scientific data

OPEN

DATA DESCRIPTOR

APExpose_DE, an air quality exposure dataset for Germany 2010–2019

Alexandre Caseiro[1,2] ✉ & Erika von Schneidemesser[1,2]

Exposure to poor air quality is considered a major influence on the occurrence of cardiovascular and respiratory diseases. Air pollution has also been linked to the severity of the effects of epidemics such as COVID-19 caused by the SARS-CoV-2 virus. Epidemiological studies require datasets of the long-term exposure to air pollution. We present the APExpose_DE dataset, a long-term (2010–2019) dataset providing ambient air pollution metrics at yearly time resolution for NO_2, NO, O_3, PM_{10} and $PM_{2.5}$ at the NUTS-3 spatial resolution level for Germany (corresponding to the *Landkreis* or *Kreisfreie Stadt* in Germany, 402 in total).

Background & Summary

Air pollution is the largest environmental risk factor for premature mortality. Exposure to air pollution has been clearly linked to the occurrence and severity of cardiovascular and respiratory diseases. The average European loses 2.2 years of life expectancy due to air pollution[1,2] A number of recent studies have shown that the impact of air pollution on the respiratory system has an adverse influence on the effects of the COVID-19 disease, with particulate air pollution contributing globally to 15 percent of COVID-19 mortality[3–8].

In order to study the effects of air pollution on human health, exposure datasets are needed. These datasets need to provide comprehensive coverage of air pollutant concentrations over a geographical area and time period. As is often the case, studies that investigate the relationship between air pollution and e.g. health outcomes or social inequalities, produce such a dataset in the context of the study[9,10] These datasets are rarely published as stand-alone papers and often rely on model data. This makes re-use of such data more difficult. There are exceptions to this, such as the air pollution datasets published by Aaron van Donkelaar and co-workers which incorporate satellite data, modelling, and observational data[11–13], the latter of which was then used in the Harvard study on the role of air pollution on COVID-19 mortality in the United States[4]. Furthermore, an appropriately high spatial resolution is often critical for such studies.

The dataset presented here was created in the context of a study investigating the role of long-term air pollution in the severity of COVID-19 outcomes for Germany. More specifically, the relationship to COVID-19 mortality, but also additional morbidity endpoints, such as hospitalization and intensive care unit therapy and/or the necessity for mechanical ventilation, were also investigated. In this context we needed a long-term air pollution dataset at the county level for Germany, which we did not find available elsewhere. The air pollutants generally treated in epidemiological studies are particulate matter ($PM_{2.5}$ and PM_{10}, particulate matter with an aerodynamic diameter smaller than 2.5 μm and 10 μm, respectively), ozone (O_3), and nitrogen dioxide (NO_2). Therefore, in the above mentioned context, we created an air pollution dataset that covers 10 years (2010–2019) at the level of county (for Germany *Landkreis* and *Kreisfreie Stadt*), for $PM_{2.5}$, PM_{10}, O_3, NO_2, and NO. County level corresponds to the third level of the Nomenclature of Territorial Unit for Statistics (NUTS-3) spatial resolution.

While air pollution monitoring is required in the European Union, as prescribed in the Air Quality Directive[14], the criteria used to locate monitoring stations (pollution levels, population and availability of funds) often result in heterogeneity in spatial coverage and representativeness[15–17]. Long-term air quality monitoring data is not evenly distributed in space, and furthermore has differing amounts of coverage depending on the air pollutant. At most, only close to one half of the 402 counties had a monitoring station for a given pollutant (Table 1). To provide comprehensive coverage, we combined observational data with model global reanalysis data from the Copernicus Atmospheric Monitoring Service (CAMS), and evaluated a variety of options considering the different types of air quality monitoring data classifications (e.g., urban, rural).

[1]Institute for Advanced Sustainability Studies, Potsdam, Germany. [2]These authors contributed equally: Alexandre Caseiro, Erika von Schneidemesser. ✉e-mail: alexandre.caseiro@iass-potsdam.de

Pollutant	minimum coverage	maximum coverage
NO_2	94	197
NO	121	194
PM_{10}	130	188
$PM_{2.5}$	60	110
O_3	121	197

Table 1. Yearly NUTS-3 coverage (of a total of 402) by the monitoring network.

