MANILA26: SIGIR 2026 Tutorial on Information Retrieval for Climate Change Impact

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

Flora D. Salim
University of New South Wales
Sydney, Australia
flora.salim@unsw.edu.au

Abstract

This three-hour tutorial uses climate change evidence synthesis as a high-stakes "stress test" for state-of-the-art IR and AI methods. Participants will explore why this domain, which is characterized by high levels of interdisciplinarity, multi-modal data (geospatial and time-series), and a lack of controlled vocabularies, challenges current state-of-the-art systems and active learning protocols. Through an architectural critique of agentic RAG systems and a deep dive into the "AMOC experiment," the tutorial distinguishes between automated literature review and expert scientific assessment. Participants will learn to diagnose specific retrieval failure modes, such as geographic coverage bias and the "tacit knowledge boundary," where human expertise remains essential. The tutorial concludes by framing a new research agenda: building unified, verifiable evidence pipelines that integrate bibliographic, geospatial, and multi-modal retrieval with rigorous provenance tracking.

CCS Concepts

• **Applied computing**; • **Information systems** → **Specialized information retrieval**; • **Computing methodologies** → *Intelligent agents*;

Keywords

Climate change impact, Systematic reviews, Deep research

ACM Reference Format:
Maarten de Rijke and Flora D. Salim. 2026. MANILA26: SIGIR 2026 Tutorial on Information Retrieval for Climate Change Impact. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3805712.3808642

1 Motivation

1.1 Background

Climate change impact refers to the specific consequences of global warming on human systems, ecosystems, and the economy, resulting from the intersection of physical hazards with societal vulnerability. It focuses on the "real-world" outcomes that occur when climate shifts disrupt the stability of natural and built environments. Such outcomes may include crop failures, health risks, and habitat loss [9].

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808642

The importance of climate change impact assessment lies in the shift from predictive physics (predicting how much the planet will warm) to societal consequence (understanding what that warming actually does to life on Earth). While climate modeling tells us the "what," impact assessment tells us the "so what." Climate change is not just about rising temperatures; it is about the intersection of a physical hazard with human vulnerability and exposure. Impact research quantifies how a 2°C increase translates into specific, non-linear failures: the collapse of a regional power grid, the total loss of a staple crop in a specific latitude, or the migration patterns of climate refugees [8, 11].

Impact data is the primary driver for adaptation and mitigation. Governments and corporations do not allocate billions of dollars based on global mean temperature; they do so based on localized impact evidence. If an evidence synthesis is flawed, due to retrieval bias or missed data dependencies, the resulting infrastructure may be built to the wrong specifications, leading to catastrophic "maladaptation" [10].

A central pillar of the field is determining whether a specific event (like a 2024 flood in the Mekong Delta) can be statistically attributed to human-induced climate change or if it falls within natural variability. This requires synthesizing decades of observational data with peer-reviewed literature. The accuracy of this synthesis determines legal liability, loss-and-damage payments between nations, and the scientific credibility of the IPCC [6].

Climate impacts are often characterized by "tipping points," thresholds where a small change in temperature leads to a massive, irreversible shift in an ecosystem (e.g., the dieback of the Amazon or the collapse of the AMOC [16]). Identifying these thresholds early requires a comprehensive "scanning" of the entire scientific horizon, making reliable information retrieval a matter of planetary security [7].

1.2 The Gap

The motivation for this tutorial is rooted in a critical gap. Current AI and IR methods are winning at "benchmarks" but failing at "assessment." While we can retrieve 100,000 papers (as seen in [3]), we struggle to determine if those papers are actually reliable enough to base a significant policy or an IPCC report on. This tutorial is motivated by four specific "domain resistances":

1.2.1 The High-Stakes Reliability Gap. In standard IR, a "relevant" result is often enough (Precision at k). In climate science, the cost of a "false positive" (basing a policy on a flawed, non-peer-reviewed, or data-dependent study) is catastrophic. The tutorial motivates participants to move from relevance to verifiability.

1.2.2 *The Interdisciplinary "Lexical Chaos".*

Climate impact is the ultimate "stress test" for retrieval because it lacks a controlled vocabulary (unlike Medicine/MeSH). A hydrologist, an economist, and an agronomist might all describe the same drought impact using entirely different terminology. The tutorial motivates the need for semantic, cross-disciplinary retrieval that does not rely on shared keywords.

1.2.3 *The "Hidden Evidence" Problem.*

The most vital climate evidence often is not available in the text. It is trapped in:

- Geospatial layers (NetCDF, GeoTIFF),
- Multi-modal figures (color-coded risk maps), and
- Grey literature (NGO and government reports).

