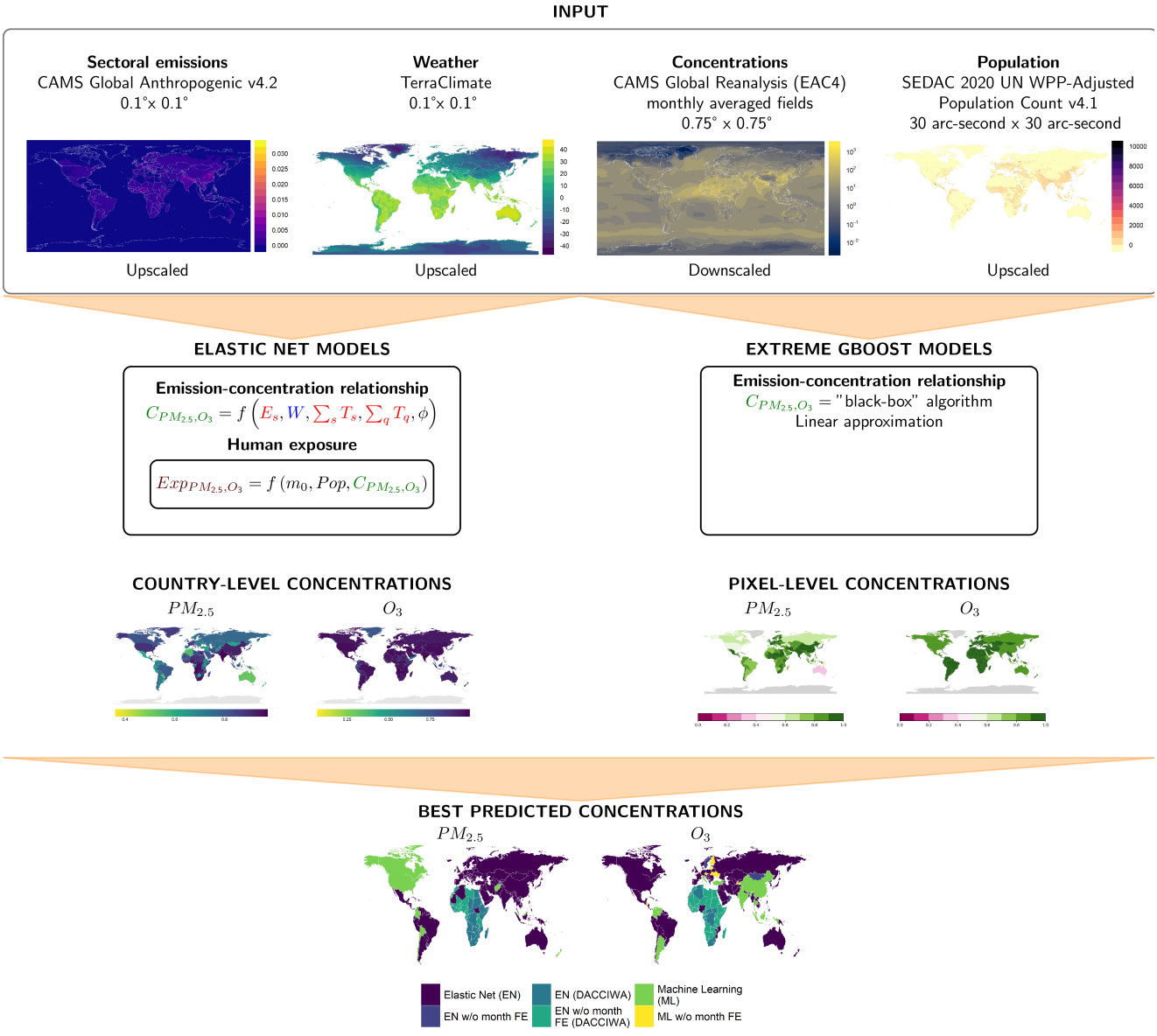
Graphical Abstract

CLAQC v1.0 – Country Level Air Quality Calculator. An empirical modelling approach.

Stefania Renna,Francesco Granella,Lara Aleluia Reis,Paulina Antipa



Highlights

CLAQC v1.0 – Country Level Air Quality Calculator. An empirical modelling approach.

Stefania Renna,Francesco Granella,Lara Aleluia Reis,Paulina Antipa

- We develop a global air quality calculator, with country level detail, based on empirical models to predict monthly and annual concentration levels of $PM_{2.5}$ and O_3 from sectoral emissions and meteorology.
- We use the CAMS emissions and reanalyses products to provide global coverage.
- We apply two different methodologies: (i) Elastic net modelling and (ii) extreme gradient boosting regressor.
- We validate our models out-of-sample and against other models.
- We run sensitivity analyses on the model fit (policy scenarios).
- For each country, we select the model that performed best, based on multiple criteria.

CLAQC v1.0 – Country Level Air Quality Calculator. An empirical modelling approach.⋆

Stefania Renna[a,*], Francesco Granella[a,b], Lara Aleluia Reis[a] and Paulina Antipa[c]

[a]RFF-CMCC European Institute on Economics and the Environment, Milan, Italy
[b]Bocconi University, Milan, Italy
[c]World Bank, Washington, DC, USA

ARTICLE INFO

Keywords:
air pollution
emission-concentration functions
$PM_{2.5}$
O_3
sectoral emissions
elastic net modelling
extreme gradient boosting regressor
sensitivity analysis

ABSTRACT

This paper describes the Country Level Air Quality Calculator (CLAQC). CLAQC is a modelling tool which predicts monthly and annual concentration levels of two major air pollutants, $PM_{2.5}$ and O_3. It takes as input sectoral emissions at the national level. The model is written in open source languages (R and Python) and stored in an open GitHub repository, to allow for transparency and community development. CLAQC builds on the recent advancements of the CAMS system, and uses CAMS global gridded emissions and CAMS reanalysis pollutant concentrations. It was designed to provide insights for national and regional policy support. It is easy-to-use and fast allowing to simulate large sets of scenarios and implementation in optimization frameworks. One of its main advantages, as compared to already existing models of the same type, is that it accounts for the sectoral contributions of several pollutants. We use and compare two different methodological approaches: Elastic net modelling and extreme gradient boosting regressor. We perform out-of-sample validation and simulate a set of conceptualized policy scenarios. We show that the model performs well for the majority of countries and that it can be used for policy support. Finally, we develop an online tool where the model results are available, and a framework for method selection per country.

1. Introduction

Exposure to air pollution is one of the major health concerns, worldwide (Murray et al., 2020). In 2019, according to HIME database,[1] 1 in 9 death worldwide were caused by fine particulate matter ($PM_{2.5}$) and ozone (O_3) air pollution. Of those deaths, 5.7% are due to O_3, and the rest to $PM_{2.5}$. To that extent, it is important to consider the air pollution heath dimension in the design of policies (Reis, Drouet, & Tavoni, 2022) and provide tools that can inform policy. Air pollution is often tackled by local and national policies, but energy and climate policies can also bring co-benefits for air pollution. To understand how to effectively intervene on air pollution, or to derive air pollution co-benefits, it is crucial to be able to estimate the effects of national or sub-national policy interventions on emission reductions. This is especially important in the fields of integrated assessment and of global policy scenario assessment. Amongst the most common models used in integrated assessment are the *TM5-FAst Scenario Screening Tool* (TM5-FASST) model (Dingenen et al., 2018), the SHERPA tool (Thunis, Degraeuwe, Pisoni, Ferrari, & Clappier, 2016), and the *Greenhouse Gas - Air Pollution Interactions and Synergies* (GAINS) model (Amann et al., 2011; Kiesewetter et al., 2015). All such models use different methods, *e.g.*, the SHERPA tool emulates scenario results from other models (such as EMEP and CHIMERE (Menut et al., 2021)), GAINS is a full integrated assessment model using emission-concentration relationships from CHIMERE and EMEP, while TM5-FASST is a source-receptor model based on the chemical transport model TM5 (Dingenen et al., 2018). None of the above can simultaneously provide sectoral contributions to $PM_{2.5}$ and O_3 concentrations for the whole globe, at the country-level. The CLAQC model fills in this gap, learning from historical variations in emissions and concentrations estimated from the CAMS emissions and the CAMS reanalysis measurements of $PM_{2.5}$ and O_3. CLAQC complements the above mentioned models, by providing an easy-to-use tool that is global but has country and sectoral detail. Reduced models trade accuracy and computational efficiency and open the possibility of optimizing policy, ultimately leading to efficient and effective health improvements. However, they are known to be less robust for highly non-linear processes, such as secondary ozone formation and secondary PM (Dingenen et al., 2018; Thunis et al., 2019). Furthermore, they are limited to the number of underlying scenarios in their train set. Our approach provides a complementary method, that relies on observed changes in concentrations given changes in emissions and meteorology. Emissions have patterns that, over the course of 17 years of meteorological variation can

⋆This document is the result of a research project funded by the World Bank.

*Corresponding author
**Principal corresponding author
stefania.renna@eiee.org (S. Renna); francesco.granella@eiee.org (F. Granella); lara.aleluia@eiee.org (L. Aleluia Reis); pschulzantipa@worldbank.org (P. Antipa)
https://laleluia.github.io/page/,www.eiee.org/team-member/lara-aleluia-reis/ (S. Renna); https://francescogranella.github.io/,www.eiee.org/team-member/francesco-granella/ (F. Granella); https://www.eiee.org/team-member/stefania-renna/ (L. Aleluia Reis)
ORCID(s): https://orcid.org/0000-0001-8096-6320 (S. Renna); https://orcid.org/0000-0002-2349-0132 (F. Granella); https://orcid.org/0000-0002-6676-7007 (L. Aleluia Reis)

[1]http://ghdx.healthdata.org/gbd-2019

"teach" us about how changes in emission activities lead to changes in observed concentrations while controlling for the changing weather and time-invariant patterns. A simple example would be the weekday/weekend traffic activity shifts, but these patterns can also be annual: *e.g.*, as local and international policies enter into force, emissions tend to decrease. In CLAQC, unlike in the source-receptor models, the perturbation level is not set (Thunis et al., 2019), while the model learns form all past variations. This allows for more "training scenarios". One disadvantage is that we cannot learn from event that have not happen. However, the COVID-19 pandemic and the consequent worldwide lockdowns have provided high disruptions in many sectors and human activity, giving our model the opportunity to "learn" from these unprecedented emission halts. Therefore, the year 2020 is included in our training set.

