Supplementary Information for

Global particulate pollution research reflects economic capacity more than pollution burden

Jiani Yang[1,*], Antong Zhang[2], Xun Jian[1], Kairui Qiu[3], Vijay Natraj[4], Hangyu Gao[5], Sally Newman[6], Joseph Pinto[7], and Yuk L.Yung[1]

[1] Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA.
[2] Department of Applied Mathematics, Brown University, Providence, RI, USA.
[3] Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA.
[4] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA.
[5]Department of Landscape Architecture, Universiti Putra Malaysia, Serdang, Malaysia.
[6]Bay Area Air Quality Management District, San Francisco, CA, USA
[7]Department of Environmental Sciences and Engineering, University of North Carolina, Chapel Hill, NC, USA.

* Correspondence: yjn@caltech.edu

Contents of this File

Supplementary Figures S1 to S13 and Table S1

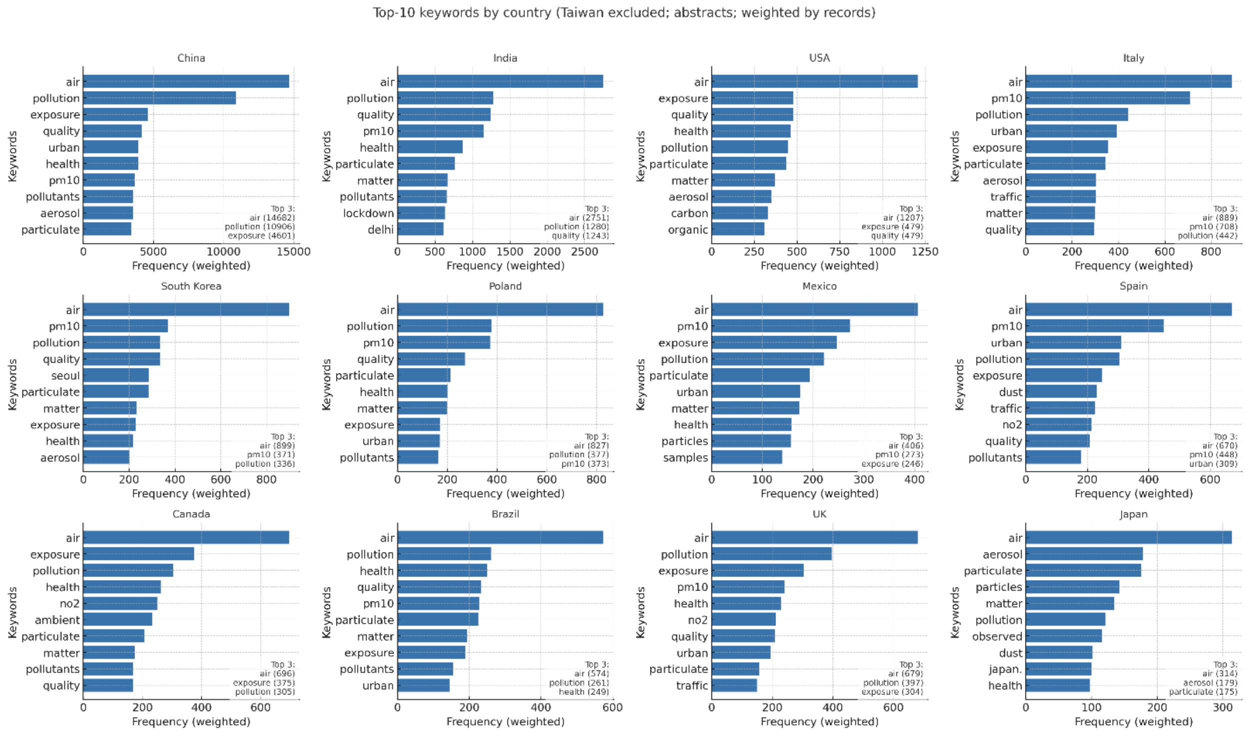

Figure S1. Country-level keyword distributions in $PM_{2.5}$-related research (1980–2025). Top ten keywords in $PM_{2.5}$ publications from the twelve leading publishing countries, derived from LLM assisted text mining of article titles and abstracts. Bars represent weighted keyword frequencies, with the three most frequent terms listed in each subplot.

