possibility of achieving a near-perfect forecast at the locations of measurement stations. It also turns out that if the pollutant concentration values predicted by the deterministic model are not used as features in the machine-learning model, the quality of the final forecast drops drastically.

1. INTRODUCTION

As required by the Environment Protection Act, the Institute of Environmental Protection–National Research Institute (IOŚ-PIB, hereafter IEP-NRI) carries out a daily nationwide air quality forecast, which is made available to the public by the Chief Inspectorate of Environmental Protection (GIOŚ) via an online information portal (http://powietrze.ios.gov.pl). This forecast serves as the basis for air quality warnings issued by the Government Centre for Security (RCB). The forecast is produced using a numerical grid-based model (GEM-AQ; Kaminski et al. [2008]) that solves a system of conservation equations describing atmospheric dynamics as well as the transport and transformation of pollutants. This model is also used for other purposes, including the periodic assessment of air quality in air quality management zones and analyses supporting strategic planning in the context of atmospheric pollution. These tasks are also carried out by IEP-NRI.

Air quality forecasting involves numerous sources of uncertainty, extensively documented in the scientific literature over several decades (Hanna [1993]; Rao et al. [2020]). These uncertainties occur at various stages of the forecasting process and include: inaccuracies in emission inventories (e.g., Zhang et al. [2012]; Holnicki and Nahorski [2015]; Pisoni et al. [2018]; Baklanov and Zhang [2020]) errors in meteorological inputs (e.g., Gilliam et al. [2015]); limited spatial and temporal model resolution (e.g., Tan et al. [2015]; Zhalehdoost and Taleai [2025]) simplifications in the representation of physical and chemical processes (e.g., Silveira et al. [2019]; Jiang et al. [2020]) and incomplete knowledge of initial conditions and boundary inflows (e.g., Liu et al. [2001]; Cheng and Sandu [2009]; Rittler et al. [2013]). In the context of the short-term forecasting with a grid spacing of several kilometres, the most significant limitation appears to be the lack of sufficiently accurate and reliable information on the spatial distribution of current emission intensities. The emission data used in the model are based on estimates derived from the previous year's statistical data, combined with emission factors. This raises questions about the usefulness of introducing improvements to the model that would substantially increase computational cost—potentially delaying forecast delivery and increasing energy consumption. This concern is particularly relevant to spatial resolution, as the computational cost increases approximately with the fourth power of the inverse grid spacing, assuming the domain size remains unchanged.

In light of the aforementioned challenges, various dynamical-statistical approaches are employed in operational forecasting systems. Historically, two classical statistical approaches were used to extend or improve meteorological forecasts. The first, known as Perfect Prog (PP; Klein et al. [1959]), assumed that statistical relationships between variables (e.g., pressure, temperature, precipitation) could be derived from observations and then applied to numerical model outputs, under the assumption that the model fields were perfect. In the early stages of numerical weather prediction (the 1950s and 1960s), this was justified mainly by the limited range of forecasted variables—primarily

Open Access. © Author(s). CC BY-NC-ND This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Table 1. Location of air quality monitoring stations used in this study*

Station Code	Address	Geographic Coordinates (WGS84)	
		Longitude [°E]	Latitude [°N]
PL0140A	Warszawa, Al. Niepodległości 227/233	21.004724	52.219298
PL0141A	Warszawa, ul. Wokalna 1	21.033819	52.160772
PL0143A	Warszawa, ul. Kondratowicza 8	21.042458	52.290864
PL0308A	Warszawa, ul. Tołstoja 2	20.933018	52.285073
PL0717A	Warszawa, ul. Bajkowa 17/21	21.176233	52.188474
PL0739A	Warszawa, ul. Chróścickiego 16/18	20.906073	52.207742

*Metadane oraz kody stacji i stanowisk pomiarowych. Główny Inspektorat Ochrony Środowiska, https://powietrze.gios.gov.pl/pjp/archives/downloadFile/584

dynamical parameters—so forecasting, for example, precipitation required additional techniques. Later, the model output statistics (MOS; Glahn and Lowry [1972]) technique was introduced, which established statistical relationships between forecasted and observed quantities based on sufficiently long data series. Here, a selected set of predictors—derived from deterministic model outputs—and a set of variables of interest whose past values known from observations are used. A wide range of statistical methods can be used for this purpose (e.g., Konovalov et al. [2009]; Neal et al. [2014]; Strużewska et al. [2016]; Petetin et al. [2022]). Currently, given the high effectiveness of supervised machine-learning (SML) techniques and the ease of their implementation due to the availability of open-source software libraries, it is worthwhile to explore the potential, methodology, and benefits of applying such techniques within a framework analogous to the traditional MOS approach.

Machine learning (ML) techniques have been widely applied in air quality research and forecasting, as reviewed by Rybarczyk and Zalakeviciute [2018] and Méndez et al. [2023]. In Poland, several studies have also employed ML, including those by Czernecki et al. [2021], Kujawska et al. [2022], Zareba et al. [2023], Kawka et al. [2023], and Gorzelnik et al. [2024]. While much of the existing prediction-oriented research focuses on employing ML as an alternative to deterministic models (as in the PP framework), this work demonstrates the superior accuracy of a hybrid, MOS-like approach that leverages deterministic forecasts as the primary input to a supervised ML model.

2. DATA AND ML SETUP

The ML models applied in this study require the definition of a specific set of features (predictors—in our case selected from the variables predicted by the deterministic model) and a target variable (predictand), which in this context is the mass concentration of 2.5 µm particulate matter ($PM_{2.5}$) at the location of a given monitoring station. In this study, we used hourly $PM_{2.5}$ measurements from the year 2024 collected at six automatic stations of the State Environmental Monitoring network in Warsaw (Table 1, Figure 1). Since these stations do not provide meteorological observations, all predictor variables (features) including both the $PM_{2.5}$ concentrations and surface-level meteorological parameters were taken from an archive of model outputs generated for air-quality assessment on a 0.18° grid (marked in Figure 1). The forecasts were initialized daily at 00 UTC and provided 24-hour predictions; their hourly fields were interpolated to the locations of monitoring stations to construct the data set used for training and validation of the SML model. The feature set comprised predicted values of wind speed and direction, air temperature, relative humidity, total cloud cover, precipitation rate, and $PM_{2.5}$ concentrations. All variables were extracted both at the four corners of the grid cell containing the location of a given monitoring station and as values interpolated to that specific point. To avoid issues related to the cyclic nature of wind direction, its velocity components were used instead.

Each forecast–observation pair representing a single training sample was analysed independently, focusing on the relationship between features and the observed target, without accounting for temporal dependencies. Seasonal, weekly, or daily cycles were also not explicitly included. Presumably, the deterministic model has already accounted for time-related factors to some extent; given the excellent results presented in Table 2, explicitly incorporating historical data into the predictor set would not be expected to provide any significant additional improvement.

The data set was divided into two subsets: training and testing, comprising 80% and 20% of the samples, respectively. Initially, the split was random; however, a chronological division was ultimately adopted, with the training set covering the first 80% of the samples. The differences in statistical verification metrics between the two approaches were negligible—below 10^{-3} or even 10^{-4} in relative terms.

