



CLAQC v1.0 – Country Level Air Quality Calculator: an empirical modeling approach

Stefania Renna^{1,2,3}, Francesco Granella^{1,2,4}, Lara Aleluia Reis^{1,2}, and Paulina Schulz-Antipa⁵

¹CMCC Foundation – Euro-Mediterranean Center on Climate Change, Lecce, Italy

²RFF-CMCC European Institute on Economics and the Environment, Milan, Italy

³Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Milan, Italy

⁴Department of Social and Political Sciences, Bocconi University, Milan, Italy

⁵World Bank, Washington, DC, USA

Correspondence: Stefania Renna (stefania.renna@cmcc.it)

Received: 2 April 2024 – Discussion started: 6 May 2024

Revised: 17 December 2024 – Accepted: 21 January 2025 – Published: 22 April 2025

Abstract. The Country Level Air Quality Calculator (CLAQC) is an open-source modeling tool that utilizes national sectoral emissions and weather data to forecast monthly and annual concentrations of PM_{2.5} and O₃. CLAQC leverages the recent advancements in the Copernicus Atmosphere Monitoring Service (CAMS) system, employing CAMS global gridded emissions and CAMS reanalysis pollutant concentrations to improve the accuracy of its predictions. One of the notable strengths of CLAQC is its ability to provide country-specific and sectoral information. We have developed two methodological approaches, namely, elastic net modeling and extreme gradient boosting regressor, that can effectively predict annual average concentrations for nearly all countries. Both methods show good performance for the country's yearly average, while sensitivity tests show less robust results at the sectoral level. The tool can simulate a vast range of policy scenarios and can be integrated into national policy assessment and optimization frameworks. Finally, we present a method selection framework for each country to optimize performance and an online tool displaying model results.¹

1 Introduction

Exposure to air pollution is a significant global health concern (Murray et al., 2020), recognized by the World Health Organization (WHO) as the first environmental health risk factor (World Health Organization, 2021). In 2019, ambient air pollution was responsible for approximately 8 % of worldwide deaths, amounting to 4.51 million deaths. The majority of these deaths (92 %) were caused by fine particulate matter (particles with a diameter smaller than 2.5 µm, PM_{2.5}), while the rest were due to tropospheric ozone (O₃) (Institute for Health Metrics and Evaluation (IHME), 2019; Fuller et al., 2022).

Policies targeting energy and environmental sectors impact airborne pollutants, leading to both co-benefits and trade-offs in air pollution (Eastham et al., 2023). Therefore, it is essential to consider the impact on air pollution when designing global and national policies (Reis et al., 2022). There is a clear need for tools that quantify the impacts associated with such policies in the field of integrated assessment models (IAMs) and national to global policy scenario assessment.

IAMs are analytical tools that aim to comprehend the interactions between the earth and human systems to assist policymakers in devising effective and cost-efficient greenhouse gas and air pollution control policies. By estimating the contributions of various sectors, it becomes possible to focus resources on problematic areas and prioritize regulations or incentives that promote cleaner production or consumption practices.

¹This document is the result of a research project funded by the World Bank. The CLAQC tool was developed as an input for the IMF-World Bank Climate Policy Assessment tool (CPAT).

Considered cutting edge for their physicochemical representation detail, chemistry-transport models (CTMs) are tools for calculating the impact of emissions on pollutant concentration levels. However, they can be computationally heavy and challenging to use. Reduced-form CTMs address such limitations in trading accuracy and process detail for computational efficiency and can be used to assess multiple scenarios and incorporated into policy optimization frameworks. The most commonly used models to evaluate air pollution policies, such as the Greenhouse Gas–Air Pollution Interactions and Synergies (GAINS) model (Amann et al., 2011; Kieseewetter et al., 2015), the SHERPA tool (Thunis et al., 2016), the TM5-FAst Scenario Screening Tool (TM5-FASST) model (Van Dingenen et al., 2018), and the Air Control Toolbox (ACT) tool (Colette et al., 2022), rely on a variety of methods.²

Recent efforts to provide more policy-specific contributions to air pollution have also emerged, such as the Climate Action Planning–Air Quality (CAP-AQ) framework proposed by Kleiman et al. (2022). This approach emulates a CTM using 100 United States sites, although limited to the city level, and integrates sectoral climate-related greenhouse gas emission reductions, air quality policy, and health-related co-benefits in policy planning. Eastham et al. (2023) employ a CTM to estimate local-specific response functions that may be used to derive sector-specific contributions, although they use a lower underlying resolution than the TM5-FASST model. The Global Intervention Model for Air Pollution (Global InMAP) is a global-scale reduced-form air quality modeling tool that simulates PM_{2.5} concentrations resulting from different source heights and sites at a heterogeneous grid size (Thakrar et al., 2022). The latter one is the most detailed, up-to-date reduced-form air pollution model on a global scale. Yet, even at its coarser temporal and spatial resolution, its runtime is still not compatible with an easy-to-implement fast policy assessment. Other studies have compared several air pollution health impact assessment tools (Anenberg et al., 2016), applied similar reduced modeling methods for local-scale air pollution modeling (Oxley et al., 2022), or assessed the sectoral and fuel-specific contributions of PM_{2.5} to mortality at various geographic scales (McDuffie et al., 2021) using CTMs in simulation mode. However, only the Response Surface Model by Eastham et al. (2023) is able to simultaneously provide annual sectoral atmospheric contributions to both PM_{2.5} and O₃ concentrations at the national and country level with global coverage.

All in all, the available reduced-form CTMs allow for the assessment of multiple scenarios, ultimately contributing to

better policy design. However, they are based on CTMs and are therefore bounded by their resolution and scope. They are also known to be less robust for highly non-linear processes, such as secondary O₃ formation and secondary PM formation (Van Dingenen et al., 2018; Thunis et al., 2019). Furthermore, they are limited by the number of underlying scenarios in their training set. The Country-Level Air Quality Calculator (CLAQC) is a statistical model that aims at filling this gap by learning from coupled historical variations in concentrations of PM_{2.5} and O₃, sectoral emissions of multiple precursors, and meteorology. Similarly to deterministic reduced-form models, empirical models are also bounded but by the underlying observed data. Unlike in the source–receptor models, the perturbation level is not set (Thunis et al., 2019) as the model learns from all past variations. Provided with adequate data, a statistical model can learn from a broad spectrum of variations in emissions, concentrations, and atmospheric conditions akin to simulating a large number of training scenarios in process-based models. CLAQC is built over 19 years of data spanning the entire global land-mass and is provided with arguably a wide spectrum of input variables. The large drop in emissions induced by COVID-19 lockdowns and disruptions worldwide further expands the range of training variables. Additionally, CLAQC allows for more flexibility than the deterministic reduced-form models by relying on the global gridded Copernicus Atmosphere Monitoring Service (CAMS) reanalysis products. As new and better data come in every year, the emulator can be updated, and a higher detail level may be possible at lower trade-off costs. For example, the sub-national detail can assist with better placement of energy, transport, and industrial infrastructure. CLAQC complements the above-mentioned models by providing an easy-to-use global tool with country and sectoral details.

The next section discusses the data used. We then present in Sect. 2.5 the different methodological approaches that have been followed: elastic net models (Sect. 2.8), and machine learning models (Sect. 2.9), while in Sect. 2.10, we compare the two methodological approaches. In Sect. 3, we present the validation of results and sensitivity analyses. In Sect. 4, we discuss the limitations of our tool. Finally, we draw conclusions in Sect. 5.

2 Data

Monitoring stations are used for air quality assessment. However, the lack of a spatially consistent large ground monitoring network in a given area is a strong constraint to achieving this objective. Despite the recent harmonization and open-access advancements in air pollution data, most publicly available global ground-level monitoring databases (e.g., OpenAQ, 2024) provide reasonable territorial coverage of the population only in developed countries, in particular in the United States and Europe. The network of ground-level

²For example, GAINS is a fully IAM using emissions–concentrations relationships from the Unified EMEP Eulerian model and CHIMERE. Similarly, the SHERPA tool emulates scenario results from EMEP and CHIMERE (Menut et al., 2021), while TM5-FASST is a source–receptor model based on the chemical transport model TM5 (Van Dingenen et al., 2018).

monitors is growing in emerging economies such as China and India, yet urban and rural areas are largely unmonitored in middle- and low-income countries. The uneven ground-level monitoring geographical coverage is problematic, as factors driving the emissions–concentrations relationship differ between monitored and unmonitored areas (e.g., population density, distance from industrial sources, and gross domestic product per capita). Emissions-to-concentrations relationships learned using ground-level monitoring yield estimates that are biased toward richer countries.

To meet the CLAQC objectives of global coverage, an alternative option to ground-level monitoring data is using reanalysis data. Global, gridded reanalysis data combine and harmonize satellite air pollution measurements and CTM output with ground-level monitors. While maintaining the quality of the monitor data at the location of the monitors, they bridge the gap in ground-level monitoring networks with satellite observations and models that span the entire globe. This principle, called data assimilation, is based on the method used in numerical weather prediction and air quality forecasting, where a previous forecast is combined with newly available observations in an optimal way to produce a new best estimate of the state of the atmosphere. Reanalysis does not have the constraint of timely forecasts, allowing for time to collect observations and allowing for the integration of improved versions of the original observations, raising the quality of the reanalysis product.

Using gridded data has many other advantages, allowing for

- weighting reductions in concentrations by population, obtaining changes in exposure to pollutants;
- better identifying the interactions of emissions with meteorology and topography;
- increasing statistical power without compromising the estimation of country-specific emissions-to-concentrations functions;
- reducing rigidity on the spatial scope (not limited to administrative borders) and keeping the sub-national modeling option flexible;
- global coverage, even in areas without ground-level monitoring.

Among the disadvantages, the need to homogenize different grids in terms of spatial resolution may lead to approximations during the data manipulation process.

2.1 Emissions

2.1.1 Precursors

The precursors of $\text{PM}_{2.5}$ included in the models are BC, OC, NH_3 , NO_x , SO_2 , and NMVOC. All are expected to increase $\text{PM}_{2.5}$ concentrations at the country level, although

local decreases on the secondary fraction may happen (Clapier et al., 2021). Data on OC and BC are almost perfectly collinear: thus, emissions from these precursors are summed into total carbon (TC) in the machine learning models where both are available.

The precursors of O_3 included in the models are its main precursors: NMVOC, NO_x , and SO_2 . NMVOC are expected to increase O_3 , whereas the relationship between NO_x and SO_2 to O_3 may be negative (Van Dingenen et al., 2018). We follow the approach of most reduced-form models, leaving out the CO precursor for its relatively minor importance (Amann et al., 2011).

2.1.2 CAMS emissions

Emission data are provided by CAMS Global Anthropogenic (CAMS-GLOB-ANT) v5.3, with a monthly temporal resolution and a spatial resolution of 0.1° . The data are originally expressed in teragrams (Tg) and are converted into kilograms (kg). The CAMS emission data are based on existing available databases, including nationally reported emissions, the Joint Research Centre's (JRC) Emissions Database for Global Atmospheric Research (EDGAR) (Huang et al., 2017; Crippa et al., 2018), the Evaluating the Climate and Air Quality Impacts of Short-Lived Pollutants (ECLIPSE) project (Stohl et al., 2015), and the Community Emissions Data System (CEDS) databases (Hoesly et al., 2018; McDuffie et al., 2020). It has the advantage of providing global gridded monthly emissions from 2000 up to 2021, although emission estimates of most recent years (2015–2021) are extrapolated by applying CEDS 2014–2019 country-level trends to gridded EDGAR v5 data and therefore are associated with higher uncertainty. See Denier van der Gon et al. (2023); Granier et al. (2019) for more detail on CAMS-GLOB-ANT data harmonization and sectoral definitions.

The CAMS emission inventory is based on business-as-usual emissions and does not take into account lockdown measures and restrictions put in place to tackle the COVID-19 pandemic. To correct this, we apply to 2020 emissions the COVID-19 adjustment Factors for eMissions (CONFORM) constructed by Doumbia et al. (2021) for the following sectors: power, industry, residential, public and commercial, and transport.³

2.1.3 DACCIWA emissions

Data quality in low-income countries can be comparably poorer due to the scarcity of measurements. Therefore, we additionally consider the Dynamics–Aerosol–Chemistry–Cloud Interactions in West Africa (DACCIWA) emission data set as model input, an Africa-specific regional emission inventory developed for providing more accurate estimations for African countries, employing updated emission factors

³CONFORM data were downloaded from the ECCAD portal at <https://eccad.aeris-data.fr> (last access: 27 March 2025).

based on in situ measurements (Keita et al., 2021). DAC-CIWA covers major human-related emission sources characterizing the African continent, such as charcoal production, wood stove combustion, and open-air garbage combustion, and classifies emissions into the following sectors: traffic, energy, residential, industry, other, and waste. It has the same spatial resolution as CAMS emissions and covers the period from 1990 to 2015.

2.1.4 Sectoral aggregation

Our focus is on identifying sectors that are likely to be directly impacted by policies aimed at reducing the use of fossil fuels. However, emissions from the various sectors, with the exception of the agricultural one, are highly collinear, making it difficult to distinguish the contribution of each individual sector to the total pollution levels (see Sect. A1 for further details). To overcome this challenge, we group together sectors with similar emission patterns, reducing the complexity of the data and improving the accuracy of our analyses. At the same time, we try to keep sectoral relevance for policy models such as the IMF–World Bank Climate Policy Assessment Tool (CPAT) (Black et al., 2023). We do not include biogenic and sectoral emissions from shipping and aviation. Furthermore, natural emissions, such as desert dust and sea salt, are not taken into account since they are less likely to be subject to policy interventions. We build the following seven sectors from CAMS-GLOB-ANT data: agriculture, industry, other (including the emissions not considered in the other sectors), off-road transportation, energy power generation, road transportation, and residential (including buildings, commercial, and services). Note that the off-road transportation sector includes railways and other types of non-road transports not typically used on public roads, such as agricultural machinery, construction equipment, and certain types of off-road vehicles used in industrial operations (e.g., tractors, telehandlers, and excavators). See Table 1 for the correspondence between the CAMS-GLOB-ANT sectors and the CLAQC ones.

Note that CAMS and DACCIWA apply slightly different sectoral classifications. For instance, we retain two transport-related sectors from CAMS after aggregation, whereas DAC-CIWA only includes one transport sector. Moreover, the final does not consider agriculture in the original data. As a result, we aggregate the available sectors into the following categories: transport, power, industry, residential, and other (with the latter one containing waste as well).

2.2 Meteorology

All meteorological data, with the exception of wind direction, comes from TerraClimate (Abatzoglou et al., 2018). TerraClimate has a wide variety of meteorological variables, good temporal coverage (1958 to 2022), high spatial resolution ($4\text{ km} \times 4\text{ km}$), and monthly time resolution. The follow-

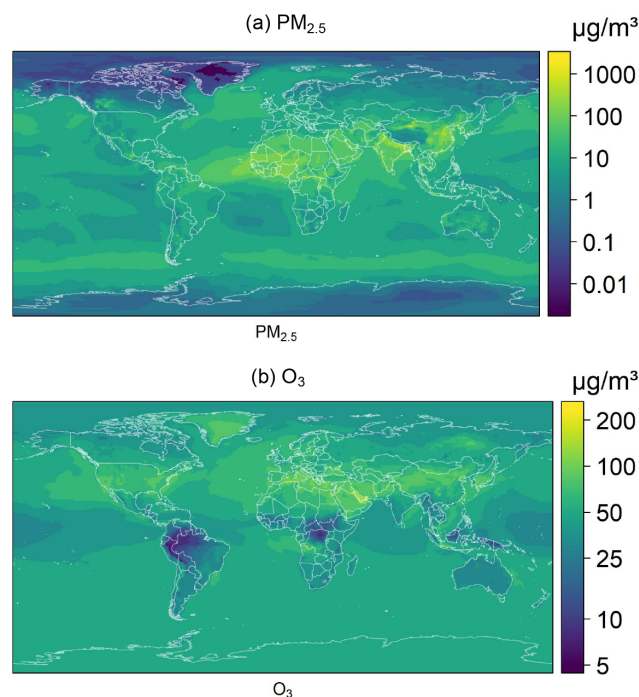


Figure 1. Level plots of EAC4 concentrations of $\text{PM}_{2.5}$ (January 2018) and O_3 (July 2018) in $\mu\text{g m}^{-3}$ with a color bar.

ing atmospheric variables are used as inputs to the models: accumulated precipitation in millimeters, maximum 2 m temperature in degrees Celsius ($^{\circ}\text{C}$), minimum 2 m temperature in $^{\circ}\text{C}$, 10 m wind speed in m s^{-1} , and mean vapor pressure deficit in kilopascals. The wind direction in degrees comes from the European Centre for Medium-Range Weather Forecast's (ECMWF's) ERA5 Reanalysis Monthly Means product by Copernicus Climate Change Service (C3S) (Copernicus Climate Change Service, 2019).

2.3 Concentrations

We obtain the ground-level ambient concentration data for $\text{PM}_{2.5}$ and O_3 air pollutants from ECMWF's Atmospheric Composition Reanalysis 4 (EAC4) monthly averaged fields (Inness et al., 2019). The data cover the period from 2003 to 2021, which is the shortest time domain of all the available datasets. Consequently, all other datasets are limited to this time period. The original spatial resolution of the data is 0.75° , which is downscaled to 0.5° . $\text{PM}_{2.5}$ and the mixing ratio of surface-level ozone (obtained from the GEMS ozone model level 60) are originally expressed in kg m^{-3} and kg kg^{-1} , respectively. To facilitate analysis, we convert them to micrograms per cubic meter ($\mu\text{g m}^{-3}$). As an example, Fig. 1 shows EAC4 concentration levels of $\text{PM}_{2.5}$ and O_3 for January 2018 and July 2018, respectively.

To transform country-level, monthly $\text{PM}_{2.5}$ concentrations into population-weighted exposure, $\text{Exp}_{k,m}$, for country k and month m , we use the 2020 UN World Popula-

Table 1. Correspondence between CAMS-GLOB-ANT and CLAQC sectoral aggregation.

CAMS-GLOB-ANT sector	CLAQC sector
Agriculture livestock (AGL) Agriculture soils (AGS) Agriculture waste burning (AWB)	Agriculture
Industrial processes (IND)	Industry
Solvent application and production (SLV) Solid waste and wastewater handling (SWD)	Other, including the emissions not considered in the other sectors
Non-road transportation (TNR)	Off-road transportation
Power generation (ENE) Fugitive emissions from solid fuels (FEF) Refineries (REF)	Energy power generation
Road transportation (TRO)	Road transportation
Residential (SER)	Residential, including buildings, commercial and services

tion Prospects (WPP) Adjusted Population Count, v4.11, at 30 arcsec spatial resolution, from the Center for International Earth Science Information Network (CIESIN) (Center For International Earth Science Information Network-CIESIN-Columbia University, 2018). We calculate monthly country-level exposure, $\text{Exp}_{k,m}$, by summing over grid cells i population weights, $\frac{\text{pop}_i}{\text{pop}_k}$, and multiplying the sum by the grid-level, monthly concentrations, $C_{i,m}$ (Eq. 1). We only use population data referring to one year, 2020, to avoid introducing another source of variation in the models. To give a sense of the data, we display 2018 country-level weighted concentrations of $\text{PM}_{2.5}$ and O_3 in Fig. 2.

$$\text{Exp}_{k,m} = \sum_{i=1}^n \frac{\text{pop}_i}{\text{pop}_k} \cdot C_{i,m} \quad (1)$$

2.4 Grid definition

All gridded data sources are re-scaled to the same ($0.5^\circ \times 0.5^\circ$) coordinate grid through linear interpolation based on the population grid and merged into a single dataset. For instance, concentration data originally at $0.75^\circ \times 0.75^\circ$ spatial resolution are downscaled to $0.5^\circ \times 0.5^\circ$, generating intermediate values that align with the reference grid. The interpolation is implemented using the `interp_like` function from the `xarray` Python package. Notice that some cells may be attributed to multiple countries in case the centroid falls exactly on the countries' borders. This should not be a source of concern as models are independently run country by country.

2.5 Methods

2.6 CLAQC rationale

We are interested in estimating the relationship between emission E of major ambient air pollutants and the respective ground-level concentration C of major pollutants c ($\text{PM}_{2.5}$, O_3). Denote such relationship f , so that

$$C_c = f(E). \quad (2)$$

The formation, transport, and dispersion of pollutants are complex natural phenomena that are highly dependent on emissions; weather, W ; and other local characteristics, such as topography. Hence, the design of pollution abatement policies in country k can benefit from the estimation of a country-specific emissions-to-concentrations function (a country-wide population-weighted average) that accounts for the interplay between emissions and weather:

$$C_{c,k} = f_k(E_k, W_k). \quad (3)$$

However, environmental and fiscal policies have heterogeneous effects across the main sectors of emissions and precursors, for instance, by inducing a rearrangement in the energy mix. Therefore, it is helpful to establish how country-wide changes in the emissions of precursor p , from a given sector s , alter ambient concentrations. The emitting sector s includes energy production, industries, buildings, transport, and agriculture. The set of pollutant precursors p includes black carbon (BC), organic carbon (OC), ammonia (NH_3), non-methane volatile organic compounds (NMVOC), nitrogen oxides (NO_x), and sulphur dioxide (SO_2). We are thus interested in estimating the following relationship:

$$C_{c,k} = f_{k,s,p}(E_{k,s,p}, W_k). \quad (4)$$

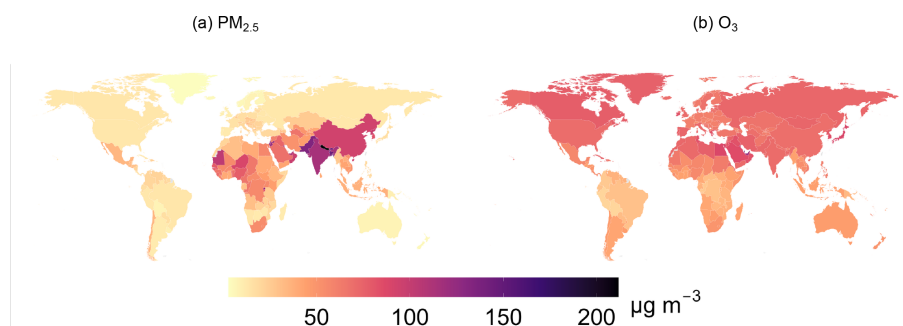


Figure 2. EAC4 concentration inputs of PM_{2.5} weighted by the population and O₃ in µg m⁻³ aggregated at the country level (2018).