A text-only RAG system is, by definition, blind to half the field. This tutorial motivates the shift toward multi-modal and geospatial-aware pipelines.

1.2.4 *The Data Dependency Trap.*

Fifty papers might all agree on an impact, but if they all used the same flawed sensor data, they represent a correlated error, not a scientific consensus. The tutorial motivates the development of systems that can perform provenance-tracking rather than just summary-counting.

2 Objectives

By the end of the three hours, participants should be able to:

(1) *Diagnose where and why standard IR methods fail* in the climate impact domain. Specifically, participants will be able to articulate cross-disciplinary citation-proximity failures, characterize the geographic coverage bias as a measurable retrieval system property (not just a corpus property), and identify which (active) learning assumptions break down under heterogeneous corpora.
(2) *Distinguish retrieval from synthesis, and review from assessment*. Specifically, participants will understand why document-level classification [3] is architecturally distinct from claim-level extraction, and why the review/assessment distinction in [2] constitutes a requirement specification for IR system design rather than a limitation caveat.
(3) *Evaluate an agentic retrieval system as an IR paper.* Participants will apply standard IR evaluation criteria (reproducibility, baseline specification, metric validity, corpus construction methodology) to a published AI-assisted assessment, and identify what evaluation would be required to validate reported efficiency claims.
(4) *Specify the non-textual evidence problem precisely and at the system design level.* Participants will understand why geospatial datasets, observational time series, and multi-modal document figures are not peripheral to climate impact evidence synthesis but are often where the primary evidence lives and where scientific disputes are actually fought out; describe the retrieval infrastructure required for each type (STAC for geospatial, repository APIs for time series, vision-language models for figures); and sketch what a unified pipeline integrating all three with bibliographic retrieval would need to do, including provenance tracking and relevance ranking across heterogeneous source types. Two compelling examples will be provided. (i) Multi-resolution climate modeling and weather forecasting, and its interaction with urban-level decision making, such as weather impact on the renewables and the grid. (ii) Scope 1, 2, and 3 carbon emission tracing from multimodal sources, including building level and corporate accounting and reports.
(5) *Identify the tacit knowledge boundary*. Participants will recognize the class of assessment judgements that require domain expertise not encoded in any retrievable corpus (correlated model errors, data provenance, methodological fitness for a given region and time scale), and explain why this boundary defines the limit of what text-based retrieval and synthesis systems can currently do regardless of their sophistication.
(6) *Formulate a tractable research contribution* at the intersection of IR methodology and climate impact evidence synthesis. Participants will be asked to zoom in on one of: a benchmark design with geographic stratification and claim-level annotation, a domain adaptation study on climate-specific citation pairs, a geographic bias measurement framework, a stopping criterion evaluation on heterogeneous corpora, or a claims-to-data provenance system integrating bibliographic and observational data.

3 Relevance

Building on the insights and connections established during the MANILA24 and MANILA25 workshops held at SIGIR 2024 and 2025, respectively, MANILA26 transitions to a tutorial format. This shift is designed to equip early-stage researchers with a foundational understanding of current challenges, empowering them to contribute directly to the field.

The most direct "competitor" for the MANILA26 tutorial is the ClimateNLP Workshop series, which has gained a permanent spot at the Annual Meeting of the Association for Computational Linguistics (ACL). ClimateNLP 2025 (ACL, Vienna) was the second iteration of the workshop. Its program explicitly targets several of the problems we have identified: automated fact-checking, environmental claim detection, and named entity mapping for spatial finance. ClimateNLP 2024 was the inaugural workshop and featured keynote speakers like Naomi Oreskes (Harvard), emphasizing the shift from "just models" to "trust and scientific consensus."

Outside of the "Big Three" NLP conferences, the Climate Change AI organization runs a specialized tutorial track (often at NeurIPS or ICLR). Recent tutorials include: (i) Question Answering over Sustainability Reports (NeurIPS 2024/25), which focuses on "information richness" and uncertainty quantification, similar to our focus on the reliability of retrieved evidence; (ii) Climate Policy Radar's Open Knowledge Graph: A technical deep-dive into the API layer and knowledge graph construction that aligns with parts of Hour 1 of our syllabus.