Such a dataset has a high potential for re-use in different types of health impact assessments, investigation of social inequalities, among other studies. There is also an intention to expand this dataset to provide further coverage for Europe, as well as to refine the use of the CAMS data (e.g. by testing the regional reanalysis besides the global reanalysis, among other possible improvements).

Methods

In this study, we consider 402 NUTS-3 units for Germany. This reflects the status as of up to November 1, 2016, when the two *Landkreise* Göttingen and Osterode am Harz merged. The sources used for the production of the dataset were Airbase, from the European Environmental Agency[18] and the CAMS global reanalysis EAC4[19].

Airbase. All the data (hourly concentrations of O_3, NO, NO_2, $PM_{2.5}$, PM_{10}) from air quality monitoring stations located in Germany were accessed between July 11, 2020 and July 20, 2020. Owing to availability, data for the years 2010–2012 were obtained from Airbase-v8, whereas for the years 2012–2018 the E1a dataset was used and for 2019 the E2a dataset was used. Any of the Airbase data used that is not yet ratified, has been flagged, to facilitate avoidance of unratified data if necessary. Airbase classifies the stations based on their type and their siting (called Area in the metadata). Because the focus of the present work is on the long term exposure, the stations of the types "Traffic" and "Industrial" were left out for being considered unrepresentative and those of the type "Background" were included. The background stations thus considered are classified as: rural, rural-nearcity, rural-regional, rural-remote, suburban and urban for Airbase E1a and E2a, and rural, suburban and urban for Airbase-v8.

The following metrics were calculated for each year and for each station where measurements of the pollutant were available and covered at least 80% of hours of the year:

- **NO_2** annual mean concentration
- number of hours of the year which have a **NO_2** concentration over 200 $\mu g/m^3$
- **NO** annual mean concentration
- **PM_{10}** annual mean concentration
- number of days of the year which have a daily average **PM_{10}** concentration over 50 $\mu g/m^3$
- **$PM_{2.5}$** annual mean concentration
- **O_3** annual mean concentration
- number of days of the year which have a daily average **O_3** concentration over 120 $\mu g/m^3$
- annual mean of the daily **O_3** maximum concentration
- maximum daily 1-h average **O_3** concentration over the entire year
- maximum daily 8-h average **O_3** concentration over the entire year.

Each station was geo-located within, and each computed yearly value associated to, a NUTS-3 unit. Within each NUTS-3 unit and for each metric, the yearly values per station were averaged in three ways, giving preference, though not exclusiveness, to certain types of stations. Each averaging strategy represents a different scenario:

average averaging the yearly values from all the stations within the NUTS-3 unit;

urban averaging of the yearly values from stations located at the most urban location types. The location types are, in order of preference: urban, suburban, rural-near city, rural-regional and rural;

remote averaging of the yearly values from stations located at the most remote location types. The location types are, in order of preference: rural, rural-regional, rural-near city, suburban and urban.

The methodology described in the previous section produces data for the NUTS-3 units and the years where monitoring data for a given pollutant is available (e.g. Figure 1 for NO_2 in 2018). Table 1 shows the number of NUTS-3 units covered by data from Airbase. The coverage is always below half of the total number of NUTS-3 units for Germany (maximum of 197 out of 402).

In order to fill any NUTS-3 units missing observational data, we considered three options: (1) use the regional (relative to the *Bundesland*, or NUTS-2) average, (2) use the nearest neighbour and (3) use the CAMS EAC4 global reanalysis[20], (see next section) after scaling.

Gap filling based on regional averages produced spatial uniformity over large areas (the *Bundesland* or NUTS-2 units), regardless of the nature of the NUTS-3. Using the value from the nearest neighbour overcomes the main drawback of the *Bundesland* approach, but produces artifacts in the form of pairs of NUTS-3 units with different typology (e.g. urban and rural) but equal exposure. Such happens mainly close to large cities: e.g. a rural NUTS-3 unit without any monitoring station adjacent to a large city, that has monitoring stations, ends up, under this strategy, with the same exposure value as the large city.