2. CLAQC Rational

We are interested in estimating the relationship between emissions E of major ambient air pollutants on the respective ground-level concentrations C of major pollutants c ($PM_{2.5}$, O_3). Denote such relationship f, so that

$$C_c = f(E). \tag{1}$$

The formation, transport and dispersion of pollutants are complex natural phenomena that are highly dependent on emissions, on weather W and other local characteristics such as topography. Hence, the design of pollution abatement policies in country k can benefit from the estimate of a country-specific emission-concentration function that accounts for interactions between emissions and weather:

$$C_{c,k} = f_k(E_k, W_k). \tag{2}$$

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

$$C_{c,k} = f_{k,s,p}(E_{k,s,p}, W_k). \tag{3}$$

We identify two methods to empirically derive $\widehat{f}_{k,s,p}$ that trade off simplicity and transparency.

The first method relies on Elastic Net models, a type of penalized linear regression that is amenable to a large number of predicting variables while preserving an intelligible structure. In this way, $\widehat{f}_{k,s,p}$ is 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 transparency, as machine learning algorithms typically do not return simple predictor-target functions. We thus approximate emission-concentration relationships with functions that are amenable to a spreadsheet-based tool. For completeness, we also present early explorations of the CLAQC tool including ordinary least squares (OLS) and spatial regression models (see Appendix section A TODO).

The next Section discusses the data that have been used (3). We then present in Section 4 the different methodological approaches that have been followed: Elastic net models (Section 4.1); and machine learning models (Section 4.2). In Section 5, we discuss their validation results and sensitivity analyses. In Section 4.4, we compare the two methodological approaches. Finally, in Section 6, we draw conclusions.

3. Data

Despite the recent harmonization and open-access advancements in air pollution data, most of the publicly available global ground-level motoring (GLM) databases are still heavily unbalanced towards the developed world. In such databases (*e.g.*, *openaq*) there is sufficient territorial coverage of the population only in industrialized countries, in particular in the USA and Europe. Monitors in the larger urban areas exist also in China and, at growing pace, in India. However, for the larger part of the developing world, the urban population, and even more so the rural population, does not live within a reasonable distance from GLM. Air quality in Africa, Central Asia, and large parts of Latin America is virtually unmonitored with regulatory-grade instruments. Uneven geographical coverage is problematic, as factors driving the emission-concentration relationship differ between monitored and unmonitored areas. To answer CLAQC's goals the only remaining option is the use of satellite data. Though, there are limitations to this approach, namely there may be discrepancies between ground measurements and satellite readings, due to their different temporal and spacial representativeness. Regulatory-grade ground-level monitors provide the best possible measurement of local concentrations of air pollutants, however they lack information on the extent of their spatial representativeness. Reanalysis data blends and harmonizes satellite measures of pollution with ground-level monitors. While it maintains the quality of monitor data at the location of monitors, it fills the gap of GLM networks with satellite observations and modelling covering the whole globe. The use of gridded data brings about many other advantages:

- 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 emission-concentration functions;
- Reducing rigidity on the spatial scope, keeping the sub-national modelling option flexible.

On the other hand, the need to homogenize different grids may imply approximations incurring in the data manipulation process.

3.1. Emissions

Emission data is provided by CAMS Global Anthropogenic v4.2, with a monthly temporal resolution, and a spatial resolution of 0.1°. Data covers the following sectors and pollutants and is originally expressed in Teragrams (Tg):

- **Sectors:** Agricultural waste burning, Road transportation, Industrial process, Solvents, Power generation, Fugitives, Residential and other sectors, Agriculture livestock, Agriculture soils, Solid waste and waste water, Off Road transportation, Ships.
- **Pollutants:** BC, OC, NH_3, NO_x, SO_2, and NMVOC.

The CAMS emission data is based on existing available databases, including nationally reported emissions, the Emissions Database for Global Atmospheric Research (EDGAR) (Crippa et al., 2018; Huang et al., 2017), Evaluating the Climate and Air Quality Impacts of Short-Lived Pollutants (ECLIPSE) (Stohl et al., 2015), and the Community Emissions Data System (CEDS) databases (Hoesly et al., 2018). It has the advantage of providing global gridded monthly emissions since 2000. See Granier et al. (2019), for more detail.

3.2. Sectoral aggregation

We are particularly interested in including the main sectors that can be directly affected by fuel policy. However, emissions from the different sectors are highly collinear (Clappier, Pisoni, & Thunis, 2015), with the exception of the agricultural sectors, thus we aggregate some sectors together while keeping the most important sectors in terms of policy implementation separated(see Granier et al. (2019) for details on CAMS sectoral definitions):

- AGR (Agriculture)
  - awb Agricultural waste burning
  - agl Agriculture livestock
  - ags Agriculture soils
- INX (Industry)
  - ind Industrial process
- POW (Energy Power generation)
  - ene Power generation
  - fef Fugitives
- OTH (Others, including the emissions not considered in the sectors above)
  - swd Solid waste and waste water
  - slv Solvents
- OTR (Off-road transportation)
  - tnr Off Road transportation
- ROA (Road transportation)
  - tro Road transportation
- SER (Buildings including residential, commercial and services)
  - res Residential and other sectors

For computational simplicity, the machine learning models do not include the "OTH" sector, however this sector is often not directly influences by fuel policies.

3.2.1. Precursors

The precursors of $PM_{2.5}$ included in models are: BC, OC, NH_3, NO_x, SO_2 and NMVOC. All are expected to increase $PM_{2.5}$ concentrations at the country level, although local decreases on the secondary fraction may happen (Clappier, Thunis, Beekmann, Putaud, & de Meij, 2021). Data on OC and BC is almost perfectly collinear. Emissions from these precursors are summed into Total Carbon (TC) in the machine learning models.

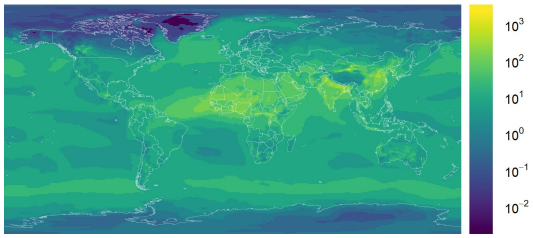
The precursors of O_3 included in the models are: NO_x, NMVOC, and SO_2. NMVOC are expected to increase O_3, whereas the relationship between NO_x and SO_2 to O_3 may be negative (van dingenen et al. 2018).

3.3. Meteorology

All meteorological data, with the exception of wind direction, comes from TerraClimate (Abatzoglou, Dobrowski, Parks, & Hegewisch, 2018). We have selected TerraClimate because it has a wide variety of meteorological variables, good temporal coverage and very high spatial resolution. Data is monthly, with spatial resolution of 0.1°. The following atmospheric variables are used: accumulated precipitation in mm; maximum 2-m temperature in $degC$; minimum 2-m temperature in $degC$; 10-m wind speed in m/s; mean vapor pressure deficit in kPa. Wind direction in $degrees$ comes from ERA-5 Reanalysis Monthly Means (Copernicus Climate Change Service, 2019). All data is converted to the same 0.5° by 0.5° grid for consistency of all the variables.

[2] Model-level 60.

(a) $PM_{2.5}$



(b) O_3

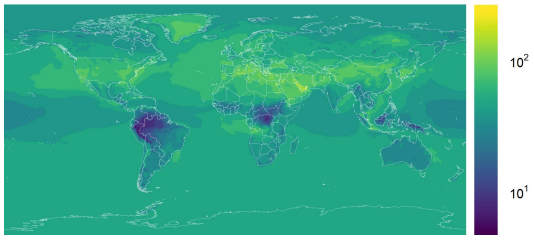


Figure 1: Level plots of EAC4 concentrations of PM2.5 (January 2018) and O3 (July 2018) in log in $\mu g/m^3$.

(a) $PM_{2.5}$ (b) O_3

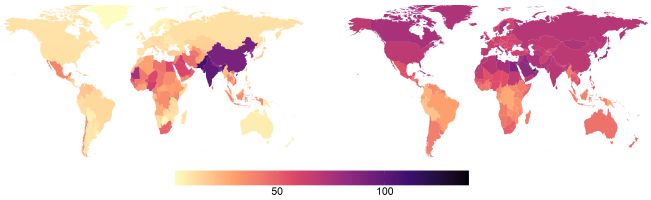


Figure 2: CLAQC weighted concentration inputs of $PM_{2.5}$ and O_3 in $\mu g/m^3$ (2018).

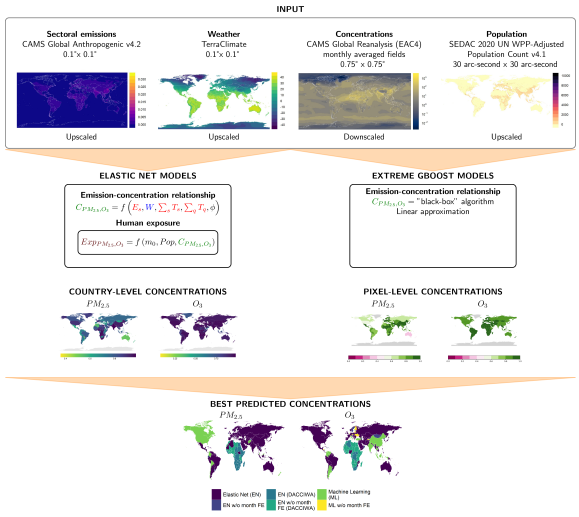


Figure 3: Methodological abstract.

3.4. Concentrations

Ambient concentration data at the ground-level[2] comes from the European Centre for Medium-Range Weather Forecast's (ECMWF) Atmospheric Composition Reanalysis 4 (EAC4) monthly averaged fields (Inness et al., 2019), for the time period of 2003-2020, and $PM_{2.5}$ and O_3 air pollutants. This is the shortest time domain from all the data sets, thus all the other data sets are constrained to such time period. The data are originally at 0.75° of spatial resolution, and are down-scaled to 0.1° with linear interpolation using weights from the Global Burden of Disease (GBD) 2019. O_3 mixing ratio (kg/kg) is converted to $\mu g/m^3$.

EAC4 reanalysis combines model data with observations, *in-situ* and satellite, from all over the world into a globally complete and consistent data set using a model of the atmosphere based on the laws of physics and chemistry. 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 time to collect observations and allowing for the integration of improved versions of the original observations, improving the quality of the reanalysis product (Inness et al., 2019).