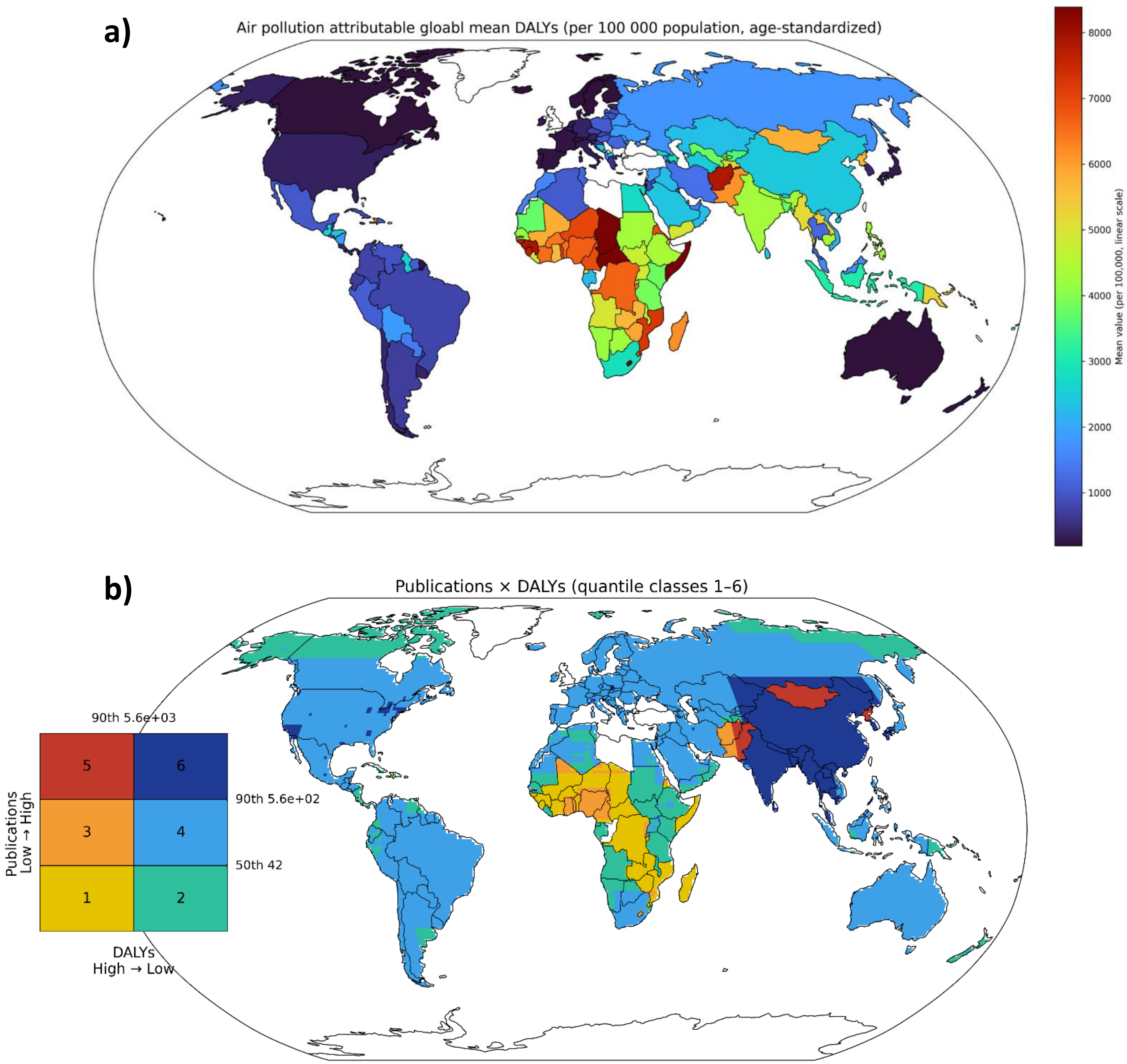

Figure S2. Global distribution of air-pollution health burden and research activity (1980–2025). (a) Country-level disability-adjusted life years (DALYs) attributable to air pollution per 100,000 population (age-standardized). Warmer colors denote greater health burdens. (b) Bivariate map showing joint quantiles of $PM_{2.5}$ publication density and DALYs. Red areas represent countries with high health burdens but low research output, whereas dark blue indicates both high DALYs and high publication activity. The 50th- and 90th-percentile thresholds for each variable are shown at left.

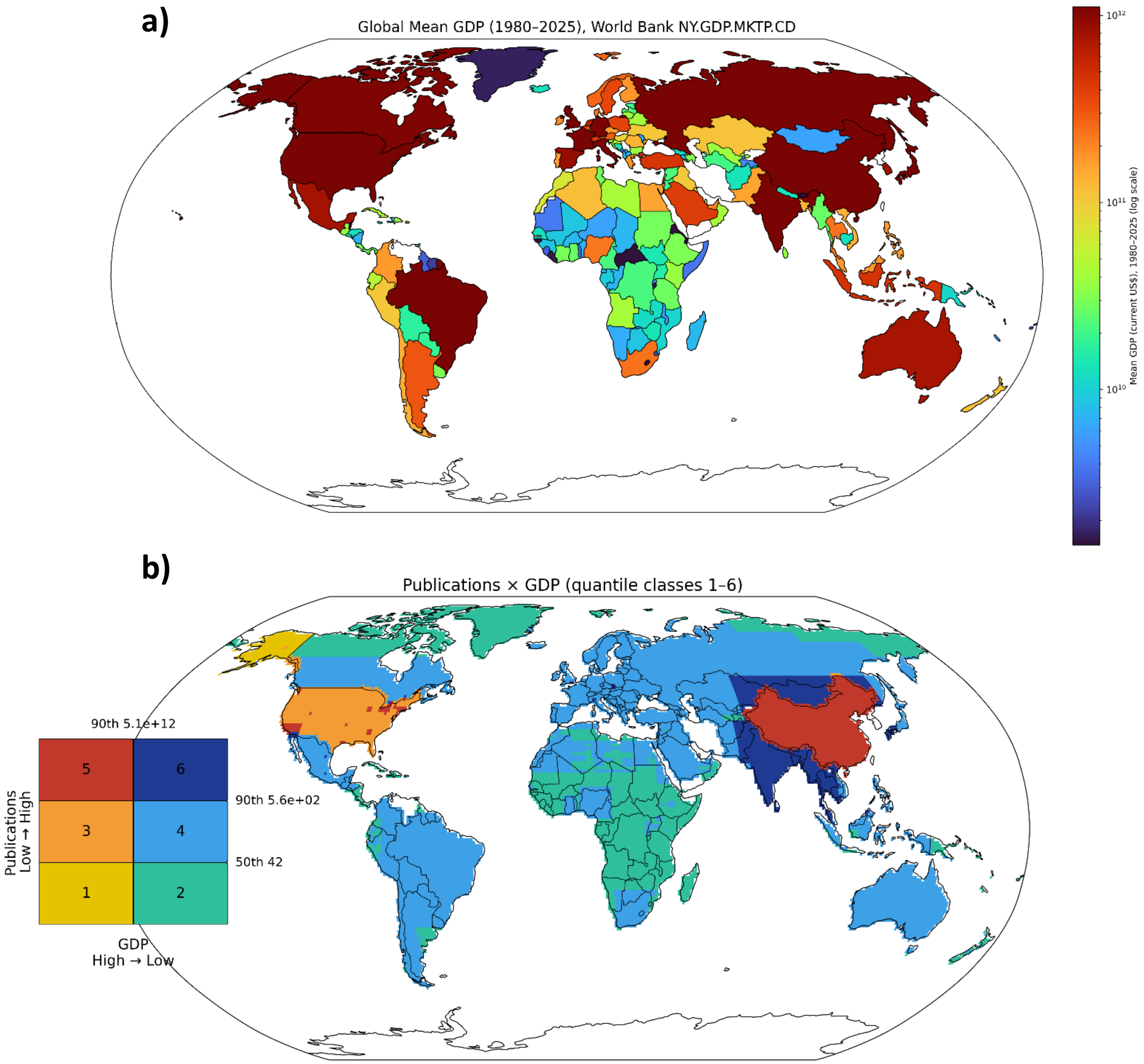

Figure S3. Relationship between national economic capacity and $PM_{2.5}$ research productivity (1980–2025). (a) Global mean gross domestic product (GDP) from 1980 to 2025 based on World Bank data (NY.GDP.MKTP.CD), shown on a logarithmic color scale. Darker red regions represent higher GDP. (b) Bivariate map showing the joint quantile distribution of national GDP and $PM_{2.5}$ publication density. Blue areas indicate countries with both high GDP and high research output, while red areas denote high-GDP but low-publication regions, and yellow–green areas mark low-income, low-research nations. The 50th- and 90th-percentile thresholds for GDP and publication density are shown at left.

