Large language models reveal big disparities in current wildfire research

Check for updates

Zhengyang Lin [1,4], Anping Chen [2,4], Xuhui Wang [1] ✉, Zhihua Liu[3] & Shilong Piao [1]

Contemporary fire-human-climate nexus has led to a surge in publication numbers across diverse research disciplines beyond the capability of experts from a single discipline. Here, we employed a generalized large language model to capture the dynamics of wildfire research published between 1980 and 2022. More than 60,000 peer-reviewed papers were scanned and analyzed. Through integrating geographical metadata extracted by the artificial intelligence and satellite wildfire datasets, we found large disparities in geographic patterns and research themes. The hottest spot of wildfire research is western United States, accounting for 15% of publications but only 0.5% of global burnt area, while the world's most widely burnt region, like Siberia and Africa are largely underrepresented by contemporary publications. Similar discrepancies are found between the fuel of wildfire and its ignition and climatic drivers, between socioeconomic development and wildfire mitigation, raising concerns on sustainable wildfire managements and calling for further artificial intelligence-aided transdisciplinary collaborations.

The escalating severity and frequency of wildfires and the risks they pose to ecosystems and society have stimulated a rising interest in wildfire research. Indeed, a search of the Web of Science (WOS) bibliographic database using wildfire-related keywords returned over 100,000 papers (see Methods). More notably, the number of publications per year has surged by more than fourfold in the last two decades (Fig. 1). This rapid growth in wildfire research has been contributed by an interdisciplinary community. However, communications across disciplines remain limited, raising the question of how research on this topic is distributed across various aspects of wildfire research and geographical regions. Addressing this issue is crucial for identifying research gaps and informing future research priorities.

On the other hand, the growing volume of literature also poses grand challenges for literature reanalysis that aims to recap knowledge consensus and identify research gaps. Traditional expert-based methods, such as literature reviews and meta-analyses[1–3], are often constrained to a limited number of publications, typically a few hundred. More recently, artificial intelligence (AI)-based approaches, including deep-learning methods, develop the capacity for analyzing large volumes of literature, offering a solution to bridge the gap between the exponentially increasing number of publications and the ability to effectively synthesize them[4]. In particular, the rapid growth in the size of Large Language Models (LLMs)[5–7] have enabled end-users to perform a variety of complex tasks beyond semantic analyses[8],

making it a potentially powerful tool for overcoming the limitations of traditional literature analyses. These LLMs, represented by ChatGPT, are known for substantially large corpus in their training stages and capable for general tasks, providing the confidence in completing tasks such as textual geographical entity recognition. Furthermore, these models have demonstrated unprecedented "emergent abilities", even reaching human-level performance in various complex tasks[9–11].

Here, we designed a literature analysis workflow (Supplementary Fig. 1) to disentangle thematic and regional patterns in recent wildfire research with gpt-3.5-turbo (updated on 13th Feb 2023), the same LLM model underlying the public version of ChatGPT. The model has already been proved to function effectively even without fine-tune, making it more convenient for the initial information extraction and classification processes. A recent study also showed comparable or even better performance in ChatGPT than conventional Natural Language Processing (NLP) method (e.g., BERT) models especially on inference and question-answering tasks[12]. We used the LLM to first categorize the collected publications into different themes related to various aspects of wildfire research, including its causes, consequences, and the methodologies used to detect or simulate wildfire dynamics (see Methods). Additionally, we extracted the geolocation information and converted it from textual descriptions in titles and abstracts into numeric forms with the maximal outside boundary

[1]Institute of Carbon Neutrality, Sino-French Institute for Earth System Science, College of Urban and Environmental Sciences, Peking University, Beijing, China. [2]Department of Biology and Graduate Degree Program in Ecology, Colorado State University, Fort Collins, CO, USA. [3]CAS Key Laboratory of Forest Ecology and Management, Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang, China. [4]These authors contributed equally: Zhengyang Lin, Anping Chen. ✉e-mail: xuhui.wang@pku.edu.cn

coordinates, also known as geoparsing. Ultimately, we were able to include 60,488 relevant articles for subsequent analysis (see Methods for paper inclusion criteria). We focused on addressing the following questions: (1) What prevailing preferences and trends have emerged in wildfire research over recent decades? (2) How do spatio-temporal variations in research paradigms manifest? (3) What are the real-world implications of disparities in research of wildfire components and actual wildfire activities for populations and socioeconomic development?

