UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Report on the 1st Workshop on Information Retrieval for Climate Impact (MANILA24) at SIGIR 2024

de Rijke, Maarten; van den Hurk, Bart; Salim, Flora; Khourdajie, Alaa Al; Bai, Nan; Calzone, Renato; Curran, Declan; Demil, Getnet; Frew, Lesley; Gießing, Noah; Gupta, Mukesh Kumar; Heuss, Maria; Hobeichi, Sanaa; Huard, David; Kang, Jingwei; Lucic, Ana; Mallick, Tanwi; Nath, Shruti; Okem, Andrew; Pernici, Barbara; Rajapakse, Thilina; Saleem, Hira; Scells, Harry; Schneider, Nicole; Spina, Damiano; Tian, Yuanyuan; Totin, Edmund; Trotman, Andrew; Valavandan, Ramamurthy; Workneh, Dereje; Xie, Yangxinyu

DOI
10.1145/3769733.3769737

Publication date
2025

Document Version
Final published version

Published in
SIGIR Forum

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment)

Link to publication

Citation for published version (APA):
de Rijke, M., van den Hurk, B., Salim, F., Khourdajie, A. A., Bai, N., Calzone, R., Curran, D., Demil, G., Frew, L., Gießing, N., Gupta, M. K., Heuss, M., Hobeichi, S., Huard, D., Kang, J., Lucic, A., Mallick, T., Nath, S., Okem, A., ... Xie, Y. (2025). Report on the 1st Workshop on Information Retrieval for Climate Impact (MANILA24) at SIGIR 2024. *SIGIR Forum*, *59*(1). https://doi.org/10.1145/3769733.3769737

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

Download date:10 Oct 2026

Report on the 1st Workshop on Information Retrieval for Climate Impact (MANILA24) at SIGIR 2024

Maarten de Rijke
University of Amsterdam
The Netherlands
m.derijke@uva.nl

Bart van den Hurk
IPCC WGII and
Deltares Water Knowledge Institute
The Netherlands
bart.vandenhurk@deltares.nl

Flora Salim
UNSW Sydney
Australia
flora.salim@unsw.edu.au

Alaa Al Khourdajie, Nan Bai, Renato Calzone, Declan Curran, Getnet Demil, Lesley Frew, Noah Gießing, Mukesh Kumar Gupta, Maria Heuss, Sanaa Hobeichi, David Huard, Jingwei Kang, Ana Lucic, Tanwi Mallick, Shruti Nath, Andrew Okem, Barbara Pernici, Thilina Rajapakse, Hira Saleem, Harry Scells, Nicole Schneider, Damiano Spina, Yuanyuan Tian, Edmund Totin, Andrew Trotman, Ramamurthy Valavandan, Dereje Workneh, Yangxinyu Xie*

Abstract

The purpose of the MANILA24 Workshop on information retrieval for climate impact was to bring together researchers from academia, industry, governments, and NGOs to identify and discuss core research problems in information retrieval to assess climate change impacts. The workshop aimed to foster collaboration by bringing communities together that have so far not been very well connected – information retrieval, natural language processing, systematic reviews, impact assessments, and climate science. The workshop brought together a diverse set of researchers and practitioners interested in contributing to the development of a technical research agenda for information retrieval to assess climate change impacts.

Date: 18 July 2024.

Website: https://sites.google.com/view/ir-for-climate-impact/2024/manila24-workshop-page.

1 Introduction

Human-induced climate change, including more frequent and intense extreme events, has caused widespread adverse impacts and related losses and damages to nature and people, beyond natural climate variability [IPCC, 2022a]. The Intergovernmental Panel on Climate Change (IPCC) is the

*Affiliation not shown for all authors due to space limitations (see Appendix A for details).

leading international body for assessment of climate change. Approximately every six years, the IPCC releases an assessment report on the different aspects, drivers, and impacts of climate change based on an assessment of the literature. The IPCC report has contributions from three different working groups. In particular, Working Group II (WGII) of the IPCC "assesses the impacts, adaptation, and vulnerabilities related to climate change, from a world-wide to a regional view of ecosystems and biodiversity, and of humans and their diverse societies, cultures and settlements. It considers their vulnerabilities and the capacities and limits of these natural and human systems to adapt to climate change and thereby reduce climate-associated risks together with options for creating a sustainable future for all through an equitable and integrated approach to mitigation and adaptation efforts at all scales" [IPCC, 2024].

Fully using the available knowledge on emerging climate change impacts is key to informing global policy processes as well as regional and local risk assessments and on-the-ground action on climate adaptation [Schleussner and Fyson, 2020]. The exponential growth in peer-reviewed scientific publications on climate change is pushing manual expert assessments to their limits [Callaghan et al., 2021, 2020; Joe et al., 2024]. While literature aggregated on the level of continents or world regions might be useful to the global policy process, informing concrete climate adaptation typically requires localized and contextualized information on climate impacts [Conway et al., 2020]. Tracking the effectiveness and progress of adaptation actions has proven difficult [Sietsma et al., 2024] – any attempt to track adaptation progress will need to be capable of rapidly handling large and varied datasets and literature sources, while acknowledging highly localized and contextualized information.

This workshop report brings together researchers from the information retrieval, natural language processing, systematic review, and climate science communities in an attempt to develop an agenda to advance information retrieval for climate impact assessment. We begin by dissecting the problem: what is the information need addressed by the IPCC WGII (Section 2)? We then switch to methodologies, and in particular systematic reviews (Section 3). We review different resources available (and/or needed) to support information retrieval for climate impact, including test sets and implementations (Section 4). We then address the question of how to make new technological advances in information retrieval work as part of the IPCC WGII assessment workflow (Section 5). The paper concludes with a broader perspective. To remain focused, our discussion and analysis is centered around IPCC WGII and its mission – we believe, however, that many of our questions and suggestions have the potential to contribute to the workflows of other IPCC working groups and task forces.

2 The Information Need

The IPCC aims to provide governments at all levels with scientific information that they can use to develop climate policies [United Nations, 2024]. The IPCC is divided into three working groups (WGs). WGI deals with the physical science basis of climate change, WGII with climate change impacts, adaptation and vulnerability, and WGIII with mitigation of climate change. The IPCC does not conduct its own research, run models, or make measurements of climate or weather phenomena. Its role is to assess the scientific, technical and socio-economic literature relevant to understanding climate change, its impacts, future risks, and options for adaptation and mitigation. Author teams assess all such information from any source that is to be included in the report.

As pointed out in the introduction, approximately every six years, the IPCC releases a series of reports on the different aspects of climate change based on large-scale assessment of all the latest literature. In early 2022, as part of the IPCC's sixth assessment report (AR6), the IPCC released the report of WGII on impacts, adaptation and vulnerability, which "assesses the impacts of climate change, looking at ecosystems, biodiversity and human communities at global and regional levels. It also reviews vulnerabilities and the capacities and limits of the natural world and human societies to adapt to climate change" [IPCC, 2022a]. The report covers ecosystems, sectors, sustainable development goals, and regions, an integrated technical summary, and a summary for policymakers. The AR6 WGII report pulls together evidence and findings from more than 34,000 journal papers and reports; it is written by 270 authors from 67 countries.

The assessment report currently in development (AR7) faces a twin challenge: assessing and synthesizing a fast-growing amount of literature related to climate change, and addressing relevant topics and areas where little data is available in the published literature. We believe that information retrieval can potentially help address both aspects, by (i) helping IPCC author teams sift through the literature, and (ii) exposing relevant information previously inaccessible to author teams due to various barriers (language, etc.).

2.1 Proposed Research

We propose an agenda driven by the ambition to produce open evidence synthesis, based on principles of transparent information gathering, curation, traceability, and information quality. Fairness and mitigating different types of bias (e.g., geographic, cultural) are important proposed research lines. Specific attention is to be paid to underrepresented sources of information and communities. Existing AI tools work well with large amounts of data, so have the potential to exacerbate imbalances in data and literature coverage [Bahri et al., 2024].

