University of Maribor Press

DOI: https://doi.org/10.18690/um.fov.5.2026.6
ISBN: 978-961-299-160-9

Machine Learning for Air Quality and CO_2 Emissions: The Role of Data Understanding

Ivan Maslov,
Agnieszka Głowacka,
Bartosz Dziewit, and
Paulina Trybek

Abstract
The emergence of machine learning (ML) has enabled sophisticated approaches to environmental prediction, yet the datasets underlying these models remain underexamined. This study investigates the role of data quality and structure in ML-based environmental applications, focusing on greenhouse gas (GHG) emissions and particulate matter concentrations. For Poland and Slovakia, a LightGBM model was trained to predict CO_2 emissions across major economic sectors: Residential, Power, Transport, Industry, and Aviation. Predictive performance was highest in sectors with regular seasonal patterns, while low-variability sectors such as Domestic Aviation posed greater challenges. For particulate matter, meteorological and time-related features were used to forecast $PM_{2.5}$ and PM_{10} during the heating season. Models captured general temporal patterns and seasonal peaks, though extreme events were partially underestimated. The findings demonstrate that predictive accuracy is strongly influenced by dataset quality, resolution, and structure, alongside emission regularity and environmental conditions. This chapter underscores the importance of careful dataset design and preprocessing in ML applications for environmental monitoring, offering practical guidance for improving the reliability of emission and air quality forecasting.

Keywords: machine learning; CO_2 emissions; dataset transparency; $PM_{2.5}$ particles; air quality forecasting

Ivan Maslov and Agnieszka Głowacka
University of Silesia, Faculty of Science and Technology, 40-007 Katowice, Poland
e-mail: ivan.maslov,agnieszka.glowacka@us.edu.pl
Bartosz Dziewit and Paulina Trybek
University of Silesia, Institute of Physics, 40-007 Katowice, Poland
e-mail: bartosz.dziewit,paulina.trybek@us.edu.pl

1 Introduction

Air pollution constitutes one of the most significant environmental health risks worldwide. According to the World Health Organization, exposure to polluted air contributes to millions of premature deaths annually, primarily due to fine particulate matter ($PM_{2.5}$) [1]. Previous global assessments have further demonstrated that ambient $PM_{2.5}$ accounts for a substantial mortality burden and that significant reductions in premature deaths could be achieved through region-specific air quality improvements aligned with WHO guidelines [2].

At the same time, atmospheric carbon dioxide (CO_2), the dominant anthropogenic greenhouse gas, is the principal driver of long-term climate change. More broadly, greenhouse gas (GHG) emissions are widely recognized as the primary force behind ongoing global warming. While particulate matter ($PM2.5$, PM_{10}), carbon monoxide (CO), and carbon dioxide (CO_2) differ in their physio-chemical properties and atmospheric lifetimes, they frequently originate from common combustion-related sources such as transport, industry, and residential heating. This overlap in emission sources highlights the close interconnection between air quality and climate policy: measures targeting fossil fuel combustion can simultaneously reduce harmful air pollutants and greenhouse gas emissions.

In 2020, the European Union adopted a plan to reduce net greenhouse gas emissions by at least 55% by 2030 compared to 1990 levels, alongside the longer-term objective of achieving climate neutrality by 2050 [3]. Such policy frameworks not only aim to mitigate climate change but also have the potential to deliver substantial public health co-benefits through improved air quality.

To address these challenges, data-driven approaches, including machine learning and artificial intelligence methods, are increasingly applied to model and predict atmospheric concentrations of pollutants such as $PM_{2.5}$ or CO_2 [4, 5, 6]. By leveraging large environmental datasets and complex nonlinear relationships among meteorological, emission, and spatial variables, these methods can improve the accuracy of air quality forecasting and support evidence-based environmental management and policy-making. However, the reliability of such predictive models strongly depends on the availability of high-quality and representative data that adequately capture the key factors influencing atmospheric concentrations of pollutants, including meteorological conditions, emission sources, and spatial–temporal variability.

Despite certain similarities in their emission sources, predicting atmospheric concentrations of $PM_{2.5}$ and CO_2 involves different methodological challenges. Particulate matter concentrations are typically characterized by strong short-term variability and pronounced spatial heterogeneity driven by local emission sources, meteorological conditions, and atmospheric processes such as dispersion and secondary aerosol formation. In contrast, atmospheric CO_2 concentrations tend to exhibit smoother spatial patterns and stronger long-term temporal trends associated with global-scale carbon cycles and cumulative greenhouse gas emissions. As a result, predictive models for $PM_{2.5}$ often require high-resolution spatiotemporal data capturing local environmental conditions, whereas CO_2 modeling more frequently emphasizes large-scale temporal dynamics and broader emission pattern.

The importance of CO_2 in climate policy can be illustrated by recent emission statistics. According to the most recent Polish National Greenhouse Inventory report (NIR) [7], prepared under the United Nations Framework Convention on Climate Change (UNFCCC), carbon dioxide accounts for 81.4% of total greenhouse gas emissions in Poland. A similar dominance of CO_2 is observed at the European level, where it represents roughly 80% of total greenhouse gas emissions [8]. Therefore, reducing CO_2 emissions plays a central role in achieving climate neutrality.

In order to keep such emissions under control, policy makers, researchers and analysts require reliable modeling approaches and prediction algorithms, which are among the key topics studied in modern computer science. In recent years, a plethora of articles has been published applying various machine learning (ML) models and techniques for the prediction of CO_2 emissions [9, 10, 11, 12, 13]. This research area is still under active development, and no single authoritative solution has yet been adopted at an intergovernmental level.

Related work explores a wide range of machine learning algorithms across different sectoral contexts. However, the data used to train these models is often accepted without sufficient scrutiny. In most cases, datasets originate from official publications or open-source repositories, are lightly processed to handle missing values, and are then passed directly to models without deeper analysis.

The characteristics and limitations of the source data—such as spatial and temporal resolution, aggregation methods, data composition, uncertainty evaluation, and even the fact that much of the data represents estimates rather than direct measurements—rarely receive adequate attention. As a result, datasets are frequently taken for granted, despite their significant influence on model outcomes.

Unlike CO_2 emissions, which are largely estimated based on activity data and emission factors, concentrations of particulate matter such as $PM_{2.5}$ and PM_{10} are typically determined through direct physical measurements. This means that these data reflect the actual state of the atmosphere at a specific place and time.

Pollutant concentration is expressed as the amount of particles per cubic meter of air. PM_{10} includes particles with a diameter of up to 10 micrometers, while $PM_{2.5}$ includes particles up to 2.5 micrometers in diameter. This distinction is important both from a health and a measurement perspective, as finer particles behave differently in the atmosphere. Various technologies are used to measure $PM_{2.5}$ and PM_{10} concentrations. The reference method remains gravimetry, which involves drawing air through a filter and determining particle concentrations based on the increase in the filter's mass. While highly accurate, this approach does not provide real-time results. In practice, automatic analyzers, such as the Beta Attenuation Monitor (BAM) and the Tapered Element Oscillating Microbalance (TEOM), are more commonly employed, enabling hourly measurements and continuous monitoring of particulate levels. Increasingly important are also low-cost sensors based on light scattering, which allow for denser monitoring networks but are characterized by greater uncertainty [14]. Technological advancements have increased the availability of real-time data and the density of observations. This is particularly important for epidemiological analyses and predictive modeling. At the same time, a larger number of devices implies greater variability in measurement quality. Under conditions of high humidity, especially during fog, microdroplets of water may be recorded as additional particulate matter, leading to overestimated results. Dif-

ferent types of instruments may also respond differently to the same atmospheric conditions, resulting in systematic differences in measurements. Additionally, sensors—especially lower-cost models—may lose accuracy over time if not regularly calibrated [15]. An important feature of PM data is their high temporal resolution. Dust concentration measurements are typically recorded at hourly intervals, and sometimes even more frequently, meaning that a single station generates thousands of observations per year. This enables the analysis of diurnal cycles, smog episodes, weekend effects, and seasonal patterns related to the heating period. However, a high number of observations does not automatically translate into greater representativeness. Measurements refer to a specific location and reflect local emission and dispersion conditions. PM data therefore have a point-based and spatially limited character, as they represent a record of the instantaneous atmospheric state at a given location rather than an averaged measure at the regional or national scale [16].

