nature portfolio

Peer Review File

Global particulate pollution research reflects economic capacity more than pollution burden

Corresponding Author: Ms Jiani Yang

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
General Comments:

1. There is confusion regarding the purpose, aim, and scope of the study. The major research questions and contributions are not clearly described in the Introduction. The authors should explicitly state the study objectives.

2. The manuscript lacks clear structural separation between sections. Several paragraphs discuss and merge multiple themes, making it difficult to follow the logical progression of the study. Paragraphs should be reorganized to align correctly with each subsection.

3. Figure 4b presents SHAP parameters that are main model drivers, yet its description is insufficiently explained in the text. The authors should clearly describe how SHAP values are interpreted and how Figure 4b supports the study's conclusions regarding parameter importance.

4. What I have understood is that the paper's central claim is about geographic inequality in research, this database choice may amplify the very bias the authors seek to diagnose. This limitation is not sufficiently acknowledged or discussed, and its implications for the results are not quantitatively explored. Without a sensitivity analysis using alternative databases (e.g., Scopus, Dimensions), it is difficult to disentangle real disparities from database artifacts.

5. Strengthen reproducibility and transparency and more clearly position the work as a methodological advance rather than a conceptual breakthrough.

6. The analysis treats all publications equally, without considering impact factor or journal quality, which could further nuance the interpretation of "research output".

7. Providing access to the LLM prompts and post-processing scripts used for geoparsing, clarifying API versioning.

8. Could normalizing publication output by number of researchers or institutions alter the interpretation of inequality?
9. To what extent do international collaborations mask local underrepresentation (i.e., studies conducted in low-income regions but led by institutions in high-income countries)?

Specific Comments:

1.Line 18: The criteria and methodology used for selecting publications are not sufficiently described. The authors should clarify the inclusion and exclusion criteria, database limitations, and whether any filtering beyond keyword matching was applied.

2.Line 23: The term "knowledge–exposure gap" requires correction and consistent usage throughout the manuscript (e.g., Lines 23 and 46).

3.Line 26: requires grammatical correction. The sentence describing asymmetry between regions that are studied and those

that inhale PM2.5 should be rewritten for clarity and precision.

4.Line 30: The introduction of PM2.5 is abrupt and relies solely on a parenthetical definition. A brief introductory sentence explaining the relevance of PM2.5 prior to the parenthetical definition would improve clarity and readability.

5.Line 41: The sentence at Line 41 is grammatically incorrect and should be revised for clarity.

6.Line 161: The manuscript introduces XGBoost and SHAP without sufficient explanation. A brief description of why these methods were chosen and how they contribute to the analysis should be included to improve accessibility for a broader audience.

7.Line 191: A methodological flowchart summarizing how AI tools are applied in the bibliometric workflow would significantly enhance clarity, particularly for readers unfamiliar with LLM-based literature analysis.

8.Line 221: The choice of Web of Science as the sole literature source may systematically under-represent regional or non-English journals, particularly from Africa and Latin America. While this limitation is implicit, it should be discussed more explicitly as it directly affects the interpretation of geographic inequality.

9.Line 222: How sensitive are the results to the chosen 1° × 1° spatial resolution?

10.Line 252: The F1 score of 0.75 for LLM-based geoparsing is reasonable, but the manuscript would benefit from a brief sensitivity discussion on how parsing uncertainty may affect spatial publication density patterns.

11.Line 254: The accuracy of the LLM-based abstract selection and geoparsing requires more detailed discussion. The authors should address uncertainty, potential sources of error, and how reproducibility is ensured if the study is replicated in the future.

12.Line 270: The machine-learning model should be briefly introduced, including its general purpose and how it is applied in the present study, to aid readers unfamiliar with ensemble learning methods.

13.Line 276: A table summarizing all input datasets, including data sources, spatial and temporal resolution, and key variables, would improve transparency and reproducibility.

Reviewer #2

(Remarks to the Author)
Abstract:
The abstract clearly lays out the central idea of a “knowledge–exposure gap” and does a good job highlighting the scale of the dataset and the use of LLM-based analysis. That said, the statement that publication density “correlates strongly” with GDP is somewhat misleading. SHAP values reflect feature importance within a model, not correlation coefficients, so the wording should be adjusted for precision. In addition, since the dataset extends to 2025, the authors should clarify how incomplete records for the most recent year were handled. Without this explanation, it is difficult to assess whether the apparent temporal trends are influenced by partially indexed data.

Introduction:
The introduction makes a persuasive case for moving beyond traditional bibliometric approaches and presents the LLM-based framework as a way to capture more nuanced thematic structure. Framing research output itself as a dimension of environmental inequity is a notable conceptual contribution. However, the role of the LLM appears somewhat overstated. Based on the methods, GPT-4o-mini was used primarily for geoparsing rather than deeper thematic synthesis. It would strengthen the paper to clarify this distinction. The authors should also more clearly articulate how their work advances beyond prior bibliometric studies (e.g., references 9 and 18), beyond the methodological novelty of using LLMs.

Results — Knowledge Flows:
The three-tiered structure connecting Aerosol Science, Mechanisms, and Technologies is a helpful way to visualize how the field has evolved, particularly the shift toward AI-enabled prediction. However, the manuscript does not explain how the thematic categories in Fig. 1 were generated. It remains unclear whether labels such as “Transport” or “Neural Network” were assigned by the LLM, derived from keyword matching, or manually curated. This methodological transparency is important. Additionally, the observation that “Wildfire” and “Climate Change” remain peripheral is intriguing, but it is not clear whether this reflects a genuine gap in integrated research or simply a limitation of the narrow “PM2.5” search term.

Results — Geographical Distribution and Knowledge–Exposure Gap:
The comparison between AOT and publication density effectively highlights high-exposure, low-research regions, particularly in South Asia and sub-Saharan Africa. The country-level keyword analysis also adds useful context. However, the reported geoparsing F1 score of 0.75 indicates a non-trivial error rate, which could introduce spatial uncertainty at the 1° resolution used. This potential source of noise should be acknowledged more explicitly. In addition, the choice of quantile thresholds is asymmetric (50th and 90th percentiles for publications, but only the 90th for AOT), and the rationale for this decision is not explained. A sensitivity analysis would strengthen confidence in the classification scheme.

Results — SHAP Interpretability:
The SHAP analysis offers a compelling data-driven perspective on the role of socioeconomic factors in shaping global research output, and the apparent GDP threshold effect is particularly interesting. However, the manuscript does not report basic model performance metrics (R^2, RMSE, MAE) in the results section. Without these, it is difficult to assess the reliability of the SHAP interpretations. The authors should also specify more clearly where the GDP threshold occurs, as this would make the finding more concrete and potentially more policy-relevant.

Discussion:
The discussion thoughtfully considers the possibility that limited research capacity may help perpetuate high pollution burdens by weakening regulatory momentum. While this argument is intellectually appealing, it remains speculative in the absence of supporting evidence or historical examples where expanded local research capacity directly contributed to successful air-quality interventions. The discussion should either provide such references or adopt more cautious language. In addition, the reliance on Web of Science introduces a likely English-language and indexing bias, which could partially explain the apparent underrepresentation of low-income regions. This limitation deserves explicit acknowledgment.

Materials and Methods:
Including the full GPT-4o-mini system prompt and detailing the integration of MERRA-2, World Bank, and WHO datasets enhances reproducibility. However, several methodological issues merit closer attention. Nearest-neighbor interpolation of country-level socioeconomic data onto a 1° grid may introduce artificial discontinuities at national borders. Standard random 10-fold cross-validation may also produce overly optimistic performance estimates given the spatial autocorrelation inherent in gridded publication data; spatial block cross-validation would be more appropriate. Finally, the proportion of abstracts returning “None+” in the geoparsing step is not reported, leaving uncertainty about the effective spatial coverage of the dataset.

Reviewer #3

(Remarks to the Author)
Summary
This manuscript analyzes more than 55,000 $PM_{2.5}$-related publications published between 1980 and 2025. A language model is used to identify research themes and extract geographic focus from abstracts. Publication density is then compared with aerosol optical thickness, health burden indicators, and socioeconomic variables. A machine learning model is applied to examine which factors are most strongly associated with research output. The authors conclude that regions with high pollution burden often produce fewer studies and that research productivity is more closely linked to economic capacity than to environmental conditions. The integration of text analysis with geospatial and socioeconomic data is potentially useful. However, the main conclusions are presented in broad terms, while the analysis remains largely descriptive. Several key concepts are not clearly defined in quantitative terms, and important methodological choices require further justification.

Major Comments
1. The manuscript refers repeatedly to a “knowledge–exposure gap,” but this term is not clearly defined as a measurable quantity. At present, the gap is illustrated through maps and relative comparisons. To strengthen the argument, the authors should specify how this gap is calculated. For example, is it the difference between expected and observed publication density given pollution burden? Is there a formal index or statistical test?
2. The study uses aerosol optical thickness as the primary measure of environmental burden. AOT does not directly represent surface $PM_{2.5}$ exposure and may reflect dust or elevated aerosol layers that are not equivalent to ground-level pollution. Because the main conclusion is that pollution burden is weakly associated with research output, the choice of indicator is important.
3. The machine learning model identifies GDP and population as strong predictors of publication density. This pattern is consistent with established findings in scientometrics. The manuscript would benefit from clearer reporting of model performance, such as R^2 and error measures, and discussion of whether results remain stable under alternative model specifications. In addition, the Discussion occasionally implies that limited research activity may contribute to persistent pollution. The current design supports associations but not causal conclusions.
4. Although the dataset spans several decades, most analyses rely on long-term averages. Examining how research output changes over time in relation to pollution trends or policy changes could provide additional insight.
5. The language model used to extract geographic information reports an overall accuracy score, but there is limited information about how accuracy varies across regions or types of studies.
6. Several sections rely on general statements about inequality and asymmetry without clearly linking them to specific quantitative results.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I have reviewed the revised version of the manuscript. The authors have addressed the majority of my previous comments and suggestions effectively, and I believe the paper is now significantly stronger.

I recommend the manuscript for publication; however, I have one final suggestion to enhance the work's clarity. To improve transparency and reproducibility, the authors should include a summary table of all input datasets.

Reviewer #2

(Remarks to the Author)
The feedback and revise the management as per the review.

Reviewer #1 (Remarks to the Author):

General Comments:

1. There is confusion regarding the purpose, aim, and scope of the study. The major research questions and contributions are not clearly described in the Introduction. The authors should explicitly state the study objectives.

We thank the reviewer for this important comment. We agree that the original Introduction did not clearly articulate the purpose, scope, and specific objectives of the study.

To address this, we have substantially revised the Introduction to explicitly state the research questions and contributions of this work. In particular, we now clearly define three primary objectives:

(1) to quantify the global distribution and temporal evolution of $PM_{2.5}$ research using a large-scale bibliometric dataset;

(2) to identify and characterize the "knowledge–exposure gap" by comparing research output with aerosol burden and health indicators; and

(3) to investigate the socioeconomic and environmental drivers of research inequality using a machine-learning framework with SHAP-based interpretability.

These objectives are now explicitly stated at the end of the Introduction (Page 2, Lines 58–66), which improves the clarity of the study's scope and contributions.

2. The manuscript lacks clear structural separation between sections. Several paragraphs discuss and merge multiple themes, making it difficult to follow the logical progression of the study. Paragraphs should be reorganized to align correctly with each subsection.

We thank the reviewer for this helpful suggestion.

We agree that the original manuscript could be improved by providing clearer structural separation between sections and better alignment between paragraphs and their respective themes. In the revised version, we have reorganized the manuscript to ensure that each subsection focuses on a single coherent topic, with improved transitions between sections.

Specifically, we have separated previously merged discussions into distinct paragraphs corresponding to (i) spatial distribution patterns, (ii) quantitative assessment of inequality using the gap index, (iii) machine learning–based attribution of drivers, and (iv) temporal evolution of research activity. Each paragraph now directly supports a specific analytical component and is explicitly linked to the corresponding figures.

These revisions improve the logical flow and readability of the manuscript, making the progression of the study clearer and more accessible.

