Mercator Research Institute on
Global Commons and Climate Change gGmbH

Potsdam Institute for
Climate Impact Research

Mapping the Climate Literature

End-to-end processes: whole systems

ICASR 09.09.2024

Tim Repke

The situation

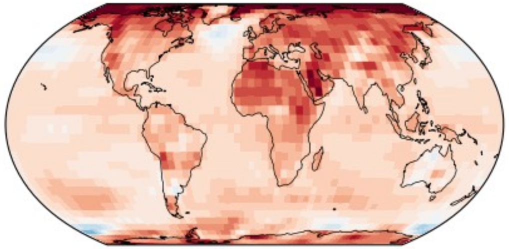
*not to scale

29.5kg
80k+
references

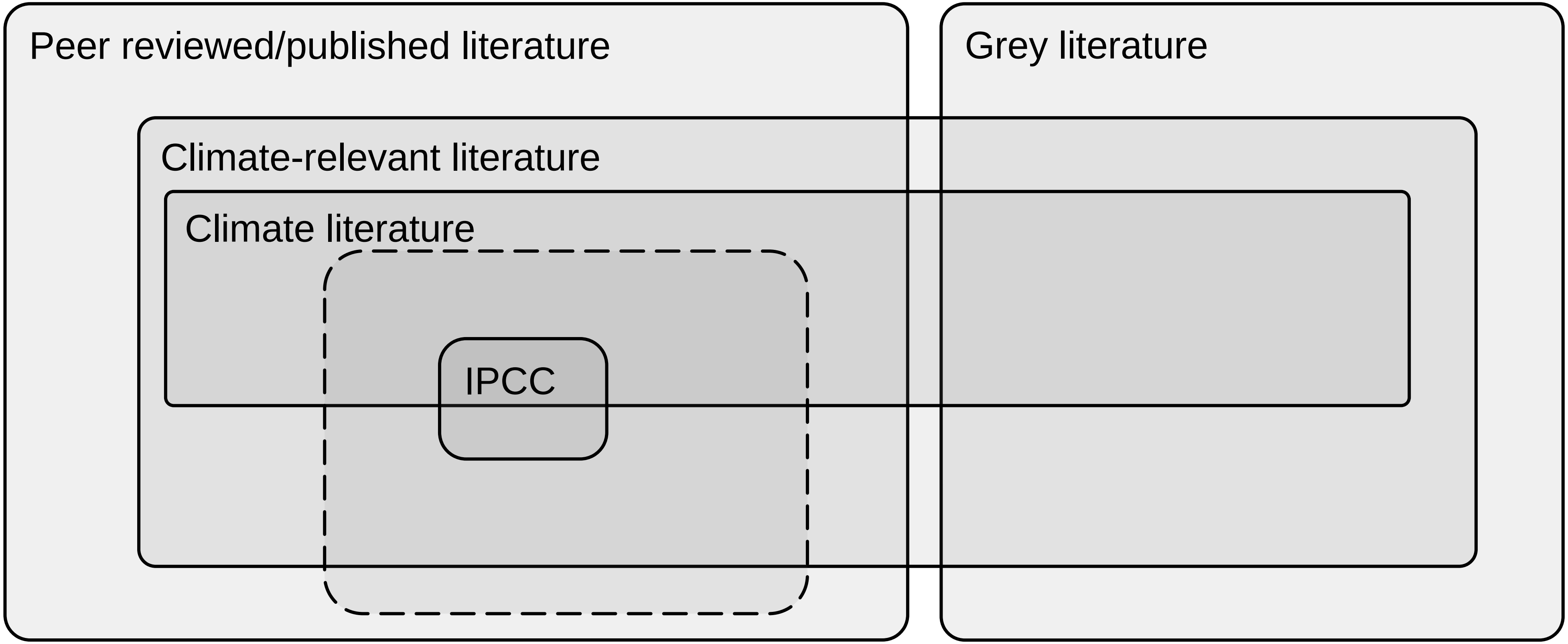
? ? ?