The MANILA26 tutorial is unique because it is:

- IR-centric: Most NLP tutorials focus on generation or summarization. Ours focuses on the retrieval architecture and provenance, which is the "bottleneck" for scientific validity.
- Assessment-focused: We are not just teaching how to build a bot. We are teaching the methodology of how to evaluate a system against the rigorous standards of an IPCC-style assessment.
- Multi-modal integration: We explicitly include geospatial (STAC) and multi-modal (figures/tables) evidence, whereas most NLP tutorials remain strictly text-based.

4 Format and Schedule

The tutorial will be delivered by two presenters. It will also include multiple brief, interactive experience reports from practitioners active in the area developing and/or deploying technology for climate change impact. Depending on availability and on facilities at the conference venue, these reports will be delivered on location, online, or through brief pre-recorded videos.

The tutorial will be organized in three blocks of one hour. The three hours are structured as a series of productive collisions between retrieval methodology on one side and the specific demands of climate impact evidence synthesis on the other. Participants will leave with a clearer picture of where their methods work, where these methods fail in non-obvious ways, and where the open problems are that require both their technical expertise and domain knowledge they will need to acquire.

Hour 1: The Climate Impact Evidence Landscape as a Retrieval Problem

1.1 Why Climate Impact Is a Hard Retrieval Domain (15 min). Climate impact retrieval presents a unique challenge for researchers because the corpus is radically interdisciplinary and lacks a controlled vocabulary, requiring systems to navigate vast, shifting synonym spaces across diverse fields like agronomy and atmospheric science. The field's dual structure, where high-stakes IPCC Assessment Reports coexist with primary research, means retrieval systems must account for how formal scientific evidence is curated and synthesized. Furthermore, critical data often resides in poorly indexed "grey literature" such as government and NGO reports, making comprehensive discovery far more difficult than in standard, more homogeneous IR benchmarks.

1.2 A Rapid Survey of Relevant Retrieval Infrastructure (30 min). Traditional citation databases like Web of Science and PubMed rely on formal Boolean logic and proprietary indexing policies, creating metadata-driven "confounds" that make cross-platform retrieval inconsistent. In contrast, OpenAlex offers a programmatic, knowledge-graph approach with a massive citation network and open API, providing the necessary infrastructure for scalable and reproducible retrieval research. Ultimately, moving beyond restrictive database interfaces to an API-first methodology, using tools like `pyalex` and pre-computed embeddings, is essential for developing modern, evaluable retrieval systems in the context of climate change impact.

> *Excursion* into non-textual evidence as a key source of climate impact evidence. How to integrate multi-modal figures, geospatial datasets, time series, and literature-based retrieval into a unified evidence pipeline?

1.3 Query Engineering as an Information Retrieval Problem (15 min). While technical Boolean retrieval is well-established, optimizing "string engineering" for systematic reviews remains a complex challenge focused on maximizing recall within manageable precision limits. Methodologies like the PECO framework provide a structured approach to this by decomposing research questions into thematic tiers, effectively framing the task as a massive, high-stakes query expansion problem [5, 12]. Finally, the strategic use of proximity operators offers a precise lexical alternative to both simple co-occurrence and rigid phrase matching, highlighting unique advantages that traditional Boolean systems still hold over neural retrieval models.

Hour 2: Where Current IR Methods Meet Domain Resistance

2.1 Dense Retrieval in Scientific Literature: What Works and What Doesn't (15 min). While SPECTER [4] and SciBERT [1] serve as primary baselines, their performance in climate impact retrieval is hindered by a citation-proximity objective that fails to bridge the gap between disparate disciplines like hydrology and economics. Furthermore, training data biases lead to denser, more accurate embedding representations for Western regions, resulting in significant retrieval "blind spot" for the highly vulnerable Global South. Consequently, a major research opportunity exists in domain adaptation, as current models like ClimateBERT focus on corporate disclosures rather than the specific, labeled scientific literature needed to correct these geographic and interdisciplinary failures [17].

2.2 Active Learning for Systematic Review Screening: Systems Perspective (30 min). Active learning for systematic reviews is a mature field, yet it faces significant systems-level challenges, particularly in defining statistically sound stopping criteria that bridge the gap between IR evaluation and real-world deployment. Platforms like ASReview demonstrate that while semantic embeddings (SBERT) generally outperform lexical methods (TF-IDF) in climate literature, the field's extreme vocabulary diversity allows traditional keyword-based models to remain surprisingly competitive. Ultimately, the lack of validated stopping heuristics for the heterogeneous evidence found in climate science, ranging from field trials to global simulations, represents a critical research gap for ensuring comprehensive evidence synthesis.

> *Practitioner experience report* on active learning in a large environmental systematic review, using multi-modal data.