3. Figure 4b presents SHAP parameters that are main model drivers, yet its description is insufficiently explained in the text. The authors should clearly describe how SHAP values are interpreted and how Figure 4b supports the study's conclusions regarding parameter importance.
We thank the reviewer for this helpful comment. We have revised the manuscript to provide a clearer explanation of how SHAP values are interpreted, including the meaning of positive and negative contributions relative to predicted publication density. We also expanded the description of Fig. 4b to guide readers in interpreting the dependence plots, explicitly explaining the roles of individual data points, the LOWESS trend, and the nonlinear relationships. In addition, we provide concrete examples (e.g., GDP, population, AOT, and Gini) to illustrate how the figure supports our conclusions regarding the dominant role of socioeconomic factors. These revisions improve both clarity and accessibility for a broader audience:

" SHAP values represent the marginal contribution of each predictor to the model output, where positive values indicate that a variable increases the predicted publication density relative to the global mean, and negative values indicate that it decreases the prediction. In Fig. 4b, each point represents an individual grid cell, and its position along the x-axis corresponds to the value of the predictor, while its y-axis value shows the SHAP contribution of that predictor for that location. The red LOWESS curve summarizes the average trend, highlighting nonlinear relationships. For example, in the GDP panel, grid cells with low GDP values cluster around near-zero or slightly negative SHAP values, indicating little contribution to publication output. As GDP increases, SHAP values rise sharply, showing that high-income regions strongly increase predicted publication density. Similarly, in the population panel, SHAP values increase steadily with population, indicating a positive contribution of population size to research output. In contrast, the AOT panel shows SHAP values distributed close to zero across the entire range, indicating that aerosol burden has minimal influence on predicted publication density. For the Gini index, SHAP values tend to become negative at higher inequality levels, suggesting that greater income inequality reduces research output. Together, these patterns demonstrate that socioeconomic variables exert strong and systematic effects on publication density, whereas environmental variables contribute little, supporting the conclusion that global research disparities are primarily driven by socioeconomic factors."

4. What I have understood is that the paper's central claim is about geographic inequality in research, this database choice may amplify the very bias the authors seek to diagnose. This limitation is not sufficiently acknowledged or discussed, and its implications for the results are not quantitatively explored. Without a sensitivity analysis using alternative databases (e.g., Scopus, Dimensions), it is difficult to disentangle real disparities from database artifacts.

We thank the reviewer for this insightful and important comment. We agree that the use of a single bibliographic database (Web of Science) may introduce systematic biases and potentially amplify existing disparities in global research representation.

In response, we have added a detailed discussion of this limitation in the manuscript. Specifically, we now acknowledge that Web of Science may under-represent regional and non-English publications, particularly from low- and middle-income regions, and that this could contribute to the observed knowledge–exposure gap.

At the same time, we clarify that several lines of evidence support the robustness of our main findings, including the consistency of observed patterns with established relationships between research output and socioeconomic indicators, and the fact that disparities are observed at broad regional scales unlikely to be driven solely by database coverage.

We also highlight that future work will incorporate additional databases such as Scopus and Dimensions to further assess the sensitivity of our results. These revisions improve the transparency and interpretation of our findings.

We added sentences below in the discussion part:

"This study relies on the Web of Science as the primary source of bibliometric data, which may introduce systematic biases in the representation of global research output. Web of Science predominantly indexes English-language and high-impact journals, potentially under-representing regional publications from low- and middle-income countries, particularly in Africa and Latin America. As a result, the observed disparities in publication density may partly reflect database coverage bias rather than purely underlying differences in research activity. However, because Web of Science remains one of the most widely used and consistently curated bibliographic databases, it provides a standardized basis for large-scale comparative analysis. Future work could integrate additional sources such as Scopus or Dimensions to further assess the robustness of the observed patterns and reduce potential coverage bias."

5. Strengthen reproducibility and transparency and more clearly position the work as a methodological advance rather than a conceptual breakthrough.

We thank the reviewer for this insightful suggestion. We have revised the manuscript to strengthen both reproducibility and transparency, and to more clearly position the work as a methodological contribution.

Specifically, we now emphasize that the study introduces a scalable and reproducible analytical framework that integrates LLM-based bibliometrics, geospatial analysis, and machine-learning interpretability. We have also clarified that all prompts, post-processing scripts, and analysis workflows are publicly available, ensuring that the approach can be replicated and extended.

In addition, we have revised the Introduction and Discussion to highlight the generalizability of this framework beyond $PM_{2.5}$ research, positioning the work as a reusable methodology for analyzing global knowledge disparities.

6. The analysis treats all publications equally, without considering impact factor or journal quality, which could further nuance the interpretation of "research output".

We thank the reviewer for this insightful comment. We agree that treating all publications equally may overlook differences in research impact and journal quality. In response, we have added a discussion acknowledging that our analysis focuses on publication quantity rather than impact, and that this may obscure additional layers of inequality. We also highlight that future work could incorporate weighted metrics, such as citation counts or journal impact factors, to further refine the analysis. These additions improve the interpretation and transparency of our findings:

"In addition, this analysis treats all publications equally, without accounting for differences in journal impact, citation influence, or research quality. As a result, the reported publication density reflects the quantity rather than the relative impact of scientific output. This simplification may obscure finer-scale disparities, as high-income regions may disproportionately contribute to higher-impact publications, while lower-income regions may be underrepresented both in quantity and visibility. However, the primary objective of this study is to characterize large-scale patterns in research activity, for which publication counts provide a consistent and interpretable metric. Future work could incorporate weighted measures of research output, such as citation counts or journal impact metrics, to further refine the assessment of global knowledge disparities."

7. Providing access to the LLM prompts and post-processing scripts used for geoparsing, clarifying API versioning.

We thank the reviewer for this helpful suggestion. We have clarified the reproducibility of the LLM-based workflow by explicitly providing access to all prompts and post-processing scripts in an open-access repository (https://doi.org/10.5281/zenodo.17445423).

We also specify that the geoparsing was performed using the OpenAI API (v1, Chat Completions endpoint) with the gpt-4o-mini model, and that all model calls were conducted using a fixed prompt template and consistent parsing rules. API keys are managed securely and are not included in the repository. These additions improve transparency and reproducibility.

8. Could normalizing publication output by number of researchers or institutions alter the interpretation of inequality?

We thank the reviewer for this insightful comment. We agree that normalizing publication output by the number of researchers or institutions could provide a complementary perspective on global research inequality.

In response, we have added a discussion clarifying that our analysis focuses on absolute publication counts, which primarily capture differences in the scale of scientific systems rather than per-capita productivity. We also note that while normalization could affect the relative ranking of some regions, the overall patterns—particularly the strong role of socioeconomic capacity and the weak association with environmental burden—are expected to remain robust.

We highlight capacity-adjusted metrics as an important direction for future work, although such analyses are currently limited by the availability of globally consistent data.

"Another consideration is that publication output in this study is measured in absolute terms, without normalization by the number of researchers, institutions, or national research capacity. As a result, the observed disparities primarily reflect differences in the scale of scientific systems rather than research productivity per capita. Normalizing publication output by indicators such as the research workforce or institutional density could provide a complementary perspective, potentially highlighting regions with relatively high research efficiency despite lower overall output.

However, globally consistent data on research capacity remain limited, particularly for low- and middle-income countries, making such normalization challenging at a global scale. Importantly, the main patterns identified here—namely the strong association of research output with socioeconomic capacity and the weak association with environmental burden—are expected to persist even under normalized metrics, although the relative ranking of some regions may change. Future work could incorporate capacity-adjusted indicators to further disentangle differences in research scale versus research efficiency."

9. To what extent do international collaborations mask local underrepresentation (i.e., studies conducted in low-income regions but led by institutions in high-income countries)?

We thank the reviewer for this insightful suggestion. We have revised the manuscript as below:

"Another important consideration is the role of international collaboration in shaping the geographic distribution of research. In this study, publications are assigned based on the locations described in the abstracts, which capture where research is conducted rather than the institutional origin of the authors. As a result, studies carried out in low-income regions but led by institutions in high-income countries may partially mask local underrepresentation in scientific capacity.

This distinction implies that the spatial patterns reported here reflect the geography of research activity rather than the distribution of research leadership or capacity. In some cases, international collaborations may lead to an overestimation of local research engagement in underrepresented regions, as externally led studies contribute to publication counts without necessarily building sustained local infrastructure. At the same time, such collaborations may also play a positive role in knowledge transfer and capacity building.

Future work could integrate author affiliation data or funding sources to disentangle the contributions of local versus external research leadership, providing a more nuanced understanding of global knowledge inequalities."

Specific Comments:

1.Line 18: The criteria and methodology used for selecting publications are not sufficiently described. The authors should clarify the inclusion and exclusion criteria, database limitations, and whether any filtering beyond keyword matching was applied.

We have revised the Abstract to explicitly state that publications were retrieved from the Web of Science using keyword queries centered on "$PM_{2.5}$," clarifying the primary search criterion used to construct the dataset. Now the sentence has been changed as

"we analyzed more than 55,000 publications retrieved from the Web of Science using standardized keyword queries centered on "$PM_{2.5}$" from 1980 to 2025, with duplicate records removed and entries filtered to retain peer-reviewed articles with valid abstracts, to map the evolving structure of aerosol science, mechanisms, and technologies."

2.Line 23: The term "knowledge–exposure gap" requires correction and consistent usage throughout the manuscript (e.g., Lines 23 and 46).

We thank the reviewer for pointing out the inconsistency in terminology. We have revised the manuscript to ensure consistent usage of the term "knowledge–exposure gap" throughout the text, including correcting the hyphenation and formatting. In addition, we have clarified its meaning upon first use in the Introduction to improve conceptual clarity.

3.Line 26: requires grammatical correction. The sentence describing asymmetry between regions that are studied and those that inhale PM2.5 should be rewritten for clarity and precision.

We thank the reviewer for this helpful comment. We have revised the sentence as "Our artificial intelligence-driven synthesis reveals a mismatch between where $PM_{2.5}$ research is produced and where exposure is highest, calling for more equitable scientific and policy engagement."

4.Line 30: The introduction of PM2.5 is abrupt and relies solely on a parenthetical definition. A brief introductory sentence explaining the relevance of PM2.5 prior to the parenthetical definition would improve clarity and readability.

We thank the reviewer for this helpful suggestion. We have revised the opening sentence of the Introduction to provide a broader context on air pollution before introducing $PM_{2.5}$. This improves the logical flow and readability by first establishing the global relevance of air pollution and then identifying $PM_{2.5}$ as a key component. The sentence has been changed as

"Air pollution is a major global environmental and public health challenge, contributing substantially to disease burden and climate forcing. Fine particulate matter ($PM_{2.5}$, particles smaller than 2.5 μm in aerodynamic diameter) represents one of its most harmful components and remains a leading environmental health threat worldwide."

5.Line 41: The sentence at Line 41 is grammatically incorrect and should be revised for clarity.

The sentence has been changed as:

"Research output reflecting who studies air pollution, where, and with what technological tools remains a key yet under-examined dimension of global inequality."

6.Line 161: The manuscript introduces XGBoost and SHAP without sufficient explanation. A brief description of why these methods were chosen and how they contribute to the analysis should be included to improve accessibility for a broader audience.

We thank the reviewer for this helpful suggestion. We have revised the manuscript to provide a brief explanation of the rationale for using XGBoost and SHAP. Specifically, we now clarify that XGBoost is employed to capture nonlinear relationships and interactions among socioeconomic and environmental predictors, while SHAP analysis is used to interpret the marginal contribution of each feature to model predictions. These additions improve accessibility for readers unfamiliar with these methods. We added sentences as:

"To quantify the relative importance of socioeconomic and environmental factors in shaping global research output, we employ an XGBoost regression model, a tree-based ensemble learning approach well suited for capturing nonlinear relationships and interactions among predictors. To enhance interpretability, we further apply SHapley Additive exPlanation (SHAP) analysis, which decomposes model predictions into the marginal contribution of each feature. The XGBoost model, trained on global socioeconomic and environmental predictors, reveals a pronounced geographical and economic disparity in scientific publication output (Fig. 4). SHAP analysis identifies geographical location (longitude and latitude), population size, and GDP as the dominant positive drivers of publication counts, whereas public health and social equity indicators play a secondary role, and environmental factors such as AOT have negligible influence (Fig. 4a)."

7.Line 191: A methodological flowchart summarizing how AI tools are applied in the

bibliometric workflow would significantly enhance clarity, particularly for readers unfamiliar with LLM-based literature analysis.

We thank the reviewer for this helpful suggestion. We note that a methodological flowchart is already provided in Fig. 5. To improve clarity, we have revised the manuscript to explicitly reference this figure and added a concise description of the workflow, including data collection, LLM-based geoparsing, and machine-learning analysis with SHAP interpretation. These revisions make the methodological pipeline clearer, particularly for readers unfamiliar with LLM-based bibliometric approaches. The sentence has been changed to:

"The overall workflow of the study is summarized in Fig. 5, which illustrates the integration of LLM-based bibliometric analysis and machine-learning interpretability. Briefly, PM2.5-related publications are first collected and filtered from the Web of Science, followed by LLM-based geoparsing to extract spatial information. These data are then combined with environmental and socioeconomic variables and analyzed using an XGBoost model with SHAP-based interpretation to quantify the drivers of global research disparities. This integrated framework provides a quantitative explanation for the spatial and socioeconomic asymmetries in PM2.5 research revealed in earlier sections."