Four SML models were tested in regression mode, using their default/recommended parameters: random forest (RF), gradient boosting (GB), extreme gradient boosting (XGBoost), and support vector regression (SVR), each briefly described below:

RF Regression (Breiman [2001]) is an ensemble learning method that builds multiple decision trees during training and outputs the average of their predictions. It reduces

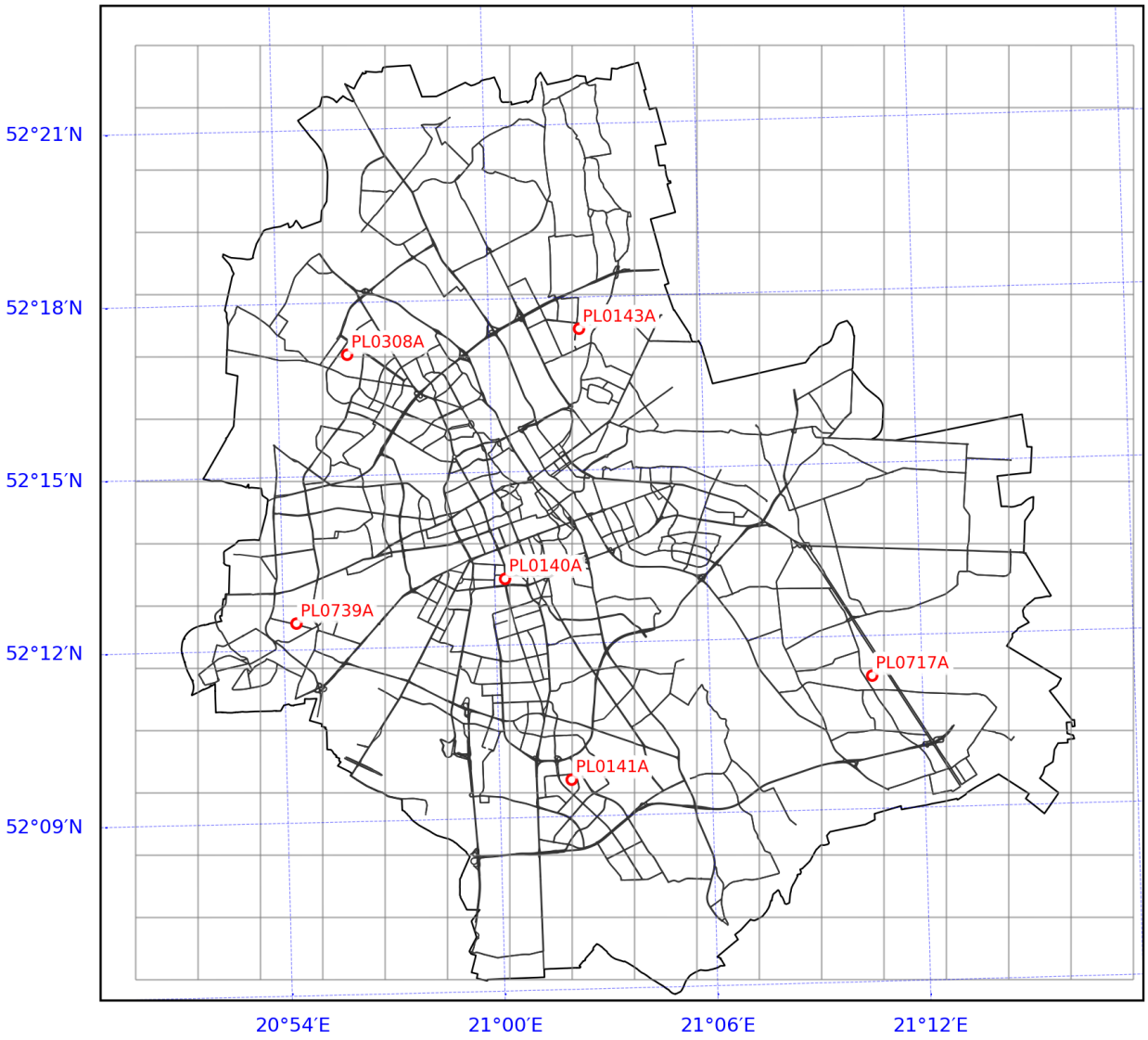

Figure 1. Location of the air-quality monitoring stations used in this study. Model grid is marked with grey lines. Station codes are explained in Table 1

Table 2. Performance metrics of the deterministic forecasts done using GEM-AQ*

Station ID	Location	RMSE	MAE	CoD (R^2)	PCC (r)
PL0140A	Al. Niepodległości	9.15	6.19	0.0585	0.5152
PL0141A	ul. Wokalna	8.39	5.86	0.0855	0.5236
PL0143A	ul. Kondratowicza	8.92	5.94	0.0325	0.5497
PL0308A	ul. Tołstoja	11.5	8.11	0.0978	0.5343
PL0717A	ul. Bajkowa	9.11	5.94	0.2719	0.6071
PL0739A	ul. Chróścickiego	8.30	5.54	0.0205	0.6077

*Values of RMSE, MAE, coefficient of determination (CoD), and PCC were computed according to Eqs. 1–4, correspondingly

overfitting by introducing randomness both in data sampling (bagging) and feature selection at each tree split. This makes the model robust to noise and capable of capturing complex, nonlinear relationships.

GB Regression (Friedman [2001]) is an ensemble technique that builds a sequence of decision trees, where each tree is trained to correct the errors made by the previous ones. The model minimizes a specified loss function using gradient descent, allowing it to capture complex patterns in the data. Although more prone to overfitting than RF, it often achieves higher predictive accuracy when properly tuned.

X*GBoost* (Chen and Guestrin [2016]) is an optimized implementation of GB that introduces system and algorithmic enhancements to improve speed and performance. Compared to standard GB, XGBoost includes features such as regularization (to reduce overfitting), parallelized tree construction, and efficient handling of missing values. As a result, it is often faster and more accurate, especially on large and complex data sets.

SVR (Smola and Schölkopf [2004]) is a regression method based on the principles of support vector machines that aims to find a function that approximates the target within a specified margin of tolerance. It uses kernel functions to model nonlinear relationships and focuses on minimizing model complexity rather than fitting all data points exactly. SVR is particularly effective for small- to medium-sized data sets with complex, high-dimensional patterns.

None of the ML models used in this study provide a simple closed-form analytical formula for prediction. Instead, they rely on algorithmic structures—such as ensembles of decision trees (in RF, GB, and XGBoost methods) or kernel-based functions involving support vectors (in SVR)—to generate outputs. While decision-tree-based models can, in principle, be translated into a series of conditional rules that resemble a decision logic, their ensemble forms (especially in RF or GB/XGBoost) result in highly complex models that are impractical to express as a single interpretable equation. As such, predictions from all tested models require access to the trained model and must be computed programmatically.

3. METRICS

To evaluate the performance of the methods under study, the following statistical metrics are used:

1. Root mean square error (RMSE) is the square root of the average squared differences between predicted and observed values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

RMSE penalizes larger errors more heavily because the errors are squared before averaging, which amplifies the impact of larger deviations.

2. Mean absolute error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

Unlike RMSE, MAE treats all errors equally, without emphasizing larger deviations.

3. Coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

R^2 indicates the proportion of variance in the observed data that is explained by the model. A value of 1 indicates perfect fit, while 0 value would indicate performance equal to that one of a naïve forecast based on a constant mean predictor. Negative values are possible, e.g., due to a large bias.

4. Pearson correlation coefficient (r)

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})\,(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{4}$$

r measures the linear correlation between predicted and observed values. The value of 1 means perfect correlation; 0, no correlation; and –1, perfect anticorrelation.

In the above formulae, y_i are the observed values, $\hat{y}_i$ are the predicted ones, overbars denote their corresponding mean values, and n is the number of observations.

4. RESULTS

The quality of the results obtained using ML models is best presented in comparison with those from the deterministic model, using the metrics discussed in Section 3. The spread of the results is illustrated in Figure 2, and the statistical indicator values are summarized in Table 2. Overall, the forecast quality can be considered limited, with coefficients of determination remaining below 0.1 at five out of six stations, indicating that the models explain only a small portion of the observed variance. At the same time, Pearson correlation coefficients (PCCs) fall within the range of 0.51 to 0.61, indicating a moderate, yet not strong, linear relationship between predictions and observations. Inspection of the scatter plots (Figure 2) further suggests a systematic underestimation of predicted values in the range of high observed $PM_{2.5}$ concentrations.