250M+ (?) “scientific” publications

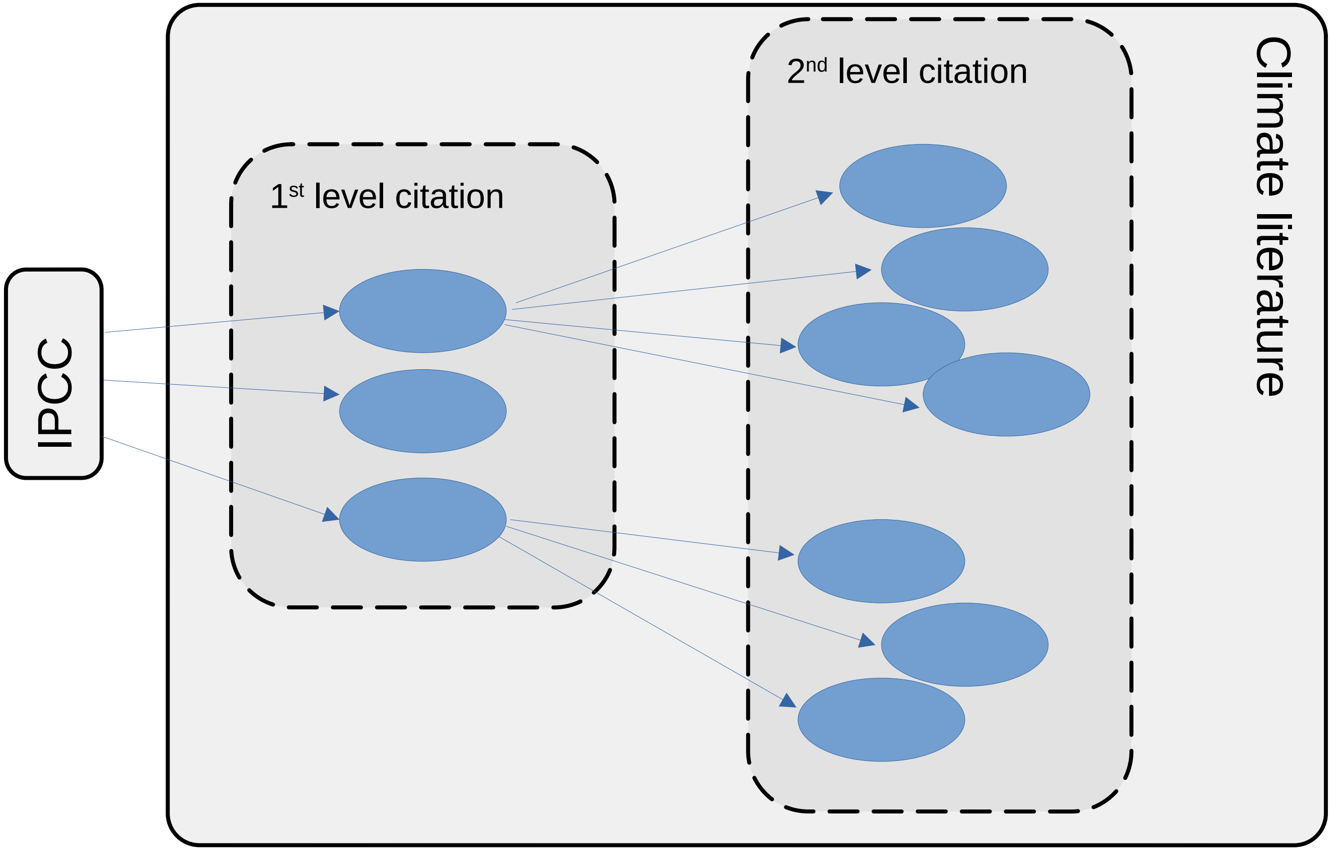
Mapping available evidence



The situation—refined

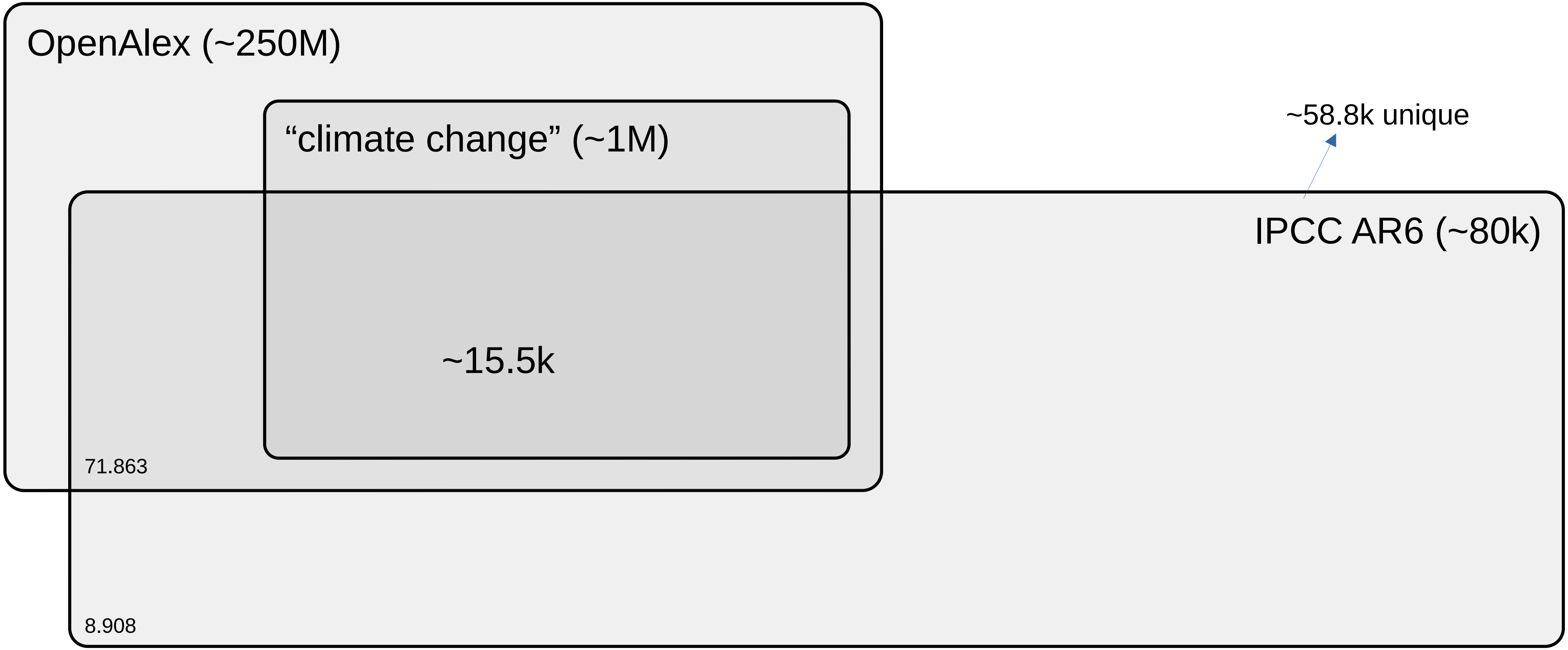


Evidence provenance and „indirect coverage"

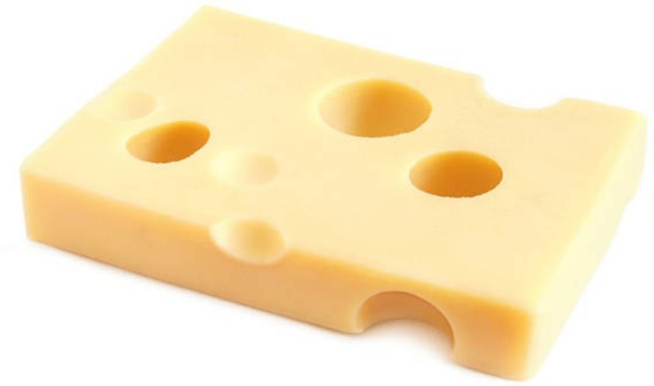



Literature Hub—Climate Policy Instruments
Home About License/Privacy
Explorer
Info
Filters
Number of documents: 78,401 / 78,401
Publication years
1990
1998
2006
2014
2022
Sector
10,612 / 10,612 AFOLU
6,447 / 6,447 Buildings
4,691 / 4,691 Industry
24,526 / 24,526 Energy
11,130 / 11,130 Transport
1,516 / 1,516 Waste
28,020 / 28,020 Cross-sectoral
Policy instrument
37,634 / 37,634 Economic
11,603 / 11,603 Regulatory
2,267 / 2,267 Information, education and training
32,628 / 32,628 Governance, strategies and targets
18,939 / 18,939 Agreements
Governance
655 / 655 Planning
31,180 / 31,180 Government administration & management
112 / 112 Institutions
Economic instrument
27,123 / 27,123 Carbon pricing
6,862 / 6,862 Subsidies
557 / 557 Non-carbon taxes
1,948 / 1,948 Direct Investment / spending
Full-text search
0 / 0
Search...
Scatterplot
0 / 0
human rights
cdm projects
forest carbon
negotiations
use change
change adaptation
adjustment
food waste
ets
clean air
waste manageme
ow carbon city
palm
loop supply
carbon intensity
environmental tax
energy efficient
energy sources

Numbers! (direct citations)



Some “gaps” are okay...



