UNIVERSITY OF LEEDS

Machine reading the science of climate change: computational tools to support evidence-based decision-making in the age of big literature

Max Callaghan

University of Leeds

School of Earth and Environment

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

December, 2021

Declaration of authorship

I confirm that the work submitted is my own, except chapters which constitute jointly-authored publications. My contribution, and the other authors to this work has been explicitly indicated below. I also confirm that appropriate credit has been given where reference has been made to the work of others.

Chapter 2 has been published as **Callaghan, M.W.**, Müller-Hansen, F.. Statistical Stopping Criteria for Automated Screening in Systematic Reviews. *Systematic Reviews.* **9** 273 (2020).
https://doi.org/10.1186/s13643-020-01521-4.

Max Callaghan designed the research and conducted the experiments. Finn Müller-Hansen contributed to the development of the statistical basis for the stopping criterion. Both authors wrote and edited the manuscript.

Chapter 3 has been published as **Callaghan, M.W.**, Minx, J.C. & Forster, P.M. A topography of climate change research. *Nat. Clim. Chang.* **10**, 118–123 (2020). https://doi.org/10.1038/s41558-019-0684-5.

Max Callaghan and Jan Minx designed the research. **Max Callaghan** performed the analysis. **Max Callaghan**, Jan Minx, and Piers Forster analysed the results. **Max Callaghan** wrote the manuscript with contributions from all authors.

Chapter 4 has been published as **Callaghan, M.W.,**, Schleussner, C.F., Nath, S, Lejeune, Q, Knutson, Thomas R., Reichstein, M. Hansen, G., Theokritoff, E., Andrijevic, M., Brecha, R., Hegarty, M., Jones, C., Lee,

K., Lucas, A., van Maanen, N., Menke, I., Pfleiderer, P., Yesil, B., Minx, J.C. AI based evidence and attribution mapping of 100,000 climate impact studies. *Nat. Clim. Chang.* **11**, 966–972 (2021) https://doi.org/10.1038/s41558-021-01168-6

M.C., J.C.M., and C-F.S. designed the research. **M.C.** developed the coding platform and machine learning pipeline to identify studies, with advice from M.R.. **M.C.**, C-F.S., G.H., Q.L., E.T. developed the codebook and coordinated screening and coding. **M.C.**, Q.L., S.N., C-F.S.. conceptualised the link to detection and attribution data. S.N. performed the univariate detection and attribution analysis of temperature and precipitation trends and assessment of internal variability, in consultation with T.R.K., who designed the methodology for these calculations. **M.C.** and S.N. designed and implemented the matching of studies with detection and attribution data. All other authors contributed to screening and coding studies. **M.C.**, C-F.S., J.C.M., Q.L., and S.N. wrote the manuscript with contributions from all authors.

Acknowledgements

I heartily thank the Heinrich Böll Stiftung for funding this PhD, and the Mercator Research Institute on Global Commons, the Priestley Center, and the School of Earth and Environment at the University of Leeds for providing stimulating and supportive environments for learning and collaboration. In particular I would like to thank my supervisors Jan Minx and Piers Forster for their support and feedback. I have been lucky to be at MCC in Jan's APSIS group since I joined as a student assistant in 2015. They have been five happy years and I have learnt a lot there. I am also indebted to all my co-authors in this and other projects. Finally I would like to thank Poppy Illsley who has been an invaluable editor, proof reader and companion, and the rest of my family and friends for creating the conditions for me to both start and finish this thesis.

Abstract

The amount of scientific literature on climate change has reached unmanageable proportions. This poses problems for researchers, especially those attempting to synthesise literature in the field. It is an even larger problem for the Intergovernmental Panel on Climate Change, whose task it is to comprehensively assess the scientific literature on climate change. This thesis explores how approaches from Natural Language Processing can be used to assist evidence synthesis, and understand and inform global environmental assessments. It uses computer assistance to ask what literature is relevant, and what it is about. First, it develops a methodology for machine learning assisted screening for systematic reviews. Second, it produces a map of the thematic content of the entire climate change literature. Finally, it uses machine learning to identify and classify tens of thousands of papers on climate impacts, and match these with model evidence on the attribution of climate trends to anthropogenic forcing.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Summary

This PhD uses the opportunities afforded by Natural Language Processing (NLP) to 1) understand the problems posed by big literature to the Intergovernmental Panel on Climate Change (IPCC), as well as evidence synthesis more generally, 2) advance synthetic knowledge on climate change, and 3) to demonstrate how NLP can assist IPCC and evidence synthesis processes. This chapter introduces the background and methods used in the thesis, and sets out the main research questions. These are answered in the following three chapters, which have been published as detailed in the declaration of authorship on page iii, and at the start of each chapter. Chapter 5 summarises the results of this thesis in context, and discusses the limitations of this approach, and sets out directions for future research.

1.1 Background

1.1.1 Big Literature

The volume of scientific publications has surpassed 200 million (Cambia, 2020). 15 million scholarly works were published in 2019 alone, amounting to more than 40,000 every day. Science is published in over 800,000 source titles using nearly four million keywords. Volume, velocity and variety are known as the 3 Vs of big data (Chen et al., 2014), and we apply these here to describe the phenomenon of "big literature" Nunez-Mir et al. (2016). Big literature means that scientists, policymakers and the public are overwhelmed by scientific publications. Gaining an overview of a topic or field is increasingly challenging. However, as is the case with big data, big literature also points towards a set of strategies for gaining new types of insights from scientific publications by employing computational techniques to process and analyse text as data. Scientific archives represent enormous accumulations of knowledge and effort. Machine reading the literature using NLP offers opportunities for benefiting from this knowledge in new ways.

Climate change is a relatively new subject in scientific literature. Although it has a history dating back to at least Arrhenius (1896), over half a million papers have been published since 1990 (99.3% of the total)[1], and growth in climate change research outpaces growth in scientific literature on the whole (Haunschild et al., 2016). Climate change is one of the most pressing challenges of our time. Meeting this challenge requires the mobilisation of vast amounts of scientific resources across all disciplines, dealing with the causes, consequences, and responses to climate change in the climate system, as well as in human and natural systems.

1.1.2 The IPCC

The Intergovernmental Panel on Climate Change (IPCC) sets out to assess the evidence on climate change "on a comprehensive, open and transparent basis". The assessment reports, which are published every 5-6 years, are mammoth undertakings which inform policymakers about the causes, consequences and potential responses to climate change. They involve thousands of authors and reviewers, and consider tens of thousands of publications. Despite the fact that the number of references in each report has increased

[1] Own calculations repeating the query in Callaghan et al. (2020)

from 1,600 in the first assessment report (FAR) in 1990 to 31,000 in the fifth assessment report (AR5) in 2014, the number of potentially relevant studies on climate change has increased from 1,500 to 110,000 in the same period. This means that ratio of IPCC citations to relevant publications has declined from 63% in the FAR, to 23% in AR5. In other words, big literature means an ever greater proportion of research on climate change is going uncited in IPCC reports.

When IPCC reports cite a smaller proportion of the relevant literature, the question of what they do and do not cite takes on a greater significance. Many researchers have investigated the strengths and weaknesses of the IPCC in reflecting the wider literature on climate change (Hulme and Mahony, 2010; Corbera et al., 2016; Bjurström and Polk, 2011). Much attention has been paid to claims that the IPCC privileges certain types of knowledge and is biased against the social sciences (Bjurström and Polk, 2011). Prominent commentaries have called for a greater role for the social sciences in IPCC reports (David G. Victor, 2015). Chapter 3 of this thesis combines machine reading with bibliometrics to investigate these claims with greater scrutiny. It provides a radical reassessment of prior ideas of IPCC bias, and uses topic modelling to understand the particular thematic content which is well or less well represented in IPCC reports.

By looking at past performance of the IPCC reports in representing the literature, Chapter 3 describes and demonstrates how NLP can contribute to the IPCC process itself. The topography of the literature we create can act as a guide to IPCC assessments, informing the process of defining an outline, and pointing to prominent themes and their interrelation before the reports are written. This process is already partially being taken up by Chapter 5 – on demand, services, and social aspects of mitigation – of working group three's contribution to the sixth assessment report. This is to be informed by a topic-model driven landscape of the literature (Creutzig et al., 2020).

1.1.3 Evidence Synthesis

The IPCC reports function as a form of evidence synthesis on a grand scale, but without formal methodology for study identification or synthesis. In evidence-based medicine though (and, increasingly, in other areas including the social and enviromental sciences), the process of evidence synthesis is much more strictly defined. In particular, systematic reviews and systematic maps offer formal methodologies designed to synthesise evidence on a given topic, while minimising bias and maintaining transparency (Haddaway and

Pullin, 2014; James et al., 2016; Higgins and Green, 2011).

A greater culture of evidence synthesis would help the IPCC to stay comprehensive, open, and transparent in the age of big literature: both in the assessment process itself, and in the production of climate-relevant meta-research (Ford et al., 2011). However, evidence synthesis is no panacea. Indeed, systematic reviews are also challenged by increasing amounts of literature. A key stage in evidence synthesis is the identification of relevant literature (Lefebvre et al., 2011), where researchers build a broad query to identify potentially relevant literature, before screening for relevance at the title then abstract level. Screening tens or even hundreds of thousands of titles and abstracts to identify a set of studies relevant to a research question becomes an onerous task. Systematic review practitioners with limited resources are faced with 3 options to deal with this challenge:

1. Limit the scope of the reviews undertaken, so that the pool of potentially relevant literature is smaller.

2. Develop more restrictive search queries, so that a greater proportion of screened studies are relevant.

3. Automate parts of this process using machine learning.

The third strategy is part an emerging area of research within the evidence synthesis community known as evidence synthesis technology. Chapter 2 of this thesis addresses a research gap within this field. Until now this gap has presented an insurmountable barrier to the use of automation in screening for systematic review in a way consistent with the principles of evidence synthesis. It defines statistical stopping criteria that allow researchers to set the maximum proportion of relevant studies they are prepared to miss, and the maximum acceptable probability of missing that many relevant studies.

1.1.4 Evidence synthesis technology beyond the identification of studies

There is mounting evidence that recent climate change is already impacting human and natural systems across the world. However, identifying the literature for each type of impact and in each world region in a systematic way has proved challenging. Understanding the role of anthropogenic climate change across multiple studies of climate

impacts is further complicated by the fact that not all studies discuss this role directly, although relevant evidence on this role may exist elsewhere. So far, attempts to assess the literature on human-attributable climate impacts have remained heuristic, and expert elicited, rather than systematic and comprehensive (Hansen and Stone, 2016).

Chapter 4 of this thesis is an AI-assisted evidence map of the literature on observed climate impacts. It develops a machine learning pipeline that not only identifies studies relevant to observed climate impacts, but also identifies the system and location being impacted, and the type of evidence provided. The database of studies assembled is synthesised in innovative ways with evidence from climate models such that plausible claims about the role of human influence on the climate in driving impacts can be made for more than 50% of the world's land area in over 23,000 studies.

1.2 Methods

1.2.1 Natural Language Processing

Natural Language Processing refers to a broad array of techniques by which computers perform tasks related to human language. It is a field with a longer history, but latterly three factors have led to huge advancements in the capabilities of statistical and latterly neural natural language processing (Manning and Schütze, 1999; Bengio et al., 2001; Mikolov et al., 2013a; Bahdanau et al., 2015; Devlin et al., 2019). Namely, the explosion of digital text archives that arrived with the world wide web, increases in computational power, and advances in deep learning. Applications are increasingly common in other fields of academic research (Tshitoyan et al., 2019; Porciello et al., 2020), and text archives are increasingly exploited by social scientists (Grimmer and Stewart, 2013), and in the field of energy and climate research (Müller-Hansen et al., 2020). This section gives a short introduction to some of the principles of NLP, before setting out the types of tasks for which NLP is used in this thesis.

NLP comprises a broad range of tasks from machine translation to question answering. In this section, I focus on the subfamily of tasks known as text classification, in which computers assign labels to texts in a way that can inform humans about the content of those texts or influence decisions about they should be processed (Minaee et al., 2020). For example, email services use text classification algorithms to identify emails that are likely to be spam, and send these to a separate folder, saving email

Text ID	Text	Spam
Text1	THIS CAN MAKE YOU EASY MONEY!	1
Text2	Can you send me the data?	0

Table 1.1: Two example texts

Text ID	this	can	make	you	easy	money	send	me	the	data
Text1	1	1	1	1	1	1	0	0	0	0
Text2	0	1	0	1	0	0	1	1	1	1

Table 1.2: Features for two example texts extracted using the "bag of words" model

users time and protecting them from the risk of fraud or exposure to viruses (Dada et al., 2019).

An algorithm in this sense is a function which takes an input, in this case a text, and returns an outcome, in this case the prediction **spam** or **not spam**. One apparently simple way to build such an algorithm is by defining rules which govern what outcome is returned.

```
def spamFilter(text):
  if "EASY MONEY" in text:
    return spam
  else:
    return notSpam
```

Subject experts could define logical rules that determine in what cases a document should be deemed spam. In the much simplified example above, documents containing the text "EASY MONEY" would be sent to the spam folder, while other documents would not. This type of rule-based algorithm development was common in the early days of Natural Language Processing (Brill and Mooney, 1997) but doing this requires a great deal of expert knowledge for each type of task, as the permutations of "spam-like" features are vast, and subject to change as spammers adapt.

Statistical natural language processing uses machine learning to *learn* how to return the right outcome. A model is *trained* with examples of texts which are labelled by humans as **spam** or **not spam**. This training process results in a model which can be used to predict the outcome for texts which have not seen before.

In order to do this, texts first have to be transformed into features, which encode

attributes of the text in numerical form. The simplest way of doing this is the so called "bag of words model", which counts the occurrence of each word in each text, as exemplified in tables 1.1 and 1.2. Table 1.2 is a numeric representation of the texts in table 1.1, or set of features X, which can be used as an input to a model, such that an outcome variable y is a function of X

$$y = f(X) \tag{1.1}$$

This training procedure aims to find a functional form (the exact way this is achieved varies according to the type of model used) that predicts the outcome variable according to the features which are associated or not associated with the outcome in our training data. In our simple case, we would find that the words "this", "make", "easy" and "money" would have a positive association with the outcome **spam**, while the words "send", "me", "the", and "data" would have a negative association. The words "can", and "you" would have neither a positive nor a negative association. With this model, we could make a prediction that the previously unseen text "Here's an easy way to make money" would be likely to be **spam**. Scaled up with thousands of examples, this type of learning can achieve good results in classifying texts.

Topic modelling

While in the spam detection example we know the labels we are interested in predicting, in some cases we may not *a priori* have a comprehensive list of labels, or we may not have labelled examples of text to train a model. In such cases we may employ *unsupervised* machine learning methods like **topic modelling**. This is especially useful when the scope of the set of documents is larger than can be comprehended by individual modellers and there may be substantively useful labels which are outside the expertise of modellers or which have emerged recently.