3.4.1. Human exposure

To transform $PM_{2.5}$ concentrations into population weighted exposure (Exp), we use the 2020 UN WPP-Adjusted Population Count, v4.11, at 30 arc-second spatial resolution, from the NASA Socioeconomic Data and Applications Center (SEDAC) (Center For International Earth Science Information Network - CIESIN - Columbia University, 2017). We calculate gridded exposure by multiplying grid-level population times grid-level, monthly concentrations (Eq. 4), while we calculate country-level exposure by summing over grid cell i the population weights times grid-level, monthly concentrations (Eq. 5).

$$Exp_{i,m} = pop_i \times C_{i,m} \tag{4}$$

$$Exp_{c,m} = \sum_{i=1}^{n} \frac{pop_i}{pop_c} \times C_{i,m} \tag{5}$$

4. Methods

CLAQC applies two methods allowing to measure method uncertainty and also to trade-off transparency and complexity. Figure 3 shows the schematic representation of the CLAQC workflow.

CLAQC is fully open-access and provides two methodological approaches: Elastic net modelling and extreme gradient boosting regressor.

4.0.1. Coefficient constraints

Specific monotonic constraints are imposed to some of the model coefficients, based on expected physical relationships.

Regarding $PM_{2.5}$, we impose monotone positive constraints between emissions and concentrations. We further impose 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 less concentration. Although wind speed generally 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 O_3, similarly, we impose that emissions of NMVOC increase its concentrations, while leave emissions of NO_x unconstrained, allowing for non-monotone relationships with O_3. Temperature is also constrained to increase O_3 concentrations. O_3 is a photo-chemical secondary pollutant, thus it increases with intensifying solar radiation: temperature is therefore used as a proxy.

The variables included in the models are the following: sectoral emissions, emission sectoral totals (*i.e.*, $E_{k,s} = \sum_s E_{k,s,m}$), emission pollutant totals (i.e., $E_{k,p} = \sum_p E_{k,s,p}$), precipitation, minimum temperature, maximum temperature, vapor pressure deficit, wind speed, wind direction, monthly fixed effects, and their interactions.

4.1. Elastic net models

First, we use Elastic net models (Zou & Hastie, 2005), a method suitable to identify the subset of best predictors obtaining a parsimonious model. It regularizes the model coefficients improving accuracy and model interpretability by decreasing variance. This prevents our models to be volatile to extreme variations and outliers. This technique avoids large errors on the one hand and, on the other it results in more conservative estimations of the concentrations obtained from the emissions reductions.

4.1.1. Motivation

Due to the high multicollinearity amongst predictors, as shown in Appendix A, a regularized linear regression method is applied, namely Elastic net regularization. Elastic net is a technique that linearly combines penalties from the Least Absolute Selection and Shrinkage Operator (LASSO) and Ridge methods. This allows us to tackle the collinearity issue amongst predictors, while controlling for coefficient signs. The disadvantage is that the method may select fewer sector predictors than the ideal amount. However, if historically they have always been correlated there is no reason to believe that highly collinear sectors will detach in behaviour in the future. This way the models keep a non-excessive number of predictors and more stable and meaningful coefficients. Additionally, the above explained selected sector aggregation attenuates this effect to some extent.

4.1.2. 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 (SSR). 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 by 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 towards 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 which do not depend on the unit of measurement of the features. For further details, see (Zou & Hastie, 2005) and Hastie, Tibshirani, and Friedman (2009).

4.1.3. Workflow

The process is represented in Figure 4 and follows the steps below, for each country:

1. To ensure reproducibility, a seed is set for the whole coding session with the set.seed R function.[3]
2. A gridded monthly data set is aggregated at the country and month level. While sectoral emissions are summed up, weighted concentrations and meteorology variables are averaged. Wind direction is treated as circular variable through the circular function from the circular R package.
3. Outliers are then identified and excluded by applying the interquartile range rule and only non-NA observations are kept in the analysis.
4. The resulting data set is randomly split without replacement stratifying by month, using the stratified function from the splitstackshape R package. Hence, randomization occurs over the temporal dimension. Four-fifths of the data sample are used as training set, and the remaining fifth is used as test set.
5. A k-fold cross-validation algorithm for tuning the Lambda (λ) parameter is applied by using the cv.glmnet function from the glmnet R package. 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 and 'gaussian' family Elastic net (Friedman et al., 2020).
6. Non-negativity constraints are set for certain predictors. See details in Paragraph 4.0.1.
7. The model is trained on the training set by applying the glmnet function from the glmnet R package.
8. Its performance is evaluated on the test set, i.e. on

[3] addTaskCallback(function(...) set.seed(86); TRUE)

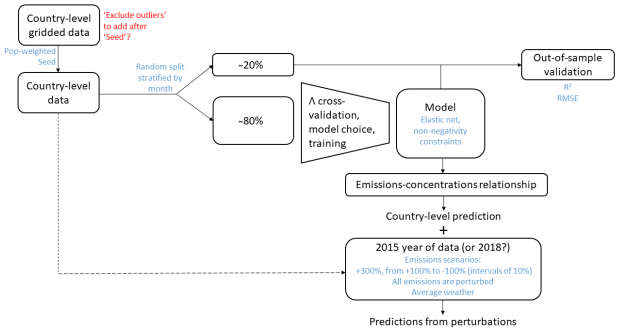


Figure 4: Visual representation of the Elastic net models workflow.

data not used to build the model itself. We report the out-of-sample R^2 and RMSE.

9. Predicted values are extracted through the predict.glmnet function from the glmnet R package.
10. Coefficients are extrapolated with the coef.glmnet function from the glmnet R package.
11. Emission-concentration relationships are derived from Elastic net shrinkage regression methods.

4.1.4. Models

The Elastic net linear regression models take the following form for each country (Eq. 6):

$$\begin{aligned} PM_{2.5m} = \alpha &+ \sum_{s,p} \beta_{s,p} E_{s,p,m} + \gamma_1 PPT_m + \gamma_2 TMIN_m + \\ &+ \gamma_3 TMAX_m + \gamma_4 VPD_m + \gamma_5 WS_m + \gamma_6 WD_m + \\ &+ \sum_s \delta_s E_{s,m} + \sum_p \lambda_p E_{p,m} + \mu E_{NO_x,m} \times E_{NH_3,m} + \\ &+ \nu E_{SO_2,m} \times E_{NH_3,m} + \xi E_{SO_2,m} \times E_{NO_x,m} + \\ &+ \sum_s \theta_s E_{s,m} \times WS_m \times WD_m + \phi_m + \epsilon_m, \\ O_{3m} = \alpha &+ \sum_{s,q} \beta_{s,q} E_{s,q,m} + \gamma_1 PPT_m + \gamma_2 TMIN_m + \\ &+ \gamma_3 TMAX_m + \gamma_4 VPD_m + \gamma_5 WS_m + \gamma_6 WD_m + \\ &+ \sum_s \delta_s E_{s,m} + \sum_q \lambda_q E_{q,m} + \mu E_{NOx,m} \times E_{NMVOC,m} + \\ &+ \nu E_{SO_2,m} \times E_{NMVOC,m} + \xi E_{SO_2,m} \times E_{NO_x,m} + \\ &+ \sum_s \theta_s E_{s,m} \times WS_m \times WD_m + \phi_m + \epsilon_m \end{aligned} \tag{6}$$

where:

s $\in \{AGR, INX, OTH, OTR, POW, ROA, SER\}$
p $\in \{BC, OC, NH_3, NO_x, NMVOC, SO_2\}$
q $\in \{NO_x, NMVOC, SO_2\}$
m = 1, ..., M
$PM_{2.5}$ = Concentration of $PM_{2.5}$ in μ/m^3 (population-weighted)
O_3 = Concentration of O_3 in μ/m^3
$E_{s,p,m}$ = Emissions of sector s and pollutant p in Tg
$E_{s,q,m}$ = Emissions of sector s and pollutant q in Tg
PPT_m = Accumulated precipitation in mm
$TMIN_m$ = Minimum 2-m temperature in $degC$
$TMAX_m$ = Maximum 2-m temperature in $degC$
VPD_m = Mean vapor pressure deficit in kPa
WS_m = 10-m wind speed in $\frac{m}{s}$
WD_m = Wind direction in $degrees$
E_p = Composite index from the sum of total emissions of pollutant p in Tg
E_s = Composite index from the sum of total emissions of sectors s in Tg
ϕ = Monthly fixed effects
ϵ = Error term

In Eq. 6, m indicates the month (January-December), s is the emission sector, p and q refer to the emitted pollutant in their respective models. $PM_{2.5}$ and O_3 concentration values, in $\mu g/m^3$, obtained from the models are country-level monthly concentration averages indexed by month m, just as all the other parameters in the equation; in particular, $PM_{2.5}$ levels are weighted by population; $E_{s,p}$ and $E_{s,q}$ are emissions of sector s and pollutant p or q, respectively, expressed in Tg; E_p and E_s are total emissions of pollutant p and of sector s, respectively; PPT_m stands for accumulated precipitation in mm; $TMIN_m$ and $TMAX_m$ are minimum 2-m temperature and maximum 2-m temperature, respectively, in $°C$; VPD_m is mean vapor pressure deficit in kPa; WS_m is 10-m wind speed in m/s; WD_m is average wind direction in $degrees$; ϕ are month fixed effects, and ϵ is the stochastic term.

4.1.5. Elastic net results post-processing

In order to have non-negative predicted values for y, a non-negativity constraint which selects the maximum value between zero and the Elastic net model prediction should be added, as specified below:

$$\hat{y}^* = \max\left\{0,\ \hat{y}_{Elastic\ Net}\right\} \tag{7}$$

Let $\hat{y}^*$ denote the final predicted value and $\hat{y}_{Elastic\ Net}$ the elastic net predicted value. According to Eq. 7, the final predicted value, $\hat{y}^*$, corresponds to $\hat{y}_{Elastic\ Net}$ if $\hat{y}_{Elastic\ Net} \geq 0$, while it equals 0 when $\hat{y}_{Elastic\ Net} < 0$.

Moreover, in order to have non-extreme predicted values for y due to input data divergence, a second safety function which selects the minimum value between the non-negative Elastic net model prediction from Eq. 7 and the double of the maximum observed concentration should also be added, as specified below:

$$\hat{y}_{Final} = \min\left\{\hat{y}^*,\ 2y_{Max}\right\} \tag{8}$$

Let $\hat{y}_{Final}$ denote the final predicted value, $\hat{y}^*$ the Elastic net predicted value after applying the formula in Eq. 7, and y_{Max} the maximum monthly concentration mean observed in a given country. According to Eq. 8, the final predicted value, $\hat{y}_{Final}$, corresponds to $\hat{y}^*$ if $\hat{y}^* \leq 2y_{Max}$, while it equals $2y_{Max}$ when $\hat{y}^* > 2y_{Max}$.