The performance statistics for the results obtained using the individual ML models are summarized in Tables 3–6. A dramatic improvement is evident—in the case of RF regression, both the PCCs and CoDs exceed 0.9999 at all but one station (where these values are still over 0.999), implying that the model has nearly reached its maximum attainable performance, without any potential for further additional enhancement.

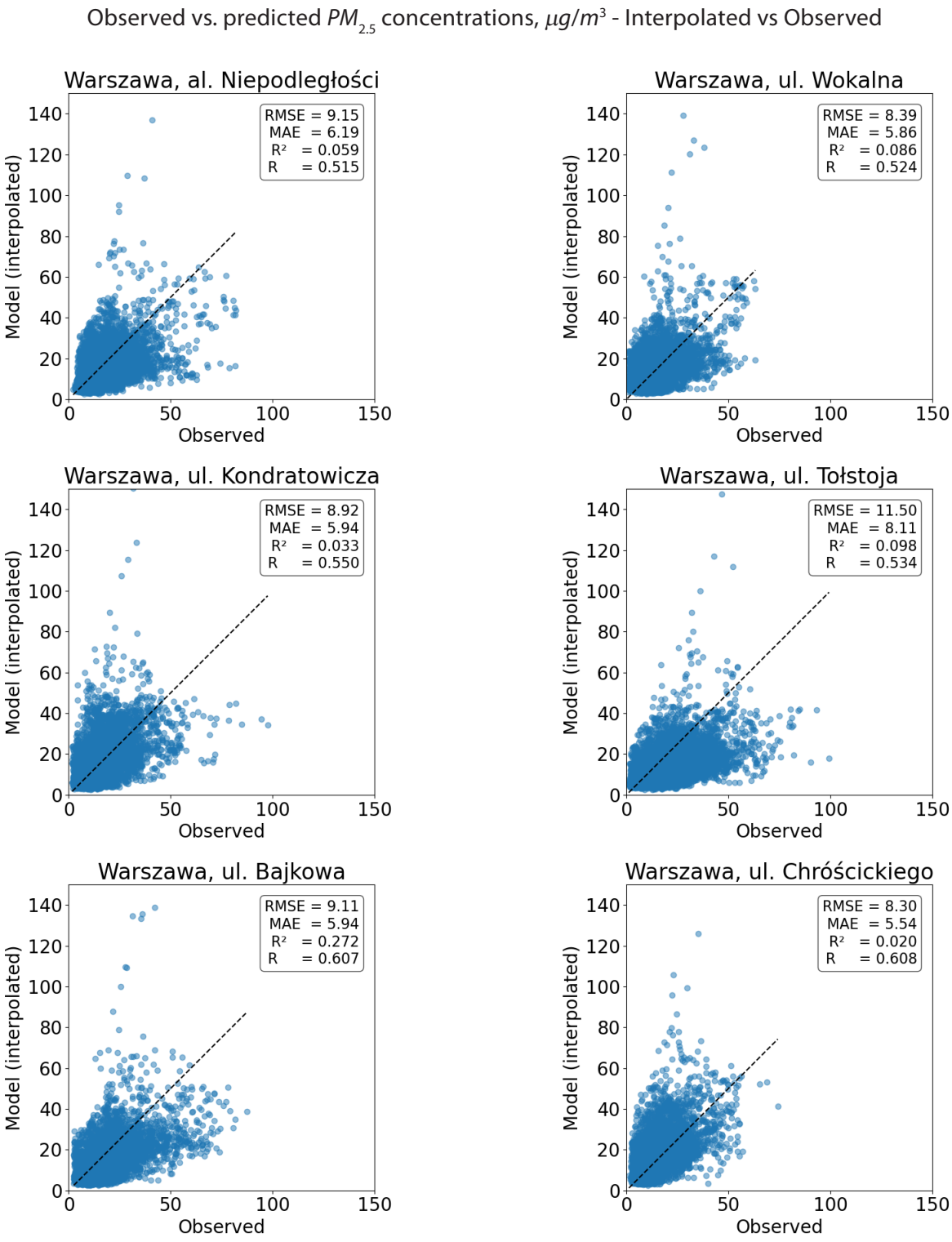

Figure 2. GEM-AQ predictions vs. observations at six air quality monitoring stations in Warsaw, during the test period in 2024

Table 3. Performance metrics of deterministic-statistical forecasts using the RF regression model*

Station ID	Location	RMSE	MAE	CoD (R^2)	PCC (r)
PL0140A	Al. Niepodległości	0.0712	0.0123	0.99995	0.99997
PL0141A	ul. Wokalna	0.0574	0.0096	0.99996	0.99998
PL0143A	ul. Kondratowicza	0.0362	0.0079	0.99998	0.99999
PL0308A	ul. Tołstoja	0.0534	0.0132	0.99998	0.99999
PL0717A	ul. Bajkowa	0.0936	0.0216	0.99995	0.99997
PL0739A	ul. Chróścickiego	0.2685	0.0199	0.99915	0.99958

*Abbreviations as in Table 2.

Table 4. Performance metrics of deterministic-statistical forecasts using the GB regression model*

Station ID	Location	RMSE	MAE	CoD (R^2)	PCC (r)
PL0140A	Al. Niepodległości	0.133	0.085	0.9998	0.9999
PL0141A	ul. Wokalna	0.115	0.069	0.9998	0.9999
PL0143A	ul. Kondratowicza	0.111	0.070	0.9999	0.9999
PL0308A	ul. Tołstoja	0.148	0.099	0.9999	0.9999
PL0717A	ul. Bajkowa	0.155	0.098	0.9999	0.9999
PL0739A	ul. Chróścickiego	0.186	0.080	0.9996	0.9998

*Abbreviations as in Table 2.

Table 5. Performance metrics of deterministic-statistical forecasts using the XGBoost regression model*

Station ID	Location	RMSE	MAE	CoD (R^2)	PCC (r)
PL0140A	Al. Niepodległości	0.432	0.137	0.9981	0.9991
PL0141A	ul. Wokalna	0.395	0.122	0.9982	0.9991
PL0143A	ul. Kondratowicza	0.472	0.117	0.9974	0.9987
PL0308A	ul. Tołstoja	0.444	0.151	0.9987	0.9994
PL0717A	ul. Bajkowa	0.951	0.295	0.9944	0.9972
PL0739A	ul. Chróścickiego	0.510	0.189	0.9969	0.9985

*Abbreviations as in Table 2.

Table 6. Performance metrics of deterministic-statistical forecasts using the SVR model*

Station ID	Location	RMSE	MAE	CoD (R^2)	PCC (r)
PL0140A	Al. Niepodległości	1.339	0.471	0.9818	0.9913
PL0141A	ul. Wokalna	1.445	0.503	0.9755	0.9885
PL0143A	ul. Kondratowicza	1.219	0.432	0.9826	0.9914
PL0308A	ul. Tołstoja	2.060	0.675	0.9727	0.9867
PL0717A	ul. Bajkowa	1.791	0.660	0.9802	0.9909
PL0739A	ul. Chróścickiego	1.614	0.525	0.9692	0.9853

*Abbreviations as in Table 2.