In topic modelling, the objective is to generate **topics**, which are distributions of **terms**. Each document is then a distribution of topics, and the objective of algorithmic approaches to topic modelling is to set the distributions of terms in topics and topics in documents such that their combination approaches the original distribution of **terms** in **documents**. In chapter 3, I apply topic modelling to a collection of over 400,000 abstracts of scientific papers about climate change. Figure 1.1 (reproduced from chapter 3) shows in the bottom right corner the "Document-Term Matrix", which corresponds to

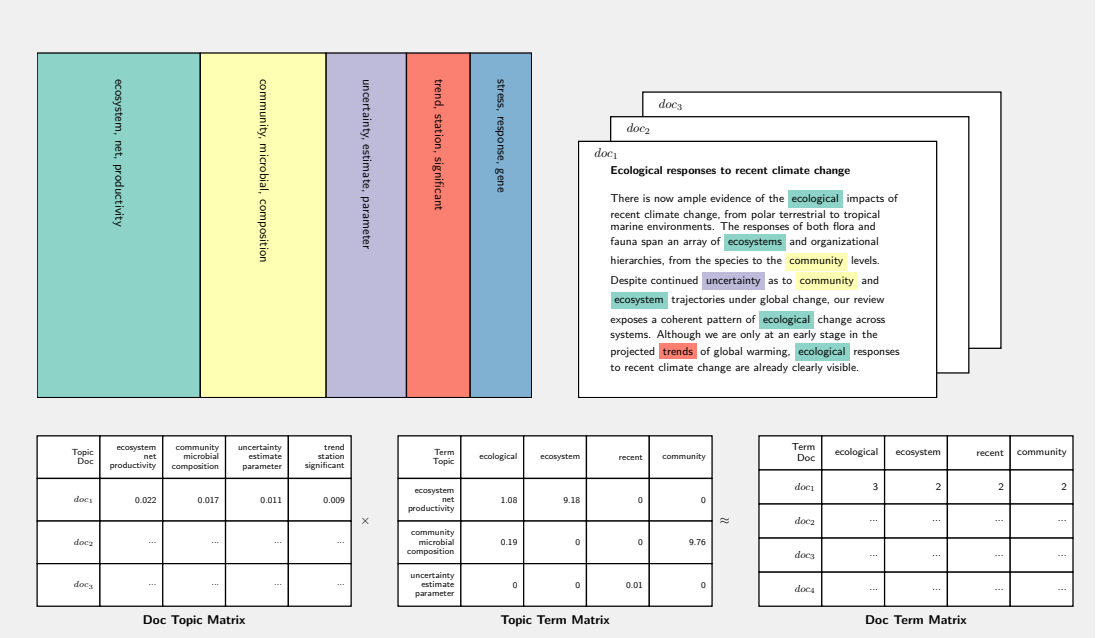


Figure 1.1: A figurative representation of a topic model

the representation of term frequencies in documents demonstrated in table 1.2. Because the number of unique terms in this corpus is close to 100,000, the Document-Term Matrix with almost 100,000 columns is not a digestible way to inform humans about the content of the documents. Instead, each document can be described as a mixture of topics, which in turn are each described as mixtures of words. In the example given, we can see that the document has the highest topic score for the topic that in turn has high scores for the words "ecosystem" and "ecological". Given that these words do indeed appear frequently in the document, we can see that the optimisation procedure has worked well. The word lists associated with the topics make them interpretable thematic features, and describing the documents by the mix of thematic features they contain provides a substantively useful description of the content of the document.

At a macro level, topic modelling allows us to make summarising statements about the proportion of documents dealing with different themes. Combining topic models with other metadata, we can say whether certain subsets of documents contain a greater proportion of documents on certain topics. In Chapter 3 I explore how topic distributions vary according to the age of documents, the discipline in which they were published, and whether or not they were cited by the IPCC. This generates insights that

inform our understanding of how the IPCC works, and highlights areas of research that may merit increased attention in future IPCC reports or increased funding allocations, particularly within the social sciences.

Dimensionality reduction

Dimensionality reduction refers to the techniques that transform data from a high- to low-dimensional space. The purpose is to preserve the properties of the data of the high-dimensional space and make them visible or interpretable in a low-dimensional space. Topic modelling is one form of dimensionality reduction, because it makes the intractable "Document-Term Matrix" intelligible by replacing the "Term" dimension with a much smaller "Topic" dimension. For visualisation purposes, it is often necessary to employ further dimensionality reduction techniques, in Chapter 3 this was done to reduce a 110-dimensional topic space to two dimensions that could be plotted or mapped.

Dimensionality reduction algorithms aim to "preserve as much of the significant structure of the high-dimensional data as possible in the low-dimensional map" (van der Maaten and Hinton, 2008). T-distributed stochastic neighbour embedding (t-SNE), which is used in this thesis, does this by converting a multidimensional dataset into a set of pairwise distances between points in the high-dimensional space. It then aims to find a low-dimensional representation of the data where documents likely to be neighbours in the original space are also likely to be neighbours in the mapped space. The method is "capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales" (van der Maaten and Hinton, 2008).

Richer numeric representations of text

In the examples given above, texts were represented by features encoding the frequency of individual words they contain. However, this way of representing texts cannot make use of similarities between words. For example, if a spammer was to replace the word "money" with the word "cash" in the sentence "This can make you easy money", a human would immediately see that the two sentences were almost identical. Our algorithm, not having seen the word "cash" in the training set, would not associate this word with an increased likelihood of being spam, as it would do with "money".

Word embeddings represent individual words as a n-dimensional numeric vectors, where similar words have similar vectors. So, in a 4-dimensional embedding space, the word "money" might be represented by the vector [0.5,0.2,1.2,3.5], and the word "cash" by the similar vector [0.4,0.2,1.3,3.6]. Word embeddings must be learned, and effective embeddings can be learnt by neural networks using large unlabelled text corpora (Mikolov et al., 2013b). The embeddings are learnt by predicting words as a function of the words in the immediate context (continuous bag of words architecture), or by predicting the context as a function of a word (skip-gram architecture).

Word embeddings can be used to represent texts in order to increase performance on downstream classification tasks. However, in early iterations they are not able to represent how the meaning of a word can be different depending on its context. For example, the word "bank" conveys a different meaning in the two following sentences.

1. "In the shade of the house, in the sunshine on the river **bank** by the boats, in the shade of the sallow wood and fig tree, Siddhartha, the handsome Brahmin's son, grew up with his friend Govinda.

2. "The typical modern Banking System consists of a Sun, namely the Central Bank, and Planets, which following American usage, it is convenient to call the Member Banks"

The human reader is immediately able to distinguish between the two meanings of bank based on the surrounding context, but traditional word embeddings provide one single vector for each word. Recent advances in machine learning have found ways to represent individual words that depends not only on the word itself but on its context (Peters et al., 2018). In the examples given above, the word "bank" would be encoded differently in each sentence.

Most recently, BERT (Bidirectional Encoder Representations from Transformers) type models, which use contextual embeddings, have advanced the state of the art across a wide range of tasks in NLP Devlin et al. (2019). BERT and similar models are pre-trained on massive text corpora on a "masked language model" task, where the individual words are randomly hidden from the input, and the objective is to predict the original word based on the context. In a further step, the model is trained to predict whether one sentence follows another given pairs of sentences that are consecutive 50% of the time and randomly selected 50% of the time. After these pre-training steps, the

model can be "fine-tuned" on downstream tasks, like classification, using labelled data. In this way, tasks with small amounts of data (1000s of labelled examples) can benefit from a rich language representation and modelling architecture that results from the pre-training procedure that uses hundreds of millions of words. Large language models like BERT currently represent the forefront of the field, but it has been noted that training such models requires computational resources that have non-trivial financial and environmental implications, and that they are subject to import limitations including the risk of harms deriving from bias and misinterpretation (Bender et al., 2021). They are therefore to be used and interpreted with great care.

Three tasks for NLP in the domain of climate science

In this thesis, I define 3 overarching tasks for NLP in understanding the science of climate change, driven by two questions: "What literature is relevant?" And "What is it about?" (Figure 1.2). The question "What is it about?" has two variants, depending on whether the thematic structure into which the texts are to be classified is known in advance or not. The distinction here is between unsupervised (topic modelling) and supervised (multilabel classifiers) machine learning approaches. Both of these overarching questions are asked in the two thematic chapters of this thesis (Chapters 3 and 4), while chapter 2 develops a methodology to better answer the first question. These questions, in their specific forms and with their specific subquestions are shown below.

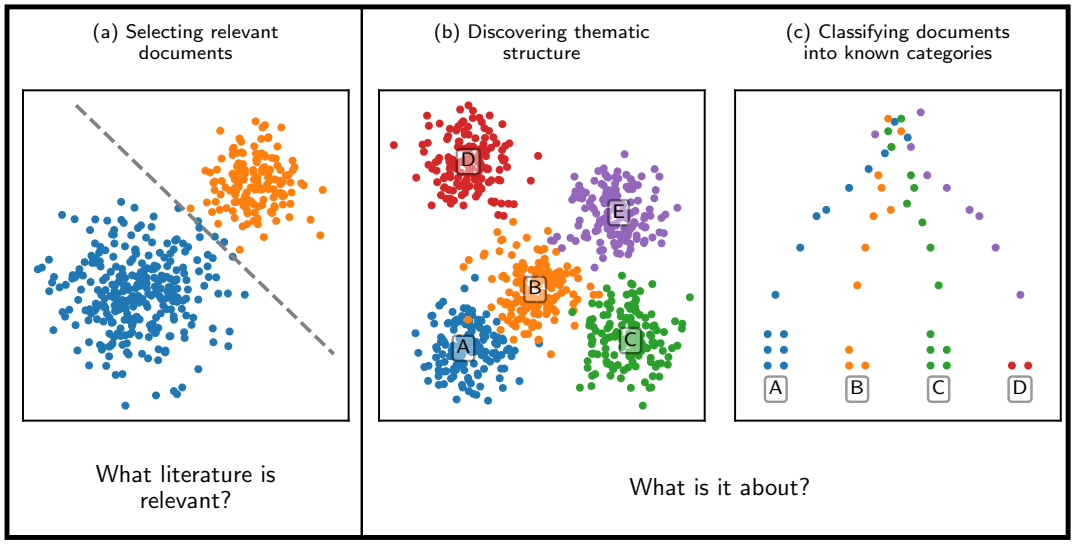


Figure 1.2: Three broad tasks in machine reading the science of climate change. In each case, dots represent documents, with their characteristics (relevant or not relevant, concerning category A or B, etc.) denoted by colour. In panels (a) and (b) documents are located in a notional 2-dimensional reduction of a multidimensional representation of text attributes (see section 1.2.1). Panel (c) describes process of sorting documents into pre-defined bins.

1.3 Research Questions

1. Chapter 2: "Statistical stopping criteria for automated screening in systematic reviews"

   (a) What literature is relevant?

   - How can we use machine learning to save time in identifying studies without sacrificing coverage?

2. Chapter 3: "A Topography of Climate Change Research"

   (a) What literature is relevant to climate change?

   - How has this grown?
   - How much literature is published in different disciplines?
   - What proportion of literature from each discipline is cited by the IPCC?

   (b) What is it about?

   - On what topics is climate literature published?
   - How are these topics related?
   - Which topics have grown recently?
   - What does the combination of topic, recent growth and discipline tell us about coverage of literature in the IPCC?

3. Chapter 4: "AI based evidence and attribution mapping of 100,000 climate impact studies"

   (a) What literature is relevant to observed impacts of climate change?

   - How has this grown?

   (b) What is it about?

   - What evidence is related to human and managed systems; terrestrial ecosystems; marine and coastal ecosystems; rivers, lakes and soil moisture; or mountains, snow and ice?
   - What type of evidence is provided? Does it attribute impacts to a trend, or merely establish sensitivity or a trend in climate variables?
   - What geographical entities do these studies give evidence on?

- What does the distribution of evidence tell us about anthropogenically attributable regional climate impacts in combination with grid cell level climate observations and model data?

In each case, the "what is it about" task serves to tag documents with meaningful, interpretable thematic categories. The distribution of these tags is then analysed with respect to either time, place, IPCC citation, or modelling evidence on anthropogenic climate change. Both approaches, and the ways in which the tagged documents are characterised and combined with other data, point to a set of strategies for new types of machine-learning-assisted evidence synthesis. Chapter 5 develops a tentative typology out of these strategies. Chapters 3 and 4 present early examples in developing machine-learning-assisted evidence syntheses.

Bibliography

Arrhenius, S. (1896). On the influence of carbonic acid in the air upon the temperature on the ground. *Philosophical Magazine*, 41:237–276.

Bahdanau, D., Cho, K. H., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Bengio, Y., Ducharme, R., and Vincent, P. (2001). A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 3:1137–1155.

Bjurström, A. and Polk, M. (2011). Physical and economic bias in climate change research: A scientometric study of IPCC Third Assessment Report. *Climatic Change*, 108(1):1–22.

Brill, E. and Mooney, R. J. (1997). An Overview of Empirical Natural Language Processing. *AI Magazine*, 18(4):13–13.

Callaghan, M., Minx, J. C., and Forster, P. (2020). A Topography of Climate Change Research. *Nature Clim. Change*, 10:118–123.

Cambia (2020). *Lens.org*. Accessed 2020-12-30.

Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209.

Corbera, E., Calvet-Mir, L., Hughes, H., and Paterson, M. (2016). Patterns of authorship in the IPCC Working Group III report. *Nature Climate Change*, 6(1):94–99.

Creutzig, F., Callaghan, M. W., Ramakrishnan, A., Javaid, A., Niamir, L., Minx, J. C., Müller-Hansen, F., Sovacool, B. K., Afroz, Z., Andor, M., Antal, M., Court, V., Diaz-Jose, J., Döbbe, F., Figueroa, M. J., Gouldson, A., Haberl, H., Hook, A., Ivanova, D., Lamb, W. F., Maizi, N., Mata, E., Nielsen, K. S., Onyige, C. D., Reisch, L. A., Roy, J., Scheelbeek, P., Sethi, M., Some, S., Sorrell, S., Tessier, M., Urmee, T., Virag, D., Wan, C., Wiedenhofer, D., and Wilson, C. (2020). A typology of 100,000 publications on demand, services and social aspects of climate change mitigation. *Environmental Research Letters*, In press.

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., and Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6):e01802.

David G. Victor (2015). Embed the social sciences in climate policy. *Nature*, 520:7–9.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 4171–4186.

Ford, J. D., Berrang-Ford, L., and Paterson, J. (2011). A systematic review of observed climate change adaptation in developed nations. *Climatic Change*, 106(2):327–336.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Haddaway, N. R. and Pullin, A. S. (2014). The Policy Role of Systematic Reviews: Past, Present and Future. *Springer Science Reviews*, 2(1-2):179–183.

Hansen, G. and Stone, D. (2016). Assessing the observed impact of anthropogenic climate change. *Nature Climate Change*, 6(5):532–537.

Haunschild, R., Bornmann, L., and Marx, W. (2016). Climate Change Research in View of Bibliometrics. *PLoS ONE*, 11(7):1–19.

Higgins, J. and Green, S., editors (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration, version 5. edition.

Hulme, M. and Mahony, M. (2010). Climate change: What do we know about the IPCC? *Progress in Physical Geography*, 34(5):705–718.

James, K. L., Randall, N. P., and Haddaway, N. R. (2016). A methodology for systematic mapping in environmental sciences. *Environmental Evidence*, 5(1):1–13.

Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M.-I., Noel-Storr, A., Rader, T., Shokraneh, F., Thomas, J., and Wieland, L. S. (2011). Cochrane Handbook for Systematic Reviews of Interventions. In Higgins, J. and Green, S., editors, *Cochrane Handbook for Systematic Reviews of Interventions*, chapter Chapter 4:. The Cochrane Collaboration, version 5. edition.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Distributed Representations of Words and Phrases and their Compositionality. pages 1–9.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013b). Efficient Estimation of Word Representations in Vector Space. *ICLR*, pages 1–12.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2020). Deep Learning Based Text Classification: A Comprehensive Review.

Müller-Hansen, F., Callaghan, M. W., and Minx, J. C. (2020). Text as big data: Develop codes of practice for rigorous computational text analysis in energy social science. *Energy Research and Social Science*, 70(July):101691.

Nunez-Mir, G. C., Iannone, B. V., Pijanowski, B. C., Kong, N., Fei, S., and Fitzjohn, R. (2016). Automated content analysis: addressing the big literature challenge in ecology and evolution. *Methods in Ecology and Evolution*, 7(11):1262–1272.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:2227–2237.