Results

Among all the keywords associated with wildfire research, "vegetation" emerged as the most frequently discussed, accounting for 47% of the papers. Within the "vegetation" category, forest fires attracted the most attention (72%; Fig. 1). There was also a large volume of wildfire studies involving "anthropogenic," "atmospheric," and "modeling" aspects.

From a temporal evolutionary perspective, the recent rapid increase in the overall number of wildfire publications (Fig. 2a) is comparable to other trending key words such as climate change[4]. Furthermore, we also observed distinct shifts in topics that have experienced the most rapid growth since 1990s (Fig. 2a). For instance, when comparing research preferences in the 2000s to that in the 1990s, we found that discussions on the "hydrological" and "atmospheric" impacts of wildfires were among the fastest-growing topics. Since the turn of the century, the widespread adoption of advanced satellite-based instrumentation and retrieval/analysis algorithms, exemplified by the deployment of the Moderate Resolution Imaging Spectroradiometer (MODIS), has catalyzed increased scholarly interest in "remote sensing" themes[13]. Similar shifts have been observed in the most recent decade, during which the domains of "climate change" and "anthropogenic" influences have gained prominence in wildfire research[4].

Using the LLM to extract and parse geolocation information from the texts (see Methods), we were able to map published papers according to their focused regions (Fig. 2b). An analysis of affiliations of full authors' lists of selected papers uncovered a consistent pattern that countries with higher levels of development dominate the research landscape. Specifically, 87.1% of the total research contributions come from the leading 20 countries by publication count, with more than half of them classified as high-income economies (Supplementary Fig. 2). In addition, 517 (70.4%) of the 732 papers dedicated to wildfire datasets were contributed from high-income countries, while less than 10% originated from lower-middle income countries. Clearly, such publication biases in wildfire research papers and datasets were primarily due to resource inequality (e.g., funding to conduct research, generate and maintain databases) between developed and developing countries. Inadequate funding could largely hamper the investment in
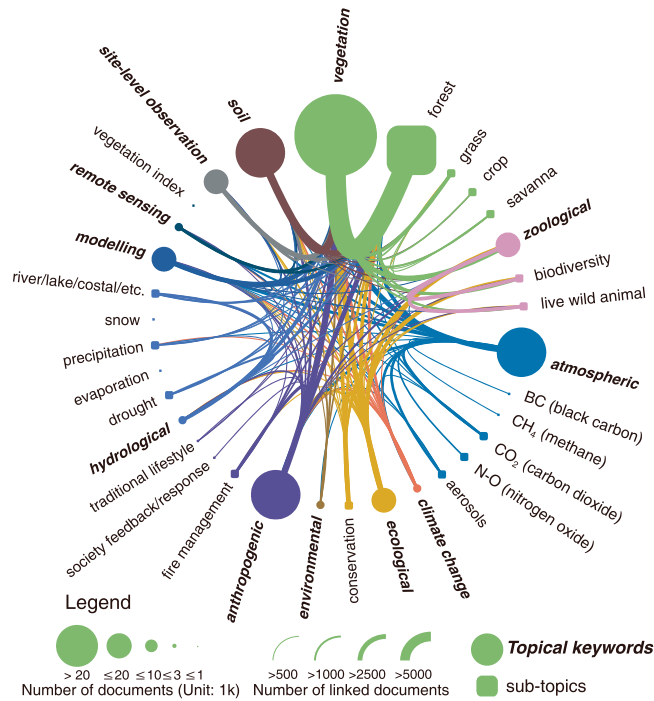

Fig. 1 | An overview of the wildfire research literature body with relationships between the selected topics. The topics are marked by circles and bold-oblique font, sub-topics by rounded rectangles and normal font. Different colors indicate different topics; while the sizes of these markers and the widths of their connecting lines represent the numbers of publications.