4.2. Machine learning models

Emission-concentration functions might not be sufficiently well approximated by a linear function due to the non-linearities of topography and secondary pollution formation (Thunis et al., 2019). Machine learning models

are powerful tools that can reproduce highly nonlinear relationships like the complex natural phenomena behind air pollution formation, transport and dispersion. Importantly, they do not require the user to impose a functional form.

4.2.1. Additional data pre-processing in ML

In addition to the pre-process description in Section 3, we perform specific additional data processing.

1. Given the very high level of collinearity between BC and OC emissions data, we sum the two precursors into a variable called Total Carbon (TC).
2. Emissions from the sector Other (OTH) are frequently missing, or otherwise highly correlated with other emissions. They are excluded from the analysis as their informative content is very low. Additionally, this sector has been seen to have high uncertainty ranges across emission databases (*e.g.*, ECLIPSE, EDGAR, CAMS, MACCity, RCP, (Lamarque et al., 2010) CEDS) .
3. Emissions are expressed in kg, and no longer in Tg.
4. In addition to year and month-of-the-year, we include an identifier of grid cells as predictor variable.

4.2.2. Workflow

We use extreme gradient boosting regressor (Chen & 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 Figure 5 and goes as follows. Separately for each country-pollutant pair,

1. A gridded data set is randomly split stratifying by grid cell. Hence, randomization occurs over the temporal dimension. Note that unlike int he Elastic Net models here we do not aggregate emissions by country but use each grid cell within a given country as an observation. Three-fourths of the data is used as training set, the remaining fourth is used as test set.
2. The model is trained on the training set.
3. Its performance is evaluated on the test set. We report the out-of-sample R^2 and RMSE.
4. We derive emission-concentration relationships from a "black-box" algorithm in a fashion similar to partial dependence plots (Friedman, 2001). Using the last 5 years of data, for emissions of precursor p from sector s,
   (a) Perturbing emissions $E_{s,p}$ by factor $P_{s,p}$, simulating a scenario in which only $E_{s,p}$ are altered by a policy.
   (b) Predict concentrations under the new scenario.
   (c) Average predictions from the grid-year level to country level.

   We repeat the process with perturbation of +300%, and from +100% to -100% at intervals of 10%.

Finally, we emulate the machine learning models with cubic functions so that they can also be used in optimization mode integrated assessment models or for easy implementation in a spreadsheet tools.

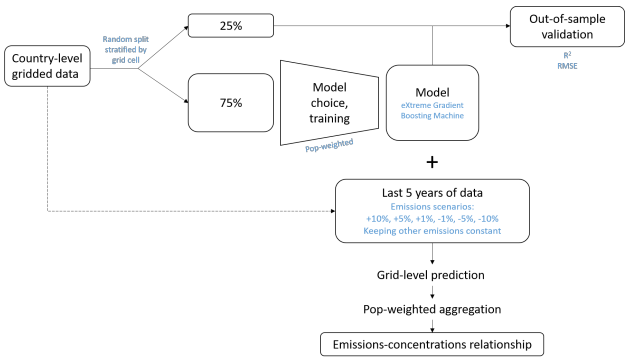


Figure 5: Visual representation of the extreme gradient boosting regressor models workflow.

4.3. Model implementation

For the countries which do not have monthly reported emissions, a monthly scheduling profile is provided for each country, in order to transform annual emissions into monthly emissions, and vice versa. We build such monthly schedule using the most recent years of emissions data (2014-2018) for each couple of pollutant and sector, representing a reference monthly emission value by sector and pollutant. The monthly weights can be multiplied by their equivalent pollutant and sector total annual precursor emissions, and then be used directly in Eqs. 6.

Additionally, we provide default meteorology fields that can be used in the models in case the input data is missing. The default meteorology variable fields are based either on the average of the last 5 years of meteorology in each country to represent current trends, or on the average of all years considered in the analysis (2003-2018) in each country to proxy long-term trends.

> We do not advise to use the model for countries with R^2 smaller than 0.5, or MAE higher than 10 and RMSE higher than 12. Such low metrics are present only in two countries.

In health and crop impact assessments of pollution due to O_3, other metrics are more common such as 6-month warm season mean of daily maximum 8-hour average (6mDMA8). We provide a post-process database that allows for the conversion from annual average O_3 concentrations to 6mDMA8 metric (see SI TODO for more details).

4.4. Method comparison

We summarize in Table 1 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 such high resolution inputs increasing the statistical power. The pixel-based approach allows for flexible spatial aggregation, although here we keep the country-level spatial resolution. On the other hand, the Gradient boosting regressor method requires emulation in

Elastic Net	Gradient boosting regressor
Simple equation	Non intelligible form (a cubic approximation is performed to overcome this)
Country-level emission totals allow for direct and fast application but trade-off flexibility	Flexible regional aggregation (the pixel-level detail allows for different regional aggregations)
Moderate sensitivity to emission changes	Less sensitive to strong emission changes
Assume that historically correlated sectors will remain correlated	

Table 1
Comparison summary between CLAQC framework methods.

order to be easily used in optimization mode or in simple models.

5. Discussion

5.1. Model results — Stylized scenarios

We run a sensitivity analysis on the model implementation to explore multiple hypothetical policy scenarios. For emissions of precursor p from sector s, we undertake the following Policy simulation:

1. We perturb emissions $E_{s,p}$ by factor $P_{s,p}$, simulating a scenario in which only $E_{s,p}$ are altered by a policy.
2. We predict concentrations under the new scenario by applying the emission-concentration relationships with perturbed emissions and a representative typical weather year. The typical weather year is calculated either by averaging the last 5 years of meteorology in each country to represent current trends, or all years considered in the analysis (2003-2018) in each country to proxy long-term trends.
3. For ML only, we average predictions from the grid-year level to the country level.
4. We repeat the process with perturbation of +300%, and from +100% to -100% at intervals of 10%.

The perturbation on emissions is applied to verify model accuracy on future data, *i.e.*, what concentrations the model would predict if emission values were not similar to the initial ones. In the case of Elastic net, after the perturbation, some predictions take on negative values.

We consider model predictions as baseline predictions, and obtain model predictions from imputing perturbed emissions into CLAQC models. We calculate annual baseline concentration predictions by averaging out monthly baseline predictions, and do the same for predictions from perturbed emissions. We then calculate the percentage difference between annual concentration predictions from perturbed sectoral emissions and baseline predictions. We also calculate the minimum, median, and maximum annual percentage variation in predictions from perturbed emissions by perturbation and sectoral level to account for the sectoral range variability.

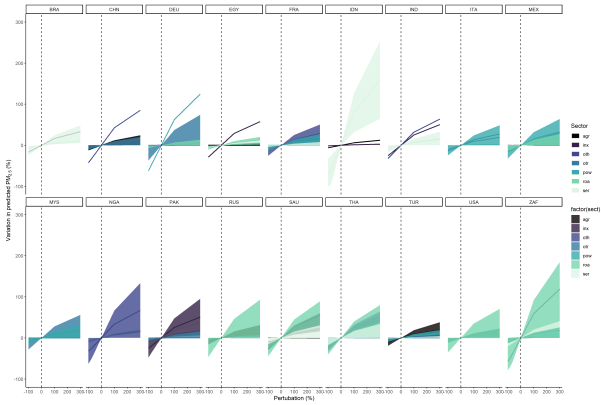


Figure 6: Percentage variation in predicted concentrations by sector and perturbation for selected countries.

Figure 6 shows the percentage variation in annual predicted concentrations and perturbations by sector, for selected countries.

5.2. Model Validation Results

A key aspect of predictive model evaluation is to verify if the models can reproduce either future or past trends. We present a two-fold validation: observed out-of-sample validation and validation against similar tools for both methods.

Figures 7 and 8 map the out-of-sample R^2 and RMSE for ML and EN models.

ML models for $PM_{2.5}$ for all countries except Australia have R^2 above 0.5, and are higher in China and India, where billions of individuals are exposed to very high concentrations of $PM_{2.5}$. Models for O_3 have systematically high R^2.

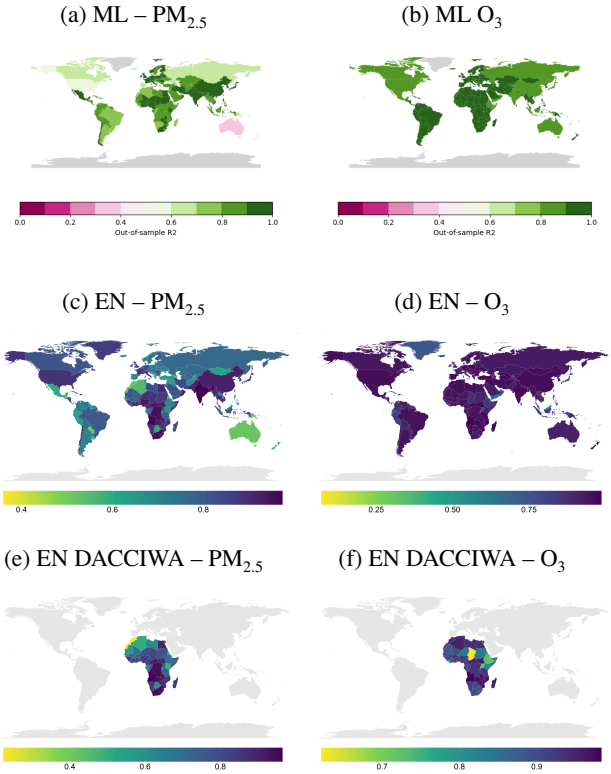


Figure 7: Out-of-sample performance metrics of ML and EN models (both from CAMS and DACCIWA data) as in Eq. 6 under Section 4.1.4: R^2.