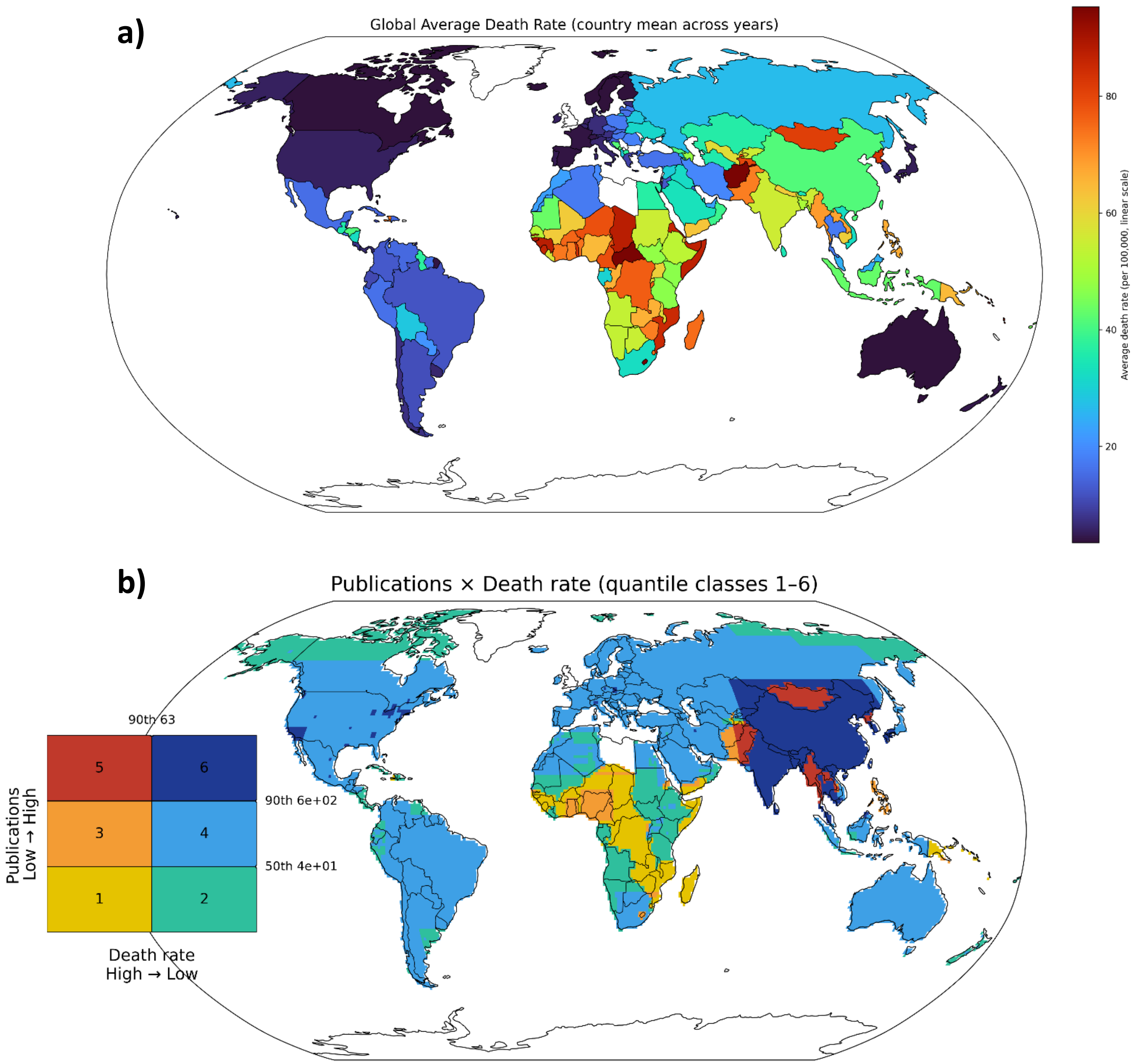

Figure S4. Global distribution of $PM_{2.5}$ attributable death rates and research activity (1980–2025). (a) Country-mean death rates (per 100,000 population) attributable to long-term $PM_{2.5}$ exposure, averaged across available years. Warmer colors indicate higher mortality burdens linked to particulate pollution. (b) Bivariate global map showing the joint quantile distribution of $PM_{2.5}$-related publication density and $PM_{2.5}$-attributable death rates. Blue areas represent countries with both high research activity and high mortality, whereas red areas denote high-mortality regions with limited research output. The 50th- and 90th-percentile thresholds for both variables are shown at left.