Fig. 2 | Spatio-temporal patterns of geoparsed documents. a Growth in the number of wildfire publications. The dotted and solid lines represent the "total wildfire publications" and "geoparsed wildfire publications". Boxes highlight three topics that have the highest increase rate during their past decades. **b** The spatial patterns of total geoparsed publication numbers during 1980 to 2022. The blank pixels represent no observed burned area according to a long-term global burned area product, AVHRR-LTDR (1982-2018)[31].
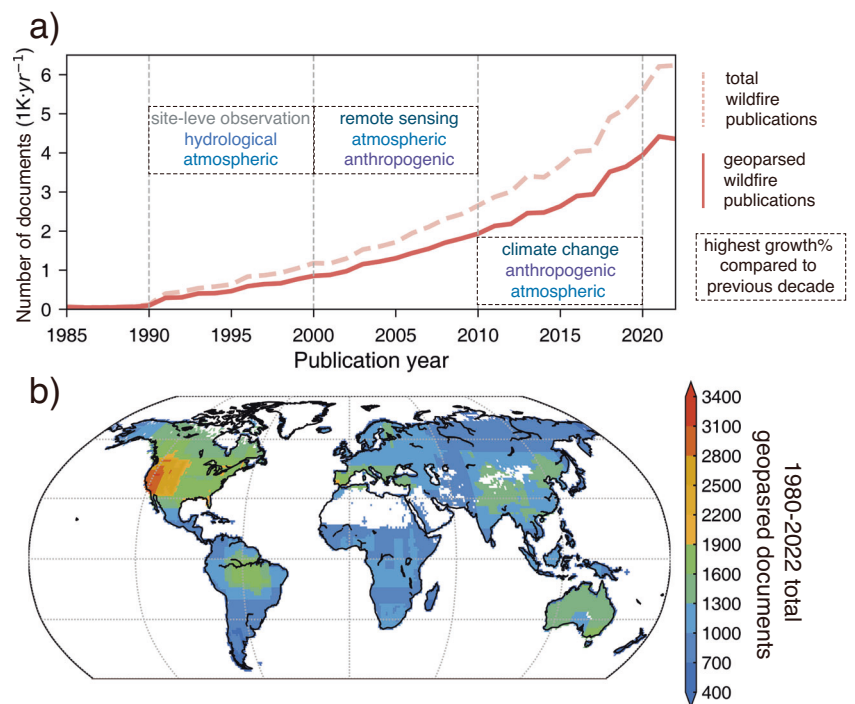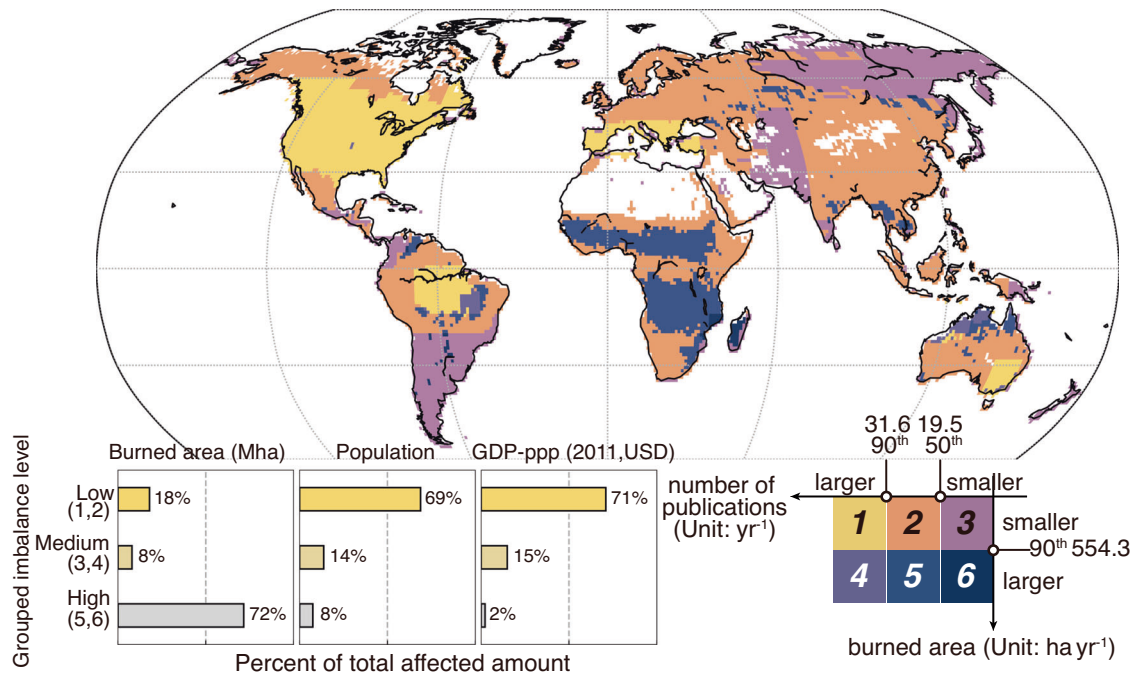
Fig. 3 | Pervasive imbalances measured by the numbers of wildfire-related publications and satellite-based burned areas. We classified the pixels into three major categories with imbalance level scales from 1 to 6. The percentile thresholds for pixels were established at the 50th and 90th percentiles for assessing the relative scale of publications, ensuring inclusion of studies spanning from 1980 to 2022. These thresholds corresponded to multi-year averages of 19.5 and 31.6 papers per year, respectively. Regarding burned area analysis, a 90th percentile threshold (equivalent to 554.3 hectares per year) and a frequency threshold (requiring at least 90% coverage across the entire temporal span of the burned area product) were applied. Blank pixels indicate areas where no burned area was observed within AVHRR-LTDR[31].