In this study, we examine the distinct characteristics and challenges of environmental datasets for both CO_2 and particulate matter ($PM_{2.5}$ and PM_{10}) in the context of machine learning applications. Unlike particulate matter, which is measured directly at specific locations and reflects short-term local atmospheric conditions, CO_2 emissions are largely estimated based on activity data and emission factors, making them subject to additional sources of uncertainty. These differences imply that predictive modeling for each pollutant type requires tailored approaches: high-resolution spatiotemporal data for particulate matter and broader temporal and sectoral information for CO_2. Accordingly, this work tests predictive models for both CO_2 and PM concentrations, highlighting how data set quality, representativeness, and the nature of the underlying measurements influence the reliability and interpretability of machine learning forecasts.

2 Analysis of carbon emission data

There are typically two types of CO_2 emission data used in practice: direct measurements and emission estimates. The former is relatively rare and is usually available either from short-term, locally focused research projects or from controlled test datasets, such as those provided by the automotive industry through the Worldwide Harmonized Light Vehicles Test Procedure (WLTP) framework [17].

Emission estimates, on the other hand, are far more common. Large-scale CO_2 emission databases typically rely on estimations derived from statistical information about human activities, such as energy production, transportation, and industrial processes.

The main reference for CO_2 emission estimation is the 2006 IPCC Guidelines for National Greenhouse Gas Inventories [18]. The Intergovernmental Panel on Climate Change (IPCC) is a United Nations body responsible for assessing scientific knowledge related to climate change. It was established in 1988 by the United Nations Environment Programme (UNEP) and the World Meteorological Organization (WMO) to provide a consistent and transparent framework for the evaluation and reporting of greenhouse gas (GHG) emissions.

The 2006 IPCC Guidelines establish a concrete methodology for estimating greenhouse gas inventories, that is, GHG datasets developed with defined quality assurance

procedures and uncertainty evaluation. The IPCC framework provides a division of emissions by major human activity sectors, namely energy, industrial processes, agriculture, forestry and other land use, and waste, along with further subdivisions within each sector. In addition, the Guidelines offer detailed guidance on how emission values should be calculated for different activities and data availability levels.

The IPCC provides a basic approach for emission estimation (1), in which emissions are calculated as the product of activity data (AD) and an emission factor (EF). Activity data represent the magnitude of a human activity—such as the amount of fuel consumed in a given process—while the emission factor defines the amount of CO_2 emitted or removed per unit of activity.

$$Emissions = AD * EF \tag{1}$$

The form of activity data varies across economic sectors. For example, it may correspond to the quantity of coal burned in industrial facilities, or to a more elaborate representation of the aviation sector that accounts for the number of flights, travel distances, aircraft types, and other relevant characteristics. Nevertheless, the underlying estimation principle remains the same, which is why activity data derived using different approaches can be combined in total emission calculations.

Regarding emission factors, the IPCC provides a database of standardized values that can be used as defaults. However, it is considered good practice to apply country-specific or locally derived emission factors whenever possible, as they better reflect local conditions and technologies. For instance, if coal used within a given country has distinct physical or chemical characteristics, locally measured emission factors can provide more accurate estimates of the resulting CO_2 emissions.

The IPCC also introduces three tiers of analysis that differ in terms of data granularity and methodological complexity. Tier 1 relies on national-level activity statistics combined with default emission factors, while Tier 3 represents the most detailed approach, using highly disaggregated local activity data together with locally derived emission factors. The Guidelines generally recommend a bottom-up approach, which focuses on analyzing individual emission sources, rather than a top-down approach that estimates emissions based on aggregated indicators such as total fuel consumption and import-export information.

It is important to note that the 2006 IPCC Guidelines also state that direct emissions monitoring alone is generally not preferred due to the high cost and difficulties related to data interpretation. In particular, direct measurements are often limited to specific locations or facilities, making it challenging to obtain representative samples and to extrapolate results to larger spatial scales, such as an entire industrial sector or a whole country. As a result, fuel-based estimation methods are often preferred for national inventories, as they provide more consistent and scalable coverage [18].

2.1 CO_2 emissions and relevant datasets

As stated above, the IPCC Guidelines serve as a common reference point for most publicly available CO_2 datasets. The scientific community generally agrees that emissions can be estimated as the product of human activity data and an emission factor. However, the approaches to data collection and the choice of how activity data are modeled vary significantly between datasets.

Figure 1 illustrates, in broad strokes, the main steps involved in the creation of a CO_2 emissions estimation dataset. It also highlights the stages at which errors or discrepancies may be introduced.

To construct such a dataset, developers require both activity data and emission factor information. Activity data can be obtained from a variety of sources, including local and national statistical offices, as well as independent international agencies such as the International Energy Agency (IEA) and the Energy Institute. Similar diversity exists for emission factor sources. In addition, data set creators may choose different estimation models and apply various data refinement techniques, for example to address gaps in time series or the lack of sufficiently detailed activity data.

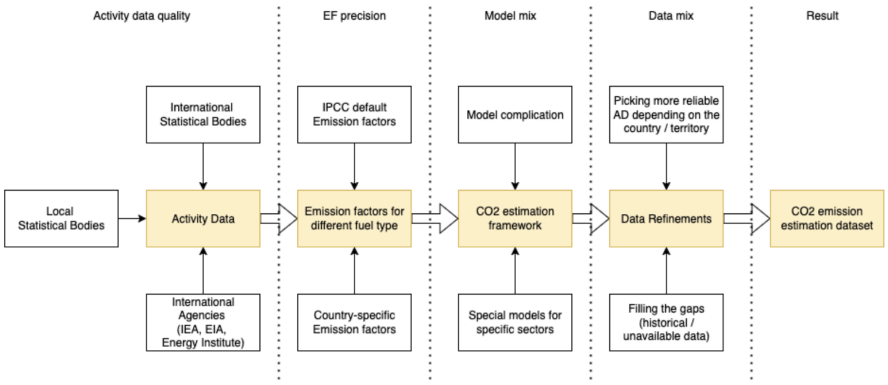

Fig. 1: Main components of a CO_2 estimation data set creation process.

As a result, emission estimates may differ from one data set to another. For example, Figure 2 compares CO_2 emission estimates for Poland and its neighboring country Slovakia, where reported emission levels are significantly lower.

The UNFCCC column represents estimates reported by national authorities and constitutes the official CO_2 emission inventory. These estimates are based on national statistical data covering energy use, industrial activity, and land use, and they typically rely on country-specific emission factors, for example those derived from information on local fuel characteristics such as coal quality.

The EDGAR dataset, developed by the European Union's Joint Research Centre, is an independent emissions database that primarily uses activity data from the International Energy Agency and, in some cases, generalized emission factors. The third data set, produced by the Global Carbon Project, is particularly interesting because it relies on United Nations data for developed countries. In principle, this should result in estimates similar to those reported under the UNFCCC framework; however, noticeable differences remain.