Porciello, J., Ivanina, M., Islam, M., Einarson, S., and Hirsh, H. (2020). Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning. *Nature Machine Intelligence*, 2(10):559–565.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Chapter 2

Statistical Stopping Criteria for Automated Screening in Systematic Reviews

Abstract

Active learning for systematic review screening promises to reduce the human effort required to identify relevant documents for a systematic review. Machines and humans work together, with humans providing training data, and the machine optimising the documents that the humans screen. This enables the identification of all relevant documents after viewing only a fraction of the total documents. However, current approaches lack robust stopping criteria, so that reviewers do not know when they have seen all or a certain proportion of relevant documents. This means that such systems are hard to implement in live reviews. This paper introduces a workflow with flexible statistical stopping criteria, which offer real work reductions on the basis of rejecting a hypothesis of having missed a given recall target with a given level of confidence. The stopping criteria are shown on test datasets to achieve a reliable level of recall, while still providing work reductions of on average 17%. Other methods proposed previously are shown to provide inconsistent recall and work reductions across datasets.

2.1 Background

Evidence synthesis technology is a rapidly emerging field that promises to change the practice of evidence synthesis work Westgate et al. (2018). Interventions have been proposed at various points in order to reduce the human effort required to produce systematic reviews and other forms of evidence synthesis. A major strand of the literature works on screening: the identification of relevant documents in a set of documents whose relevance is uncertain O'Mara-Eves et al. (2015). This is a time consuming and repetitive task, and in a research environment with constrained resources and increasing amounts of literature, this may limit the scope of the evidence synthesis projects undertaken. Several papers have developed Active Learning (AL) approaches Miwa et al. (2014); Wallace et al. (2010b,a); Jonnalagadda and Petitti (2013); Przybyła et al. (2018) to reduce the time required to screen documents. This paper sets out how current approaches are unreliable in practice, and outlines and evaluates modifications that would make AL systems ready for live reviews.

Active learning is an iterative process where documents screened by humans are used to train a machine learning model to predict the relevance of unseen papers Settles (2009). The algorithm chooses which studies will next be screened by humans, often those which are likely to be relevant or about which the model is uncertain, in order to generate more labels to feed back to the machine. By prioritising those studies most likely to be relevant, a human reviewer most often identifies all relevant studies – or a given proportion of relevant studies (described by recall: the number of relevant studies identified divided by the total number of relevant studies) – before having seen all the documents in the corpus. The proportion of documents not yet seen by the human when they reach the given recall threshold is referred to as the work saved. This represents the proportion of documents that they do not have to screen, which they would have had to without machine learning.

Machine learning applications are often evaluated using sets of documents from already completed systematic reviews for which inclusion or exclusion labels already exist. As all human labels are known *a priori*, it is possible to simulate the screening process, recording when a given recall target has been achieved. In live review settings, however, recall remains unknown until all documents have been screened. In order for work to really be saved, reviewers have to stop screening while uncertain about recall. This is particularly problematic in systematic reviews because low recall increases the

risk of bias Lefebvre et al. (2011). The lack of appropriate stopping criteria has therefore been identified as a research gap Bannach-Brown et al. (2019); Marshall and Wallace (2019), although some approaches have been suggested. These have most commonly fallen into the following categories:

- **Sampling criteria:** Reviewers estimate the number of relevant documents by taking a random sample at the start of the process. They stop when this number, or a given proportion of it, has been reached Shemilt et al. (2014)
- **Heuristics:** Reviewers stop when a given number of irrelevant articles are seen in a row Jonnalagadda and Petitti (2013); Przybyła et al. (2018).
- **Pragmatic criteria:** Reviewers stop when they run out of time Miwa et al. (2014).
- **Novel automatic stopping criteria:** Recent papers have proposed more complicated novel systems for automatically deciding when to stop screening Yu and Menzies (2019); Di Nunzio (2018); Howard et al. (2020)

We review the first three classes of these methods in the following section and discuss their theoretical limitations. They are then tested on several previous systematic review datasets. We demonstrate theoretically and with our experimental results, that these three classes of methods can not deliver consistent levels of work savings or recall - particularly across different domains, or datasets with different properties O'Mara-Eves et al. (2015). We also discuss the limitations of novel automatic stopping criteria, which have all demonstrated promising results, but do not achieve a given level of recall in a reliable or reportable way. Without the reliable or reportable achievement of a desired level of recall, deployment of AL systems in live reviews remains challenging.

This study proposes a system for estimating the recall based on random sampling of remaining documents. We use a simple statistical method to iteratively test a null hypothesis that the recall achieved is less than a given target recall. If the hypothesis can be rejected, we conclude that the recall target has been achieved with a given confidence level and screening can be stopped. This allows AL users to predefine a target in terms of uncertainty and recall, so that they can make transparent, easily communicable statements like "We reject the null hypothesis that we achieve a recall of less than 95% with a significance level of 5%".

In the remainder of the paper, we first discuss in detail the shortcomings of existing stopping criteria. Then, we introduce our new criteria based on a hypergeometric test. We evaluate our stopping criteria, and compare their performance with heuristic and sampling based criteria on real-world systematic review datasets on which AL systems have previously been tested Cohen et al. (2006); Yu and Menzies (2019); Terasawa et al. (2009); Castaldi et al. (2009).

2.2 Methods Review

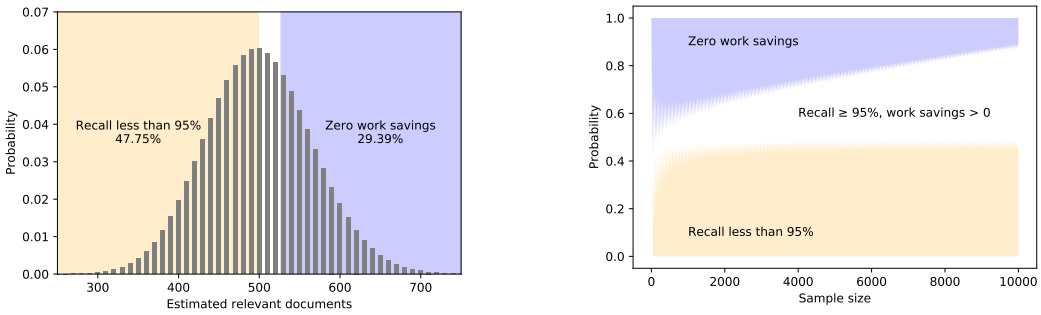
We start by explaining the sampling and heuristic based stopping criteria and discussing their methodological limitations.

2.2.1 Sampling Based Stopping Criteria

The stopping criterion suggested by Shemilt et al. Shemilt et al. (2014) involves establishing the Baseline Inclusion Rate (BIR), by taking a random sample at the beginning of screening. The BIR is used to estimate the number of relevant documents in the whole dataset. Reviewers continue to screen until this number, or a proportion of it corresponding to the desired level of recall, is reached.

However, the estimation of the BIR fails to correctly take into account sampling uncertainty [1]. This uncertainty is crucial, as errors can have severe consequences. Let us assume that users will stop screening when they have identified 95% of the relevant number of documents. If the estimated number of relevant documents is more than the true number of relevant documents divided by 0.95, then the users will never see 95% of the estimated number. This means that they will keep screening until they have seen all documents, and no work savings will be achieved. Conversely, if the number of relevant documents is underestimated by even a single unit, then the recall achieved will be lower than the target.

[1] Although Shemilt et al. Shemilt et al. (2014) employ a method to choose a sample size based on uncertainty, they fail to acknowledge the potential implications for recall of their choice. Their margin of error of 0.0025 and observed proportion of relevant studies of 0.0005 translate to estimates of 400 ± 451 relevant results. To reduce the margin of error to $\pm 5\%$ of estimated relevant studies, they would have had to screen 638,323 out of 804,919 results. See the notebook https://github.com/mcallaghan/rapid-screening/blob/master/analysis/bir_theory.ipynb that accompanies this paper for a detailed discussion.



(a) Error types after 1,000 documents

(b) Error types across sample sizes

Figure 2.1: Distribution of under- or over-estimation errors using the BIR sampling method in a dataset of 20,000 documents of which 500 are relevant. Panel (a) shows the probability distribution of the estimated number of relevant documents after a sample of 1,000 documents. Panel (b) shows the probability of each type of error according to the sample size.

The number of relevant documents drawn without replacement from a finite sample of documents follows the hypergeometric distribution. Figure 2.1a shows the distribution of the predicted number of documents after drawing 1,000 documents from a total of 20,000 documents, where 500 documents (2.5%) are relevant. The left shaded portion of the graph shows all the cases where the recall will be less than 95%. This occurs 48% of the time. The right shaded portion of the graph shows the cases where the number of relevant documents is overestimated so much that no work savings could be made to achieve a target recall of 95%. This occurs 29% of the time. In only 23% of cases can work savings be achieved while still achieving a recall of at least 95%.

Figure 2.1b shows the probability distribution of these errors according to the sample size. Even in very large samples both types of error remain frequent. This shows how baseline estimation inevitably offers poor reliability, either in terms of recall or in work saved.

Heuristic Stopping Criteria

Some studies give the example of heuristic stopping criteria based on drawing a given number of irrelevant articles in a row Jonnalagadda and Petitti (2013); Przybyła et al. (2018). We take this as a proxy for estimating that the proportion of documents remaining in the unseen documents is low, as the probability of observing 0 relevant

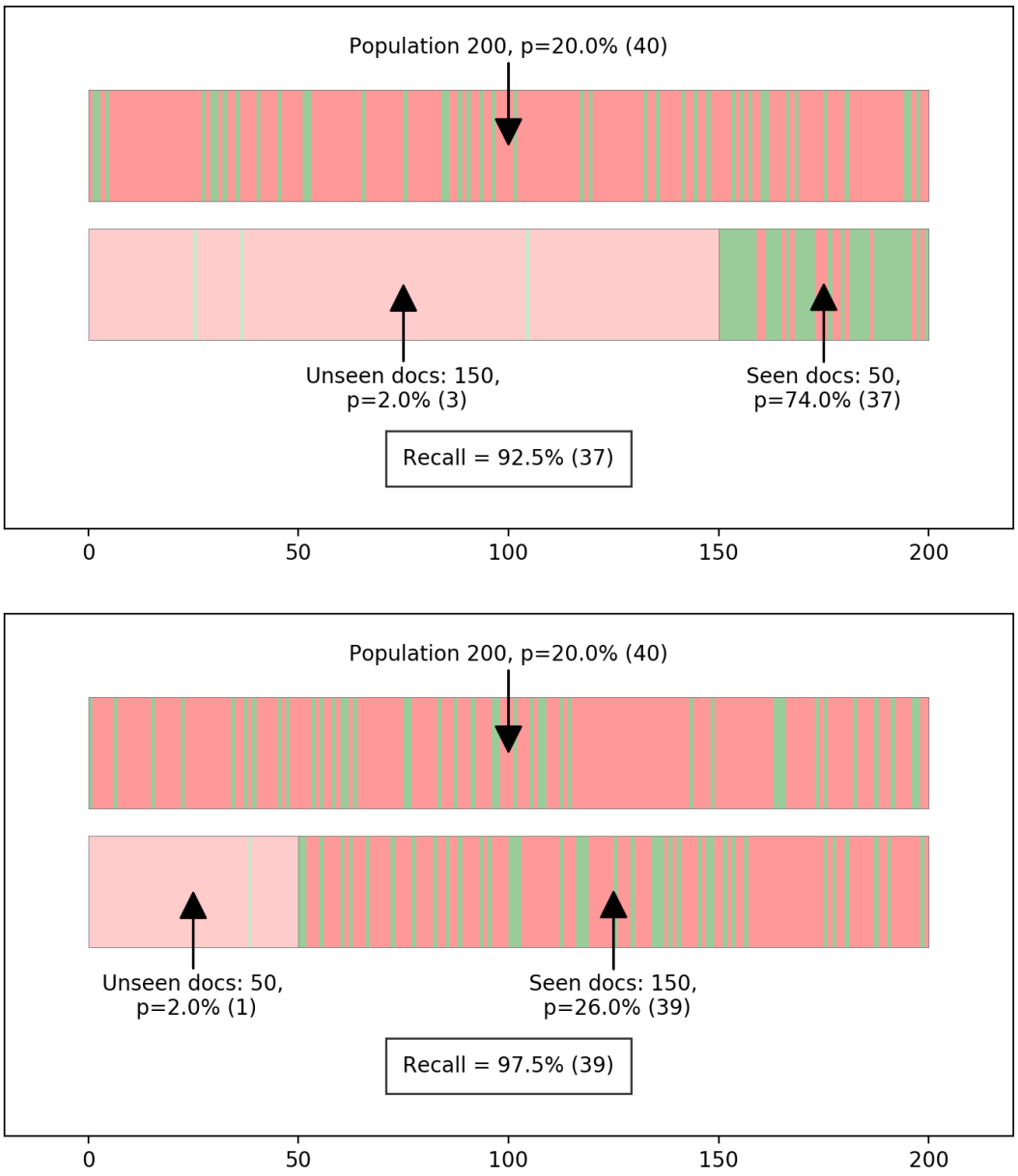


Figure 2.2: Similar low proportions of relevant documents in unseen documents with different consequences for recall. The top bar shows a random distribution of relevant documents (green) and irrelevant documents (red) at a given proportion of relevance. The bottom bar shows distributions of relevant and irrelevant documents in hypothetical sets of seen (right) and unseen (left - transparent) documents.

documents in a given sample (analogous to a set of consecutive irrelevant results) is a decreasing function of the number of relevant documents in the population. We find this a promising intuition, but argue that 1) it ignores uncertainty, as discussed in relation to the previous method; 2) it lacks a formal description that would help to find a suitable threshold for the criterion; and 3) it misunderstands the significance of a low proportion of relevant documents in estimating the recall.

Figure 2.2 illustrates this third point. We show two scenarios with identical low proportions of relevant documents observed in the unseen documents. In the top figure, machine learning (ML) has performed well, and 74% of the screened documents were relevant. In the bottom figure, ML has performed less well, and only 26% of the screened documents were relevant. In both cases, only 2% of unseen documents are relevant, but 2% of a larger number means more relevant documents are missed. Recall is not simply a function of the proportion of unseen relevant documents, but also of the number of unseen documents. This also means that where ML has performed well (as in the top figure), a low proportion of relevant documents in those not yet checked is indicative of lower recall than where ML has performed less well. Likewise, where the proportion of relevant documents in the whole corpus is low, a similarly low proportion of relevant documents is likely to be observed, even when true recall is low. This shows us that even a perfect estimator of the proportion of unseen relevant documents is insufficient on its own to provide sufficient information about when to stop screening. To estimate recall reliably, it is necessary to take into account the total number of unseen relevant documents (or their proportion times the number of unseen documents).

Pragmatic stopping criteria

Wallace et al. Wallace et al. (2010b) develop a "simple, operational stopping criterion": stopping after half the documents have been screened. Although the criterion worked in their experiment, it is unclear how this could be generalised, and its development depended on knowledge of the true relevance values. Jonnalagadda and Petitti Jonnalagadda and Petitti (2013) note that "the reviewer can elect to end the process of classifying documents at any point, recognizing that stopping before reviewing all documents involves a trade-off of lower recall for reduced workload", although clearly the reviewer lacks information about probable recall.

Novel automatic stopping criteria

Two examples come from the information retrieval literature. Di Nunzio Di Nunzio (2018) presents a novel automatic stopping criterion based on BM25, although recall reported is "often between 0.92 and 0.94 and consistently over 0.7". Yu and Menzies Yu and Menzies (2019) also present a stopping criterion based on BM25 which allows the user to target a specific level of recall. However, reviewers are not given the opportunity to specify a confidence level, and for two of the four datasets in which they tested their criteria, the median achieved recall at a stopping criteria targeting 95% recall was below 95%. In each case, the reliability of the estimate is dependent on the performance of the model.

Finally, Howard et al. Howard et al. (2020) present a method to estimate recall based on the number of irrelevant documents D observed in a list of documents since the δth previous relevant document. They reason that this should follow the negative binomial distribution based on the proportion of remaining relevant documents p, and use this information to estimate $\hat{p}$, and with this, the total number of relevant articles and the estimated recall.