We identify two methods to empirically derive $\hat{f}_{k,s,p}$, that trade off simplicity and transparency with prediction power, as we explain in what follows. The first method relies on elastic net models, a penalized linear regression amenable to a large number of predicting variables while preserving an intelligible structure. $\hat{f}_{k,s,p}$ is modeled as a linear function of emissions and weather variables that can be easily reproduced.

The second method relies on machine learning algorithms that are better suited than linear models to learn highly non-linear relationships, such as those between precursors and weather conditions. Better performance comes, however, at the cost of interpretability, as machine learning algorithms typically do not return simple predictor–target functions. For this reason, we also provide approximate emissions-to-concentrations relationships with functions that are suitable for simpler spreadsheet-style use.

We follow two approaches that trade off interpretability and predictive performance. We first estimate an Elastic Net model, a linear model with selection and shrinkage. The linear form allows for easy interpretation of coefficients, whereas selection and shrinkage give more weight to the variables of the highest importance and address multicollinearities in the data. Second, we fit an extreme gradient boosting regressor, a decision tree-based machine learning algorithm. Figure 3 shows the schematic representation of the CLAQC workflow.

2.7 Coefficient constraints

We impose monotonic constraints on certain model coefficients to align with expected physicochemical relationships. These constraints specify how input variables should affect the target, ensuring interpretable and physically plausible results. For instance, a positive monotonic constraint enforces a non-negative relationship, ensuring that as an input variable increases, the predictor output does not decrease.

In the presence of noise, complex interactions in the data, or predictor cross-correlation, models may otherwise learn patterns that are not realistic or physically plausible. Additionally, monotonic constraints help prevent overfitting, en-

hancing robustness when input data are limited or uncertain. For example, it is not expected that an increase in BC emissions would lead to a decrease in PM_{2.5} concentrations.

While at the local scale, reducing certain precursors of secondary inorganic aerosols might not always lead to a decrease in PM_{2.5} levels – due to non-linear atmospheric reactions noted by Thunis et al. (2019) and Ding et al. (2021) – our national-scale models focus on broader trends. To avoid giving undue importance to cases where local emission reductions might result in increased levels of inorganic PM_{2.5}, we apply monotonic constraints between emissions and concentrations.

Rather than directly including secondary inorganic aerosols, the models incorporate interactions between PM precursors – specifically NH₃, NO_x, and SO₂ – as proxies for secondary reactions.

It is crucial to understand that in situations where secondary reactions substantially affect the overall mass of PM_{2.5} within a country, our models are designed to omit these precursors from the list of predictors, thereby not reflecting a decrease in PM_{2.5} levels.

We further require that greater precipitations and temperatures decrease PM_{2.5}. Precipitation lowers PM_{2.5} by wet deposition, while temperature is a proxy for inversion layer height; i.e., high temperature generally means high inversion layer heights and therefore a lower concentration (Seinfeld, 2016). Although wind speed normally facilitates pollutant dispersion, we impose no constraint on its role as long-distance transportation of suspended particles may increase PM_{2.5}. All other coefficients are unbound.

Regarding O₃, similarly, we impose that emissions of NMVOC increase its concentrations while leaving emissions of NO_x unconstrained, allowing for non-monotone relationships with O₃ (Ding et al., 2021). We also constrain temperature to increase O₃ concentrations (Jhun et al., 2015; Lu et al., 2019). O₃ is a photo-chemical secondary pollutant (Seinfeld, 2016), which increases with intensifying solar radiation. Temperature is therefore used as a proxy.

We include the following variables in both models: sectoral emissions, precipitation, minimum temperature, maximum temperature, vapor pressure deficit, wind speed, and

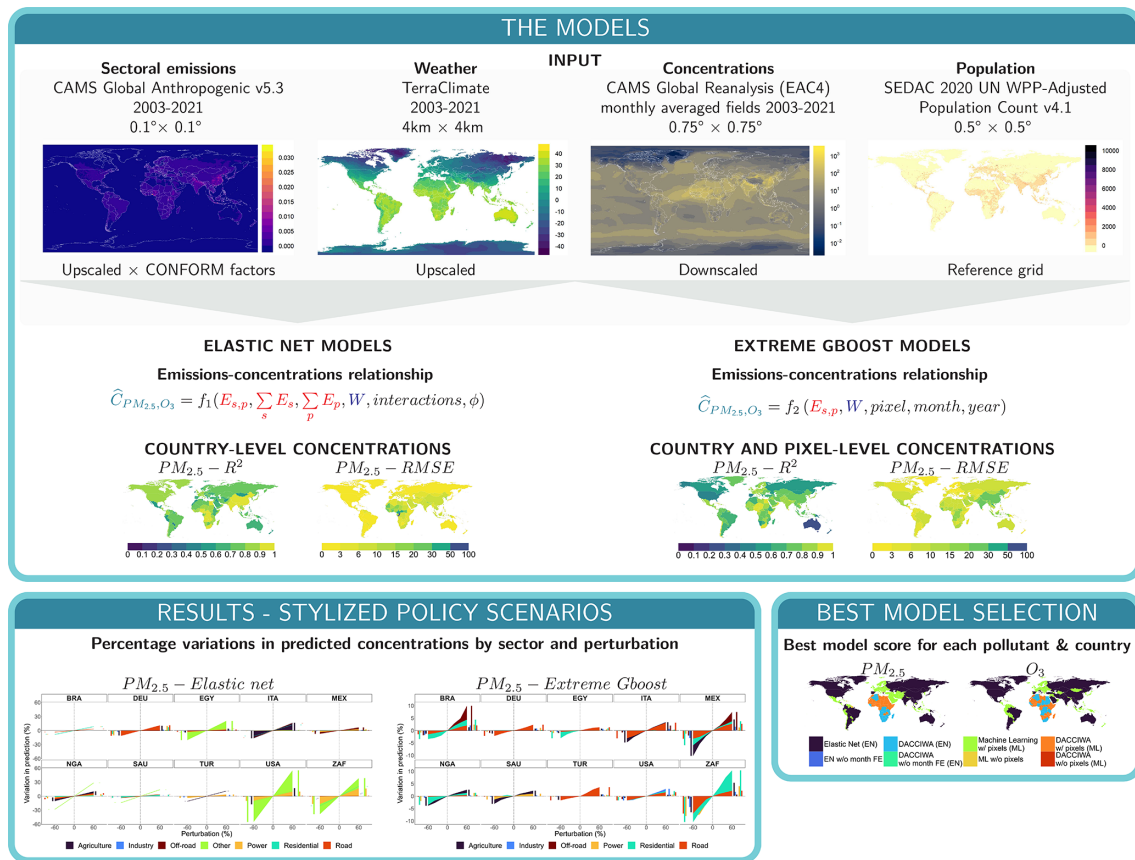


Figure 3. Methodological abstract.

wind direction. In the case of EN, we also add monthly emission sectoral totals (i.e., $\sum_s E_{k,s,m}$) and monthly emission pollutant totals (i.e., $\sum_p E_{k,p,m}$) to increase the chances that the models capture variations in emissions as well as monthly fixed effects and interaction terms (see Sect. 2.8 for further details on EN model specifications).

2.8 Elastic net models

Due to the high multicollinearity among predictors, as shown in Appendix A1, ordinary least-squares (OLS) regression may fail to yield reliable parameter estimates. Penalized linear regression maintains the interpretability of coefficients of linear models while selecting the variables with the greatest predictive power. We use elastic net models (Zou and Hastie, 2005), a method suitable for identifying the subset of best predictors obtaining a parsimonious model. It solves the following minimization problem for the model parameters β_0 and β , where β_0 is the model's intercept and β represents the coefficients of the input variables:

$$\min_{\beta_0, \beta} \frac{1}{2N} \sum_{i=1}^N (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \left[\frac{(1-\alpha) \|\beta\|_2^2}{2} + \alpha \|\beta\|_1 \right]. \quad (5)$$

Combining the penalty elements of the Least Absolute Selection and Shrinkage Operator (LASSO) regression ($\|\beta\|_1$) and Ridge regression ($\|\beta\|_2^2$) on the basis of the alpha (α) parameter, the penalization parameter lambda (λ) selects variables like the former and shrinks them as it does the latter. It regularizes the model coefficients, improving the model's accuracy and interpretability by decreasing the input variables' space. This prevents our models from being volatile to extreme variations and outliers. Such a technique avoids large errors on the one hand, and, on the other, it results in more conservative estimations of the concentrations obtained from the emission reductions.

We perform elastic net modeling in R statistical language (R Core Team, 2020), version 4.0.2 (22 June 2020), on 64 bit Windows 10 (build 22621). To allow reproducing the R environment, we employ the `renv` package (Ushey, 2022). The elastic net workflow is represented in Fig. 4 and follows the steps below. For each country, the steps are as follows:

1. To ensure reproducibility, a seed is set with the `set.seed` R function.

2. We average the gridded monthly dataset to the country–year–month level.⁴
3. We then identify and exclude outliers by applying the interquartile range rule and listwise deletion.
4. We randomly split the 2003–2021 data into training (84 % of observations) and test sets (16 %) stratifying by month.⁵ We run sensitivity tests on the splitting ratio, obtaining robust results across splittings: for further insights, see Sect. A5.
5. We apply a k -fold cross-validation algorithm for tuning the λ regularization parameter using the `cv.glmnet` function from the `glmnet` R package (Friedman et al., 2010). We apply the following specifications: 30 folds, $\alpha = 0.5$ corresponding to elastic net regularization with no optimization of the alpha parameter, 'deviance' type.measure for specifying the mean squared error loss function, and 'gaussian' family (Friedman et al., 2020).
6. Monotonic constraints are imposed for certain predictors. See details in Sect. 2.7.
7. We train the model on the training set by applying the `glmnet` function from the `glmnet` R package.⁶
8. We evaluate its performance on the test set, i.e., on data not used to build the model itself. We report the out-of-sample R squared (R^2) and root mean square error (RMSE), calculated as in Eqs. (6) and (7). y_i^{test} is the test set actual value for observation i , $\hat{y}_i^{\text{test}}$ is the test set predicted value for observation i , $\bar{y}^{\text{test}}$ is the mean value of the test set actual values, and n_{test} is the number of observations in the test set.
9. Finally, we predict concentrations for varying levels of emissions and derive empirical emissions-to-concentrations relationships. More specifically, we simulate perturbations of emissions from -60% to $+60\%$ at 10% steps based on the last 5 years of data. This time-frame is selected to reflect recent trends, offering more policy-relevant insights into the empirical relationship between emissions and concentrations. Notice that the user could choose another time period for simulations.

⁴While we sum up sectoral emissions, we average weighted concentrations and meteorology variables. We treat wind direction as a circular variable through the `circular` function from the `circular` R package (Agostinelli and Lund, 2022).

⁵This is done using the `stratified` function from the `splitstackshape` R package (Mahto, 2019). Note that an 80–20 train–test splitting is applied for models based on DACCIWA emissions.

⁶Notice that the `glmnet` function standardizes by default all the variables, removing the influence of their scales.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i^{\text{test}} - \hat{y}_i^{\text{test}})^2}{\sum_{i=1}^n (y_i^{\text{test}} - \bar{y}^{\text{test}})^2} \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i^{\text{test}} - \hat{y}_i^{\text{test}})^2}{n_{\text{test}}}} \quad (7)$$

The elastic net linear regression models take the following form for each country (Eqs. 8 and 9) (Seinfeld, 2016):

$$\begin{aligned} \text{PM}_{2.5t} = & \alpha + \sum_{s,p_1} \beta_{s,p_1} E_{s,p_1,t} + \gamma_1 \text{PPT}_t + \gamma_2 \text{TMIN}_t \\ & + \gamma_3 \text{TMAX}_t + \gamma_4 \text{VPD}_t + \gamma_5 \text{WS}_t + \gamma_6 \text{WD}_t + \\ & + \sum_s \delta_s E_{s,t} + \sum_{p_1} \lambda_{p_1} E_{p_1,t} + \mu E_{\text{NO}_x,t} \cdot E_{\text{NH}_3,t} \\ & + \nu E_{\text{SO}_2,t} \cdot E_{\text{NH}_3,t} + \xi E_{\text{SO}_2,t} \cdot E_{\text{NO}_x,t} + \\ & + \sum_s \theta_s E_{s,t} \cdot \text{WS}_t \cdot \text{WD}_t + \phi_t + \varepsilon_t, \end{aligned} \quad (8)$$

$$\begin{aligned} \text{O}_{3t} = & \alpha + \sum_{s,p_2} \beta_{s,p_2} E_{s,p_2,t} + \gamma_1 \text{PPT}_t + \gamma_2 \text{TMIN}_t \\ & + \gamma_3 \text{TMAX}_t + \gamma_4 \text{VPD}_t + \gamma_5 \text{WS}_t + \gamma_6 \text{WD}_t + \\ & + \sum_s \delta_s E_{s,t} + \sum_{p_3} \lambda_{p_3} E_{p_3,t} + \mu E_{\text{NO}_x,t} \cdot E_{\text{NMVOC},t} \\ & + \nu E_{\text{SO}_2,t} \cdot E_{\text{NMVOC},t} + \xi E_{\text{SO}_2,t} \cdot E_{\text{NO}_x,t} + \\ & + \sum_s \theta_s E_{s,t} \cdot \text{WS}_t \cdot \text{WD}_t + \phi_t + \varepsilon_t, \end{aligned} \quad (9)$$

where

- $s \in \{\text{agriculture, industry, other, off-road transportation, energy power generation, road transportation, residential}\}$
- $p_1 \in \{\text{BC, NH}_3, \text{NMVOC, NO}_x, \text{OC, SO}_2\}$
- $p_2 \in \{\text{NMVOC, NO}_x\}$
- $p_3 \in \{\text{NMVOC, NO}_x, \text{SO}_2\}$
- $\text{PM}_{2.5t}$ is the monthly concentration of $\text{PM}_{2.5}$ in $\mu\text{g m}^{-3}$ (population-weighted)
- O_{3t} is the monthly concentration of O_3 in $\mu\text{g m}^{-3}$
- $E_{s,p_1,t}$ is the monthly emissions of sector s and pollutant p_1 in kilograms
- $E_{s,p_2,t}$ is the monthly emissions of sector s and pollutant p_2 in kilograms
- $E_{s,p_3,t}$ is the monthly emissions of sector s and pollutant p_3 in kilograms
- PPT_t is the monthly accumulated precipitation in millimeters

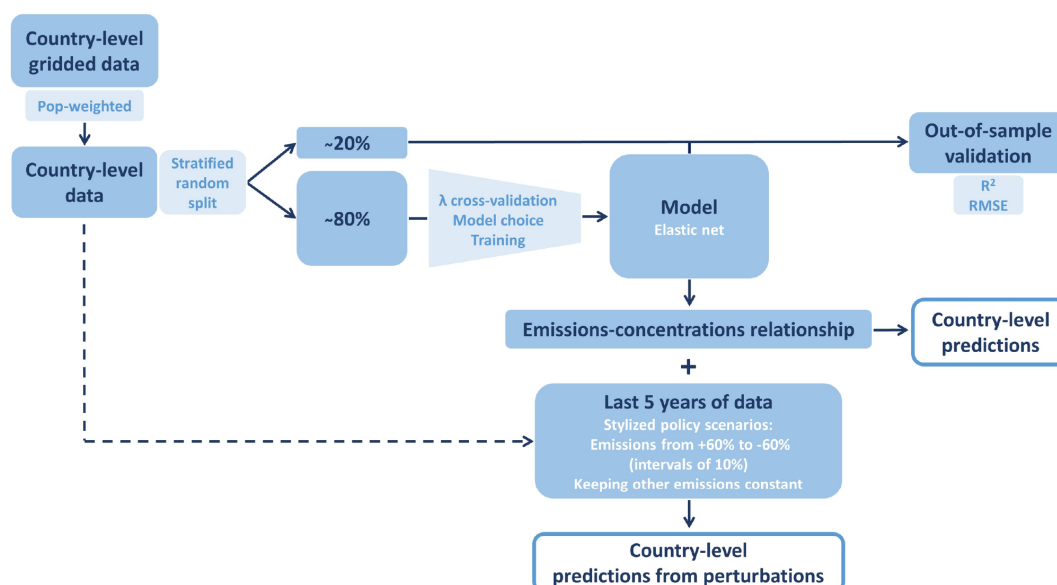


Figure 4. Visual representation of the elastic net models' workflow.

- $TMIN_t$ is the monthly minimum 2 m temperature in $^{\circ}C$
- $TMAX_t$ is the monthly maximum 2 m temperature in $^{\circ}C$
- VPD_t is the monthly mean vapor pressure deficit in kilopascals
- WS_t is the monthly 10 m wind speed in $m s^{-1}$
- WD_t is the monthly wind direction in degrees
- $E_{p1,t}$ is the monthly composite index from the sum of total emissions of pollutant p_1 in kilograms
- $E_{p2,t}$ is the monthly composite index from the sum of total emissions of pollutant p_2 in kilograms
- $E_{p3,t}$ is the monthly composite index from the sum of total emissions of pollutant p_3 in kilograms
- $E_{s,t}$ is the monthly composite index from the sum of total emissions of sectors s in kilograms
- ϕ_t is the monthly fixed effects
- ε_t is the error term.

In Eqs. (8) and (9), t indicates time, s is the emission sector, and $p_{[n]}$ refers to the sector-related emitted pollutants in their respective models. $PM_{2.5}$ and O_3 concentration values obtained from the models in $\mu g m^{-3}$ are country-level monthly concentration averages indexed by time t , just as all the other parameters in the equation; notice that $PM_{2.5}$ levels are weighted by population as explained in Sect. 2.3. α is the

model intercept; λ , β , γ_i , δ , μ , ν , ϵ , and θ are the predictors' coefficients; $E_{s,p[n]}$ are emissions of sector s and pollutant $p_{[n]}$, respectively; and E_p and E_s are total emissions of pollutant p and of sector s , respectively. All emission variables are expressed in kilograms. PPT_t stands for accumulated precipitation in millimeters; $TMIN_t$ and $TMAX_t$ are minimum 2 m temperature and maximum 2 m temperature, respectively, in $^{\circ}C$; VPD_t is mean vapor pressure deficit in kilopascals; WS_t is 10 m wind speed in $m s^{-1}$; WD_t is average wind direction in degrees; ϕ_t are month fixed effects; and finally, ε_t is the stochastic term. Note that the emission terms in the equations differ due to their different atmospheric reactions. In both equations, we include multiple emission terms to increase the chances that models capture variations in emissions. In Eq. (8), to model the secondary inorganic aerosol formation, we interact total emissions of NO_x and NH_3 , SO_2 and NH_3 , and NO_x and SO_2 , respectively. Similarly, in Eq. (9), we interact total emissions of NO_x and $NMVOC$, SO_2 and $NMVOC$, and SO_2 and NO_x . As before, this attempts to capture the reactions between the precursors of O_3 since the presence of at least two of these precursors is necessary for its formation. While $NMVOC$ and NO_x are O_3 main precursors, reacting in the presence of solar radiation, SO_2 plays an indirect role in O_3 formation (Baird and Cann, 2013; Seinfeld, 2016). SO_2 is typically emitted by industrial sources. It is involved in secondary PM formation, which can reduce the radiative properties and oxidative capacity of the atmosphere, indirectly affecting O_3 formation. In both equations, we also create an interaction between sectoral emissions and wind speed and direction to proxy transport and dispersion of pollutants. We include total sectoral emissions to reflect the fact that sector-specific policies typ-

ically impact multiple pollutants through dedicated emission offset protocols. Additionally, we consider total emissions from individual pollutants because variations in total pollutant emissions may result from not only specific sectors but also inter-sector changes, transported emissions, and chemical reactions. Refer to Sect. A2.2 for the EN model specifications with DACCIWA emissions and to Sect. A4 for EN model implementation.

In order to have non-negative predicted values for y , we add a non-negativity constraint that selects the maximum value between zero and the EN model prediction, $\hat{y}_{\text{EN}}$. Moreover, in order to have non-extreme predicted values for y , e.g., due to input data divergence, we apply a second safety function that caps predicted values to 3 times the observed country-level concentrations under no perturbations, $3y$:

$$\hat{y}^* = \min\{\max\{0, \hat{y}_{\text{EN}}\}, 3y\}. \quad (10)$$

That is, the final prediction $\hat{y}^*$ is equal to 0 if $\hat{y}_{\text{EN}} < 0$, and equal to $\hat{y}_{\text{EN}}$ or at most $3y$ if $\hat{y}_{\text{EN}} \geq 0$.