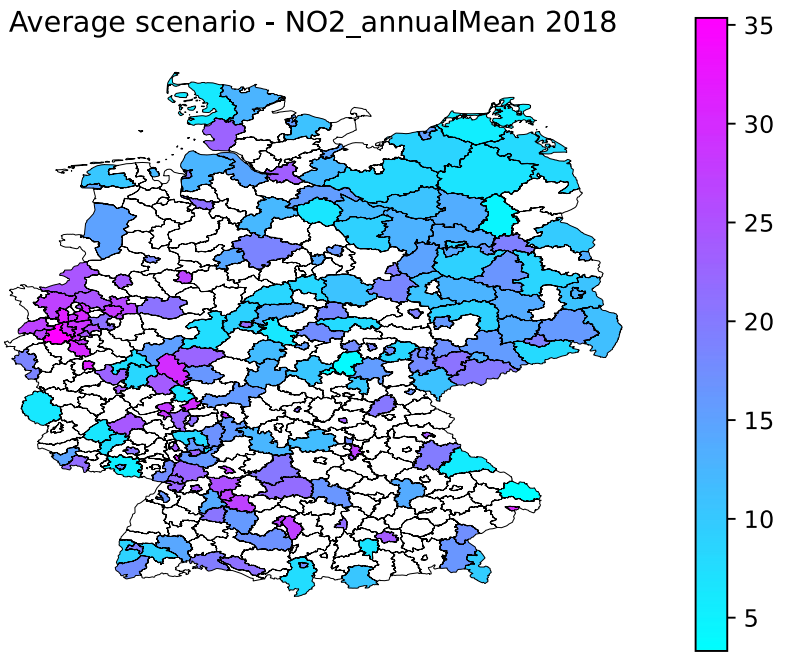

Fig. 1 NO_2 estimated yearly concentration ($\mu g\ m^{-3}$) for the average scenario at the NUTS-3 level for 2018.

CAMS. Due to those limitations we opted to explore the use of CAMS global reanalysis data to do the gap filling.

The CAMS reanalysis was checked for specific regional bias over Germany. The main biases impacting the present dataset for the latitudinal belt 40–50° N in Europe of the product are a ±15% bias for tropospheric ozone, with a seasonal cycle, and an underestimation of wintertime NO_2 columnar concentrations over part of Europe[21]. We estimate these biases to be either acceptable or compensated, at least partly, by the scaling procedure used (see below). The fields over Germany and for the study period were accessed on September 14, 2020 from the CAMS Atmosphere Data Store.

The procedure followed to compute the CAMS-based metrics is outlined in Fig. 2. Each metric with a time resolution equal to or longer than one day was computed for each cell, after averaging the 3-hourly output time-step to daily values. Those metrics were: NO_2 annual mean concentration, NO annual mean concentration, PM_{10} annual mean concentration, number of days of the year which have a daily average PM_{10} concentration over $50\,\mu g\ m^{-3}$, $PM_{2.5}$ annual mean concentration, O_3 annual mean concentration, number of days of the year which have a daily average O_3 concentration over $120\,\mu g\ m^{-3}$.

The resulting rasters (e.g. Figure 3 for the NO2 2018 yearly average concentration) were clipped (area weighted mean) to each NUTS-3 unit area after upscaling the spatial resolution from 0.75° to 0.01° (each smaller cell having the same value as its parent, larger, cell), resulting in verctorized metrics (one value for each NUTS-3 unit, each metric and each year). For each metric and under each scenario a scaling function between the vectorized CAMS values and the respective Airbase-based values was derived. Figure 4 shows the scaling function used to produce, for the NUTS-3 units where monitoring data was not available with satisfactory temporal coverage, CAMS-derived $PM_{2.5}$ from the clipped rasters and the monitoring-based NUTS-3 annual average.

The scaling function was then used to produce CAMS-derived values for the NUTS-3 units and the years where no monitoring data is available or is available with insufficient temporal coverage (Fig. 2).

Despite the considerable amount of scatter between the CAMS-derived data and the monitoring-derived data (e.g. Fig. 4 for $PM_{2.5}$), the approach using CAMS for gap filling overcomes the limitations that arose when using the nearest-neighbour or the *Bundesland* approaches.