wildfire research, monitoring systems, and policy measures for mitigation[14,15].

Furthermore, we also found a notable discrepancy in the attentions of study areas (Fig. 2a). For instance, while the burned area in the western United States accounts for less than 0.5% of the global burned area over the past two decades, this region emerged as a primary global research hotspot for wildfire, accounting for 15.0% of the total published papers. Other well-known wildfire-prone regions, including Canada, the Amazon, Australia, Mediterranean regions, northern India, and northeast/south China, also received substantial attention, each being the subject of more than one thousand papers (Fig. 2a). On the other hand, despite accounting for more than 1.3% of global burned area and 4% of global fire emissions, vast Siberia has been largely underrepresented in the current literature[16]. Using the ratio of publication rates to the observed burned areas or fire emissions as indicators, it is evident that the African continent also emerges as a major understudied region (Fig. 3 and Supplementary Fig. 2).

At the biome level, our analyses reveal a large disparity in research attention and the actual occurrence and impact of wildfires in grasslands and savannas in Africa and northern Australia (Fig. 3). Despite constituting just 8% of the global population, these biomes bear the burden of 72% of the global total burned area. In contrast, 14% and 69% of the global population experience medium (categories 3 and 4) and low (categories 1 and 2) levels of disparity, with only 8% and 18% of the burned area, respectively. Together, areas characterized by a high level of under-representation in wildfire research account for only 2% of the global Gross Domestic Product (GDP, in 2011 purchasing power parity international dollars), a demonstration of low economic capacity for both wildfire management and research. In particular, the high costs of implementing state-of-the-art fire management strategies pose a substantial challenge for these communities with limited economic resources[17]. Furthermore, developing economies in the tropics are also vulnerable to rapid land-use changes[18,19]. These changes can trigger a positive feedback loop involving wildfires, land use, and climate, further increasing the likelihood and severity of wildfires. In addition to burned areas, it is important to note that fire emissions may impact or extend to places beyond local ecosystems via atmospheric transportation (Supplementary Fig. 3). Especially, public health expenditures associated with exposure to open fires impose a considerable burden on regions marked by large disparities in wildfire research[14], often due to inadequate study and research. In our analysis, we found that 39% of the population and 24% of socioeconomic development were exposed to high imbalance level, most of which lack of adequate representation, with 77% of total fire-induced carbon emissions. In contrast, 41% of the population and 48% of the GDP were found in regions with low imbalance levels, with only 5% of total wildfire emissions. It should also be mentioned that our analysis may still fall short in fully assessing complexity and scale of imbalances, given the diverse significance and impacts of wildfires across different regions and ecosystems. For example, a large proportion of burned area in Africa comes from routine seasonal burning in savanna and cropland[20]. On the other hand, the high-latitude forests act as essential carbon sink but are concurrently highly susceptible to severe wildfires, which can lead to profound disruptions in their carbon sequestration and release processes[21].