Identify gaps
- Evidence gaps
- Coverage gaps
- Scope(?)

Artifacts and “biased densities”
- Incomplete/inconsistent sources
- Artificially boosted source evidence

Literature Hub

https://climateliterature.org/

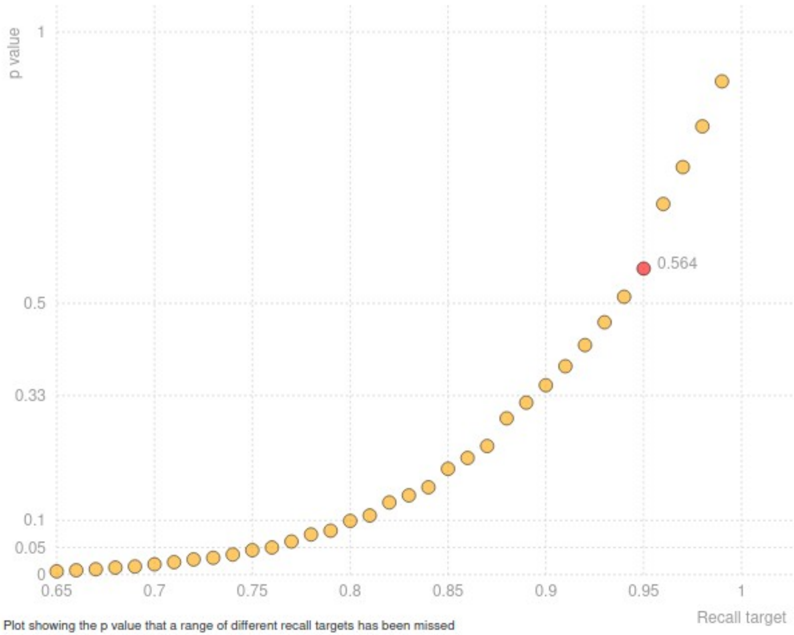
BUSCAR: Biased Urns for efficient Stopping Criteria in technology-Assisted Reviews

When we use machine learning to assist in screening documents for a systematic review, we often identify a majority of the relevant records before we have seen every record. However, if we want to save work, we need to stop early, and we have no way of knowing for sure how many relevant records we might have missed.

If we see a large number of irrelevant records in a row, this is a good sign that the proportion of remaining records that are relevant is low.

Before we use this intuition as a basis for stopping screening, we can consider the number of relevant records we have seen, as well as the number of records we have not yet screened, and calculate the implications of this estimated low prevalence for recall, or the proportion of relevant records we actually identify.

The calculator below does this on the basis of the stopping criteria documented in Callaghan and Müller-Hansen (2020). Please cite this paper if you use this calculator.



Plot showing the p value that a range of different recall targets has been missed

How many documents did your queries identify after deduplication?
2000

How many have you screened by hand?
1000

How many relevant records have you identified?
95

How many consecutive irrelevant records have you identified?
100

What level of recall (in %) would you like to achieve?
95

With p=**0.564** you have reached your desired recall target

This is based on assessing the chances of seeing **0** relevant records from a sample of **100** rec
without replacement from a population of **1100** documents that contained at least **6** relevant dc
calculated using the hypergeometric distribution function.

In a review, you could state "using the stopping criteria defined in Callaghan and Müller-Hanse
p=0.564, we reject the null hypothesis that we have not achieved our recall target of 95%".

Note that the hypergeometric distribution assumes that records are retrieved at random. In fact, the use of machine learning *generally* means we are more likely to retrieve relevant documents than random chance - this is why we use machine learning prioritisation after all. This means that, as long as machine learning is **no worse than random chance**, the stopping criterion will be **conservative**, meaning that if we stop with **p=0.05** in 100 different reviews, we would have stopped too early in **less than** 5% of cases. This is good, since we want to minimise the risk of missing relevant studies, but it does mean we often stop later than we could have done. We are currently working on ways to account for this non-random sampling procedure using biased urn theory.

In this online calculator, we can only calculate a p score based on the number of consecutive irrelevant records. In fact, we can calculate this for any sample of records containing 0 or more relevant records. In our paper, we calculate the score for all possible samples made up of sequences of previously screened records. You can implement this using our R package or Python package.

The Buscar app was built by Max Callaghan: Github

We are hiring!

MCC

Potsdam Institute for Climate Impact Research

PIK

You get tingly feelings when you get to evaluate digital evidence synthesis tools?

Come work for us in our [redacted] project funded by [redacted] !