8.Line 221: The choice of Web of Science as the sole literature source may systematically under-represent regional or non-English journals, particularly from Africa and Latin America. While this limitation is implicit, it should be discussed more explicitly as it directly affects the interpretation of geographic inequality.

We thank the reviewer for this important comment. We agree that reliance on the Web of Science may introduce systematic biases, particularly in under-representing regional and non-English publications from low- and middle-income countries. We have now added an explicit discussion of this limitation in the manuscript, including its potential impact on the interpretation of geographic disparities in research output. We also highlight that future work could incorporate additional databases such as Scopus or Dimensions to further assess the robustness of our findings. We added the following sentences in the Discussion:

"This study relies on the Web of Science as the primary source of bibliometric data, which may introduce systematic biases in the representation of global research output. Web of Science predominantly indexes English-language and high-impact journals, potentially under-representing regional publications from low- and middle-income countries, particularly in Africa and Latin America. As a result, the observed disparities in publication density may partly reflect database coverage bias rather than purely underlying differences in research activity. However, because Web of Science remains one of the most widely used and consistently curated bibliographic databases, it provides a standardized basis for large-scale comparative analysis. Future work could integrate additional sources such as Scopus or Dimensions to further assess the robustness of the observed patterns and reduce potential coverage bias."

9.Line 222: How sensitive are the results to the chosen 1° × 1° spatial resolution?

We thank the reviewer for this insightful comment. To address this concern, we conducted additional sensitivity analyses by aggregating the 1° × 1° publication and AOT fields to coarser spatial resolutions (5° × 5° and 10° × 10°) and repeating the classification used in Fig. 3b. The results are shown in the revised Supplementary Fig.9-10.

We find that the key spatial patterns remain robust across resolutions. In particular, South Asia consistently emerges as a high-exposure, low-research hotspot (class 1), even after aggregation to 5° and 10° grids. Sub-Saharan Africa also remains predominantly classified as low- or medium-publication with elevated AOT (classes 1–3), indicating persistent underrepresentation relative to pollution burden.

While spatial aggregation smooths local variability and reduces small-scale heterogeneity—particularly in regions such as Southeast Asia and parts of South America—the large-scale mismatch between pollution exposure and research output remains clearly visible. These results indicate that the main conclusions of Fig. 3b are not driven by grid-scale noise or geoparsing uncertainty, but reflect robust, large-scale patterns.

We have also clarified in the manuscript the rationale for the asymmetric quantile thresholds. Publication density is highly skewed and benefits from a three-tier classification (low, medium, high), whereas the upper decile of AOT is used to isolate the most pollution-burdened regions. Additional tests using alternative thresholds (not shown) yield consistent spatial patterns.

These additions strengthen confidence that the identified knowledge–exposure gap is robust to both spatial resolution and classification choices.

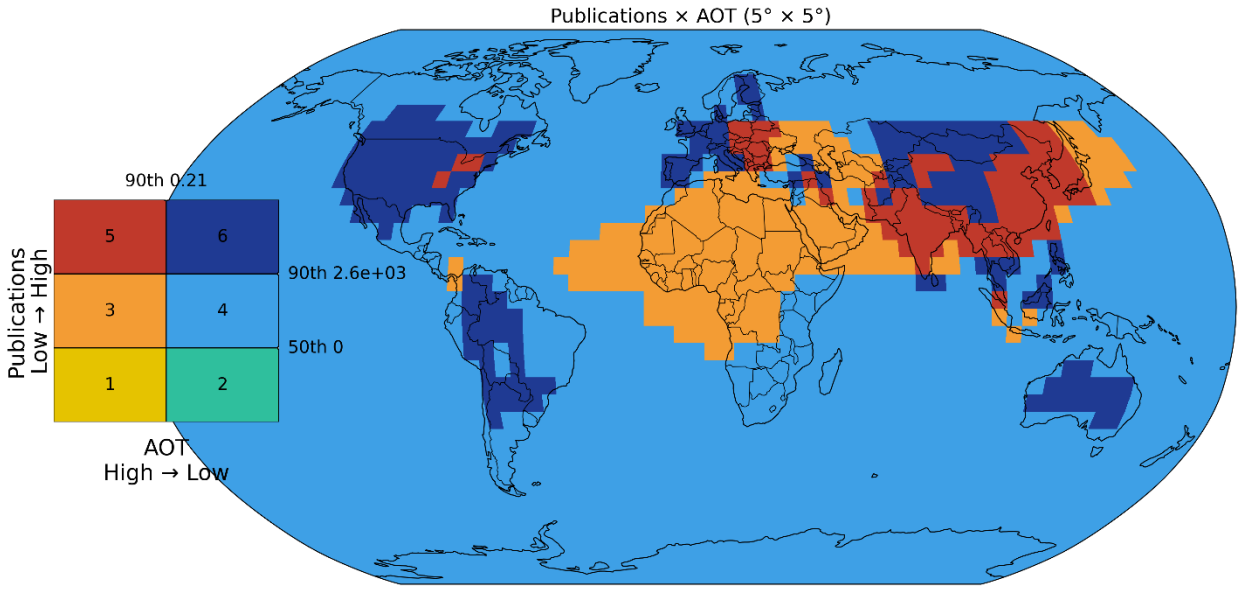

Figure. S9. Sensitivity of the knowledge–exposure gap to spatial aggregation (5° × 5° resolution). Global classification of grid cells based on publication density and aerosol optical thickness (AOT) after aggregating the original 1° × 1° fields to 5° × 5° resolution. Publication density is categorized 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.

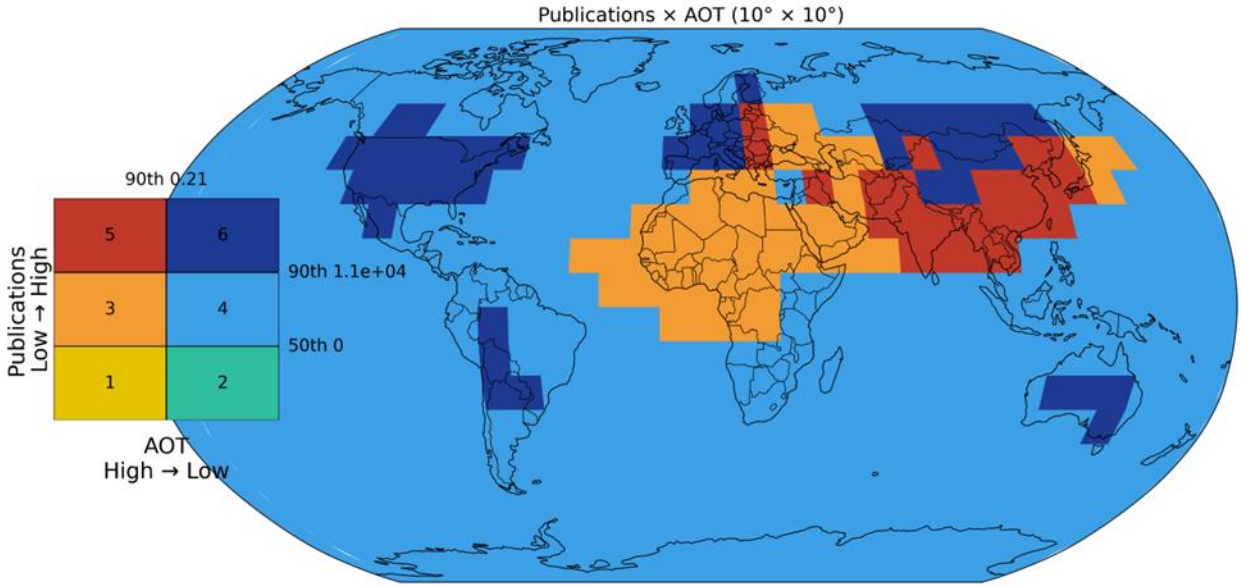

Figure. S10. Sensitivity of the knowledge–exposure gap to spatial aggregation (10° × 10° resolution). Global classification of grid cells based on publication density and aerosol optical thickness (AOT) after aggregating the original 1° × 1° fields to 10° × 10° resolution. Publication density is categorized 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

"To assess the sensitivity of our results to spatial resolution, we conducted additional analyses using coarser grids (e.g., 5° × 5° and 10° × 10°). The large-scale spatial patterns of publication density and the identified knowledge–exposure gap remain robust across resolutions, although finer grids provide greater spatial detail. These results suggest that our conclusions are not sensitive to the specific choice of 1° × 1° resolution (Fig.S9-S10)."

10.Line 252: The F1 score of 0.75 for LLM-based geoparsing is reasonable, but the manuscript would benefit from a brief sensitivity discussion on how parsing uncertainty may affect spatial publication density patterns.

We thank the reviewer for this insightful comment. We have added a discussion on the potential impact of geoparsing uncertainty on spatial publication density patterns. Specifically, we clarify that while the F1 score of 0.75 indicates moderate accuracy, the resulting errors are expected to introduce localized noise rather than systematic bias, and are unlikely to affect large-scale spatial patterns. We also note that finer-scale patterns in data-sparse regions may be more sensitive to such uncertainty. This clarification has been incorporated into the manuscript.

"While an F1 score of 0.75 indicates moderate accuracy in geoparsing, the resulting uncertainties are expected to introduce localized spatial noise rather than systematic bias in large-scale patterns. Because publication density is aggregated over a 1° × 1° grid and analyzed at regional to continental scales, random parsing errors are unlikely to substantially alter the overall spatial distribution of research activity. However, finer-scale patterns in data-sparse regions may be more sensitive to geoparsing uncertainty, and this limitation should be considered when interpreting localized results."

11.Line 254: The accuracy of the LLM-based abstract selection and geoparsing requires more detailed discussion. The authors should address uncertainty, potential sources of error, and how reproducibility is ensured if the study is replicated in the future.

We thank the reviewer for this important comment. We have added a detailed discussion addressing uncertainty, potential sources of error, and reproducibility in the LLM-based abstract processing and geoparsing.

Specifically, we now describe key sources of uncertainty, including ambiguity in geographic descriptions, variability in LLM outputs, and spatial aggregation effects. We also clarify that these uncertainties are expected to introduce localized noise rather than systematic bias, and are unlikely to affect large-scale spatial patterns.

In addition, we have strengthened the reproducibility of the study by providing the full prompt, data-processing workflow, and analysis code in an open-access repository, and by describing the use of consistent parsing rules and fixed prompt templates. These revisions improve the transparency and robustness of the methodological framework:

"The accuracy of the LLM-based abstract processing and geoparsing introduces several sources of uncertainty. First, ambiguity in geographic descriptions within abstracts (e.g., regional names, multi-location studies, or implicit references) may lead to incomplete or imprecise location extraction. Second, the probabilistic nature of large language models can result in variability in parsing outputs, particularly for texts with limited contextual clarity. Third, the use of a fixed 1° × 1° grid may further propagate minor location errors into neighboring cells.

Despite these uncertainties, several factors support the robustness of our results. The geoparsing performance, evaluated against human-annotated data, achieves an F1 score of 0.75, indicating reasonable accuracy for large-scale analysis. Moreover, because publication density is aggregated at regional to continental scales, random extraction errors are expected to introduce localized noise rather than systematic bias in the overall spatial patterns.

To enhance reproducibility, we provide the full LLM prompt, data-processing workflow, and analysis code in an open-access repository. All model calls were performed using a fixed prompt template and consistent parsing rules to ensure methodological consistency. While minor variability in LLM outputs may occur across different API versions, the large sample size and aggregation strategy reduce sensitivity to such variations. Future work may further improve robustness by incorporating ensemble parsing strategies or benchmarking across multiple LLM models."

12.Line 270: The machine-learning model should be briefly introduced, including its general purpose and how it is applied in the present study, to aid readers unfamiliar with ensemble learning methods.

We thank the reviewer for this helpful suggestion. We have revised the manuscript to provide a clearer introduction to the machine-learning model, including both its general purpose and its specific role in this study. In particular, we now clarify that XGBoost is used to model gridded publication density as a function of socioeconomic and environmental variables, enabling identification of the relative importance of key drivers. This improves accessibility for readers unfamiliar with ensemble learning methods:

"To quantify how socioeconomic and environmental factors contribute to global disparities in $PM_{2.5}$ research output, we applied an XGBoost regression model, a tree-based ensemble learning method capable of capturing nonlinear relationships and interactions among predictors. In this study, the model is used to predict gridded publication density from multiple explanatory variables, enabling identification of the relative importance of factors such as GDP, population, and aerosol optical thickness."