How do we make sure AI can help with topics or areas where less information is available? Another line of research has to do with the data sources used for evidence synthesis: (i) (capturing and) using domain-specific information, (ii) the usage of grey literature, i.e., the diverse and heterogeneous body of material available outside, and not subject to, traditional academic peer-review processes such as preprints and policy documents [Adams et al., 2017], (iii) Indigenous Knowledge and local knowledge, especially if it is not available online, and (iv) (qualitative) data collected through grass roots and/or citizen data science efforts.

It is important that retrieval and analysis models can handle multi-modal data (i.e., text, numerical, images, etc.). Only published data currently goes into the IPCC review process, and this is will not cover, e.g., remote sensing data that has no description and interpretation in the literature. The final direction concerns important guardrails that apply to the evidence synthesis process. One is that there is zero tolerance for black boxes – this is important to avoid undermining the credibility of the synthesis with policy makers and governments. An important requirement for "unconventional" sources of information is its quality assessment; IPCC authors need to defend the appropriateness of the use of non-peer reviewed literature; we need to elaborate on this quality assessment of non peer-reviewed literature. And another concerns the carbon emissions when choosing information retrieval (IR) approaches (i.e., energy requirements of using large language models (LLMs)). Do benefits of the model/data/research outweigh the carbon costs?

2.2 Research Challenges

These general research questions manifest themselves along the entire evidence synthesis chain and motivate a broad range of concrete research directions to be investigated concerning the information need that IPCC addresses. Some of these have to do with the scope of the evidence synthesis, some concern data acquisition and quality assessment, and some with contextual factors.

Data gathering and sharing. If more diverse sources of evidence are integrated into the evidence synthesis process, how can we ensure open, transparent data collection and sustainable sharing?

Data quality. A common suggestion is that evidence should include more diverse sources. Evidence based on solely grey literature should be supported by other evidence lines. However, combining datasets of different structure or manually annotating unstructured data is time-intensive, and grey literature in particular is difficult to work with [Sietsma et al., 2024]. How can local expert knowledge be embedded in the evidence synthesis process, while ensuring that data/literature imbalances are not exacerbated? How can acceptable data be created from unconventional sources?

Climate adaptation. It has been claimed that robust synthesis of climate *adaptation* literature and insights was under-represented in climate impact assessments, including the most recent version of the WGII report [Berrang-Ford et al., 2021]. How to address and capture successful adaptations, whether technological, institutional, behavioral, or nature-based [O'Neill et al., 2022]. How to surface and flag information that helps avoid "maladaptations" that may increase vulnerability, lock-in, or unequal impacts [Barnett and O'Neill, 2013]? How to measure and mine outcomes of adaptation actions and aggregate them, e.g., based on similarity of contexts, precision, causality?

Timeliness. A single IPCC assessment cycle lasts approximately six years and the resulting syntheses are difficult to keep up to date. How can more recent statistics, figures, and findings from research papers lead to "updated" reports? What would acceptable "living evidence" in the context of the IPCC look like [Elliott et al., 2017]?

Evaluation. How can we evaluate "fit to information need" of assessment reports or of the various summaries, such as the summary for policymakers [IPCC, 2022b] or the technical summary [Pörtner et al., 2022], that are generated based on the reports? The IPCC uncertainty guidances notes [Mastrandrea et al., 2010] are meant to assist in the consistent treatment of uncertainties in developing expert judgments and in communication – how can they be integrated with traditional evaluation guidelines from the IR community?

2.3 Obstacles and Risks

To enable this research we need broad collaborations between IR researchers and climate impact researchers. Finding effective ways of working together and finding a shared vocabulary requires considerable effort that some researchers may not be able to afford or have insufficient institutional

support, resources or recognition for [Allan et al., 2018]. An important risk of using information retrieval based literature collection concerns the continuation or even exacerbation of imbalances in coverage in IPCC reports by focusing on areas and topics that have large amounts of data available for training. A potential obstacle for bringing recent advances in IR technology to bear on the IPCC's evidence synthesis workflow is the broad – and successful – uptake of technologies that are often referred to as "opaque" or "black-box" technologies such as LLMs and generative AI. Another key constraint is that the IPCC report production process is guided by a set of principles agreed by the panel,[1] potentially limiting the integration of important innovations in the evidence synthesis process.

3 Methodology

The logic of the IPCC assessments is somewhat similar to the workflow of systematic reviews. Traditional systematic reviews are structure following a number of, more or less, standardized steps [Cooper et al., 2018]: (i) querying (which includes (a) protocol definition and (b) search strategy development), (ii) screening (which includes (a) study abstract screening and (b) study full-text screening), and (iii) aggregating (which includes (a) study synthesis and results preparation and (b) dissemination of systematic review). IPCC's approach can be characterized using the following dimensions:

Rigorous assessment: Use established methods for comprehensive reviews, aligning with IPCC's mission to assess climate literature.

Database utilization: Employ robust academic databases such as Scopus and Web of Science for extensive literature search.

Literature screening: Implement a systematic selection process to ensure the inclusion of relevant studies.

Publication coding: Use qualitative analysis software to code and categorize information effectively.

Scholarly perspectives: Evaluate the gathered facts and identify research gaps to provide a well-rounded scholarly insight.

Data visualization: Create visual representations such as diagrams, word clouds, and tables to present the synthesized data clearly.

Connect with practitioners: Summarize key findings to inform policy implications and engage with practitioners for real-world interpretation.

IPCC's current approach to evidence synthesis has several shortcomings: (i) the growth of the literature: human-driven information collection skews search relevance, hence tool support is needed to manage and streamline new literature; (ii) information comes in diverse formats, both

[1] https://www.ipcc.ch/site/assets/uploads/2018/09/ipcc-principles-appendix-a-final.pdf

structured and unstructured: fragmented information hinders understanding and analysis, hence tool support is needed to integrate structured and unstructured data; and (iii) there is a significant time lag due to the six year cycle, for which *living* systematic reviews may be a way forward [Elliott et al., 2017].

3.1 Proposed Research

Developing evidence synthesis methodologies that help address challenges in IPCC's assessment process requires new research on a broad range of information retrieval topics subject to a number of constraints and guardrails. In particular, information retrieval methods are needed for identifying, integrating, understanding, and selecting relevant literature and information extraction methods for (i) efficiently synthesizing relevant data and (ii) improving accuracy and relevance of extracted information. The main tasks requirements that these methods need to address are (i) eliminating hallucination in case LLM-based methods are used; (ii) mitigating bias by enhancing the selection process and promoting fairness, inclusivity, and representativity; and (iii) maintaining human accountability and ensure transparency in AI-assisted processes while securing a comprehensive coverage of the literature to be assessed.

3.2 Research Challenges

Challenges are faced on each of the areas that the proposed research covers. The proposed research touches on the core of evidence synthesis: the way of working, the data, data enrichment, the synthesis and summarization process, its timeliness, and the role of people in the process.

Protocol development. How should evidence synthesis protocols be co-designed to facilitate collaborative research, adhering to Indigenous protocols, and including under-represented, culturally and linguistically diverse groups [Roberts et al., 2021]? What are effective ways of engaging with these communities? How can we use generative AI to produce effective communication material for different communities [Lansbury et al., 2023]?

Characterizing the literature. If the input to the synthesis process diversifies, what is the process of search? How, using which credibility criteria, should we determine corpora of scientific and "non-traditional" research outputs (e.g., Indigenous Knowledge), grey literature (e.g., government reports), and select sources to include?

The labeling bottleneck. Traditionally, assessment reports relied on domain experts who identified and selected, often in teams, documents for relevance and for specific aspects of climate impact as part of the evidence gathering and synthesis process. The rapid growth of scientific literature has made this increasingly challenging. Can evidence gathering and screening be automated with recent advances in the use of LLMs for relevance labeling [Thomas et al., 2024]? Can information extraction tasks be off-loaded to LLMs [Mallick et al., 2024]? And to which degree can LLMs be used to compare and synthesize extracted evidence [Joe et al., 2024]?