Overall, no clear systematic bias can be observed across the datasets — for example, EDGAR estimates are not consistently higher or lower than others. All of these databases are widely regarded as credible and are commonly used in scientific analysis and policy development.

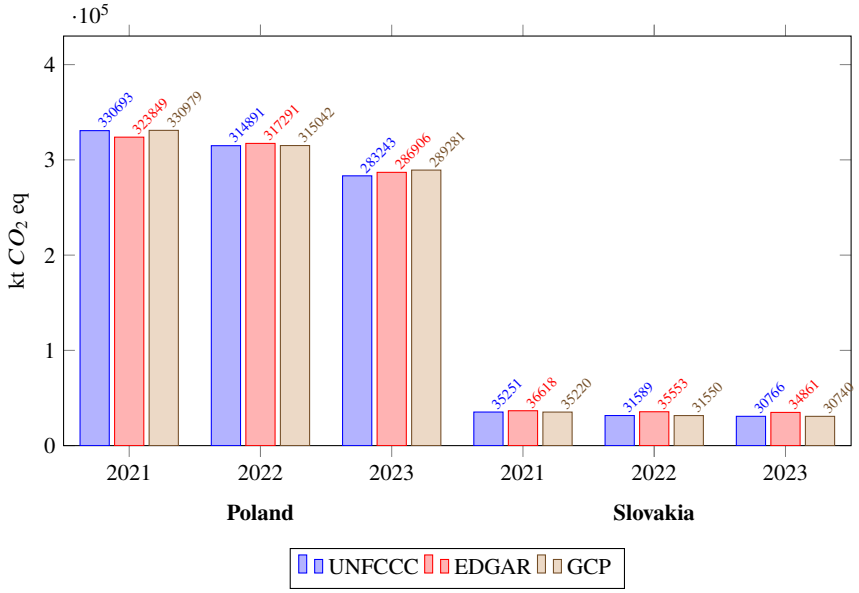

Fig. 2: Annual CO_2 emission estimation in kt CO_2 eq by country and source (2021 - 2023).

Nevertheless, these inconsistencies are important to consider when developing any kind of modeling or prediction project, especially when combining data from multiple sources in machine learning applications.

In the following section, we characterize several publicly available datasets that are widely used in AI and machine learning research.

2.2 Datasets review

The United Nations Framework Convention on Climate Change (UNFCCC) [19] is an international environmental treaty adopted in 1992 in response to growing concerns about climate change. Today, the Convention has been adopted by 197 Parties. Under the UNFCCC framework, countries are expected to report national greenhouse gas inventories in strict accordance with IPCC Guidelines.

The dataset contains annual data for greenhouse gases for each country, currently extending up to 2021. Although the database [20] provides aggregated data on all parties, important distinction exists between different groups of countries. Parties listed in Annex I, that are primarily developed countries and major greenhouse gas emitters, are required to submit detailed national inventory reports on an annual basis. In contrast, developing countries report their emissions through National Communications (NCs) or Biennial Update Reports (BURs), which are submitted less frequently and often with a delay of several years. For example, Chad reports emissions data only for the year 2010.

Using data from different Annex 1 countries could also present challenges, as countries may choose different ways of interpreting activity data and different precise models for emission estimation on the same activity sector.

Another important characteristic of data reported under the United Nations Framework Convention on Climate Change (UNFCCC) is that national inventories are largely based on country-specific data sources and reporting practices. These inventories rely heavily on local governmental activity data, such as national energy and industrial statistics, as well as locally determined fuel properties and emission factors. While this approach can better reflect national conditions, it also means that the quality and transparency of the underlying data may vary between countries. In some cases, this may raise questions about the consistency and reliability of reported statistics.

Global Carbon Budget data: The Global Carbon Project (GCP) is an international research initiative focused on studying the global carbon cycle. Each year, the GCP publishes the Global Carbon Budget report, which provides estimates of global carbon emissions and carbon sinks based on the latest available data and models [21].

In addition to the summary report, the GCP releases several openly available datasets that cover fossil fuel CO_2 emissions, land-use change emissions, and carbon uptake by land and ocean systems. These datasets are provided in both national and gridded formats, include uncertainty estimates, and are updated annually, making them widely used in climate research and data-driven modeling.

Previously, the data set was taken from US-based CDIAC, the Carbon Dioxide Information Analysis Center of the Oak Ridge National Laboratory. CDIAC project was closed in September 30, 2017. Nowdays GCP uses CDIAC-FF data as a base and extend it by 2–3 years using energy growth rates derived from data published by the Energy Institute.

As described in GCP documentation, CDIAC applied standardized emission factors to apparent energy consumption derived from United Nations energy statistics, including emissions from gas flaring. These estimates were then extrapolated for additional

years using growth rates previously obtained from BP's Statistical Review of World Energy. Apparent consumption is calculated from data on energy production, imports, exports, and stock changes, and differs from observed consumption, which is based on direct industry reporting or sales data. Within the IPCC framework, this approach corresponds to the Reference Approach, used for cross-checking the national inventories, while the Sectoral Approach relies on alternative activity data sources.

In recent years, the GCP has increasingly shifted toward using country-specific and sector-specific energy data sources where available. So sometimes and for some countries estimates derived from CDIAC-FF may be replaced by data reported under the UNFCCC. In another example, emissions from natural gas consumption are now partly derived from data provided by the Joint Organisations Data Initiative (JODI). GCP researchers acknowledge that this results in the use of multiple data sources and extrapolation methods, but argue that these approaches aim to estimate the same underlying quantities and therefore do not significantly alter the overall emission estimates.

Data from the Global Carbon Budget are used by other projects, notably by the Our World in Data project, which in turn serves as a popular open-source solution of CO_2 and greenhouse gas emissions data for machine learning applications [22].

EDGAR, which stands for the Emissions Database for Global Atmospheric Research, is a project developed by the Joint Research Centre of the European Commission [23]. As stated on the project website, EDGAR provides global estimates of past and present anthropogenic emissions of greenhouse gases and air pollutants at the country level and on a spatial grid [24].

EDGAR applies a distinct emission estimation model 2 that combines country-specific emission factor (EF) and activity data (AD), largely derived from assessments by the International Energy Agency, with information on national technological mixes (TECH). The model incorporates technology-dependent emission (EOP) factors and accounts for emission reductions achieved through the use of abatement systems (RED). This approach allows EDGAR to reflect differences in technologies and mitigation measures across countries while maintaining a consistent global framework.

$$EM_c(y,x) = \sum_{i,j,k} \Big[AD_{c,i}(y) \cdot TECH_{c,i,j}(y) \cdot EOP_{c,i,j,k}(y) \cdot EF_{c,i,j}(y,x) \cdot (1 - RED_{c,i,j,k}(y,x)) \Big] \tag{2}$$

It is also important to note that EDGAR extends its emissions time series for the most recent years using projected or preliminary energy statistics, including growth rates published by the Energy Institute. This means that the most recent years in the EDGAR dataset are based on extrapolated data rather than fully reported actual activity statistics. For machine learning applications, this introduces additional complexity, as models may be trained on data compiled using different sources and estimation methods across time, which should be taken into account when interpreting results.

2.3 Challenges for machine learning applications

The current state of open-source CO_2 dataset composition presents significant challenges for researchers aiming to apply machine learning algorithms to carbon dioxide emission analysis and emission prediction. In the following, we highlight several of these challenges.