For each scenario, a linear relationship between the metrics with a time resolution shorter than one day and a metric of the same pollutant with a time resolution equal or larger than one day was derived from the Airbase-based values and used with the CAMS-derived values to produce CAMS-based data for those metrics. The number of hours of the year which have a NO_2 concentration over $200\,\mu g\ m^{-3}$ was therefore computed from the NO_2 annual mean concentration. The annual mean of the daily O_3 maximum concentration, the maximum daily 1-h average O_3 concentration over the entire year and the maximum daily 8-h average O_3 concentration over the entire year were computed from the O_3 annual mean concentration. An example of such a scaling function is shown in Fig. 5.

Data Records

As a final step, the Airbase and CAMS derived data are combined to produce the APExpose_DE dataset. As an example, Fig. 6 shows the decadal average of the NO_2 yearly averages.

The APExpose_DE dataset is available in the form of an ASCII file: *APExpose_DE__2010–2019.csv*[22]. Each record (each line in the file) corresponds to a NUTS-3 unit (identified by its name and its code), and a scenario,

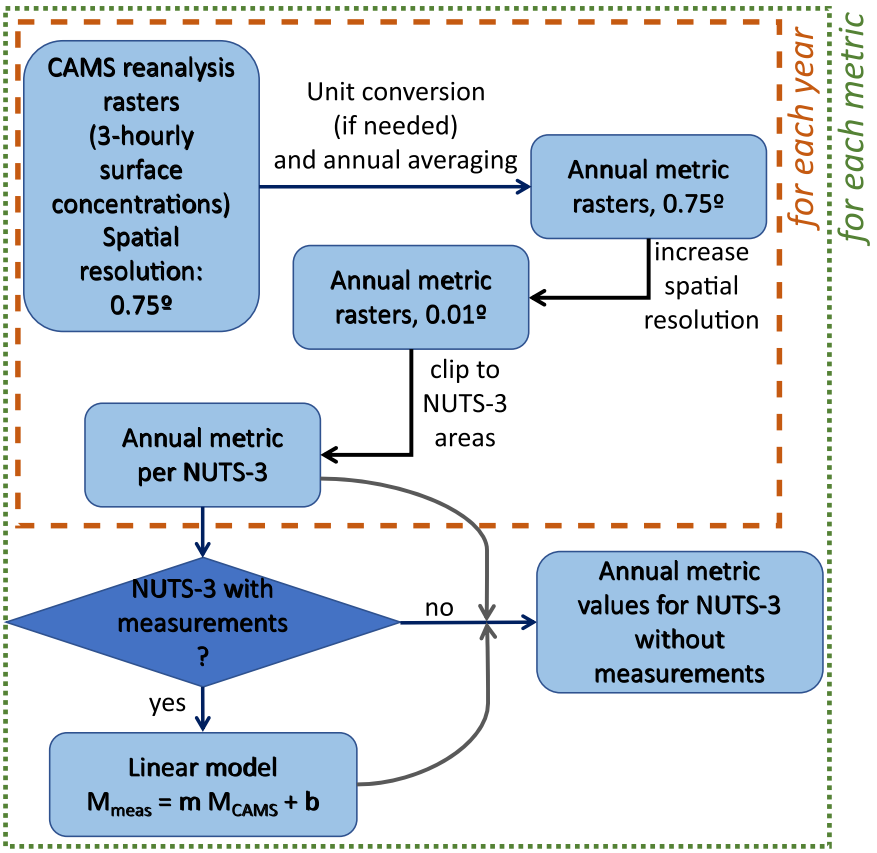

Fig. 2 Flowchart for the CAMS data: from the CAMS rasters to the annual metrics for the NUTS-3 units which do not have measurements (or do not have measurements with sufficient spatial coverage) on a given year. The procedure is used for the following metrics: NO_2 annual mean concentration, NO annual mean concentration, PM_{10} annual mean concentration, number of days of the year which have a daily average PM_{10} concentration over 50 $\mu g\ m^{-3}$, $PM_{2.5}$ annual mean concentration, O_3 annual mean concentration, number of days of the year which have a daily average O_3 concentration over 120 $\mu g\ m^{-3}$.