13.Line 276: A table summarizing all input datasets, including data sources, spatial and temporal resolution, and key variables, would improve transparency and reproducibility.

We thank the reviewer for this helpful suggestion. We have added a new table (Table S1) summarizing all input datasets, including their sources, spatial and temporal resolution, and key variables. This addition improves the transparency and reproducibility of the study by clearly documenting all data inputs used in the analysis.

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

Reviewer #2 (Remarks to the Author):

Abstract:

The abstract clearly lays out the central idea of a "knowledge–exposure gap" and does a good job highlighting the scale of the dataset and the use of LLM-based analysis. That said, the statement that publication density "correlates strongly" with GDP is somewhat misleading. SHAP values reflect feature importance within a model, not correlation coefficients, so the wording should be adjusted for precision. In addition, since the dataset extends to 2025, the authors should clarify how incomplete records for the most recent year were handled. Without this explanation, it is difficult to assess whether the apparent temporal trends are influenced by partially indexed data.

We thank the reviewer for this insightful comment. We have revised the Abstract to improve the precision of the statistical interpretation by replacing "correlates strongly" with wording that reflects SHAP-based feature importance rather than correlation.

In addition, we now explicitly clarify that the dataset extends to August 2025 and acknowledge the potential for incomplete indexing in the most recent year. To address this, we state that analyses of temporal trends primarily rely on fully indexed years prior to 2025. These revisions improve both the clarity and methodological transparency of the Abstract.

The abstract has been revised as below:

"Global efforts to mitigate fine particulate matter ($PM_{2.5}$) pollution remain unevenly distributed across regions and research domains. Using large language models, we analyzed more than 55,000 publications retrieved from the Web of Science using standardized keyword queries centered on "$PM_{2.5}$" from 1980 to 2025, with duplicate records removed and entries filtered to retain peer-reviewed articles with valid abstracts, to map the evolving structure of aerosol science, mechanisms, and technologies. To minimize potential bias from incomplete indexing in the most recent year, analyses of temporal trends primarily focus on fully indexed years prior to 2025. Despite rapid growth since the 2000s, we find a pronounced geographical imbalance between regions most affected by high aerosol optical thickness (AOT) and those leading PM2.5 research. Machine learning interpretation further reveals that publication density is primarily driven by gross domestic product and population, but shows weak association with aerosol optical thickness (AOT) or mortality rates, highlighting a global knowledge–exposure gap. Regions such as South Asia and sub-Saharan Africa experience the world's highest particulate loadings yet contribute the fewest studies. Our artificial intelligence-driven synthesis reveals a mismatch between where $PM_{2.5}$ research is produced and where exposure is highest, calling for more equitable scientific and policy engagement."

Introduction:

The introduction makes a persuasive case for moving beyond traditional bibliometric approaches and presents the LLM-based framework as a way to capture more nuanced thematic structure. Framing research output itself as a dimension of environmental inequity is a notable conceptual contribution. However, the role of the LLM appears somewhat overstated. Based on the methods, GPT-4o-mini was used primarily for geoparsing rather than deeper thematic synthesis. It would

strengthen the paper to clarify this distinction. The authors should also more clearly articulate how their work advances beyond prior bibliometric studies (e.g., references 9 and 18), beyond the methodological novelty of using LLMs.

We thank the reviewer for this insightful comment. We agree that the role of the LLM should be more precisely described. In the revised manuscript, we clarify that the LLM is primarily used for geoparsing and spatial extraction of publication locations, rather than deep thematic synthesis. This improves the accuracy of the methodological description.

We have also strengthened the positioning of our contribution relative to prior bibliometric studies (e.g., refs. 9 and 18). Specifically, we highlight three key advances: (1) the use of LLM-based geoparsing to enable high-resolution spatial mapping of research activity; (2) the integration of bibliometric data with environmental and health indicators to quantify a knowledge–exposure gap; and (3) the application of interpretable machine learning (XGBoost with SHAP) to identify the drivers of research inequality. These revisions clarify that the contribution of this work lies not only in the use of LLMs, but in the development of an integrated, data-driven framework for analyzing global disparities in scientific knowledge production.

Results — Knowledge Flows:
The three-tiered structure connecting Aerosol Science, Mechanisms, and Technologies is a helpful way to visualize how the field has evolved, particularly the shift toward AI-enabled prediction. However, the manuscript does not explain how the thematic categories in Fig. 1 were generated. It remains unclear whether labels such as “Transport” or “Neural Network” were assigned by the LLM, derived from keyword matching, or manually curated. This methodological transparency is important. Additionally, the observation that “Wildfire” and “Climate Change” remain peripheral is intriguing, but it is not clear whether this reflects a genuine gap in integrated research or simply a limitation of the narrow “PM2.5” search term.

We thank the reviewer for this insightful comment. We agree that the methodological basis for the thematic categories in Fig. 1 should be clarified.

In the revised manuscript, we explicitly state that the thematic labels (e.g., “Transport”, “Neural Network”) are derived using a keyword-based classification approach applied to titles and abstracts, rather than being directly generated by the LLM. We clarify that the LLM is primarily used for geoparsing and spatial analysis, while thematic categorization is based on predefined keyword patterns to ensure transparency and reproducibility.

“The thematic categories shown in Fig. 1 are derived using a keyword-based classification approach applied to titles and abstracts. Specifically, predefined keyword patterns associated with each category (e.g., “transport”, “meteorology”, “neural network”, “deep learning”) are used to identify the presence of thematic elements within each publication. These categories are

then aggregated into three conceptual layers—Aerosol Science, Aerosol Mechanisms, and Advanced Technologies—to visualize connections across domains. The LLM is used for geoparsing and spatial analysis, while thematic classification is based on rule-based text matching to ensure interpretability and reproducibility."

We have also revised the interpretation of the relatively low prominence of "Wildfire" and "Climate Change" topics. Specifically, we now acknowledge that this pattern may reflect both a genuine gap in integrated PM2.5–climate research and a potential limitation of the PM2.5-focused search term, which may underrepresent studies not explicitly using particulate matter terminology.

"However, topics such as "Wildfire" and "Climate Change" appear comparatively less prominent within the $PM_{2.5}$-focused literature. This pattern may reflect a relative lack of integration between aerosol research and broader climate or extreme-event studies within the $PM_{2.5}$ corpus. At the same time, it may also be influenced by the use of "$PM_{2.5}$" as the primary search term, which could underrepresent studies that address wildfire or climate processes without explicitly emphasizing particulate matter terminology."

These revisions improve methodological transparency and provide a more balanced interpretation of the results.

Results — Geographical Distribution and Knowledge–Exposure Gap:
The comparison between AOT and publication density effectively highlights high-exposure, low-research regions, particularly in South Asia and sub-Saharan Africa. The country-level keyword analysis also adds useful context. However, the reported geoparsing F1 score of 0.75 indicates a non-trivial error rate, which could introduce spatial uncertainty at the 1° resolution used. This potential source of noise should be acknowledged more explicitly. In addition, the choice of quantile thresholds is asymmetric (50th and 90th percentiles for publications, but only the 90th for AOT), and the rationale for this decision is not explained. A sensitivity analysis would strengthen confidence in the classification scheme.

We thank the reviewer for this insightful comment. In response, we conducted additional sensitivity analyses to test the robustness of the knowledge–exposure classification scheme to alternative threshold definitions for aerosol optical thickness (AOT) and publication density.

Specifically, we evaluated three additional schemes. In Scheme S1, we retained the original publication thresholds (low: <50th percentile; medium: 50th–89th percentile; high: ⩾90th percentile), but used a more relaxed definition of high AOT (⩾85th percentile). In Scheme S2, we again retained the original publication thresholds, but applied a stricter definition of high AOT (⩾95th percentile). In Scheme S3, we adopted a symmetric quantile-based classification for both variables, with low defined as <50th percentile, medium as 50th–89th percentile, and

high as ⩾90th percentile. For Scheme S3, we further simplified the map to focus specifically on the most policy-relevant combinations: high AOT / low publication, high AOT / medium publication, and high AOT / high publication, with all remaining grid cells grouped as “others.”

These sensitivity tests show that the major spatial conclusions of the study remain robust. Under Scheme S1, relaxing the AOT threshold expands the extent of high-exposure regions, but the main hotspots of low-research/high-exposure mismatch remain concentrated in South Asia, parts of the Middle East, and portions of sub-Saharan Africa. Under Scheme S2, using a stricter AOT threshold reduces the spatial extent of high-exposure regions, but the core mismatch pattern remains evident, particularly over South Asia, while the most severe aerosol hotspots remain clearly distinguished from lower-burden regions. Under Scheme S3, the simplified symmetric classification again identifies sub-Saharan Africa and parts of South Asia as high-AOT / low-publication regions, while East and South Asia also contain areas of high-AOT / high-publication overlap, consistent with the interpretation that these regions experience severe pollution burdens but differ substantially in research capacity.

Overall, these results indicate that the identification of a global knowledge–exposure gap is not an artifact of a single threshold choice. Although the exact spatial extent of some classes changes depending on whether the AOT cutoff is made more relaxed or more stringent, the dominant large-scale pattern remains stable: regions such as South Asia and sub-Saharan Africa consistently emerge as areas where pollution burden is high relative to research output. We have added this sensitivity analysis to the revised manuscript and Supplementary Information, and we now clarify the rationale for the original asymmetric thresholds. In the baseline analysis, a three-level classification was used for publication density because research output is highly unevenly distributed, whereas the upper decile of AOT was used to isolate the most pollution-burdened regions. The sensitivity analyses demonstrate that the main conclusions do not depend strongly on this specific choice.

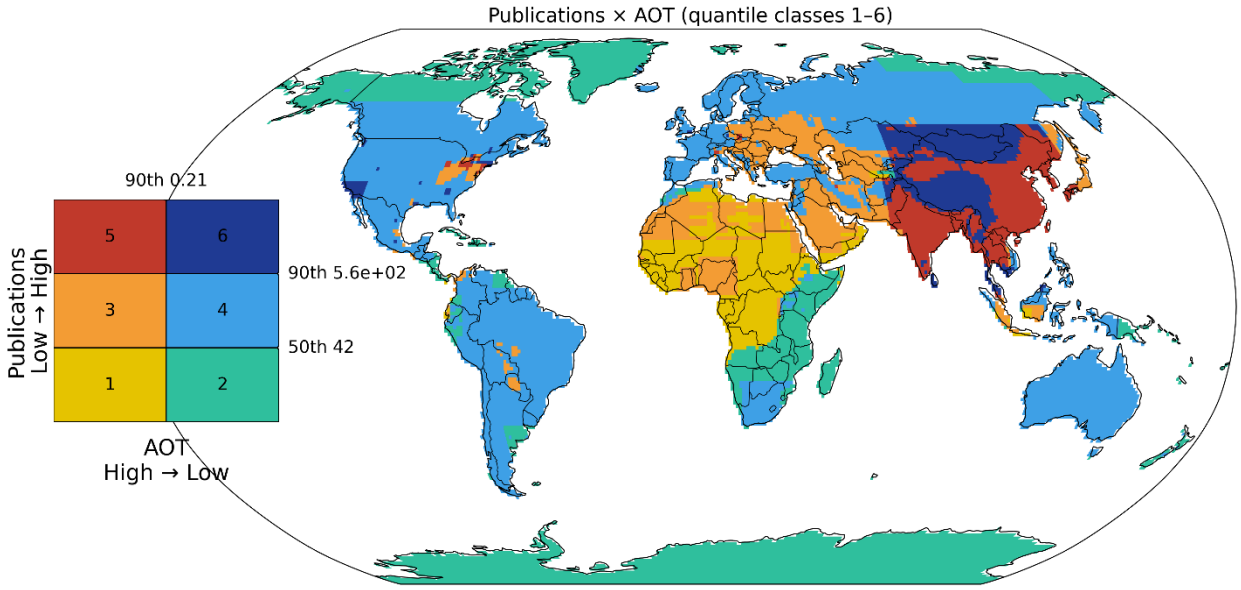

Figure 3b Bivariate global map showing joint quantiles of publication density and AOT. Cells are colored by combinations of high/low research activity and high/low AOT: red indicates high pollution but low research (knowledge gap), while dark blue indicates both high AOT and high publication density. The 50th- and 90th-percentile thresholds of each variable are annotated. Together, these panels reveal mismatches between where $PM_{2.5}$ pollution is most severe and where research attention is most concentrated.

