Bulgarian Journal of Meteorology and Hydrology

Национален институт по метеорология и хидрология

National Institute of Meteorology and Hydrology

Bul. J. Meteo & Hydro 29 /2 (2025)

Development of hourly PM10 stochastic models at Pavlovo station in Sofia

Neyko M. Neykov, Nadya Neykova, Hristomir Branzov, Valery Nikolov

National Institute of Meteorology and Hydrology,
Tsarigradsko shose 66, 1784 Sofia, Bulgaria

Abstract: Air quality management in Sofia city relies heavily on time series data obtained from air monitoring stations as a basis for determining population exposure to airborne pollutants. Hourly air pollution concentrations data of particulate matter (PM10) have been collected at six monitoring stations of the Executive Environment Agency in Sofia city for the period 01.01.2014-31.12.2025. Accurate forecasting of concentrations of the pollutants will allow early warning procedures to be improved, which is useful for safety reasons and enables, for example, traffic restrictions or a decision to make public transport free. Therefore, the development of forecasting air quality tool is an important issue for public authorities. This paper addresses several methodological issues, including deterministic forecasts derived from meteorological models and approaches to forecasting air quality up to 72 hours in advance. The focus is on the joint use of the deterministic Weather Research and Forecasting (WRF) model, which characterizes the atmospheric conditions in Sofia from ground surface up to 5500 m height, and delivers meteorological derivatives serving as input into a stochastic model. A stochastic model based on a mixture of dynamic time series regressions with a log-normal distribution is considered and discussed. The model is fitted using data from 2014–2024, and forecast accuracy is evaluated on 2025 test data following a rolling forecasting origin cross-validation procedure. Standard forecast accuracy metrics such as mean absolute error, mean squared error, root mean square error and correlation coefficients between observed and forecasted values are discussed.

Keywords: air pollution, forecasting of PM10 concentrations, mixture of time series regressions, WRF model.

1. INTRODUCTION TO AIR POLLUTION MODELS

Air pollution is of growing health concern in many urban areas around the world. One of the pollutants of greater concern is PM10 (particulate matter). The main sources of anthropogenic emissions of PM10 are residential heating, motor vehicles and industrial activities, particularly in urban areas such as Sofia, the capital of Bulgaria. Exposure to high levels of PM10 poses significant risks to public health. Therefore, forecasting hourly PM10 concentrations and providing early warnings are of great importance, as they allow the general public to be informed in a timely manner to adjust their mobility plans accordingly.

Only few papers related to the subject of our study will be considered because the forecasting pollution technology is well known. It is a synthesis between weather prediction models and stochastic models. Numerical prediction models provide the weather forecast a few days ahead whereas the pollution forecast is made by means of the stochastic models trained on the past pollution and weather forecast data. More accurate weather forecasts and higher-quality air pollution data lead to more reliable air pollution forecasts. Standard stochastic model techniques that have been used are based on the ordinary methodologies: 1) multiple linear regression; 2) principal component regression; 3) nonparametric regression models using splines, kernels or wavelets to account nonlinearities of the predictors; 4) time series autoregression models of types ARX, ARMAX; 5) generalized linear models and generalized additive models in order to handle non Gaussian distributed data; 6) various neural network types. Many of these models are well accepted and validated in practice for nearly stationary and homogeneous air pollution data. However, in case of highly heterogeneous nonstationary data these models are not adequate. In order to overcome this deficiency, the so called stochastic ensemble models techniques have been developed. Details can be found in McLachlan and Peel (2000), Seni and Elder (2010), and Zhou (2012). Mixture-of-experts is an ensemble learning methodology developed in the field of statistics and neural networks by Carvalho and Tanner (2005), Brownlee (2021). This methodology explicitly addresses the predictive modeling problem by considering multiple models that adapt locally to data subsamples. As a result, nonlinearities and other data complexities can be effectively accounted for.