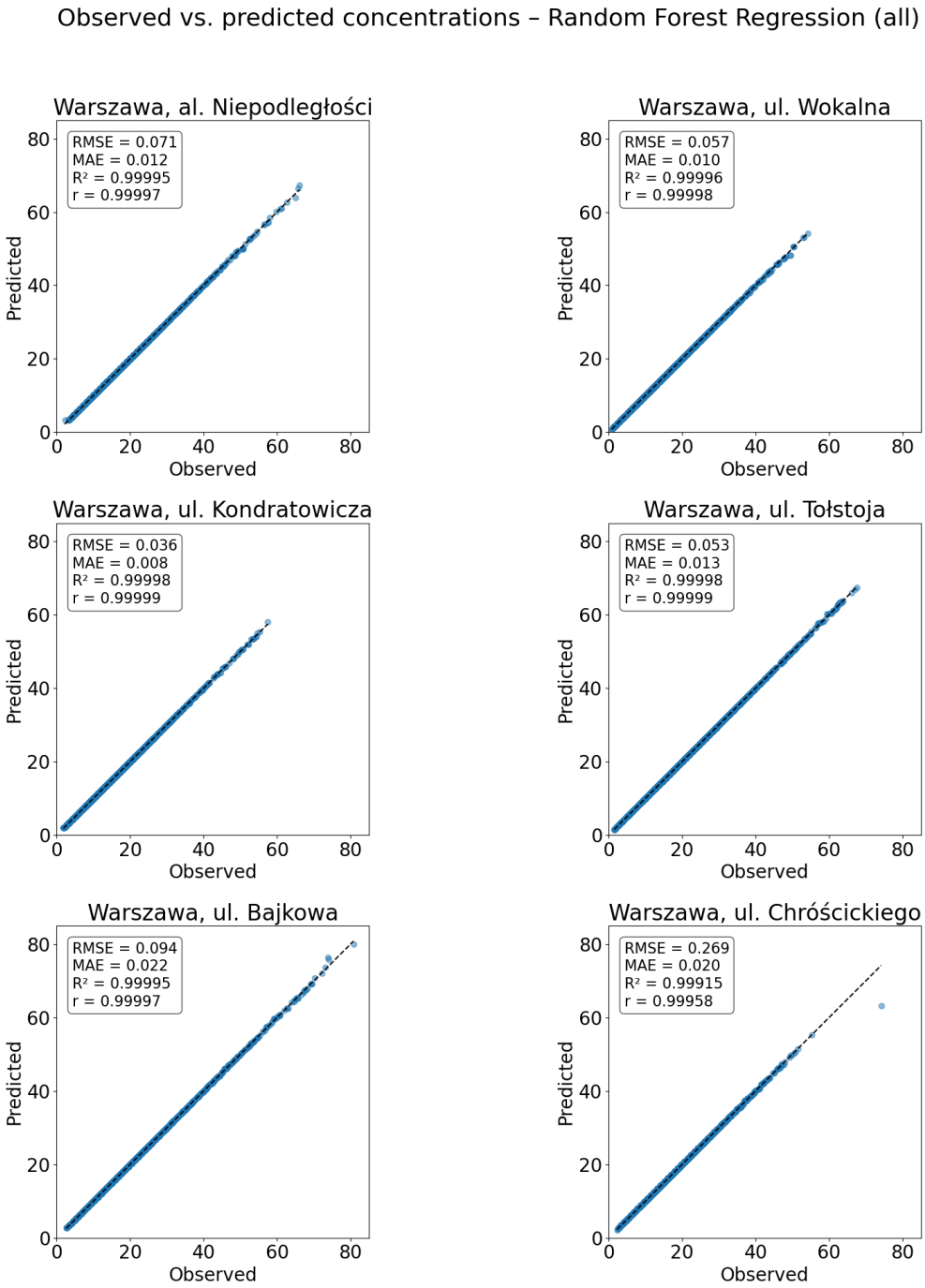

Figure 3. The hybrid deterministic-SML predictions vs. observations at six air quality monitoring stations in Warsaw, during the test period in 2024. All the features listed in Section II were used for training

Based on the comparison in Tables 3–6, RF regression yielded the most accurate results, despite all models performing at an almost perfect level. Consequently, the following analysis will focus exclusively on the RF model.

Figure 3 presents a comparison between forecasts and observations in the form of scatter plots for individual stations. While nearly all points align along the 1:1 line, a few deviations can be observed in the range of high values—most of them small, with only one notable exception. This effect is presumably due to the relatively low number of such cases in the training sample and merits attention in future, broader, and more detailed studies.

Figures 4 and 5 show histograms of residuals (errors) for the RF model and the deterministic forecast. Overall, the application of RF reduces errors by more than 2 orders of magnitude and also corrects distribution asymmetry

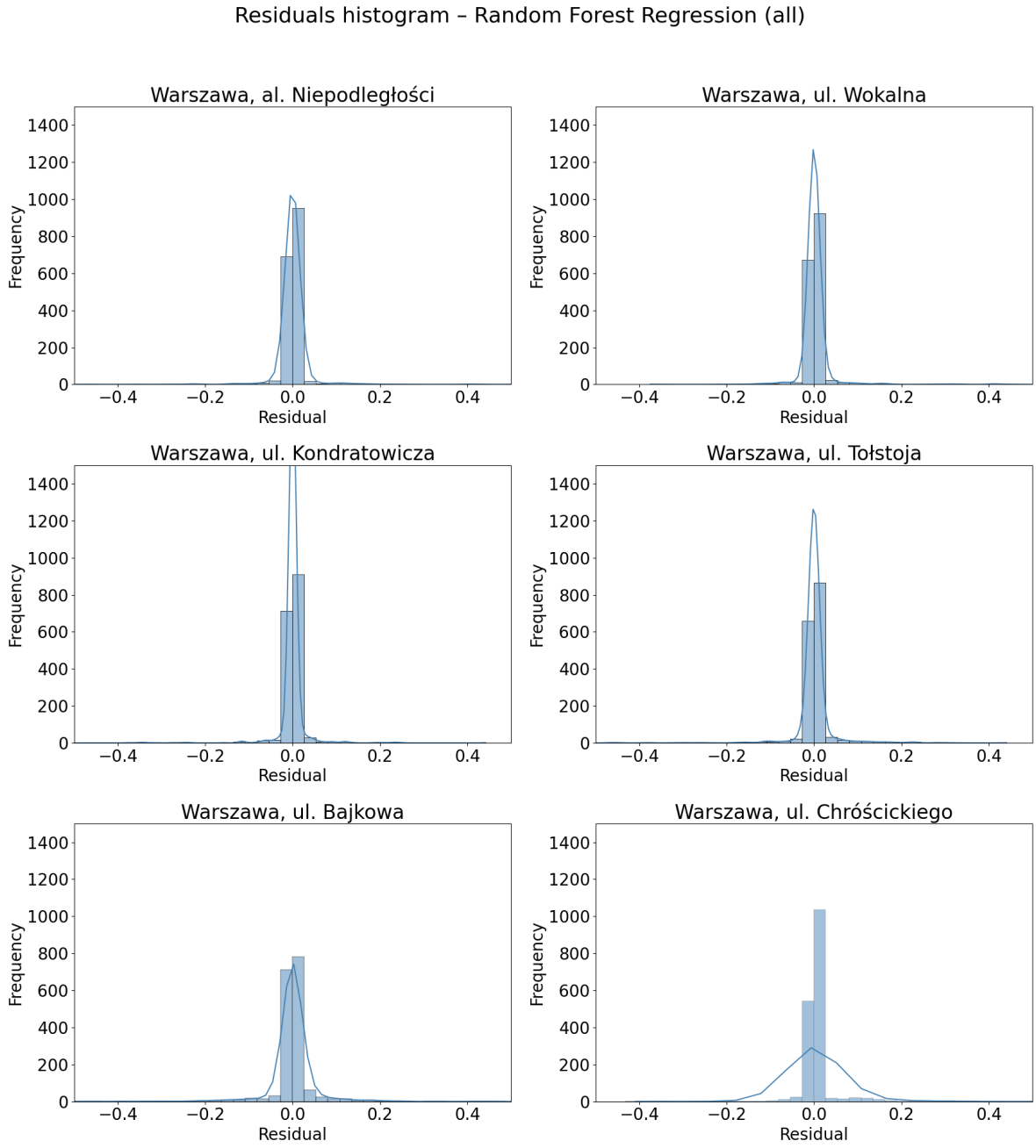

Figure 4. Residuals (errors) histograms in the hybrid deterministic-SML predictions at six air-quality monitoring stations in Warsaw during the test period in 2024

(removes bias), which is particularly evident at the stations located on Bajkowa Street and Tołstoja Street.