Building on the conceptual framework of fire research components known as the "fire triangle" model[22], we further examined changes in research attention to different wildfire topics since 1980s (Fig. 4). Overall, "vegetation" attracted the most attention in terms of the amount of literature, not only due to its direct ecological and socioeconomic impacts caused by vegetation fires but also its critical role as "fuel" in the "fire triangle." With rapidly changing fire weather driven by climate change, the aspect of "climate" in fire research has gradually become a focus of recent wildfire-related research. Regarding ignition sources, our analysis revealed changes in both the quantity and varying levels of attention given to human versus natural causes across different regions (Supplementary Fig. 4). Human-caused fires are increasingly dominating wildfire research in South America, Europe, and numerous parts of Asia. Although in regions like the Amazon and Southeast Asia, commodity-driven deforestation remains the primary cause of forest loss[23,24], fires ignited for land clearing purposes to create cropland,
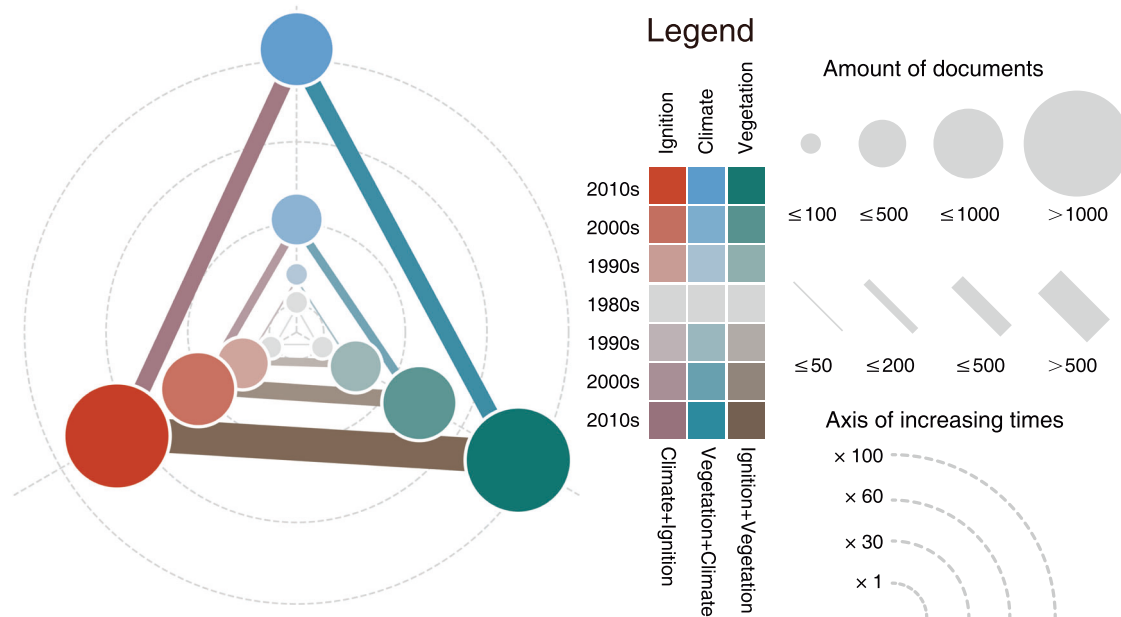
Fig. 4 | The dynamics of wildfire research paradigms based on the conceptual "fire triangles". The changes of research paradigms were assessed by the relative increase in the number of publications primarily concerning one of the three "fire triangles" (ignition, climate, fuel (vegetation))[22]. Connection lines denote publications addressing two of the triangle points (aspects). Each different aspect is marked by a different color. The chronological sequence of these research focuses is also illustrated with color changes. The sizes of circles and widths of lines indicate the number of annual publications within a specific decade. The background is depicted by light gray dashed lines, representing the axis of increasing frequency of publications.

pastures, and plantations are common practices[23]. These fires sometimes spiral out of human control and can escalate into numerous active fire spots and extensive burned area[25,26]. Among the research with inferred ignition sources, "Vegetation," "Atmospheric" are ranked as the top topics for both the human and natural-caused fires in most areas. China and India show different preferences in ignition studies, likely due to the lower percentage of burned areas, higher instances of crop residual burning, and heightened population exposure[27,28].