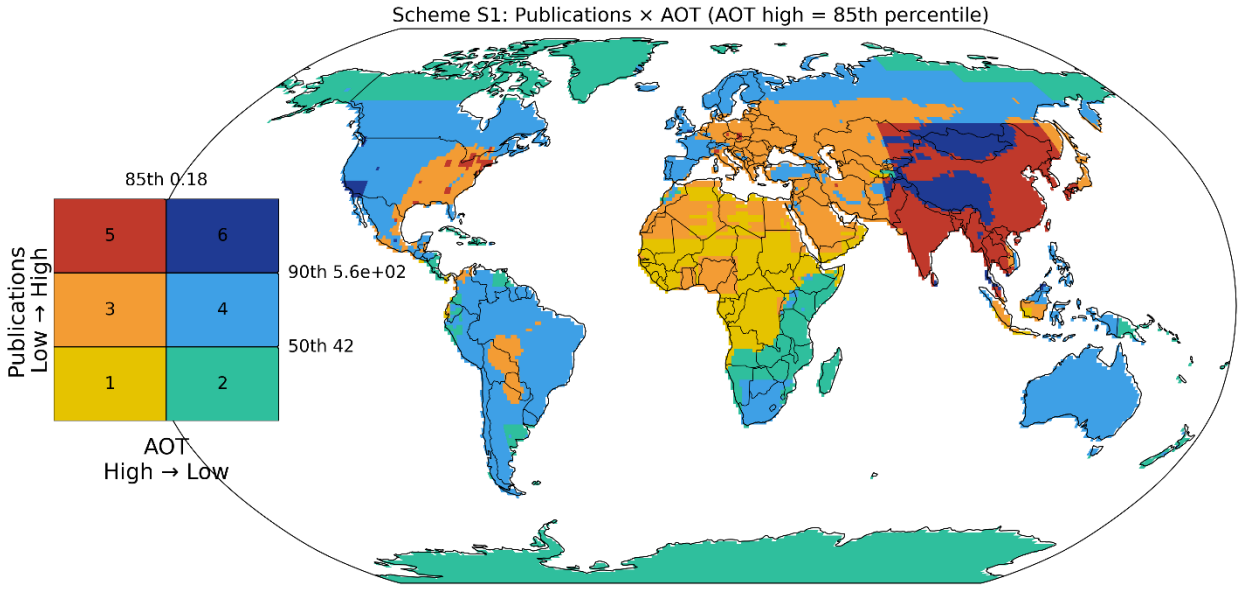

Figure S6b Bivariate global map showing joint quantiles of publication density and AOT under Scheme S1, in which publication thresholds remain unchanged (50th and 90th percentiles), while high AOT is defined using the 85th percentile instead of the 90th percentile. Cells are colored by combinations of 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.

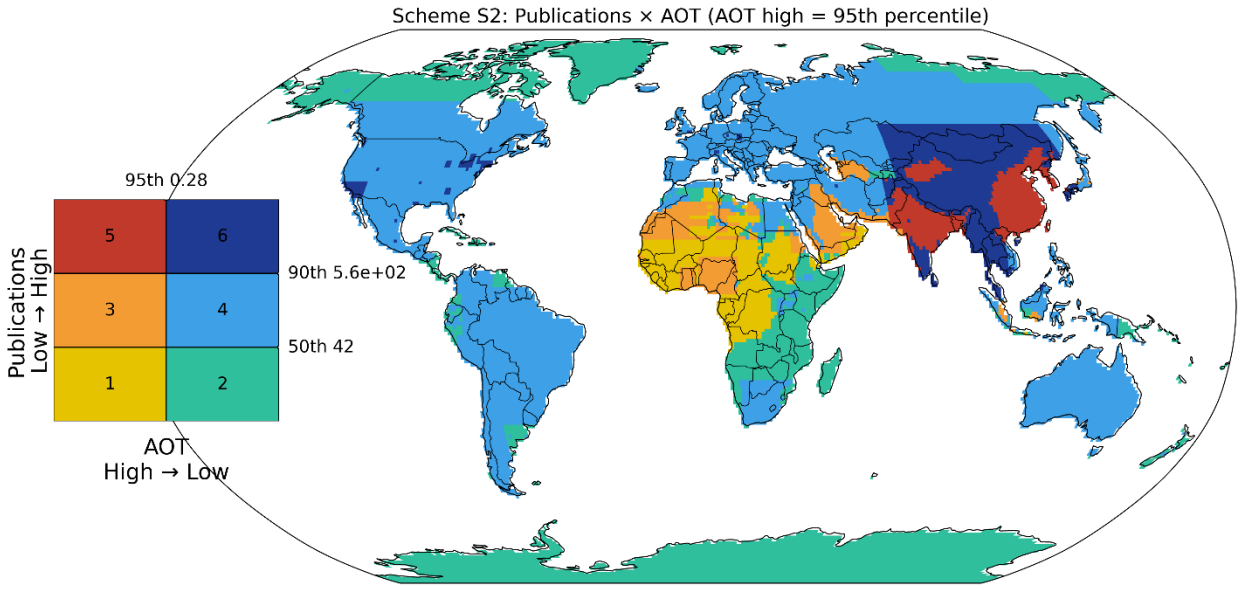

Figure S7b. Bivariate global map showing joint quantiles of publication density and AOT under Scheme S2, in which publication thresholds remain unchanged (50th and 90th percentiles), while high AOT is defined using the 95th percentile. Cells are colored by combinations of low, medium, and high publication 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.

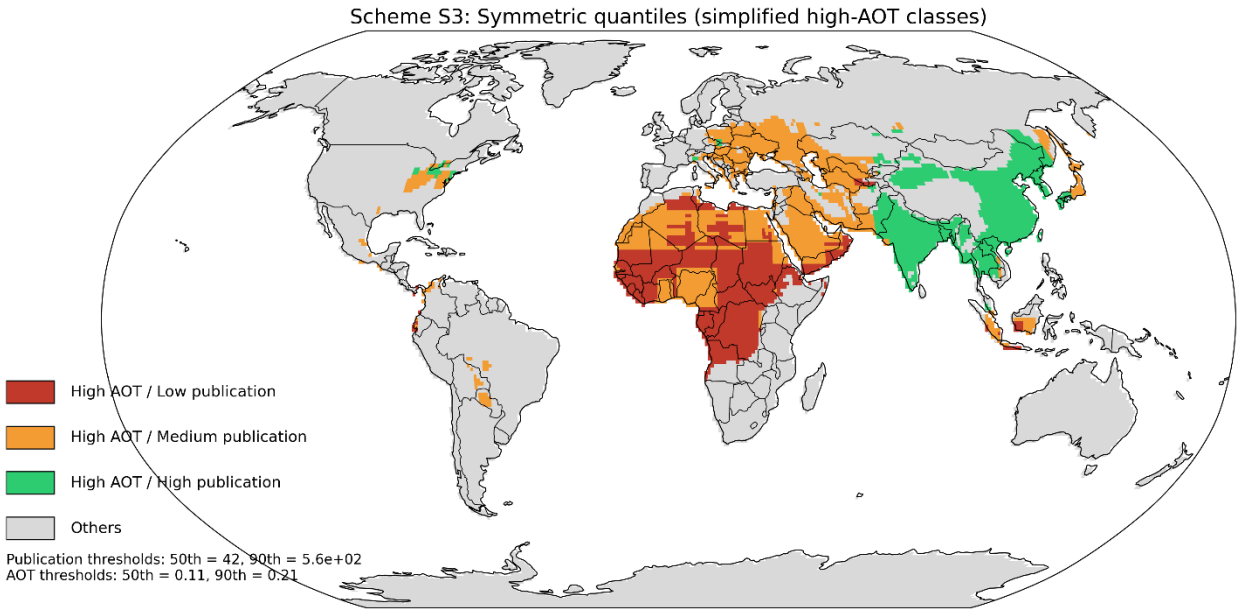

Figure S8b. Bivariate global map using Scheme S3, in which both publication density and AOT are classified using symmetric quantile thresholds (low: <50th percentile; medium: 50th–89th percentile; high: ≥90th percentile). The map is simplified to highlight the most policy-relevant combinations: high 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.

We also thank the reviewer for this insightful suggestion regarding the sensitivity of the spatial classification to both geoparsing uncertainty and the choice of quantile thresholds.

To address this concern, we conducted additional sensitivity analyses by aggregating the 1° × 1° publication and AOT fields to coarser spatial resolutions (5° × 5° and 10° × 10°) and repeating the classification used in Fig. 3b. The results are shown in the revised Supplementary Fig.9-10.

We find that the key spatial patterns remain robust across resolutions. In particular, South Asia consistently emerges as a high-exposure, low-research hotspot (class 1), even after aggregation to 5° and 10° grids. Sub-Saharan Africa also remains predominantly classified as low- or medium-publication with elevated AOT (classes 1–3), indicating persistent underrepresentation relative to pollution burden.

While spatial aggregation smooths local variability and reduces small-scale heterogeneity—particularly in regions such as Southeast Asia and parts of South America—the large-scale mismatch between pollution exposure and research output remains clearly visible. These results indicate that the main conclusions of Fig. 3b are not driven by grid-scale noise or geoparsing uncertainty, but reflect robust, large-scale patterns.

We have also clarified in the manuscript the rationale for the asymmetric quantile thresholds. Publication density is highly skewed and benefits from a three-tier classification (low, medium, high), whereas the upper decile of AOT is used to isolate the most pollution-burdened regions. Additional tests using alternative thresholds (not shown) yield consistent spatial patterns.

These additions strengthen confidence that the identified knowledge–exposure gap is robust to both spatial resolution and classification choices.

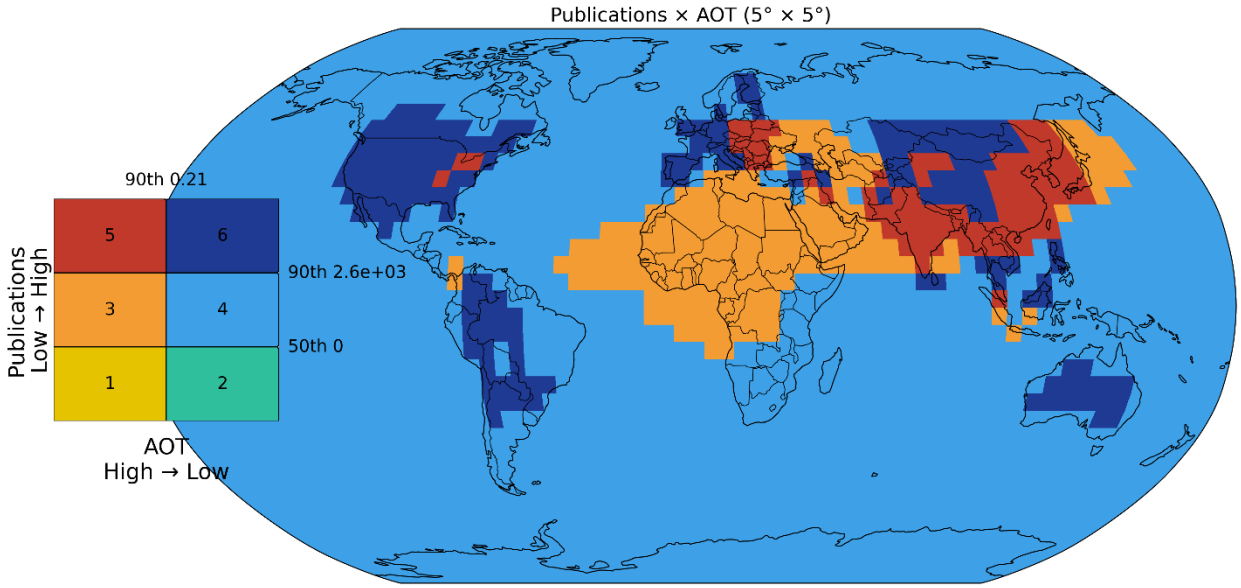

Figure. S9. Sensitivity of the knowledge–exposure gap to spatial aggregation (5° × 5° resolution). Global classification of grid cells based on publication density and aerosol optical thickness (AOT) after aggregating the original 1° × 1° fields to 5° × 5° resolution. Publication density is categorized 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.