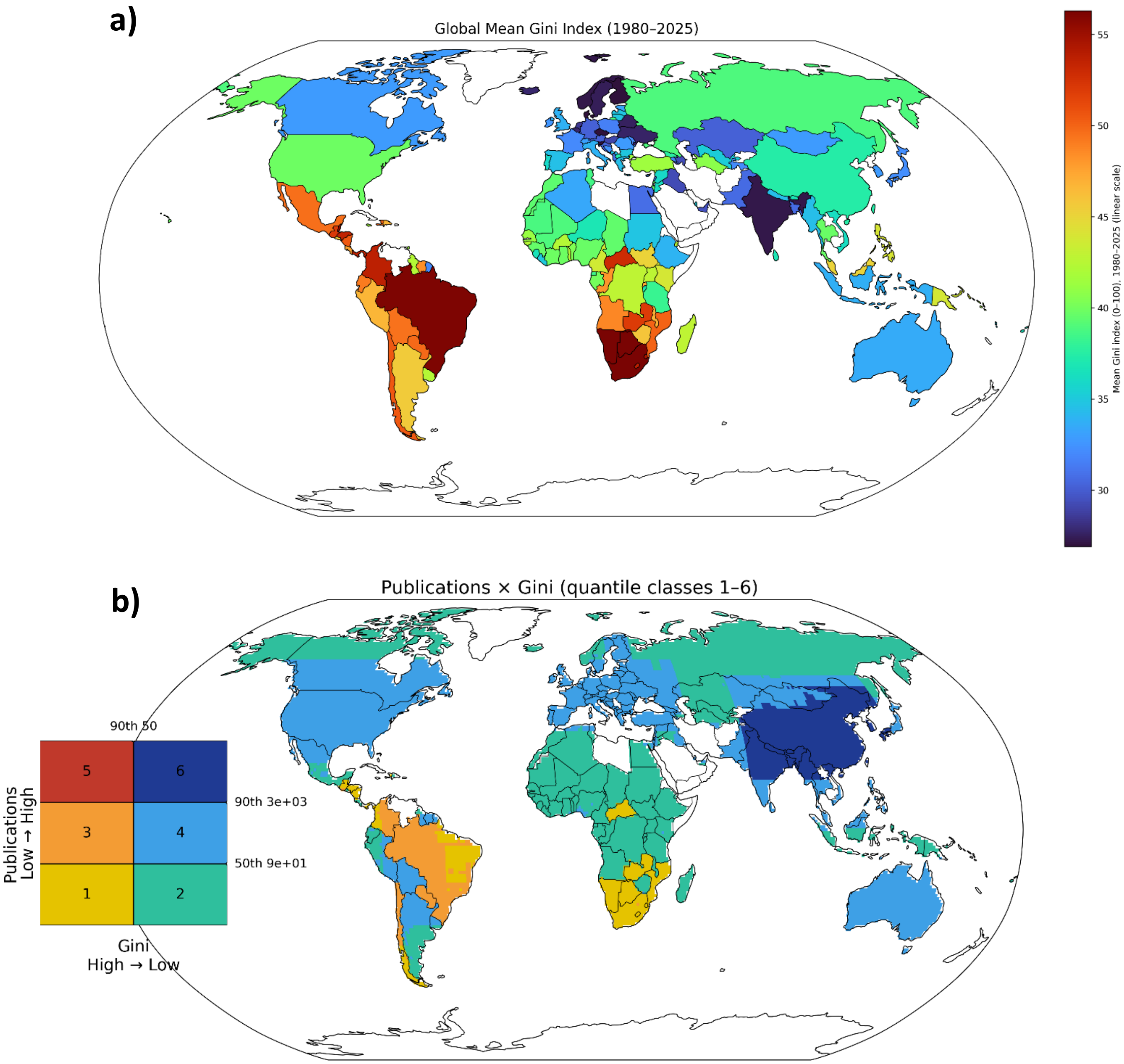

Figure S5. Global inequality and its relationship to $PM_{2.5}$ research distribution (1980–2025). (a) Global mean Gini index averaged from 1980 to 2025, representing income inequality by country (World Bank data). Warmer colors indicate higher inequality. (b) Bivariate map showing the joint quantile distribution of the national Gini index and $PM_{2.5}$ publication density. Blue areas correspond to countries with low inequality and high research activity, whereas red areas indicate high inequality but limited publication output. The 50th- and 90th-percentile thresholds for both variables are shown at left.

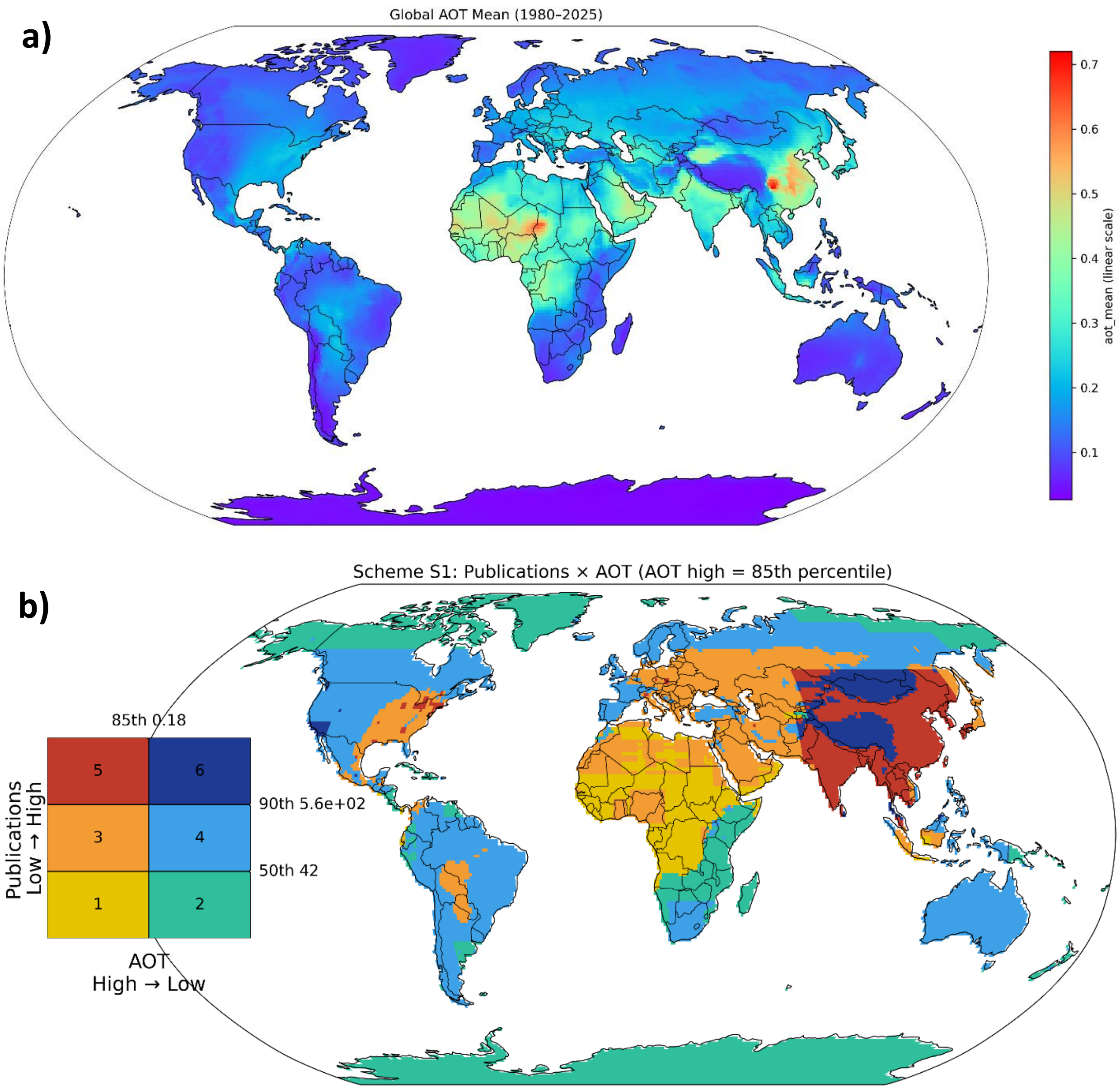

Figure. S6. Sensitivity of the global $PM_{2.5}$ research–exposure classification to a relaxed AOT threshold (1980–2025). (a) Mean monthly AOT from 1980–2025, derived from MERRA-2. Warmer colors denote regions with higher long-term aerosol loading. **(b)** Bivariate global map showing joint quantiles of publication density and AOT under Scheme S1, in which publication thresholds remain unchanged (50th and 90th percentiles), while high AOT is defined using the 85th percentile instead of the 90th percentile. Cells are colored by combinations of low, medium, and high publication density with low or high AOT. Red indicates high pollution but low research (knowledge gap), whereas dark blue indicates high AOT and high publication density. The annotated thresholds show the 50th- and 90th-percentile cutoffs for publication density and the 85th-percentile cutoff for AOT. Together, these panels show that relaxing the AOT threshold expands the spatial extent of high-exposure regions, while preserving the major mismatch patterns between pollution burden and research intensity.