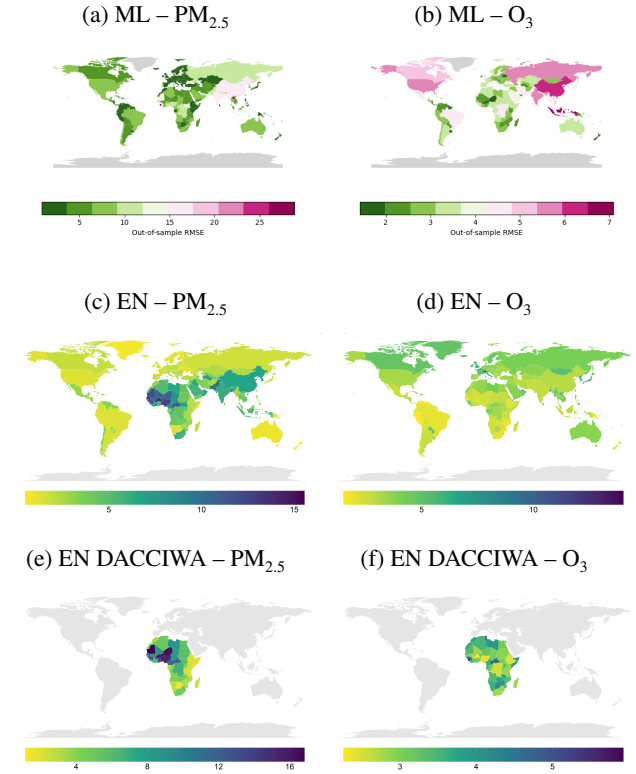


Figure 8: Out-of-sample performance metrics of ML and EN models (both from CAMS and DACCIWA data) as in Eq. 6 under Section 4.1.4: RMSE.

5.3. Additional remarks

We include an identifier of grid cell as 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 emissions scenarios that are not excessively different from the baseline. For instance, recurrent trans-boundary pollution can be modeled by the interaction of cell identifiers and month.

The improvement in geographic precision might come at the cost, on the other hand, of more bias in the case of extreme perturbations. For robustness, we also estimate the models without the identifier and supply the results in a separate Excel spreadsheet.

Models with grid cell identifiers perform better, as expected (Figure 9). Figure 10 compares the changes in concentrations predicted with and without 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.

5.4. Model selection

We construct several models, using two different methods (Elastic net and XGBoost), two different sector-detailed emission databases (CAMS and DACCIWA), and several variations of our model predictors. This allows us to span a possibility space and do sensitivity on normative assumptions. We propose a systematic model selection based on two criteria:

- The model error/reliability, measured by out-of-sample R^2 and RMSE.
- Expert informed emission database source reliability. We take the informed decision to score the maximum for all the countries where DACCIWA database is available due to the more consistent methods that were used to develop it, which preferred *in-situ* measurements as opposed to large data proxies and source profiles.

We also analyse variation of the output, measured by the standard deviation of the concentrations when emissions are perturbed. For Elastic net models, perturbed emissions are country aggregates as compared to within country pixel of the Machine Learning models. Due to this miss-match between methods we drop this criterion in the decision for model selection.

We re-scale all the elements of our decision criteria between 0 and 1, with 1 being the maximum score (Fig. 11). We obtain the ensemble score by weighting each of the criteria as in Eq. 7. We choose the best model by taking the maximum of the score (Eq. 8).

$$s_v = \frac{1}{2.5}C_{R2} + \frac{1}{5}C_{RMSE} + \frac{1}{2.5}C_{Source} \tag{9}$$

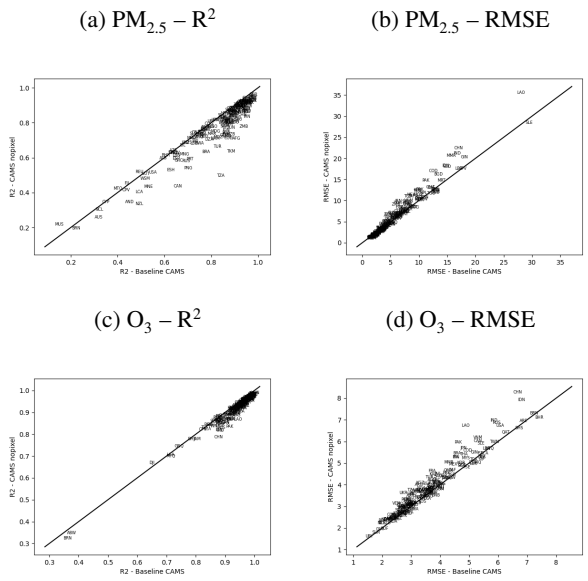


Figure 9: Comparison of performance of models with (x-axis) and without (y-axis) grid cell identifier. Black lines indicate equality.

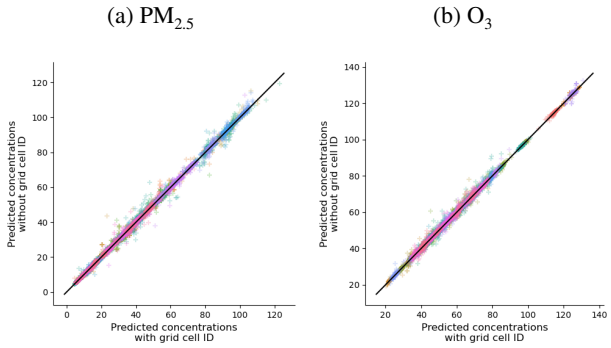


Figure 10: Concentrations under the simulation of perturbations for models with and without grid cell identifier. Each cross is a country-sector-precursor-perturbation combination. Black lines indicate equality, colors indicate countries.

Where $C_{R2}, C_{RMSE}, C_{SD}, C_{Source}$ are the scores of each criteria of model version v.

$$\max_{v} \left(s_v\right) \tag{10}$$

Figure 11 shows, for each of the criteria, which model is estimated to provide the best outcome. For $PM_{2.5}$, in terms of R^2, the XGBoost and the Elastic net models share the map, however the other error component, RMSE, performs better in the Elastic net models. Regarding O_3, the two methods share the map both for R^2 and RMSE. The Elastic net models, especially those without monthly fixed effects, provide higher variation in the output, performing better in the responsiveness to emissions reductions for both O_3 and $PM_{2.5}$. In terms of the emissions database source, the models with DACCIWA data are selected.

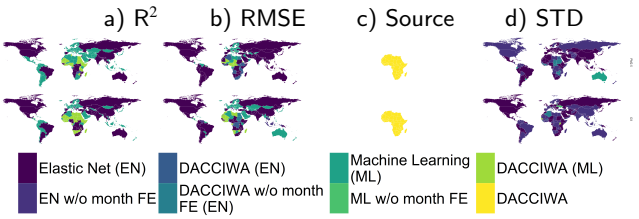


Figure 11: Best model score for each pollutant, each country and for all the decision criteria.

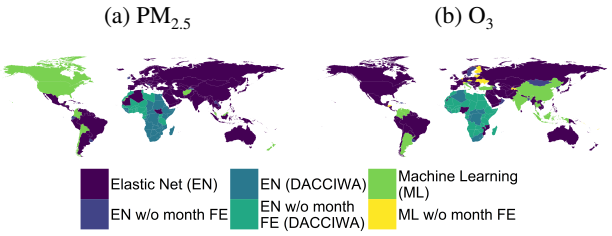


Figure 12: Best model score for each pollutant and each country.

Finally, the best model suggestion of each country is found through Eq. 8, and presented in Figure 12. Figure 12 summarizes our recommendations for model use for each country and pollutant. Once again, XGBoost and Elastic Net models share the map. For the region of Africa, the DACCIWA models both with XGBoost and Elastic Net are the most suitable ones according to our criteria.

All the maps and country-specific values can be explored in our web-tool `https://datashowb.shinyapps.io/CLAQC/`.

6. Conclusions

We develop an innovative air quality model framework for policy assessment, through empirical models. We use two methods that can be exploited both together allowing for the estimation of method uncertainty, and separately with flexible input data requirements. We provide global country-level coverage and sectoral detail and discuss on their validation results. Our aim is to deploy a fast tool at the global scale to derive concentrations from emissions in order to effectively intervene with emission abatement strategies on relevant sectors and air pollution exposure at the global level.

Though, important drawbacks need to be taken into account. In our framework, transboundary pollution is not modelled; the number of sectors is limited to the data availability and not all macro-sectors are included in the models (*e.g.*, ships); robustness heavily depends on the available temporal domain; data limitations do exist, such as the difference in magnitude of available emission data sets, and may be overcome through ensemble models; CAMS data are used both for emissions and concentrations, therefore there may be patterns in how the data is constructed; our models are not suitable for extremely disruptive meteorological and emission scenarios, since they are trained on business-as-usual trends; uncertainty ranges are missing in both models' estimates. Furthermore, the

models are data intensive and computationally demanding. EN drawbacks: Functions of emission variables or other controls not included, does not capture non-linearities.

Moreover, it makes use of various sources of data which might require some harmonization assumptions; results are highly dependent on data quality; such framework is very ambitious in terms of detail, some sector and/or countries might lack robustness.

CLAQC is a complementary model to the modelling community, proving empirically based estimates and added value for global scale sectoral and country level analysis. Moreover, its dynamic architecture allows for integration of new yearly data when relevant; it makes use of the ever more used gridded data (recent advancements on reanalysis and satellite data services).

Our models, based on empirical data, allow for sectoral differentiation at the country and eventually at the sub-national level (gridded estimations are possible, *i.e.* sub-country analysis), focusing on the pollutants with higher impact on human health. Several modelling methods are considered and tested. They can be useful tools for integrated assessment models and for supporting policy makers in understanding which sectors contribute the most to national air pollution and to what extent, to most effectively tackle air pollution. Furthermore, this work sheds new light on the centrality of reliable sectoral emission data.

Software availability

Software name: CLAQC v1.0;
Contact address: ;
Program language: R, Python;
Website: `https://datashowb.shinyapps.io/CLAQC/`

Declaration of interest

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Acknowledgments

This work has been funded by the World Bank. The authors would like to thank Dirk Heine, Simon Black, Martin Heger and Christian Schoder for their collaboration and comments. We acknowledge the RFF-CMCC European Institute on Economics and the Environment (EIEE) for providing the logistical platforms to perform this work, and the ECCAD-AERIS portal for the archiving and distribution of the emission data.

This work presents prediction models generated using modified Copernicus Atmosphere Monitoring Service information [2003-2018], downloaded from the Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS) (`https://cds.climate.copernicus.eu/cdsapp#!/dataset/cams-global-reanalysis-eac4-monthly?tab=overview`).Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

Figure 13: Geographical distribution of cells with non-null emissions. The color of a grid cell gives the percentage of observations with non-null emissions.

Review statement

Code and data availability statement

All data sets used in CLAQC applications are freely available online.