Following Shahraiyni and Sodoudi's (2016) extensive review of PM10 modeling, additional contributions are mentioned here for the sake of completeness. For instance, Vlachogianni et al. (2011) developed forecast models based on stepwise multiple linear regression (MLR) for Athens and Helsinki to forecast the maximum hourly and the daily average PM10 concentrations for the next day. Russo et al. (2015) present a simple neural network and data preselection framework, discriminating the most essential input data for accurately forecasting the concentrations of PM10, based on five years observations in the metropolitan region of Lisbon, Portugal. The daily values of PM10 concentrations measured at 12 monitoring stations in the agglomeration of Lisbon between 2002 and 2006, were considered. Bertaccini et al. (2012) consider vehicular traffic as important predictor in air-quality forecasting models. These authors develop Generalized Additive Models (GAMs), (Wood, 2017) to analyze the behavior of PM10 concentrations, collected at the environmental monitoring stations distributed throughout the city of Turin, Italy, for the period 2003-2005. Although GAMs are highly flexible, the authors reported higher correlations between forecasted and observed values when the models were developed separately for each season. Cusano et al. (2025) developed models for hourly PM10 concentrations in Terni, Umbria, by integrating GAMs, neural networks, and weather prediction models using data from the city's monitoring network. Misiti et al. (2015) investigated the capabilities of mixture of linear regression models for statistically forecasting daily mean PM10 concentrations, with regards to the influence of synoptic-scale circulation patterns. Several meteorological variables as the daily mean temperature, atmospheric pressure and wind speed, maximum gradient of temperature and the average concentration measured on the previous day are included as predictors. The evaluation of these models is based on the forecast accuracy for several horizons starting from some hours to a day ahead. Garcia-Nieto et al. (2014) build PM10 nonparametric regression model based on multivariate adaptive regression splines (MARS), a particular type of GAMs, using 3 years (2006-2008) dataset of nitrogen oxides, carbon monoxide, sulfur dioxide and ozone in order to obtain a preliminary estimate of the dependence between pollutants in the Oviedo urban area. Lots of machine learning methods, including gradient boosting machine (GBM), are widely used for modeling hourly PM2.5 and PM10 concentrations, e.g. Park et al. (2021) and Wallek et al. (2024).

The aim of this paper is to present several stochastic models for forecasting hourly PM10 concentrations up to 72 hours ahead. The analysis uses data from the ExEA monitoring station located in Pavlovo, Sofia. The proposed models include GAMs, gradient boosting machines (GBMs), multivariate adaptive regression splines (MARS), and mixture-of-experts (MoE) approaches. These modeling techniques are widely applied in PM10 studies due to their ability to capture complex nonlinear relationships between air pollutant concentrations and meteorological predictors. Consequently, they often outperform traditional statistical techniques in forecasting PM10 levels and in high-resolution spatial mapping. A key advantage of MoE models lies not only in their flexibility and capacity to model nonlinearities, but also in their ability to adapt to data heterogeneity arising from varying meteorological conditions. As input predictors, we use meteorological variables derived from the Weather Research and Forecasting (WRF) model that characterize atmospheric conditions over Pavlovo from the surface up to 5,500 m altitude, along with their lagged values, lagged PM10 concentrations spanning 72 to 114 hours, and selected interaction terms. The PM10 response variable is modeled assuming a log-normal distribution.

2. DATA DESCRIPTION AND EXPLORATORY DATA ANALYSIS

The data set consists of both PM10 concentrations and meteorological variables for the period 01.01.2014-28.12.2025. Air pollutant include hourly average concentrations of PM10 measured at Pavlovo station. The meteorological data are based on the numerical weather prediction model WRF output.

To examine the PM10 distribution at Pavlovo station, Figure 1 presents boxplots of hourly average concentrations across different time scales: years, months, hours of the day, and days of the week. From these plots one can see that PM10 concentrations are higher during the cold half of the year and exhibit strong seasonal and diurnal variations. It is seen that PM10 values during the week days are higher than those measured during the weekends. The histograms of PM10 values presented in Figure 2 (right) are right-skewed; while the logarithmic transformation is more symmetric, the distribution remains far from normal.