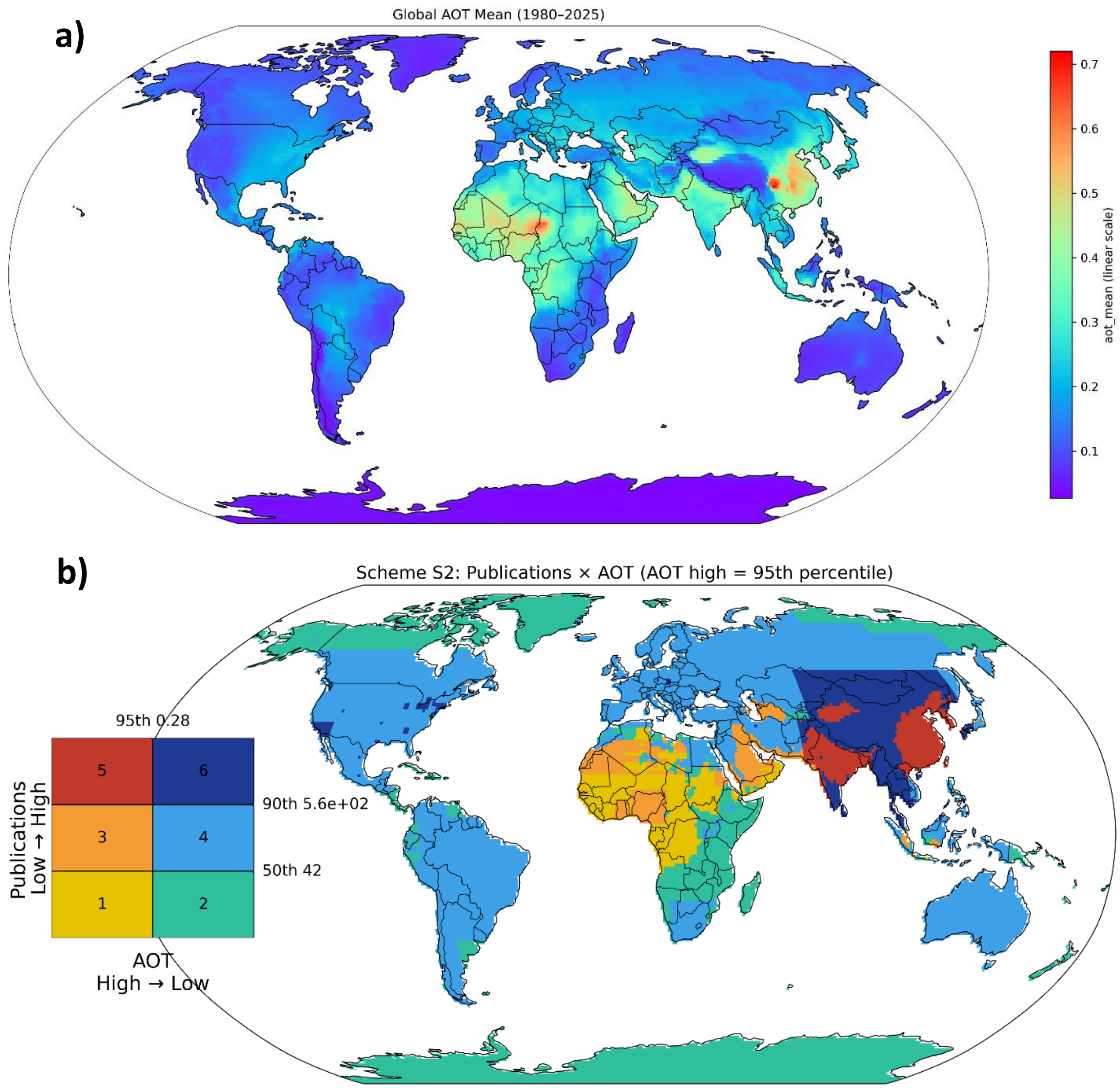

Figure. S7. Sensitivity of the global PM2.5 research–exposure classification to a stricter AOT threshold (1980–2025). (a) Mean monthly AOT from 1980–2025, derived from MERRA-2. Warmer colors denote regions with higher long-term aerosol loading. **(b)** Bivariate global map showing joint quantiles of publication density and AOT under Scheme S2, in which publication thresholds remain unchanged (50th and 90th percentiles), while high AOT is defined using the 95th percentile. Cells are colored by combinations of low, medium, and high publication density with low or high AOT. Red indicates high pollution but low research (knowledge gap), whereas dark blue indicates both high AOT and high publication density. The annotated thresholds show the 50th- and 90th-percentile cutoffs for publication density and the 95th-percentile cutoff for AOT. Together, these panels show that applying a stricter AOT threshold reduces the spatial extent of high-exposure regions while preserving the core mismatch patterns between pollution burden and research intensity.