Data are preprocessed. Every CO_2 emissions dataset consists of estimates derived from underlying statistical activity data. This activity data may originate from different institutions that apply distinct methods of data collection and aggregation. Furthermore, missing values in these datasets are often filled using extrapolation based on historical averages or replaced with values from alternative data sources.

This inherent preprocessing can complicate machine learning research, as the data used for training, validation, or reference may not be uniform or fully comparable across datasets.

Low temporal resolution. Most publicly available CO_2 emission datasets provide data at an annual resolution at best, which significantly limits their suitability for many machine learning applications. The lack of higher-frequency observations restricts the ability of models to capture short-term dynamics and temporal variability.

Some authors argue [13] that this limitation can also influence methodological choices, encouraging the use of scaled data and evaluation metrics such as RMSE or MAE computed on normalized values rather than raw observations. As a result, reported model performance may be overstated and less representative of real-world predictive accuracy.

Changing nature of activity data. In some cases, statistical agencies and dataset developers revise historical emission estimates when new or improved activity data become available. In addition, the underlying activity data itself may change due to social, economic, or political actions that are independent of scientific modeling practices. Such changes can introduce structural breaks in emission time series that are not related to actual changes in emission behavior.

A clear example can be found in the most recent Polish National Greenhouse Gas Inventory submitted to the UNFCCC [7]. It is stated that the main cause of significant increase of GHGs emissions in 2016 - 2017 was ”substantial rise of fuels use in road transport driven by effective combat against grey-zone at fuel market started in 2016”. This suggests that a portion of fuel consumption may not have been fully captured in official statistics prior to this intervention.

From a data analysis perspective, such changes complicate trend estimation and prediction. Emission values before and after 2016 may not be directly comparable, and statistical correction or segmentation of the time series may be required before applying machine learning models for forecasting or trend analysis.

Unaddressed uncertainties. Every CO_2 emission estimation report is subject to uncertainty evaluation. Uncertainty is defined in the IPCC Guidelines [18] as the range of possible true values together with their likelihood. The Intergovernmental Panel on Climate Change (IPCC) does not prescribe a single acceptable uncertainty threshold for emission estimates; instead, it requires Parties to quantify and transparently report uncertainties associated with their inventories.

Uncertainties arise from multiple sources, most notably from the evaluation of emission factors and the aggregation of activity data, and their magnitude varies significantly across sectors and activities. For example, uncertainties in CO_2 emissions from fossil fuel combustion in Annex I UNFCCC countries are typically estimated to be on the order of ±5% to ±10%, a range that is also reported for corresponding estimates in the EDGAR dataset [25]. In contrast, activities that are less well characterized exhibit substantially higher uncertainty. Within EDGAR, uncertainty estimates reach approximately ±50% for biogas combustion, ±35% for cement-related emissions, and up to ±100% for domestic aviation emissions [26].

Uncertainties may amplify when combined with estimation models, resulting in possible variation in final estimated values. That is why models trained on CO_2 inventories may achieve good performance with respect to reported values while still failing to accurately represent real-world emissions, particularly if the underlying data contain substantial and heterogeneous uncertainties.

Without explicitly accounting for these uncertainties during model evaluation, machine learning models risk producing overconfident predictions and misleading conclusions. Consequently, careful consideration of data uncertainty should be an integral part of both model development and evaluation when applying machine learning to CO_2 emission analysis.

Lack of domain-specific knowledge. The development of CO_2 emission prediction models is an inherently cross-disciplinary task. Proper interpretation of results requires not only technical expertise in machine learning, but also domain-specific knowledge of climate science and emission accounting.

Figure 3 illustrates a simplified representation of the carbon cycle. Carbon dioxide is emitted both as a result of anthropogenic activities and through natural processes, such as forest fires. At the same time, significant amounts of CO_2 are absorbed by land and ocean systems through various sinks and removal processes. These removals are typically accounted for under Land Use, Land-Use Change and Forestry (LULUCF).

Emissions and removals related to LULUCF are particularly difficult to estimate, which is why this sector is sometimes excluded when carbon emissions are reported. However, the magnitude of these fluxes is far from negligible. According to the 2025 Polish National Inventory [7], net removals from land use in 2023 are estimated at approximately 32 656 kt CO_2 eq, a value comparable to the total annual emissions of Slovakia in the same year.

This example highlights the risk of misinterpretation when domain knowledge is lacking. Machine learning practitioners working with CO_2 datasets should therefore treat emission figures with caution and, where possible, collaborate with domain experts in climate science or ecology to ensure correct interpretation and use of the data.

2.4 Example of prediction of CO_2

To investigate the predictability of CO_2 emissions across different sectors of the economy, separate models were developed for each sector. This sector-based approach allows the models to capture sector-specific relationships and emission dynamics. The

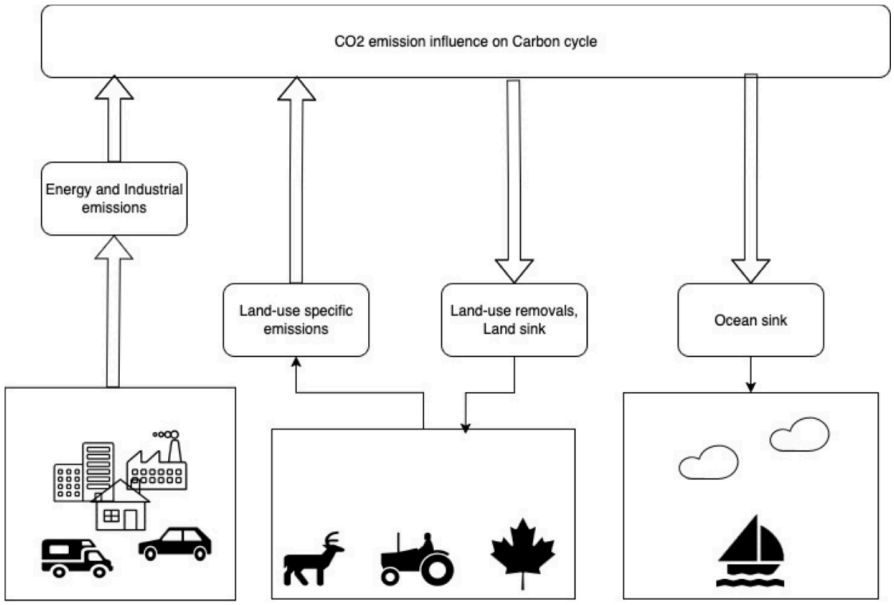

Fig. 3: Basic representation of the carbon cycle.

LightGBM algorithm, a gradient boosting method based on decision trees, was used for modeling. Light Gradient Boosting Machine (LightGBM) is an efficient and scalable gradient boosting framework that is widely used for classification and regression tasks [27]. The data used in this section come from [28] and are further characterized in article Piyu Ke et al. [29].

The data (covering from 2023-2025 for Poland and Slovakia) were split chronologically. The first 80% of observations were used to train the model and learn the relationships present in the data, while the remaining 20% were reserved for evaluating the model's ability to forecast new, unseen values.

The model used several types of features:

- lagged values from previous days (e.g., 1, 7, 14, and 30 days earlier),
- 7-day and 14-day moving averages,
- calendar variables such as day of the week and month,
- an overall time trend indicating whether emissions increase or decrease over a longer period.

Summarized prediction results for Slovakia and Poland are presented in Table 1. The corresponding visualizations of the predictions are shown in Figures 4 and 5.