Conclusion

In conclusion, the utilization of LLM allowed us to efficiently and quickly monitor research trends across numerous specific research questions and objectives, while also in a transparent and upgradable manner. By deploying an AI-aided approach, we uncovered large disparities among various wildfire research components and quantified their associated levels of imbalance with an extensive database of tens of thousands of peer-reviewed papers. This becomes especially relevant in light of the ongoing rapid global changes that are expected to heighten fire risks[29,30]. We found that the pronounced imbalances in many fire-prone regions (e.g., Africa and the Amazon) coincides with the less-developed and less-resilient to increasing fire activities. Such disparity undermines our ability to understand the historical role of fire in local ecosystems and society, and to design mitigation strategies coping with anticipated future global changes.

Methods

Query from bibliographic database

To identify and gather as much relevant literature as possible on the topic of "wildfires," we conducted a thorough search of the Web of Science bibliographic database. The detail of searching and query results are demonstrated as below:

1. Temporal range: 1900-13th Feb. 2023
2. Keywords: see details in Supplementary Table 1

A total of 103,720 peer-reviewed papers and conference reports were selected from the query after duplicates removed.

Information extraction

We implemented the gpt-3.5-turbo model by utilizing the OpenAI API, which is identical to the model used in the ChatGPT product and requires no fine-tuning steps. Based on 1,569 independently screened results (with 899 identified as "related to wildfire research" through human classification), we employed a cross-validation approach to test the model's ability to make "rational" decisions and automatically eliminate items with little relevance to the key topic. Binary comparisons between the model outputs and human-supervised results achieved an average F1 score of 0.85 through bootstrapping with 1000 repetitions, a value comparable to that of previous research based on BERT. Initially, 72,352–80,297 papers were selected as belonging to the category "related to wildfire research" based on various descriptions of prompt instructions. We systematically designed a set of advanced prompts to extract relevant information from the literature, such as the major and minor disciplines, study area, study period, stage of fire (whole fire process/pre-fire/actively burning/post-fire), and other key information summarized from abstracts. To ensure maximum coverage, we carefully tested and refined the prompts. Additionally, we utilized the geoparser capabilities of the model to extract location information from the text and converted it to sets of coordinates describing the maximum boundaries of geographical entities at 1° spatial resolution. Country codes under the standard ISO-3166 were also returned where applicable. The valid geoparsing results within the scope of wildfire research returned a total of 60,488 articles, making up the primary database for subsequent analysis. The workflow, including the geolocation parsing process, is illustrated in Supplementary Fig. 1.

Prompt examples

In contrast to feature engineering, data cleaning, and other machine-learning processing approaches, LLMs prefer prompting interfaces as the method to interact with them. The quality of prompts directly affects the effectiveness of results. Here, we provide the prompt lines that used in this study to extract information:

Key information extraction

"You are a perfect classifier. List all the following contents and start with a new line. I would like you to (not specific or not applicable marked as N/Az):

(1) determine if the text is related to wildfire research (1 for yes and 0 for no, item name: relaty);

(2) if not (relaty eq 0), marked as NANA_zxy and then stop this dialogue;

(3) if so (relaty eq 1), select at least one topic from the following pool: vegetation/zoological/atmospheric/climate change/ecological/environmental/anthropogenic/hydrological/modeling/remote sensing/site-level observation/soil, list the type of article (research/review/opinion/policy/letters), study area, study period (e.g., 2005, 2001–2010), stage of fire (whole fire process/pre-fire/actively burning /post-fire). List the above information in different lines.

Notice: medical-relevant papers should be marked as NANA_zxy and not be considered wildfire research."

Geolocation parser

"You are an accurate geoparser with outputs in 1° spatial resolution (per grid). Parse the following text into separate regions. Start with a new paragraph for every single region. Define the maximum outer boundary as MOB ([left-top location, right-bottom location]) and the central point as CP. All the MOB and CP are recorded in the pairs of (latitude, longitude) and integers.