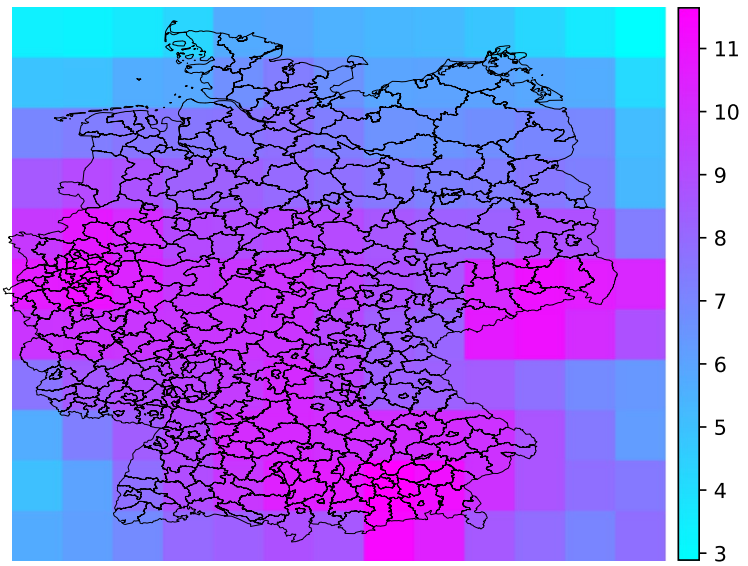

Fig. 3 NO_2 yearly (2018) average concentration ($\mu g\ m^{-3}$) from the CAMS EAC4 global reanalysis.

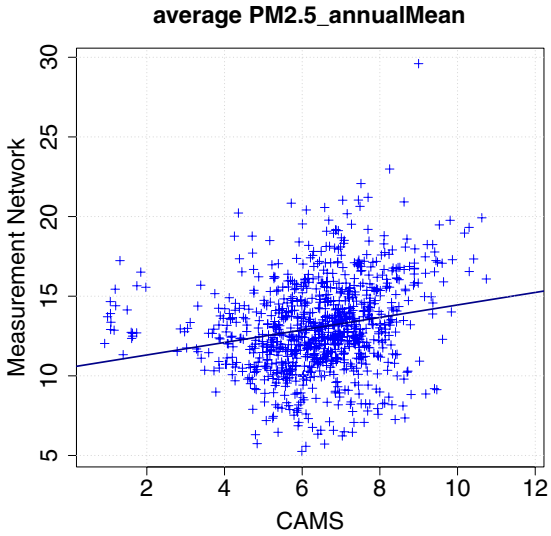

Fig. 4 An example of a relationship between an Airbase-derived metric and a CAMS-derived metric (shown here $PM_{2.5}$ annual average) for the NUTS-3 units where monitoring data was available with satisfactory temporal coverage. The linear scaling function thus computed (solid line in the plot) is used to produce the CAMS-derived metrics for the NUTS-3 units where monitoring data was not available with satisfactory temporal coverage.

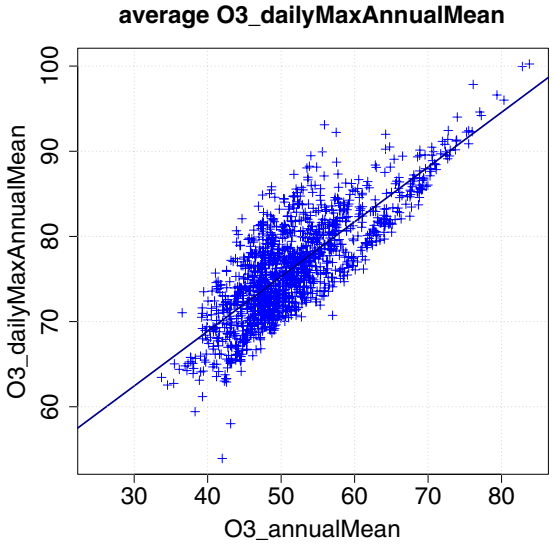

Fig. 5 Relationship between the monitoring-based O_3 daily maximum annual mean and the monitoring-based O_3 annual mean for the average scenario. The linear function (solid line in the plot) will be used to derive the CAMS-based O_3 daily maximum annual mean for the NUTS-3 units where monitoring data was not available with satisfactory temporal coverage.

for a given year. There are 402 NUTS-3 units in Germany and 3 scenarios were developed, the total number of records in the dataset is 1206 per year, or 12060 for the entire study period. Each record includes a numeric value for each metric considered.