In the case of Poland (see Figure 4) the predictability of CO_2 emissions varies across economic sectors. The best model performance was achieved in the Residential sector ($R^2 = 0.97$). This indicates that emissions in this area follow a very clear and repetitive

Table 1: Prediction performance for CO_2 emissions by sector in Slovakia and Poland.

Country	Sector	MAE	RMSE	R^2
Slovakia	Domestic Aviation	0.000007	0.000010	-0.020222
	International Aviation	0.000071	0.000088	0.834955
	Power	0.000743	0.001193	0.849839
	Residential	0.001150	0.001599	0.936084
	Ground Transport	0.001583	0.002228	0.690569
	Industry	0.005361	0.007676	0.407599
Poland	Domestic Aviation	0.000107	0.000136	0.312053
	International Aviation	0.000862	0.001029	0.569176
	Residential	0.004399	0.006512	0.969262
	Ground Transport	0.008854	0.013357	0.646320
	Industry	0.010669	0.013829	0.601022
	Power	0.020775	0.027573	0.733910

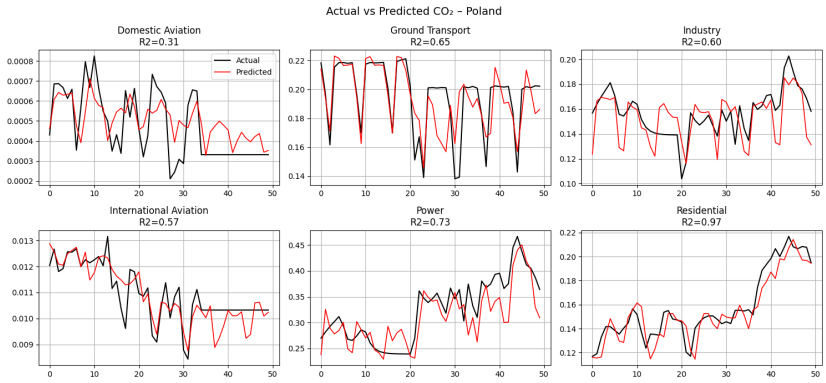

Fig. 4: Results of the LightGBM model used for the prediction of CO_2 in Poland.

pattern over time. The model accurately captures both short-term fluctuations and the overall upward trend observed at the end of the analyzed period.

Good results were also obtained in the Power ($R^2 = 0.73$), Ground Transport ($R^2 = 0.65$), and Industry ($R^2 = 0.60$) sectors. In the energy and road transport sectors, emissions change in a relatively regular manner, which allows the model to forecast them with reasonable accuracy. In the industrial sector, the fit is slightly weaker, which may be due to greater irregularity in emissions and the influence of external factors not directly included in the model (e.g., changes in production levels).

The weakest performance was observed in the aviation sectors. For Domestic Aviation ($R^2 = 0.31$) and International Aviation ($R^2 = 0.57$), the quality of predictions is lower. The data contain long periods during which emissions remain constant. This is not due to missing data, but rather to the specific characteristics of the sector. Low variability means there is less information available for the model, making it more difficult to learn patterns and accurately predict future values.

In the case of Slovakia (see Figure 5), the model also performed best in sectors where emissions follow a clear and repetitive pattern over time. The highest prediction quality was achieved in the Residential sector ($R^2 = 0.94$) and the Power sector ($R^2 = 0.85$). This means that emissions in these areas are relatively stable and show clear seasonality, which makes them easier to predict.

Good results were also obtained in International Aviation ($R^2 = 0.83$) and Ground Transport ($R^2 = 0.69$). In these sectors, emissions follow more regular patterns, which helps the model make better forecasts.

Weaker results were observed in the Industry sector ($R^2 = 0.41$). Emissions in industry are more irregular and may depend on external factors, such as changes in production, which were not directly included in the model.

The weakest predictive performance was observed in the Domestic Aviation sector, where the model's forecasts did not fully capture the observed temporal patterns in the data. Similar to Poland, the data show long periods when emissions stay at the same level. This means there is very little change over time. When the data hardly change, the model has little information to learn from, so predictions in this sector are not very accurate.

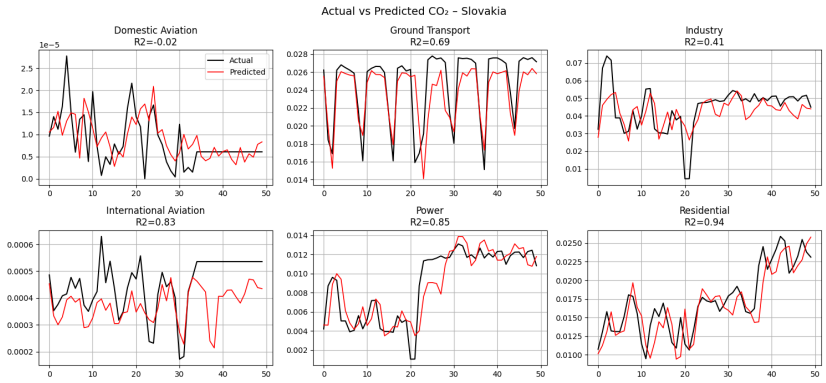

Fig. 5: Results of the LightGBM model used for the prediction of CO_2 in Slovakia.

In both countries, the highest predictability was observed in the Residential and Power sectors. This suggests that emissions in these areas follow the most regular and seasonal patterns, which makes modeling easier. In both countries, Domestic Aviation was the most problematic sector because low variability makes forecasting difficult. The main difference concerns International Aviation. In Poland, the model performance was moderate, while in Slovakia the results were much better. In the Industry sector, predictability is lower in both countries compared to sectors with strong seasonality, although Poland achieved slightly better results than Slovakia. Overall, the models work best in sectors with clear and repetitive emission patterns, and worse in sectors where emissions are low, stable, or irregular.

3 Analysis of $PM_{2.5}$ and PM_{10}

3.1 Factors influencing $PM_{2.5}$ and PM_{10} concentrations

$PM_{2.5}$ and PM_{10} concentrations result from the interaction of emission sources, chemical processes occurring in the atmosphere, and meteorological and spatial conditions. Emission sources can be divided into natural and anthropogenic. Natural sources include, among others, soil erosion, mineral dust transport, sea spray, wildfires, and volcanic eruptions. These phenomena are usually episodic in nature and strongly dependent on weather conditions. However, in many regions anthropogenic sources dominate, such as fuel combustion in the residential and municipal sector, transport, and power generation. Emissions from dispersed, low–height sources favor the accumulation of pollutants in the lower layers of the atmosphere, especially during periods of limited air exchange. Road transport generates both exhaust emissions and non-exhaust emissions (tire and road surface wear, as well as re-suspension of dust). A significant portion of $PM_{2.5}$ mass consists of secondary particles formed in the atmosphere from gaseous precursors such as SO_2, NO_x, and ammonia. Through chemical reactions, sulfates and nitrates are formed, and their share depends on temperature and air humidity. These processes give particulate concentrations a distinct seasonal pattern [16].

Meteorological conditions play a crucial role in determining $PM_{2.5}$ and PM_{10} levels, as they influence the dispersion and removal of particles from the atmosphere. Even with similar emission levels, changes in weather can cause significant fluctuations in observed concentrations. Low wind speed favors the accumulation of pollutants, whereas stronger winds lead to their dilution and transport to other areas. Precipitation contributes to the removal of particles from the air, with larger PM_{10} particles typically being eliminated more effectively than finer $PM_{2.5}$ particles. Temperature and the vertical structure of the atmosphere are also important. Temperature inversions limit air mixing, promoting the accumulation of pollutants near the ground. Relative humidity can affect particle properties, increasing their mass and modifying the measurement signal. Spatial factors are also significant, such as building density, the presence of major transportation routes, and topography. PM_{10} responds more strongly to local mechanical sources, whereas $PM_{2.5}$ more often reflects the influence of regional transport [16]. As a result, $PM_{2.5}$ and PM_{10} concentrations are not a simple function of emissions but rather the outcome of a complex and nonlinear interaction among emission processes, atmospheric chemistry, and meteorological and spatial conditions. Their analysis requires consideration of temporal and spatial dependencies as well as multiple interacting environmental variables [16].