However, their method does not quantify uncertainty, but can only claim that the method "*tends* to result in a conservative estimate of recall" (emphasis ours). This is not guaranteed by the criterion itself but rather a finding of the simulation with example datasets. Further, the authors do not give sufficient information to reproduce their results, providing neither code (they describe their own proprietary software), nor an equation for $\hat{p}$. Additionally, the criterion requires a tuning parameter δ, which users may have insufficient information to set optimally. Lastly, because screening is a form of sampling without replacement, the negative hypergeometric distribution should be preferred to the negative binomial, even though the latter can be a good approximation for cases with large numbers of documents.

These last examples are promising developments, but they all fail to take into account the needs of live systematic reviews, where the reliability of and ease of communication about recall are paramount, and the results are independent of model performance. In the following, we explain our own method, which provides clearly communicable estimates of recall, and which manage uncertainty in a way robust to model performance.

2.2.2 Methods

2.2.3 A Statistical Stopping Criterion for Active Learning

In our screening setup, we start off with N_{tot} documents that are potentially relevant. ρ_{tot} of these documents are actually relevant, but we don't know this value *a priori*. As we screen relevant documents we include them, so ρ_{seen} represents the number of relevant documents screened, and recall τ is given by

$$\tau = \frac{\rho_{seen}}{\rho_{tot}} \tag{2.1}$$

We set a target recall τ_{tar} and a confidence level α. We want to keep screening until $\tau \geq \tau_{tar}$, and devise a hypothesis test to estimate whether this is the case with a given level of confidence. We do this based on interrupting the active-learning process and drawing a random sample from the remaining unseen documents. We first describe this test, before showing how a variation on the test can be used to decide when to begin drawing a random sample.

Random Sampling

At the start of the sample, N_{AL} is the number of documents seen during the active learning process, and N is the number of documents remaining, so that

$$N = N_{tot} - N_{AL} \tag{2.2}$$

We refer to the number of relevant documents seen during active learning as ρ_{AL}, and the number of remaining relevant documents as K. We do not know the value of K but know that it is given by the total number of relevant documents minus the number of relevant documents seen during active learning.

$$K = \rho_{tot} - \rho_{AL} \tag{2.3}$$

We now take random draws from the remaining N documents, and denote the number of documents drawn with n and the number of relevant documents drawn with k. The number of relevant documents seen is updated by adding the number of relevant documents seen since sampling began to the number of relevant documents seen during active learning.

$$\rho_{seen} = \rho_{AL} + k \tag{2.4}$$

We proceed to form a null hypothesis that the true value of recall is less than our target recall:

$$H_0 : \tau < \tau_{tar} \tag{2.5}$$

Accordingly, the alternative hypothesis is that recall is equal to or greater than our target:

$$H_1 : \tau \geq \tau_{tar} \tag{2.6}$$

Because we are sampling without replacement, we can use the hypergeometric distribution to find out the probability of observing k relevant documents in a sample of n documents from a population of N documents of which K are relevant. We know that k is distributed hypergeometrically:

$$k \sim Hypergeometric(N, K, n) \tag{2.7}$$

We introduce a hypothetical value for K, which we call K_{tar}. This represents the minimum number of relevant documents remaining at the start of sampling compatible with our null hypothesis that recall is below our target.

$$K_{tar} = \lfloor \frac{\rho_{seen}}{\tau_{tar}} - \rho_{AL} + 1 \rfloor \tag{2.8}$$

This equation is derived by combining Eqs. 2.1 and 2.4. Because k can only take integer values, K_{tar} is the smallest integer that satisfies the inequality in Eq. 2.5. With K_{tar}, we can reformulate our null hypothesis: the true number of relevant documents in the sample is greater than or equal to our hypothetical value.

$$H_0 : K \geq K_{tar} \tag{2.9}$$

We test this by calculating the probability of observing k or fewer relevant documents from the hypergeometric distribution given by K_{tar}, using the cumulative probability mass function.

$$p = P(X \leq k), \text{ where } X \sim Hypergeometric(N, K_{tar}, n) \tag{2.10}$$

Because the cumulative probability mass function $P(X \leq k)$ is decreasing with increasing K, this gives the maximum probability of observing k for all values of K compatible with our null hypothesis. Similar arguments have been made to derive confidence intervals for estimating the parameter K in the hypergeometric distribution function (Buonaccorsi, 1987; Sahai and Khurshid, 1995) and the derivation of an equivalent criterion could use the upper limit of such a confidence interval of an estimated K from the observation of k.

We can reject our null hypothesis and stop screening if the maximum probability of obtaining our observed results given our null hypothesis p is below $1 - \alpha$ [1]. To further investigate the accuracy of the test, we perform an experiment drawing 1 million random samples in 6 scenarios with different characteristics. We vary the value of ρ_{AL} to simulate starting random sampling with different levels of recall achieved.

Figure 2.3 shows that in each case, as long as recall is lower than the target recall when sampling begins, the percentage of times that the criteria is triggered too early is within two tenths of a percentage point of 5% and the the 5th percentile of achieved recall values is within two tenths of a percentage point of the target recall 95%.

Ranked quasi-sampling

We now proceed to describe a special case of the method described above which we (1) use as a heuristic in order to decide when to begin random sampling; and (2) test as an independent stopping criterion. The method works by treating batches of previously screened documents as if they were random samples.

We calculate p as above for subsets of the already screened documents. Concretely, we use subsets of documents A_i by looking back to the last i documents, $A_i = \{d_{N_{seen}-1}, ..., d_{N_{seen}-i}\}$, where the documents d are indexed in the order in which they have been screened. For a specific i, this corresponds to random sampling beginning after seeing i documents in the section above. Thus, we set N_{AL} to i, n to $N_{seen}-i$, ρ_{AL} to the number of relevant documents seen when i documents had been seen, and k to the number of relevant documents seen since i documents had been seen, and calculate p according to Eq. 2.10. We compute p for all sets A_i with $i \in N_{seen} - 1 \ldots 1$. This

[1] The notebook, https://github.com/mcallaghan/rapid-screening/blob/master/analysis/hyper_criteria_theory.ipynb, in the github repository accompanying this paper contains a step by step explanation of this method with code and examples

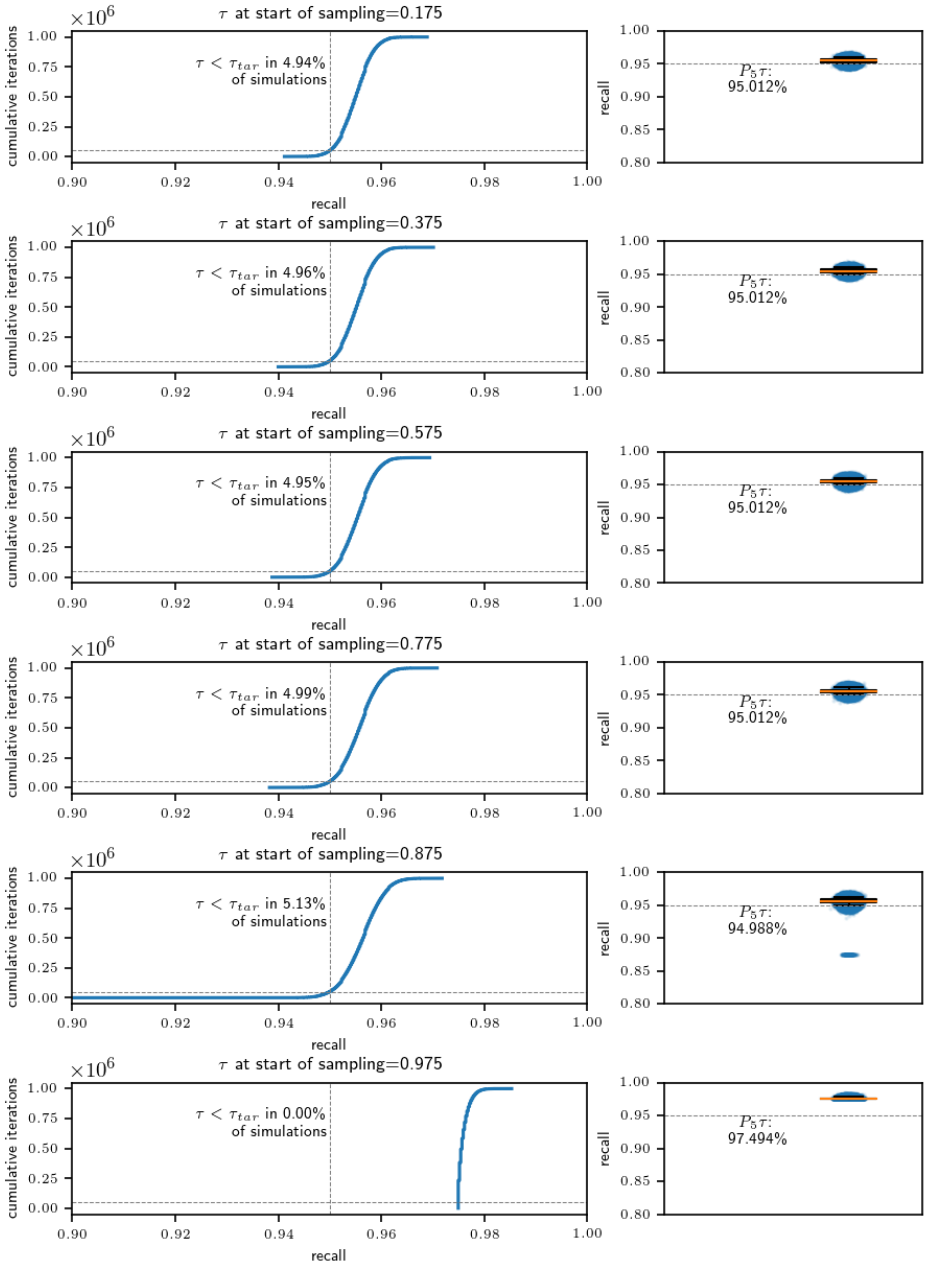


Figure 2.3: The distribution of achieved recall values given our random sampling stopping criterion for 6 scenarios with different recall values at the start of sampling.

gives us a vector $\boldsymbol{p}$, representing the values of p which would have been estimated at each point at which we could have stopped active learning and began random sampling. The point at which the p-value for our null hypothesis is lowest is given by p_{min}. With the vectorized implementation included in our accompanying code, these calculations are completed in less than the time it would take a human to code the next document.

First, we use this method as a useful heuristic for deciding when to stop active learning, and switch to random sampling. For this, we choose a higher threshold for the likelihood, $p_{min} < 1 - \frac{\alpha}{2}$. Second, we use the same ranked quasi-sampling as an independent stopping criterion, by continuing screening with active learning until $p_{min} < 1 - \alpha$. We present the results of this second procedure separately below.

Given that the documents seen during active learning are ranked according to predicted relevance, they do not in fact represent a random sample. This means that the test is unlikely to be accurate. It would be reasonable to assume that the proportion of relevant documents in each ranked quasi-sample is as high if not higher than the proportion of relevant documents in the unseen documents. This assumption would make this estimator conservative. As such it works in a similar way to the criterion proposed by Howard et al. Howard et al. (2020), although it makes use of more information and provides hypothesis testing rather than just a point estimate of recall.

2.2.4 Evaluation

We evaluate each of the criteria discussed on real world test data, operationalising the heuristic stopping criteria with 50, 100, and 200 consecutive irrelevant records. We run 100 iterations on each dataset and record the following measures.

- **Actual Recall**: The recall when the stopping criteria was met
- **WS-SC**: Work saved when the stopping criteria was met
- **Additional Burden**: the work saved when the criterion was triggered subtracted from the work saved when the recall target was actually achieved.

For simplicity, we use a basic SVM model Cortes and Vapnik (1995); Pedregosa et al. (2011), with 1-2 word n-grams taken from the document abstracts used as input data. We start with random samples of 200 documents (we do not employ Shemilt et al's methods for identifying the "optimal" sample size, as we showed these in the methods

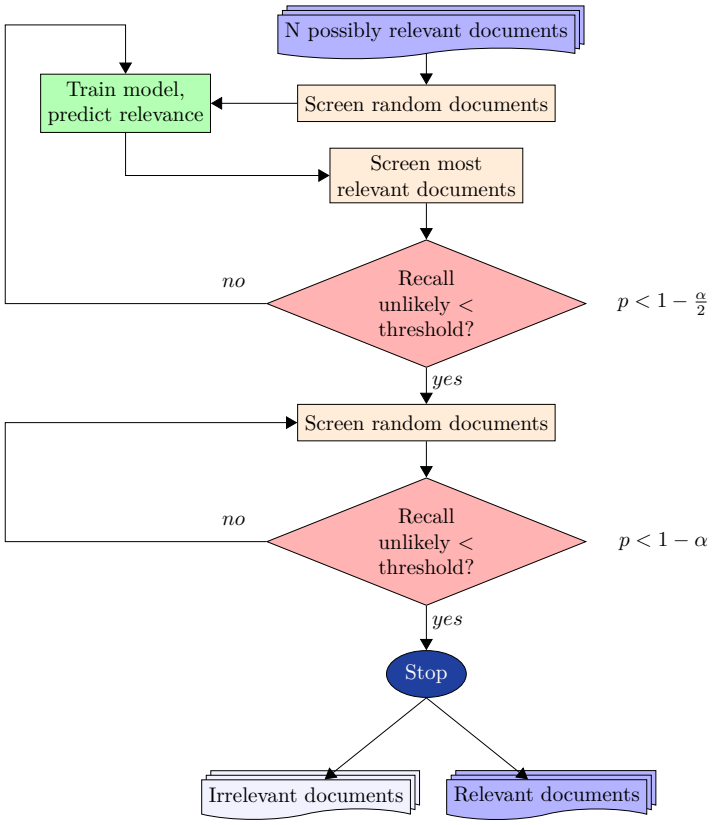


Figure 2.4: A workflow for active learning in screening with a statistical stopping criterion

section to be unhelpful). Subsequently, we "screen", that is, we reveal the labels of, batches of the 20 documents with the highest predicted relevance scores, retraining the model after each batch. Theoretically, using smaller batch sizes could mean that the recall target is achieved more quickly, but this is a trade-off between computational time spent training, and the speed at which the algorithm can "learn". However this is a modelling choice which may affect work saved, but not recall. Each criterion is evaluated after each document is "screened". For our criteria, we set the target recall value to 95% and the confidence level to 95%.

The systematic review datasets used for testing are described in table 2.1. We use the seminal collection of systematic reviews used to develop machine learning applications for document screening by Aaron Cohen and co-authors in 2006 Cohen et al. (2006), along with the widely used Proton Beam Terasawa et al. (2009) and COPD Castaldi et al. (2009) datasets, and computer science datasets used to test FASTREAD Yu and Menzies (2019). Testing on datasets with different properties and from different domains is key to establishing criteria appropriate for general use. Choosing as broad as possible data also prevents us from being able to "tune" our machine learning approach in ways that may work well for specific datasets but not generalise well. Work savings, even maximum work savings are therefore below the state of the art recorded for each of these datasets. In this way we can show how well the criteria perform even when the model performs badly.