2.9 Machine learning models

Emissions-to-concentrations functions might not be sufficiently well approximated by a linear function due to the nonlinearities of topography and secondary pollution formation (Thunis et al., 2019). Machine learning models are powerful tools that can reproduce highly non-linear relationships such as the complex natural phenomena behind air pollution formation, transport, and dispersion. Importantly, they do not require the user to impose a functional form. Differently from the modeling with elastic net, data are not spatially aggregated. In addition to the pre-processing steps described in Sect. 2, we perform specific data processing.

1. Given the very high level of collinearity between BC and OC emission data, we sum the two precursors into a variable called total carbon (TC).
2. Emissions from the other sector are excluded. These emissions are frequently missing or otherwise highly correlated with other emissions. Moreover, their informative content is very low.
3. In addition to year and month of the year, we include an identifier of grid cells as a predictor variable.

We use extreme gradient boosting regressor (Chen and Guestrin, 2016), a tree-based algorithm that has been shown to perform very well in supervised tasks with structured data (e.g., Ma et al., 2020, in the context of air pollution).

The process is represented in Fig. 5 and is as follows.

1. For each country–pollutant pair, the input data for ML are a grid panel dataset, composed of (N) grid cells and observed over (T) time periods. For each grid cell, we randomly assign the T observations (a time series) to the

train or test set. Hence, we stratify by grid cell, and randomization occurs over the temporal dimension. This stratified randomization ensures equal spatial representation in both datasets. Given unobservable but time-constant characteristics of cells (such as topography) and the desire for equal spatial representation, we prefer this method to simple randomized allocation, which might allocate the entire time series for a cell to either set. We use three-fourths of the data as the training set and the remaining fourth as the test set. As for EN models, we conduct sensitivity tests on the train–test splitting ratios, achieving consistent results across splits. For more details, refer to Sect. A5.

2. We train the model on the training set.
3. We evaluate its performance on the test set. We report the out-of-sample R^2 and RMSE calculated as in Eqs. (6) and (7).
4. We derive emissions-to-concentrations relationships from the extreme gradient boosting algorithm in a fashion similar to partial dependence plots (Friedman, 2001). Section 3.1 describes this step in more detail.

For ML models, we include an identifier of grid cell as an input variable, similar to what cell fixed effects would be in a regression framework. This increases the fit of models to geographical variation in emissions, concentrations, weather, and their interactions, especially in emission scenarios that are not excessively different from the baseline. For instance, recurrent transboundary pollution can be modeled by the interaction of cell identifiers and months. The improvement in geographic precision might come at the cost, however, of a higher bias in the case of extreme perturbations. For robustness, we also estimate the models without the identifier.

2.10 Method comparison

We summarize in Table 2 the advantages and disadvantages of the methods used in the CLAQC framework. While the elastic net models do not perform well using pixel-detail data and use country-level aggregate data instead, for most of the countries, the gradient boosting regressor method delivers reasonable results with high-resolution inputs increasing the statistical power. The pixel-based approach allows for flexible spatial aggregation, although we only discuss the country-level spatial resolution here.

3 Discussion

3.1 Model results – emission scenarios

We simulate perturbations in emissions to simulate hypothetical policy scenarios. Separately for every precursor, we perturb emissions by a factor P and predict concentrations

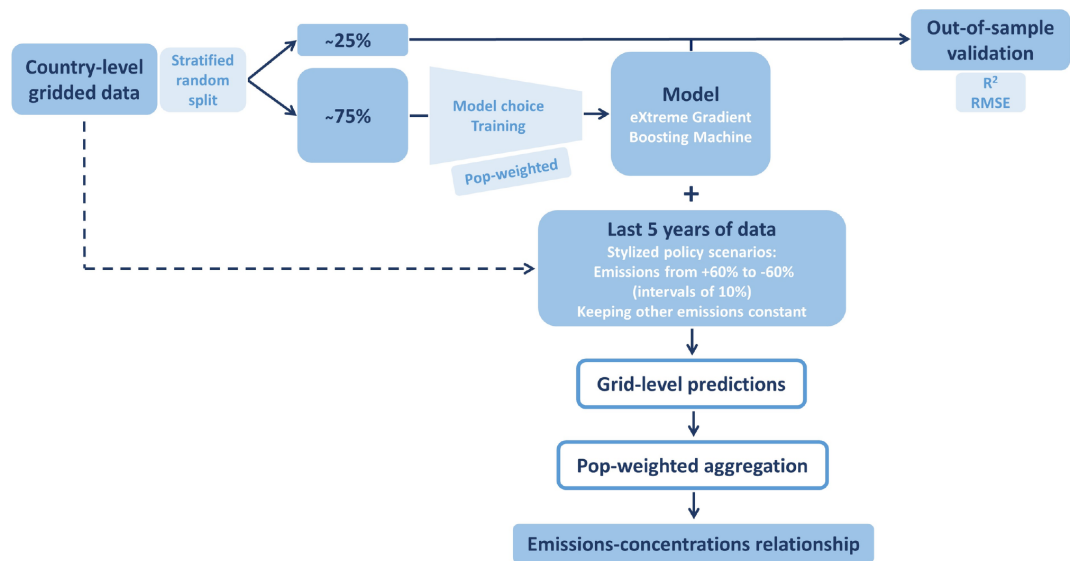


Figure 5. Visual representation of the extreme gradient boosting regressor models workflow.

Table 2. Comparison summary between CLAQC framework methods.

Elastic net	Extreme gradient boosting regressor
Simple equation	Non intelligible form
Country-level emission totals allow for direct and fast application but trade off flexibility in spatial aggregation	Flexible regional aggregation
Moderate sensitivity to emission changes	Low sensitivity to strong emission changes
Assume that historically correlated sectors will remain correlated	

under the average monthly weather of the five most recent years (2017–2021). Monthly predictions are then averaged to yearly ones and, for ML, from grid-level to country-level predictions. The process is performed with perturbation from +60 % to –60 % at intervals of 20 %. While policies generally aim to reduce emissions, including emission-increase scenarios is crucial for a comprehensive understanding of potential air quality outcomes of a wide range of possible future conditions, for example, the persistent investments in coal in India or the investment in gas fracking. It is important to showcase that these policy interventions may lead to exposure increases. For these reasons, we have looked at decreases and increases in emissions.

We consider model predictions as baseline predictions, $\widehat{\text{Concentrations}}_{\text{Baseline},y}$.

The predicted relative change in concentrations for a perturbation P of emissions of precursor p in sector s is

$$\% \Delta \widehat{\text{Concentrations}}_{A,y,s,p} = \frac{\widehat{\text{Concentrations}}_{A,y,s,p} - \widehat{\text{Concentrations}}_{\text{Baseline},y,s,p}}{\widehat{\text{Concentrations}}_{\text{Baseline},y,s,p}} \cdot 100. \quad (11)$$

Given that the model algorithms may include multiple emission variables within a sector, e.g., both NO_x and BC emissions from the road sector, to account for the sectoral range variability, we calculate the minimum and maximum annual percentage variation in predictions from perturbed emissions by perturbation and sectoral level (Eqs. 12 and 13).

$$\min_{A,s} \left(\widehat{\Delta \text{Concentrations}}_{A,y,s,p} \right) \quad (12)$$

$$\max_{A,s} \left(\widehat{\Delta \text{Concentrations}}_{A,y,s,p} \right) \quad (13)$$

Figures 6 and 7 plot the relative variation in annual predicted concentrations of PM_{2.5} and O₃, respectively, against perturbations by sector ranging from −60 % to +60 % in the selected major economies or populous countries: Brazil (BRA), Germany (DEU), Egypt (EGY), Italy (ITA), Mexico (MEX), Nigeria (NGA), Saudi Arabia (SAU), Turkey (TUR), the United States (USA), and South Africa (ZAF). Sectors are color-coded. To consult other countries' results, refer to the CLAQC online tool at <https://datashowb.shinyapps.io/CLAQC-App/> (last access: 27 March 2025).

In general, EN and ML models detect approximately the same number of sectors. However, the sectors selected vary according to the method. Moreover, EN models show greater variability in predictions compared to ML ones. This is expected as linear methods are less accurate, based on only a single estimator per predictor, and do not fully capture non-linear relationships.

Regarding PM_{2.5}, in 8 out of 10 EN models from Fig. 6, the other sector is selected, affecting predictions the most in the USA, ZAF, NGA, and EGY. In ITA, DEU, the USA, and NGA, agriculture plays a relevant role as well. Also, the road and residential sectors emerge as relevant in contributing to particle formation, though we find the greatest impacts in DEU, ITA, and BRA. Such sectors are often detected in ML as well.

In the case of O₃ predictions, positive perturbations may lead to a decrease in predicted concentration levels: this happens because NO_x consumes O₃ in NO_x-rich regimes, such as for DEU's road sector in the ML model. The industrial sector is picked up in both EN and ML models, particularly for EN models in ITA, EGY, MEX, USA, SAU, and BRA. In ML models, the agricultural sector is often associated with major variations (TUR, NGA, ITA, USA, and MEX). The power sector is another relevant sector, appearing in all selected countries.

Mainly, we find that EN models are good for predicting the total mass of ambient pollutants, while for some countries, they are not reliable for sectoral attribution. Therefore, in such a case, we suggest opting for OLS coefficients to attribute sector shares to pollutant totals (see Sect. A2.3 in the Appendix for further details).

See Sect. A5 in the Appendix to consult sensitivity tests on the train–test splitting ratio and Sects. A7.1 and A7.2 to consult other model specifications' results.

3.2 Model internal validation results

A key aspect of predictive model evaluation is to verify if the models can reproduce events well based on past trends.

We present out-of-sample validation for both methods while we perform validation against similar tools, i.e., ECLIPSE-GAINS and TM5-FASST, for EN models only (for the latter one, see Sect. A6 in the Appendix).

Figures 8 and 9 map the out-of-sample R^2 and RMSE for EN and ML models obtained from both CAMS and DACCIWA emissions. We do not advise using the models for countries with an R^2 smaller than 0.5 or RMSE higher than 12.

Results vary by prediction target (PM_{2.5} vs. O₃) and input type (e.g., CAMS vs. DACCIWA emissions), which may reflect inconsistencies in the emission or concentration data. Additionally, local factors such as unique orography and micro-meteorological conditions can significantly impact predictions in some areas, even country-level averages. Both elastic net and machine learning models are generally better at predicting O₃ than PM_{2.5} as the former is highly correlated with incoming radiation or temperature, while predicting PM_{2.5} is more challenging due to its complex secondary chemistry, local sources, and particle composition. Chemistry-transport models predict better O₃ than PM as well, due to the more complex mixture of particles and local effects from more sources of the latter one (Guérette et al., 2020).

Among the EN models with CAMS emissions for PM_{2.5}, 13 countries have an R^2 below 0.5, and 40 have RMSE above 12. Only three countries have an R^2 below 0.5, and four have RMSE above 12 for O₃. EN models for PM_{2.5} with DACCIWA emissions perform comparably; R^2 is smaller than 0.5 in 4 countries, and RMSE is greater than 12 in 19. All EN models for O₃ with DACCIWA emissions have an R^2 above 0.5 and RMSE below 12.

The ML models without a pixel identifier perform poorly in 10 and 2 countries regarding R^2 for PM_{2.5} and O₃, respectively, while we find an RMSE above 12 in 21 countries for PM_{2.5} and none for O₃. As in the EN models, PM_{2.5} predictions appear to be less accurate than those for O₃, while the ML models from DACCIWA emissions perform better in predicting both PM_{2.5} and O₃ than the EN models. In general, ML models with grid cell identifiers perform better than those without them both in terms of explained linear variation and error. For more details on the ML models with grid cell identifier, see Sect. A3.1, and for the validation metrics of the other model specifications, see Sect. A3.2.

3.3 Model selection

Having estimated multiple models that rely on different algorithms and data inputs, we set out to select the ones with

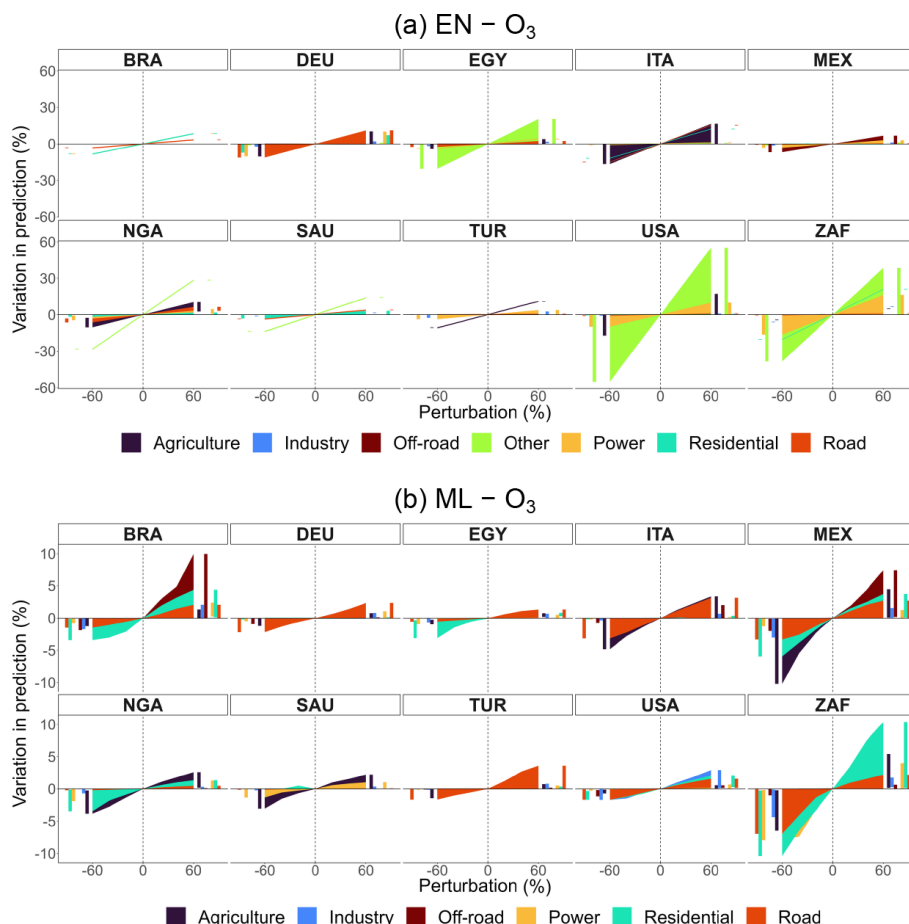


Figure 6. Percentage variation in predicted concentrations by sector and perturbation for selected countries in EN and ML models for weighted $\text{PM}_{2.5}$. Bar charts on the sides of each subplot help visualize overlapping variations.

the best predictive power.⁷ We propose a systematic model selection based on two criteria:

- *Model error/reliability.* It is measured by out-of-sample R^2 and RMSE.
- *Reliability of emission input data.* DACCIWA is preferred to CAMS as a source of data over Africa only as the former is developed with more consistent methods, which prefer in situ measurements as opposed to large data proxies and source profiles. Where DACCIWA is available, we assign a Source score of 1 to models using DACCIWA and 0 to models using CAMS. Where DACCIWA is unavailable, Source takes the value 0.

We re-scale all the elements of our decision criteria between 0 and 1, with 1 being the maximum score. We obtain

⁷More specifically, we consider four possible model specifications for elastic net and four for machine learning: $\text{EN} \times \{\text{CAMS emissions, DACCIWA emissions}\} \times \{\text{with month-fixed effects, without month-fixed effects}\}$ and $\text{ML} \times \{\text{CAMS emissions, DACCIWA emissions}\} \times \{\text{with grid identifier, without grid identifier}\}$.

the ensemble score, s_v , by weighting each of the criteria as in Eq. (14). For each country, we choose the model that maximizes (Eq. 14).

$$\frac{1}{6} \cdot R^2 + \frac{1}{6} \cdot \text{RMSE} + \frac{1}{3} \cdot \text{Source} \quad (14)$$

We note that other weighting criteria for performance decisions are possible. Figure 10 shows which model maximizes each criteria (R^2 , RMSE, Source). As expected, elastic net models have greater predictive power with month-fixed effects than without. Similarly, machine learning models perform better with pixel identifiers. On a general level, machine learning models perform better than elastic net models in Europe; in Africa, the two methods share the map, while the elastic net one is preferable in the remaining regions. We note that, by construction, DACCIWA is always preferable to CAMS as a source of input data over Africa only. To clarify further, we discuss the case of South Africa. The $\text{PM}_{2.5}$ results in Fig. 10 show that the ML model (pixel version) using DACCIWA emissions maximizes R^2 (panel a). For RMSE, the EN model with DACCIWA emissions achieves the best

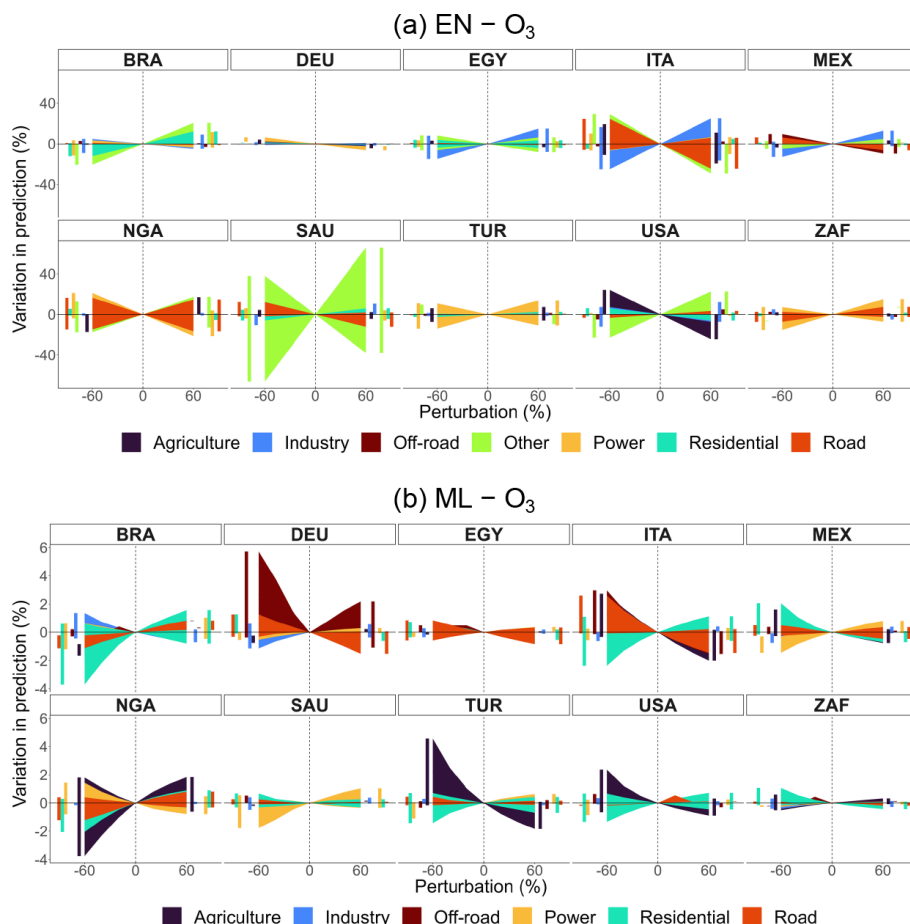


Figure 7. Percentage variation in predicted concentrations by sector and perturbation for selected countries in EN and ML models for O_3 . Bar charts on the sides of each subplot help visualize overlapping variations.

performance (panel b), and DACCIWA emissions are identified as the preferred input source (panel c). For O_3 , the EN model using CAMS-GLOB-ANT emissions maximizes R^2 (panel a), while the one with DACCIWA emissions minimizes RMSE (panel b). In terms of input preference, DACCIWA remains the preferred emission input (panel c).

Figure 11 maps instead the model specifications that maximize the composite criteria of Eq. (14). With some exceptions, the patterns highlighted in Fig. 10 are replicated.

All the maps and country-specific model results can be explored through the CLAQC-App web tool.

4 Limitations

The CLAQC framework is very ambitious in terms of detail. However, it is important to note the limitations stemming from the input data, model characteristics, and specifications.