(1) if it is a qualified geographical location:
(a) no smaller than the area of 1 grid, list the parsed name and MOB;
(b) smaller than the area of 1 grid, list the parsed name and CP;
(2) if not, only return 'None + '; no extra notes/raw texts nor explanations should be returned."

Ignition classifications

"does the given text contain any descriptions of ignition sources?

if does, select from the below type: natural-caused/human-caused/mixed; default as mixed; if not, return N/A; do not return any extra/further words."

Multisource data

To ensure comparable spatial and temporal coverage with the evidence synthesis map, we applied different sources of supplementary data to depict the comparisons between publication attention and observed wildfire occurrence or emission. The burned area data were from AVHRR-LTDR from 1982 to 2018 at 0.05-degree spatial resolution[31]. This global burned area product was derived from a long-term data record generated from advanced very high-resolution radiometer images. Fire emission data were obtained from GFED4.1 s from 1997 to 2021, with inclusion of contributions from small fires and revised fuel consumption parameterizations optimized[16]. We used the Gridded Population of the World, Version 4 (GPWv4, Population Count Adjusted to Match 2015 Revision of UN WPP Country Totals, Revision 11) as the human population inputs. This product was adjusted to match the 2015 Revision of the United Nation's World Population Prospects (UN WPP) country totals for the years from 2000 to 2020 at 5-year intervals[32]. Socioeconomic development indicators were obtained from Gridded global datasets for Gross Domestic Product from 1990 to 2015[33]. We resampled and reprojected the results from above datasets to align with the same preferences as the publication patterns.

Data availability

Global Detection of Long-Term (1982–2017) Burned Area with AVHRR-LTDR Data is publicly available from ESA Climate Change Initiative (Fire_cci) (https://climate.esa.int/en/projects/fire/). The global fire emission database (GFED4.1 s) is available from its repository: https://www.globalfiredata.org/data.html. The Gridded Population of the World (GPWv4) is publicly available from NASA's Data Center, Socioeconomic Data and Applications Center (SEDAC): https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-count-adjusted-to-2015-unwpp-country-totals-rev11. Gridded global datasets for Gross Domestic Product can be accessed through https://doi.org/10.5061/dryad.dk1j0. The selected literature and processed data to create the figures are available through Zenodo: https://zenodo.org/records/10859331.

Code availability

The python programming codes used for the generation of ChatGPT outputs of this study are publicly available through the open-access repository https://zenodo.org/records/10811211.

References

1. Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. Meta-analysis and the science of research synthesis. *Nature* **555**, 175–182 (2018).
2. Bowman, D. M. J. S. et al. Vegetation fires in the Anthropocene. *Nat. Rev. Earth Environ.* **1**, 500–515 (2020).
3. He, T. & Lamont, B. B. Baptism by fire: the pivotal role of ancient conflagrations in evolution of the Earth's flora. *Nat. Sci. Rev.* **5**, 237–254 (2018).
4. Callaghan, M. et al. Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nat. Clim. Chang.* **11**, 966–972 (2021).
5. Thoppilan, R. et al. LaMDA: Language models for dialog applications. Preprint at https://doi.org/10.48550/arXiv.2201.08239 (2022).
6. OpenAI. GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).
7. Zhao, W. X. et al. A Survey of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2303.18223 (2023).
8. Wang, J. et al. Global evidence of expressed sentiment alterations during the COVID-19 pandemic. *Nat. Human Behav.* **6**, 349–358 (2022).
9. Hong, Z. ChatGPT for computational materials science: a perspective. *Energy Mat. Adv.* **4**, 0026 (2023).
10. Chatterjee, J. & Dethlefs, N. This new conversational AI model can be your friend, philosopher, and guide… and even your worst enemy. *Patterns* **4**, 100676 (2023).
11. Patel, S. B. & Lam, K. ChatGPT: the future of discharge summaries? *Lancet Dig. Health* **5**, e107–e108 (2023).
12. Zhong, Q., Ding, L., Liu, J., Du, B. & Tao, D. Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT. Preprint at http://arxiv.org/abs/2302.10198 (2023).
13. Chuvieco, E. et al. Satellite remote sensing contributions to wildland fire science and management. *Curr. Forestry Rep.* **6**, 81–96 (2020).
14. Petersen, O. H. Inequality of research funding between different countries and regions is a serious problem for global science. *Function* **2**, zqab060 (2021).
15. Rich countries must align science funding with the SDGs. *Nature* **621**, 444 (2023).
16. van der Werf, G. R. et al. Global fire emissions estimates during 1997–2016. *Earth Syst. Sci. Data* **9**, 697–720 (2017).
17. Spreading like wildfire. *Nat. Clim. Change* **7**, 755 (2017).
18. Xu, Y. et al. Recent expansion of oil palm plantations into carbon-rich forests. *Nat. Sustain.* https://doi.org/10.1038/s41893-022-00872-1 (2022).
19. Zalles, V. et al. Near doubling of Brazil's intensive row crop area since 2000. *Proc. Nat. Acad. Sci.* **116**, 428–435 (2019).
20. Andela, N. et al. A human-driven decline in global burned area. *Science* **356**, 1356–1362 (2017).
21. Fan, L. et al. Siberian carbon sink reduced by forest disturbances. *Nat. Geosci.* https://doi.org/10.1038/s41561-022-01087-x (2022).
22. Moritz, M. A., Morais, M. E., Summerell, L. A., Carlson, J. M. & Doyle, J. Wildfires, complexity, and highly optimized tolerance. *Proc. Nat. Acad. Sci.* **102**, 17912–17917 (2005).
23. Andela, N. et al. Tracking and classifying Amazon fire events in near real time. *Sci. Adv.* **8**, eabd2713 (2022).