All computational steps required to reproduce this analysis are documented online at `https://github.com/mcallaghan/rapid-screening`.

	dataset	data_source	N	r_docs	p
0	UrinaryIncontinence	cohen	284	68	0.24
1	Antihistamines	cohen	287	90	0.31
2	Estrogens	cohen	349	79	0.23
3	NSAIDS	cohen	358	83	0.23
4	OralHypoglycemics	cohen	475	135	0.28
5	Triptans	cohen	594	205	0.35
6	ADHD	cohen	803	83	0.10
7	AtypicalAntipsychotics	cohen	1030	333	0.32
8	CalciumChannelBlockers	cohen	1103	257	0.23
9	ProtonPumpInhibitors	cohen	1210	227	0.19
10	SkeletalMuscleRelaxants	cohen	1348	30	0.02
11	COPD	copd_pb	1443	179	0.12
12	Kitchenham	fastread	1700	45	0.03
13	Opiods	cohen	1769	43	0.02
14	BetaBlockers	cohen	1872	270	0.14
15	ACEInhibitors	cohen	2234	168	0.08
16	Statins	cohen	2743	152	0.06
17	ProtonBeam	copd_pb	4108	240	0.06
18	Radjenovic	fastread	5999	47	0.01
19	Wahono	fastread	7002	62	0.01
20	Hall	fastread	8911	104	0.01

Table 2.1: Dataset properties

2.3 Results

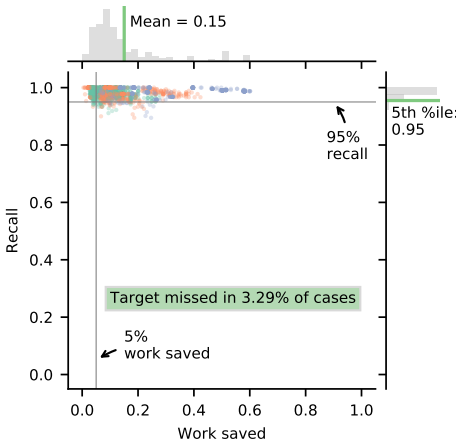
Figure 2.5 shows the actual recall and work savings achieved when each stopping criterion has been satisfied. For comparison, we also include the results that would have been achieved with *a priori* knowledge of the data, that is, the work saved when the 95% recall target was actually reached. In a live systematic review, reviewers would never know when this had been reached, but these are the work savings most often reported in machine learning for systematic review screening studies.

Both the random sampling and the ranked sampling criteria achieve the target threshold of 95% in more than 95% of cases. That this is greater than 95% is accounted for by the fact that random sampling sometimes begins after the target recall has been achieved, in which case the null hypothesis would be a priori impossible. The ranked quasi-sampling criterion outperforms the random sampling criterion with respect to both recall and work savings, saving a mean of 17% of the work compared to 15%, and missing the target in only 0.95% compared to 3.29% of cases. In theory, the ranked sampling criteria is conservative if the assumption holds that documents chosen by machine learning are not less likely to be relevant than those chosen at random. Based on our experiments, this assumption seems reasonable, and accounts for the higher recall. Because the ranked quasi-sampling criterion can flexibly choose its sample, whereas the random criterion has to wait for a random sample to be triggered, the criterion is also triggered earlier, as it can make use of more data. This accounts for the higher work savings.

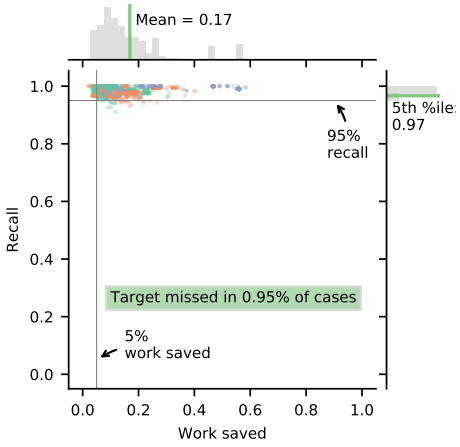
The baseline sampling criteron (Figure 2.5c) misses the 95% recall target in 39.67% of cases, while the most common work saving is 0%. This is in line with our expectations that, due to random sampling error, the expected number of documents will often be over-estimated or under-estimated, resulting in zero work savings or poor recall.

The Heuristic stopping criteria, both for 50 consecutive irrelevant results (Figure 2.5d - IH50), and for 200 irrelevant results (Figure 2.5e) also perform unreliably. Although the mean work saved for IH50 is 41%, the target is missed in 39% of cases. The cases below the horizontal grey line indicate instances where work has been saved at the expense of achieving the recall target.

In figure 2.6 we rescale the x axis, calling it additional burden, which is simply the work saved when the criterion is triggered minus the work saved when the recall target was actually achieved. This measure indicates whether the stopping criterion

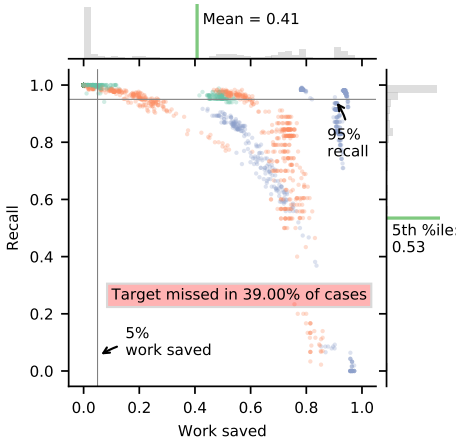


(a) Hypergeometric sampling

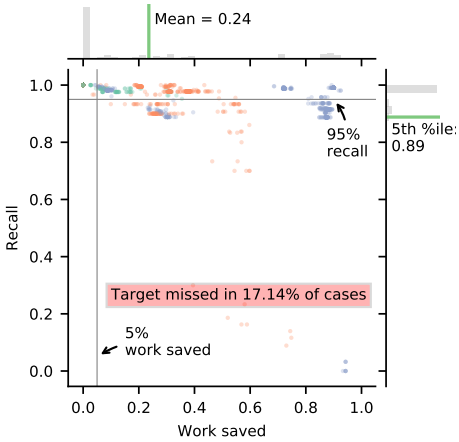


(b) Ranked quasi-sampling

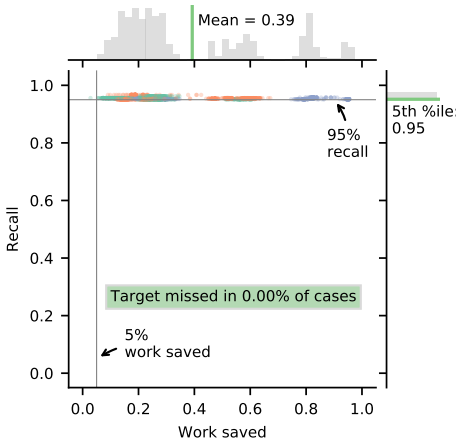
(c) Baseline Sampling



(d) 50 consecutive irrelevant results

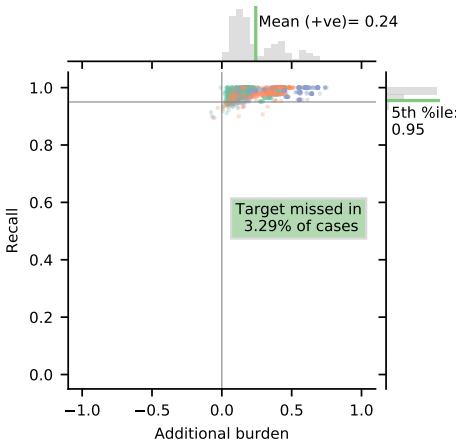


(e) 200 consecutive irrelevant results

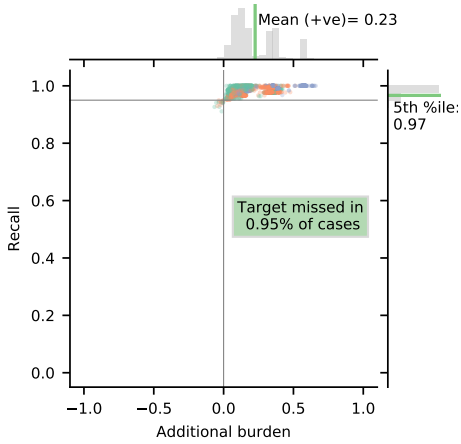


(f) *a priori* knowledge

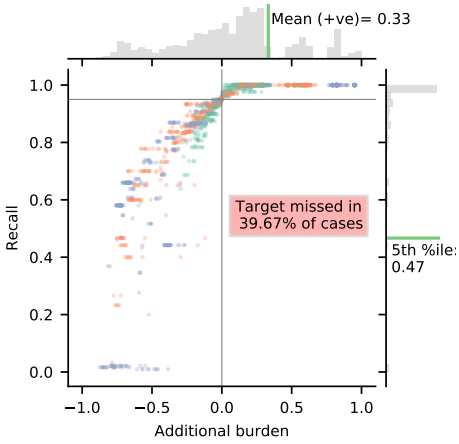
Figure 2.5: Distribution of recall and work saved after each stopping criteria. Green dots show results for datasets with less than 1,000 documents, orange dots show datasets with 1,000 - 2,000 documents, and blue dots show datasets with more than 2,000 documents.



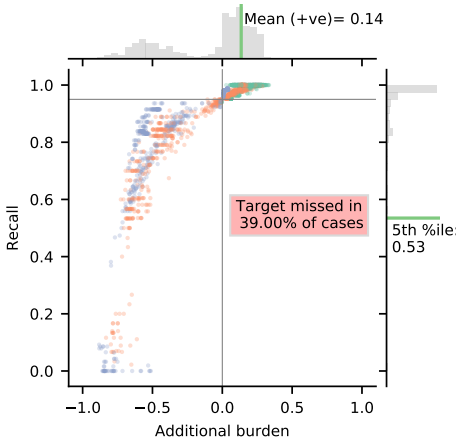
(a) Hypergeometric sampling



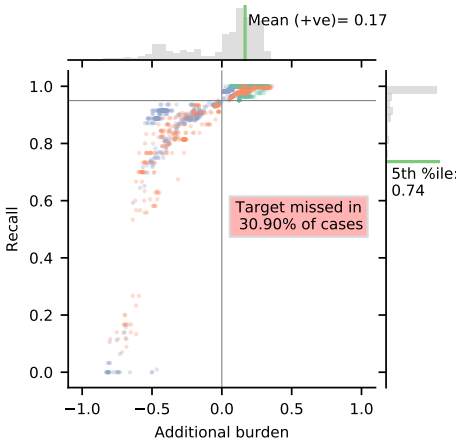
(b) Ranked quasi-sampling



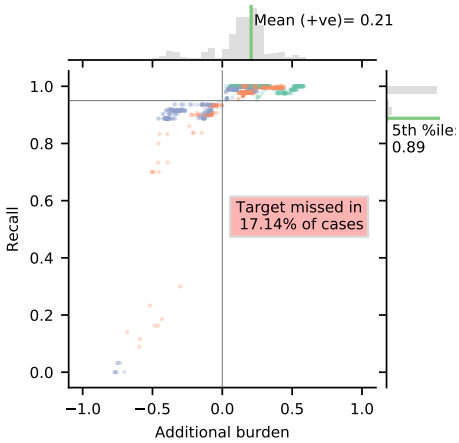
(c) Baseline Sampling



(d) 50 consecutive irrelevant results



(e) 100 consecutive irrelevant results



(f) 200 consecutive irrelevant results

Figure 2.6: Distribution of recall and additional burden after each stopping criterion. Additional burden is the work saved when the criterion was triggered minus the work saved when the target was reached. Coloring of data points as in Fig. 2.5.

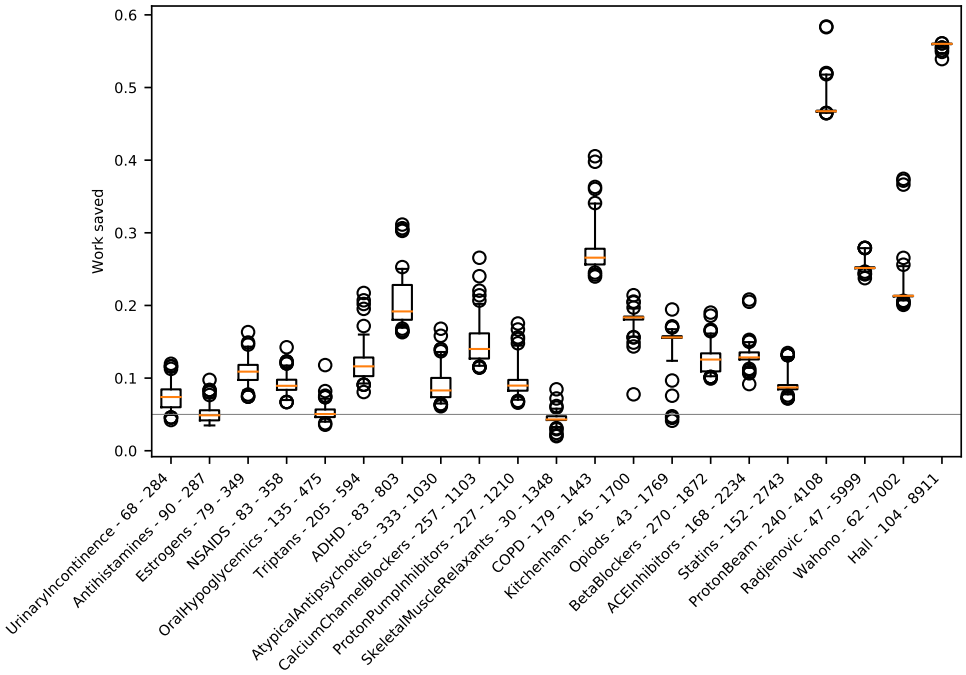
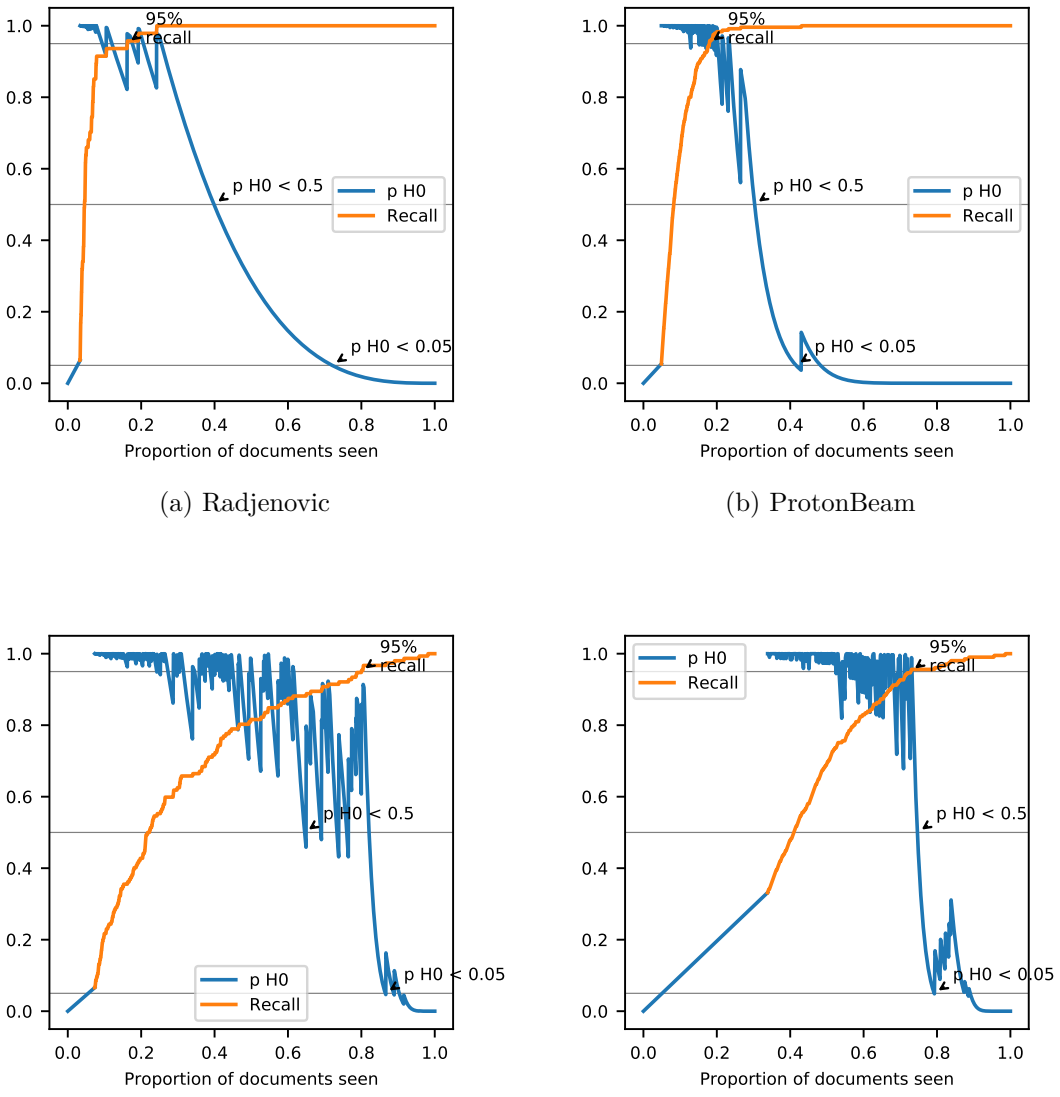


Figure 2.7: Work saved for the ranked quasi-sampling method in each dataset. Labels show the number of relevant documents and the total number of documents. The datasets are presented in order of the number of documents. The whiskers represent the 5th and 95th percentiles. The grey line shows work savings of 5%.