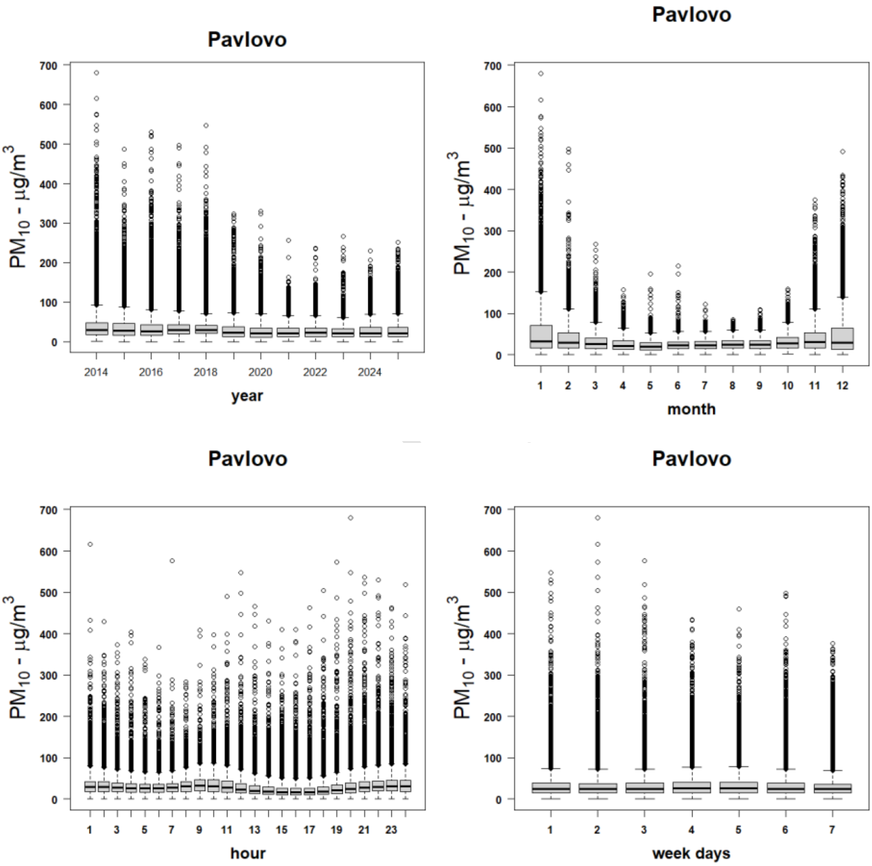

Fig. 1. Distributions of PM10 values at Pavlovo station - yearly, monthly, diurnal and week days data for the period 01.01.2014-28.12.2025

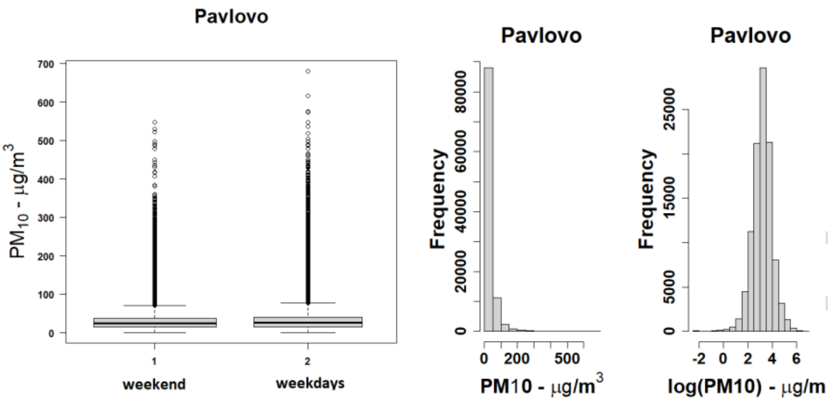

Fig. 2. Pavlovo station PM10 data series for the period 01.01.2014-28.12.2025: (left) distributions of PM10 values at weekends and week days; (right) histograms of PM10 and their logarithmic transformed values

3. MIXTURES-OF-EXPERTS TIME SERIES REGRESSION MODELS

A great deal of research on nonlinear time series regression models has been done during the last 30 years. Among many models proposed in the literature, one can find an important class of finite mixture of distributions denoted as mixtures-of-experts (MoE). This class of models has been used in many different areas in order to account for nonlinearities and other complexities such as heterogeneity in the data. Details can be found in McLachlan and Peel (2001), Carvalho and Tanner (2005).