Q–Q plots for the RF model and the deterministic forecast are compared in Figures 6 and 7. Once again, the results from the RF model align almost perfectly along the 1:1 line. In contrast, the deterministic forecast exhibits both underestimation and overestimation of the frequency of specific value ranges. In all stations, overpredictions were observed in the upper range of values—a feature also visible in Figure 2.

Figure 8 shows the temporal variability over a continuous 300-hour period selected from the test data set. Notable differences in the temporal patterns across various locations in the city are likely related to spatial differences in local emissions (e.g., traffic, residential heating). At some stations (e.g., Chróścickiego Street), this variability is fairly well captured by the deterministic forecast, while at others (e.g., Bajkowa Street), the agreement is significantly poorer. In contrast, the forecast produced using RF (P-ML) aligns almost perfectly with the observed values.

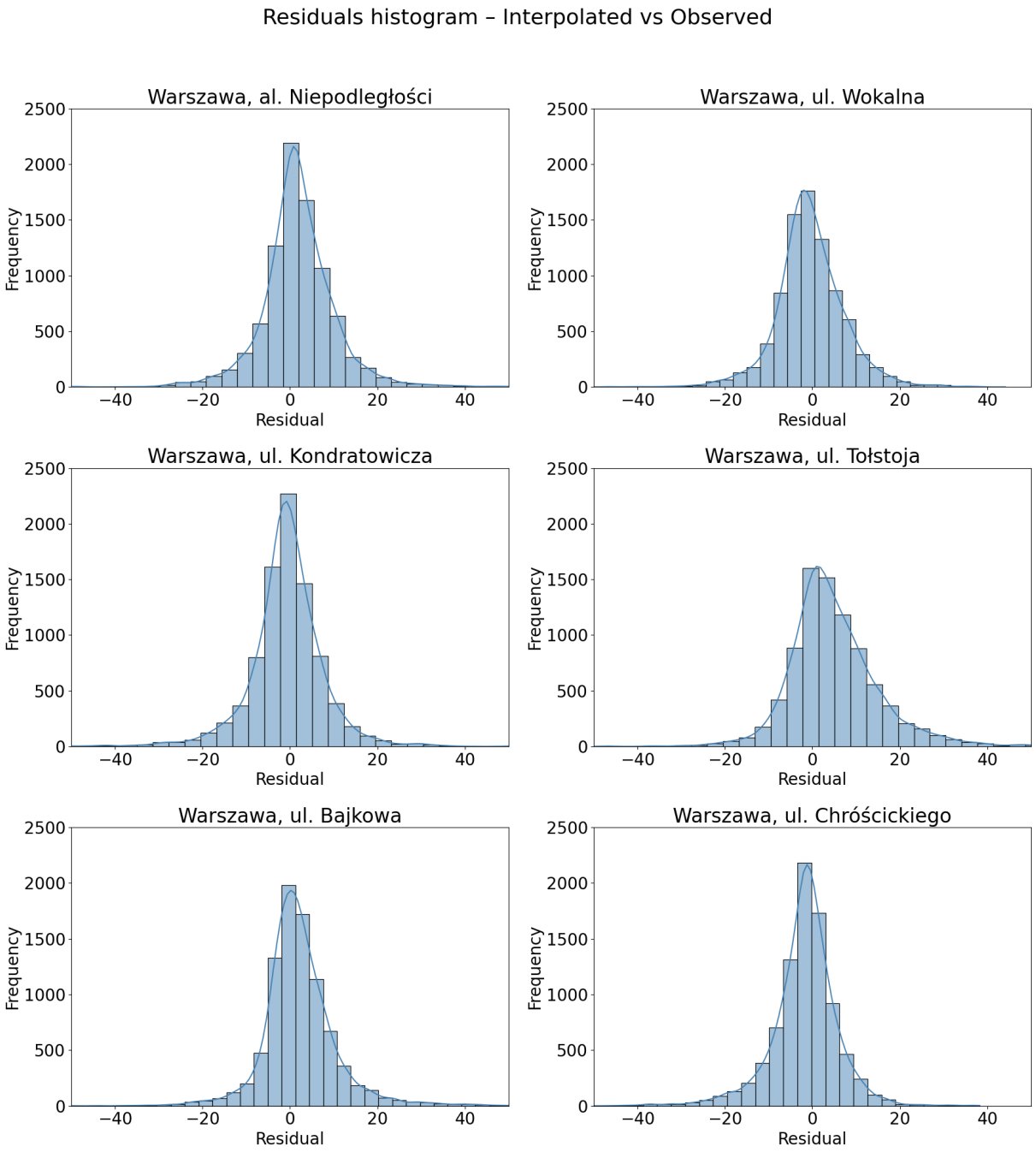

Figure 5. Residuals (errors) histograms in the deterministic (GEM-AQ) predictions at six air-quality monitoring stations in Warsaw during the test period in 2024

Given the near-perfect performance achieved by enhancing the deterministic forecast with the RF model, one might ask whether similar results could be obtained using ML alone, relying solely on meteorological forecasts. To address this question, $PM_{2.5}$ concentration values from the deterministic forecast were excluded from the input features. The resulting performance metrics are presented in Table 7. As shown, this exclusion led to a substantial decline in overall accuracy, highlighting the crucial role of the deterministic air-quality forecast in the hybrid modelling approach; this forecast, however, implicitly reflects the influence of other factors, including meteorological processes, which shape the grid-resolved concentration fields. Conversely, when meteorological variables are excluded from the input while retaining the predicted concentrations, the performance metrics