(a) Radjenovic

(b) ProtonBeam

(c) Statins

(d) Triptans

Figure 2.8: The path of recall (yellow) and the p-value of H0 for four different datasets

was triggered too early (negative values), or too late (positive values). The figure directly highlights the tradeoffs involved in deciding when to stop screening: For our criteria, there is mostly a small additional burden which comes with the necessity to make sure the desired recall target has been reached and reject the null hypothesis that this has not been the case. For the other criteria, there are many cases in which additional burden is negative, i.e. the criterion has been triggered too early. In these cases, however, the desired recall is hardly ever reached.

To help explain the different work savings that were observed in our experiments, we show the distribution of work savings from our ranked quasi-sampling criterion for each dataset in figure 2.7. In general, higher work savings are possible when the total number of documents is larger. However, in datasets with a low proportion of relevant documents, many documents need to be screened to achieve a high confidence that there are only few relevant documents remaining in the unseen ones. Therefore, smaller work savings are possible.

Figure 2.8 shows the recall and the p-value for the null hypothesis for the the iteration where the recall target is reached first for four datasets. Although the 95% recall target is achieved very quickly in the Radjenovic dataset, the null hypothesis cannot be excluded until much later. This is because the dataset has only 47 relevant documents out of a population of 5,999. After the 95% recall target was achieved, 45 out of 47 relevant documents had been seen and 5,029 documents remained. The null hypothesis was therefore that 3 or more of these 5,029 documents were relevant, which requires a lot of evidence to disprove. The burden of proof was smaller in the case of the Proton Beam dataset: at the point that the 95% recall threshold was reached, the null hypothesis to disprove was that a minimum of 13 out of 3,369 remaining documents were relevant.

The Statins and Triptans datasets show how the criterion performs when the machine learning model has performed poorly in predicting relevant results. In each case, 95% recall is achieved with close to 20% of documents remaining. With fewer documents remaining, it takes fewer screening decisions to rule out the possibility that the number of relevant documents left is incompatible with the achievement of the recall target.

2.4 Discussion

Our results show that it is possible to use machine learning to achieve a given level of recall with a given level of confidence. The tradeoff for achieving recall reliably is that the work saving achieved is less than the maximum possible work saving. However, for large datasets with a significant proportion of relevant documents, the additional effort required to satisfy the criterion will be small compared to the work saved by using machine learning. This makes the approach well suited to broad topics with lots of literature. In other words, it is precisely where machine learning will be most useful that the additional effort will be small.

Different use cases for machine learning enhanced screening may also carry different requirements for recall, or different tolerances for uncertainty. These can be flexibly accommodated within our stopping criterion. Importantly, the ability to make statements about the authors' confidence in achieving a given recall target makes it possible to clearly communicate the implications of using machine learning enhanced screening to readers and reviewers who are not machine learning specialists. This is extremely important in live systematic reviews.

Our criteria have the further advantage that they are independent of the choice or performance of the machine learning model. If a model performs badly at discerning relevant from irrelevant results, the only consequence will be that the work saved will be low. With other criteria this may result in poor recall. When using machine learning for screening, poor recall can result in biased results, while low work savings represent no loss to the reviewer as compared to not using machine learning.

One caveat in the derivation of our criteria is that we did not address the problem of multiple testing formally. Such a derivation is mathematically challenging and beyond the scope of this paper. However, the performance of the criteria shows that this is of limited practical concern. Formally describing screening procedures with iterative testing should be a next step towards even more rigorous stopping criteria and should be fully worked out in future research.

So far, systematic review standards have no way of accommodating screening with machine learning. We hope that the reliability and clarity of reporting offered by our stopping criteria make them suitable for incorporation into standards, so that machine learning for systematic review screening can fulfil its promise of reducing workload and making more ambitious reviews tractable.

2.5 Conclusion

This paper demonstrates the drawbacks of existing stopping criteria for machine learning approaches to document screening, particularly with regard to reliability. We propose a simple method that delivers reliable recall, independent of machine learning approach or model performance. Our statistical stopping criteria allow users to easily communicate the implications of their use of machine learning, making machine learning enhanced screening ready for live reviews.

Bibliography

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S. C., Ananiadou, S., Liao, J., and Macleod, M. R. (2019). Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews*, 8(1):1–12.

Buonaccorsi, J. P. (1987). A Note on Confidence Intervals for Proportions in Finite Populations. *The American Statistician*, 41(3):215–218.

Castaldi, P. J., Cho, M. H., Cohn, M., Langerman, F., Moran, S., Tarragona, N., Moukhachen, H., Venugopal, R., Hasimja, D., Kao, E., Wallace, B., Hersh, C. P., Bagade, S., Bertram, L., Silverman, E. K., and Trikalinos, T. A. (2009). The COPD genetic association compendium: A comprehensive online database of COPD genetic associations. *Human Molecular Genetics*, 19(3):526–534.

Cohen, A. M., Hersh, W. R., Peterson, K., and Yen, P. Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. In *Machine Learning*, pages 273–297.

Di Nunzio, G. M. (2018). A study of an automatic stopping strategy for technologically assisted medical reviews. In Pasi, G., Piwowarski, B., Azzopardi, L., and Hanbury, A., editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10772 LNCS, pages 672–677, Cham. Springer International Publishing.

Howard, B. E., Phillips, J., Tandon, A., Maharana, A., Elmore, R., Mav, D., Sedykh, A., Thayer, K., Merrick, B. A., Walker, V., Rooney, A., Shah, R. R., Llc, S., Durham, D. D., Toxicology, N., Ntp, P., Sciences, H., and Rtp, T. W. A. D. (2020). SWIFT-Active Screener : Accelerated document screening through active learning and integrated recall estimation. *Environment International*, 138(April 2019):105623.

Jonnalagadda, S. and Petitti, D. (2013). A new iterative method to reduce workload in systematic review process. *International Journal of Computational Biology and Drug Design*, 6(1/2):5.

Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M.-I., Noel-Storr, A., Rader, T., Shokraneh, F., Thomas, J., and Wieland, L. S. (2011). Cochrane Handbook for Systematic Reviews of Interventions. In Higgins, J. and Green, S., editors, *Cochrane Handbook for Systematic Reviews of Interventions*, chapter Chapter 4:. The Cochrane Collaboration, version 5. edition.

Marshall, I. J. and Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8(1):1–10.

Miwa, M., Thomas, J., O'Mara-Eves, A., and Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics*, 51:242–253.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., and Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, 4(1):1–22.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python Fabian. *Journal of Machine Learning Research*, 12:2825–2830.

Przybyła, P., Brockmeier, A. J., Kontonatsios, G., Le Pogam, M. A., McNaught, J., von Elm, E., Nolan, K., and Ananiadou, S. (2018). Prioritising references for systematic reviews with RobotAnalyst: A user study. *Research Synthesis Methods*, 9(3):470–488.

Sahai, H. and Khurshid, A. (1995). A note on confidence intervals for the hypergeometric parameter in analyzing biomedical data. *Comput Biol Med.*, 25(1):35–38.

Settles, B. (2009). Active Learning Literature Survey. Technical report, University of Wisonsin-Madison.

Shemilt, I., Simon, A., Hollands, G. J., Marteau, T. M., Ogilvie, D., O'Mara-Eves, A., Kelly, M. P., and Thomas, J. (2014). Pinpointing needles in giant haystacks: Use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*, 5(1):31–49.

Terasawa, T., Dvorak, T., Ip, S., Raman, G., Lau, J., and Trikalinos, T. (2009). Review Annals of Internal Medicine Systematic Review : Charged-Particle Radiation Therapy for Cancer. *Annals of Internal Medicine*, (5).

Wallace, B. C., Small, K., Brodley, C. E., and Trikalinos, T. A. (2010a). Active learning for biomedical citation screening. (July):173.

Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., and Schmid, C. H. (2010b). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, 11.

Westgate, M. J., Haddaway, N. R., Cheng, S. H., McIntosh, E. J., Marshall, C., and Lindenmayer, D. B. (2018). Software support for environmental evidence synthesis. *Nature Ecology & Evolution*, 2:588–590.

Yu, Z. and Menzies, T. (2019). FAST 2 : An intelligent assistant for finding relevant papers. *Expert Systems with Applications*, 120:57–71.

Chapter 3

A Topography of Climate Change Research

Abstract

The massive expansion of scientific literature on climate change Minx et al. (2017) poses challenges for global environmental assessments and our understanding of how these assessments work. Big data and machine learning can help us deal with large collections of scientific text, making the production of assessments more tractable, and giving us better insights about how past assessments have engaged with the literature. We use topic modelling to draw a topic map, or topography, of over 400,000 publications from the Web of Science (WoS) on climate change. We update current knowledge on the Intergovernmental Panel on Climate Change (IPCC), showing that, when compared to the baseline of the literature identified, the social sciences are in fact over-represented in recent assessment reports. Technical, solutions-relevant knowledge - especially in agriculture and engineering - is under-represented. We suggest a variety of other applications of such maps, and our findings have direct implications for addressing growing demands for more solution-oriented climate change assessments that are also more firmly rooted in the social sciences Kowarsch et al. (2017); David G. Victor (2015). The perceived lack of social science knowledge in assessment reports

does not necessarily imply a IPCC bias, but rather suggests a need for more social science research with a focus on "technical" topics on climate solutions.

3.1 Introduction

We live in an age of "Big Literature" Nunez-Mir et al. (2016); Minx et al. (2017), where the science of climate change is expanding exponentially Grieneisen and Zhang (2011); Haunschild et al. (2016). In the five years since the publication of the last IPCC assessment report IPCC (2014), 202,000 papers were published in the Web of Science (WoS) (see Table 3.1). This is almost as much as the 205,000 papers published during the first five assessment periods; a period of nearly 30 years. A total of around 350,000 new publications can be expected for the current sixth assessment cycle of the IPCC, based on current growth patterns (Figure 3.1). Moreover, the literature has also become more diverse. This is reflected in the expansion of the literature's vocabulary - from 2,000 unique words in the first assessment period to 95,000 words so far in the sixth - indicating that the field has incorporated new content. For example, the zika virus, which was mentioned in 182 articles from 2014-2018, had never before been discussed in the titles or abstracts of articles relating to climate change. Yet it has emerged as a topic of high relevance: the incidence of the virus, the outbreak of which in Brazil in 2016 was declared a public health emergency by the WHO, is set to increase under rising global temperatures Rao et al. (2019). Similar rapid emergence patterns can be seen for INDCs and SDGs in AR6, and Biochar and REDD in AR5, among others[1].

Big literature poses at least three challenges for scientific policy advice and science itself: *First*, established procedures in scientific assessments like those conducted by the IPCC fail to address the exploding literature base. For example, the ratio of studies cited in IPCC reports to the number or relevant studies has declined from 60% to 20% Minx et al. (2017), posing a rapidly growing risk of selection bias. More generally, the provision of comprehensive, objective, open and transparent assessments of the available scientific literature, as defined in the principles governing IPCC work IPCC (2013), is no longer possible for authors or author teams by traditional means.Machine reading and learning methods as well as other data science applications are required to enable an understanding of the field of climate change research at scale. *Second*,

[1]The glossary in SI contains a complete list of the acronyms shown in the table

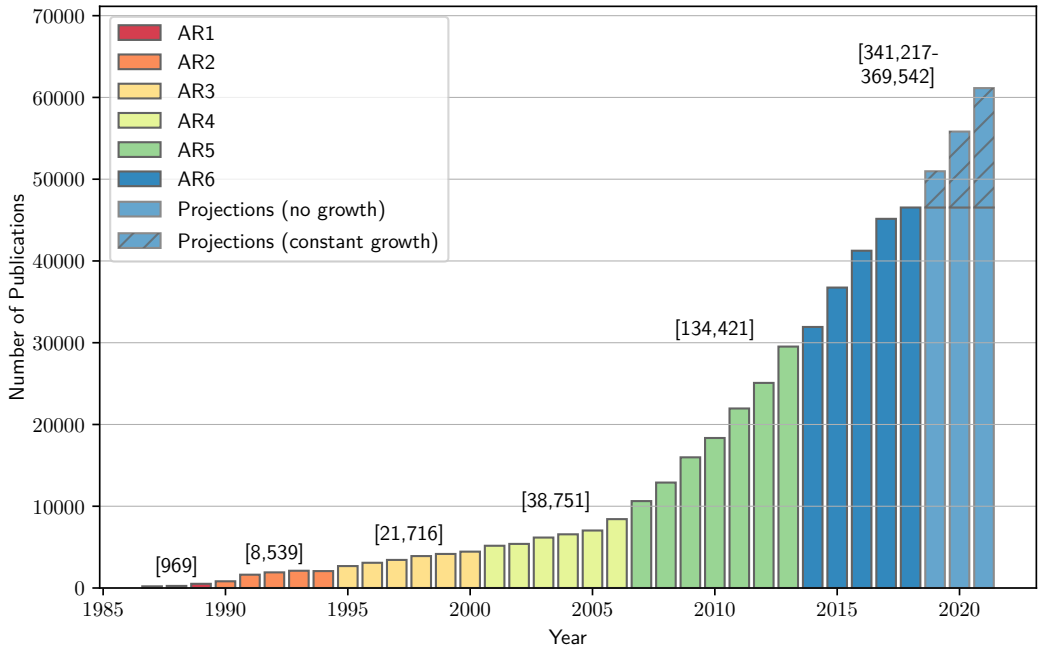


Figure 3.1: The number of climate change documents in the Web of Science in each year. For 2019-21 we project the number of papers assuming there is no more growth, and assuming that growth continues at the same rate as over the past five years

	AR1	AR2	AR3	AR4	AR5	AR6
Years	1986-1989	1990-1994	1995-2000	2001-2006	2007-2013	2014-
Documents	1,167	8,539	21,716	38,750	134,413	201,606
Unique words	2,000	12,480	23,346	34,637	71,867	94,746
New words	change (560)	oil (287)	downscaling (217)	sres (234)	biochar (1,791)	mmms (313)
	climate (428)	deltac (283)	degreesc (187)	petm (95)	redd (1,113)	cop21 (234)
	co2 (318)	whole (256)	ncep (130)	amf (88)	cmip5 (679)	c3n4 (214)
	climatic (289)	tax (254)	fco (107)	sf5cf3 (86)	cmip3 (587)	sdg (187)
	model (288)	landscape (249)	pfc (98)	clc (81)	mofs (299)	zika (182)
	atmospheric (281)	alternative (243)	otcs (98)	embankment (81)	sdm (297)	ndcs (168)
	effect (280)	availability (242)	dtr (95)	cwd (79)	mof (275)	indc (164)
	global (224)	life (239)	nee (89)	etm (75)	biochars (252)	indcs (134)

Table 3.1: Growth of Literature on Climate Change. A glossary of acronyms is provided in SI

evidence synthesis - the enterprise of reviewing the literature based on a formal and systematic set of methods Chalmers et al. (2002) - becomes increasingly important for aggregating and consolidating the rapidly emerging knowledge and enabling scientific assessments to do their job. Yet traditional methods of evidence synthesis themselves are pushed to their limits by the large amount of scientific publications. The field of evidence synthesis technology, which tries to streamline human tasks through machine learning at the different stages of the review process, is still in its infancy Beller et al. (2018). *Finally*, overwhelming amounts of literature may be a major reason why studies of scientific assessments Bjurström and Polk (2011); Hulme and Mahony (2010); David G. Victor (2015) do not offer robust quantifications, for claims about the relationship between report citations and the underlying literature.