Trustworthy synthesis. IPCC's assessment reports inform policy-makers. They provide a scientific basis for governments at all levels to develop climate related policies. The importance of understanding how to synthesize information with varying degrees of fidelity cannot be overstated. For example, in the climate model intercomparison project[2] organizations can submit their climate models, which others can use to do climate simulations, usually by compute averages over all the different models. In CMIP6, there were 50+ models; some were "too hot" [Hausfather et al., 2022] and averaging would therefore overestimate temperature increase. IPCC6 acknowledged this, stating that models should be weighted according to fidelity, not just blindly synthesized.

IPCC's reports, and the processes and systems underlying them, need to be trustworthy [Wang et al., 2024]. From a technological point of view, a process or technology can be trustworthy for *intrinsic* reasons, e.g., because a sufficient level of transparency of the process and technology are provided. Or it can be trustworthy for *extrinsic* reasons, i.e., because we have ways of probing it for properties such as accuracy, reliability, repeatability, resilience, and safety. How can these intrinsic and extrinsic approaches be developed for large-scale evidence synthesis?

Living evidence. Evidence changes over time. What are effective update protocols? The temporal dimension plays a big part in evidence outcomes. Is the amount of literature that is to be assessed in a given IPCC cycle still manageable and can it be expected to be representative for the target time period? What would the requirements be for self-updating synthesis systems? How should we preserve previous versions [Simmonds et al., 2017]?

Human in the loop synthesis. How can we measure the cost, and benefit, of having humans in the evidence synthesis loop [Thomas et al., 2017]? How should we provide feedback to the systems and who provides this feedback? And vice versa, how do we transfer tasks from systems to people, or trigger interventions from people in complex evidence synthesis tasks? Can interactive or collaborative frameworks from other tasks be adopted (e.g., [Zhang et al., 2021])?

3.3 Obstacles and Risks

We currently know very little about the extent to which automation of key evidence synthesis steps such as query automation, screening automation, and automation of the aggregation results are effective for evidence synthesis in the climate impact domain, and what little we know has usually been learned on English language scientific publications. An important obstacle is the lack of data and benchmarks (especially Indigenous Knowledge, and low-resource languages) and lack of training corpus for these languages. An important risk is limited success in community engagement, across disciplines and across different types of stakeholders. It is important to do this in the right way by adhering to responsible practices, even though they may not have been established for all types of (non-traditional) domain experts and underrepresented groups [see, e.g., Lewis et al., 2020].

[2]CMIP, https://wcrp-cmip.org/cmip-phases/cmip6/

4 Resources

Progress in the development of effective methods to automate all steps in systematic reviews has been greatly facilitated by the availability of shared resources. In evidence synthesis in healthcare and medicine, for instance, we have witnessed an explosion of methods due to shared tasks and datasets. Prominent examples are the CLEF Technology Assisted Reviews task [Kanoulas et al., 2019] and the CSMeD: Meta-dataset for systematic review automation evaluation [Kusa et al., 2023]. In information retrieval for climate impact there is an urgent need for similar standardized datasets and test collections.

4.1 Proposed Research

Support for the creation of shared resources (data, code, evaluation methods) requires an understanding of the information and steps that are needed to accomplish the assessment goals, of the (inter)actions of evidence reviewers working towards the goals. This has two sides: (i) collecting information on current assessment practices, and (ii) analyzing data, process and options for automation.

It is instructive to revisit the main steps in the systematic review process [Cooper et al., 2018; Scells, 2021] – *querying*, *screening*, and *aggregating* – and see what progress has been enabled using shared resources and large language models in recent years.

The *querying stage* typically involves steps such as query formulation, query refinement, and query representation. As an example of recent progress using LLMs and building shared resources in the context of querying, Wang et al. [2023] explore the use of LLMs in the medical domain to follow instructions and generate queries with high-precision, while trading this off for recall. Using resources from the CLEF initiative[3] for evaluation, the authors find that LLMs are a valuable tool for rapid reviews where time is a constraint and trading off higher precision for lower recall is acceptable.

The *screening* stage typically involves screening prioritization, cut-off prediction, and document classification. As another example of recent progress facilitated by the CLEF e-Health,[4] the Text Retrieval Conference (TREC) Total Recall benchmarking activity,[5] and the TREC Legal[6] resources, Stevenson and Bin-Hezam [2023] propose a stopping method based on point processes and test the robustness of a wide range of statistical and neural stopping methods in a series of contrastive experiments.

Finally, the *aggregating* stage typically involves information extraction, result synthesis, statistical analyses, and dissemination. As an example of recent progress that uses LLMs and builds on shared resources, Shaib et al. [2023] examine the potential of LLMs to synthesize (that is, aggregate the findings presented in) multiple medical articles in a way that accurately reflects the evidence. Using articles indexed in the Trialstreamer database[7] and meta-analyses from the Cochrane Li-

[3] https://www.clef-initiative.eu
[4] https://github.com/CLEF-TAR
[5] http://plg.uwaterloo.ca/~gvcormac/total-recall/
[6] https://trec-legal.umiacs.umd.edu/
[7] https://trialstreamer.ieai.robotreviewer.net

brary,[8] the authors conclude that human-written summaries in the (bio)medical domain require a level of synthesis that is not yet captured by today's LLMs.

Do these findings generalize from the medical domain to information retrieval for climate impact and to all aspects of the systematic review workflow? What are the resources needed to facilitate similar advances in information retrieval for climate impact and how do we create them?

4.2 Research Challenges

Resources in support of information retrieval for climate impact can be addressed at two levels – at the level of individual resources and at the aggregate level of multiple resources. First level, at the level of individual resources, we see the following research challenges.

4.2.1 Individual Resources

Document collections. Several collections relevant to support research into information for climate impact have been used in recent years. White literature collections used include a subset of the Semantic Scholar Open Research Corpus consisting of 600K climate-related articles [Mallick et al., 2024], subsets (abstracts, titles and metadata (no full text)) from the Web of Science Core Collections databases consisting of 375K [Callaghan et al., 2020] and 565K [Sietsma et al., 2021] articles. OpenAlex is another option.[9] Grey literature of different sorts has also been used in the context of climate science, even though for IPCC not all types of grey literature (e.g., tweets and media coverage) are considered eligible. Bai et al. [2024] share a collection of 60K (multi-modal) tweets covering different climate change stances. The GDELT climate news narrative datasets, covering 2009 to 2020, represent television, language, the global perspective and context dating back from half a decade to a decade in all.[10] Frew et al. [2024] use web archives and query logs to demonstrate how conflicting climate change stances between government administrations were correlated with content drift and lack of public access to climate information webpages over time. How can we create a large-scale, heterogeneous corpus for climate impact, a corpus, moreover that would support a heterogeneous benchmark containing diverse climate-related tasks (like the BEIR dataset in ad-hoc retrieval [Thakur et al., 2021])? How do we determine the usefulness of including non-peer reviewed scientific reports, theses, policy documents in corpus and benchmark development? How can we create benchmarks, for both static and tracking tasks?

Knowledge sources. Knowledge sources play an important role in making sense of both white and grey climate literature. Example knowledge sources used to understand climate science literature at scale include KnowWhereGraph [Janowicz et al., 2022], a geo-knowledge graph that is based on existing standards like RDF,[11] OWL[12] and GeoSPARQL,[13] incorporates custom ontologies, and uses a hierarchical grid for spatial representations; it is built upon many open-source

[8] https://www.cochranelibrary.com/

[9] https://openalex.org/.

[10] https://blog.gdeltproject.org/four-massive-datasets-charting-the-global-climate-change-news-narrative-2009-2020/

[11] https://www.w3.org/RDF/

[12] https://www.w3.org/OWL/

[13] https://www.ogc.org/publications/standard/geosparql/

spatial datasets, including an expert knowledge graph built using scientific publication/papers on disaster relief.[14] Two other important knowledge sources for anomaly and unusual event detection are the iNaturalist biological species dataset[15] and the local environmental observer (LEO) network.[16] How can we use these resources to enrich white and grey literature at scale and enable different types of semantic search [Balog, 2018]?