First, results are highly dependent on input data availability and quality. Using gridded data at different resolutions requires harmonization on a common global grid. As data values are estimated at locations that were not origi-

nally measured in the raw data, the interpolation process can introduce measurement errors of unknown distribution. We source data for both emissions and concentrations from the CAMS services. This choice allows for a smoother integration. However, biases in the emissions dataset may propagate to the concentration ones and vice versa. Additionally, while reanalysis is considered a state-of-the-art and complete method, it will surely not yield the same results as having institutionally approved ground monitoring stations in each grid cell as it involves the use of data assimilation and CTM extrapolation to regions without ground or satellite monitoring. However, given the high disparities in the available ground monitoring data across the globe, we believe that CAMS reanalysis products, such as CAMS-GLOB-ANT and EAC4, are the next state-of-the-art available solution for these regions. Finally, we note that the model predictions have been evaluated with the support of observed levels of emissions and concentrations. Although we find a good out-of-sample model performance on annual concentration levels, caution is needed for the extrapolation of extreme perturbations. Furthermore, we did not analyze the reasons be-

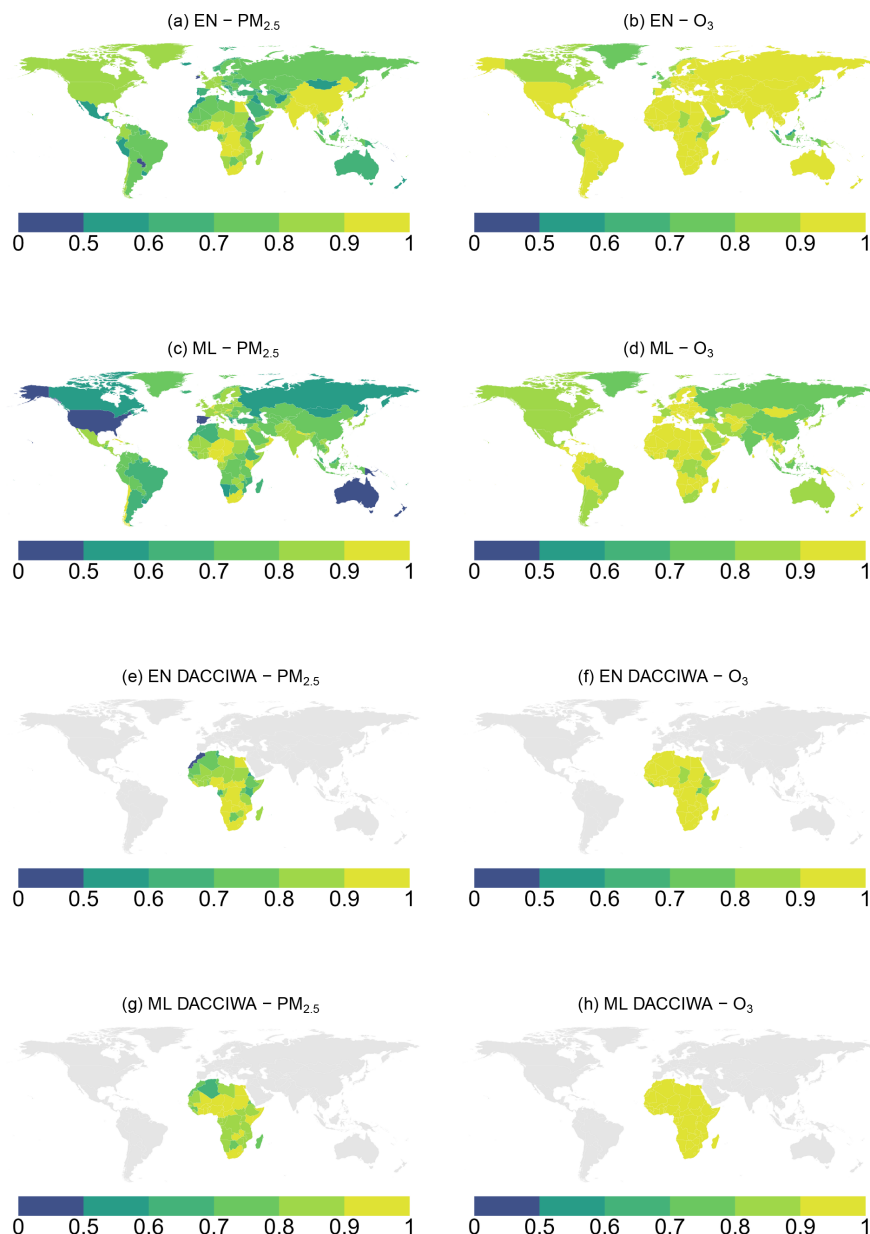


Figure 8. Out-of-sample performance metrics of ML (no pixel) and EN models (from both CAMS and DACCIIWA data) as in Eqs. (8) and (9) under Sect. 2.8: R^2 .

hind countries' poor performance, which limits our ability to interpret differences in results. In general, the model should be used under the fit-for-purpose principle, i.e., for country-level policy roll-out purposes.

Second, the COVID-19 pandemic provided disruptive emission perturbations that are of key importance for this model. However, they represent only one set of large perturbations, which may differ from the real-world implementation of country-by-country policies and may differ spatially, meteorologically, or seasonally.

Third, sectoral emissions include a very large variability across publicly available databases. The results presented

here are therefore sensitive to these uncertainties. If the baseline sectoral emission distribution used as input into CLAQC is substantially different from the CAMS baseline sectoral emissions, we recommend a rescaling of the total pollutant emissions to the CAMS sectoral emission profile.

Fourth, the EN model estimators may present higher variability as they rely on fewer observations relative to the ML models.

Fifth, as CLAQC was built to support policy impact evaluation, the approaches presented here do not explicitly model transboundary movements of pollution and biogenic emissions, such as desert dust and sea salt. Only averages enter the

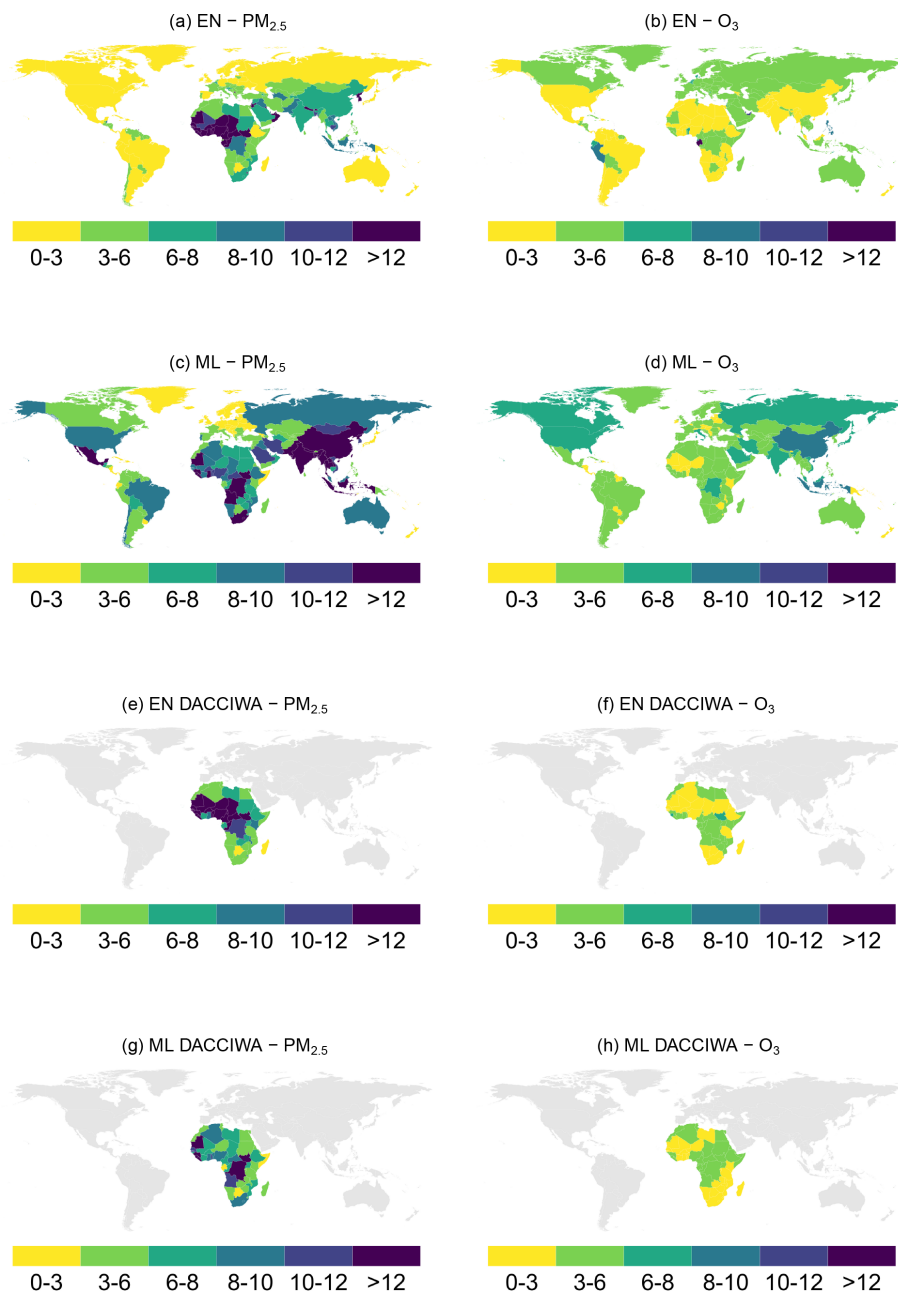


Figure 9. Out-of-sample performance metrics of ML and EN models (from both CAMS and DACCIIWA data) as in Eqs. (8) and (9) under Sect. 2.8: RMSE.

models through the time and place identifiers (month-fixed effects in EN and grid identifiers in ML). Secondary inorganic aerosols (SIAs) are also not directly modeled. Instead, their effects are approximated through interactions among emissions' predictors. In addition, we have not performed sensitivity analyses to assess the impact of excluding SIAs on the final estimates.

5 Conclusions

We have developed CLAQC, a tool that provides fast simulations of emissions changes with national and sub-national resolution and global coverage. Based on statistical methods, it aims at supporting policy assessments in a timely fashion. The user sets the emission reduction for a given precursor from a given sector (or a combination thereof), and CLAQC simulates the implied change in concentrations of $\text{PM}_{2.5}$ and O_3 . A possible application is, for instance, the calculation of

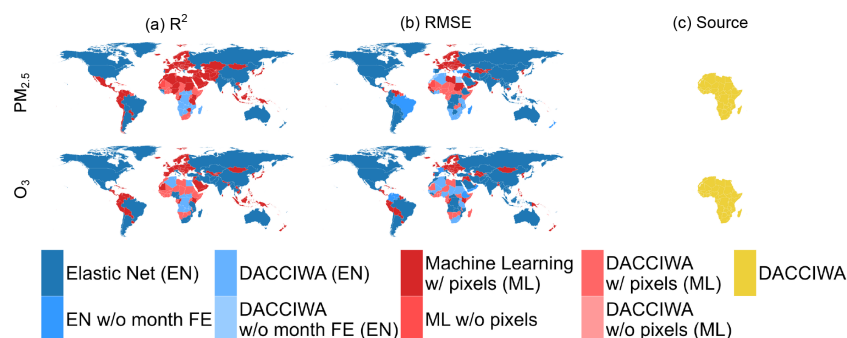


Figure 10. Best model score for each pollutant, country, and decision criterion.

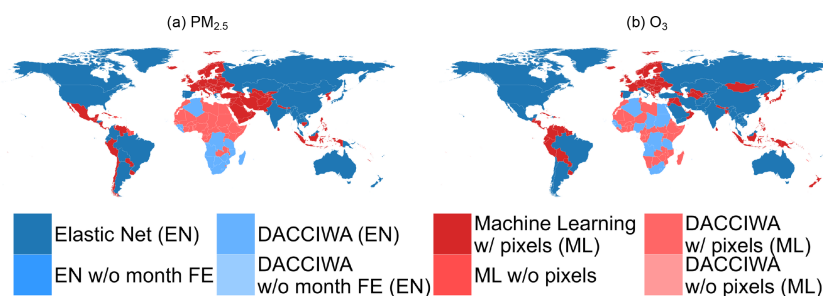


Figure 11. Best model score for each pollutant and each country.

co-benefits of climate policies. CLAQC can also be a tool for the scientific community and complement instruments such as IAMs.

CLAQC is grounded on two different methods that trade off transparency and predictive performance. Both methods perform well, with predictive performance above reasonable levels for most countries. Elastic net models generally well estimate total annual exposure, although they are less reliable for pinpointing the contributions of individual sectors. For such a task, the machine learning approach should be preferred.

The CLAQC framework lends itself to multiple developments. It is a complementary tool to the modeling and policy scenario community, providing empirically based estimates and added value for global-scale sectoral and country-level analyses. Its dynamic architecture makes it simple to update with more recent data, and the framework can be extended to both new data sources and methods. One potential evolution is to transform it into an ensemble model to enhance its accuracy, robustness, and reliability.

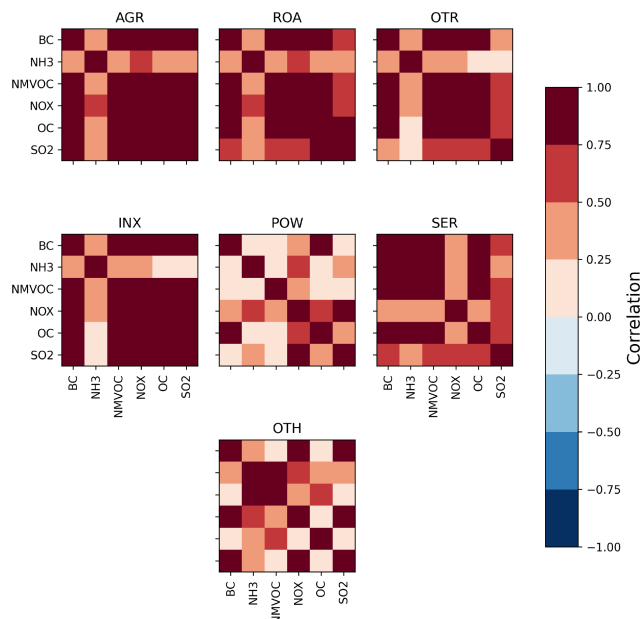


Figure A1. Correlation matrix among CAMS emission predictors.

Appendix A

A1 Multicollinearity

Differently from concentrations, emissions of pollutants and precursors are not always directly measured, but they can also be inferred using activity data and highly detailed emis-

sion factors. Emission data display a high correlation, even within grid cells, which is plausibly attributable to the correlation of emissions of pollutants within sectors and in economic activity across sectors. Figure A1 displays the cross-pollutant correlations within sectors.

A2 Elastic net models

A2.1 Method description

Shrinkage regression methods, such as elastic net, were developed to tackle some OLS limitations, in particular concerning the model interpretation and prediction accuracy. In OLS, the linear equation coefficients are estimated by minimizing the sum of squared residuals. Though, when there are many predictors, OLS models generally show high variance and unstable coefficients. The elastic net method minimizes such variance. In fact, shrinkage regression may improve prediction accuracy by either shrinking regression coefficients towards zero or setting them to zero or both. However, a trade-off is produced: as the variance is reduced, the bias may increase, in this case, a bias toward more conservative outcomes. Moreover, in the OLS approach, when a large number of predictors is present, it may not be straightforward to identify those representing the most relevant influence. In CLAQC elastic net models, a penalization parameter lambda (λ) is introduced; in OLS, this parameter is zero. In CLAQC, λ is selected using cross-validation to minimize divergence so that for each country, the most optimized penalization parameter of the coefficients is identified. In such a procedure, predictors are also standardized in order to identify solutions that do not depend on the unit of measurement of the features. For further details, see Zou and Hastie (2005) and Hastie et al. (2009).

A2.2 Elastic net models with DACCIWA emission data

The EN linear regression models obtained using DACCIWA emission data (Keita et al., 2017, 2021) take the following form for each country:

$$\begin{aligned} \text{PM}_{2.5t} = & \alpha + \sum_{s,p_1} \beta_{s,p_1} E_{s,p_1,t} + \gamma_1 \text{PPT}_t + \gamma_2 \text{TMIN}_t \\ & + \gamma_3 \text{TMAX}_t + \gamma_4 \text{VPD}_t + \gamma_5 \text{WS}_t + \gamma_6 \text{WD}_t + \\ & + \sum_s \delta_s E_{s,t} + \sum_{p_1} \lambda_{p_1} E_{p_1,t} + \xi E_{\text{SO}_{2,t}} \cdot E_{\text{NO}_{x,t}} \\ & + \sum_s \theta_s E_{s,t} \cdot \text{WS}_t \cdot \text{WD}_t + \phi_t + \varepsilon_t, \end{aligned} \quad (\text{A1})$$

$$\begin{aligned} \text{O}_{3t} = & \alpha + \sum_{s,p_2} \beta_{s,p_2} E_{s,p_2,t} + \gamma_1 \text{PPT}_t + \gamma_2 \text{TMIN}_t \\ & + \gamma_3 \text{TMAX}_t + \gamma_4 \text{VPD}_t + \gamma_5 \text{WS}_t + \gamma_6 \text{WD}_t + \\ & + \sum_s \delta_s E_{s,t} + \sum_{p_3} \lambda_{p_3} E_{p_3,t} + \mu E_{\text{NO}_{x,t}} \cdot E_{\text{NMVOC},t} \\ & + \nu E_{\text{SO}_{2,t}} \cdot E_{\text{NMVOC},t} + \xi E_{\text{SO}_{2,t}} \cdot E_{\text{NO}_{x,t}} + \\ & + \sum_s \theta_s E_{s,t} \cdot \text{WS}_t \cdot \text{WD}_t + \phi_t + \varepsilon_t, \end{aligned} \quad (\text{A2})$$

where

- $s \in \{\text{transport, power, industry, residential, other}\}$

- $p_1 \in \{\text{BC, NMVOC, NO}_x, \text{OC, SO}_2\}$
- $p_2 \in \{\text{NMVOC, NO}_x\}$
- $p_3 \in \{\text{NMVOC, NO}_x, \text{SO}_2\}$
- $\text{PM}_{2.5t}$ is the monthly concentration of $\text{PM}_{2.5}$ in $\mu\text{g m}^{-3}$ (population-weighted)
- O_{3t} is the monthly concentration of O_3 in $\mu\text{g m}^{-3}$
- $E_{s,p_1,t}$ is the monthly emissions of sector s and pollutant p_1 in kilograms
- $E_{s,p_2,t}$ is the monthly emissions of sector s and pollutant p_2 in kilograms
- $E_{s,p_3,t}$ is the monthly emissions of sector s and pollutant p_3 in kilograms
- PPT_t is the monthly accumulated precipitation in millimeters
- TMIN_t is the monthly minimum 2 m temperature in $^{\circ}\text{C}$
- TMAX_t is the monthly maximum 2 m temperature in $^{\circ}\text{C}$
- VPD_t is the monthly mean vapor pressure deficit in kilopascals
- WS_t is the monthly 10 m wind speed in m s^{-1}
- WD_t is the monthly wind direction in degrees
- $E_{p_1,t}$ is the monthly composite index from the sum of total emissions of pollutant p_1 in kilograms
- $E_{p_2,t}$ is the monthly composite index from the sum of total emissions of pollutant p_2 in kilograms
- $E_{p_3,t}$ is the monthly composite index from the sum of total emissions of pollutant p_3 in kilograms
- $E_{s,t}$ is the monthly composite index from the sum of total emissions of sectors s in kilograms
- ϕ_t is the monthly fixed effects
- ε_t is the error term.

Unlike the equations in (8) and (9) from Sect. 2.1.4, in Eqs. (A1) and (A2), all NH_3 predictors⁸, and agriculture sectoral emissions are not present, while the transport sector's predictors are not split into road and off-road transportation.

$\text{PM}_{2.5}$ and O_3 concentration values, in $\mu\text{g m}^{-3}$, are obtained from the models are country-level monthly concentration averages indexed by time t as all the other parameters in the equation; emissions are in kilograms; weather variables' units of measurement are expressed as specified in Sect. 2.8.

⁸This includes interactions containing NH_3 totals.

A2.3 Sectoral attribution

Given that in some cases, elastic net results are not suitable for sectoral attribution, to tackle such an issue, we run constrained OLS models with sector totals and other controls only. In this way, elastic net results can be used to predict the total mass of our pollutants of interest, while OLS coefficients can be exploited to distribute concentration contributions by sector. We follow the procedure explained in Sect. 2.8 with some modifications. In particular, in step 2, we first aggregate gridded sectoral emissions, $E_{s,p[n],t}$, at the country, year, and month level and then normalize them to range from 0 to 1 (min–max normalization) as follows:

$$\text{norm}E_{s,p,t} = \frac{E_{s,p,t} - \text{Min}_{s,p}}{\text{Max}_{s,p} - \text{Min}_{s,p}}, \quad (\text{A3})$$

where $\text{norm}E_{s,p,t}$ is the normalized monthly sectoral emission, $\text{Min}_{s,p}$ is the minimum emission level across months for pollutant p and sector s , and $\text{Max}_{s,p}$ is the maximum. Finally, we sum them sector-wise by country, year, and month to obtain monthly sector total, $\text{norm}E_{s,t}$, as specified:

$$\text{norm}E_{s,t} = \sum_{p,t} \text{norm}E_{s,p,t}. \quad (\text{A4})$$

In step 7 (model fitting), within the `glmnet` function, we set the lambda parameter and all penalty factors related to the emission variables equal to 0 and set the threshold parameter for interrupting convergence to the solution, `thresh`, to 10^{-14} to get the OLS results. For the same reason, in step 8 (model evaluation), we set the penalty parameter `lambda`, `s`, to 0.

For each pollutant, poll_t , the model specification only includes the total sectoral emission, $E_{s,p1,t}$; weather variables; and month-fixed effects, ϕ_t , and takes the following form:

$$\text{poll}_t = \alpha + \gamma_1 \text{PPT}_t + \gamma_2 \text{TMIN}_t + \gamma_3 \text{TMAX}_t + \gamma_4 \text{VPD}_t + \gamma_5 \text{WS}_t + \gamma_6 \text{WD}_t + \sum_s \delta_s E_{s,t} + \phi_t + \varepsilon_t. \quad (\text{A5})$$

However, it is important to acknowledge that such OLS models may have limitations:

- Due to multicollinearity among certain sectors, it is likely that OLS models will not include all sectoral predictors.
- OLS models may introduce bias since relevant variables, such as biogenic emissions, interactions, non-linear terms, and others, are excluded from the model specification.
- Assuming non-linear relationships between sectoral emissions, weather conditions, and concentration levels, if only linear variables are used, the OLS models may incorrectly attribute sector shares.
- Mostly, OLS models will differ from elastic net models in terms of variable selection and coefficient estimation.