3.2 $PM_{2.5}$ and PM_{10} data characteristics

In analyses of $PM_{2.5}$ and PM_{10} concentrations, a wide range of factors could potentially be considered, such as sectoral emissions, road traffic intensity, or industrial activity. In practice, however, the availability and usability of such data are significantly limited.

Sectoral emission data are typically published as annual inventories and at an administrative scale (e.g., national or regional), which makes it impossible to directly link them with concentration measurements at an hourly resolution. Similarly, traffic intensity data are usually point-based, cover selected road sections, and do not always spatially coincide with the locations of air quality monitoring stations. As a result, the use of these variables would require numerous assumptions regarding spatial and temporal interpolation, which could increase analytical uncertainty. For this reason, the following analysis focuses on meteorological variables, which are available in a consistent format and at high temporal resolution, and can be directly integrated with PM measurement data.

3.3 OpenAQ platform

OpenAQ [30]is a global platform that collects air quality data from multiple countries and makes them available in a standardized format. It does not conduct its own measurements; instead, it integrates information from public monitoring systems, such as national environmental protection agencies and regional monitoring networks. In this way, the platform aggregates dispersed data from around the world and harmonizes them by standardizing measurement units (e.g., $\mu g/m^3$), parameter names, and time formats (most commonly UTC). OpenAQ does not modify the measurement values themselves; rather, it organizes their technical structure and makes them accessible to a broad range of users, enabling further use in analysis and research.

The data provided by OpenAQ primarily consist of real-time measurements of air pollutant concentrations, most commonly particulate matter ($PM_{2.5}$ and PM_{10}) and selected gaseous pollutants such as NO_2, SO_2, and O_3. Values are typically expressed in $\mu g/m^3$ or ppb, depending on the pollutant type, and are assigned to specific monitoring stations. Each observation is linked to a geographic location, a timestamp, and the type of measured parameter. From an analytical perspective, these data are tied to a specific station and moment in time, allowing both temporal analyses (e.g., diurnal and seasonal patterns) and spatial comparisons across different locations. The platform is open and transparent, which facilitates integration with other datasets, such as meteorological data. At the same time, users should account for potential gaps in time series, differences in reporting frequency, and changes in station operation, all of which may affect data completeness and continuity [30].

3.4 Meteostat

Meteostat is an open-source meteorological data platform based on measurements from stations operated by national meteorological services. The service does not generate its own observations; instead, it aggregates and organizes data provided by official institutions such as NOAA (National Oceanic and Atmospheric Administration) or DWD (Deutscher Wetterdienst), presenting them in a standardized format. Meteostat provides

data through several technical interfaces. These include a JSON-based API, a Python library that enables data analysis (e.g., using the Pandas library), and the option to download bulk datasets for individual meteorological stations. The range of available variables includes key atmospheric parameters, such as air temperature, relative humidity, wind speed and direction, precipitation, and atmospheric pressure. Depending on the location, additional information may also be available, such as cloud cover or visibility. Data are published, among others, at hourly resolution, enabling their direct integration with $PM_{2.5}$ and PM_{10} concentration measurements [31].

3.5 Integration of environmental data

The integration of air quality data and meteorological data is based on spatial and temporal matching of observations. In practice, this means linking measurements using the geographic coordinates of stations and their timestamps. Each PM monitoring station is assigned data from the nearest meteorological station, assuming it is representative of the given area, and observations are synchronized at the same temporal resolution–most commonly hourly.

However, it should be emphasized that meteorological stations and PM monitoring stations are not always located in the exact same place. This constitutes a simplification that may introduce additional uncertainty, particularly in areas with complex topography. The dataset used is not merely a straightforward record of environmental conditions. It is also shaped by the measurement technologies applied, the adopted procedures, and the organization of the monitoring system. Different devices may measure particulate matter slightly differently, stations are situated in specific types of locations, and data are further processed and standardized before being made available. This means that the final structure of the dataset is influenced not only by atmospheric conditions themselves, but also by the way they are observed. Predictive models therefore learn not only the relationships between meteorological conditions and particulate concentrations, but also characteristics of the measurement system from which the data originate. If the dataset contains gaps, differences between stations, or technological changes over time, the model may partially "learn" these patterns as well. A high number of observations and a standardized data format do not guarantee full comparability of all measurements. Therefore, when interpreting model results, it is important to remember that the data represent an organized record of observations rather than a perfect and entirely neutral reflection of reality.

3.6 Example of PM prediction

The analysis combined hourly $PM_{2.5}$ and PM_{10} concentration data from the OpenAQ (3.3) platform with meteorological variables obtained from Meteostat (3.4), including air temperature, relative humidity, wind speed, precipitation, and atmospheric pressure. Both datasets covered the same location. The datasets were integrated by matching ob-

servations in space and time. Each air quality monitoring station was linked to the nearest meteorological station based on geographic coordinates. The data were then aligned to the same hourly resolution, creating a single dataset for modeling.

The goal of the experiment was to build two predictive models to forecast hourly $PM_{2.5}$ and PM_{10} concentrations during the heating season. In each case, the dependent variable was the concentration of $PM_{2.5}$ or PM_{10} (in $\mu g/m^3$), while the independent variables included selected meteorological factors and time-related variables.

The models used current meteorological values as well as their time lags and moving averages (e.g., 6-hour and 12-hour averages). Calendar variables, such as hour of the day and month, were also included to capture daily and seasonal patterns.

The dataset was split chronologically into training and testing parts to reflect real forecasting conditions. Model performance was evaluated using time-series cross-validation. This method provides a more reliable estimate of predictive performance than a single train–test split while preserving the time order of observations.

The reported root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) values are averages from multiple sequential splits. The input data were not standardized or normalized because tree-based gradient boosting models, such as LightGBM, are not sensitive to the scale of input variables. Therefore, feature scaling was not required.

The prediction results are summarized in Table 2.

Table 2: Predictive performance of the LightGBM models for $PM_{2.5}$ and PM_{10} concentrations.

	RMSE [$\mu g/m^3$]	MAE [$\mu g/m^3$]	R^2
$PM_{2.5}$	4.08	2.83	0.93
PM_{10}	3.08	2.34	0.78

The model predicting $PM_{2.5}$ concentrations achieved a clearly higher R^2 value (0.93) than the model predicting PM_{10} concentrations (0.78). This suggests that $PM_{2.5}$ concentrations are more strongly influenced by the meteorological and time-related variables included in the model.

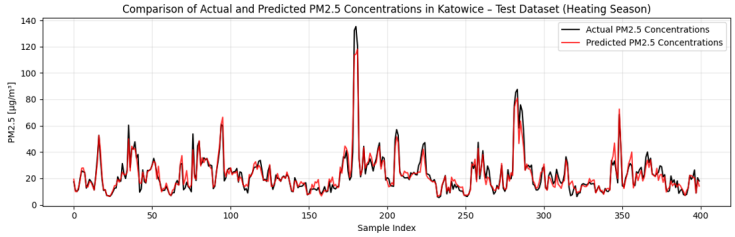

Fig. 6: Results of the LightGBM model used for the $PM_{2.5}$ prediction.