Tools and technology. Various domain-specific resources and tools have been developed for information retrieval for climate impact recently. For example, ClimateBert is transformer-based language model that is further pretrained on over 2 million paragraphs of climate-related texts, crawled from various sources such as common news, research articles, and climate reporting of companies [Webersinke et al., 2021]. WildfireGPT is an LLM designed to support the analysis of wildfire data, by a diverse set of end users, including researchers and engineers. WildfireGPT is an interesting example of a domain-specific resource not just because of its use cases but also because it uses a broad range of domain-specific data [Xie et al., 2024]. Thulke et al. [2024] introduce ClimateGPT, a family of large language models designed for understanding and responding to information about climate change; the paper demonstrates the effectiveness of adapting a strong general-purpose LLM through continued pre-training on a curated climate corpus and fine-tuning it with high-quality, human-expert-involved instruction data.

Tian et al. [2025] propose a retrieval-reranking framework that uses LLMs to identify and recommend semantically and spatiotemporally similar unusual environmental events described in news articles and web posts. By integrating embedding-based semantic search with a geo-time re-ranking strategy that integrates multi-faceted criteria including spatial proximity, temporal association, semantic similarity, and category-instructed similarity, the work demonstrates how LLMs can support climate event mining in a real-world platform setting. Mallick et al. [2024] describe an end-to-end decision-support tool that accurately identifies location information within textual documents, enabling extraction of geographic-specific climate trends and topics, to provide a concise understanding of geographic-specific climate change trends. How can tools and technology with such far-reaching potential be made accessible to everyone, not just the technologically privileged? Rajapakse et al. [2024] describe a widely used library designed to simplify the training, evaluation, and usage of transformer models that operationalizes the idea of "open-source for all" with adoption in materials science, environmental science, and healthcare. Can we put this perspective to work for all tools and technology (or technology components) that help make sense of information related to climate impact?

4.2.2 Aggregating Across Multiple Resources

There are multiple research challenges that concern the development or usage at different levels of aggregation.

Transfer & generalization. The first challenge concerns the use of the broad range of advances and resources realized in language technology in recent years. How can we transfer these advances

[14] https://knowwheregraph.org/graph/
[15] https://www.inaturalist.org/observations?place_id=any&subview=map
[16] https://www.leonetwork.org/en/docs/about/about

to the domain and tasks of information retrieval for climate impact with minimal effort to reduce costs and accelerate the uptake *while* achieving reliable performance? Can we take the findings from evidence synthesis in technology assisted reviews in the medical or legal domain and transfer these to the climate domain? Can these questions be turned into few-shot learning problems [Wang et al., 2025]?

Benchmarking. The lack of standardized benchmarks and datasets for validating and comparing information retrieval models for climate impact poses significant challenges to the development, assessment, and comparison of information retrieval technologies in this field. Can we use (very) limited human labeling and rely on LLM-generated labels where large quantities of labels are needed [Thomas et al., 2024] for test collection creation? To which degree would a weak-to-strong alignment perspective [Lyu et al., 2025] be helpful for creating the required benchmarking resources?

Stacking up. Recent years have witnessed great progress in the performance of individual components that make up a systematic review workflow – in querying, screening, and aggregating. Do improvements in the components lead to improvements in the overall assessment process? How do issues such as bias, reproducibility, and performance emerge in complex retrieval for impact pipelines [Hocking et al., 2023]? How can the full review pipeline be optimized effectively? Is the field ready to explore end-to-end learnable approaches to information retrieval for climate impact such as generative information retrieval [Tang et al., 2024]?

Scaling up. Managing and processing the vast amounts of heterogeneous data from diverse sources such as sensors, climate models, literature and satellite data poses significant challenges. In the medical domain, organizations such as Cochrane and the Joanna Briggs Institute offer extensive methodological guidance for dealing with large-scale collections, putting a large emphasis on the use of software such as Covidence[17] or JBI Sumari.[18] How do such tools scale up to the size and heterogeneity of materials related to climate impact? Furthermore, linking text data from white and grey literature that describes a climate event to relevant physical measurements about the event, like temperature, rainfall, audio-visual recordings, etc. at a world-wide scale is a challenge that goes far beyond the entity linking challenges that the information retrieval community addresses today [Meij et al., 2013].

4.3 Obstacles and Risks

Collaboration among researchers in information retrieval, natural language processing, data engineering, and climate science, developers, and methodologists is essential for advancing the field of information retrieval for climate impact. The fact that these communities and sub-communities – with different sharing and evaluation cultures – are largely disconnected, is a risk that may hinder progress in the development and uptake of resources and benchmarks.

[17] https://www.covidence.org
[18] https://sumari.jbi.global

5 Usage

Recall the core steps from IPCC's overall evidence synthesis process – querying, screening, and aggregating. In each of these steps it employs provenance tracking as a core component of the workflow. By documenting the origin and reliability of the evidence, WGII provides a foundation for informed decision-making on climate change impacts, adaptation, and vulnerability.

5.1 Proposed Research

IPCC's way-of-working (with a strong emphasis on provenance tracking) comes with several key challenges. First of all, the exponential growth and increasing complexity of scientific literature complicates the IPCC's mandate to conduct "comprehensive, objective, open and transparent" assessment in line with the IPCC principles, an issue that is exacerbated by the workload required from scientists to engage with assessment [De-Gol et al., 2023; Minx et al., 2017]. Second, the length of the IPCC reports poses significant challenges to access the detailed underlying evidence in the main reports. For example, in the most recent assessment round (AR6), the two main reports exceed 2,000 pages each, with one exceeding 3,000 pages. Including additional specialized reports and summaries, a total of ∼12,000 pages was reached [Al Khourdajie, 2024]. The sheer volume hinders the effective dissemination of important scientific findings that do not feature in the summaries.

Assuming that the community manages to make progress along the lines suggested in Section 2 (on understanding IPCC's information need), Section 3 (on methods for addressing IPCC's information needs), and Section 4 (on resources to implement and assess those methods), how and where can the resulting advances be made to work for IPCC's assessments, given the challenges outlined above? We need to navigate the vast and ever-expanding corpus of white and grey literature on climate in an efficient manner to enable comprehensive and credible consensus building processes. How to use tools to effectively communicate the extensive findings in the lengthy IPCC reports while maintaining the integrity of the assessments? How to determine the weight of heterogeneous evidence, complexity, uncertainty and the level of confidence?

5.2 Research Challenges

Al Khourdajie [2024] summarize several current lines of work to structure the assessment work, in a way that addresses the questions listed above and that provides connecting points for the agendas in Section 2, 3 and 4. The IPCC scoping process could use *evidence maps* to highlight key topics. Once outlines are agreed, *topic discovery* could help authors of assessment reports plan drafts by clustering publications and topics. A *hybrid exploration* approach could then be used to inform assessment drafting. *Collaborative platforms* could then be used to expand engagement, while *exploration* tools would help to expand coverage.

Synthesis and evidence maps. An evidence map is a visual representation of the available research on a particular topic. It provides an overview of the extent, range, and nature of evidence, highlighting where research exists and, importantly, where gaps remain [Hempel et al., 2014]. In the area of climate adaptation research gaps are present in the assessment of adaptation

effectivity; there is regional skewness in literature coverage (with substantially more literature in global north countries), and there is poor coverage of quantitative assessments of adaptation performance [IPCC, 2022a]. How can LLMs be used to identify and classify white and grey literature publications and provide rapid and comprehensive coverage?

Topic and event discovery. How can traditional information retrieval and LLM-based methods be used to identify and track topics and events in white and grey literature so as to reveal gaps?