A3 Machine learning models

A3.1 XGBoost additional remarks

As stated in Sect. 2.9, we consider two ML model specifications, with and without the grid cell feature. As expected, models with a grid cell identifier perform better than those without it (Fig. A2). Figure A3 compares the changes in concentrations predicted with and without a grid cell identifier in the extreme scenario of 100 % reduction in emissions. The results of both sets of models are similar. We thus prefer models with the identifier for their greater out-of-sample performance.

A3.2 XGBoost models with DACCIWA emission data

A4 Model implementation

The models presented in this framework follow different implementation procedures. While elastic net models can be implemented using a linear equation and temporal profiles, machine learning models can be implemented after emulation through a spreadsheet-style format.

A4.1 Elastic net models

We provide a monthly scheduling profile for each country to transform annual emissions into monthly emissions to be fed into the models. We build such a monthly schedule using the most recent years of emission data (2017–2021) for each couple of pollutant and sector, $E_{s,p}$, representing a reference monthly emission value by sector and pollutant. The monthly weights can be multiplied by the equivalent pollutant and sector total annual precursor emissions and then be used directly in Eqs. (8) and (9).

Additionally, we provide default meteorology fields that can be used in the models in case the input data are missing. The default meteorology variable fields are based on each country's average of the last 5 years of meteorology to represent current trends.

In health and crop impact assessments of air pollution due to O_3 , other metrics, such as the 6-month warm season mean of daily maximum 8 h average (6mDMA8), are more common. We provide a post-process dataset that allows for converting from annual average O_3 concentrations to a 6mDMA8 metric, obtained by Tropospheric Ozone Assessment Report (TOAR) surface ozone data products. See Schultz et al. (2017) for more details.

A4.2 Machine learning — guide to the Excel spreadsheet

A machine learning model was built for every country and ambient pollutant ($\text{PM}_{2.5}$ and O_3), and an empirical emissions-to-concentrations relationship was constructed. The Excel spreadsheet `Results.xlsx` contains the data

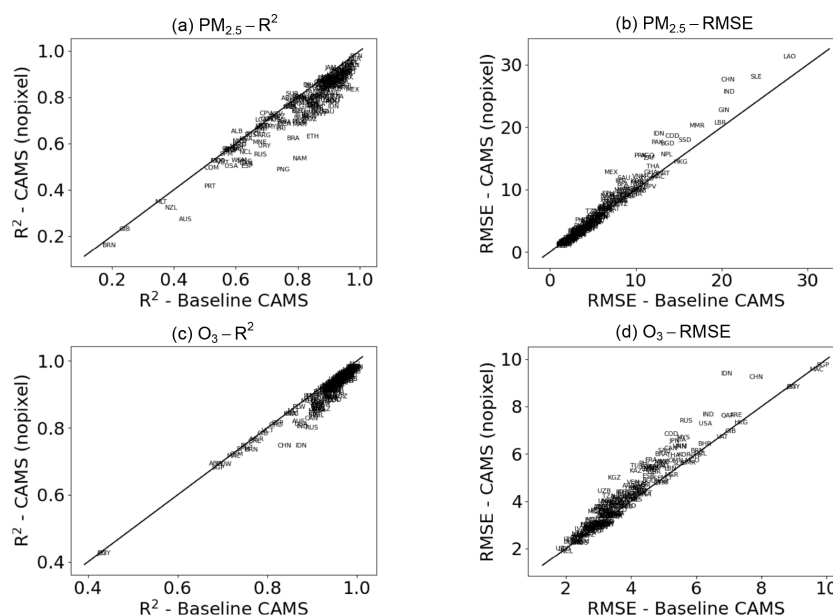


Figure A2. Comparison of performance of models with (x axis) and without (y axis) grid cell identifier. Black lines indicate equality.

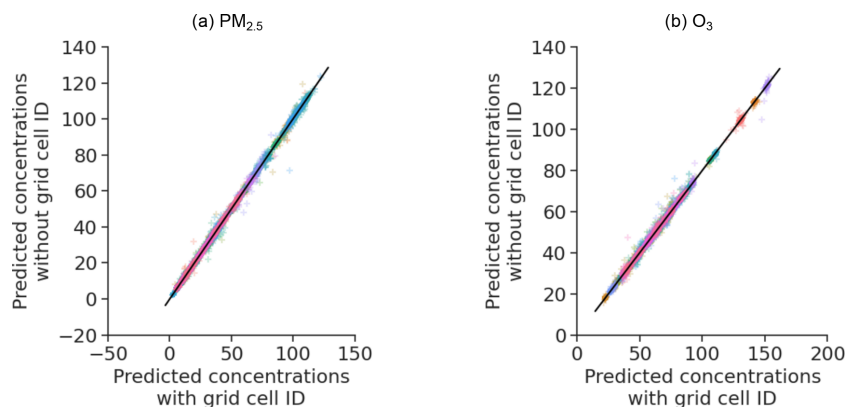


Figure A3. Concentrations under the simulation of perturbations for models with and without grid cell identifier. Each cross is a country–sector–precursor–perturbation combination. Black lines indicate equality, colors indicate countries.

required to derive the changes and the levels of concentrations of pollutants under emission scenarios supplied by the user. Each row is defined by the combination of country, pollutant, sector, and precursor. The goodness of fit of each country–pollutant model, as measured by out-of-sample R^2 and RMSE, is reported as well.

For easier implementation within a spreadsheet, the relationships have been approximated with piecewise linear functions that map perturbations of emissions to concentrations. A perturbation P is the relative difference in emissions between the baseline scenario and a chosen scenario expressed in 100 percentage points.

Omitting subscripts for the country and emitted pollutant for ease of notation, we call $E_{\text{Baseline},s,p}$ the baseline emission from sector s of precursor p and $E_{A,s,p}$ the emission

under the alternative scenario A . Then, for every country, pollutant, sector, and precursor, the perturbation $P_{A,s,p}$ is

$$P_{A,s,p} = \left(\frac{E_{A,s,p}}{E_{\text{Baseline},s,p}} - 1 \right) \cdot 100. \quad (\text{A6})$$

Assuming all other emissions are constant, the concentrations under the alternative scenario A are

$$\text{Concentrations}_{A,s,p} = \begin{cases} a_{-100,s,p} & \text{if } P_{A,p,t} < -100 \\ a_{-100,s,p} + b_{-100,s,p} \cdot (P_{A,p,t} + 100) & \text{if } -100 \leq P_{A,p,t} < -80 \\ a_{-80,s,p} + b_{-80,s,p} \cdot (P_{A,p,t} + 80) & \text{if } -80 \leq P_{A,p,t} < -60 \\ a_{-60,s,p} + b_{-60,s,p} \cdot (P_{A,p,t} + 60) & \text{if } -60 \leq P_{A,p,t} < -40 \\ a_{-40,s,p} + b_{-40,s,p} \cdot (P_{A,p,t} + 40) & \text{if } -40 \leq P_{A,p,t} < -20 \\ a_{-20,s,p} + b_{-20,s,p} \cdot (P_{A,p,t} + 20) & \text{if } -20 \leq P_{A,p,t} < 0 \\ a_{0,s,p} + b_{0,s,p} \cdot (P_{A,p,t}) & \text{if } 0 \leq P_{A,p,t} < 20 \\ a_{20,s,p} + b_{20,s,p} \cdot (P_{A,p,t} - 20) & \text{if } 20 \leq P_{A,p,t} < 40 \\ a_{40,s,p} + b_{40,s,p} \cdot (P_{A,p,t} - 40) & \text{if } 40 \leq P_{A,p,t} < 60 \\ a_{60,s,p} + b_{60,s,p} \cdot (P_{A,p,t} - 60) & \text{if } 60 \leq P_{A,p,t} < 80 \\ a_{80,s,p} + b_{80,s,p} \cdot (P_{A,p,t} - 80) & \text{if } 80 \leq P_{A,p,t} < 100 \\ a_{100,s,p} + b_{100,s,p} \cdot (P_{A,p,t} - 100) & \text{if } P_{A,p,t} > 100 \end{cases} \quad (\text{A7})$$

The coefficient $a_{j,s,p}$ is the level of concentrations when emissions of precursor p from sector s are perturbed by $j\%$. The coefficient $b_{j,s,p}$ is the slope of the piecewise function in the interval starting at j . The coefficients $a_{-100,s,p}, \dots, a_{100,s,p}$ and $b_{-100,s,p}, \dots, b_{100,s,p}$ are reported in the spreadsheet in columns L to AG. The coefficient $a_{0,s,p}$ is the value that the function takes when the perturbation is null. Thus, it is a generally close approximation of the concentration level given baseline emissions. Emissions are in kilograms, while concentrations of $\text{PM}_{2.5}$ are expressed in $\mu\text{g m}^{-3}$, and concentrations of O_3 are in $\text{6mDMA8 } \mu\text{g m}^{-3}$. Baseline concentrations, in column H, are the average concentrations (over the entire country) from 2017 to 2021. In column I, baseline emissions are the average precursor emissions from a given sector over the same period.⁹ Scenario emissions, in column J, are set by the user. The perturbation in column K is automatically computed. Concentrations under the alternative scenario are computed in column AK following Eq. (A7). It should be noted that the calculation assumes that only emissions of the row sector–precursor pair are perturbed. All other emissions are assumed to be constant. The change in concentrations attributable to the perturbation $P_{A,s,p}$ is calculated in column AI as the difference between baseline concentrations and concentrations under the alternative scenario. Again, this is the change in concentrations assuming all other sectoral emissions are kept constant. The change is computed as follows:

$$\begin{aligned} \Delta \text{Concentrations}_{A,s,p} &= \text{Concentrations}_{\text{Baseline}} \\ &\quad - \text{Concentrations}_{A,s,p} \\ &= a_{0,s,p} - [a_{j,s,p} + b_{j,s,p} \cdot (P_{A,s,p} - j)] \\ &= a_{0,s,p} - a_{j,s,p} - b_{j,s,p} \cdot (P_{A,s,p} - j), \end{aligned} \quad (\text{A8})$$

where $P_{A,s,p}$ is inside an interval starting at j . When scenario emissions are set to zero, the change in concentrations

⁹Averages are weighted by population in models for $\text{PM}_{2.5}$, but not in models for O_3 .

gives the (opposite of the) estimated contribution of each sector–precursor to the total concentrations in 2017–2021. The approximation of the emissions–concentrations relationship functions is best for small and moderate perturbations and larger under scenarios of extreme perturbations. We suggest applying perturbing emissions only in the $\pm 60\%$ range based on the fitness-for-purpose principle and given the limitations discussed in Sect. 4. To avoid that approximation error reverses the relationship between emissions and concentrations of $\text{PM}_{2.5}$, which is known to be positive, we impose in column P that negative perturbations cannot result in an increase in concentrations, and vice versa. The total change in concentrations under emission scenario A is computed in column AJ summing across sectors and precursors:

$$\begin{aligned} \Delta \text{Concentrations}_{A,s,p} &= \sum_{s,p} \Delta \text{Concentrations}_{A,s,p} \\ &= \sum_{s,p} a_{0,s,p} - a_{j,s,p} - b_{j,s,p} \cdot (P_{A,s,p} - j). \end{aligned} \quad (\text{A9})$$

The level of concentrations under scenario A is then reported in column AJ as

$$\text{Concentrations}_A = \text{Concentrations}_{\text{Baseline}} + \sum_{\text{Country}} \Delta \text{Concentrations}_{A,s,p}. \quad (\text{A10})$$

It should be noted that, differently from the other columns, the total change in concentrations $\sum_{\text{Country}} \Delta \text{Concentrations}_{A,s,p}$ (column AJ) and the level of concentrations Concentrations_A (column AK) are invariant within a country–pollutant pair. Therefore, the same value appears in multiple rows.

Comparing the two scenarios

It is possible to compare concentrations in two scenarios in the following way. Consider two scenarios, A and B. The difference in concentrations attributable to changes in precursor p from sectors s is

$$\begin{aligned} \Delta \text{Concentrations}_{A,s,p} - \Delta \text{Concentrations}_{B,s,p} &= a_{0,s,p} - a_{j_A,s,p} - b_{j_A,s,p} \cdot (P_{A,s,p} - j) - [a_{0,s,p} \\ &\quad - a_{j_B,s,p} - b_{j_B,s,p} \cdot (P_{B,s,p} - j)] \\ &= a_{j_B,s,p} + b_{j_B,s,p} \cdot (P_{B,s,p} - j) - a_{j_A,s,p} - b_{j_A,s,p} \\ &\quad \cdot (P_{A,s,p} - j), \end{aligned} \quad (\text{A11})$$

whereas the difference in the total change in concentrations (and the difference in levels of concentrations) is

$$\begin{aligned} \Delta \text{Concentrations}_{A,s,p} - \Delta \text{Concentrations}_{B,s,p} &= \text{Concentrations}_A - \text{Concentrations}_B \\ &= \sum_{s,p} a_{j_B,s,p} + b_{j_B,s,p} \cdot (P_{B,s,p} - j) - a_{j_A,s,p} \\ &\quad - b_{j_A,s,p} \cdot (P_{A,s,p} - j). \end{aligned} \quad (\text{A12})$$

Example. All emissions are set to zero in scenario A and set uniformly at 90 % of baseline emissions in scenario B .

$$\begin{aligned} \Delta_{\text{Country}} \text{Concentrations}_{A,s,p} - \Delta_{\text{Country}} \text{Concentrations}_{B,s,p} \\ = \text{Concentrations}_A - \text{Concentrations}_B \\ = \sum_{s,p} a_{-20,s,p} + b_{-10,s,p} \cdot (-10 + 20) - a_{-100,s,p} \\ - b_{-100,s,p} \cdot (-100 + 100) \end{aligned} \quad (\text{A13})$$

A5 Train–test splitting sensitivity analysis

The training and test dataset splits differ between EN and ML models due to differences in their data samples. EN models are trained and tested on country-level aggregated data, while ML models use gridded country-level data, resulting in a larger sample size. Given these differences, we do not harmonize data splitting across methods. Instead, we ensure a sufficiently large training set for EN models to reduce variance in parameter estimates.

To validate model configurations more robustly, we conduct additional runs with varied train–test splits for models obtained from CAMS emissions: 75–25 and 70–30 for EN models and 80–20 and 70–30 for ML models, alongside their original splits. Figures A4 and A5 present emission policy scenarios derived from EN models for PM_{2.5} and O₃, comparing splits of 84–16, 75–25, and 70–30. Similarly, Figs. A6 and A7 compare policy scenarios for ML models using 80–20, 75–25, and 70–30 splits. These sensitivity analyses confirm that model predictions remain stable across different train–test splits, showing only minor variations.

A6 Model external validation results

For the elastic net methodology only, we evaluate the previously presented CLAQC models against different global data sources: namely, the Evaluating the Climate and Air Quality Impacts of Short-Lived Pollutants (ECLIPSE) scenarios (<https://iiasa.ac.at/models-tools-data/global-emission-fields-of-air-pollutants-and-ghgs>, last access: 27 March 2025, Stohl et al., 2015) provided within the GAINS model (Amann et al., 2011; Kiesewetter et al., 2015) and the TM5-FASST model (Van Dingenen et al., 2018).

A6.1 Comparison with the GAINS model

To evaluate CLAQC models against GAINS, we obtain the anthropogenic emission data from ECLIPSE CLE (Current legislation)¹⁰ v5, 1990–2050, quinquennial, at 0.5° spatial resolution, focusing on 2020, 2025, and 2030. Annual gridded sectoral emission data cover the following sectors and are originally expressed in kt yr^{−1} (these have been converted to kilograms to implement them into the CLAQC model): agriculture (waste burning on fields), industry (combustion

and processing), power plants, energy conversion, extraction¹¹, residential and commercial, waste, and surface transportation.¹²

We download the GAINS PM_{2.5} concentrations from the GAINS Online (<https://gains.iiasa.ac.at/models/index.html>, last access: 27 March 2025, IIASA, 2009) tool, measured in µg m^{−3}.¹³ To acknowledge initial differences between datasets, we compare 2015 CAMS PM_{2.5} concentrations with 2015 GAINS PM_{2.5} concentrations (see Fig. A8).

We compare the model outcomes after having applied population weights to the GAINS reported concentrations (see Fig. A9). The weighted CAMS concentrations from almost all considered countries are above the line of equality. Given that starting emissions and concentrations show different values between the two approaches, CLAQC and ECLIPSE-GAINS models, we expect that also their outcomes will yield different results. While CAMS concentrations range between 19.3 and 199.7 µg m^{−3} (median of 54.3 µg m^{−3}), GAINS concentrations span between 9.5 and 74.9 µg m^{−3} (median of 30.6 µg m^{−3}).

Regarding the comparison between CAMS and GAINS emissions, we find that while annual pollutant totals in the two datasets reflect similar magnitudes, annual sectoral emissions diverge substantially in their order of magnitude, e.g., up to the order of thousands. More broadly, such heterogeneity is confirmed when comparing other sectoral emission sources present in the literature (Li et al., 2017; Kurokawa and Ohara, 2020). Thus, sector-specific air quality models face the infamous problem of drastic input source uncertainty when it comes to delivering sectoral details. Importantly, even a few orders of magnitude of differences between the emission input data and the emission data used during the model training may generate non-realistic concentrations as the resulting coefficients are scaled to the order of magnitude of the underlying training data. There are two ways of overcoming this issue: (i) using CLAQC for scenario comparison, i.e., a reference scenario and a policy scenario are simulated, and the difference between the two scenarios may be used for policy analysis instead of the absolute values, and (ii) implementing the emissions into CLAQC by applying the CAMS sectoral emission shares to the total of the emissions inputted. Here we apply the latter.

We then input ECLIPSE emissions into the CLAQC model. Firstly, we pre-process the data to match the input requirements of CLAQC; in particular, we do the following:

- We aggregate sectoral emissions to make them match CLAQC’s variables.
- We re-scale the surface transportation sector into off-road and road transportation based on CAMS shares.

¹¹Including gas flaring.

¹²Shipping sectoral emissions are not considered.

¹³Notice that for O₃, GAINS reports a different metric, i.e., SOMO35; therefore, we make such an evaluation for PM_{2.5} only.

¹⁰This is the baseline scenario.

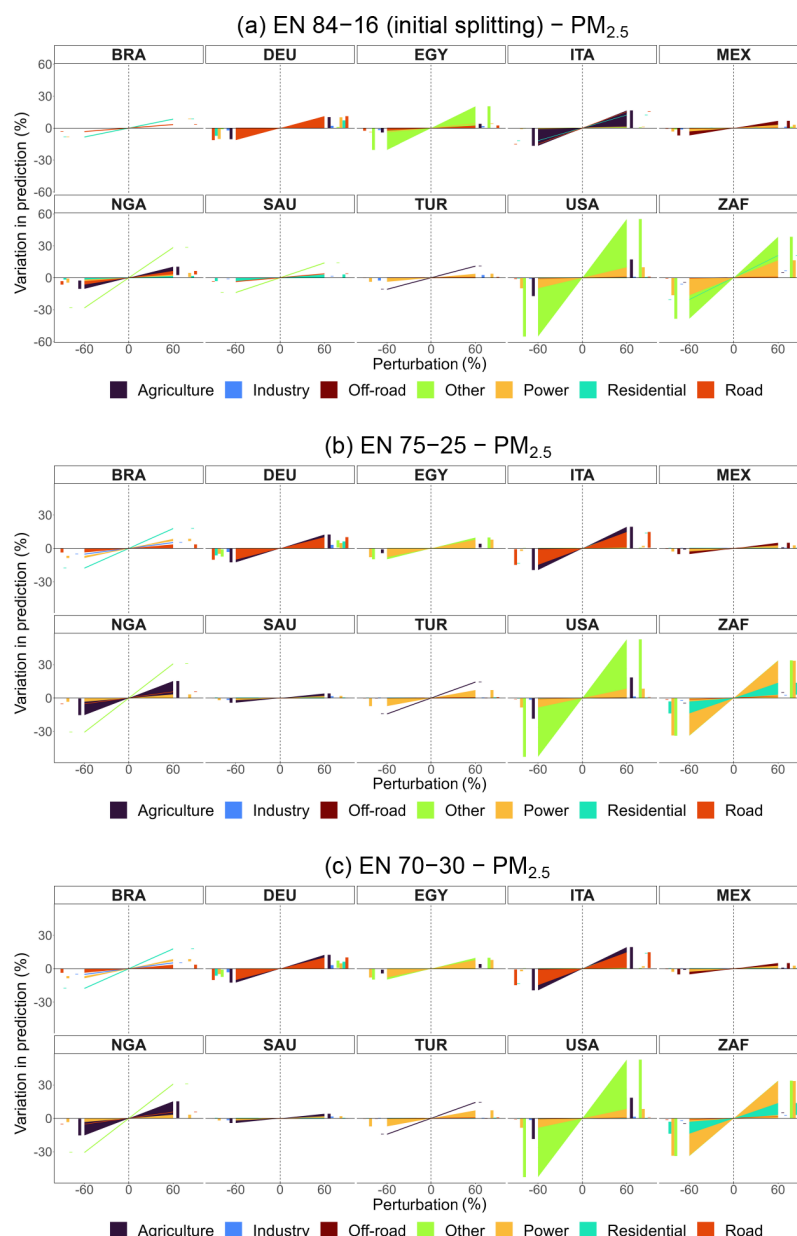


Figure A4. Percentage variation in predicted concentrations of EN models for population-weighted $PM_{2.5}$ obtained from CAMS emissions, by sector and perturbation for selected countries: 84–16 (initial splitting), 75–25, and 70–30 train–test splitting. Bar charts on the sides of each subplot help visualize overlapping variations.