Let $y_t = \log(\mathrm{PM10}_t)$ be the logarithm of the hourly average values of $\mathrm{PM10}_t$ for $t = 1, \dots, T$ and $x_t = (y_{t-r}, \cdots, y_{t-d}, x_{t1}, \cdots, x_{tm})^T$ be a vector of predictors including lags of transformations of the observed response y_t and relevant external meteorological predictors, their lags or derivatives. The distribution of y_t is highly complex due to the weather conditions (states) x_t and can be approximated by the following class of MoE:

$$\psi(y_t, x_t, \Psi) = \sum_{j=1}^{J} \pi_j(z_t, \beta_j)\varphi(y_t, \mu_j, \sigma_j)$$

where $\varphi(y_t, \mu_j, \sigma_j)$ is the density of the normal distribution with mean $\mu_j(t) = \theta_{j0} + \theta_{j1}x_{t1} + \cdots + \theta_{jm}x_{tm}$ and dispersion σ_t^2, $\pi_j(z_t, \beta_j) \geq 0$ is the jth probability, the so called mixing proportions which satisfy the conditions $\sum_{j=1}^{J} \pi_j(z_t, \beta_j) = 1$, $\Psi = (\beta_1^T, \cdots, \beta_J^T, \theta_1^T, \cdots, \theta_J^T, \sigma_1, \cdots, \sigma_m)^T$ is the vector of unknown coefficients, $\beta_j^{\mathrm{T}} = (\beta_{j0}, \cdots, \beta_{jq})$ and $\theta_j^{\mathrm{T}} = (\theta_{j0}, \cdots, \theta_{jm})$, the coordinates of vector z_t might be part of the x_t coordinates.

The maximum likelihood method is used to estimate the vector of coefficients Ψ. The likelihood function of the time series $y_{d+1}, \cdots, y_T$ is defined by:

$$L(\Psi) = \prod_{t=d+1}^{T} \psi(y_t, x_t, \Psi) = \prod_{t=d+1}^{T} \sum_{j=1}^{J} \pi_j(z_t, \beta_j)\varphi(y_t, \mu_j, \sigma_j)$$

In general, the probabilities $\pi_j(z_t, \beta_j)$ are assumed to have a multinomial linear logistic (classification neural network) regression form:

$$\pi_j(z_t, \beta_j) = \frac{exp(z_t^T \beta_j)}{\sum_{i=1}^{J} exp(z_t^T \beta_i)} \quad \text{for} \quad \text{j} = 1, \ldots, \text{J}.$$

Therefore the maximum likelihood estimate $\widehat{\Psi}$ of Ψ is defined as the solution of the optimization problem:

$$\max_{\Psi} L(\Psi) \quad \text{under constraints} \quad \pi_j(z_t, \beta_j) \geq 0 \quad \text{and} \quad \sum_{j=1}^{J} \pi_j(z_t, \beta_j) = 1 \qquad (1)$$

The optimization is performed with the so called lasso type penalization on the coordinates of Ψ (an additional constraint $\sum_{l=1}^{L} |\Psi_l| \leq \varepsilon$, where $\varepsilon \geq 0$ and L is the length of Ψ in order to select the most significant predictors and overcome multicollinearity problem, according to Städler et al. (2010). The expectation-maximization (EM) algorithm is a standard technique to obtain $\widehat{\Psi}$ (McLachlan and Peel, 2001).

Standard software based on the EM algorithm can be used to handle the computation. For instance, the R packages *flexmix* developed by Grün et al. (2025) and *MoEClust* of Murphy and Murphy (2020) can fit MoE.

In real applications the number of mixture components J is unknown and has to be estimated. A classical approach is to fit models with an increasing number of components and then to compare them. The optimal number of mixture components is determined as minimum of the Bayesian Information Criteria (BIC) defined by:

$$BIC = -2\log\left(L(\widehat{\Psi})\right) + klog(T - d)$$

where *T–d* is the data sample size, and *d* is the maximal lag of x_t (McLachlan and Peel, 2001).

Selecting a model that minimizes BIC provides a relatively parsimonious model that fits the data well. However, BIC should be used as a guide only as the final decision must be evaluated in terms of physical realism.

The R packages *earth*, *gbm* and *relgam* developed by Milborrow et al. (2024), Ridgeway (2024) and Tay and Tibshirani (2024), respectively, are used to estimate PM10 MARS, GBM and GAMs models. More details about these methods can be found in Friedman (1991), Ridgeway (2024), and Tay and Tibshirani (2020).