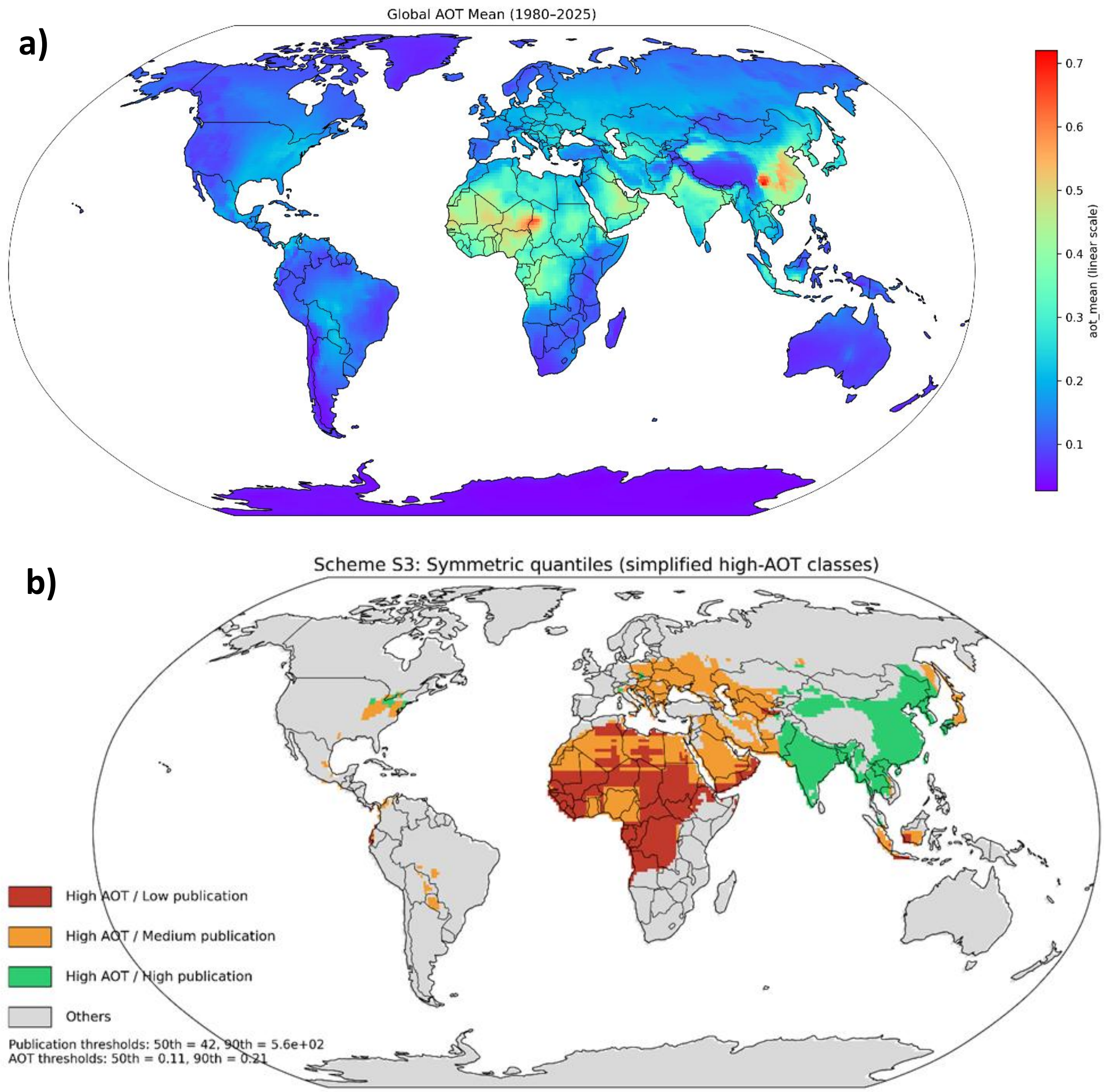

Figure. S8. Sensitivity of the global $PM_{2.5}$ research–exposure classification to symmetric quantile thresholds (1980–2025). (a) Mean monthly AOT from 1980–2025, derived from MERRA-2. Warmer colors denote regions with higher long-term aerosol loading. **(b)** Bivariate global map using **Scheme S3**, in which both publication density and AOT are classified using symmetric quantile thresholds (low: <50th percentile; medium: 50th–89th percentile; high: ≥90th percentile). The map is simplified to highlight the most policy-relevant combinations: high AOT / low publication (red), high AOT / medium publication (orange), high AOT / high publication (green), and all other regions (grey). Threshold values are annotated for both variables. These results confirm that regions such as sub-Saharan Africa and parts of South Asia consistently emerge as high-exposure, low-research areas, while regions with both high pollution and high research activity (e.g., parts of East Asia) remain distinct. The persistence of these patterns indicates that the identified knowledge–exposure gap is robust to alternative classification schemes.

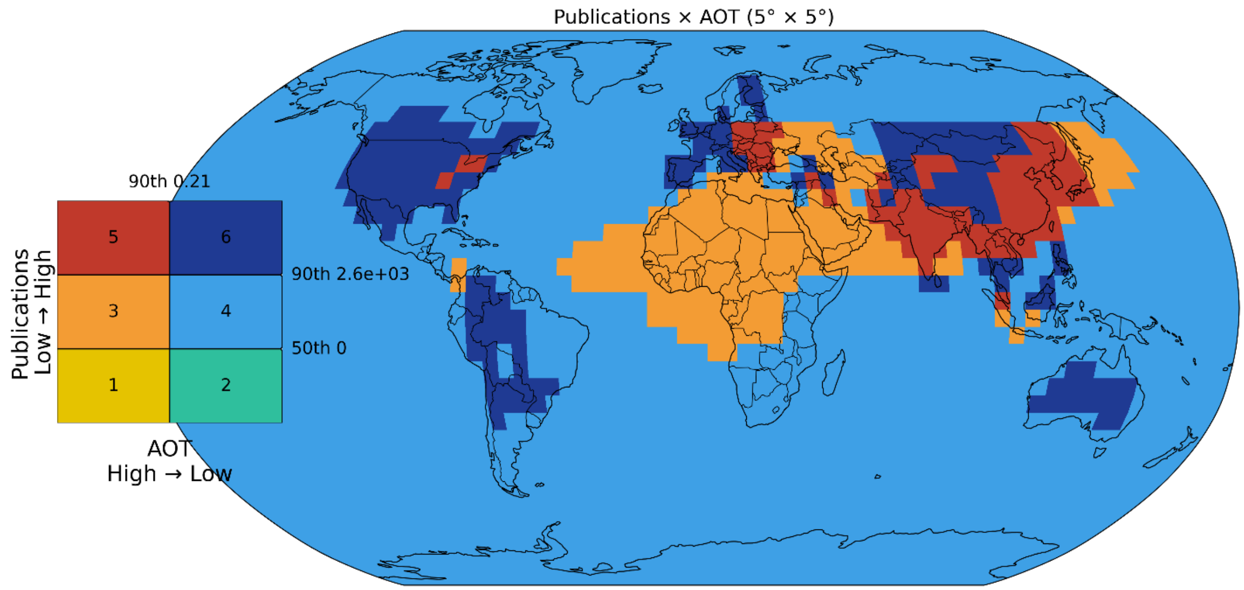

Figure. S9. Sensitivity of the knowledge–exposure gap to spatial aggregation (5° × 5° resolution). Global classification of grid cells based on publication density and aerosol optical thickness (AOT) after aggregating the original 1° × 1° fields to 5° × 5° resolution. Publication density is categorized into low (<50th percentile), medium (50th–89th percentile), and high (≥90th percentile), while AOT is classified using the 90th percentile threshold to identify high-exposure regions. The six classes correspond to combinations of publication intensity and pollution burden, as shown in the legend.

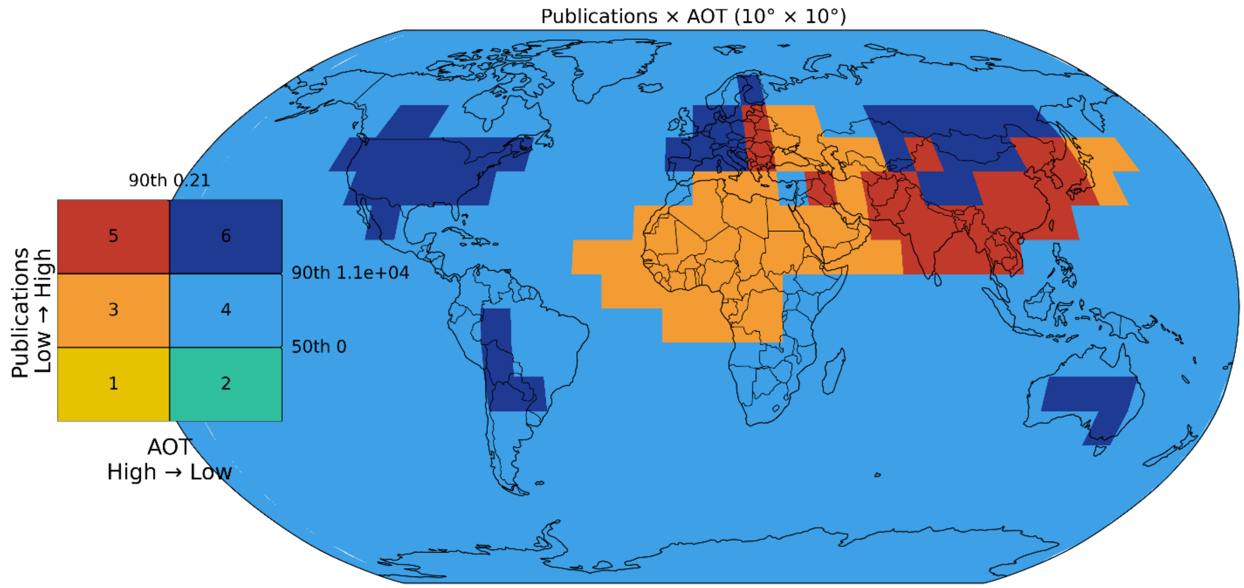

Figure. S10. Sensitivity of the knowledge–exposure gap to spatial aggregation (10° × 10° resolution). Global classification of grid cells based on publication density and aerosol optical thickness (AOT) after aggregating the original 1° × 1° fields to 10° × 10° resolution. Publication density is categorized into low (<50th percentile), medium (50th–89th percentile), and high (≥90th percentile), while AOT is classified using the 90th percentile threshold to identify high-exposure regions. The six classes correspond to combinations of publication intensity and pollution burden, as shown in the legend