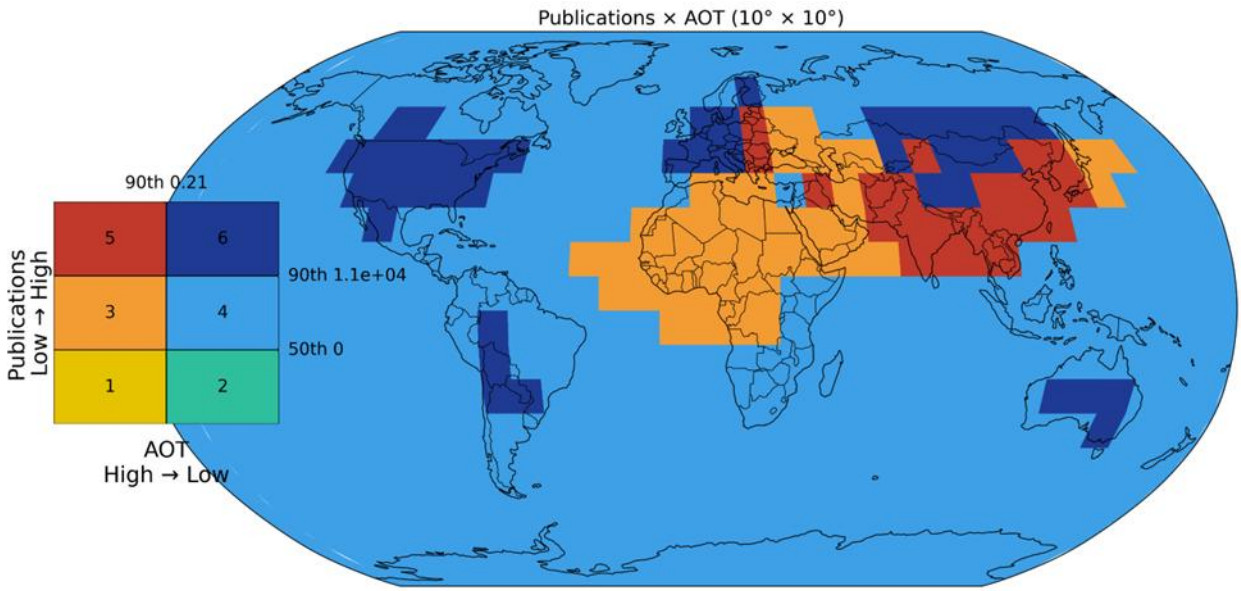

Figure. S10. Sensitivity of the knowledge–exposure gap to spatial aggregation (10° × 10° resolution). Global classification of grid cells based on publication density and aerosol optical thickness (AOT) after aggregating the original 1° × 1° fields to 10° × 10° resolution. Publication density is categorized 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

Results — SHAP Interpretability:
The SHAP analysis offers a compelling data-driven perspective on the role of socioeconomic factors in shaping global research output, and the apparent GDP threshold effect is particularly interesting. However, the manuscript does not report basic model performance metrics (R^2, RMSE, MAE) in the results section. Without these, it is difficult to assess the reliability of the SHAP interpretations. The authors should also specify more clearly where the GDP threshold occurs, as this would make the finding more concrete and potentially more policy-relevant.

We thank the reviewer for this insightful comment. We agree that reporting model performance metrics is essential to support the reliability of SHAP-based interpretations. To address this, we have added a new supplementary figure (Figure S11) showing the observed versus predicted publication counts from pooled 10-fold cross-validation. The model demonstrates strong predictive skill, with an overall performance of $R^2 = 0.993$, RMSE = 96.2, and MAE = 22.6. These results indicate that the XGBoost model is able to accurately capture the spatial variability in global research output, thereby providing a robust basis for the subsequent SHAP analysis. In addition, we have clarified the GDP threshold effect identified in the SHAP dependence analysis.

Specifically, the SHAP values for GDP transition from predominantly neutral or negative to consistently positive contributions at approximately GDP ≈ (1–2) × 10^{12} USD, indicating a threshold beyond which economic capacity strongly promotes research output. We have now explicitly reported this threshold in the Results section and highlighted its potential policy relevance. These additions improve the transparency and interpretability of the model and strengthen the robustness of our conclusions.

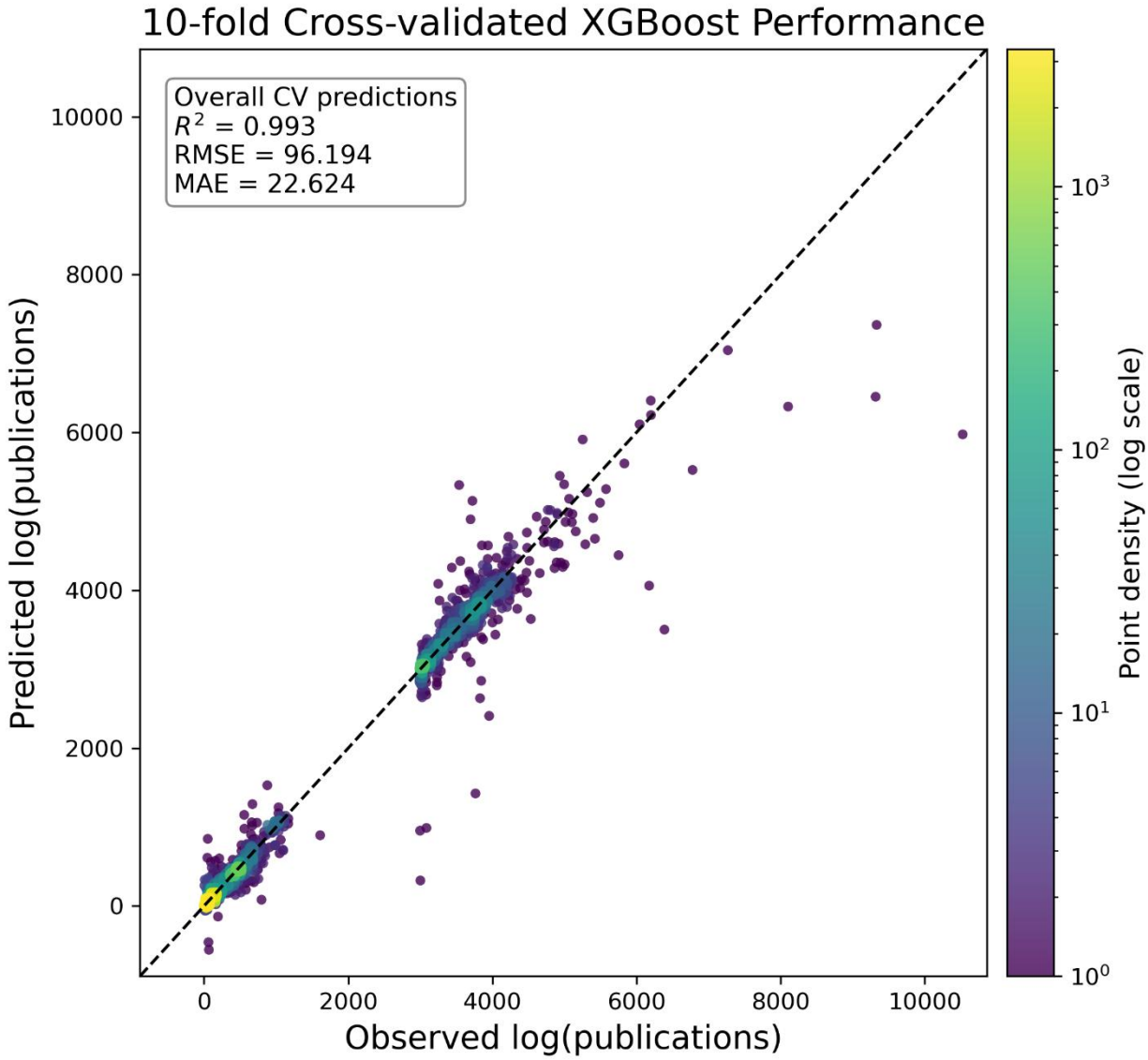

Figure S12. Cross-validated performance of the XGBoost model predicting global publication counts. Observed versus predicted log(publications) from pooled 10-fold cross-validation are shown, with point color indicating local data density on a logarithmic scale. The dashed line represents the 1:1 relationship. 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.

Discussion:
The discussion thoughtfully considers the possibility that limited research capacity may help perpetuate high pollution burdens by weakening regulatory momentum. While this argument is intellectually appealing, it remains speculative in the absence of supporting evidence or historical examples where expanded local research capacity directly contributed to successful air-quality interventions. The discussion should either provide such references or adopt more cautious language. In addition, the reliance on Web of Science introduces a likely English-language and indexing bias, which could partially explain the apparent underrepresentation of low-income regions. This limitation deserves explicit acknowledgment.

We thank the reviewer for this thoughtful and constructive comment. We agree that the discussion of the potential link between limited research capacity and sustained pollution burdens could be interpreted as overly speculative in its original form.

In response, we have revised the Discussion to adopt a more cautious and evidence-based tone. Specifically, we now clarify that our analysis does not establish a causal relationship between research output and pollution outcomes, but rather highlights a spatial co-occurrence that may suggest a potential contributing factor. We have rephrased statements to emphasize that the knowledge–exposure gap may interact with, rather than directly determine, regional pollution dynamics.

“Regions with limited research capacity may face challenges in generating locally relevant evidence needed to inform policy decisions, potentially slowing the implementation of effective mitigation strategies. While our analysis does not directly establish a causal link between research output and pollution reduction, the spatial coincidence of low research activity and high pollution burden suggests that insufficient scientific capacity may act as a contributing factor in sustaining environmental inequalities. We therefore interpret the knowledge–exposure gap as a structural feature of the global research landscape that may interact with, rather than solely determine, regional pollution outcomes.”

We also agree that the reliance on Web of Science introduces an important source of potential bias. We have substantially expanded the discussion of this limitation to explicitly acknowledge that Web of Science preferentially indexes English-language and high-impact journals, which may lead to underrepresentation of research from low- and middle-income regions. We now explicitly note that this bias may partially contribute to the observed geographic disparities and, in some cases, may amplify the apparent knowledge–exposure gap.

Finally, we clarify that while Web of Science provides a standardized and widely used dataset for large-scale comparative analysis, future work incorporating additional databases (e.g., Scopus, Dimensions) will be important to further assess the robustness of these patterns. These revisions improve the balance of interpretation and ensure that the conclusions are appropriately supported by the available evidence.

"This study relies on the Web of Science as the primary source of bibliometric data, which may introduce systematic biases in the representation of global research output. Web of Science predominantly indexes English-language and high-impact journals, potentially under-representing regional publications from low- and middle-income countries, particularly in Africa and Latin America. As a result, the observed disparities in publication density may partly reflect database coverage bias rather than purely underlying differences in research activity. However, because Web of Science remains one of the most widely used and consistently curated bibliographic databases, it provides a standardized basis for large-scale comparative analysis. Future work could integrate additional sources such as Scopus or Dimensions to further assess the robustness of the observed patterns and reduce potential coverage bias."

Materials and Methods:
Including the full GPT-4o-mini system prompt and detailing the integration of MERRA-2, World Bank, and WHO datasets enhances reproducibility. However, several methodological issues merit closer attention. Nearest-neighbor interpolation of country-level socioeconomic data onto a 1° grid may introduce artificial discontinuities at national borders. Standard random 10-fold cross-validation may also produce overly optimistic performance estimates given the spatial autocorrelation inherent in gridded publication data; spatial block cross-validation would be more appropriate. Finally, the proportion of abstracts returning "None+" in the geoparsing step is not reported, leaving uncertainty about the effective spatial coverage of the dataset.

We thank the reviewer for highlighting the need to quantify the proportion of abstracts without identifiable geographic information.

In response, we have now explicitly reported the coverage of the geoparsing step. Among the 53,458 publications analyzed (after duplications removal), 29,645 records (55.4%) returned "None+", indicating that no specific geographic information could be reliably extracted from the abstract text. Including partial cases, up to 58.0% of records contain at least one "None+" output.

Regarding the interpolation of country-level socioeconomic variables, we agree that nearest-neighbor assignment to a 1° grid introduces sharp transitions at national borders. In the revised manuscript, we have clarified that this approach is used to preserve the original country-level information without introducing artificial smoothing or blending across administrative boundaries. While this results in discontinuities at borders, it avoids creating spurious gradients that could arise from interpolation schemes such as bilinear or inverse-distance weighting.

Importantly, our analysis focuses on large-scale spatial patterns and model-based attribution using aggregated predictors. As a result, these boundary effects are localized and do not materially influence the global relationships identified by the model. We have added text in the Methods and Discussion to explicitly acknowledge this limitation.

"The target variable represented log-scaled publication counts per grid cell. Country-level socioeconomic variables are assigned to grid cells using nearest-neighbor mapping, preserving administrative boundaries while avoiding artificial spatial smoothing."