This study uses topic modelling Blei et al. (2010) to map out the vast body of evidence on climate change. Topic modelling is an unsupervised machine-learning technique, where patterns of word co-occurrences in documents are used to learn a set of topics which can be used to describe the corpus. The word topic derives from the Greek word for place (topos), and by *situating* the documents in a reduced-form projection of

their thematic content (see Figure 3.3), we create a *topographic map* of the literature on climate change. Such a systematic engagement with the thematic content of the climate science is missing from the literature so far.

We apply this map in a second step to understand how the IPCC reports have represented the available climate change literature and re-evaluate claims of bias based on a more comprehensive understanding of the available climate science. We enrich the discussion of representation in the literature by discussing topics as well as disciplines.

3.2 Methods

3.2.1 Data

This study reproduces the query developed by (Grieneisen and Zhang, 2011), which is carried out on the Web of Science core collection. We downloaded the results of the query on March 19, 2019. Though not exhaustive, the Web of Science gives a good coverage of the literature in major peer-reviewed journals. The Web of Science data gives us a disciplinary classification (based on the journal) and publication year, among other metadata, for each document. Each document is assigned to an assessment period according to the timeline shown in table 1.

We also tested the query documented in Haunschild et al. (2016), by checking a random sample of documents exclusive to it. We found that the majority of additional documents were not relevant, and decided to use only the query from Grieneisen and Zhang (2011).

We use the references scraped from IPCC assessment reports from (Minx et al., 2017), and attempt to match these with the results from the Web of Science. We use doc2vec similarity scores Le and Mikolov (2014) to identify the 500 most similar titles for each reference, and count the document as a match if the jaccard similarity score of the two word shingles of the reference title and the document title is greater than 0.5 Khabsa and Giles (2014). Extended Table 3.2 shows the percentage of IPCC citations matched in each working group for each assessment report. This is significantly lower in earlier periods, as data coverage and quality of citation databases is lower for earlier periods. Matching in WG III is also lower, suggesting a greater share of non-peer review literature, or literature not directly mentioning climate change, but related to its mitigation (for example on energy policy).

We analysed by hand a sample of 100 IPCC references which could not be matched and found that 46% of these references were not in the Web of Science at all, 53% were in the Web of Science but not in our query, and 1 document was in our query but had mistakenly been identified as not being so. This was due a different version of the title appearing in the IPCC citation and the Web of Science record.

3.2.2 Pre-processing

Data quality in earlier Web of Science results is poorer, and some documents have missing abstracts. In the quantification of the size of the literature and its vocabulary in Table 3.1, titles are substituted for abstracts where they are not available. The words of the documents are lemmatized, replacing different forms of the same word (i.e. word/words) with a single instance. Commonly occurring words, or "stopwords" are removed, as are all words shorter than 3 characters, and all words containing only punctuation or numbers.

The documents are transformed into a document-term matrix, where each row represents a document, and each column represents a unique word. Each cell contains the number of that column's terms in that document. Only terms which occur more than once are considered.

For the calculation of the topic model, documents with missing abstracts are ignored, and the document term matrix is transformed into a document frequency-inverse document frequency (tf-idf) matrix, where scores are scaled according to the frequency of their occurrence in the corpus. This gives more weight to terms which appear in few documents, and less weight to those which appear in many.

$$tf(t,d) = f_{t,d}, \quad idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{3.1}$$

3.2.3 Topic Model

We use non-negative Matrix Factorisation (NMF) Lee and Seung (1999), an approach to topic modelling which factorises the term-frequency-inverse document frequency matrix V into the matrices W, the topic-term matrix, and H the document-topic matrix, whose product approximates V:

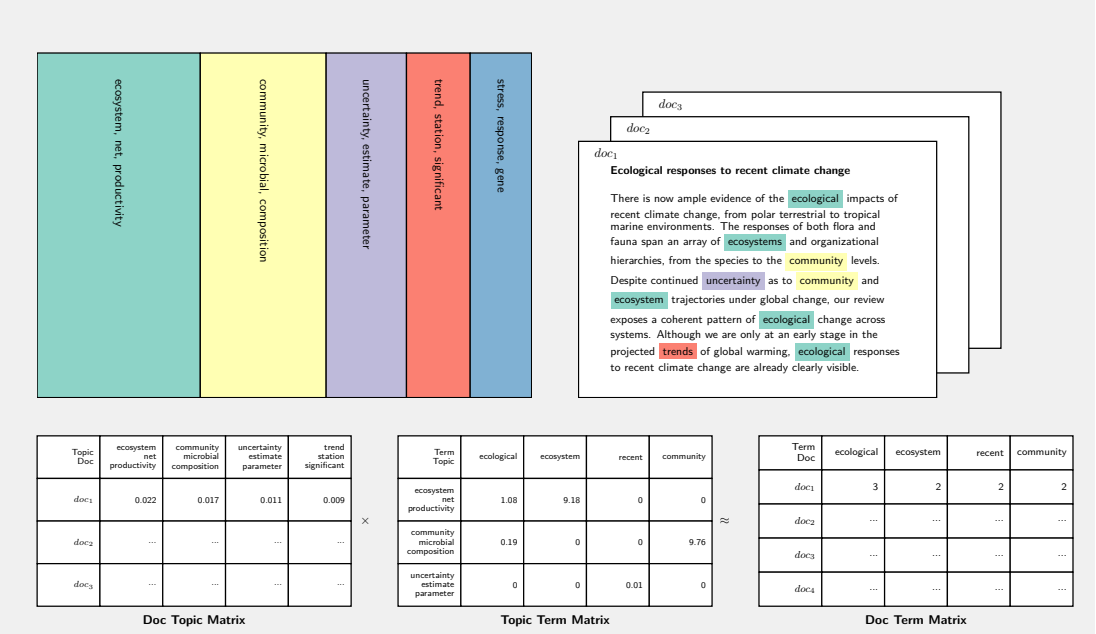


Figure 3.2: Topic make up of a single document. The Doc Term Matrix shows the number of occurrences of each term in the document. The Topic Term Matrix shows the topic score of each term-topic combination. The Doc Topic Matrix shows the document-topic score for each topic. This topic makeup of the document shown is illustrated by the bars in the top left. Words highly associated with each topic that occur in the document are highlighted. All values are real, although the doc-term matrix is scaled by the inverse-document frequency before being used in the model.

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu} \tag{3.2}$$

As demonstrated in Figure 3.2, each topic is represented as a set of word scores, and each document a set of topic scores. The combination of the two approach the word scores in the document. For clarity in the figure, these are shown as simple counts, but in the model these are scaled according to each term's frequency within the corpus as explained above.

Topics are calculated using the scikitlearn library Pedregosa et al. (2011), and are saved in a database and topic visualisation system based on Chaney and Blei (2012) [1].

[1]The system adds new functionality to Chaney and Blei (2012) and combines it with a system for

Model selection

Topic models are calculated for 70, 80, 90, 100, 110, 120, 130, 140 and 150 topics. The run with 150 topics was discarded as it contained a topic to which no terms or documents were assigned. The relative usefulness of each model was assessed subjectively by the authors, based on inspection of the online visualisation tool, and the spreadsheet **topic_comparison.xlsx** accompanying the supporting information. The spreadsheet shows each set of topics in adjacent columns. Topics from each model are placed next to the topics with the largest number of each topic's 10 highest scoring words in common. This helps authors to find an appropriate level of granularity for the analysis. Statistical methods for the selection of topic model parameters are available but they do not necessarily align with human perceptions of topic model quality Chang et al. (2009). We make a judgement based on subjective criteria, but for transparency publish the results of the analysis for different numbers of topics in Extended Figure 3.9. The main conclusions drawn about the rapid growth and under-representation of solutions-relevant topics are stable across models.

Topic assignment to working groups

A topic's score for each working group is calculated by summing the document-topic scores for all documents cited by that working group. We call the topic's primary working group that working group for which the above sum is the highest, but in some cases, where there are very few IPCC citations of documents related to a topic this can be misleading . For example, the word "capacity" is relevant to the adsorption topic, so documents talking about adaptive capacity receive a low score for the topic. Because only very few documents highly relevant to the topic (in that they talk about adsorption or adsorptive capacity) are cited by the IPCC, and many of the weakly relevant documents are cited by the IPCC, the sum of the topic scores of the weakly relevant documents outweighs the sum of the topic scores of the strongly relevant documents, meaning that the topic is mistakenly assigned to working group II when it is more properly relevant to working group III. We point out that topics are in any case mixtures of documents cited by different working groups, and stress that the colouring of the topics by working group is merely illustrative.

managing sets of documents and queries. The code and additional information is published online at https://github.com/mcallaghan/tmv

Topic Representation and Newness

To calculate topic representation in IPCC reports we divide each topic's share in the subsample of documents cited by IPCC reports by its share in the whole corpus (excluding documents published after the last assessment report). Disciplinary representation is calculated in the same way.

We calculate a topic's total score as the sum of document-topic scores. A topic's window score is the sum of document-topic scores considering only documents in the given time window. To represent a topic's newness, we multiply each assessment period number by the share of it's total score occurring in that window, and take the mean of these scores. A topic in which 100% of documents which make it up occurred in assessment period 1 (6) would thereby receive a score of 1 (6), while a topic evenly distributed across all assessment periods would receive a score of 3.5.

Disciplinary Entropy

Disciplinary Entropy inverts the measurement of a conference's topical diversity suggested in Hall et al. (2008), by measuring a topic z's entropy H, where

$$H(f|z) = -\sum_{i=1}^{K} \hat{p}(f|z) \log \hat{p}(f|z) \tag{3.3}$$

based on the empirical distribution of a field f in the documents d in each topic:

$$\hat{p}(f|z) = \sum_{d:z_d=z} \hat{p}(f|d)\hat{p}(d|z) \tag{3.4}$$

It is an indication of the diversity of disciplines within the set of documents related to a topic.

Topic Map

The topic model gives us the location of each document in a 140 dimensional topic space, with each dimension corresponding to a that document's *topic-ness* in a given topic. t-Distributed Stochastic Neighbour Embedding (t-SNE) is a dimensionality reduction technique which we use to represent each document's topic scores in 2 dimensions van der Maaten and Hinton (2008). Documents are placed on the map such that documents with similar combinations of topics are close together.

3.3 Mapping out the landscape of climate change literature

Figure 3.3 shows a *thematic* or *topopographic map* of the 400,000 publications on climate change in our dataset with a total number of 140 topics. The number of topics must be defined exogenously, but the results are robust to different specifications. Using non-negative matrix factoriziation Lee and Seung (1999), the topics are machine-learned from the papers' abstracts (see methods for details, examples of different model specifications, and a thorough explanation of model selection), and the topic scores of each document are reduced to the two dimensions shown through t-distributed stochastic neighbour embedding van der Maaten and Hinton (2008) [1].

The map shown covers a broad range of topics, with related topics shown in clusters. In general, topics related to climate science and impacts are in the West, while solution-oriented topics are in the East. More fine-grained research areas can also be distinguished. For example, publications related to urban infrastructure (**buildings**, **energy**, **cement**, **waste**) are located in the East, physical climate impacts such as **sea-level**, **droughts** or [crop] **yield** are in the South-West and energy systems are in North-East. There are larger groups of documents at the fringes of the map that relate mainly to one or two specific topics such as **biochar**, **coral**, or **CO2 storage**. Interestingly, scenarios feature centrally in the map, at the interface between different scientific communities. This corresponds to their integrative nature in IPCC reports Moss et al. (2010). This map of the thematic structure of the literature could be useful for individual communities or for climate change assessments.

The disciplinary composition of this research topography indicated by the different colours in Figure 3.3 highlights the dominance of natural sciences in climate change research. More than 60% of the literature is published in natural science journals. Similarly, in 115 out of 140 topics the contribution of publications in natural science journals is greater than any other discipline. We calculate disciplinary entropy of topics as a measure of their degree of interdisciplinarity (see Extended Figure 3.6 and methods for details). This shows how research on **health**, **food**, or **policy** comes from a range of disciplines, while research on **ice** and **oceans** comes almost exclusively from the natural sciences).

[1] A full list of topics and related words, and a list of documents, their positions on the map, and their related topics are given in the SI

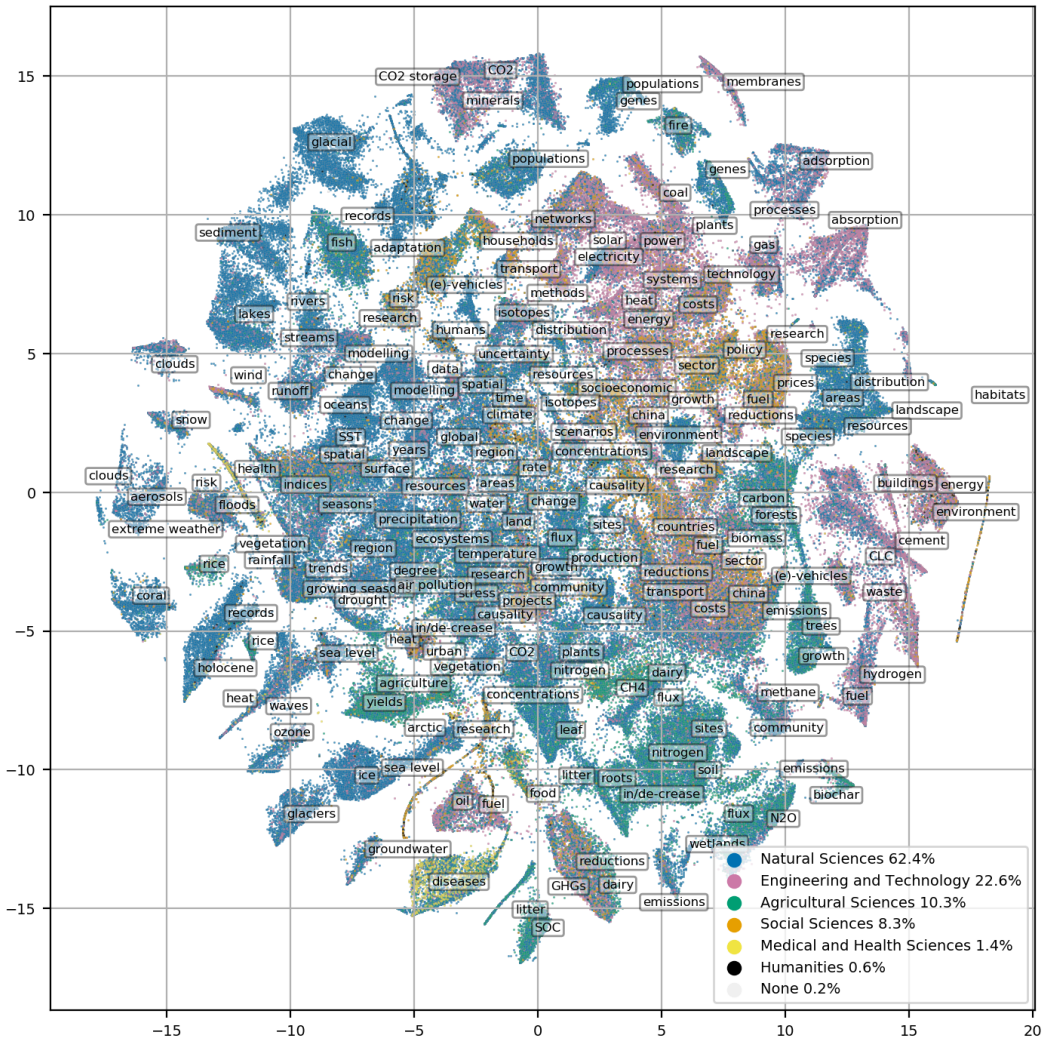


Figure 3.3: A map of the literature on climate change. Document positions are obtained by reducing the topic scores to two dimensions via t-SNE Documents are coloured by web of science discipline category. Topic labels are placed in the center of each of the large clusters of documents associated with each topic.