- We re-scale ECLIPSE CLE V5 annual sectoral emissions to monthly sectoral emissions by applying CLAQC monthly emission profiles based on 2014–2018 data.
- We use the resulting gridded emissions with typical meteorology data¹⁴ and aggregate them at the country level to be implemented into the CLAQC model.

¹⁴Here we refer to an average of all years.

Finally, after running ECLIPSE emissions into CLAQC, we aggregate the obtained country-level monthly concentrations at the annual level to compare them with GAINS country-level annual concentrations.

As shown in Fig. A10, the GAINS model underestimates population-weighted concentrations of $PM_{2.5}$ in several Asian countries for the year 2020.¹⁵ CLAQC predicted values ranging between 2.9 and $154.2 \mu g m^{-3}$ (median of

¹⁵The same pattern is repeated in the years 2025 and 2030, so we omit such graphs.

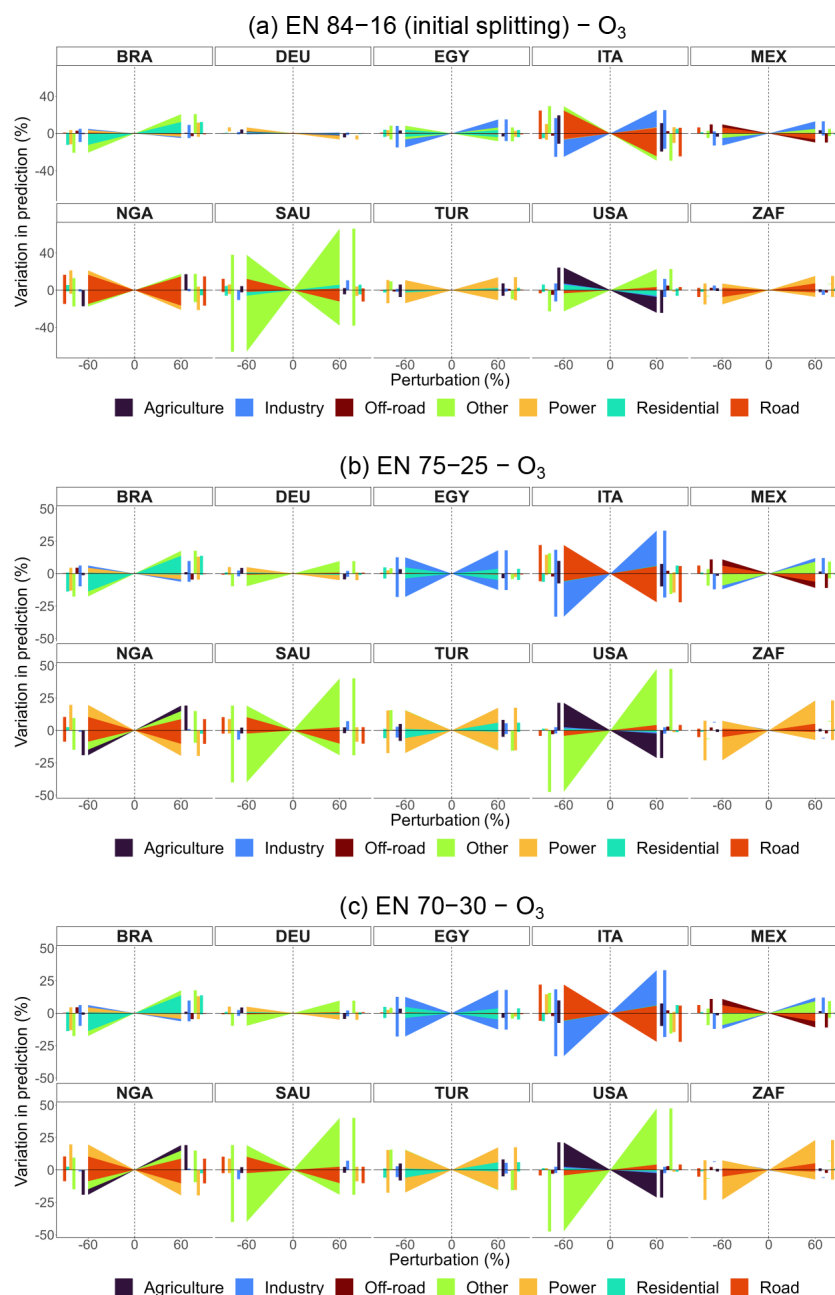


Figure A5. Percentage variation in predicted concentrations of EN models for O_3 obtained from CAMS emissions, by sector and perturbation for selected countries: 84–16 (initial splitting), 75–25, and 70–30 train–test splitting. Bar charts on the sides of each subplot help visualize overlapping variations.

$67.7 \mu\text{g m}^{-3}$) as opposed to GAINS concentrations ranging between 8.5 and 80.6 (median of $27.2 \mu\text{g m}^{-3}$). These differences can be explained by the different sectoral emission aggregations, spatial and temporal resolutions, and different approaches followed in calculating concentrations. While CAMS concentrations are derived from a combination of multiple sources, including measurements taken from monitoring stations, satellite observed data, and modeled atmospheric data (from an ensemble of models), GAINS uses rela-

tionships from EMEP and CHIMERE models (see Sect. 2.3). Since the CAMS data use satellite imagery, they include many natural sources that may not be easily observed by models, such as sea salt, desert dust, and wildfires.

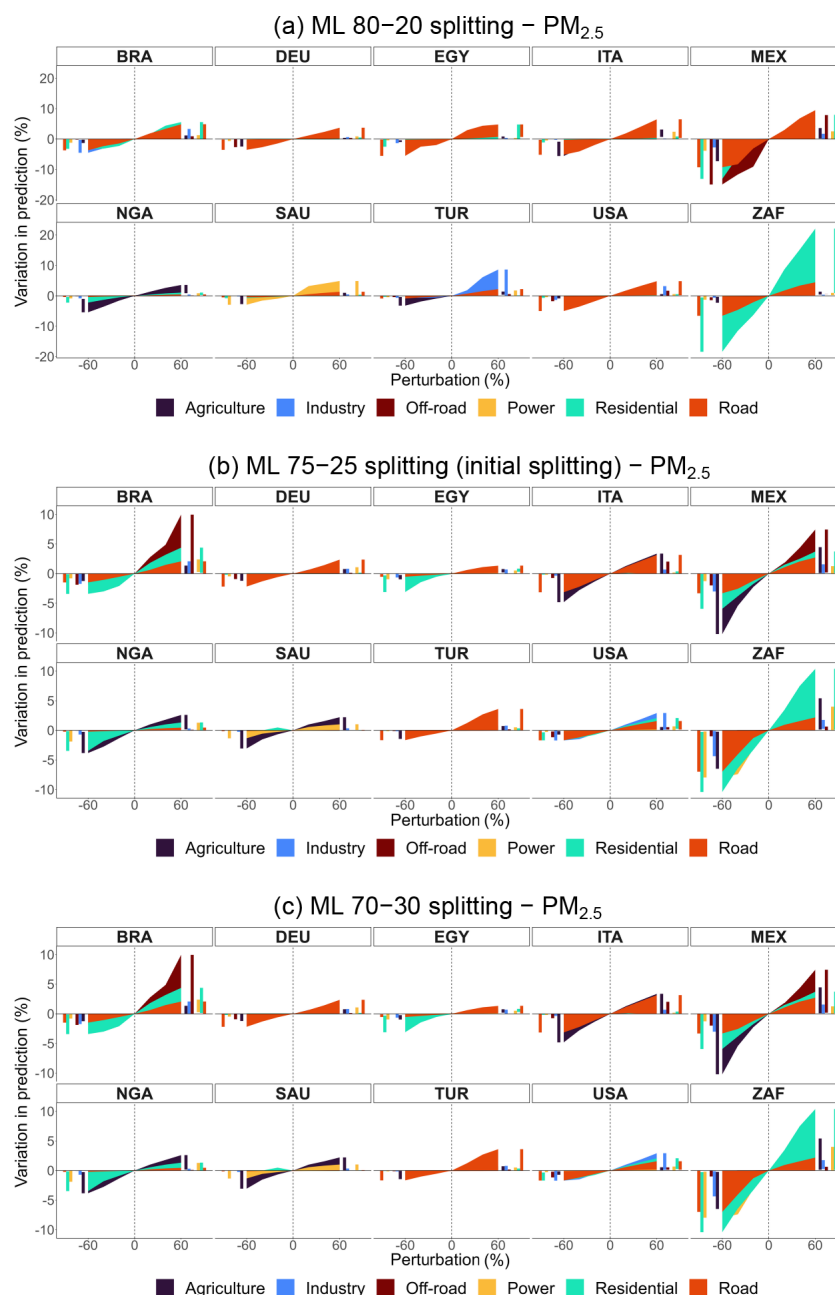


Figure A6. Percentage variation in predicted concentrations of ML models for weighted $\text{PM}_{2.5}$ obtained from CAMS emissions, by sector, pollutant, and perturbation for selected countries: 80–20, 75–25 (initial splitting), and 70–30 train–test splitting.

A6.2 Comparison with the TM5-FASST model

The TM5-FASST model¹⁶ is a reduced-form air quality source–receptor model at the global scale constructed by the JRC. CLAQC model comparison is applied only to TM5-FASST single-country regions among the 56 regions available.

¹⁶This model is derived from "spreadsheet FASST V1.2 NORMALIZED".

TM5-FASST concentrations are expressed as population-weighted $\text{PM}_{2.5}$ in $\mu\text{g m}^{-3}$, including dust and sea salt. Thus, they are directly comparable with CAMS-CLAQC yearly population-weighted concentrations, i.e., CLAQC model's outcomes aggregated at the annual level. We use TM5-FASST annual pollutant total emissions.¹⁷

¹⁷Total emissions consist of IPCC Fifth Assessment's Representative Concentration Pathways (RCP) (Lamarque et al., 2010).

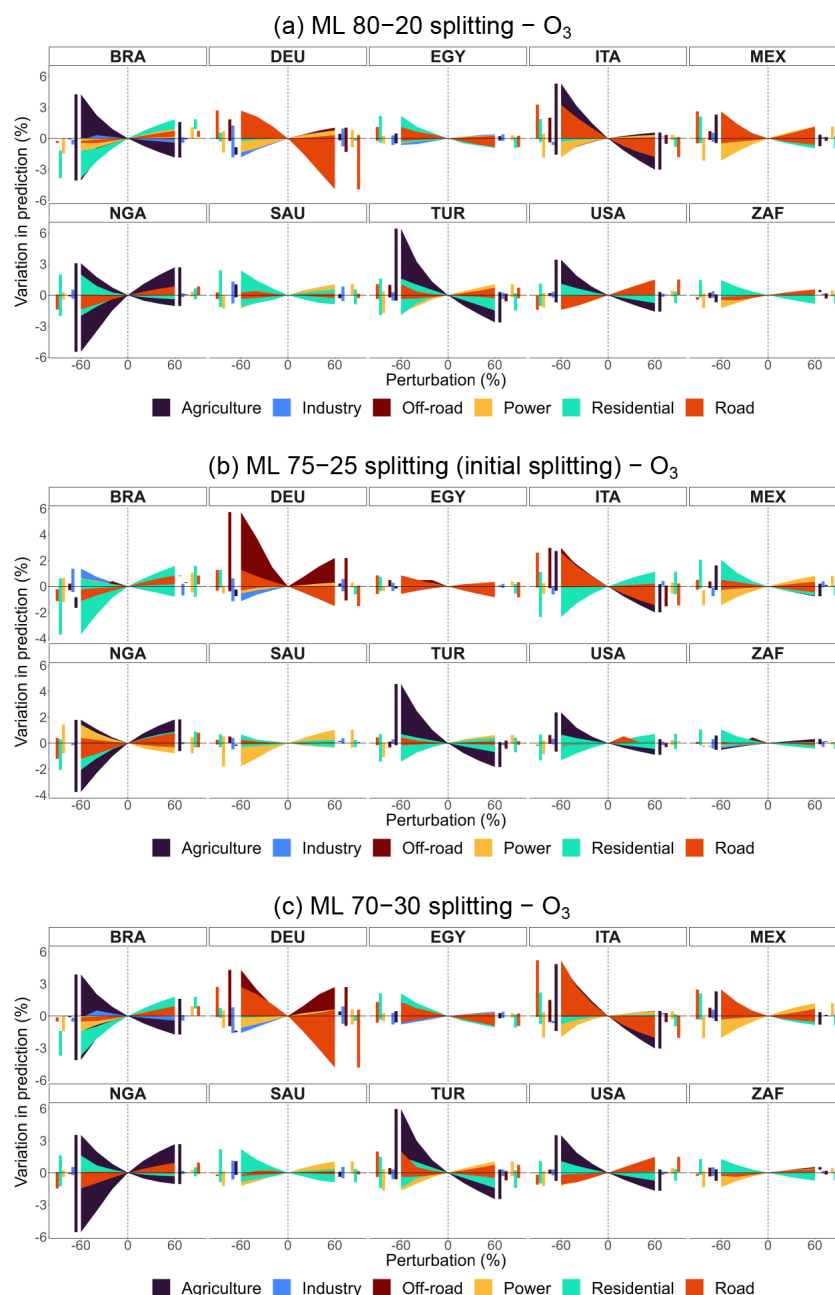


Figure A7. Percentage variation in predicted concentrations of ML models for O_3 obtained from CAMS emissions, by sector, pollutant, and perturbation for selected countries: 80–20, 75–25 (initial splitting), and 70–30 train–test splitting.

We implement the 2015 CAMS emissions into the TM5-FASST scenario, aggregating the CAMS monthly sectoral emissions by pollutant and year and comparing its predictions with CAMS emissions inputted into CLAQC models. As a result, in the case of $\text{PM}_{2.5}$ exposure, TM5-FASST with CAMS-CLAQC emissions predicts lower values compared to CLAQC emissions into the CLAQC model (see Fig. A11a). The latter ones range between 5.7 and $55.7 \mu\text{g m}^{-3}$, with a median of $26.6 \mu\text{g m}^{-3}$, while predictions from CAMS-CLAQC emissions into TM5-FASST

range between 1.3 and $24 \mu\text{g m}^{-3}$, with a median of $7.4 \mu\text{g m}^{-3}$. Differently, in the case of O_3 exposure¹⁸, CAMS-CLAQC emissions into the CLAQC model predict lower concentrations compared to the TM5-FASST model, as detailed in Fig. A11b. Specifically, CLAQC predicted exposures from CAMS-CLAQC emissions range between 5.2 and $55.7 \mu\text{g m}^{-3}$ (median of $22.1 \mu\text{g m}^{-3}$), while TM5-

¹⁸ O_3 exposure is converted from 6mDMA8 expressed in parts per billion into O_3 mean exposure in $\mu\text{g m}^{-3}$.

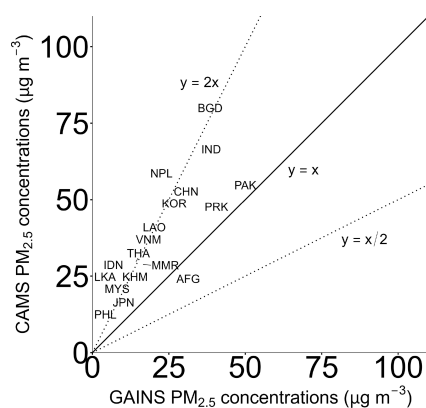


Figure A8. Country-level annual concentrations of $\text{PM}_{2.5}$ in Asia in 2015 from CAMS and GAINS datasets ($\mu\text{g m}^{-3}$). The dotted lines represent the following factor differences between models: $y = 2x$ and $y = \frac{x}{2}$.

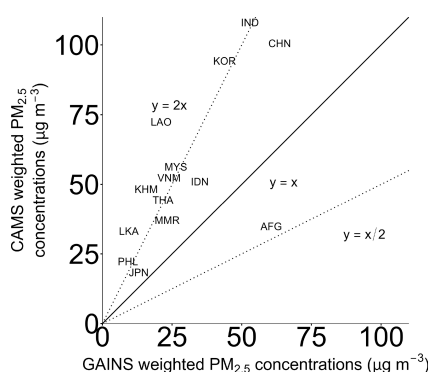


Figure A9. Country-level annual population-weighted concentrations of $\text{PM}_{2.5}$ in Asia in 2015 from CAMS and GAINS datasets ($\mu\text{g m}^{-3}$).

FASST values range between 47.8 and $140 \mu\text{g m}^{-3}$ (median of $114 \mu\text{g m}^{-3}$).

A7 Emission scenarios

A7.1 Emission scenarios of models from DACCIWA emissions

In this section, we present the stylized emission scenarios generated using DACCIWA emissions in both EN and ML models for a subset of countries: the Democratic Republic of the Congo (COD), Egypt (EGY), Ethiopia (ETH), Kenya (KEN), Nigeria (NGA), Tanzania (TZA), Uganda (UGA), and South Africa (ZAF) (see Figs. A12, A13, and A14).

As in the case of models derived from CAMS emissions, EN models from DACCIWA emissions shown in Fig. A12 exhibit greater sectoral variability compared to ML models. In both methods, the residential sector emerges as a significant contributor to $\text{PM}_{2.5}$ concentrations. The transport and industry sectors are consistently present in all ML models

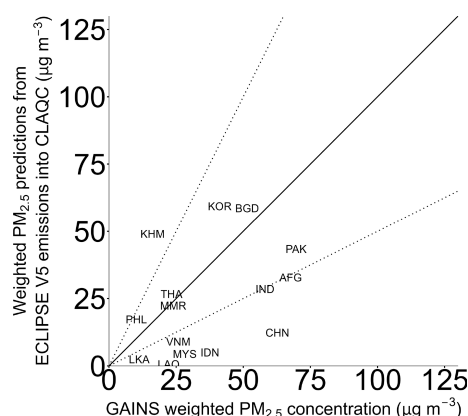


Figure A10. Country-level annual population-weighted concentrations of $\text{PM}_{2.5}$ in Asia in 2020 from CLAQC and re-scaled GAINS datasets ($\mu\text{g m}^{-3}$).

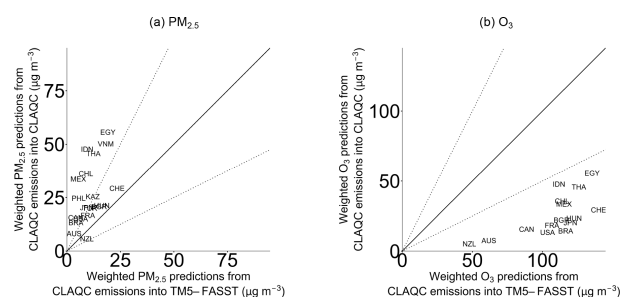


Figure A11. Country-level annual population-weighted concentrations of $\text{PM}_{2.5}$ and O_3 from CLAQC emissions into CLAQC and CLAQC emissions into TMS-FASST ($\mu\text{g m}^{-3}$).

considered, though with relatively lower weights than the residential sector. Also, the power sector has a minimal impact on concentrations, except for ZAF. Among the EN models, the power sector is influential in five out of eight countries, while in ML models, it notably contributes to concentrations in only one country (ZAF). Regarding O_3 , the industry and power sectors exhibit higher contributions in COD, EGY, KEN, and ZAF in EN models, while ML models consistently include the residential, transport, and industry sectors. In most cases, both EN and ML models capture the same relationship between emissions and concentrations, though with different magnitudes. However, there are instances where ML and EN models present contrasting associations between sectors and concentrations. For example, in the EN model for COD, an increase in industrial emissions corresponds to an increase in concentrations, while the ML model indicates a decrease.

In Figs. A13 and A14, we show the variation in percentages in predicted concentrations of EN and ML models (without pixel detail) obtained from CAMS emissions, not only by sector and perturbation but also by precursor, for selected countries.