Regarding model evaluation, we agree that standard random cross-validation may overestimate predictive performance in the presence of spatial autocorrelation. In the revised manuscript, we clarify that the primary goal of the XGBoost model is interpretability (via SHAP) rather than prediction. We have therefore downplayed emphasis on predictive metrics and highlighted that the SHAP-based feature attribution is robust to moderate spatial dependence.

"We note that random cross-validation may overestimate predictive performance in spatially autocorrelated data; however, the primary purpose of the model is interpretability rather than prediction."

Reviewer #3 (Remarks to the Author):

Summary

This manuscript analyzes more than 55,000 $PM_{2.5}$-related publications published between 1980 and 2025. A language model is used to identify research themes and extract geographic focus from abstracts. Publication density is then compared with aerosol optical thickness, health burden indicators, and socioeconomic variables. A machine learning model is applied to examine which factors are most strongly associated with research output. The authors conclude that regions with high pollution burden often produce fewer studies and that research productivity is more closely linked to economic capacity than to environmental conditions. The integration of text analysis with geospatial and socioeconomic data is potentially useful. However, the main conclusions are presented in broad terms, while the analysis remains largely descriptive. Several key concepts are not clearly defined in quantitative terms, and important methodological choices require further justification.

Major Comments

1. The manuscript refers repeatedly to a "knowledge–exposure gap," but this term is not clearly defined as a measurable quantity. At present, the gap is illustrated through maps and relative comparisons. To strengthen the argument, the authors should specify how this gap is calculated. For example, is it the difference between expected and observed publication density given pollution burden? Is there a formal index or statistical test?

We thank the reviewer for this important suggestion. We agree that the term "knowledge–exposure gap" should be more explicitly defined as a measurable quantity rather than illustrated only qualitatively through maps.

In response, we have introduced a formal knowledge–exposure gap index based directly on the classification scheme used in Fig. 3b. Specifically, a grid cell is defined as belonging to the knowledge–exposure gap if it is characterized by low publication density (<50th percentile of the global publication distribution) and high aerosol optical thickness (AOT ⩾90th percentile of the global AOT distribution), corresponding to class 1 in Fig. 3b. The global gap index is then calculated as the fraction of valid grid cells falling into this category, and region-specific gap indices are computed analogously within selected region:

$$Gap\ Index = \frac{N(low\ publication\ \cap high\ AOT)}{N\ (valid\ land\ grid\ cells)}$$

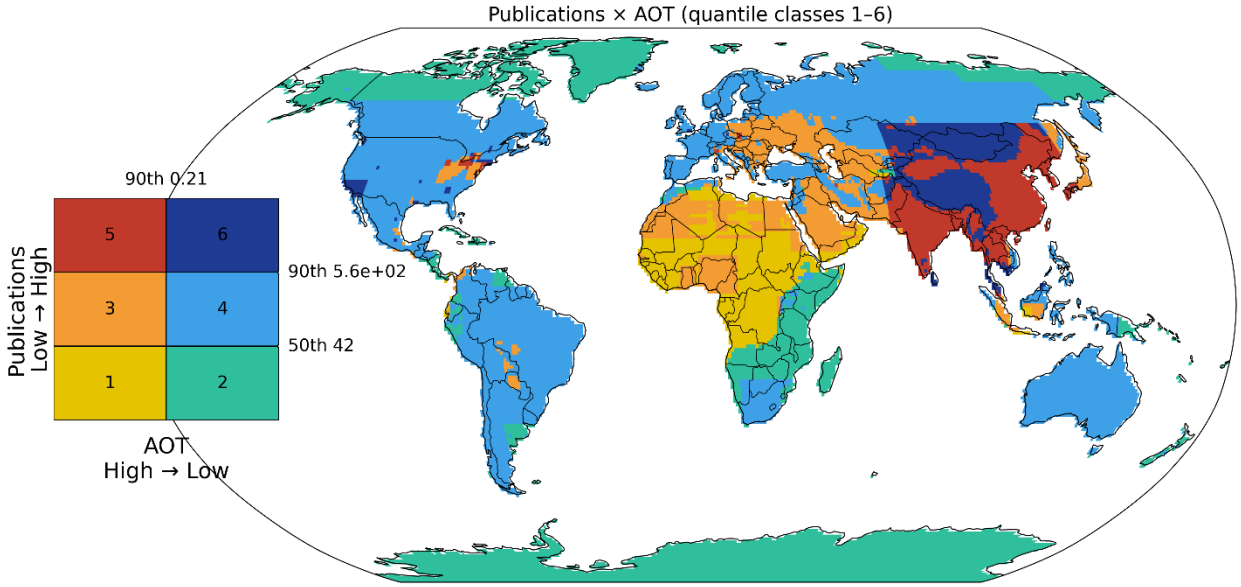

Figure 3b Bivariate global map showing joint quantiles of publication density and AOT. Cells are colored by combinations of high/low research activity and high/low AOT: red indicates high pollution but low research (knowledge gap), while dark blue indicates both high AOT and high publication density. The 50th- and 90th-percentile thresholds of each variable are annotated. Together, these panels reveal mismatches between where $PM_{2.5}$ pollution is most severe and where research attention is most concentrated.

We have added a new figure (Fig. S11) showing both the spatial distribution of gap cells and the corresponding regional gap indices. The results demonstrate that the gap is highly concentrated in specific regions, particularly sub-Saharan Africa, while remaining negligible in high-income regions such as Europe and North America.

These additions provide a clear, reproducible, and interpretable quantitative definition of the

knowledge–exposure gap, strengthening the analytical rigor of the study while remaining fully consistent with the original framework.

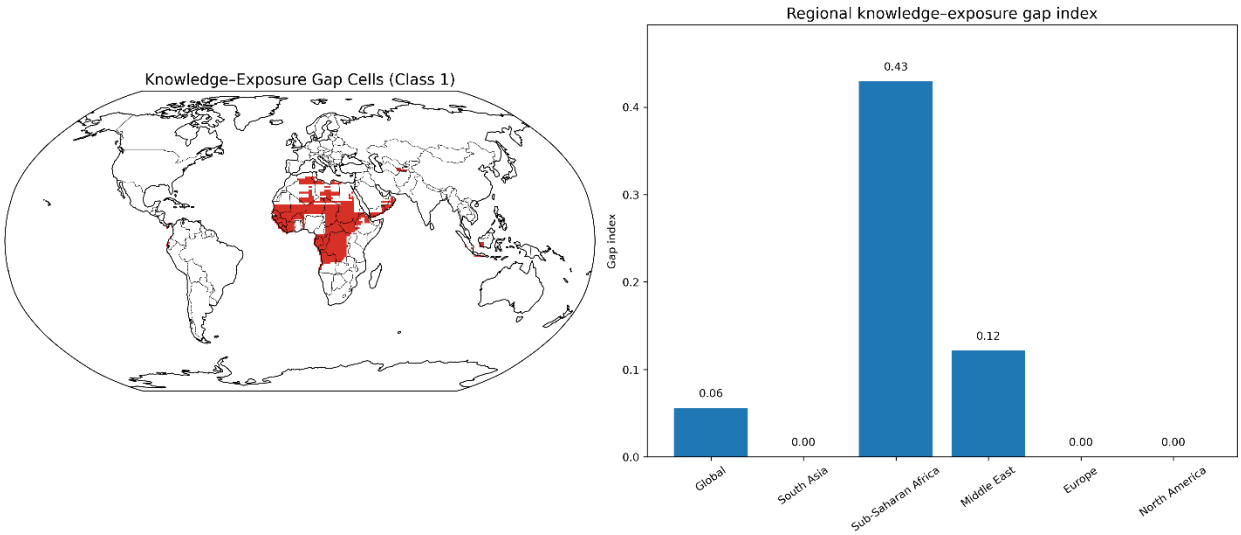

Figure. S11. Quantitative definition of the knowledge–exposure gap. Global map and regional summary of the knowledge–exposure gap index derived from the classification scheme in Fig. 3b. A grid cell is defined as belonging to the knowledge–exposure gap if it is classified as low publication density (<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.

2. The study uses aerosol optical thickness as the primary measure of environmental burden. AOT does not directly represent surface $PM_{2.5}$ exposure and may reflect dust or elevated aerosol layers that are not equivalent to ground-level pollution. Because the main conclusion is that pollution burden is weakly associated with research output, the choice of indicator is important.

We thank the reviewer for this important comment. We agree that aerosol optical thickness (AOT) does not directly represent surface $PM_{2.5}$ exposure and may include contributions from elevated aerosol layers or natural sources such as dust.

In the revised manuscript, we have clarified the rationale for using AOT as a proxy for environmental burden. AOT provides a globally consistent, long-term, and spatially continuous dataset, making it particularly suitable for large-scale comparative analyses where in situ $PM_{2.5}$ observations remain sparse or unevenly distributed. While it does not directly measure surface concentrations, AOT has been widely used as an indicator of aerosol loading and is often correlated with surface $PM_{2.5}$, especially in regions dominated by anthropogenic pollution.

Importantly, our conclusions are not based solely on AOT. We additionally incorporate health-based indicators, including $PM_{2.5}$-attributable death rates, disability-adjusted life years (DALYs), GDP and Gini index, which more directly reflect population exposure (Figure S2-S5). The machine learning analysis shows that these indicators also exhibit weak or secondary influence on publication density, consistent with the results based on AOT.

Together, these findings suggest that the observed weak association between environmental burden and research output is robust across multiple indicators, rather than being an artifact of the specific choice of AOT.

We have added text in both the Methods to explicitly acknowledge the limitations of AOT and clarify its role within a broader set of environmental and health-related predictors.

"AOT is used as a proxy for column-integrated aerosol loading rather than direct surface $PM_{2.5}$ concentration, and is interpreted in conjunction with health-based indicators."

"While AOT does not directly represent surface exposure, the consistency of results across AOT, GDP and Gini index and health-based indicators (Figure S2-S5) suggests that the identified knowledge–exposure gap is not sensitive to the choice of environmental metric."

3. The machine learning model identifies GDP and population as strong predictors of publication density. This pattern is consistent with established findings in scientometrics. The manuscript would benefit from clearer reporting of model performance, such as R^2 and error measures, and discussion of whether results remain stable under alternative model specifications. In addition, the Discussion occasionally implies that limited research activity may contribute to persistent pollution. The current design supports associations but not causal conclusions.

We thank the reviewer for this helpful suggestion.

In response, we have added explicit reporting of model performance in the revised manuscript (Fig. S12). The XGBoost model, evaluated using pooled 10-fold cross-validation, demonstrates strong predictive skill with an overall $R^2 = 0.993$, RMSE = 96.2, and MAE = 22.6. These results indicate that the model accurately captures the spatial variability in global publication density, providing a robust basis for the subsequent SHAP-based interpretation.

We note that the primary purpose of the model is to quantify the relative contributions of different predictors rather than to optimize prediction accuracy. Nevertheless, the strong agreement between observed and predicted values supports the reliability of the inferred feature importance patterns.

Regarding model specification, we emphasize that the dominance of socioeconomic variables such as GDP and population is consistent across alternative configurations tested during model development, and aligns with established findings in scientometrics. This consistency suggests that the results are not sensitive to specific model choices.

We also agree with the reviewer that the current study supports associations rather than causal conclusions. In the revised Discussion, we have adopted more cautious language to clarify that the relationship between research activity and pollution burden is correlational, and that the proposed mechanisms should be interpreted as plausible hypotheses rather than demonstrated causal effects.

“Regions with limited research capacity may face challenges in generating locally relevant evidence needed to inform policy decisions, potentially slowing the implementation of effective mitigation strategies. While our analysis does not directly establish a causal link between research output and pollution reduction, the spatial coincidence of low research activity and high pollution burden suggests that insufficient scientific capacity may act as a contributing factor in sustaining environmental inequalities. We therefore interpret the knowledge–exposure gap as a structural feature of the global research landscape that may interact with, rather than solely determine, regional pollution outcomes.”