The ratification status of the Airbase data used for each NUTS-3, year, scenario and metric is given in the file *APExpose_DE__2010–2019__Ratified.csv*. The station types used for each NUTS-3, year, scenario and metric are listed in the file *APExpose_DE__2010–2019__StationTypes.csv*. These two metadata files have the same structure as the main file.

While we plan to expand the dataset to cover further European countries, this was beyond the current scope of this study. As updates or expansions to the dataset are carried out, these will be noted in the open access dataset.

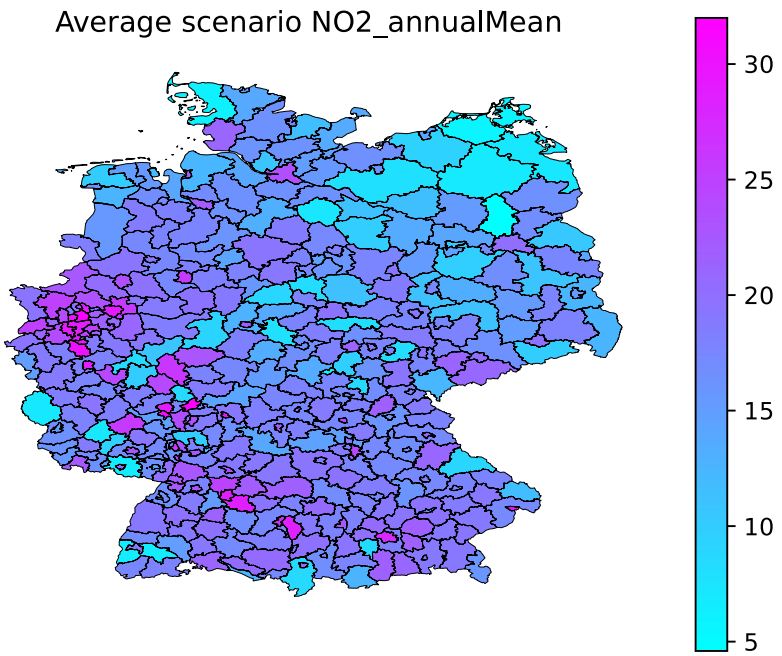

Fig. 6 NO_2 decadal (2010–2019) average concentration ($\mu g\ m^{-3}$) from the APExpose_DE dataset.

Technical Validation

The air quality data used to generate this dataset goes through quality assurance and quality control before being made officially available. In addition, we only used data from sites where 80 percent or more of the hourly data was available, so as to not introduce any seasonal or other bias.

Three different averaging options, corresponding to three scenarios (average, remote and urban), were evaluated for determining the concentration based on monitoring data for those NUTS-3 units where data was available. The comparison of these options showed that while differences did result, they were minor: the 95^{th} quantile of the relative difference between the rural or the urban scenario relative to the average scenario was 7.6% and 6.5%, respectively. The different options/scenarios are furthermore provided in the dataset and can be further evaluated and selected based on the use case where they are to be implemented.

Usage Notes

The ASCII format of the provided dataset enables a simple access and workup. The NUTS-3 code, provided for each record, enables linking the dataset to other, possibly vectorized, datasets at the NUTS-3 or coarser level.

Code availability

The code used to generate the dataset can be obtained under the same doi[22].