The visualization of the predicted data is presented in Figure 6. The model correctly captured the general pattern of the $PM_{2.5}$ time series, including short-term changes and periods of high concentrations typical of the heating season. However, it slightly underestimated extreme values, which may be due to the tendency of tree-based models to smooth out very high observations.

A similar experiment was carried out for the PM_{10} fraction during the heating season (see Figure 7).

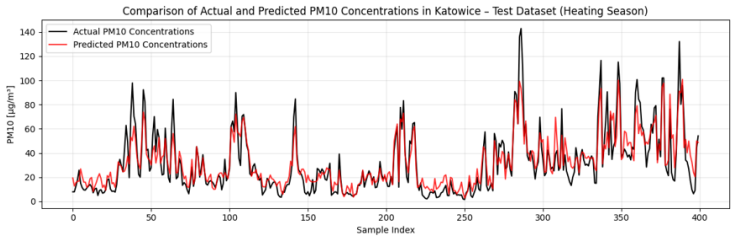

Fig. 7: Results of the LightGBM model used for the PM_{10} prediction.

The model correctly reproduced the overall pattern of the time series, including periods of higher concentrations. Compared to $PM_{2.5}$, the PM_{10} predictions showed greater short-term variability and responded more clearly to sudden increases. However, similar to the finer fraction, extreme values were still partly underestimated, especially during sharp spikes in concentrations.

4 Discussion

Data-driven analyses of greenhouse gas and particulate matter emissions have become increasingly feasible due to the availability of open-source datasets and high-resolution environmental measurements. These resources provide researchers with the opportunity to investigate emission dynamics, develop predictive models, and explore the interactions between anthropogenic activities, atmospheric processes, and meteorological conditions.

Despite their limitations, open-source CO_2 emission estimation datasets play an important role in enabling data-driven research. They provide computer scientists with accessible inputs for developing and testing machine learning models that incorporate diverse forms of activity data and use estimated emissions as target variables. As such, these datasets have contributed significantly to methodological exploration and comparative analysis in emission modeling.

At the same time, the limitations identified in this work strongly support the case for expanded and more consistent monitoring coverage. Advances in sensing technologies and large-scale data acquisition systems offer the potential to collect higher-resolution and more direct measurements of CO_2 emissions. Such data could improve our under-

standing of emission dynamics and support the development of more advanced predictive models that rely less on indirect statistical activity data and strong estimation assumptions. Improving measurement coverage and data quality is therefore a key step toward more reliable machine learning applications in carbon emission analysis.

The results highlight that the predictability of CO_2 emissions strongly depends on the economic sector. In both Slovakia and Poland, the Residential and Power sectors exhibited the highest predictability, reflecting clear, repetitive, and seasonal emission patterns. In contrast, Domestic Aviation consistently showed the weakest performance due to low variability and long periods of nearly constant emissions, which limit the information available for model learning. As described in [26], the aviation sector also tends to exhibit higher levels of error in the estimation process. Intermediate results were observed for Ground Transport and Industry, where emissions are more irregular or influenced by external factors such as production changes. A notable difference between countries was found in International Aviation: predictability was moderate in Poland but considerably higher in Slovakia, suggesting differences in sector-specific emission dynamics. Overall, these findings indicate that sector-specific modeling is essential for accurately forecasting CO_2 emissions, and that model performance is largely determined by the regularity and variability of emissions within each sector.

Similarly, the predictive modeling of $PM_{2.5}$ and PM_{10} concentrations demonstrates the crucial role of meteorological and temporal factors in shaping particulate levels. The models successfully captured general temporal patterns, including short-term fluctuations and seasonal peaks, although extreme values were somewhat underestimated. The results further suggest that finer particles ($PM_{2.5}$) are more strongly influenced by the predictors used, while both fractions benefit from the integration of high-resolution meteorological data. These findings underscore the importance of considering complex, nonlinear interactions among emissions, atmospheric chemistry, and weather conditions when modeling particulate matter concentrations.

Taken together, the analyses of both CO_2 and particulate matter highlight that predictive modeling is most effective when the underlying environmental patterns are regular, seasonal, and well-characterized. At the same time, low variability, episodic events, and irregular emission dynamics pose challenges that must be addressed through improved monitoring, data integration, and methodological refinement. These insights point to future directions for enhancing the reliability and accuracy of machine learning approaches in air quality including carbon emission research.

Acknowledgement

This work was supported by the AI2SEP project (No. 2023-1-PL01-KA220-HED-000166765). We gratefully acknowledge the project's support and funding, which enabled this research. The authors declare that they have no competing interests relevant to the content of this chapter.

References

1. World Health Organization. (2024). *Ambient (outdoor) air pollution* [Fact sheet]. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health
2. Apte, J. S., Marshall, J. D., Cohen, A. J., & Brauer, M. (2015). Addressing global mortality from ambient PM2.5. *Environmental Science & Technology*, *49*(13), 8057–8066. https://doi.org/10.1021/acs.est.5b01236
3. European Commission. (2020). *Impact assessment – Stepping up Europe's 2030 climate ambition: Investing in a climate-neutral future for the benefit of our people* (SWD(2020) 176 final). European Commission.
4. Harishkumar, K., Yogesh, K., & Gadicha, M. (2020). Forecasting air pollution particulate matter (PM2.5) using machine learning regression models. *Procedia Computer Science*, *171*, 2057–2066. https://doi.org/10.1016/j.procs.2020.04.221
5. Masood, A., & Ahmad, K. (2020). A model for particulate matter (PM2.5) prediction for Delhi based on machine learning approaches. *Procedia Computer Science*, *167*, 2101–2110. https://doi.org/10.1016/j.procs.2020.03.258
6. Kumari, S., & Singh, S. K. (2023). Machine learning-based time series models for effective CO2 emission prediction in India. *Environmental Science and Pollution Research*, *30*(55), 116601–116616. https://doi.org/10.1007/s11356-023-30514-x
7. National Centre for Emission Management (KOBiZE). (2025). *Poland. 2025 National Inventory Document (NID): National greenhouse gas inventory submission under the UNFCCC*. KOBiZE.
8. European Environment Agency. (2025). *Annual European Union greenhouse gas inventory 1990–2023 and inventory document 2025* (EEA/PUBL/2025/024). European Environment Agency.
9. Ma, N., Shum, W. Y., Han, T., & Lai, F. (2021). Can machine learning be applied to carbon emissions analysis: An application to the CO2 emissions analysis using Gaussian process regression. *Frontiers in Energy Research*, *9*, Article 756311. https://doi.org/10.3389/fenrg.2021.756311
10. Ghorbal, A. B., Grine, A., Elbatal, I., Al-Mofleh, H., & El-Saeed, A. R. (2025). Predicting carbon dioxide emissions using deep learning and Ninja metaheuristic optimization algorithm. *Scientific Reports*, *15*(1), Article 4021. https://doi.org/10.1038/s41598-025-83214-z
11. Salem, K. M., Rey-Hernández, J. M., Rey-Martínez, F. J., & Elgharib, A. O. (2025). Assessing the accuracy of AI approaches for CO2 emission predictions in buildings. *Journal of Cleaner Production*, *513*, Article 145692. https://doi.org/10.1016/j.jclepro.2025.145692
12. Al Nuaimi, H. S., Acquaye, A., & Mayyas, A. (2025). Machine learning applications for carbon emission estimation. *Resources, Conservation and Recycling Advances*, *27*, Article 200263. https://doi.org/10.1016/j.rcradv.2025.200263
13. Begum, A. M., & Mobin, M. A. (2025). A machine learning approach to carbon emissions prediction of the top eleven emitters by 2030 and their prospects for meeting Paris agreement targets. *Scientific Reports*, *15*(1), Article 19469. https://doi.org/10.1038/s41598-025-04236-5
14. European Parliament & Council of the European Union. (2008). Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal of the European Union*, *152*, 1–44.
15. Jayaratne, R., Liu, X., Thai, P., Dunbabin, M., & Morawska, L. (2018). The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog. *Atmospheric Measurement Techniques*, *11*(8), 4883–4890. https://doi.org/10.5194/amt-11-4883-2018
16. World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization.
17. United Nations Economic Commission for Europe. (2021). UN Regulation No 154 – Uniform provisions concerning the approval of light duty passenger and commercial vehicles with regards to criteria emissions, emissions of carbon dioxide and fuel consumption and/or the measurement of electric energy consumption and electric range (WLTP) [2021/2039]. *Official Journal of the European Union*, *423*, 1–210. http://data.europa.eu/eli/reg/2021/2039/oj
18. Eggleston, H. S., Buendia, L., Miwa, K., Ngara, T., & Tanabe, K. (Eds.). (2006). *2006 IPCC guidelines for national greenhouse gas inventories*. Institute for Global Environmental Strategies.