4. MODEL PREDICTORS

The WRF model has been used to deliver 72 hour forecasts at 1 $km \times 1$ km spatial and 8 second temporal resolution to characterize the atmospheric conditions in Sofia area for the next 3 days from ground surface up to 5500 m height. The model forecasts are: t2[C] - surface temperature at 2m, td2[C] - dew point temperature, ts[C] - soil temperature, tst2[C] - surface temperature ratio, U10[deg] - wind direction and

V10[m/s] - speed at 10 m, Q2[g/kg] - specific humidity, Us[m/s] - roughness velocity, Moi[1/m] - inverse Monin - Obukhov length, PBLH[m] - Planetary Boundary Layer height, PBLT[m^2/s^2] - kinetic turbulent energy, PBLL[m] - horizontal scale, CL - low cloudiness level (ceiling level), Pasq - Pasquil stability parameter, W[m/s] - mean vertical velocity in the lowest 50 m layer, G20, G50, G120, G200, G300, G500, G800[C/100m] - temperature gradients in 7 layers (0-20 m, 20-50 m, 50-120 m, 120-200 m, 200-300 m, 300-500 m and 500-800 m), RAINC [mm] - convective precipitation and RAINNC [mm] - nonconvective precipitation.

These derivatives serve as input predictors to MoE, GAMs, GBM and MARS time series regression models. Moreover, in order to account strongly the meteorological impact on the pollution process dynamics lags of transformations of the WRF predictors at the following discrete times in hours 6, 12, 18, 24, 48 and 72 are also included in the models e.g. $t2_{t-6}, t2_{t-12}, t2_{t-18}, t2_{t-24}, t2_{t-48}, t2_{t-72}$ for the lags of surface temperature at 2 m.

Finite Fourier terms are included as deterministic predictors in the MoE model in order to account various hidden periodicity and trends of the pollution process such as seasonality and diurnally.

Lags of transformation of the observed values of PM10 are included in the corresponding models as well as predictors at the following discrete times in hours 72, 78, 84, 90, 96, 102, 108 and 114 in order to account the evolution of the pollution process. As the horizon of the PM10 concentrations forecast is 72 hours this means that past hourly values of PM10 with a delay at least 72 hours could be included in the model. Pairs of interactions between some of these derivatives e.g. t2*PBLH, t2*G800, G800*RAINNC are also included as model predictors in order to improve the quality of the fit. Thus a total number of model predictors become 364.

5. DATA SPLITTING FOR TIME SERIES VALIDATION

Dataset is split into training and testing sets corresponding to the periods 01.01.2014-31.12.2024 (95723 hours) and 01.01.2025-28.12.2025 (5192 hours). The test data set is split into 72≈5192/72 folds consisting of 72 consecutive hours each in order to assess the forecasts accuracy. The "rolling forecasting origin" cross-validation procedure offered by Hyndman and Athanasopoulos (2013) that moves the training and test sets in time is followed. The procedure starts with fitting the model to the training data set and then continues by model evaluation on the current fold of testing data set. The current fold is moved to the training data set after its model evaluation while the fold forecasts are saved for further analysis such as the accuracy measures discussed in the next paragraph. In this way the training set is incremented while the testing set is decremented by a fold at a time. The procedure continues until the folds are exhausted.

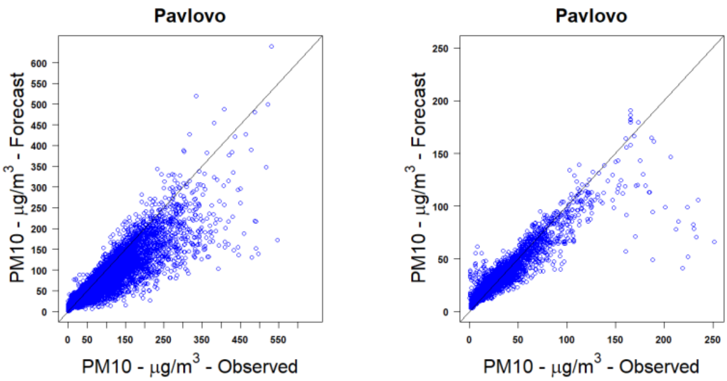

Fig. 3. Performance: predicted versus observed PM10 plots, from training (on the left) to test (on the right) data based on MoE models with 6 components.