Hybrid exploration. Expert judgment remains crucial for assessing literature that is meant to help inform policy-makers. Solely relying on human expert judgment introduces big risks in skewed assessments, as the volume of literature gets simply too large to be reviewed systematically. Hence, it makes sense to move forward with hybrid approaches, in order to meet the IPCC goals of comprehensive and unbiased assessments. How can we organize effective iterative human-in-the-loop (or machine-in-the-loop) approaches that alternate between query reformulation and retrieval steps and are likely to identify important thematic clusters and yield content-based insights?

Collaborative platforms. De-Gol et al. [2023] describe their experience with a collaborative technology platform, tailored to support an updated process of elaborating IPCC reports. Among other things, the platform, ScienceBrief, featured a living map of evidence that used motion and spatial reasoning through animation and physical simulation to visualize the scientific consensus around a topic. De-Gol et al. argue that such platforms could greatly enhance IPCC assessments by making them more open and accessible, further increasing transparency. What sort of content-based recommendation methods, based on ideas in Sections 3 and 4, could help to create engagement?

Exploration. How can modern information retrieval methods, in combination with visualization and semantic tools, support the discovery of publication networks, enable swift navigation and simplify subsequent literature synthesis?

5.3 Obstacles and Risks

Beyond the data, technological, and methodological challenges listed above, the uptake of outcomes from the research suggested in Section 2, 3, and 4 in climate impact assessments faces ethical and social challenges [Ghosh et al., 2022]. Ensuring that the increased use of advanced information retrieval technology promotes equity and justice is essential – often, the people most affected by climate change are the least involved in the creation of the technologies, or the assessments of climate change. Avoiding the creation of new inequalities, e.g., in terms of access to information, contribution to the assessment, usefulness of the IPCC products, or exacerbating existing ones, is a key consideration. Ensuring transparency and accountability in the use of new information retrieval technologies is crucial for building trust and legitimacy. This includes clearly documenting the methods, data, and assumptions used in the assessment. Finally, engaging stakeholders in the development and use of new information retrieval technologies is essential for ensuring that assessments are relevant and useful. This requires effective communication and collaboration.

6 Final Note

Climate change is real. The information retrieval community has a unique opportunity to inform and foster a collaborative ecosystem of researchers and practitioners that contribute to effective information retrieval solutions in support of research and decision-making to help address the reality of climate change impacts [Sietsma et al., 2024]. The agenda setting activities of the MANILA24 workshop were meant to do just that. Preparations for MANILA25, the second edition of the workshop, to be held at SIGIR 2025, are well under way, continuing the goal of developing, maintaining, and executing an effective research agenda for information retrieval for climate change impacts.

Acknowledgements

We would like to thank the organizers of SIGIR 2024 for hosting the MANILA24 workshop.

AAK was supported by the European Union's Horizon Europe research and innovation programme (101056306, IAM COMPACT). GD was funded by the Academy of Finland Digital Water Flagship Project, European Chist Era Waterline Project. LF was funded by the Old Dominion University Division of Student Engagement & Enrollment Services, Graduate Student Travel Award. MH was funded by the Dutch Research Council (024.004.022). SH acknowledges the support of the Australian Research Council Centre of Excellence for the Weather of the 21st Century (CE230100012). JK was supported by the Dutch Research Council (KICH3.LTP.20.006). TM acknowledges support by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy (DEAC02-06CH11357). TR acknowledges the support of the Dreams Lab, a collaboration between Huawei Finland, the University of Amsterdam, and the Vrije Universiteit Amsterdam. MdR was funded by the Dutch Research Council (024.004.022, NWA.1389.20.183, KICH3.LTP.20.006) and the European Union's Horizon Europe research and innovation program (101070212, FINDHR). HS is supported by SmartSat CRC. FS acknowledges the support of the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S, CE20010000) and SmartSat CRC. DS was funded by the Australian Research Council (DE200100064) and the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S, CE200100005). YT was funded by the U.S. National Science Foundation (2230034, 2120943, and travel award for SIGIR), Google.org's Impact Challenge for Climate Innovation Program, and OpenAI's Researcher Access Program.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

A Authors and Affiliations

This is a list of authors of this publication. Workshop organizers:

- Maarten de Rijke (University of Amsterdam, The Netherlands)
- Bart van den Hurk (IPCC WGII and Deltares Water Knowledge Institute, The Netherlands)
- Flora Salim (UNSW Sydney, Australia)

Others who contributed to this publication as well:

- Alaa Al Khourdajie (Imperial College London, UK and International Institute for Applied System Analysis (IIASA), Austria; workshop participant)
- Nan Bai (Delft University of Technology and Wageningen University and Research, The Netherlands; workshop participant)
- Renato Calzone (ILUSTRE/Tilburg University, The Netherlands; program committee member)
- Declan Curran (UNSW Sydney, Australia; workshop participant)
- Getnet Demil (University of Oulu, Finland; workshop participant)
- Lesley Frew (Old Dominion University, USA; workshop participant)
- Noah Gießing (FIZ Karlsruhe/MARDI, Germany; workshop participant)
- Mukesh Kumar Gupta (IPCC WGII and Singapore Management University, Singapore; workshop participant)
- Maria Heuss (University of Amsterdam, The Netherlands; program committee member)
- Sanaa Hobeichi (UNSW Sydney, Australia; program committee member)
- David Huard (Ouranos, Canada; workshop participant)
- Jingwei Kang (University of Amsterdam, The Netherlands; workshop participant)
- Ana Lucic (University of Amsterdam, The Netherlands; program committee member)
- Tanwi Mallick (Argonne National Laboratory, USA; workshop participant)
- Shruti Nath (AOPP/University of Oxford, UK; program committee member)
- Andrew Okem (IPCC, The Netherlands; workshop participant)
- Barbara Pernici (Politecnico di Milano, Italy; workshop participant)
- Thilina Rajapakse (University of Amsterdam, The Netherlands; workshop participant)
- Hira Saleem (UNSW Sydney, Australia; workshop participant)
- Harrisen Scells (University of Tübingen, Germany; workshop participant)
- Nicole Schneider (University of Maryland, USA; workshop participant)
- Damiano Spina (RMIT University, Australia; workshop participant)
- Yuanyuan Tian (Arizona State University, USA; workshop participant)
- Edmund Totin (IPCC, Benin; workshop participant)
- Andrew Trotman (University of Otago, New Zealand; workshop participant)
- Ramamurthy Valavandan (IPCC WGII, The Netherlands; workshop participant)
- Dereje Workneh (Howard University, USA; workshop participant)
- Yangxinyu Xie (University of Pennsylvania, USA; workshop participant)

B Process

This paper grew out of the half-day SIGIR 2024 workshop on Information Retrieval for Climate Impact [van den Hurk et al., 2024], which was held at SIGIR 2024 on July 18, 2024. The workshop was organized by BvdH, MdR, FS, who invited a number of researchers from the information retrieval and climate science domains to form the program committee. The workshop itself consisted of two parts. The first part was focused on gaining a better understanding of the information retrieval challenges faced by IPCC WGII, with four presentations on the topics underlying the four core sections of this paper (Sections 2–5) as well as eight presentations that were selected based

on an open call for contributions. The second part of the workshop consisted of a collaborative writing session, with four groups working on the topics underlying the four core sections of the paper, using "seed topics" contributed by members of the program committee. The workshop organizers used the seed topics, presentations, and outputs of the writing session to draft a first version of the paper, which was then shared with the author team for review and editing.

References

Richard J. Adams, Palie Smart, and Anne Sigismund Huff. Shades of grey: Guidelines for working with the grey literature in systematic reviews for management and organizational studies. *International Journal of Management Reviews*, 19(4):432–454, 2017.

Alaa Al Khourdajie. The role of artificial intelligence tools in climate change scientific assessments. *SSRN preprint 4747126*, 2024.