2.3 Automated Claim Extraction: The Hard Open Problem (15 min). The core research frontier in climate impact retrieval lies in moving beyond document-level classification toward structured, claim-level evidence extraction that captures granular data like geographic scope, impact magnitude, and emissions scenarios. This requires a sophisticated multi-modal pipeline capable of integrating PDF layout analysis, table extraction, and figure caption parsing to synthesize quantitative data and text into a coherent record. Because current models like ClimateBERT lack this capability, the primary bottleneck remains a multi-disciplinary knowledge engineering challenge: defining annotation schemas that can resolve the linguistic and visual ambiguities inherent in complex scientific claims.

Hour 3: Agentic Systems for Evidence Synthesis

3.1 Agentic Retrieval Architecture: Current State and Open Problems (15 min). The architectural evaluation of agentic retrieval systems, such as Elicit, reveals that while they wrap sophisticated LLM planning and synthesis around traditional dense retrieval, they fundamentally inherit existing failures like geographic bias and interdisciplinary gaps. These "agentic wrappers" do not resolve underlying retrieval errors; instead, they risk amplifying them through stochastic query generation and multi-step reasoning chains that are difficult to audit. For IR researchers, this creates a significant reproducibility crisis, as the combination of evolving indices, non-deterministic model outputs, and unversioned production weights makes these systems nearly impossible to evaluate by traditional scientific standards.

3.2 The AMOC Case Study: An IR Evaluation of an Agentic Scientific Assessment (30 min). This session will be organized like a seminar. The specific frame for the audience is to treat [2] as a systems paper and subject it to the kind of evaluation critique they would apply to a retrieval system paper at SIGIR or ECIR. The paper describes a Gemini-based agentic environment that synthesized 79 papers through 104 human-AI revision cycles in approximately 46 person-hours, with 13 climate scientists, producing an assessment of AMOC stability, with self-reported efficiency gains of 3–17×.

> *Practitioner experience report* on time series data in scientific assessment.

3.3 Failure Modes, Open Problems, and the Research Agenda (15 min). The closing session emphasizes that verifiability must transcend simple grounding, requiring a multi-modal provenance chain that traces every synthesized claim back to specific data points in text, tables, or figures. Addressing systemic coverage bias requires moving beyond qualitative descriptions to develop formal evaluation frameworks that quantitatively measure the performance gap across different geographic and linguistic strata. Ultimately, the grand challenge is the technical integration of disparate evidence types, bibliographic text, geospatial layers, and observational time series, into a unified pipeline that can navigate the domain's inherent data heterogeneity without smoothing over scientific uncertainties.

5 Materials

Participants will be provided with the following materials: (i) slides, (ii) bibliography file, and (iii) list of key resources (datasets, APIs).

Acknowledgments

This research was (partially) supported by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, the European Union under grant agreement No. 101201510 (UNITE), and the ARC Training Centre for Whole Life Design of Carbon Neutral Infrastructure (IC230100015). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of their respective employers, funders and/or granting authorities.

References

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. doi:10.18653/v1/D19-1371

[2] Christian Buck, Levke Caesar, Michelle Chen Huebscher, Massimiliano Ciaramita, Erich M. Fischer, Zeke Hausfather, Özge Kart Tokmak, Reto Knutti, Markus Leippold, Joseph Ludescher, Katharine J. Mach, Sofia Palazzo Corner, Kasra Rafiezadeh Shahi, Johan Rockström, Joeri Rogelj, and Boris Sakschewski. 2026. AI-Assisted Scientific Assessment: A Case Study on Climate Change. *arXiv preprint arXiv:2602.09723* (February 2026).

[3] Max Callaghan, Jan C. Minx, Piers M. Forster, Christopher J. Smith, William F. Lamb, Sabine Fuss, Christoph von Stechow, Nicolas Koch, Susanne Kadner, Jérôme Hilaire, et al. 2021. Machine-learning-based Evidence and Attribution Mapping of 100,000 Climate Impact Studies. *Nature Climate Change* 11, 11 (2021), 966–972. doi:10.1038/s41558-021-01168-6

[4] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2270–2282. doi:10.18653/v1/2020.acl-main.207

[5] Collaboration for Environmental Evidence. 2013. Guidelines for Systematic Review and Evidence Synthesis in Environmental Management. Version 4.2. Environmental Evidence: https://www.environmentalevidence.org/Documents/Guidelines/Guidelines4.2.pdf.