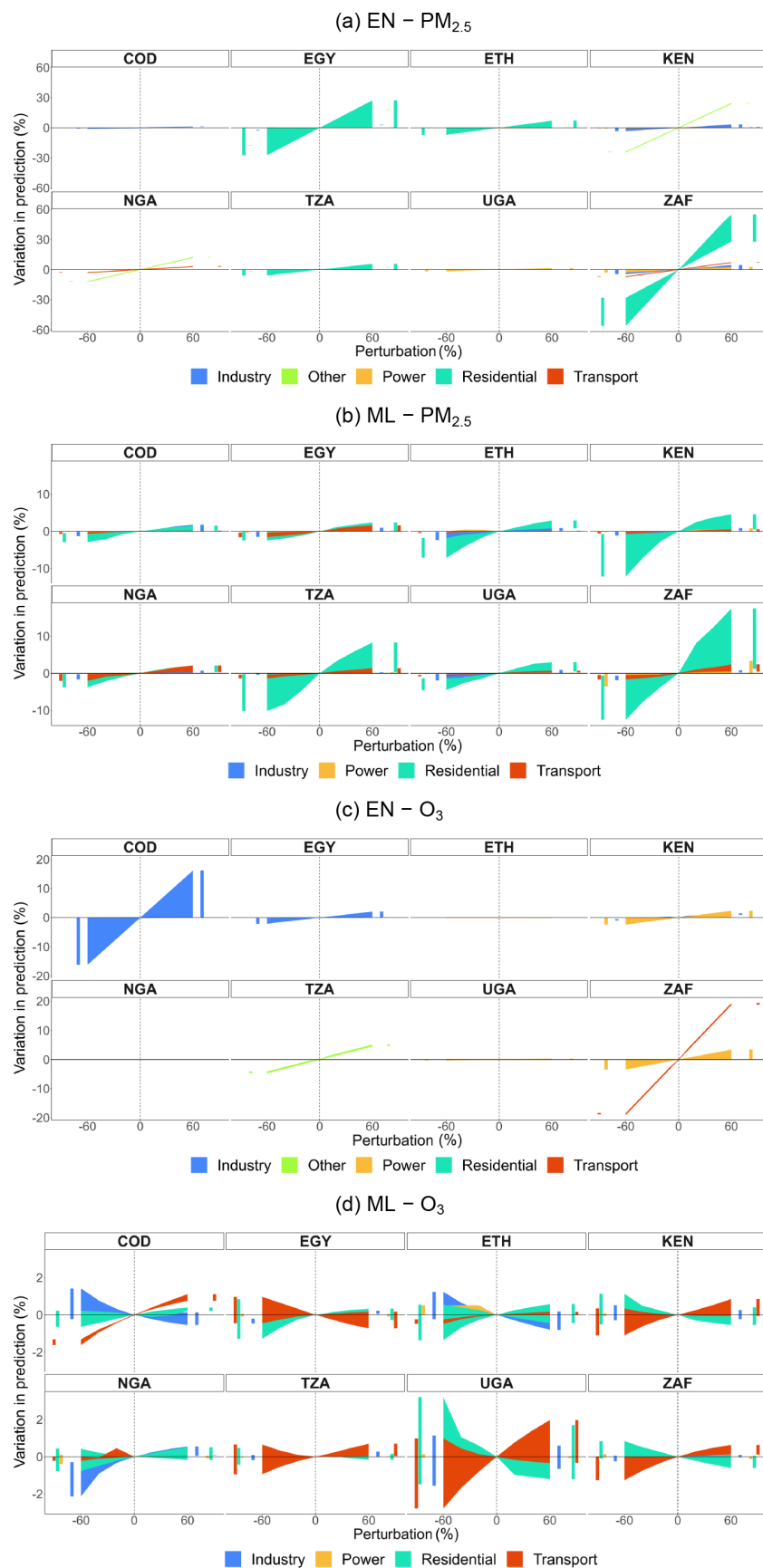


Figure A12. Percentage variation in predicted concentrations by sector and perturbation for selected countries in EN and ML models from DACCIWA emissions for PM_{2.5} and O₃ (without pixel detail). Bar charts on the sides of each subplot help visualize overlapping variations.

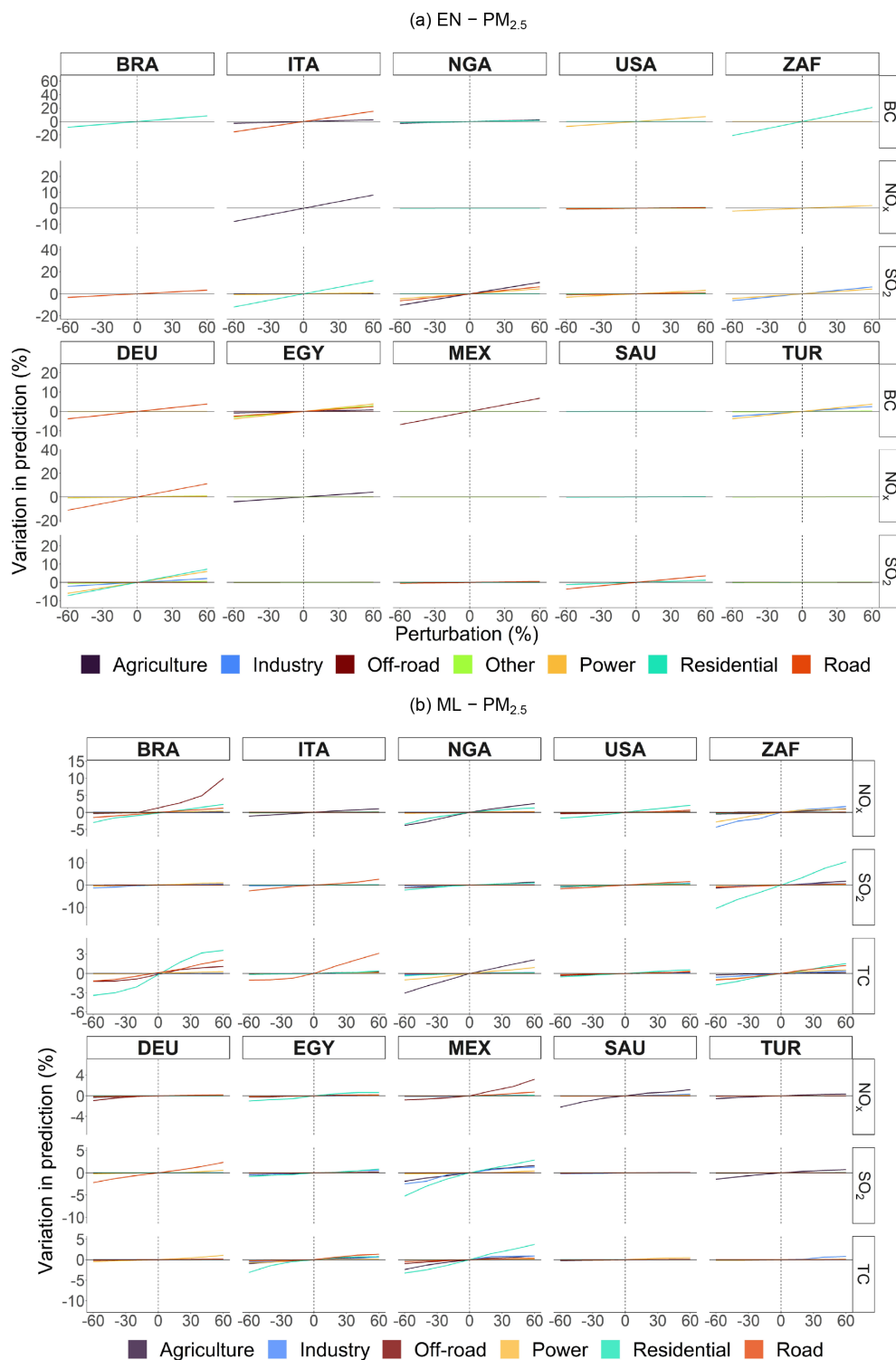


Figure A13. Percentage variation in predicted concentrations of EN and ML models (without pixel detail) for PM_{2.5} obtained from CAMS emissions by sector, pollutant, and perturbation for selected countries.

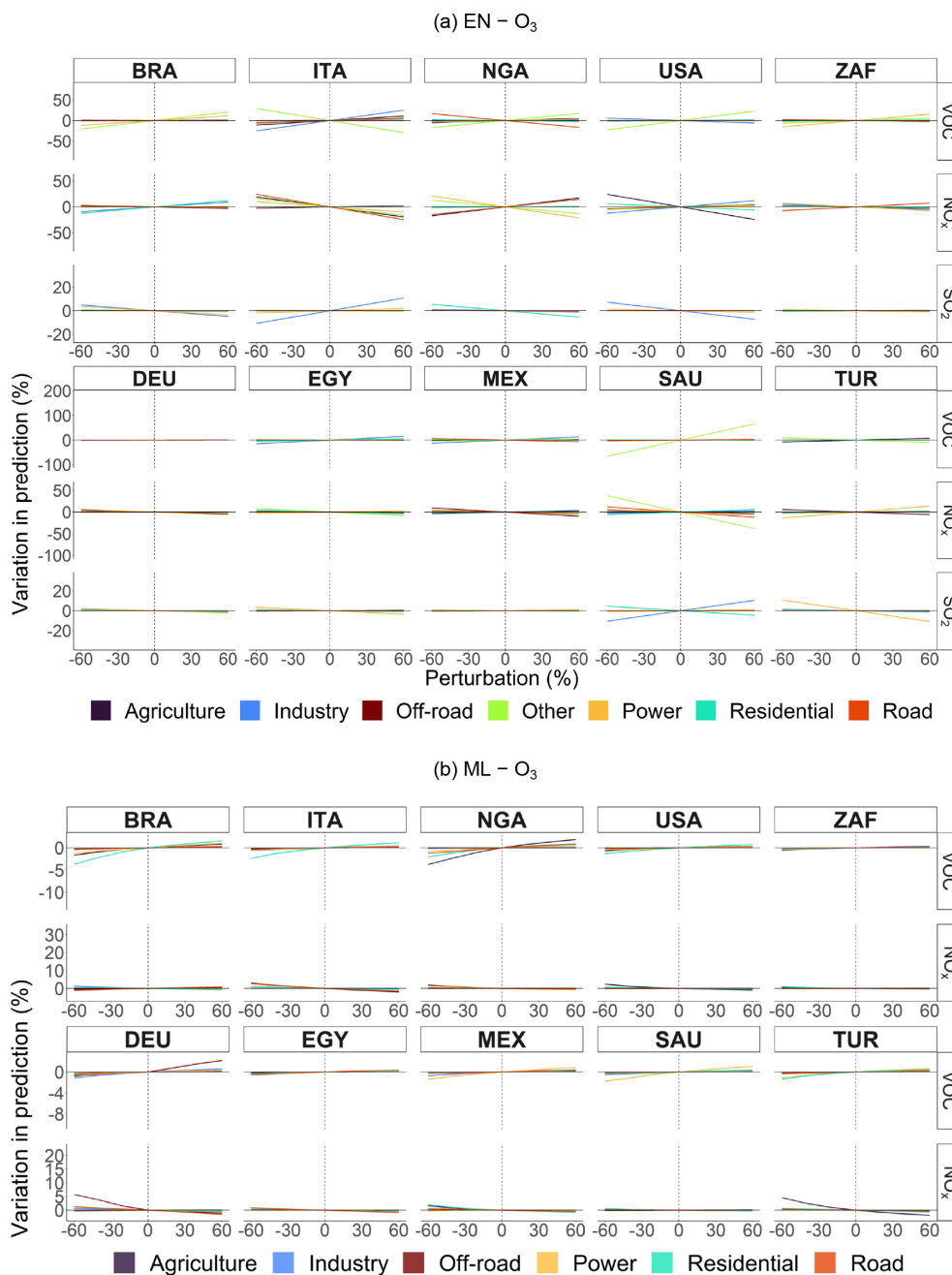


Figure A14. Percentage variation in predicted concentrations of EN and ML models (without pixel detail) for O₃ obtained from CAMS emissions by sector, pollutant, and perturbation for selected countries.

A7.2 Emission scenarios of ML models with pixel detail

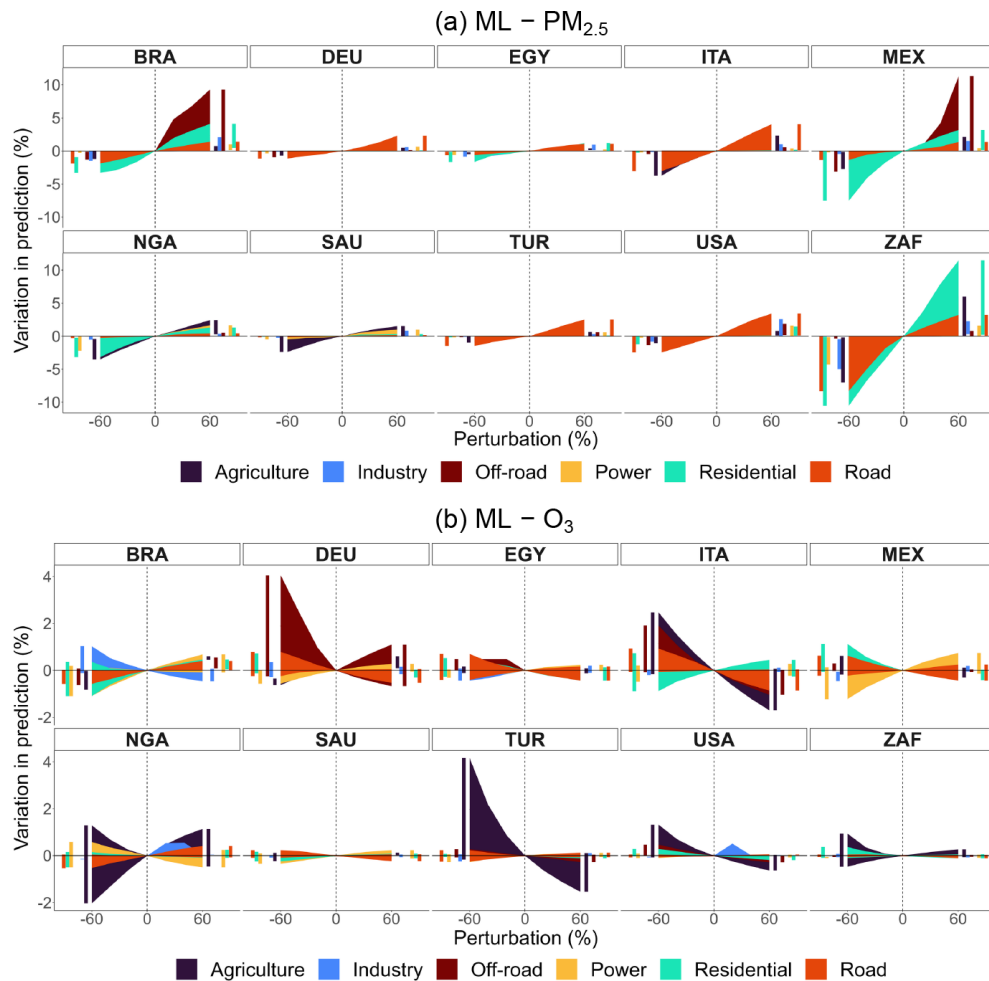


Figure A15. Percentage variation in predicted concentrations by sector and perturbation for selected countries in ML models (with pixel detail) obtained from CAMS emissions. Bar charts on the sides of each subplot help visualize overlapping variations.

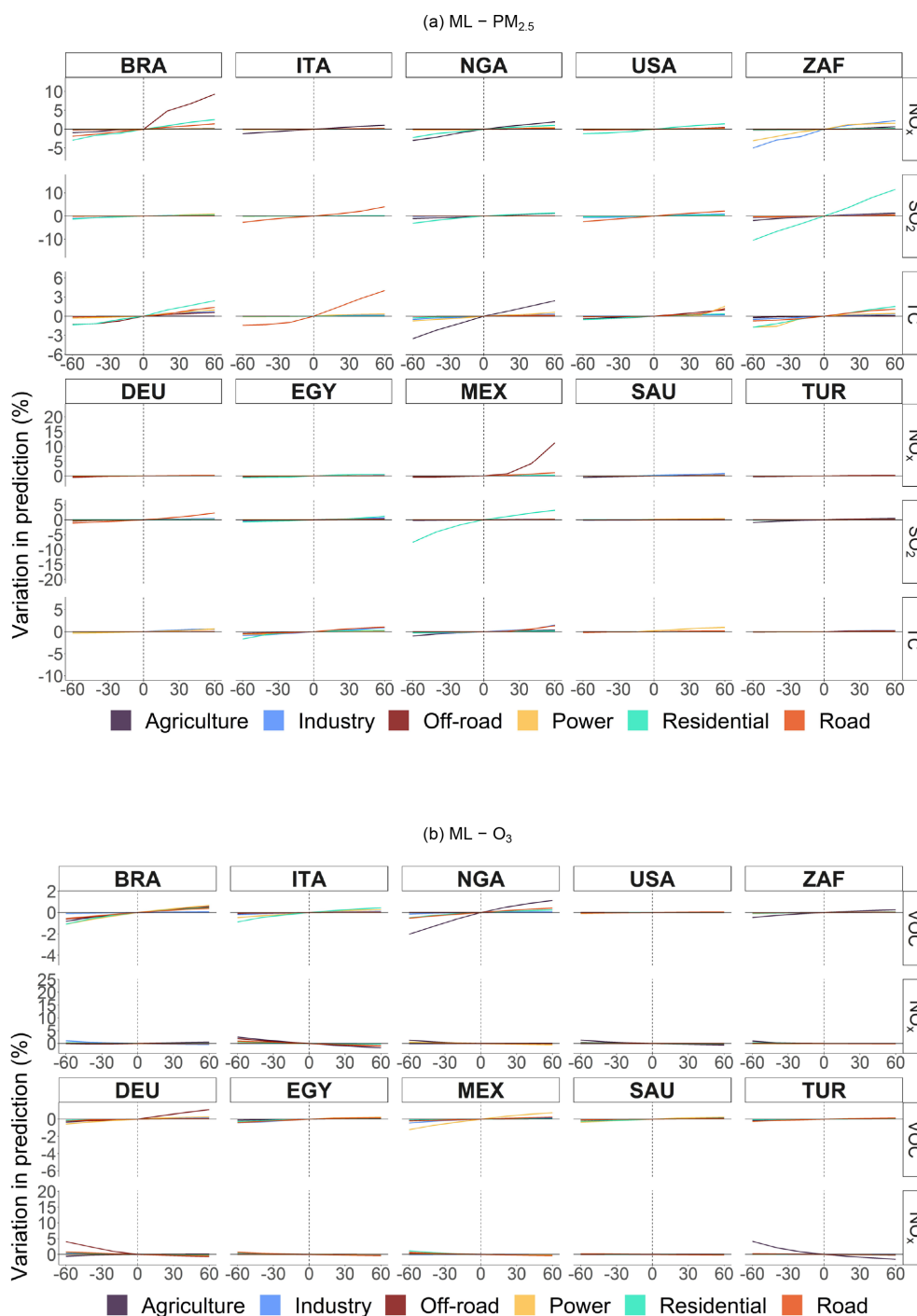


Figure A16. Percentage variation in predicted concentrations of ML models (with pixel detail) obtained from CAMS emissions by sector, pollutant, and perturbation for selected countries.

In Figs. A15 and A16, we show perturbation plots for ML models with sub-national-level detail. When considering PM_{2.5}, it is observed that both versions of the ML models generally select the same precursors. However, there are slight variations in the prediction outcomes based on changes in these precursors. For instance, in the USA model, all available sectors are chosen in both model versions. However, in

the pixel-level model, the residential and road sectors carry more weight compared to the aggregate version, particularly for positive perturbations. This pattern holds true for models predicting O₃ levels as well. Continuing with the example of the USA case, while for most precursors the contribution seems to be similar among models, the pixel-level model exhibits a reduced relevance of the residential and power sec-

tors in comparison to the country-level model, especially for negative perturbations. On the other hand, the agriculture sector makes a similar contribution in both model versions. Overall, it appears that the ML aggregate version serves as a reliable approximation of the spatially heterogeneous models with pixel detail, indicating its effectiveness in capturing the underlying dynamics.

Code and data availability. All datasets used in CLAQC applications are freely available online. The modeling scripts and the output datasets are openly downloadable at <https://doi.org/10.5281/zenodo.14177055> (Renna et al., 2024).

Author contributions. SR and FG developed the model code and contributed to data curation. SR, FG, and LAR conducted the formal analysis, methodology, validation, and visualization. SR prepared the initial manuscript with input from all co-authors. LAR and PSA led the conceptualization, supervised the project, and provided resources. All authors contributed to manuscript review and editing.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Neither the European Commission nor the ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. The authors would like to thank Dirk Heine, Simon Black, Martin Heger, and Christian Schoder, who were part of the World Bank when this work started, for their collaboration and comments. We also thank two anonymous reviewers for their constructive feedback that significantly improved the analyses following the initial submission. We acknowledge and thank the RFF-CMCC European Institute on Economics and the Environment (RFF-CMCC EIEE) and Bocconi University for providing the logistical platforms to perform this work and its collaborators for all the useful comments and advice, the Emissions of Atmospheric Compounds and Compilation of Ancillary Data (ECCAD-AERIS) portal for the archiving and distribution of the emission and CONFORM data, the Tropospheric Ozone Assessment Report (TOAR) initiative for providing the surface ozone data and analyses shown in Schultz et al. (2017), the International Institute for Applied Systems Analysis (IIASA) for providing ECLIPSE and GAINS data, and the Joint Research Centre (JRC) for TM5-FASST data. This work presents prediction models generated using modified Coper-

nicus Atmosphere Monitoring Service information (2003–2021) downloaded from the Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS) and from the Copernicus Climate Change Service (C3S) information (2003–2021), Climate Data Store (CDS) Catalogue, and using TerraClimate meteorology data.

Financial support. This research has been supported by the World Bank Group and has received funding from the European Union's Horizon Europe research and innovation program under the AdJUST project (grant agreement no. 101069880 – CUP C53C21000280006) and from European Union National Recovery and Resilience Plan (PNRR; Missione 4 – Componente 2 – Avviso 341 del 15/03/2022 – NextGenerationEU) in the framework of the project GRINS – Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP C83C22000890001).

Review statement. This paper was edited by Samuel Remy and reviewed by two anonymous referees.