James Allan, Jaime Arguello, Leif Azzopardi, Peter Bailey, Tim Baldwin, Krisztian Balog, Hannah Bast, Nick Belkin, Klaus Berberich, Bodo von Billerbeck, Jamie Callan, Rob Capra, Mark Carman, Ben Carterette, Charles L. A. Clarke, Kevyn Collins-Thompson, Nick Craswell, W. Bruce Croft, J. Shane Culpepper, Jeff Dalton, Gianluca Demartini, Fernado Diaz, Laura Dietz, Susan Dumais, Carsten Eickhoff, Nicola Ferro, Norbert Fuhr, Shlomo Geva, Claudia Hauff, David Hawking, Hideo Joho, Gareth Jones, Jaap Kamps, Noriko Kando, Diane Kelly, Jaewon Kim, Julia Kiseleva, Yiqun Liu, Xiaolu Lu, Stefano Mizzaro, Alistair Moffat, Jian-Yun Nie, Alexandra Olteanu, Iadh Ounis, Filip Radlinski, Maarten de Rijke, Mark Sanderson, Falk Scholer, Laurianne Sitbon, Mark Smucker, Ian Soboroff, Damiano Spina, Torsten Suel, James Thom, Paul Thomas, Andrew Trotman, Ellen Voorhees, Arjen P. de Vries, Emine Yilmaz, and Guido Zuccon. Report from the third strategic workshop on information retrieval in Lorne (SWIRL 2018). *SIGIR Forum*, 52:34–90, June 2018.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

Nan Bai, Ricardo da Silva Torres, Anna Fensel, Tamara Metze, and Art Dewulf. Inferring climate change stances from multimodal tweets. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2467–2471. Association for Computing Machinery, 2024.

Krisztian Balog. *Entity-Oriented Search*, volume 10 of *Synthesis Lectures on Information Concepts, Retrieval, and Services*. Morgan & Claypool Publishers, 2018.

Jon Barnett and Saffron J. O'Neill. Minimising the risk of maladaptation. In *Climate Adaptation Futures*, chapter 7, pages 87–93. John Wiley & Sons, Ltd, 2013.

Lea Berrang-Ford, A. R. Siders, Alexandra Lesnikowski, Alexandra Paige Fischer, Max W. Callaghan, Neal R. Haddaway, Katharine J. Mach, Malcolm Araos, Mohammad Aminur Rahman Shah, Mia Wannewitz, Deepal Doshi, Timo Leiter, Custodio Matavel, Justice Issah Musah-Surugu, Gabrielle Wong-Parodi, Philip Antwi-Agyei, Idowu Ajibade, Neha Chauhan, William

Kakenmaster, Caitlin Grady, Vasiliki I. Chalastani, Kripa Jagannathan, Eranga K. Galappaththi, Asha Sitati, Giulia Scarpa, Edmond Totin, Katy Davis, Nikita Charles Hamilton, Christine J. Kirchhoff, Praveen Kumar, Brian Pentz, Nicholas P. Simpson, Emily Theokritoff, Delphine Deryng, Diana Reckien, Carol Zavaleta-Cortijo, Nicola Ulibarri, Alcade C. Segnon, Vhalinavho Khavhagali, Yuanyuan Shang, Luckson Zvobgo, Zinta Zommers, Jiren Xu, Portia Adade Williams, Ivan Villaverde Canosa, Nicole van Maanen, Bianca van Bavel, Maarten van Aalst, Lynée L. Turek-Hankins, Hasti Trivedi, Christopher H. Trisos, Adelle Thomas, Shinny Thakur, Sienna Templeman, Lindsay C. Stringer, Garry Sotnik, Kathryn Dana Sjostrom, Chandni Singh, Mariella Z. Siña, Roopam Shukla, Jordi Sardans, Eunice A. Salubi, Lolita Shaila Safaee Chalkasra, Raquel Ruiz-Díaz, Carys Richards, Pratik Pokharel, Jan Petzold, Josep Penuelas, Julia Pelaez Avila, Julia B. Pazmino Murillo, Souha Ouni, Jennifer Niemann, Miriam Nielsen, Mark New, Patricia Nayna Schwerdtle, Gabriela Nagle Alverio, Cristina A. Mullin, Joshua Mullenite, Anuszka Mosurska, Mike D. Morecroft, Jan C. Minx, Gina Maskell, Abraham Marshall Nunbogu, Alexandre K. Magnan, Shuaib Lwasa, Megan Lukas-Sithole, Tabea Lissner, Oliver Lilford, Steven F. Koller, Matthew Jurjonas, Elphin Tom Joe, Lam T. M. Huynh, Avery Hill, Rebecca R. Hernandez, Greeshma Hegde, Tom Hawxwell, Sherilee Harper, Alexandra Harden, Marjolijn Haasnoot, Elisabeth A. Gilmore, Leah Gichuki, Alyssa Gatt, Matthias Garschagen, James D. Ford, Andrew Forbes, Aidan D. Farrell, Carolyn A. F. Enquist, Susan Elliott, Emily Duncan, Erin Coughlan de Perez, Shaugn Coggins, Tara Chen, Donovan Campbell, Katherine E. Browne, Kathryn J. Bowen, Robbert Biesbroek, Indra D. Bhatt, Rachel Bezner Kerr, Stephanie L. Barr, Emily Baker, Stephanie E. Austin, Ingrid Arotoma-Rojas, Christa Anderson, Warda Ajaz, Tanvi Agrawal, and Thelma Zulfawu Abu. A systematic global stocktake of evidence on human adaptation to climate change. *Nature Climate Change*, 11(11):989–1000, November 2021.

Max Callaghan, Carl-Friedrich Schleussner, Shruti Nath, Quentin Lejeune, Thomas R. Knutson, Markus Reichstein, Gerrit Hansen, Emily Theokritoff, Marina Andrijevic, Robert J. Brecha, Michael Hegarty, Chelsea Jones, Kaylin Lee, Agathe Lucas, Nicole van Maanen, Inga Menke, Peter Pfleiderer, Burcu Yesil, and Jan C. Minx. Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nature Climate Change*, 11:966–972, 2021.

Max W. Callaghan, Jan C. Minx, and Piers M. Forster. A topography of climate change research. *Nature Climate Change*, 10:118–123, 2020.

Declan Conway, Robert J. Nicholls, Sally Brown, Mark G. L. Tebboth, William Neil Adger, Bashir Ahmad, Hester Biemans, Florence Crick, Arthur F. Lutz, Ricardo Safra De Campos, Mohammed Said, Chandni Singh, Modathir Abdalla Hassan Zaroug, Eva Ludi, Mark New, and Philippus Wester. The need for bottom-up assessments of climate risks and adaptation in climate-sensitive regions. *Nature Climate Change*, 9:503–511, 2020.

Chris Cooper, Andrew Booth, Jo Varley-Campbell, Nicky Britten, and Ruth Garside. Defining the process to literature searching in systematic reviews: A literature review of guidance and supporting studies. *BMC Medical Research Methodology*, 18(1):85, 2018.

Anthony Jude De-Gol, Corinne Le Quéré, Adam J. P. Smith, and Marianne Aubin Le Quéré. Broadening scientific engagement and inclusivity in IPCC reports through collaborative technology platforms. *npj Climate Action*, 2(1):49, 2023.