Several MoE time series regression models with different components number were fitted to the training data set. The model with 6 components is minimal as its BIC equals 146998.6 whereas the corresponding BIC values of the models with 1 to 5 and 7 components are with greater values. From the left plot of Figure 3 one can see that the fit is not perfect although the value of the majority data follows diagonal line, the correlation coefficients presented in Table 1 is 0.91. For instance, the model underestimates the extreme PM10 values on the other hand overestimate lower PM10 values.

6. ACCURACY MEASURES

Common metrics to evaluate the forecast accuracy include: mean absolute error (MAE) $(1/T)\sum_{t=1}^{T}|y_t^{test} - \hat{y}_t^{test}|$, mean squared error (MSE) $(1/T)\sum_{t=1}^{T}(y_t^{test} - \hat{y}_t^{test})^2$, root mean square error (RMSE) $\sqrt{(1/T)\sum_{t=1}^{T}(y_t^{test} - \hat{y}_t^{test})^2}$ and Pearson correlation coefficient. These measures are evaluated on the test data y_t^{test} for t=1,…, 5192, the forecast values $\hat{y}_t^{test}$ are based on the test data set which is independent from the training data set. MAE ranges from 0 to infinity and a perfect fit is obtained when MAE equals 0. The smaller value of RMSE indicates the better performance of the model. The MSE is the average forecast error representing the systematic error of a forecast model to under or overforecast and it takes values on the whole real line. Value closer to zero implies a perfect fit. The Pearson coefficient of correlation takes on values between -1 and 1, with values closer to 1 implying a better fit.

Table 1. Performance measures based on PM10 training and test data

Type	Training data				Test data			
Model	MAE	MSE	RMSE	cor	MAE	MSE	RMSE	cor
MoE	5.89	125.47	11.20	0.93	6.37	184.51	13.58	0.91
GAMs	13.32	676.46	26.00	0.68	11.77	386.96	19.67	0.67
GBM	12.61	435.16	20.86	0.64	12.21	523.48	22.88	0.62
MARS	11.77	499.45	22.35	0.65	12.27	431.83	20.78	0.61

Performance measures based on MoE, GMB and MARS models are given in Table 1. MoE model is optimal in comparison with GAMs, GMB and MARS models according to estimated accuracy measures MAE, MSE, RMSE and the correlation coefficients within the training and and test data given Table 1. The accuracy measures based on MoE are reasonably low whereas the correlation coefficients equal to 0.93 and 0.91, respectively.

7. CONCLUSION

Several models are developed to forecast hourly PM10 concentrations 72 hours ahead using data collected by the ExEA monitoring station located at Pavlovo for the period 01.01.2014-28.12.2025. These models are hybrid between WRF model and the so called MoE time series regression models. The WRF output delivers 72 hour forecast which characterizes the atmospheric conditions at Pavlovo. The WRF output derivatives, the lags of these derivatives, lags of PM10 series serve as input predictors into the MoE time series regression models. The models provide reasonably good fits on test data set in terms of the standard forecast accuracy metrics such as MAE, MSE and RMSE and correlation coefficients between observed and forecasted values. The paper presents a new methodology for estimation of the relationship between air pollution ingredients PM10 and the numerical weather prediction output meteorological parameters for Pavlovo station of the urban area of Sofia, Bulgaria. Similar results are obtained for the other monitoring stations of the ExEA in Sofia located at Druzhba, Hipodruma and Nadezhda.

ACKNOWLEDGEMENTS

This work has been done as a collaboration between NIMH and Sofia Municipality project on air pollution. The authors wish to express their gratitude to Prof. DSc Syrakov for the fruitful discussions.

REFERENCES

Bertaccini, P., Dukic, V. and Ignaccolo, R. (2012) Modeling the short-term effect of traffic and meteorology on air pollution in Turin with Generalized Additive Models. Advances in Meteorology, Volume 2012, Article ID 609328, 16 pages, https://doi.org/10.1155/2012/609328

Carvalho, A. and Tanner, M. A. (2005). Modeling nonlinear time series with local mixtures of generalized linear models. The Canadian Journal of Statistics, vol. 33(1), pp. 97-113.