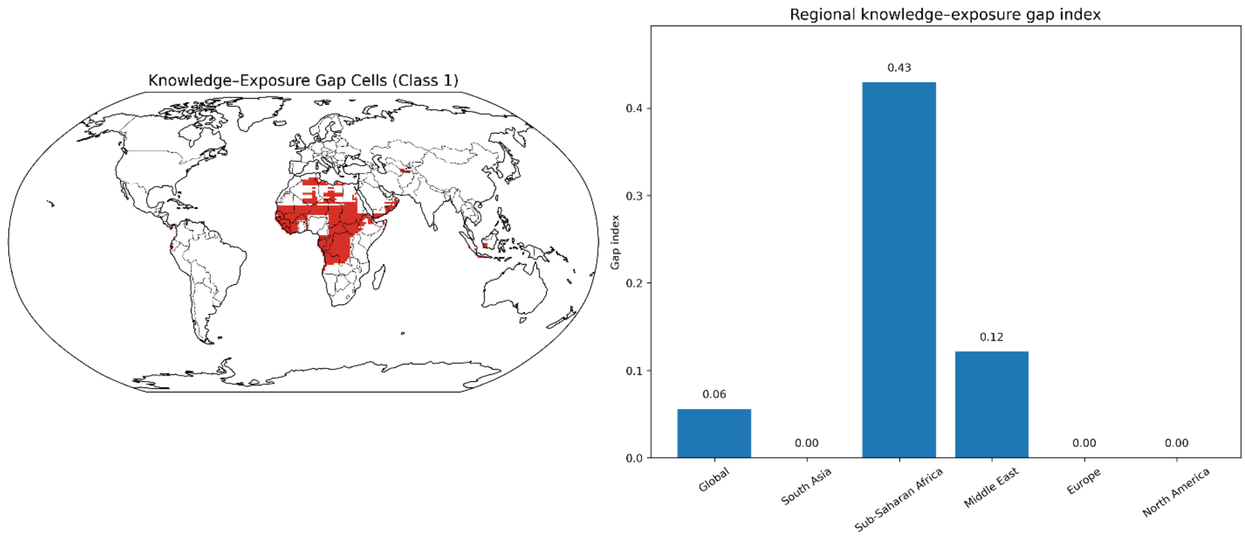

Figure. S11. Quantitative definition of the knowledge–exposure gap. Global map and regional summary of the knowledge–exposure gap index derived from the classification scheme in Fig. 3b. A grid cell is defined as belonging to the knowledge–exposure gap if it is classified as low publication density (<50th percentile) and high aerosol optical thickness (AOT ≥90th percentile), corresponding to class 1 in Fig. 3b. The left panel shows the spatial distribution of gap cells, and the right panel summarizes the fraction of such cells within selected regions. This index provides a quantitative measure of the mismatch between scientific attention and pollution burden.