Julian H. Elliott, Anneliese Synnot, Tari Turner, Mark Simmonds, Elie A. Akl, Steve McDonald, Georgia Salanti, Joerg Meerpohl, Harriet MacLehose, John Hilton, David Tovey, Ian Shemilt, James Thomas, Thomas Agoritsas, John Hilton, Caroline Perron, Elie Akl, Rebecca Hodder, Charlotte Pestridge, Lauren Albrecht, Tanya Horsley, Joanne Platt, Rebecca Armstrong, Phi Hung Nguyen, Robert Plovnick, Anneliese Arno, Noah Ivers, Gail Quinn, Agnes Au, Renea Johnston, Gabriel Rada, Matthew Bagg, Arwel Jones, Philippe Ravaud, Catherine Boden, Lara Kahale, Bernt Richter, Isabelle Boisvert, Homa Keshavarz, Rebecca Ryan, Linn Brandt, Stephanie A. Kolakowsky-Hayner, Dina Salama, Alexandra Brazinova, Sumanth Kumbargere Nagraj, Georgia Salanti, Rachelle Buchbinder, Toby Lasserson, Lina Santaguida, Chris Champion, Rebecca Lawrence, Nancy Santesso, Jackie Chandler, Zbigniew Les, Holger J. Schünemann, Andreas Charidimou, Stefan Leucht, Ian Shemilt, Roger Chou, Nicola Low, Diana Sherifali, Rachel Churchill, Andrew Maas, Reed Siemieniuk, Maryse C. Cnossen, Harriet MacLehose, Mark Simmonds, Marie-Joelle Cossi, Malcolm Macleod, Nicole Skoetz, Michel Counotte, Iain Marshall, Karla Soares-Weiser, Samantha Craigie, Rachel Marshall, Velandai Srikanth, Philipp Dahm, Nicole Martin, Katrina Sullivan, Alanna Danilkewich, Laura Martínez García, Anneliese Synnot, Kristen Danko, Chris Mavergames, Mark Taylor, Emma Donoghue, Lara J. Maxwell, Kris Thayer, Corinna Dressler, James McAuley, James Thomas, Cathy Egan, Steve McDonald, Roger Tritton, Julian Elliott, Joanne McKenzie, Guy Tsafnat, Sarah A. Elliott, Joerg Meerpohl, Peter Tugwell, Itziar Etxeandia, Bronwen Merner, Alexis Turgeon, Robin Featherstone, Stefania Mondello, Tari Turner, Ruth Foxlee, Richard Morley, Gert van Valkenhoef, Paul Garner, Marcus Munafo, Per Vandvik, Martha Gerrity, Zachary Munn, Byron Wallace, Paul Glasziou, Melissa Murano, Sheila A. Wallace, Sally Green, Kristine Newman, Chris Watts, Jeremy Grimshaw, Robby Nieuwlaat, Laura Weeks, Kurinchi Gurusamy, Adriani Nikolakopoulou, Aaron Weigl, Neal Haddaway, Anna Noel-Storr, George Wells, Lisa Hartling, Annette O'Connor, Wojtek Wiercioch, Jill Hayden, Matthew Page, Luke Wolfenden, Mark Helfand, Manisha Pahwa, Juan José Yepes Nuñez, Julian Higgins, Jordi Pardo Pardo, Jennifer Yost, Sophie Hill, and Leslea Pearson. Living systematic review: 1. introduction—the why, what, when, and how. *Journal of Clinical Epidemiology*, 91:23–30, 2017.

Lesley Frew, Michael L. Nelson, and Michele C. Weigle. Retrogressive document manipulation of US federal environmental websites. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3762–3766, 2024.

Arunabha Ghosh, Nandini Harihar, and Prayank Jain. Co-development of technologies of the future. Technical report, Stockholm+50 background paper series. Stockholm Environment Institute, 2022.

Zeke Hausfather, Kate Marvel, Gavin A. Schmidt, John W. Nielsen-Gammon, and Mark Zelinka. Climate simulations: recognize the 'hot model' problem. *Nature*, 605:26–29, 2022.

Susanne Hempel, Stephanie L. Taylor, Nell J. Marshall, Isomi M. Miake-Lye, Jessica M. Beroes, Roberta Shanman, Michele R. Solloway, and Paul G. Shekelle. Evidence map of mindfulness. Technical report, Department of Veterans Affairs, 2014.

Lucy Hocking, Sarah Parkinson, Avery Adams, Emmanuel Molding Nielsen, Cecilia Ang, and Helena de Carvalho Gomes. Overcoming the challenges of using automated technologies for public health evidence synthesis. *BMC Public Health*, 28(45):2300183, 2023.

IPCC. *Climate Change 2022: Impacts, Adaptation and Vulnerability*. Cambridge University Press, 2022a. H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem and B. Rama, editors.

IPCC. Summary for policymakers. In H.-O. Pörtner, D.C. Roberts, E.S. Poloczanska, K. Mintenbeck, M. Tignor, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, and A. Okem, editors, *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 3–33. Cambridge University Press, 2022b.

IPCC, 2024. Working group II impacts, adaptation and vulnerability. https://www.ipcc.ch/working-group/wg2/, 2024.

Krzysztof Janowicz, Pascal Hitzler, Wenwen Li, Dean Rehberger, Mark Schildhauer, Rui Zhu, Cogan Shimizu, Colby Fisher, Ling Cai, Gengchen Mai, et al. Know, Know Where, KnowWhereGraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence. *AI Magazine*, 43(1):30–39, 2022.

Elphin Tom Joe, Sai Koneru, and Christine J. Kirchhoff. Assessing the effectiveness of GPT-4o in climate change evidence synthesis and systematic assessments: Preliminary insights. In *ClimateNLP 2024: ACL 2024 Workshop on Natural Language Processing meets Climate Change*. ACL, August 2024.

Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. CLEF 2019 technology assisted reviews in empirical medicine overview. In *Working Notes of CLEF 2019 — Conference and Labs of the Evaluation Forum*, volume 2380, page 250, 2019. URL https://ceur-ws.org/Vol-2380/paper_250.pdf.

Wojciech Kusa, Oscar E. Mendoza, Matthias Samwald, Petr Knoth, and Allan Hanbury. CSMeD: Bridging the dataset gap in automated citation screening for systematic literature reviews. *arXiv preprint arXiv:2311.12474*, 2023.

Nina Lansbury, Brad Moggridge, Sandra Creamer, Lillian Ireland, Lisa Buckley, Geoff Evans, Olivia Milsom, Gretta Pecl, and Vinnitta Mosby. Aboriginal and Torres Strait Islander Peoples' voices and engagement in the Intergovernmental Panel on Climate Change: Advice to inform the Australian Government towards IPCC Assessment Report 7. An independent report commissioned by the Australian Government (Department of Climate Change, Energy, the Environment and Water), Canberra. https://public-health.uq.edu.au/files/25162/IPCC-Voices-Report.pdf, 2023.

Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, Kūpono Duncan, Sergio Garzon, D. Fox Harrell, Peter-Lucas Jones, Kekuhi Kealiikanakaoleohaililani, Megan Kelleher, Suzanne Kite, Olin Lagon, Jason Leigh, Maroussia Levesque, Keoni Mahelona, Caleb Moses, Isaac ('Ika'aka) Nahuewai, Kari Noe, Danielle Olson, 'Ōiwi Parker Jones, Caroline Running Wolf, Michael Running Wolf, Marlee Silva, Skawennati Fragnito, and Hēmi Whaanga. Indigenous protocol and artificial intelligence position paper. Technical report, Aboriginal Territories in Cyberspace, Honolulu, HI, 2020. Edited by Jason Edward Lewis. English Language Version of "Ka'ina Hana 'Ōiwi a me ka Waihona 'Ike Hakuhia Pepa Kūlana" available at: https://spectrum.library.concordia.ca/id/eprint/990094/.

Yougang Lyu, Lingyong Yan, Zihan Wang, Dawei Yin, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. MACPO: Weak-to-strong alignment via multi-agent contrastive preference optimization. In *The Thirteenth International Conference on Learning Representations*, April 2025.

Tanwi Mallick, John Murphy, Joshua David Bergerson, Duane R. Verner, John K. Hutchison, and Leslie-Anne Levy. Analyzing regional impacts of climate change using natural language processing techniques. *arXiv preprint arXiv:2401.06817*, 2024.

Michael D. Mastrandrea, Christopher B. Field, Thomas F. Stocker, Ottmar Edenhofer, Kristie L. Ebi, David J. Frame, Hermann Held, Elmar Kriegler, Katharine J. Mach, Patrick R. Matschoss, Gian-Kasper Plattner, Gary W. Yohe, and Francis W. Zwiers. Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. Technical report, Intergovernmental Panel on Climate Change (IPCC), 2010. https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf.