References

- Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., and Hegewisch, K. C.: TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015, *Scientific Data*, 5, 170191, <https://doi.org/10.1038/sdata.2017.191>, 2018.
- Agostinelli, C. and Lund, U.: R package *circular*: Circular Statistics (version 0.4-94), CA: Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University, Venice, Italy. UL: Department of Statistics, California Polytechnic State University, San Luis Obispo, California, USA, <https://r-forge.r-project.org/projects/circular/> (last access: 27 March 2025), 2022.
- Amann, M., Bertok, I., Borken-Kleefeld, J., Cofala, J., Heyes, C., Höglund-Isaksson, L., Klimont, Z., Nguyen, B., Posch, M., Rafaj, P., Sandler, R., Schöpp, W., Wagner, F., and Winiwarter, W.: Cost-effective control of air quality and greenhouse gases in Europe: Modeling and policy applications, *Environ. Modell. Softw.*, 26, 1489–1501, <https://doi.org/10.1016/j.envsoft.2011.07.012>, 2011.
- Anenberg, S. C., Belova, A., Brandt, J., Fann, N., Greco, S., Guttikunda, S., Heroux, M.-E., Hurley, F., Krzyzanowski, M., Medina, S., Miller, B., Pandey, K., Roos, J., and Dingenen, R. V.: Survey of Ambient Air Pollution Health Risk Assessment Tools, *Risk Anal.*, 36, 1718–1736, <https://doi.org/10.1111/risa.12540>, 2016.
- Baird, C. and Cann, M.: *Chimica ambientale*. Terza edizione italiana condotta sulla quinta edizione americana, Zanichelli, ISBN 9788808173782, 2013.
- Black, S., Parry, I., Mylonas, V., Vernon, N., and Zhunussova, K.: The IMF-World Bank Climate Policy Assessment Tool (CPAT): A Model to Help Countries Mitigate Climate Change, IMF Working Papers, ISBN 9798400242342/1018-5941, 2023.
- Center For International Earth Science Information Network-CIESIN-Columbia University: Gridded Population of the World, Version 4 (GPWv4): Population Count Adjusted to Match

- 2015 Revision of UN WPP Country Totals, Revision 11, NASA Earthdata [data set], <https://www.earthdata.nasa.gov/data/projects/gpw> (last access: 7 April 2025), 2018.
- Chen, T. and Guestrin, C.: XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, USA, 13–17 August 2016, 785–794, <https://doi.org/10.1145/2939672.2939785>, 2016.
- Clappier, A., Thunis, P., Beekmann, M., Putaud, J., and de Meij, A.: Impact of SO_x, NO_x and NH₃ emission reductions on PM_{2.5} concentrations across Europe: Hints for future measure development, *Environ. Int.*, 156, 106699, <https://doi.org/10.1016/j.envint.2021.106699>, 2021.
- Colette, A., Rouil, L., Meleux, F., Lemaire, V., and Raux, B.: Air Control Toolbox (ACT_v1.0): a flexible surrogate model to explore mitigation scenarios in air quality forecasts, *Geosci. Model Dev.*, 15, 1441–1465, <https://doi.org/10.5194/gmd-15-1441-2022>, 2022.
- Copernicus Climate Change Service: ERA5 monthly averaged data on single levels from 1979 to present, Climate Data Service [data set], <https://doi.org/10.24381/CDS.F17050D7>, 2019.
- Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Dentener, F., van Aardenne, J. A., Monni, S., Doering, U., Olivier, J. G. J., Pagliari, V., and Janssens-Maenhout, G.: Gridded emissions of air pollutants for the period 1970–2012 within EDGAR v4.3.2, *Earth Syst. Sci. Data*, 10, 1987–2013, <https://doi.org/10.5194/essd-10-1987-2018>, 2018.
- Denier van der Gon, H., Gauss, M., Granier, C., Arellano, S., Benedictow, A., Darras, S., Dellaert, S., Guevara, M., Jalkanen, J.-P., Krueger, K., Kuenen, J., Liaskoni, M., Liousse, C., Markova, J., Prieto Perez, A., Quack, B., Simpson, D., Sindelarova, K., and Soulie, A.: Documentation of CAMS emission inventory products, Copernicus Atmosphere Monitoring Service, <https://doi.org/10.24380/q2si-ti6i>, 2023.
- Ding, D., Xing, J., Wang, S., Dong, Z., Zhang, F., Liu, S., and Hao, J.: Optimization of a NO_x and VOC Cooperative Control Strategy Based on Clean Air Benefits, *Environ. Sci. Technol.*, 56, 739–749, <https://doi.org/10.1021/acs.est.1c04201>, 2021.
- Doumbia, T., Granier, C., Elguindi, N., Bouarar, I., Darras, S., Brasseur, G., Gaubert, B., Liu, Y., Shi, X., Stavrakou, T., Tilmes, S., Lacey, F., Deroubaix, A., and Wang, T.: Changes in global air pollutant emissions during the COVID-19 pandemic: a dataset for atmospheric modeling, *Earth Syst. Sci. Data*, 13, 4191–4206, <https://doi.org/10.5194/essd-13-4191-2021>, 2021.
- Eastham, S. D., Monier, E., Rothenberg, D., Paltsev, S., and Selin, N. E.: Rapid Estimation of Climate – Air Quality Interactions in Integrated Assessment Using a Response Surface Model, *ACS Environmental Au*, 3, 153–163, <https://doi.org/10.1021/acsenvironau.2c00054>, 2023.
- Friedman, J., Hastie, T., and Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent, *J. Stat. Softw.*, 33, 1–22, <https://doi.org/10.18637/jss.v033.i01>, 2010.
- Friedman, J. H.: Greedy function approximation: A gradient boosting machine, *Ann. Stat.*, 29, 1189–1232, <https://doi.org/10.1214/aos/1013203451>, 2001.
- Friedman, J. H., Hastie, T. J., Tibshirani, R., Narasimhan, B., Tay, K., and Simon, N.: Lasso and Elastic-Net Regularized Generalized Linear Models, R package glmnet version 4.0-2, <https://cran.r-project.org/web/packages/glmnet/glmnet.pdf> (last access: 27 March 2025), 2020.
- Fuller, R., Landrigan, P. J., Balakrishnan, K., Bathian, G., Bose-O'Reilly, S., Brauer, M., Caravanos, J., Chiles, T., Cohen, A., Corra, L., Cropper, M., Ferraro, G., Hanna, J., Hanrahan, D., Hu, H., Hunter, D., Janata, G., Kupka, R., Lanphear, B., Lichtveld, M., Martin, K., Mustapha, A., Sanchez-Triana, E., Sandilya, K., Schaeffli, L., Shaw, J., Seddon, J., Suk, W., Téllez-Rojo, M. M., and Yan, C.: Pollution and health: a progress update, *The Lancet Planetary Health*, 6, e535–e547, [https://doi.org/10.1016/s2542-5196\(22\)00090-0](https://doi.org/10.1016/s2542-5196(22)00090-0), 2022.
- Granier, C., Darras, S., Denier van der Gon, H., Doubalova, J., Elguindi, N., Galle, B., Gauss, M., Guevara, M., Jalkanen, J.-P., Kuenen, J., Liousse, C., Quack, B., Simpson, D., and Sindelarova, K.: The Copernicus Atmosphere Monitoring Service global and regional emissions (April 2019 version), Copernicus Atmosphere Monitoring Service, <https://doi.org/10.24380/D0BN-KX16>, 2019.
- Guérette, E.-A., Chang, L. T.-C., Cope, M. E., Duc, H. N., Emmerston, K. M., Monk, K., Rayner, P. J., Scorgie, Y., Silver, J. D., Simmons, J., Trieu, T., Utembe, S. R., Zhang, Y., and Paton-Walsh, C.: Evaluation of Regional Air Quality Models over Sydney, Australia: Part 2, Comparison of PM_{2.5} and Ozone, *Atmosphere*, 11, 233, <https://doi.org/10.3390/atmos11030233>, 2020.
- Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning, Springer New York, <https://doi.org/10.1007/978-0-387-84858-7>, 2009.
- Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., Seibert, J. J., Vu, L., Andres, R. J., Bolt, R. M., Bond, T. C., Dawidowski, L., Kholod, N., Kurokawa, J.-I., Li, M., Liu, L., Lu, Z., Moura, M. C. P., O'Rourke, P. R., and Zhang, Q.: Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the Community Emissions Data System (CEDS), *Geosci. Model Dev.*, 11, 369–408, <https://doi.org/10.5194/gmd-11-369-2018>, 2018.
- Huang, G., Brook, R., Crippa, M., Janssens-Maenhout, G., Schieberle, C., Dore, C., Guizzardi, D., Muntean, M., Schaaf, E., and Friedrich, R.: Speciation of anthropogenic emissions of non-methane volatile organic compounds: a global gridded data set for 1970–2012, *Atmos. Chem. Phys.*, 17, 7683–7701, <https://doi.org/10.5194/acp-17-7683-2017>, 2017.
- IIASA: GAINS Online: Tutorial for advanced users, International Institute for Applied Systems Analysis, <http://web.archive.iiasa.ac.at/rains/reports/GAINS-tutorial.pdf> (last access: 27 March 2025), 2009.
- Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M., and Suttie, M.: The CAMS reanalysis of atmospheric composition, *Atmos. Chem. Phys.*, 19, 3515–3556, <https://doi.org/10.5194/acp-19-3515-2019>, 2019.
- Institute for Health Metrics and Evaluation (IHME): Global Health Data Exchange – Global Burden of Disease Study 2019 (GBD 2019), <https://ghdx.healthdata.org/gbd-2019> (last access: 27 March 2025), 2019.
- Jhun, I., Coull, B. A., Schwartz, J., Hubbell, B., and Koutrakis, P.: The impact of weather changes on air quality and health in

- the United States in 1994–2012, *Environ. Res. Lett.*, 10, 084009, <https://doi.org/10.1088/1748-9326/10/8/084009>, 2015.
- Keita, S., Liousse, C., Assamoi, E.-M., Doumbia, T., N'Datchoh Touré, E., Gnamien, S., Elguindi, N., Granier, C., and Yoboué, V.: African Anthropogenic Emissions Inventory for gases and particles from 1990 to 2015, DACCWA, <https://doi.org/10.25326/56>, 2017.
- Keita, S., Liousse, C., Assamoi, E.-M., Doumbia, T., N'Datchoh, E. T., Gnamien, S., Elguindi, N., Granier, C., and Yoboué, V.: African anthropogenic emissions inventory for gases and particles from 1990 to 2015, *Earth Syst. Sci. Data*, 13, 3691–3705, <https://doi.org/10.5194/essd-13-3691-2021>, 2021.
- Kiesewetter, G., Borken-Kleefeld, J., Schöpp, W., Heyes, C., Thunis, P., Bessagnet, B., Terrenoire, E., Fagerli, H., Nyiri, A., and Amann, M.: Modelling street level PM₁₀ concentrations across Europe: source apportionment and possible futures, *Atmos. Chem. Phys.*, 15, 1539–1553, <https://doi.org/10.5194/acp-15-1539-2015>, 2015.
- Kleiman, G., Anenberg, S. C., Chafe, Z. A., Appiah, D. C., Assefa, T., Bizberg, A., Coombes, T., Cuestas, D., Henze, D. K., Kessler, A., Kheirbek, I., Kinney, P., Mahlatji, M., Marshall, J. D., Naidoo, S., Potwana, N., Rodriguez, A., Tessum, C. W., and Thomas, C.: Enhanced Integration of Health, Climate, and Air Quality Management Planning at the Urban Scale, *Frontiers in Sustainable Cities*, 4, 934672, <https://doi.org/10.3389/frsc.2022.934672>, 2022.
- Kurokawa, J. and Ohara, T.: Long-term historical trends in air pollutant emissions in Asia: Regional Emission inventory in ASIA (REAS) version 3, *Atmos. Chem. Phys.*, 20, 12761–12793, <https://doi.org/10.5194/acp-20-12761-2020>, 2020.
- Lamarque, J.-F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z., Lee, D., Liousse, C., Mieville, A., Owen, B., Schultz, M. G., Shindell, D., Smith, S. J., Stehfest, E., Van Aardenne, J., Cooper, O. R., Kainuma, M., Mahowald, N., McConnell, J. R., Naik, V., Riahi, K., and van Vuuren, D. P.: Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application, *Atmos. Chem. Phys.*, 10, 7017–7039, <https://doi.org/10.5194/acp-10-7017-2010>, 2010.
- Li, M., Liu, H., Geng, G., Hong, C., Liu, F., Song, Y., Tong, D., Zheng, B., Cui, H., Man, H., Zhang, Q., and He, K.: Anthropogenic emission inventories in China: a review, *Natl. Sci. Rev.*, 4, 834–866, <https://doi.org/10.1093/nsr/nwx150>, 2017.
- Lu, X., Zhang, L., and Shen, L.: Meteorology and Climate Influences on Tropospheric Ozone: a Review of Natural Sources, Chemistry, and Transport Patterns, *Current Pollution Reports*, 5, 238–260, <https://doi.org/10.1007/s40726-019-00118-3>, 2019.
- Ma, J., Ding, Y., Cheng, J. C., Jiang, F., Tan, Y., Gan, V. J., and Wan, Z.: Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, *J. Clean. Prod.*, 244, 118955, <https://doi.org/10.1016/j.jclepro.2019.118955>, 2020.
- Mahto, A.: splitstackshape: Stack and Reshape Datasets After Splitting Concatenated Values, r package version 1.4.8, <https://CRAN.R-project.org/package=splitstackshape> (last access: 27 March 2025), 2019.
- McDuffie, E. E., Smith, S. J., O'Rourke, P., Tibrewal, K., Venkataraman, C., Marais, E. A., Zheng, B., Crippa, M., Brauer, M., and Martin, R. V.: A global anthropogenic emission inventory of atmospheric pollutants from sector- and fuel-specific sources (1970–2017): an application of the Community Emissions Data System (CEDS), *Earth Syst. Sci. Data*, 12, 3413–3442, <https://doi.org/10.5194/essd-12-3413-2020>, 2020.
- McDuffie, E. E., Martin, R. V., Spadaro, J. V., Burnett, R., Smith, S. J., O'Rourke, P., Hammer, M. S., van Donkelaar, A., Bindle, L., Shah, V., Jaeglé, L., Luo, G., Yu, F., Adeniran, J. A., Lin, J., and Brauer, M.: Source sector and fuel contributions to ambient PM_{2.5} and attributable mortality across multiple spatial scales, *Nat. Commun.*, 12, 3594, <https://doi.org/10.1038/s41467-021-23853-y>, 2021.
- Menut, L., Bessagnet, B., Briant, R., Cholakian, A., Couvidat, F., Mailler, S., Pennel, R., Siour, G., Tuccella, P., Turquet, S., and Valari, M.: The CHIMERE v2020r1 online chemistry-transport model, *Geosci. Model Dev.*, 14, 6781–6811, <https://doi.org/10.5194/gmd-14-6781-2021>, 2021.
- Murray, C. J. L., Aravkin, A. Y., Zheng, P., et al.: Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019, *The Lancet*, 396, 1223–1249, [https://doi.org/10.1016/s0140-6736\(20\)30752-2](https://doi.org/10.1016/s0140-6736(20)30752-2), 2020.
- OpenAQ: OpenAQ, <https://openaq.org/> (last access: 27 March 2025), 2024.
- Oxley, T., Vieno, M., Woodward, H., ApSimon, H. M., Mehlig, D., Beck, R., Nemitz, E., and Reis, S.: Reduced-Form and Complex ACTM Modelling for Air Quality Policy Development: A Model Inter-Comparison, *Social Science Research Network (SSRN)*, <https://doi.org/10.2139/ssrn.4158187>, 2022.
- R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/> (last access: 27 March 2025), 2020.
- Reis, L. A., Drouet, L., and Tavoni, M.: Internalising health-economic impacts of air pollution into climate policy: a global modelling study, *The Lancet Planetary Health*, 6, e40–e48, [https://doi.org/10.1016/s2542-5196\(21\)00259-x](https://doi.org/10.1016/s2542-5196(21)00259-x), 2022.
- Renna, S., Granella, F., Aleluia Reis, L., and Schulz-Antipa, P.: Code and output data of “CLAQC v1.0 – Country Level Air Quality Calculator. An empirical modeling approach”, Zenodo [code and data set], <https://doi.org/10.5281/zenodo.14177055>, 2024.
- Schultz, M. G., Schröder, S., Lyapina, O., Cooper, O., Galbally, I., Petropavlovskikh, I., Schneidmesser, E. V., Tanimoto, H., Elshorbany, Y., Naja, M., Seguel, R., Dauert, U., Eckhardt, P., Feigenspahn, S., Fiebig, M., Hjellbrekke, A.-G., Hong, Y.-D., Kjeld, P. C., Koide, H., Lear, G., Tarasick, D., Ueno, M., Wallasch, M., Baumgardner, D., Chuang, M.-T., Gillett, R., Lee, M., Molloy, S., Moolia, R., Wang, T., Sharps, K., Adame, J. A., Ancellet, G., Apadula, F., Artaxo, P., Barlasina, M., Bogucka, M., Bonasoni, P., Chang, L., Colomb, A., Cuevas, E., Cupeiro, M., Degorska, A., Ding, A., Fröhlich, M., Frolova, M., Gadhavi, H., Gheusi, F., Gilge, S., Gonzalez, M. Y., Gros, V., Hamad, S. H., Helmig, D., Henriques, D., Hermansen, O., Holla, R., Huber, J., Im, U., Jaffe, D. A., Komala, N., Kubistin, D., Lam, K.-S., Laurila, T., Lee, H., Levy, I., Mazzoleni, C., Mazzoleni, L., McClure-Begley, A., Mohamad, M., Murovic, M., Navarro-Comas, M., Nicodim, F., Parrish, D., Read, K. A., Reid, N., Ries, L., Saxena, P., Schwab, J. J., Scorgie, Y., Senik, I., Simmonds, P., Sinha, V., Skorokhod, A., Spain, G., Spangl, W., Spoor, R., Springston, S. R., Steer, K., Steinbacher, M., Suhar-

- guniyawan, E., Torre, P., Trickl, T., Weili, L., Weller, R., Xu, X., Xue, L., and Zhiqiang, M.: Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations, *Elementa: Science of the Anthropocene*, 5, 58, <https://doi.org/10.1525/elementa.244>, 2017.
- Seinfeld, J. H. and Pandis, S. N.: *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, WILEY, ISBN 1118947401, https://www.ebook.de/de/product/25599491/john_h_seinfeld_spyros_n_pandis_atmospheric_chemistry_and_physics_from_air_pollution_to_climate_change.html (last access: 27 March 2025), 2016.
- Stohl, A., Aamaas, B., Amann, M., Baker, L. H., Bellouin, N., Berntsen, T. K., Boucher, O., Cherian, R., Collins, W., Daskalakis, N., Dusinska, M., Eckhardt, S., Fuglestad, J. S., Harju, M., Heyes, C., Hodnebrog, Ø., Hao, J., Im, U., Kanakidou, M., Klimont, Z., Kupiainen, K., Law, K. S., Lund, M. T., Maas, R., MacIntosh, C. R., Myhre, G., Myriokefalitakis, S., Olivie, D., Quaas, J., Quennehen, B., Raut, J.-C., Rumbold, S. T., Samset, B. H., Schulz, M., Seland, Ø., Shine, K. P., Skeie, R. B., Wang, S., Yttri, K. E., and Zhu, T.: Evaluating the climate and air quality impacts of short-lived pollutants, *Atmos. Chem. Phys.*, 15, 10529–10566, <https://doi.org/10.5194/acp-15-10529-2015>, 2015.
- Thakrar, S. K., Tessum, C. W., Apte, J. S., Balasubramanian, S., Millet, D. B., Pandis, S. N., Marshall, J. D., and Hill, J. D.: Global, high-resolution, reduced-complexity air quality modeling for PM_{2.5} using InMAP (Intervention Model for Air Pollution), *PLOS ONE*, 17, e0268714, <https://doi.org/10.1371/journal.pone.0268714>, 2022.
- Thunis, P., Degraeuwe, B., Pisoni, E., Ferrari, F., and Clappier, A.: On the design and assessment of regional air quality plans: The SHERPA approach, *J. Environ. Manage.*, 183, 952–958, <https://doi.org/10.1016/j.jenvman.2016.09.049>, 2016.
- Thunis, P., Clappier, A., Tarrason, L., Cuvelier, C., Monteiro, A., Pisoni, E., Wesseling, J., Belis, C., Pirovano, G., Janssen, S., Guerreiro, C., and Peduzzi, E.: Source apportionment to support air quality planning: Strengths and weaknesses of existing approaches, *Environ. Int.*, 130, 104825, <https://doi.org/10.1016/j.envint.2019.05.019>, 2019.
- Ushey, K.: renv: Project Environments, *r* package version 0.15.2, <https://CRAN.R-project.org/package=renv> (last access: 27 March 2025), 2022.
- Van Dingenen, R., Dentener, F., Crippa, M., Leitao, J., Marmer, E., Rao, S., Solazzo, E., and Valentini, L.: TM5-FASST: a global atmospheric source–receptor model for rapid impact analysis of emission changes on air quality and short-lived climate pollutants, *Atmos. Chem. Phys.*, 18, 16173–16211, <https://doi.org/10.5194/acp-18-16173-2018>, 2018.
- World Health Organization: WHO global air quality guidelines. Particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, Tech. rep., <https://www.who.int/publications/i/item/9789240034228> (last access: 27 March 2025), 2021.
- Zou, H. and Hastie, T.: Regularization and variable selection via the elastic net, *J. R. Stat. Soc. B*, 67, 301–320, <https://doi.org/10.1111/j.1467-9868.2005.00503.x>, 2005.