A. Appendix/Supplement

A.1. Data description

A.1.1. Sparsity, skewness

Anthropogenic emissions of pollutants or their precursors concentrate in urban and suburban areas, where most economic activities take place. The data employed covers approximately the entire landmass, a great share of which is source to minimal to no emissions. Data for sparsely populated areas is less relevant and of possibly inferior quality at the same time. In addition, the very skewed distribution of emissions and concentrations can be problematic in a regression framework. For these reasons, in machine learning models we limit the analysis of the input data to a subset of data for which emissions are non-null.

A.1.2. 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 emission factors. Emission data displays high correlation, even within grid cells, plausibly attributable to correlation of emissions of pollutants within sectors; and in economic activity across sectors. Figure **??** displays the cross-pollutant correlations within sectors.

A.2. Guide to the Excel spreadsheet

For every country and ambient pollutant ($PM_{2.5}$ and O_3), a machine learning model was built and an empirical emissions-concentrations relationship was constructed. The Excel spreadsheet `Results.xlsx` contains the data required to derive the changes and the levels of concentrations of pollutants under emissions 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 emis-

sions between the baseline scenario and a chosen scenario, expressed in 100 percentage points.

Omitting subscripts for country and pollutant for ease of notation, call $Emissions_{Baseline,s,p}$ the baseline emissions from sector s of precursor p and $Emissions_{A,s,p}$ the emissions under the alternative scenario A. Then, for every country, pollutant, sector, and precursor, the perturbation $P_{A,s,p}$ is

$$P_{A,s,p} = \left(\frac{Emissions_{A,s,p}}{Emissions_{Baseline,s,p}} - 1 \right) \cdot 100. \tag{11}$$

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

$$Concentrations_{A,s,p} = \begin{cases} a_{-100,s,p} & \text{if } P_{A,p,t} < -100 \\ a_{-100,s,p} + b_{-100,s,p} \cdot \left(P_{A,p,t} - -100\right) & \text{if } -100 \leq P_{A,p,t} < -80 \\ a_{-80,s,p} + b_{-80,s,p} \cdot \left(P_{A,p,t} - -80\right) & \text{if } -80 \leq P_{A,p,t} < -60 \\ a_{-40,s,p} + b_{-40,s,p} \cdot \left(P_{A,p,t} - -40\right) & \text{if } -40 \leq P_{A,p,t} < -20 \\ a_{-60,s,p} + b_{-60,s,p} \cdot \left(P_{A,p,t} - -60\right) & \text{if } -60 \leq P_{A,p,t} < -40 \\ a_{-20,s,p} + b_{-20,s,p} \cdot \left(P_{A,p,t} - -20\right) & \text{if } -20 \leq P_{A,p,t} < 0 \\ a_{0,s,p} + b_{0,s,p} \cdot \left(P_{A,p,t} - 0\right) & \text{if } 0 \leq P_{A,p,t} < 20 \\ a_{40,s,p} + b_{40,s,p} \cdot \left(P_{A,p,t} - 40\right) & \text{if } 40 \leq P_{A,p,t} < 60 \\ a_{20,s,p} + b_{20,s,p} \cdot \left(P_{A,p,t} - 20\right) & \text{if } 20 \leq P_{A,p,t} < 40 \\ a_{60,s,p} + b_{60,s,p} \cdot \left(P_{A,p,t} - 60\right) & \text{if } 60 \leq P_{A,p,t} < 80 \\ a_{80,s,p} + b_{80,s,p} \cdot \left(P_{A,p,t} - 80\right) & \text{if } 80 \leq P_{A,p,t} < 100 \\ a_{100,s,p} + b_{100,s,p} \cdot \left(P_{A,p,t} - 100\right) & \text{if } P_{A,p,t} > 100 \end{cases} \tag{12}$$

The coefficient $a_{j,s,p}$ is the level of concentrations when emissions of precursor p from sector s are perturbed by an amount within the interval starting at j, *e.g.* $j = -20$ for a perturbation of -10%. The coefficient $b_{j,s,p}$ is the slope of the piecewise function in the interval starting at j. The coefficients $a_{-100,s,p} \ldots a_{100,s,p}$ and $b_{-100,s,p} \ldots b_{100,s,p}$ are reported in the spreadsheet in columns K to AF. 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 baseline emissions.

> Emissions are in kg. Concentrations of $PM_{2.5}$ are expressed in $\mu g/m^3$, concentrations of ozone are in 6mDMA8 $\mu g/m^3$.

Baseline concentrations, in column G, are the average concentrations (over the entire country) from 2014 to 2018. Baseline emissions, in column H, are the average emissions of a precursor from a given sector over the same period.[4]

Scenario emissions, in column I, are *set by the user*. The perturbation, in column J, is automatically computed.

Concentrations under the alternative scenario are computed in column G following Eq. 12. It should be noted that the calculation *assumes that only emissions of the row sector-precursor pair are perturbed*. All other emissions are assumed constant.

The change in concentrations attributable to the perturbation $P_{A,s,p}$ is calculated in column AH 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:

$$\begin{aligned} \Delta\, Concentrations_{A,s,p} &= Concentrations_{Baseline} - Concentrations_{A,s,p} \\ &= a_{0,s,p} - \left[a_{j,s,p} + b_{j,s,p} \cdot (P_{A,s,p} - j)\right] \\ &= a_{0,s,p} - a_{j,s,p} - b_{j,s,p} \cdot (P_{A,s,p} - j) \end{aligned} \tag{13}$$

where $P_{A,s,p}$ is inside an interval starting at j.

When scenario emissions are set to zero, the change in concentrations gives the (opposite of the) estimated contribution of each sector-precursor to the total concentrations in 2014-2018.

> The approximation of the emissions-concentrations relationship functions is best for small and moderate perturbations, and larger under scenarios of extreme perturbations. To avoid that approximation error reverses the relationship between emissions and concentrations of $PM_{2.5}$, 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 emissions scenario A is computed in column AI summing across sectors and precursors:

$$\begin{aligned} \underset{Country}{\Delta} Concentrations_{A,s,p} &= \sum_{s,p} \Delta\, Concentrations_{A,s,p} \\ &= \sum_{s,p} a_{0,s,p} - a_{j,s,p} - b_{j,s,p} \cdot (P_{A,s,p} - j) \end{aligned} \tag{14}$$

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

$$\begin{aligned} Concentrations_{A} = Concentrations_{Baseline} + \\ \underset{Country}{\Delta} Concentrations_{A,s,p} \end{aligned} \tag{15}$$

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

Comparing 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:

$$\begin{aligned} &\Delta\, Concentrations_{A,s,p} - \Delta\, Concentrations_{B,s,p} = \\ &= a_{0,s,p} - a_{j_A,s,p} - b_{j_A,s,p} \cdot (P_{A,s,p} - j) - \left[a_{0,s,p} - a_{j_B,s,p} - b_{j_B,s,p} \cdot (P_{B,s,p} - j)\right] \\ &= a_{j_B,s,p} + b_{j_B,s,p} \cdot (P_{B,s,p} - j) - a_{j_A,s,p} - b_{j_A,s,p} \cdot (P_{A,s,p} - j) \end{aligned}$$

[4] Averages are weighted by population in models for $PM_{2.5}$, but not in models for O_3.

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

$$\begin{aligned}&\underset{Country}{\Delta} Concentrations_{A,s,p} - \underset{Country}{\Delta} Concentrations_{B,s,p} = \\ &= Concentrations_A - Concentrations_B = \\ &= \sum_{s,p} a_{j_B,s,p} + b_{j_B,s,p} \cdot (P_{B,s,p} - j) - a_{j_A,s,p} - b_{j_A,s,p} \cdot (P_{A,s,p} - j)\end{aligned}$$

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

$$\begin{aligned}&\underset{Country}{\Delta} Concentrations_{A,s,p} - \underset{Country}{\Delta} Concentrations_{B,s,p} = \\ &= Concentrations_A - Concentrations_B = \\ &= \sum_{s,p} a_{-20,s,p} + b_{-10,s,p} \cdot (-10+20) - a_{-100,s,p} - b_{-100,s,p} \cdot (-100+100)\end{aligned}$$

A.3. Model comparison

The CLAQC models previously presented are evaluated against different global data sources: namely, the *Evaluating the Climate and Air Quality Impacts of Short-Lived Pollutants* (ECLIPSE) scenarios [5] (Stohl et al., 2015) provided by the GAINS model (Amann et al., 2011; Kiesewetter et al., 2015), and the TM5-FASST (Dingenen et al., 2018).

A.3.1. Comparison with the GAINS model

To evaluate CLAQC models against GAINS the anthropogenic emission data are obtained from ECLIPSE CLE (Current legislation)[6] 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 $kt/year$:[7]

- **Sectors:** Agriculture (waste burning on fields), Industry (combustion and processing), Power plants, energy conversion, extraction,[8] Residential and commercial, Waste, Surface transportation, Ships.[9]

The GAINS $PM_{2.5}$ concentrations are downloaded from GAINS Online[10] tool and measured in $\mu g/m^3$.[11]

In order to acknowledge initial differences between data sets, 2015 CAMS $PM_{2.5}$ concentrations are compared with 2015 GAINS $PM_{2.5}$ concentrations.

We compare the models outcomes after having applied population-weights to the GAINS reported concentrations. 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, it is expected that also their outcomes will yield different results. While CAMS concentrations range between

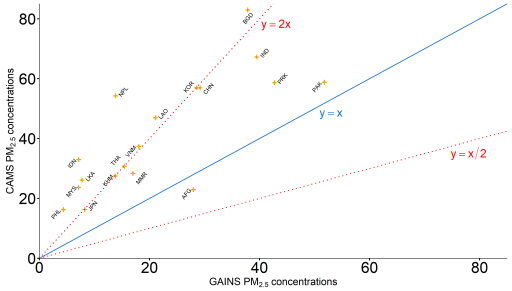


Figure 14: 2015 country-level annual concentrations of $PM_{2.5}$ in Asia from CAMS and GAINS datasets ($\mu g/m^3$). The red dotted lines represent the following factor differences between models: $y = 2x$ and $y = \frac{x}{2}$.

20 and 97.7 $\mu g/m^3$ (median of 44.9 $\mu g/m^3$), GAINS concentrations span between 8.2 and 60.7 $\mu g/m^3$ (median of 23.5 $\mu g/m^3$).