Received: 15 March 2021; Accepted: 16 September 2021;
Published online: 28 October 2021

References

1. Health Effects Institute. State of Global Air 2020. https://www.stateofglobalair.org/sites/default/files/documents/2020-10/soga-2020-report-10-26_0.pdf (2020).
2. Lelieveld, J. *et al.* Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective. *Cardiovascular Research* **116**, 1910–1917 (2020).
3. Pozzer, A. *et al.* Regional and global contributions of air pollution to risk of death from COVID-19. *Cardiovascular Research* **116**, 2247–2253 (2020).
4. Wu, X., Nethery, R. C., Sabath, M. B., Braun, D. & Dominici, F. Air pollution and covid-19 mortality in the united states: Strengths and limitations of an ecological regression analysis. *Science Advances* **6** (2020).
5. Cole, M. A., Ozgen, C. & Strobl, E. Air Pollution Exposure and Covid-19 in Dutch Municipalities. *Environmental and Resource Economics* **76**, 581–610 (2020).
6. Konstantinoudis, G. *et al.* Long-term exposure to air-pollution and covid-19 mortality in england: A hierarchical spatial analysis. *Environment International* **146**, 106316 (2021).
7. López-Feldman, A., Heres, D. & Marquez-Padilla, F. Air pollution exposure and covid-19: A look at mortality in mexico city using individual-level data. *Science of The Total Environment* **756**, 143929 (2021).
8. Travaglio, M. *et al.* Links between air pollution and covid-19 in england. *Environmental Pollution* **268**, 115859 (2021).
9. Khomenko, S. *et al.* Premature mortality due to air pollution in European cities: a health impact assessment. *The Lancet Planetary Health* **5**, e121–34 (2021).
10. Richardson, E. A., Pearce, J., Tunstall, H., Mitchell, R. & Shortt, N. K. Particulate air pollution and health inequalities: a europe-wide ecological analysis. *International Journal of Health Geographics* **12**, 34 (2013).
11. van Donkelaar, A. *et al.* Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. *Environmental Health Perspectives* **118**, 847–855 (2010).

12. van Donkelaar, A. *et al.* Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology* **50**, 3762–3772 (2016).
13. van Donkelaar, A., Martin, R. V., Li, C. & Burnett, R. T. Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology* **53**, 2595–2611 (2019).
14. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX (2008).
15. Duyzer, J., van den Hout, D., Zandveld, P. & van Ratingen, S. Representativeness of air quality monitoring networks. *Atmospheric Environment* **104**, 88–101 (2015).
16. Munir, S., Mayfield, M., Coca, D. & Jubb, S. A. Structuring an integrated air quality monitoring network in large urban areas â€“ discussing the purpose, criteria and deployment strategy. *Atmospheric Environment: X* **2**, 100027 (2019).
17. Li, H. Z. *et al.* Spatially dense air pollutant sampling: Implications of spatial variability on the representativeness of stationary air pollutant monitors. *Atmospheric Environment: X* **2**, 100012 (2019).
18. *EEA AirBase public air quality database* https://www.eea.europa.eu/themes/air/explore-air-pollution-data (2021).
19. *ECMWF Copernicus Atmosphere Monitoring Service Global Reanalysis (EAC4)* https://www.ecmwf.int/en/forecasts/dataset/cams-global-reanalysis (2021).
20. Inness, A. *et al.* CAMS global reanalysis (EAC4) https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-reanalysis-eac4?tab=overview (2019).
21. Bennouna, Y. *et al.* Validation report of the CAMS global reanalysis of aerosols and reactive gases, years 2003–2019 https://atmosphere.copernicus.eu/sites/default/files/2020-04/CAMS84_2018SC2_D5.1.1-2019.pdf (2020).
22. Caseiro, A. & von Schneidemesser, E. APExpose_DE. *Zenodo* https://doi.org/10.5281/zenodo.5093950 (2021).

Acknowledgements

We acknowledge the Copernicus Atmosphere Monitoring Service: the CAMS EAC4 reanalysis data was downloaded from the Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS) https://ads.atmosphere.copernicus.eu/#!/home using the dedicated API under the Licence Agreement Version 1.2 (November 2019). The work of the authors is supported by IASS Potsdam, with financial support provided by the Federal Ministry of Education and Research of Germany (BMBF) and the Ministry for Science, Research and Culture of the State of Brandenburg (MWFK).

Author contributions

Conception: E.v.S. and A.C.; dataset production: A.C.; results analysis: E.v.S. and A.C.; manuscript preparation: A.C.; manuscript revision: E.v.S.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.