Edgar Meij, Krisztian Balog, and Daan Odijk. Entity linking and retrieval. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1127, 2013.

Jan C. Minx, Max Callaghan, William F. Lamb, Jennifer Garard, and Ottmar Edenhofer. Learning about climate change solutions in the IPCC and beyond. *Environmental Science & Policy*, 77: 252–259, 2017.

Brian C. O'Neill, Maarten van Aalst, Zelina Zaiton Ibrahim, Lea Berrang-Ford, Suruchi Bhadwal, Halvard Buhaug, Delane Diaz, Katja Frieler, Matthias Garschagen, Alexandre K. Magnan, Guy Midgley, Alisher Mirzabaev, Adelle Thomas, and Rachel Warren. Key risks across sectors and regions. In H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, and B. Rama, editors, *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 2411–2538. Cambridge University Press, 2022.

Hans Pörtner, Debra C. Roberts, Helen Adams, Ibidun Adelekan, Carolina Adler, Rita Adrian, Paulina Aldunce, Elham Ali, Rawshan Ara Begum, Birgit Bednar-Friedl, Rachel Bezner Kerr,

Robbert Biesbroek, Joern Birkmann, Kathryn Bowen, Martina Angela Caretta, Jofre Carnicer, Edwin Castellanos, Tae Sung Cheong, Winston Chow, Guéladio Cissé, Susan Clayton, Andrew Constable, Sarah R. Cooley, Mark John Costello, Marlies Craig, Wolfgang Cramer, Richard Dawson, David Dodman, Jackson Efitre, Matthias Garschagen, Elisabeth Gilmore, Bruce Glavovic, David Gutzler, Marjolijn Haasnoot, Sherilee Harper, Toshihiro Hasegawa, Bronwyn Hayward, Jeffrey Hicke, Yukiko Hirabayashi, Cunrui Huang, Kanungwe Kalaba, Wolfgang Kiessling, Akio Kitoh, Rodel Lasco, Judy Lawrence, Maria Fernanda Lemos, Robert Lempert, Christopher Lennard, Debora Ley, Tabea Lissner, Qiyong Liu, Emma Liwenga, Salvador Lluch-Cota, Sina Loeschke, Simone Lucatello, Yong Luo, Brendan Mackey Katja Mintenbeck, Alisher Mirzabaev, Vincent Moeller, Mariana Moncassim Vale, Mike Morecroft, Linda Mortsch, Aditi Mukherji, Tero Mustonen, Michelle Mycoo, Johanna Nalau, Mark New, Andrew Okem, Jean Pierre Ometto, Brian O'Neill, Rajiv Pandey, Camille Parmesan, Mark Pelling, Patricia Fernanda Pinho, John Pinnegar, Elvira S. Poloczanska, Anjal Prakash, Benjamin Preston, Marie-Fanny Racault, Diana Reckien, Aromar Revi, Steven K. Rose, E. Lisa F. Schipper, Daniela Schmidt, David Schoeman, Rajib Shaw, Nicholas P. Simpson, Chandni Singh, William Solecki, Lindsay Stringer, Edmond Totin, Christopher Trisos, Yongyut Trisurat, Maarten van Aalst, David Viner, Morgan Wairiu, Rachel Warren, Philippus Wester, David Wrathall, and Zelina Zaiton Ibrahim. Technical summary. In H.-O. Pörtner, D.C. Roberts, E.S. Poloczanska, K. Mintenbeck, M. Tignor, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, and A. Okem, editors, *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 37–118. Cambridge University Press, 2022.

Thilina Rajapakse, Andrew Yates, and Maarten de Rijke. Simple transformers: Open-source for all. In *SIGIR-AP 2024: Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 209–215. ACM, December 2024.

Zac Roberts, Bronwyn Carlson, Sandy O'Sullivan, Madi Day, Jo Rey, Tristan Kennedy, Tetei Bakic, and Andrew Farrell. A guide to writing and speaking about Indigenous People in Australia. Technical report, Macquarie University, 2021. https://doi.org/10.25949/5tfk-5113.

Harrisen Scells. *Query Automation for Systematic Reviews*. PhD thesis, School of Information Technology and Electrical Engineering, The University of Queensland, 2021.

Carl-Friedrich Schleussner and Claire L. Fyson. Scenarios science needed in UNFCCC periodic review. *Nature Climate Change*, 10:272, 2020.

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron Wallace. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1387–1407. ACL, July 2023.

Anne J. Sietsma, James D. Ford, Max W. Callaghan, and Jan C. Minx. Progress in climate change adaptation research. *Environmental Research Letters*, 16(5):054038, April 2021.

Anne J. Sietsma, James D. Ford, and Jan C. Minx. The next generation of machine learning for tracking adaptation texts. *Nature Climate Change*, 14:31–39, 2024.

Mark Simmonds, Georgia Salanti, Joanne McKenzie, Julian Elliott, and On behalf of the Living Systematic Review Network. Living systematic reviews: 3. Statistical methods for updating meta-analyses. *Journal of Clinical Epidemiology*, 91:38–46, 2017.

Mark Stevenson and Reem Bin-Hezam. Stopping methods for technology-assisted reviews based on point processes. *ACM Trans. Inf. Syst.*, 42(3), December 2023.

Yubao Tang, Ruqing Zhang, Zhaochun Ren, Jiafeng Guo, and Maarten de Rijke. Recent advances in generative information retrieval. In *SIGIR 2024: 47th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3005–3008. ACM, July 2024.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

James Thomas, Anna Noel-Storr, Iain Marshall, Byron Wallace, Steven McDonald, Chris Mavergames, Paul Glasziou, Ian Shemilt, Anneliese Synnot, Tari Turner, Julian Elliott, and on behalf of the Living Systematic Review Network. Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology*, 91:31–37, 2017.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. In *SIGIR 2024: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940. ACM, July 2024.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, Evgeny Matusov, Mudar Yaghi, Mohammad Shihadah, Hermann Ney, Christian Dugast, Jonathan Dotan, and Daniel Erasmus. ClimateGPT: Towards AI synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*, January 2024.

Yuanyuan Tian, Wenwen Li, Lei Hu, Xiao Chen, Michael Brook, Michael Brubaker, Fan Zhang, and Anna K Liljedahl. Advancing large language models for spatiotemporal and semantic association mining of similar environmental events. *Transactions in GIS*, 29(1):e13282, 2025.

United Nations. About the IPCC. https://www.ipcc.ch/about/, 2024.

Bart van den Hurk, Maarten de Rijke, and Flora D. Salim. MANILA24: SIGIR 2024 workshop on information retrieval for climate impact. In *SIGIR 2024: 47th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3044–3046. ACM, July 2024.

Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. Can ChatGPT write a good boolean query for systematic review literature search? In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1426–1436. Association for Computing Machinery, July 2023.

Shuai Wang, Harrisen Scells, Shengyao Zhuang, Martin Potthast, Bevan Koopman, and Guido Zuccon. Zero-shot generative large language models for systematic review screening automation. In *Advances in Information Retrieval*, pages 403–420. Springer, April 2024.

Zihan Wang, Ziqi Zhao, Yougang Lyu, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. A cooperative multi-agent framework for zero-shot named entity recognition. In *The Web Conference 2025*. ACM, April 2025.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. ClimateBert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, October 2021.

Yangxinyu Xie, Bowen Jiang, Tanwi Mallick, Joshua David Bergerson, John K. Hutchison, Duane R. Verner, Jordan Branham, M. Ross Alexander, Robert B. Ross, Yan Feng, Leslie-Anne Levy, Weijie Su, and Camillo J. Taylor. WildfireGPT: Tailored large language model for wildfire analysis. *arXiv preprint arXiv:2402.07877*, February 2024.

Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. A human-machine collaborative framework for evaluating malevolence in dialogues. In *ACL-IJCNLP 2021: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, August 2021.