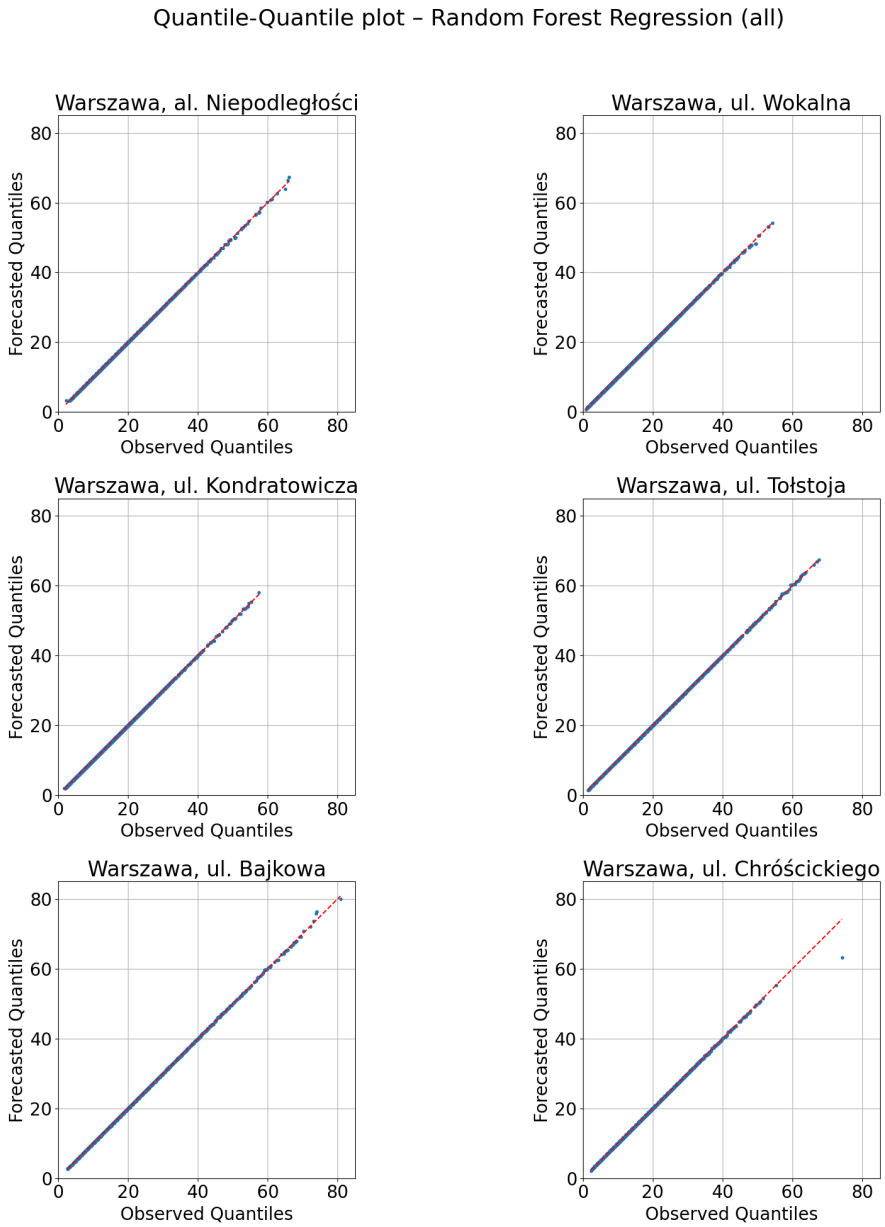

Figure 6. Quantile-Quantile comparison of distributions of the hybrid deterministic-SML predictions vs. observations at six air quality monitoring stations in Warsaw, during the test period in 2024

Table 7. Performance metrics of statistical forecasts using RF regression model, based upon the deterministic meteorological forecast (without consideration of the model-predicted $PM_{2.5}$ concentration)*

Station ID	Location	RMSE	MAE	CoD (R^2)	PCC (r)
PL0140A	Al. Niepodległości	9.12	6.70	0.154	0.461
PL0141A	ul. Wokalna	8.29	6.19	0.198	0.497
PL0143A	ul. Kondratowicza	8.91	6.40	0.070	0.471
PL0308A	ul. Tołstoja	11.64	8.59	0.129	0.464
PL0717A	ul. Bajkowa	11.39	7.90	0.198	0.547
PL0739A	ul. Chróścickiego	8.890	6.43	0.066	0.496

*Abbreviations as in Table 2

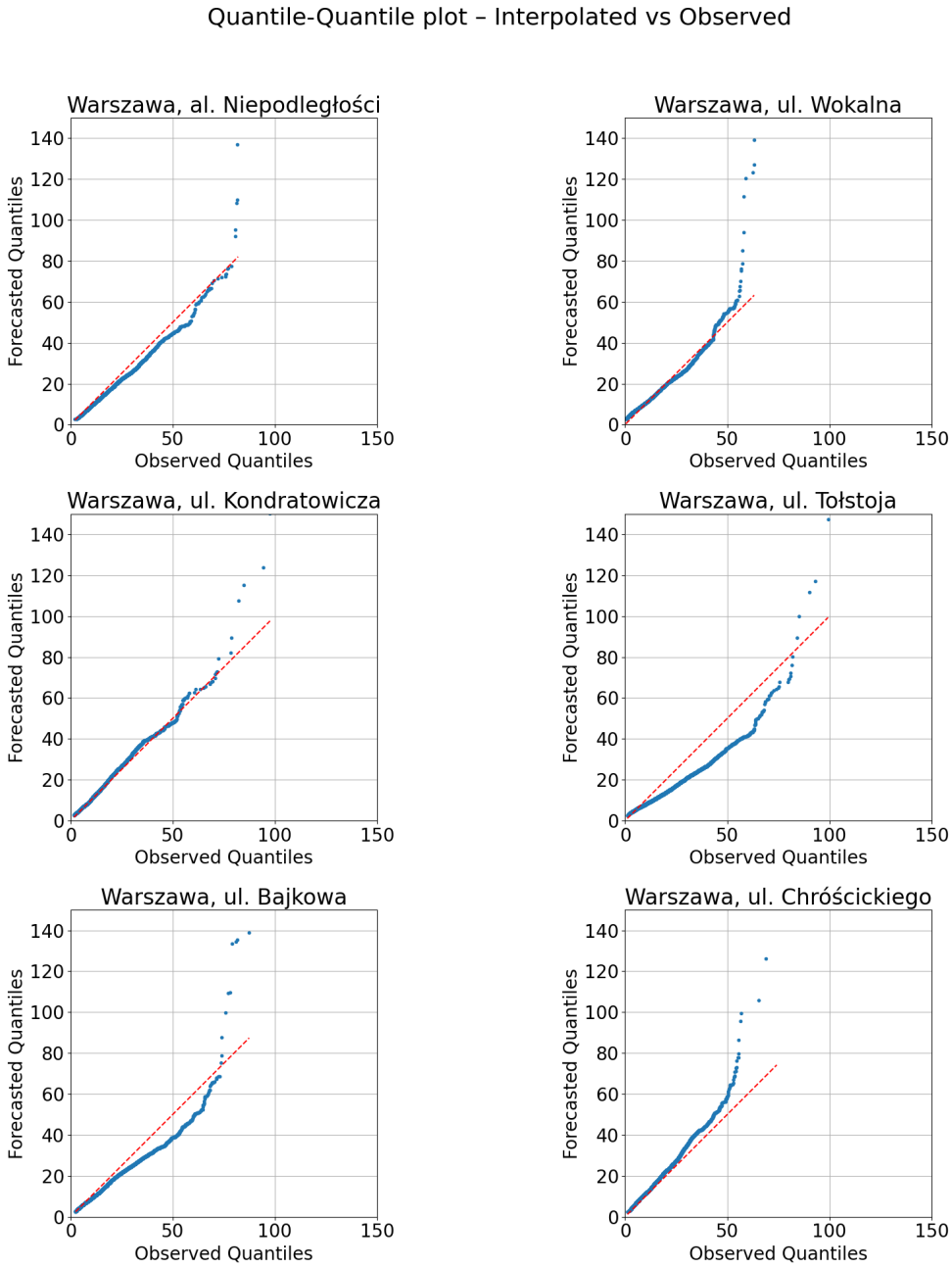

Figure 7. Quantile-Quantile comparison of distributions of the deterministic GEM-AQ predictions vs. observations at six air quality monitoring stations in Warsaw, during the test period in 2024

are roughly equal to those obtained with the full hybrid approach (Table 8). While the answer may sound negative, it should be noted that this simple test is not sufficient to address such a broadly stated question because only a limited set of meteorological variables was included in the training. Although this selection is similar to what is commonly used in other studies, several potentially important factors are only indirectly represented through the presence of the forecasted grid-resolved $PM_{2.5}$ concentration in the feature set.

5. CONCLUSIONS

The present study demonstrates the enormous potential of SML techniques when combined with deterministic forecasting. Still, the scope of this work is limited, and it should be regarded as exploratory in nature, representing an initial assessment of the hybrid approach. Nevertheless, the resulting performance metrics indicate a near-perfect level of accuracy, leaving little need—or room—for further refinement. Furthermore, the computational costs are