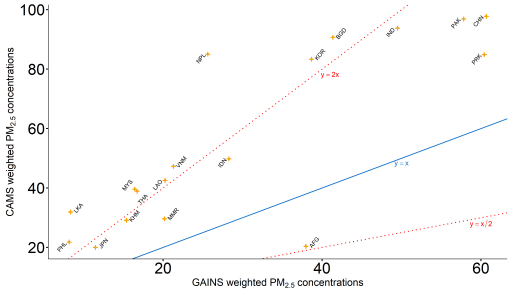


Figure 15: 2015 country-level annual population-weighted concentrations of $PM_{2.5}$ in Asia from CAMS and GAINS datasets ($\mu g/m^3$).

Regarding the comparison of the CAMS and GAINS emissions, we found that while annual pollutant totals in the two data sets reflect similar magnitudes, the 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 (Kurokawa & Ohara, 2020; Li et al., 2017). Thus, sectoral-specific air quality models face the infamous problematic of drastic input source uncertainty when it comes down to deliver sectoral detail. Importantly, even few orders of magnitude difference between the emissions input data and the emissions 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; 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.

[5] https://iiasa.ac.at/web/home/research/researchPrograms/air/ECLIPSEv5.html
[6] Baseline scenario.
[7] Converted to Tg to implement them into CLAQC model.
[8] Including gas flaring.
[9] Shipping sectoral emissions are not considered.
[10] https://gains.iiasa.ac.at/models/index.html (IIASA, 2009)
[11] Notice that for O_3 GAINS reports a different metric *i.e.* $SOMO35$ therefore this evaluation is done only for $PM_{2.5}$.

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

- Sectoral emissions are aggregated to make them match with CLAQC's variables.
- Surface transportation sector is re-scaled into Off-road and Road transportation, based on CAMS shares.
- ECLIPSE CLE V5 annual sectoral emissions are re-scaled to monthly sectoral emissions by applying ECLIPSE V5 temporal profiles, downloaded from IIASA and provided as monthly grid shares.
- The resulting gridded emissions are used with 2015 meteorology data and aggregated at the country-level, to be implemented into CLAQC model.

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

As shown in Figure 16, GAINS model underestimates population-weighted concentrations of $PM_{2.5}$ in several Asian countries for year 2020.[12] CLAQC predicted values range between 2.9 and 154.2 $\mu g/m^3$ (median of 67.7 $\mu g/m^3$), as opposed to GAINS concentrations ranging between 8.5 and 80.6 (median of 27.2 $\mu g/m^3$). These differences can be explained by the different sectoral aggregations in emissions and by spatial and temporal resolutions, but also by the 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 modelled atmospheric data (from an ensemble of models), GAINS is based only on ground level data and modelled relationships from EMEP and CHIMERE models (see Section 3.4). Because the CAMS data uses satellite imagery, it includes many natural sources that may not be easily observed by models, such as sea salt, desert dust and wildfires.

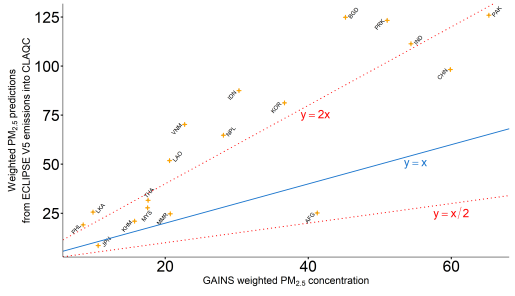


Figure 16: 2020 country-level annual population-weighted concentrations of $PM_{2.5}$ in Asia from CLAQC and re-scaled GAINS datasets ($\mu g/m^3$).

[12]The same pattern is repeated in years 2025 and 2030.

A.3.2. Comparison with the TM5-FASST model

TM5-FASST model[13] is a reduced-form air quality source-receptor model at global scale. CLAQC model comparison is applied only to TM5-FASST single-country regions, amongst the 56 regions available.

TM5-FASST concentrations are expressed as population-weighted $PM_{2.5}$ in $\mu g/m^3$, 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 used the Emissions[14]. As before, we start by comparing the CAMS concentrations range between 5.9 and 58.5$\mu g/m^3$ with a median value of 15.7$\mu g/m^3$, while TM5-FASST exposure ranges between 1.3 and 53.6 $\mu g/m^3$ with a median value of 7.64$\mu g/m^3$, showing some degrees of heterogeneity. Additionally, we also compare TM5-FASST and GAINS predicted concentrations for the year 2015. Importantly, the models differ in terms of their emission assumptions, however a model comparison is useful to understand the range of uncertainly of the input variables and the methods. Figure 17 shows that for most of the countries CAMS tends to deliver concentrations levels higher than TM5-FASST.

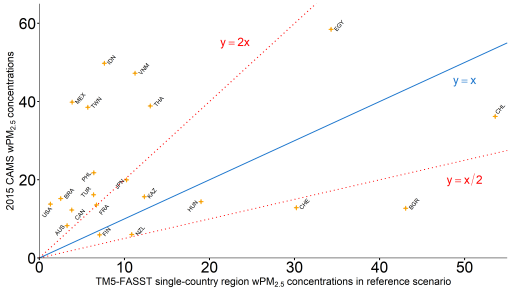


Figure 17: Country-level annual population-weighted concentrations of $PM_{2.5}$ from CAMS and TM5-FASST data ($\mu g/m^3$).

We implement the 2015 CAMS emissions into TM5-FASST scenario aggregating the CAMS monthly sectoral emissions by pollutant and year. As a result, in the case of $PM_2.5$ exposure, TM5-FASST with CAMS-CLAQC emissions predicts higher values compared to TM5-FASST concentrations from reference scenario (see Figure 18). In fact, the latter ones range between 1.3 and 53.6 $\mu g/m^3$, with a median of 9.8 $\mu g/m^3$, while CAMS-CLAQC emissions range between 8 and 92.6 $\mu g/m^3$, with a median of 29.3 $\mu g/m^3$. Differently, in the case of O_3 exposure, there is a higher level of convergence in predictions, as detailed in Figure 19. TM5-FASST predicted exposures from CAMS-CLAQC emissions range between 31.2 and 62.1 *ppb* (median of 49.2 *ppb*), and, very similarly, TM5-FASST values between 31.1 and 66.6 *ppb* (median of 48 *ppb*).

[13]Derived from `"spreadsheet FASST V1.2 NORMALIZED"`

[14]Consisting in IPCC Fifth Assessment's Representative Concentration Pathways (RCP) (Lamarque et al., 2010)

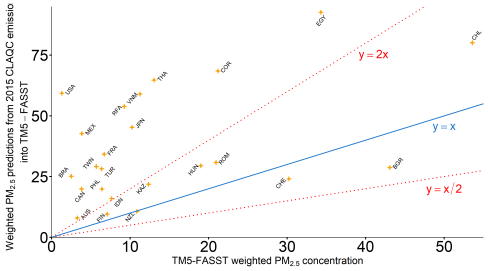


Figure 18: Country-level annual population-weighted concentrations of $PM_{2.5}$ from CLAQC and TM5-FASST data ($\mu g/m^3$).

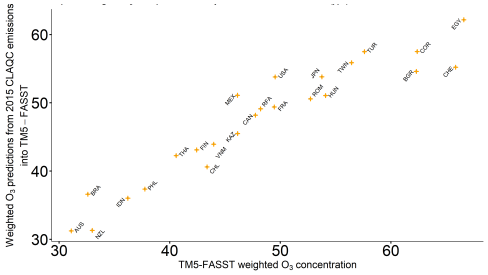


Figure 19: Country-level population-weighted 6mDMA8h concentrations of O_3 from CLAQC and TM5-FASST data (ppb).

A.4. Elastic net models with DACCIWA emission data

In Figure **??**, performance metrics of Elastic net models with DACCIWA emission data are proposed.

For O_3 models, R^2 ranges between 0.62 and 0.98 (median of 0.92); RMSE between 2.2 and 5.9 (median of 3.4); and MAE between 1.6 and 4.04 (median of 2.55); no countries with either R^2 below 0.4, or MAE greater than 10, or RMSE greater than 12. While, for $PM_{2.5}$, R^2 ranges between 0.26 and 0.96, (median of 0.79); RMSE between 1.2 and 17.1 (median of 5.1); and MAE between 0.94 and 12 (median of 4); 11 countries with either R^2 below 0.4, or MAE greater than 10, or RMSE greater than 12.

A.4.1. How to use the models

The linear regression models obtained using DACCIWA emission data take the following form for each country:

$$
\begin{aligned}
PM_{2.5m} &= \alpha + \sum_{s,p} \beta_{s,p} E_{s,p,m} + \gamma_1 PPT_m + \gamma_2 TMIN_m + \\
&+ \gamma_3 TMAX_m + \gamma_4 VPD_m + \gamma_5 WS_m + \gamma_6 WD_m + \\
&+ \sum_s \delta_s T_{s,m} + \sum_p \lambda_p E_{p,m} + \xi E_{SO_2,m} \times E_{NO_x,m} + \\
&+ \sum_s \theta_s E_{s,m} \times WS_m \times WD_m + \phi_m + \varepsilon_m, \\
O_{3m} &= \alpha + \sum_{s,q} \beta_{s,q} E_{s,q,m} + \gamma_1 PPT_m + \gamma_2 TMIN_m + \\
&+ \gamma_3 TMAX_m + \gamma_4 VPD_m + \gamma_5 WS_m + \gamma_6 WD_m + \\
&+ \sum_s \delta_s E_{s,m} + \sum_q \lambda_q E_{q,m} + \mu E_{NO_x,m} \times E_{NMVOC,m} + \\
&+ \nu E_{SO_2,m} \times E_{NMVOC,m} + \xi E_{SO_2,m} \times E_{NO_x,m} + \\
&+ \sum_s \theta_s E_{s,m} \times WS_m \times WD_m + \phi_m + \varepsilon_m
\end{aligned}
\tag{16}
$$