24. Curtis, P. G., Slay, C. M., Harris, N. L., Tyukavina, A. & Hansen, M. C. Classifying drivers of global forest loss. *Science* **361**, 1108–1111 (2018).

25. Page, S. E. et al. The amount of carbon released from peat and forest fires in Indonesia during 1997. *Nature* **420**, 61–65 (2002).

26. Xu, W. et al. Active fire dynamics in the amazon: new perspectives from high-resolution satellite observations. *Geophys. Res. Lett.* **48**, e2021GL093789 (2021).

27. Shyamsundar, P. et al. Fields on fire: Alternatives to crop residue burning in India. *Science* **365**, 536–538 (2019).

28. Das, B. et al. A model-ready emission inventory for crop residue open burning in the context of Nepal. *Environ. Pollut.* **266**, 115069 (2020).

29. Turco, M. et al. Exacerbated fires in Mediterranean Europe due to anthropogenic warming projected with non-stationary climate-fire models. *Nat. Commun.* **9**, 3821 (2018).

30. Chen, Y. et al. Future increases in Arctic lightning and fire risk for permafrost carbon. *Nat. Clim. Chang.* **11**, 404–410 (2021).

31. Otón, G., Lizundia-Loiola, J., Pettinari, M. L. & Chuvieco, E. Development of a consistent global long-term burned area product (1982–2018) based on AVHRR-LTDR data. *Int. J. Appl. Earth Observ. Geoinform.* **103**, 102473 (2021).

32. Center for International Earth Science Information Network - CIESIN - Columbia University. *Gridded Population of the World, Version 4 (GPWv4): Population Count Adjusted to Match 2015 Revision of UN WPP Country Totals, Revision 11* (NASA Socioeconomic Data and Applications Center (SEDAC), 2018).

33. Kummu, M., Taka, M. & Guillaume, J. H. A. Gridded global datasets for gross domestic product and human development index over 1990–2015. *Sci. Data* **5**, 180004 (2018).

Author contributions
Xuhui Wang & Zhengyang Lin conceived and designed the study. Zhengyang Lin performed the analyses and drew the figures. Zhengyang Lin, Anping Chen and Xuhui Wang wrote the first version of the manuscript, with inputs and revisions from Zhihua Liu and Shilong Piao. All authors contributed to interpretation of the results.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-024-01341-7.

Correspondence and requests for materials should be addressed to Xuhui Wang.

Peer review information *Communications Earth & Environment* thanks Ioannis Prapas and Roberto Cilli for their contribution to the peer review of this work. Primary Handling Editor: Aliénor Lavergne. A peer review file is available.

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

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