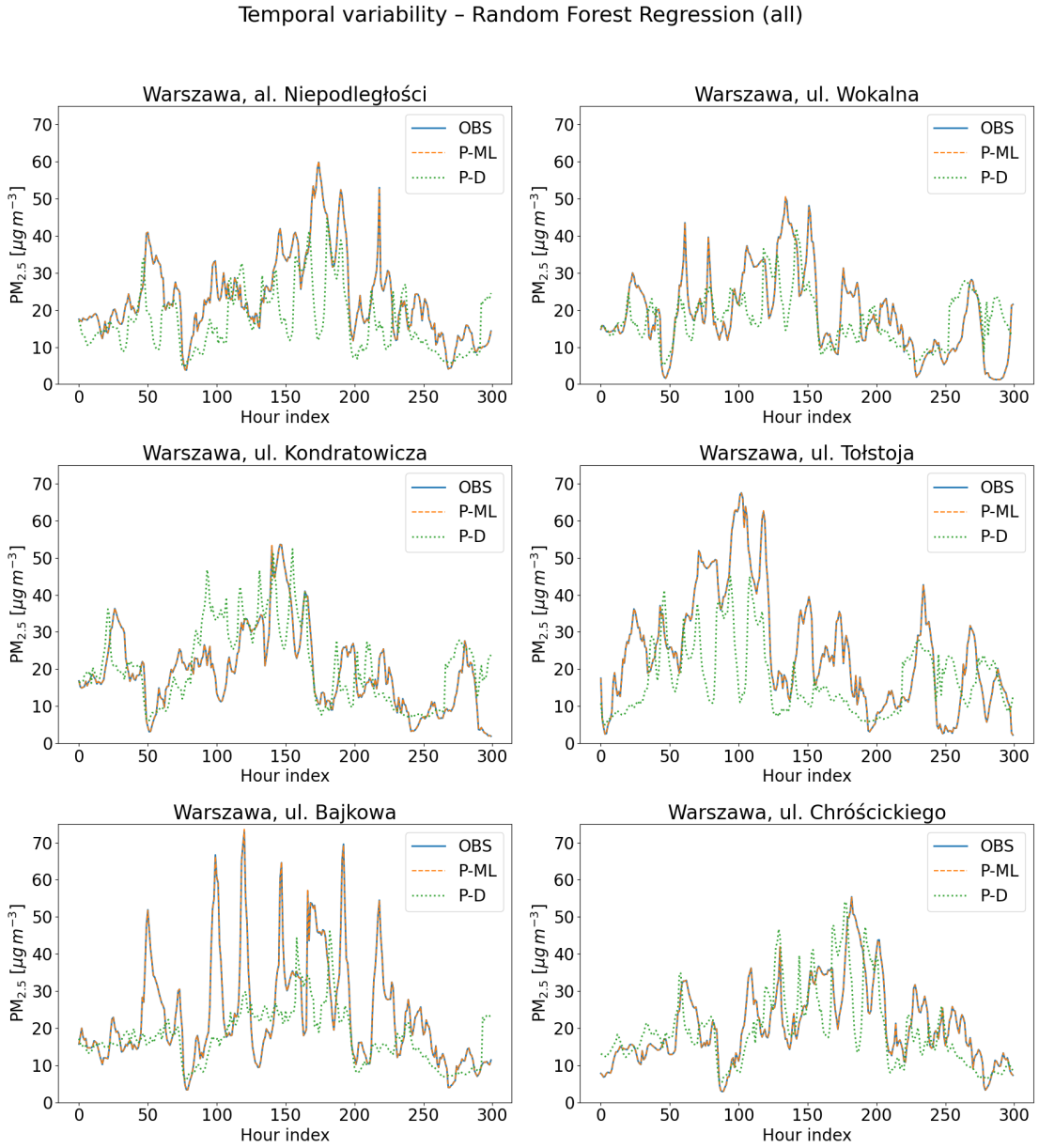

Figure 8. Observed and predicted $PM_{2.5}$ concentrations during the first 300 hours of the test period. OBS, observations; P-ML, prediction using hybrid deterministic-SML; P-D, deterministic forecast only

Table 8. Performance metrics of statistical forecasts using RF regression model, without consideration of meteorological variables in RF*

Station ID	Location	RMSE	MAE	CoD (R^2)	PCC (r)
PL0140A	Al. Niepodległości	0.0592	0.0099	0.99996	0.999982
PL0141A	ul. Wokalna	0.0397	0.0066	0.99998	0.999991
PL0143A	ul. Kondratowicza	0.0287	0.0059	0.99999	0.999995
PL0308A	ul. Tołstoja	0.0382	0.0089	0.99999	0.999995
PL0717A	ul. Bajkowa	0.0903	0.0179	0.99995	0.999975
PL0739A	ul. Chróścickiego	0.2305	0.0161	0.99937	0.999693

*Abbreviations as in Table 2

low—a single training-prediction cycle involving ~10^4 records in a data sample takes less than 20 seconds of a single CPU execution time of a modest desktop computer (equipped with 3.1GHz Intel® Core™ i5-3570S CPU). Therefore, despite the preliminary nature of this work, immediate implementation at least on a local scale is strongly recommended.

That said, several limitations and caveats warrant discussion. Despite the high performance demonstrated by the hybrid modelling approach, several limitations must be acknowledged:

Dependence on deterministic forecast quality. The SML model relies heavily on the input derived from the deterministic forecast of air pollutant concentration fields. As shown in the ablation experiments, excluding these inputs leads to a substantial drop in accuracy.

Limited generalizability. The ML model was trained and tested using data from a specific year (2024) and six monitoring stations within a single urban area (Warsaw), for a single pollutant. Its ability to generalize to other years, regions, or pollution regimes remains untested and will require confirmation within the scope of a broader study.

Limited spatio-temporal generalization. SML models are inherently restricted to the space-time domain defined by the available training data—that is, specific monitoring station locations and their historical records. Any significant, sudden, and unmodelled change in this domain (such as the construction, removal, or modification of emission sources)—if not captured by the deterministic component—may invalidate predictions. Thus, continuous performance monitoring is essential despite having the already established relationships. Moreover, unlike fixed meteorological network stations, some air-quality monitoring sites are relocated or operate only for limited periods. When a station is commissioned at a new location, SML-based forecasting becomes feasible only after a sufficient amount of local training data has been collected.

Limited operational usefulness. Although the hybrid approach demonstrates excellent predictive accuracy for short-term concentrations at monitoring sites, its practical utility remains limited to real-time applications—such as public warning systems or immediate emission control responses. It is not suitable for *post hoc* air-quality assessments, where the predicted variable is already known at the time of evaluation, rendering ML unnecessary in this context. Furthermore, in scenario-based analyses of air-pollution control strategies, changes in emission patterns—whether due to policy, infrastructure, or socioeconomic shifts—can be explicitly accounted for in the deterministic component. However, such changes are unlikely to be correctly reflected in the ML model, which may distort or ignore their impact if the new conditions fall outside the scope of its training data.

Interpretability and operational integration. While ensemble-based models such as RF offer high predictive power, they lack transparency and reproducibility compared to simpler statistical models. This may complicate integration into existing operational workflows and reduce trust among stakeholders unfamiliar with ML techniques.

Considering the above, further extended testing is necessary. Future research should address the following topics:

- **Pollutants other than $PM_{2.5}$.** Since different pollutants follow distinct emission patterns and undergo different transport and transformation processes in the atmosphere, dedicated training–testing cycles will likely be required for at least groups of pollutants, and possibly for individual species.
- **Identification of the set of features** necessary to ensure high-quality forecasts. This may become important in full-scale applications, where computational efficiency matters. A single train–test cycle with the methods used here does not provide information on the importance of individual predictors. This question can be addressed using statistical analyses, unsupervised ML techniques, or by excluding specific variables from the training set with the help of a RF, which in this study was done by removing the predicted $PM_{2.5}$ concentrations. The situation is complicated by interdependencies among variables and by their possible dependence on factors not included in the chosen feature set. In this particular case, the predicted $PM_{2.5}$ field likely acts as a proxy for several meteorological and emission-related factors not explicitly represented in the feature set, including vertical mixing, grid-resolved transport, dry deposition, wash-out and rain-out processes, transformation of aerosol, as well as the diurnal, weekly, and annual variability of emissions assumed in the model.
- **Other regions,** with differing emission characteristics, topography, land use, and local climate conditions, may show distinct generalization properties. In the present study, conducted for an urban area, substantial spatial gradients in emissions and pollutant concentrations are expected due to the complex structure of sources and surface conditions. In rural areas, variability is generally lower; however, monitoring stations are usually spaced much farther apart.
- **Sensitivity** of the model performance to the length of the training data set, in order to determine whether the amount of data used for training is optimal.
- **Potential integration with data assimilation schemes** to help mitigate the spatio-temporal limitations of SML-based forecasting.
- **Combining classification and regression SML approaches.** While regression SML provides numerical predictions that can be interpreted as expected values of the target variable, classification SML can estimate the probability of exceeding predefined thresholds, thus offering an objective basis for decision making.
- **Focus on outliers and air-pollution episodes.** A single year of data may be insufficient for reliably identifying extreme states. Since public warning is a primary goal in this context, detailed evaluation of such cases is of the utmost importance.