19. United Nations Framework Convention on Climate Change. (1992). *United Nations Framework Convention on Climate Change*. UNFCCC Secretariat.
20. UNFCCC Secretariat. (2025). *GHG data from UNFCCC*. https://unfccc.int/topics/mitigation/resources/registry-and-data/ghg-data-from-unfccc
21. Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Hauck, J., Olsen, A., Peters, G. P., Peters, W., Pongratz, J., Schwingshackl, C., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Anthoni, P., Barbero, L., Bates, N. R., Becker, M., ... Zhu, D. (2025). Global Carbon Budget 2024. *Earth System Science Data*, *17*(3), 965–1039. https://doi.org/10.5194/essd-17-965-2025
22. Ritchie, H., Rosado, P., & Roser, M. (2023). *CO2 and greenhouse gas emissions*. Our World in Data. https://ourworldindata.org/co2-and-greenhouse-gas-emissions
23. Directorate-General for Joint Research Centre. (2025). *EDGAR - Emissions Database for Global Atmospheric Research*. European Commission. https://edgar.jrc.ec.europa.eu/methodology
24. Crippa, M., Guizzardi, D., Pagani, F., Banja, M., Muntean, M., Schaaf, E., Monforti-Ferrario, F., Becker, W. E., & Vignati, E. (2024). *GHG emissions of all world countries* (JRC138862). Publications Office of the European Union. https://doi.org/10.2760/4002897
25. Banja, M., Crippa, M., Guizzardi, D., Muntean, M., Pagani, F., & Pisoni, E. (2025). A comparative analysis of EDGAR and UNFCCC GHG emissions inventories: Insights on trends, methodology and data discrepancies. *Earth System Science Data*, *17*(11), 6461–6486. https://doi.org/10.5194/essd-17-6461-2025
26. Solazzo, E., Crippa, M., Guizzardi, D., Muntean, M., Choulga, M., & Janssens-Maenhout, G. (2021). Uncertainties in the Emissions Database for Global Atmospheric Research (EDGAR) emission inventory of greenhouse gases. *Atmospheric Chemistry and Physics*, *21*(7), 5655–5683. https://doi.org/10.5194/acp-21-5655-2021
27. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.
28. Ke, P., Deng, Z., Zhu, B., Zheng, B., Wang, Y., Boucher, O., Arous, S. B., Zhou, C., Dou, X., Sun, T., Li, Z., Yan, F., Cui, D., Hu, Y., Huo, D., Pierre, J., Engelen, R., Davis, S. J., Ciais, P., & Liu, Z. (2022). Carbon Monitor Europe, near-real-time daily CO2 emissions for 27 EU countries and the United Kingdom. *Scientific Data*, *9*(1), Article 687. https://doi.org/10.1038/s41597-022-01761-2
29. Ke, P., Deng, Z., Zhu, B., Zheng, B., Wang, Y., Boucher, O., Arous, S. B., Zhou, C., Andrew, R. M., Dou, X., Sun, T., Song, X., Li, Z., Yan, F., Cui, D., Hu, Y., Huo, D., Chang, J.-P., Engelen, R., ... Liu, Z. (2023). Carbon Monitor Europe near-real-time daily CO2 emissions for 27 EU countries and the United Kingdom. *Scientific Data*, *10*(1), Article 374. https://doi.org/10.1038/s41597-023-02284-y
30. OpenAQ. (2025). *OpenAQ documentation*. https://docs.openaq.org
31. Meteostat. (2025). *Meteostat developer documentation*. https://dev.meteostat.net/overview
32. Gilfillan, D., & Marland, G. (2021). CDIAC-FF: Global and national CO2 emissions from fossil fuel combustion and cement manufacture: 1751–2017. *Earth System Science Data*, *13*(4), 1667–1880. https://doi.org/10.5194/essd-13-1667-2021
33. W3C. (2008). *Web content accessibility guidelines (WCAG) 2.0*. https://www.w3.org/TR/WCAG20/
34. Alam, G. M. I., Arfin Tanim, S., Sarker, S. K., Hasan, M. M., & Islam, M. S. (2025). Deep learning model based prediction of vehicle CO2 emissions with eXplainable AI integration for sustainable environment. *Scientific Reports*, *15*, Article 3655. https://doi.org/10.1038/s41598-025-87233-y
35. Andrew, R. M., & Peters, G. P. (2024). *The Global Carbon Project's Fossil CO2 Emissions Dataset* (v18) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14106218
36. Anita, W. M., Ueda, K., Uttajug, A., Seposo, X. T., & Takano, H. (2023). Association between long-term ambient PM2.5 exposure and under-5 mortality: A scoping review. *International Journal of Environmental Research and Public Health*, *20*(4), Article 3270. https://doi.org/10.3390/ijerph20043270
37. Ayaz, I. (2024). Forecasting CO2 emissions with machine learning methods: Türkiye example and future trends. *Naturengs*, *5*(2), 82–87.

About authors

Ivan Maslov is a student of Applied Computer Science at the University of Silesia in Katowice. He focuses on software development and modern IT technologies. He has experience working with digital tools, Python, as well as creating and editing multimedia content. His interests include software systems, media, and emerging technologies, combining a technical background with experience in data-driven environments.

Agnieszka Głowacka is a first-year Master's student of Micro- and Nanotechnology at the University of Silesia in Katowice. Her interests focus on the application of modern technologies in the analysis and processing of scientific data. She is particularly interested in interdisciplinary approaches combining elements of physics, chemistry, and informatics, as well as practical applications of advanced technologies in science and industry.

Bartosz Dziewit is an assistant professor at the Faculty of Science and Technology of the University of Silesia in Katowice, affiliated with the Institute of Physics, and currently serves as the Director of the Applied Computer Science program. His research focuses on particle physics (especially neutrino physics), data analysis, and computer science, and he is actively involved in teaching and supervising students in areas such as computer systems, networks, and cybersecurity

Paulina Trybek is an assistant professor at the University of Silesia in Katowice, affiliated with the Institute of Physics, where she specializes in the analysis of biomedical time series. She is actively involved in numerous student projects, supporting the development of data analysis competencies. She is also the coordinator of the project "Developing Talents in Artificial Intelligence to Solve Disruptive Environmental Problems".