[6] Kristie L. Ebi, Nicholas H. Ogden, Jan C. Semenza, and Alistair Woodward. 2020. Using Detection and Attribution to Quantify How Climate Change Is Affecting Health. *Health Affairs* 39, 12 (2020), 2168–2174. doi:10.1377/hlthaff.2020.01004

[7] James D. Ford, Robbert Biesbroek, Lea Berrang Ford, Felix Creutzig, Neal Haddaway, Sherilee Harper, Jan C. Minx, Mark New, Anne J. Sietsma, Carol Zavaleta-Cortijo, and Max Callaghan. 2025. Recommendations for Producing Knowledge Syntheses to Inform Climate Change Assessments. *Nature Climate Change* 15, 7 (2025), 698–708. doi:10.1038/s41558-025-02354-6

[8] Global Commission on Adaptation. 2019. *Adapt Now: A Global Call for Leadership on Climate Resilience.* Technical Report. Global Center on Adaptation and World Resources Institute, Rotterdam, The Netherlands. https://gca.org/reports/adapt-now-a-global-call-for-leadership-on-climate-resilience/

[9] IPCC. 2022. Summary for Policymakers. In *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, 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 (Eds.). Cambridge University Press, Cambridge, UK and New York, NY, USA, 3–33. doi:10.1017/9781009325844.001

[10] Alexandre K. Magnan, E. Lisa F. Schipper, Maxime Burkett, John Campbell, Katy Cremin, Neal J. Enright, Juan P. Guerschman, Mohamed Hamza, Margot Hurlbert, Patrick Kariuki, et al. 2016. Addressing the Risk of Maladaptation to Climate Change. *Wiley Interdisciplinary Reviews: Climate Change* 7, 5 (2016), 646–665. doi:10.1002/wcc.409

[11] Michael D. Morecroft, Simon Duffield, Mike Harley, James W. Pearce-Higgins, Nick J. B. Isaac, Lawrence S. Walker, Pippa K. Gillingham, Joseph Camfield, Leonardo Barbieri, Ian Dickie, Richard James, Richard J. Smithers, Kevin Anderson, Mary C. Bell, Tom Brereton, Stephen J. Brooks, Geoffrey J. Cary, Christopher Christie, Mike Christie, Annie Cornick, James Curson, Adriana De Palma, Tom Drage, John L. Edmondson, Harriet Gadsdon, James Gascoigne, G. Godi, Dave Goulson, Rachel Harris, Steven I. Higgins, Dave Hodgson, David Hollands, James Isaac, Peter Kettlewell, Robert A. King, Alex Langford, Nick A. Littlewood, Nicholas A. Macgregor, Philip J. K. McGowan, Tom H. Oliver, Tim O'Riordan, Mark Pullan, Nick Sotherton, Jerry R. Tallowin, Chris B. Thaxter, Juliet Vickery, David Walker, Andrew Walmsley, Samantha Westaway, Natalie Weyer, Julian Williams, and Robert J. Wilson. 2019. Measuring the Success of Climate Change Adaptation and Mitigation in Terrestrial Ecosystems. *Science* 366, 6471 (2019), eaaw9256. doi:10.1126/science.aaw9256

[12] Andrew S. Pullin and Gavin B. Stewart. 2006. Guidelines for Systematic Review in Conservation and Environmental Management. *Conservation Biology* 20, 6 (2006), 1647–1656. doi:10.1111/j.1523-1739.2006.00485.x

[13] Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix J. Weigdels, Anthony Isaacs, Jeroen Hu, Max van Haastrecht, Gijs Admiraal, et al. 2021. ASReview: Open Source Software for Efficient and Transparent Active Learning for Systematic Reviews. *Nature Machine Intelligence* 3, 2 (2021), 125–133. doi:10.1038/s42256-020-00287-7

[14] Bart van den Hurk, Maarten de Rijke, and Flora D. Salim. 2024. 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.* ACM, 3044–3046.

[15] Bart van den Hurk, Maarten de Rijke, and Flora D. Salim. 2025. MANILA25: SIGIR 2025 Workshop on Information Retrieval for Climate Impact. In *SIGIR 2025: 48th international ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 4192–4194.

[16] René M. van Westen, Henk A. Kliphuis, and Henk A. Dijkstra. 2024. Physics-based Early Warning Signal Shows that AMOC is on Tipping Course. *Science Advances* 10, 6 (2024), eadk1189. doi:10.1126/sciadv.adk1189

[17] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBERT: A Pretrained Language Model for Climate-Related Disclosures. *AAAI Conference on Artificial Intelligence (AAAI)* 36, 11 (2022), 11439–11447. doi:10.1609/aaai.v36i11.21396