ACKNOWLEDGEMENTS

The data used in this study originate from the CIEP archive (accessed via its public portal), the OpenStreetMap database, and the IEP-NRI air quality archives. Computations were performed on a Linux Fedora 42 platform using Python 3.13 and a range of software libraries and tools, including Scikit-Learn, XGBoost, RPNPy, NumPy, Pandas, Seaborn, Matplotlib, SciPy, CartoPy, OSMnx, Jupyter, as well as OpenAI's LLM models, which were employed to accelerate coding and enhance text quality. The author is indebted to Dr. Jacek Kamiński and Dr. Joanna Strużewska for their valuable comments and discussions during the course of this work. This research was funded by the research potential support fund of the Ministry of Science and Higher Education, Poland, granted to the Warsaw University of Technology.

REFERENCES

BAKLANOV A., ZHANG Y. 2020. Advances in air quality modeling and forecasting. Global Transitions 2: 261–270.

BREIMAN L. 2001. Random forests. Machine Learning 45, 1: 5–32.

CHEN T., GUESTRIN C. 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794, doi: 10.1145/2939672.29397.

CHENG H., SANDU A. 2009. Uncertainty quantification and apportionment in air quality models using the polynomial chaos method. Environmental Modelling & Software 24, 917–925.

CZERNECKI B., MAROSZ M., JĘDRUSZKIEWICZ J. 2021. Assessment of machine learning algorithms in short-term forecasting of PM10 and PM2.5: Concentrations in selected Polish agglomerations. Aerosol and Air Quality Research 21, 7: 200586.

FRIEDMAN J. 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics 29, 5: 1189–1232.

GILLIAM R.C., HOGREFE C., GODOWITCH J.M., NAPELENOK S., MATHUR R., RAO S.T. 2015. Impact of inherent meteorology uncertainty on air quality model predictions. Journal of Geophysical Research: Atmospheres 12, 259–280.

GLAHN H.R., LOWRY D.A. 1972. The use of model output statistics in objective weather forecasting, Journal of Applied Meteorology and Climatology 11, 8: 1203–1211.

GORZELNIK T., BOGACKI M., OLENIACZ R. 2024. Identification of factors influencing episodes of high PM10 concentrations in the air in Krakow (Poland) using random forest method. Sustainability 16, 20: 9015.

HANNA S.R. 1993. Uncertainties in air quality model predictions. Boundary-Layer Meteorology 62, 3–20.

HOLNICKI P., NAHORSKI Z. 2015. Emission data uncertainty in urban air quality modeling: Case study. Environmental Modeling & Assessment 20: 583–597.

JIANG L., BESSAGNET B., MELEUX F., TOGNET F., COUVIDAT F. 2020. Impact of physics parameterizations on high-resolution air quality simulations over the Paris region. Atmosphere 11: 618.

KAMINSKI J.W., NEARY L., STRUZEWSKA J., MCCONNELL J.C., LUPU A., JAROSZ J., TOYOTA K., GONG S.L., COTE J., LIU X., CHANCE K., RICHTER A. 2008. GEM-AQ, an on-line global multiscale chemical weather modelling system: Model description and evaluation of gas phase chemistry processes. Atmospheric Chemistry and Physics 8, 12: 3255–3281.

KAWKA M., STRUZEWSKA J., KAMINSKI J.W. 2023. Downscaling of regional air quality model using Gaussian plume model and random forest regression. Atmosphere 14, 7: 1171.

KLEIN W.H., LEWIS B., ENGER I. 1959. Objective prediction of five-day mean temperatures during winter. Journal of the Atmospheric Sciences 16, 6: 672–682.

KONOVALOV I.B., BEEKMANN M., MELEUX F., DUTOT A., FORET G. 2009. Combining deterministic and statistical approaches for PM_{10} forecasting in Europe. Atmospheric Environment 43: 6425–6434.

KUJAWSKA J., KULISZ M., OLESZCZUK P., CEL W. 2022. Machine learning methods to forecast the concentration of PM10 in Lublin, Poland. Energies 15, 17: 6428.

LIU T.H., JENG F-T., HUANG H.-C., BERGE E., CHANG J.S., 2001. Influences of initial conditions and boundary conditions on regional and urban scale Eulerian air quality transport model simulations. Chemosphere: Global Change Science 3: 175–183.

MÉNDEZ M., MERAYO M.G., NÚÑEZ M. 2023. Machine learning algorithms to forecast air quality: A survey. Artificial Intelligence Review 56, 9: 10031–10066.

NEAL L.S., AGNEW P., MOSELEY S., ORDÓÑEZ C., SAVAGE N.H. TILBEE M. 2014. Application of a statistical post-processing technique to a gridded, operational, air quality forecast. Atmospheric Environment, 98: 385–393.

PETETIN H., BOWDALO D., BRETONNIÈRE P.A., GUEVARA M., JORBA O., ARMENGO J.M., SAMSO CABRE M., SERRADELL K., SORET A., GARCIA-PANDO C.P., 2022. Model output statistics (MOS) applied to Copernicus Atmospheric Monitoring Service (CAMS) O_3 forecasts: Trade-offs between continuous and categorical skill scores. Atmospheric Chemistry and Physics 22: 11603–11630.

PISONI E., ALBRECHT D., MARAB T.A., ROSATI R., TARANTOLA S., THUNIS P. 2018. Application of uncertainty and sensitivity analysis to the air quality SHERPA modelling tool. Atmospheric Environment 183: 84–93.

RAO S.T., LUO H., ASTITHA M., HOGREFE C., GARCIA V., MATHUR R. 2020. On the limit to the accuracy of regional-scale air quality models. Atmospheric Chemistry and Physics 20: 1627–1639.

RITTER M., MÜLLER M.D., JORBA O., PARLOW E., SALLY LIU, L.-J. 2012. Impact of chemical and meteorological boundary and initial conditions on air quality modeling: WRF-Chem sensitivity evaluation for a European domain. Meteorology and Atmospheric Physics 119: 59–70.

RYBARCZYK Y., ZALAKEVICIUTE R. (2018). Machine learning approaches for outdoor air quality modelling: a systematic review. Applied Sciences 8: 2570.

SILVEIRA C., FERREIRA J., MIRANDA A.I. 2019. The challenges of air quality modelling when crossing multiple spatial scales. Air Quality, Atmosphere & Health 12: 1003–1017.

SMOLA A.J., SCHÖLKOPF B. 2004. A tutorial on support vector regression. Statistics and Computing 14, 3: 199–222.

STRUZEWSKA J., KAMINSKI J.W., JEFIMOW M., 2016. Application of model output statistics to the GEM-AQ high resolution air quality forecast. Atmospheric Research 181: 186–199.

TAN J., ZHANG Y., MA W., YU Q., CHEN L. 2015. Impact of spatial resolution on air quality simulation: A case study in a highly industrialized area in Shanghai, China. Atmospheric Pollution Research 6: 322–333.

ZAREBA M., DLUGOSZ H., DANEK T., WEGLINSKA E. 2023. Big-data-driven machine learning for enhancing spatiotemporal air pollution pattern analysis. Atmosphere 14, 4: 760.

ZHALEHDOOST A., AND TALEAI M. 2025. Unravelling the importance of spatial and temporal resolutions in modeling urban air pollution using a machine learning approach. Scientific Reports 15: 27708.

ZHANG Y., SEIGNEUR C., BOCQUET M., MALLET V., BAKLANOV A., 2012. Real-time air quality forecasting. Part I. history, techniques, and current status. Atmospheric Environment 60: 632–655.