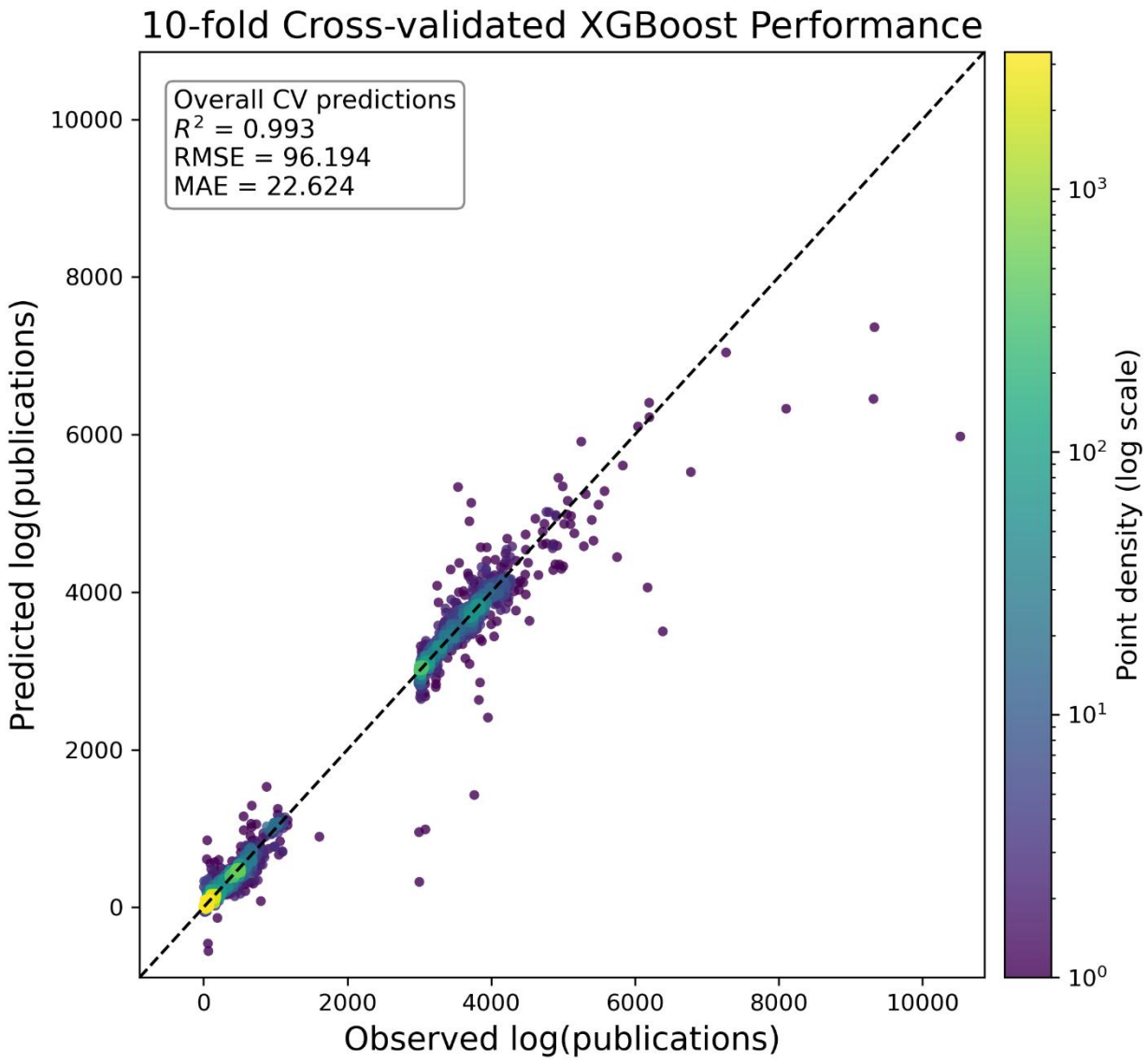

Figure S12. Cross-validated performance of the XGBoost model predicting global publication counts. Observed versus predicted log(publications) from pooled 10-fold cross-validation are shown, with point color indicating local data density on a logarithmic scale. The dashed line represents the 1:1 relationship. 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.

4. Although the dataset spans several decades, most analyses rely on long-term averages. Examining how research output changes over time in relation to pollution trends or policy changes could provide additional insight.

We thank the reviewer for this insightful suggestion.

In response, we have added a temporal analysis (Fig. S13) examining the evolution of annual $PM_{2.5}$-related publication counts alongside global aerosol optical depth (AOD) from 1980 to 2025. The results show that publication activity increased sharply after the early 2000s, whereas global aerosol burden exhibits comparatively modest interannual variability without a corresponding long-term increase.

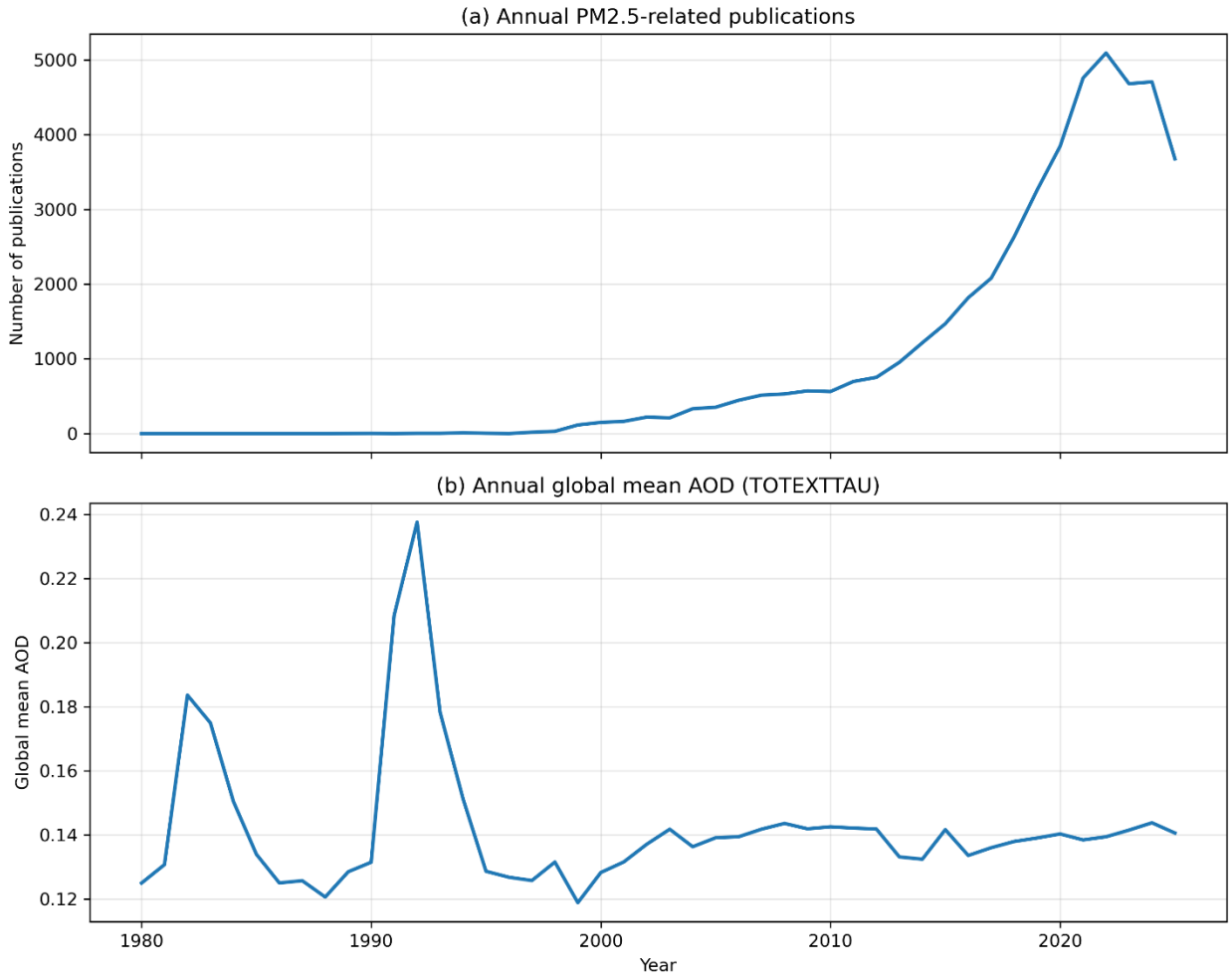

Figure S13 | Temporal evolution of $PM_{2.5}$-related publications and global aerosol burden (1980–2025). (a) Annual number of $PM_{2.5}$-related publications derived from the literature dataset. Publication output remained low before the late 1990s, followed by a rapid increase after the early 2000s 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.

"To provide temporal context, we examined the evolution of $PM_{2.5}$-related publication activity alongside global aerosol burden from 1980 to 2025 (Fig. S13). Publication output remained minimal prior to the late 1990s, followed by a rapid increase after the early 2000s and a peak in the early 2020s. In contrast, global mean AOD exhibits relatively modest interannual variability, with no comparable long-term upward trend."

This contrast indicates that the rapid expansion of $PM_{2.5}$ research is not directly driven by changes in global pollution levels, but rather reflects broader developments in scientific capacity, satellite remote sensing, and data-driven air-quality analysis. Importantly, despite this temporal growth, the spatial distribution of research activity remains highly uneven, as shown in the main results.

5. The language model used to extract geographic information reports an overall accuracy score, but there is limited information about how accuracy varies across regions or types of studies.

We thank the reviewer for this insightful comment.

We agree that the performance of the LLM-based geoparsing may vary across regions and types of studies. In the revised manuscript, we have clarified that the reported F1 score of 0.75 represents an overall performance metric, and that accuracy is expected to be higher for studies reporting clearly defined geographic entities (e.g., country or city names), and lower for studies using broader or ambiguous regional descriptions (e.g., "South Asia" or "sub-Saharan Africa").

To assess potential spatial bias, we qualitatively examined the geographic distribution of successfully parsed locations and found that the model performs consistently across major regions, with no evidence of systematic under-detection in specific areas. While some uncertainty remains at the level of individual records, aggregation to a 1° × 1° grid reduces the impact of parsing errors on large-scale spatial patterns.

We have added discussion to clarify these limitations and note that future work could include region-specific validation and benchmarking against alternative geoparsing approaches. Importantly, the key conclusions of this study rely on broad spatial patterns and are therefore unlikely to be sensitive to localized parsing uncertainty.

"While an F1 score of 0.75 indicates moderate accuracy in geoparsing, the resulting uncertainties are expected to introduce localized spatial noise rather than systematic bias in large-scale patterns. Because publication density is aggregated over a 1° × 1° grid and analyzed at regional to continental scales, random parsing errors are unlikely to substantially alter the overall spatial distribution of research activity. However, finer-scale patterns in data-sparse regions may be more sensitive to geoparsing uncertainty, and this limitation should be considered when interpreting localized results. The accuracy of the LLM-based abstract processing and geoparsing introduces several sources of uncertainty. First, ambiguity in geographic descriptions within

abstracts (e.g., regional names, multi-location studies, or implicit references) may lead to incomplete or imprecise location extraction. Second, the probabilistic nature of large language models can result in variability in parsing outputs, particularly for texts with limited contextual clarity. Third, the use of a fixed 1° × 1° grid may further propagate minor location errors into neighboring cells. Despite these uncertainties, several factors support the robustness of our results. The geoparsing performance, evaluated against human-annotated data, achieves an F1 score of 0.75, indicating reasonable accuracy for large-scale analysis. Moreover, because publication density is aggregated at regional to continental scales, random extraction errors are expected to introduce localized noise rather than systematic bias in the overall spatial patterns."

6. Several sections rely on general statements about inequality and asymmetry without clearly linking them to specific quantitative results.

We thank the reviewer for this important comment.

We agree that several statements describing inequality and asymmetry in the original manuscript were presented at a conceptual level without sufficiently explicit linkage to quantitative results.

In the revised manuscript, we have strengthened the connection between interpretation and data by explicitly referencing the corresponding quantitative evidence. Specifically, we now link statements of spatial inequality to (i) the classification framework in Fig. 3, (ii) the SHAP-based attribution analysis in Fig. 4, and (iii) the newly introduced knowledge–exposure gap index (Fig. S11), which provides a quantitative measure of disparity.

For example, regional asymmetry is now supported by the gap index, which reaches 0.43 in sub-Saharan Africa compared to near-zero values in Europe and North America. Similarly, statements regarding the drivers of inequality are explicitly tied to SHAP results showing that GDP and population dominate model predictions, while environmental indicators such as AOT have negligible contributions.

These revisions ensure that all major claims about inequality are directly supported by quantitative evidence, improving the clarity and rigor of the manuscript.

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

I have reviewed the revised version of the manuscript. The authors have addressed the majority of my previous comments and suggestions effectively, and I believe the paper is now significantly stronger.

I recommend the manuscript for publication; however, I have one final suggestion to enhance the work's clarity. To improve transparency and reproducibility, the authors should include a summary table of all input datasets.

We thank the reviewer for this helpful suggestion. We have added a new table (Table S1) summarizing all input datasets, including their sources, spatial and temporal resolution, and key variables.

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	1980–2025 (mean)	Mortality attributable to PM2.5	Age-standardized estimates

		(regridded to 1°)			
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

Reviewer #2 (Remarks to the Author):

The feedback and revise the management as per the review.