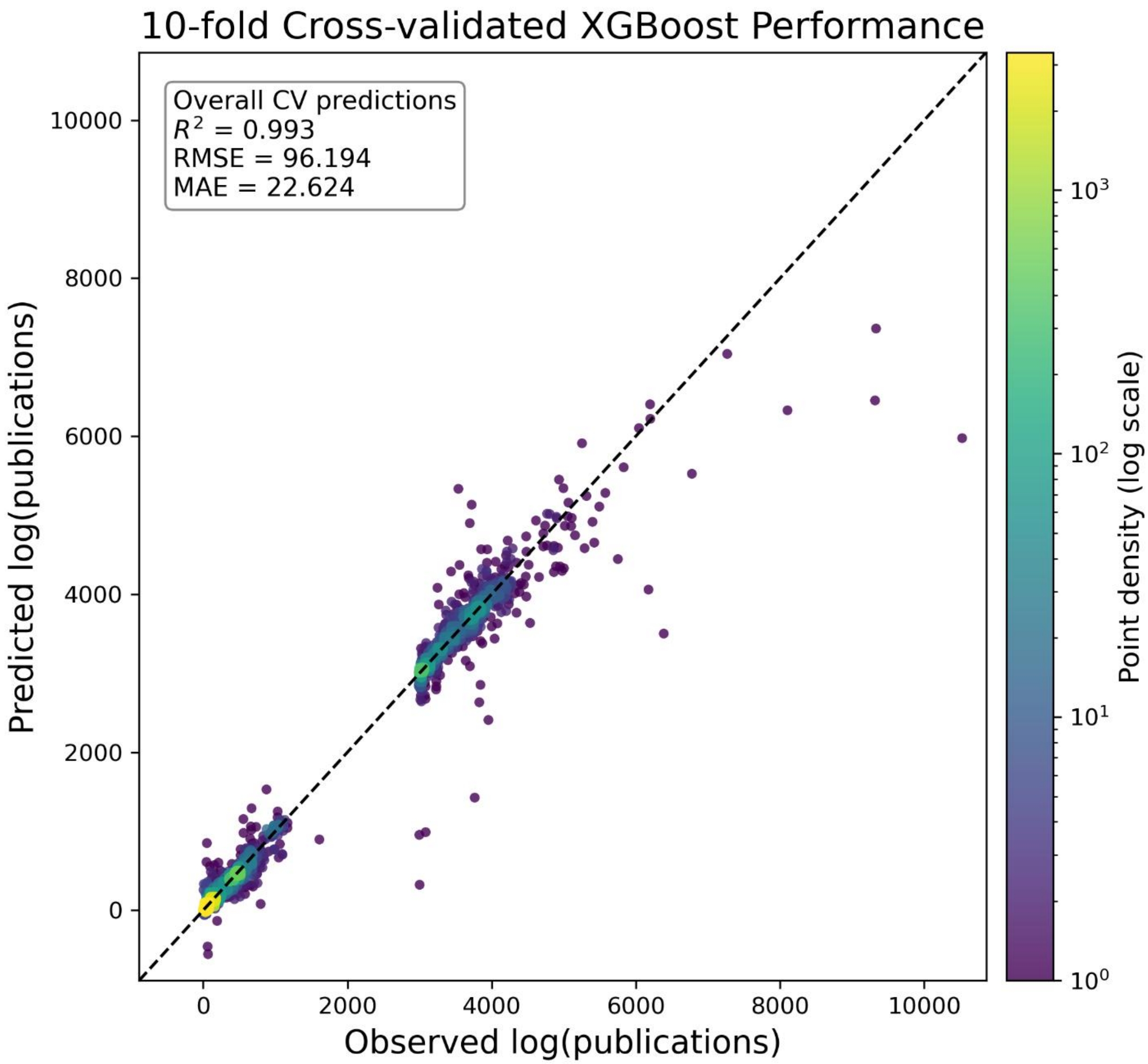

Figure S12. Cross-validated performance of the XGBoost model predicting global publication counts. Observed versus predicted log(publications) from pooled 10-fold cross-validation are shown, with point color indicating local data density on a logarithmic scale. The dashed line represents the 1:1 relationship. The model demonstrates strong predictive skill ($R^2 = 0.993$, RMSE = 96.2, MAE = 22.6), indicating that the machine learning framework reliably captures the spatial variability in publication intensity. High-density clusters correspond to regions with moderate publication activity, while sparse points at higher values reflect a small number of highly productive grid cells.

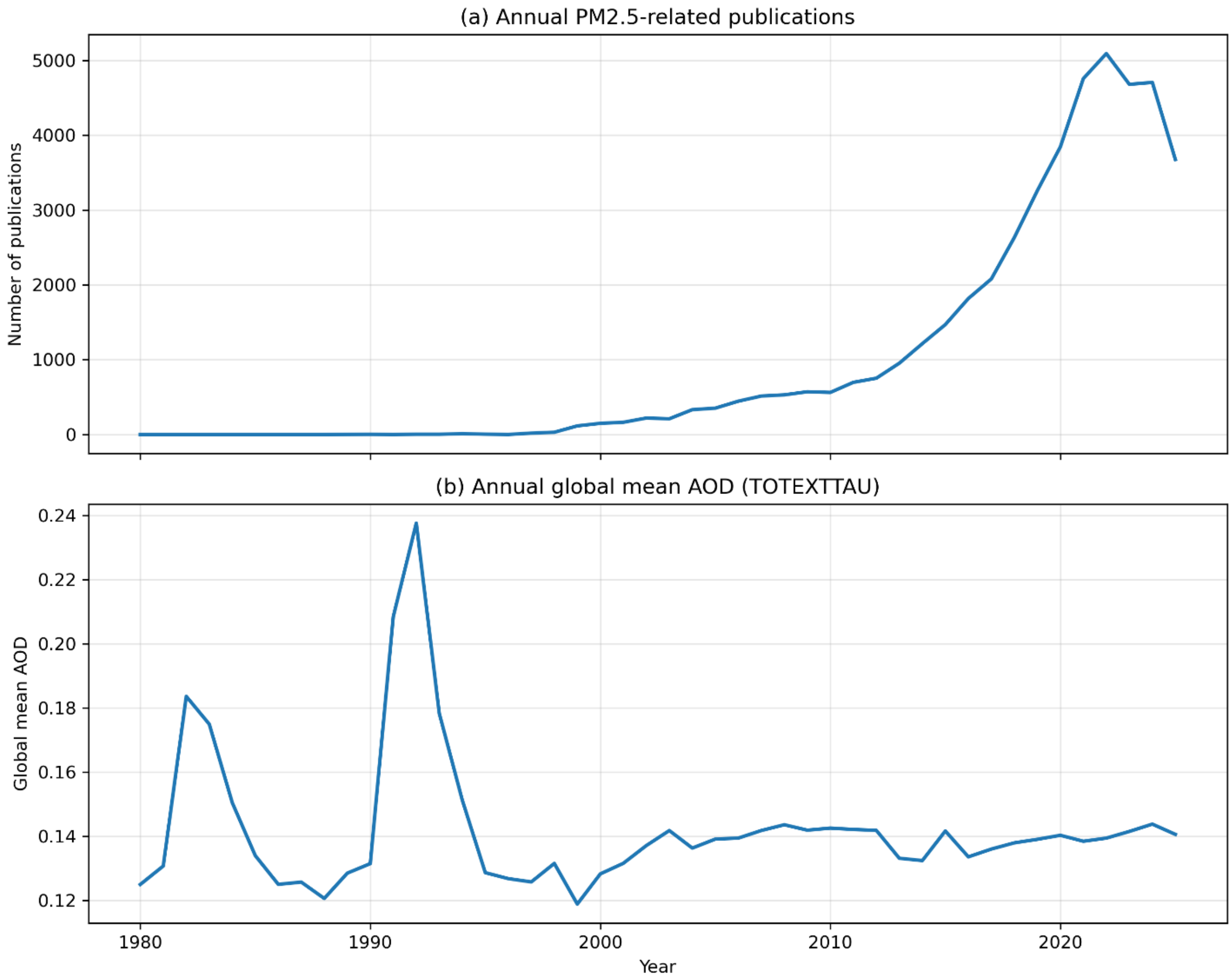

Figure S13. Temporal evolution of $PM_{2.5}$-related publications and global aerosol burden (1980–2025). (a) Annual number of $PM_{2.5}$-related publications derived from the literature dataset. Publication output remained low before the late 1990s, followed by a rapid increase after the early 2000s and a peak in the early 2020s. (b) Annual global mean aerosol optical depth (AOD) derived from reanalysis data. In contrast to the sharp increase in publication activity, global mean AOD exhibits relatively modest interannual variability, with no comparable long-term increase.

Table S1 Summary of datasets used in this study

Dataset/Variable	Source	Spatial Resolution	Temporal Coverage	Key Variables	Notes
$PM_{2.5}$ Publications	Web of Science	1° × 1° (LLM-mapped)	1980–Aug 2025	Publication counts (log-transformed)	Retrieved using keyword "PM2.5"; duplicates removed; valid abstracts only
Aerosol Optical Thickness (AOT)	MERRA-2 reanalysis	0.5° × 0.625° (regridded to 1°)	1980–2025 (mean)	AOT	Represents long-term aerosol burden
DALYs	WHO Global Health Estimates	Country-level (regridded to 1°)	1980–2025 (mean)	Disability-adjusted life years	Health burden indicator
Death Rate (PM2.5-related)	WHO Global Health Estimates	Country-level (regridded to 1°)	1980–2025 (mean)	Mortality attributable to PM2.5	Age-standardized estimates
GDP	World Bank	Country-level (regridded to 1°)	1980–2025 (mean)	Gross domestic product	Economic capacity indicator
Gini Index	World Bank	Country-level (regridded to 1°)	1980–2025 (mean)	Income inequality	Social equity indicator
Population	World Bank	Country-level (regridded to 1°)	1980–2025 (mean)	Total population	Demographic scale
Geographic Coordinates	Derived	1° × 1°	Static	Latitude, Longitude	Included as spatial predictors