Cusano, M., Gaeta, A., Morelli, R., Cattani, G., Canepari, S., Massimi, L. and Leone, L. (2025) Spatial Modeling of Trace Element Concentrations in PM10 Using Generalized Additive Models (GAMs), Atmosphere , vol. 16, 464. https://doi.org/10.3390/atmos16040464

Friedman, J. H. (1991) Multivariate Adaptive Regression Splines. Ann. Statist. vol. 19(1), pp. 1-67. https://doi.org/10.1214/aos/1176347963

Grün, B, Leisch, F., Sarkar, D., Mortier, F. and Picard, N. (2025) flexmix: Flexible Mixture Modeling, https://CRAN.R-project.org/package=flexmix

Hyndman, R.J., Athanasopoulos, G. (2023) Forecasting: Principles and Practice, 3rd ed., OTexts: Melbourne, Australia, OTexts.com/fpp3
Misiti, M., Misiti, Y., Poggi, J-M, and Poritier, B. (2015) Mixture of linear regression models for short term PM10 forecasting in Haute Normandie (France). Societe Française de STATISTIQUE, vol. 6(1), pp. 47-60.
McLachlan, G.J. and Peel, D. (2000) Finite Mixture Models.Wiley, NewYork.
Milborrow, St., Hastie, T., Tibshirani, R., Miller, A. and Lumley, T. (2024) R package earth for Multivariate Adaptive Regression Splines, http://www.milbo.users.sonic.net/earth/.
Nieto, P.J.G., Anton, J.C.A. and Garcia-Gonzalo, E. (2014) PM10 modeling in the Oviedo urban area (Northern Spain) by using multivariate adaptive regression splines. AIP Conf. Proc. 1618, 1013-1016, https://doi.org/10.1063/1.4897906
Park, S., Son, S., Bae, J., Lee, D., Kim, J.-J., Kim, J. (2021) Robust Spatiotemporal Estimation of PM Concentrations Using Boosting-Based Ensemble Models. Sustainability, vol. 13, 13782. https://doi.org/10.3390/su132413782
Ridgeway, G. (2024) The R package gbm Generalized Boosted Regression Models. https://github.com/gbm-developers/gbm
R Development Core Team, R: A Language And Environment for Statistical Computing, 2025, R Foundation for Statistical Computing, Vienna, Austria, http://www.r-project.org./.
Ridgeway, G. (2024) Generalized Boosted Models: A guide to the gbm package. http://127.0.0.1:28309/library/gbm/doc/index.html
Ridgeway, G. (1999). The state of boosting. Computing Science and Statistics, vol. 31, 172-181.
Russo, A., Lind, P.G., Raischel, F., Trigo, R. and Mendes, M. (2015) Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales. Atmospheric Pollution Research, vol. 6, pp. 540-549.
Shahraiyni, H. and Sodoudi, S. (2016) Statistical Modeling Approaches for PM10 Prediction in Urban Areas: A Review of 21st-Century Studies, Atmosphere, vol. 7(2), https://doi.org/10.3390/atmos7020015
Seni, G. and Elder, J. (2010). Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Springer.
Städler, N., Bühlmann, P. and van de Geer, S. (2010). L1-penalization for mixture regression models. Test, vol. 19, 209-256.
Tay, J. K. and Tibshirani, R. (2020). Reluctant Generalised Additive Modelling. International Statistical Review, 88(S1). https://doi.org/10.1111/insr.12429
Tay, J. K. and Tibshirani, R. (2024) relgam: Reluctant Generalized Additive Models. https://CRAN.R-project.org/package=relgam
Vlachogianni, A., Kassomenos, P., Karppinen, A., Karakitsios, S. and Kukkonen, J. (2011). Evaluation of a Multiple Regression Model for the Forecast of the Concentrations of NOx and PM10 in Athens and Helsinki. Sci. Total Environ. vol. 409, pp. 1559-1571.
Wallek, S., Langner, M., Schubert, S., Franke, R., Sauter, T. (2024) Hourly Particulate Matter (PM10) Concentration Forecast in Germany Using Extreme Gradient Boosting. Atmosphere vol. 15, 525. https:// doi.org/10.3390/atmos15050525
Wood, S. (2017) Generalized Additive Models. Chapman and Hall/CRC, NewYork.
Zhou, Z-H. (2012). Ensemble Methods: Foundations and Algorithms (Chapman & Hall/CRC Machine Learning & Pattern Recognition).