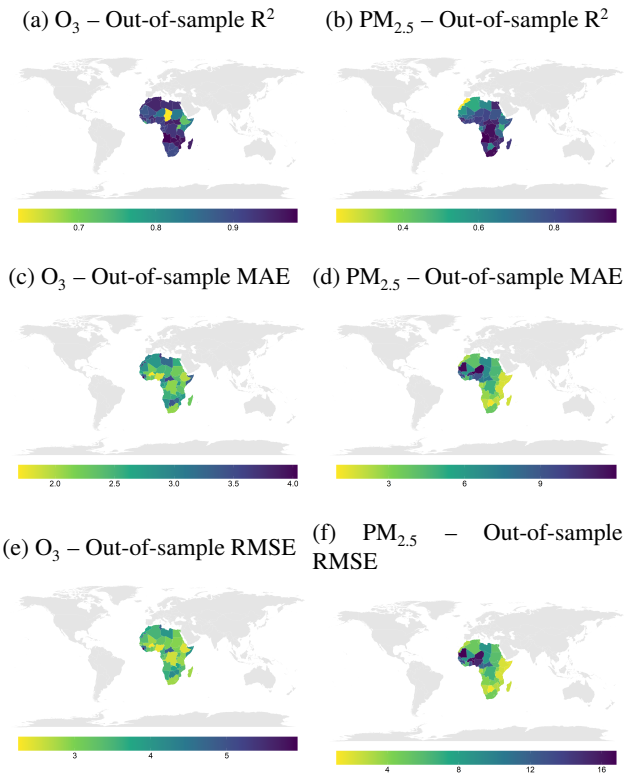


Figure 20: Out-of-sample performance metrics of models with DACCIWA emissions in Eq. 16 under Section A.4.1: R^2, MAE, and RMSE.

where:

s	$\in \{Transportation,\ Energy,\ Industries,\ Residential,\ Other\}$
p	$\in \{BC,\ OC,\ NO_x,\ NMVOC,\ SO_2\}$
q	$\in \{NO_x,\ NMVOC,\ SO_2\}$
m	= 1, ..., M
$PM_{2.5}$	= Concentration of $PM_{2.5}$ in μ/m^3 (population-weighted)
O_3	= Concentration of O_3 in μ/m^3
$E_{s,p,m}$	= Emissions of sector s and pollutant p in kg
$E_{s,q,m}$	= Emissions of sector s and pollutant q in kg
PPT_m	= Accumulated precipitation in mm
$TMIN_m$	= Minimum 2-m temperature in $degC$
$TMAX_m$	= Maximum 2-m temperature in $degC$
VPD_m	= Mean vapor pressure deficit in kPa
WS_m	= 10-m wind speed in $\frac{m}{s}$
WD_m	= Wind direction in $degrees$
$E_{p,m}$	= Composite index from the sum of total emissions of pollutant p in Tg
$E_{s,m}$	= Composite index from the sum of total emissions of sectors s in kg
ϕ	= Monthly fixed effects
ε_m	= Error term

Unlike the Eqs. in 6, all NH_3 predictors[15], and Agriculture sectoral emissions are not present, while Transportation sector's predictors are not split into Road and Off-road transportation, as specified in Section **??**.[16]

$PM_{2.5}$ and O_3 concentration values, in $\mu g/m^3$, obtained from the models are country-level monthly concentration averages indexed by month m, as all the other parameters in

[15] Including interactions containing NH_3 totals.

[16] The Other sector also contains the Waste sector.

the equation; emissions are in kg; weather variables' units of measurement are expressed as specified in Section 4.1.4.

CRediT authorship contribution statement

Stefania Renna: Conceptualization of this study, Investigation, Project administration, Supervision, Data curation, Methodology, Software, Validation, Writing — Original draft preparation, Writing — Review & Editing. **Francesco Granella:** Investigation, Data curation, Methodology, Software, Validation, Writing — Review & Editing. **Lara Aleluia Reis:** Investigation, Data curation, Methodology, Software, Validation, Writing — Review & Editing. **Paulina Antipa:** Investigation, Data curation, Software, Validation, Writing — Review & Editing.

References

Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., & Hegewisch, K. C. (2018, January). TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. *Scientific Data*, *5*(1), 170191. Retrieved 2021-04-12, from https://www.nature.com/articles/sdata2017191 doi: 10.1038/sdata.2017.191

Amann, M., Bertok, I., Borken-Kleefeld, J., Cofala, J., Heyes, C., Höglund-Isaksson, L., … Winiwarter, W. (2011, dec). Cost-effective control of air quality and greenhouse gases in europe: Modeling and policy applications. *Environmental Modelling & Software*, *26*(12), 1489–1501. doi: 10.1016/j.envsoft.2011.07.012

Center For International Earth Science Information Network - CIESIN - Columbia University. (2017). *Gridded population of the world, version 4 (gpwv4): Population density, revision 11.* Palisades, NY: Socioeconomic Data and Applications Center (SEDAC). doi: 10.7927/H49C6VHW

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Clappier, A., Pisoni, E., & Thunis, P. (2015, dec). A new approach to design source–receptor relationships for air quality modelling. *Environmental Modelling & Software*, *74*, 66–74. doi: 10.1016/j.envsoft.2015.09.007

Clappier, A., Thunis, P., Beekmann, M., Putaud, J., & de Meij, A. (2021, nov). Impact of SOx, NOx and NH3 emission reductions on PM2.5 concentrations across europe: Hints for future measure development. *Environment International*, *156*, 106699. doi: 10.1016/j.envint.2021.106699

Copernicus Climate Change Service. (2019). *Era5 monthly averaged data on single levels from 1979 to present.* ECMWF. doi: 10.24381/CDS.F17050D7

Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Dentener, F., van Aardenne, J. A., … Janssens-Maenhout, G. (2018, oct). Gridded emissions of air pollutants for the period 1970–2012 within EDGAR v4.3.2. *Earth System Science Data*, *10*(4), 1987–2013. doi: 10.5194/essd-10-1987-2018

Dingenen, R. V., Dentener, F., Crippa, M., Leitao, J., Marmer, E., Rao, S., … Valentini, L. (2018, nov). TM5-FASST: a global atmospheric source–receptor model for rapid impact analysis of emission changes on air quality and short-lived climate pollutants. *Atmospheric Chemistry and Physics*, *18*(21), 16173–16211. doi: 10.5194/acp-18-16173-2018

Friedman, J. H. (2001, oct). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). doi: 10.1214/aos/1013203451

Friedman, J. H., Hastie, T. J., Tibshirani, R., Narasimhan, B., Tay, K., & Simon, N. (2020). Lasso and elastic-net regularized generalized linear models [r package glmnet version 4.0-2]..

Granier, C., Darras, S., Denier van der Gon, H., Doubalova, J., Elguindi, N., Galle, B., … Sindelarova, K. (2019). The copernicus atmosphere monitoring service global and regional emissions (april 2019 version). doi: 10.24380/D0BN-KX16

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York. doi: 10.1007/978-0-387-84858-7

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., … Zhang, Q. (2018, jan). Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the community emissions data system (CEDS). *Geoscientific Model Development*, *11*(1), 369–408. doi: 10.5194/gmd-11-369-2018

Huang, G., Brook, R., Crippa, M., Janssens-Maenhout, G., Schieberle, C., Dore, C., … Friedrich, R. (2017, jun). Speciation of anthropogenic emissions of non-methane volatile organic compounds: a global gridded data set for 1970–2012. *Atmospheric Chemistry and Physics*, *17*(12), 7683–7701. doi: 10.5194/acp-17-7683-2017

IIASA. (2009). Gains online: Tutorial for advanced users [Computer software manual]. Retrieved from http://webarchive.iiasa.ac.at/rains/reports/GAINS-tutorial.pdf

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., … Suttie, M. (2019, March). The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*, *19*(6), 3515–3556. Retrieved 2021-04-12, from https://acp.copernicus.org/articles/19/3515/2019/ doi: 10.5194/acp-19-3515-2019

Kiesewetter, G., Borken-Kleefeld, J., Schöpp, W., Heyes, C., Thunis, P., Bessagnet, B., … Amann, M. (2015, feb). Modelling street level PM&ltsub&gt10&lt/sub&gt concentrations across europe: source apportionment and possible futures. *Atmospheric Chemistry and Physics*, *15*(3), 1539–1553. doi: 10.5194/acp-15-1539-2015

Kurokawa, J., & Ohara, T. (2020, nov). Long-term historical trends in air pollutant emissions in asia: Regional emission inventory in ASia (REAS) version 3. *Atmospheric Chemistry and Physics*, *20*(21), 12761–12793. doi: 10.5194/acp-20-12761-2020

Lamarque, J.-F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z., … van Vuuren, D. P. (2010, aug). Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application. *Atmospheric Chemistry and Physics*, *10*(15), 7017–7039. doi: 10.5194/acp-10-7017-2010

Li, M., Liu, H., Geng, G., Hong, C., Liu, F., Song, Y., … He, K. (2017, nov). Anthropogenic emission inventories in china: a review. *National Science Review*, *4*(6), 834–866. doi: 10.1093/nsr/nwx150

Ma, J., Ding, Y., Cheng, J. C., Jiang, F., Tan, Y., Gan, V. J., & Wan, Z. (2020). Identification of high impact factors of air quality on a national scale using big data and machine learning techniques. *Journal of Cleaner Production*, *244*, 118955. Retrieved from https://www.sciencedirect.com/science/article/pii/S0959652619338259 doi: https://doi.org/10.1016/j.jclepro.2019.118955

Menut, L., Bessagnet, B., Briant, R., Cholakian, A., Couvidat, F., Mailler, S., … Valari, M. (2021, nov). The CHIMERE v2020r1 online chemistry-transport model. *Geoscientific Model Development*, *14*(11), 6781–6811. doi: 10.5194/gmd-14-6781-2021

Murray, C. J. L., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., … Lim, S. S. (2020, oct). 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*(10258), 1223–1249. doi: 10.1016/s0140-6736(20)30752-2

Reis, L. A., Drouet, L., & Tavoni, M. (2022, jan). Internalising health-economic impacts of air pollution into climate policy: a global modelling study. *The Lancet Planetary Health*, *6*(1), e40–e48. doi: 10.1016/s2542-5196(21)00259-x

Stohl, A., Aamaas, B., Amann, M., Baker, L. H., Bellouin, N., Berntsen, T. K., … Zhu, T. (2015, sep). Evaluating the climate and air quality impacts of short-lived pollutants. *Atmospheric Chemistry and Physics*, *15*(18), 10529–10566. doi: 10.5194/acp-15-10529-2015

Thunis, P., Clappier, A., Tarrason, L., Cuvelier, C., Monteiro, A., Pisoni, E., … Peduzzi, E. (2019, sep). Source apportionment to support air quality planning: Strengths and weaknesses of existing approaches. *Environment International*, *130*, 104825. doi: 10.1016/j.envint.2019.05.019

Thunis, P., Degraeuwe, B., Pisoni, E., Ferrari, F., & Clappier, A. (2016, dec). On the design and assessment of regional air quality plans: The SHERPA approach. *Journal of Environmental Management*, *183*, 952–958. doi: 10.1016/j.jenvman.2016.09.049

Zou, H., & Hastie, T. (2005, apr). Regularization and variable selection via the elastic net. , *67*(2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x