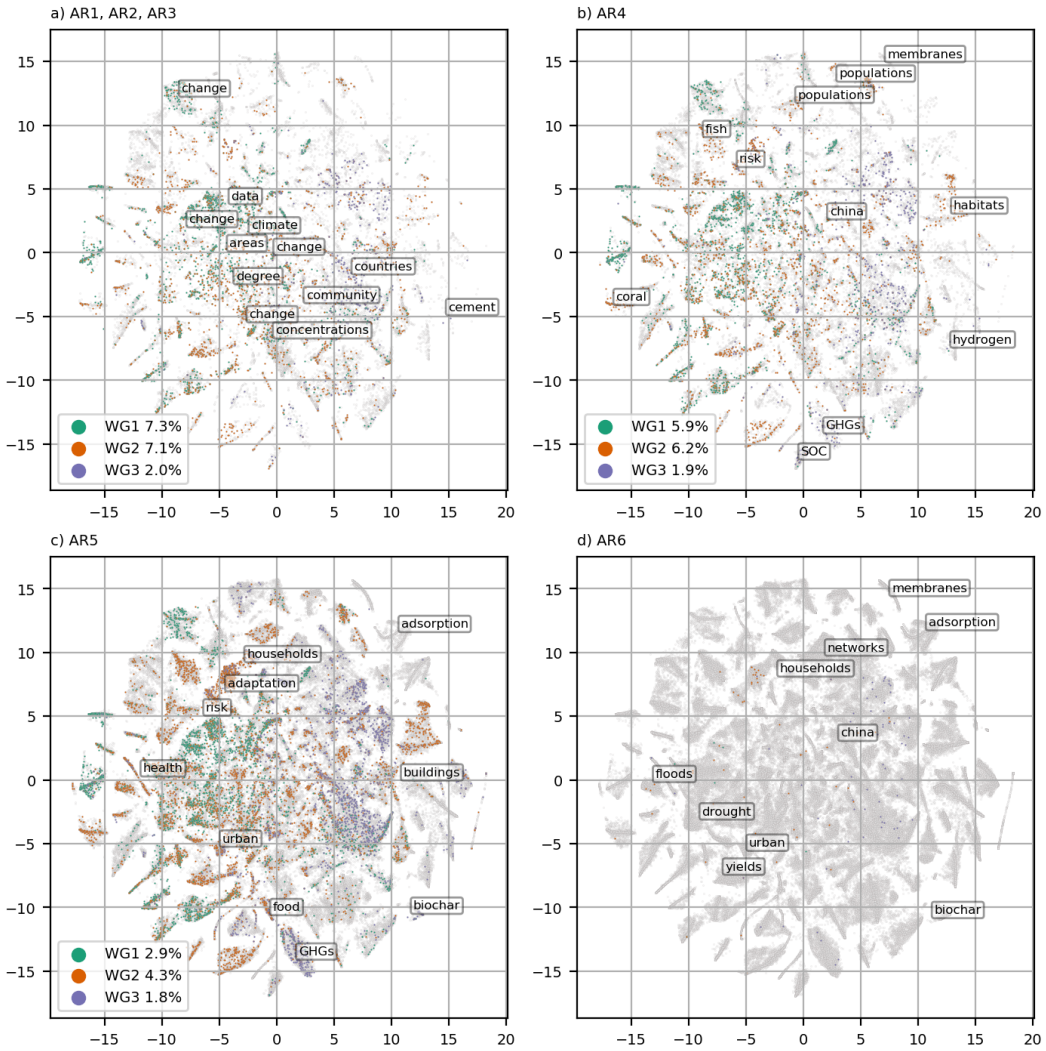


Figure 3.4: Evolution of the landscape of climate change literature. In each period, the 10 fastest growing topics are labelled. Where documents could be matched to IPCC citations, they are coloured by the working group citing them.

Finally, the topography shows the thematic evolution of the literature (Figure 3.4), with topics exhibiting distinct patterns of growth. Fast-growing topics in the last three assessment periods have included, among others, **coral**, **risks**, **adaptation**, **hydrogen**, **buildings**, **CO2 removal**, **networks** and **biochar**. **Biochar** is particularly remarkable in that the sizeable literature which emerged in AR5 was completely absent from the climate change literature beforehand.

The identification of new topics as they emerge, particularly as these are identified without prior knowledge of the literature, can help researchers and assessment-makers to keep abreast of a quickly evolving field.

3.4 Research representation in IPCC reports

We apply our topic map to understand the representation of science in IPCC assesssments and how it manages to respond to demands for more solution-oriented knowledge Kowarsch et al. (2017). Several studies have identified, made, or repeated claims of a disciplinary bias of IPCC assessments towards the natural sciences, and within the social sciences towards economics Bjurström and Polk (2011); David G. Victor (2015); Hulme and Mahony (2010); Corbera et al. (2016). Where these claims were based on an analysis of IPCC citations Bjurström and Polk (2011), they fail to assess this claim against a measurable benchmark. We argue here that the composition of the climate change literature as a whole provides such a benchmark, in view of the organisation's mandate to provide "comprehensive, objective, open and transparent" assessment of the available science IPCC (2013). Our database of publications allows us to study representation with such a benchmark, and over time rather than for single assessment cycles.

Figure 3.5.a shows that the social sciences were indeed under-represented in the third assessment report, but by the fifth assessment report were over-represented. Likewise, other social sciences than economics have become better represented since AR3 (see figure 3.7f) with social & economic geography (4.3% of the literature), political science (1.0%), and sociology (0.8%) showing improved representation in AR5 compared to AR3, and social and economic geography, political science, and other social sciences better represented than economics.

This challenges what we think we know about the IPCC. The social sciences, by now, are actually the best represented field, with a share in the literature cited by

IPCC reports 1.32 times as high as in the literature at large. On the other hand the Agricultural Sciences and Engineering & Technology have been consistently under-represented, with 2.27 and 3.49 times the share of studies in the wider literature than in the literature cited by the IPCC in AR5 respectively. Humanities are also under-represented, although they make up a very small proportion of the total literature.

The topography allows us to delve deeper into the subject matter that receives more or less attention in the IPCC. Figures 3.5b and 3.5c plot the representation of the topics shown in the map. Figure 3.5c shows that topics more commonly cited by IPCC working group I are older and largely better represented in IPCC reports. These topics, for example **ozone**, **oceans**, **clouds**, **aerosols** and **sea levels** make up some of the core topics of the physical science of climate change.

The topics in the lower right of the graph are the most pertinent to the question of whether the IPCC is well representing knowledge on climate change. These topics are newer and until now have been under-represented in IPCC reports. Because they are new areas of knowledge, they may be highly salient in a periodic assessment process. These topics are primarily in working group III, on mitigation [1].

The difference between these under-represented new topics and other new topics that are better represented is intriguing. This difference is visible in figure 3.4, where in AR5, the clusters of documents around the **adsorption**, **buildings**, and **biochar** topics contain few IPCC citations, whereas the clusters around **food**, **health**, **adaptation**, and **GHGs** contain more. As shown in figure 3.5c, **adsorption**, **buildings** and **biochar** are 4.08, 3.34 and 3.61 times more prevalent in the literature than in IPCC citations, while **food** is 1.22 times more prevalent in the literature and **health** and **adaptation** are 1.02 and 2.22 times more prevalent in IPCC citations respectively. The IPCC, has been better at integrating new knowledge from these topics, and in general better at integrating new knowledge from WG II than WG III topics.

Further, within WG III topics, those that are well represented contain a greater proportion of social science research (figure 3.5b). The topics **countries**, **policy**, and **prices** are close to a proportional representation and are made up of around 30% social science research. **Waste**, **biochar**, **cement** and **coal**, are more than 3 times more prevalent in the wider literature than in the literature cited by the IPCC, and

[1] see methods for a discussion of the categorisation of topics, including CLC, adsorption and hydrogen, which may more properly be described as relevant to WGIII

are made up of around 5% social science research. This pattern is not visible in other working groups (see Extended Figure 3.8), and complicates the perception of the under-representation of the social sciences.

Recalling policymakers' demands for more solution-oriented assessments Kowarsch et al. (2017), we could also interpret the topics that are newer and under-represented as "solutions-relevant". However, while policymakers' demands for solutions-oriented knowledge were rather about policy options, these under-represented new topics deal with more technical solutions and are found rather in technical disciplines within engineering & technology and the agricultural sciences.

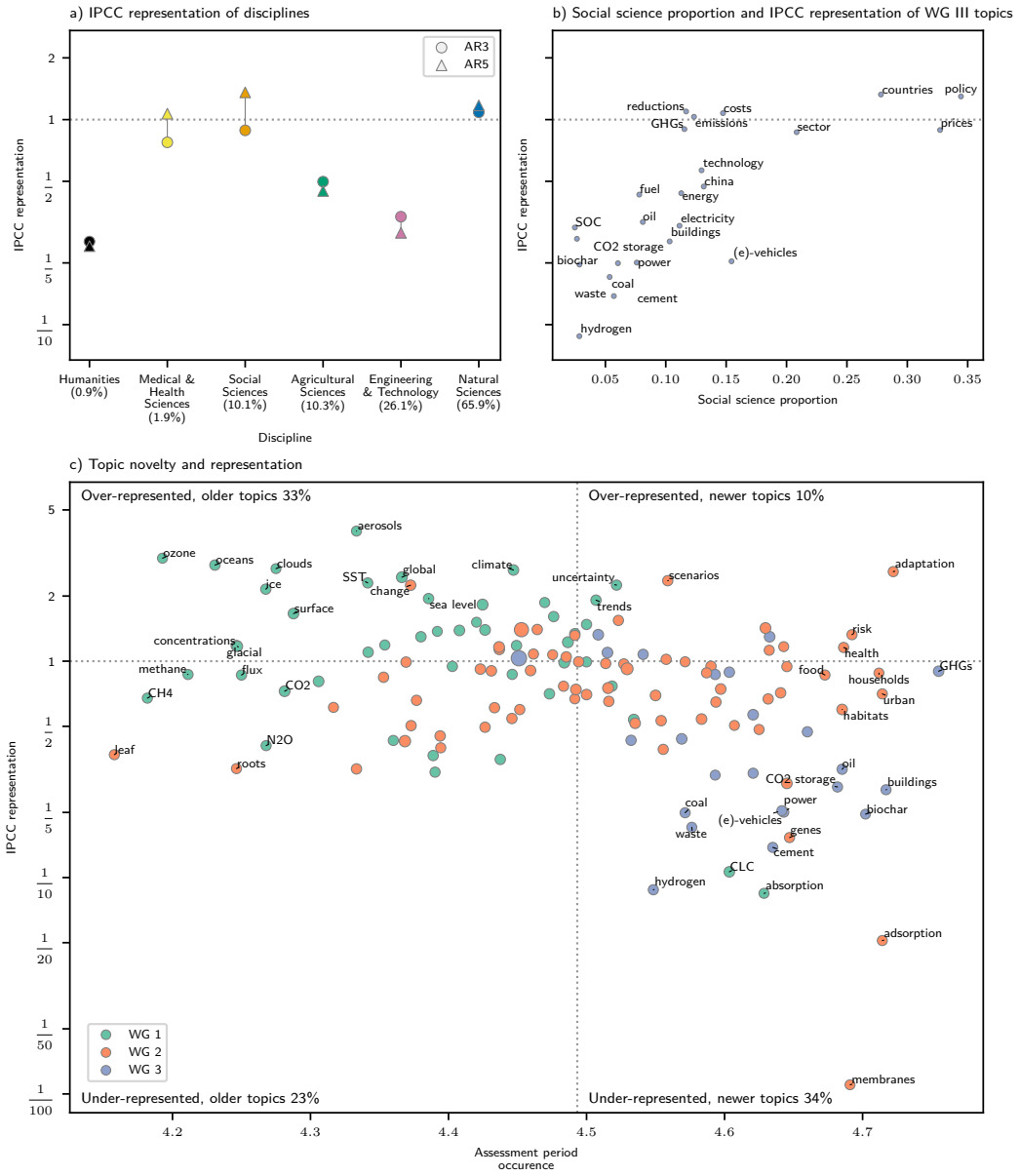
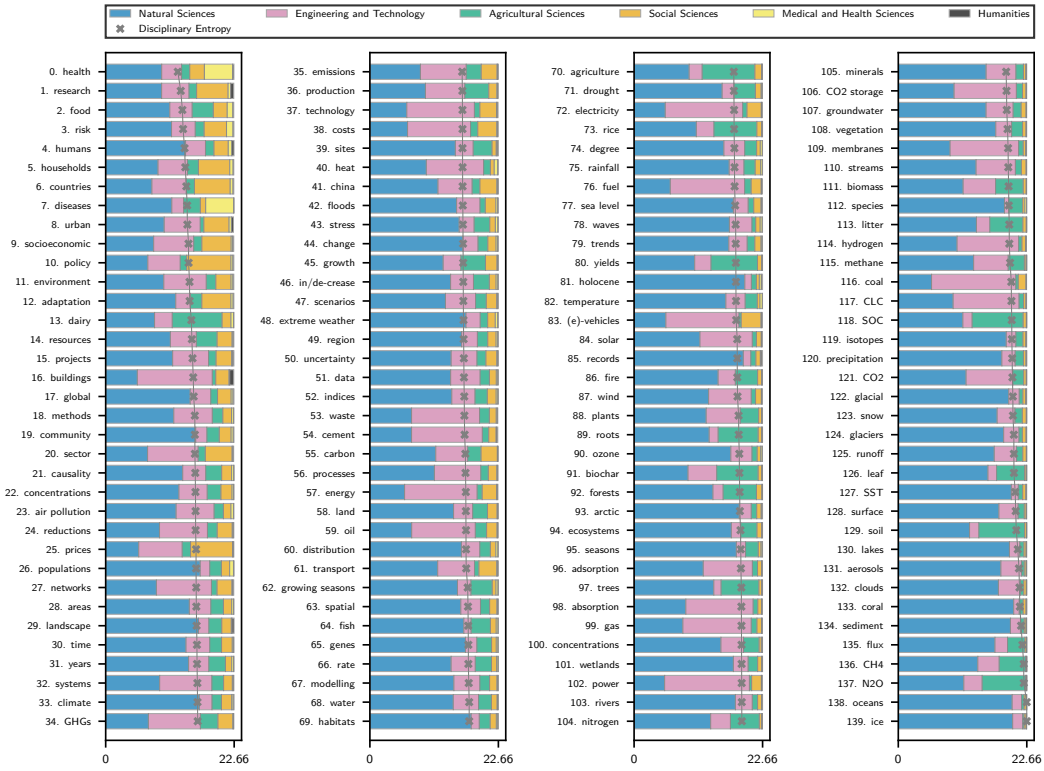


Figure 3.5: Representation in IPCC reports: **a)** by discipline, **b)** by social science proportion of WG 3 topics, **c)** and novelty of all topics, where topics in the highest and lowest 10% of either axis are labelled. Topics are coloured according to the working group from which they receive the most citations. Representation is the share of the subset of documents being cited by the IPCC divided by the share of the subset in the whole literature. We plot on a log scale so that 0.5 is equally distant to 1 as 2; plot labels show real values.

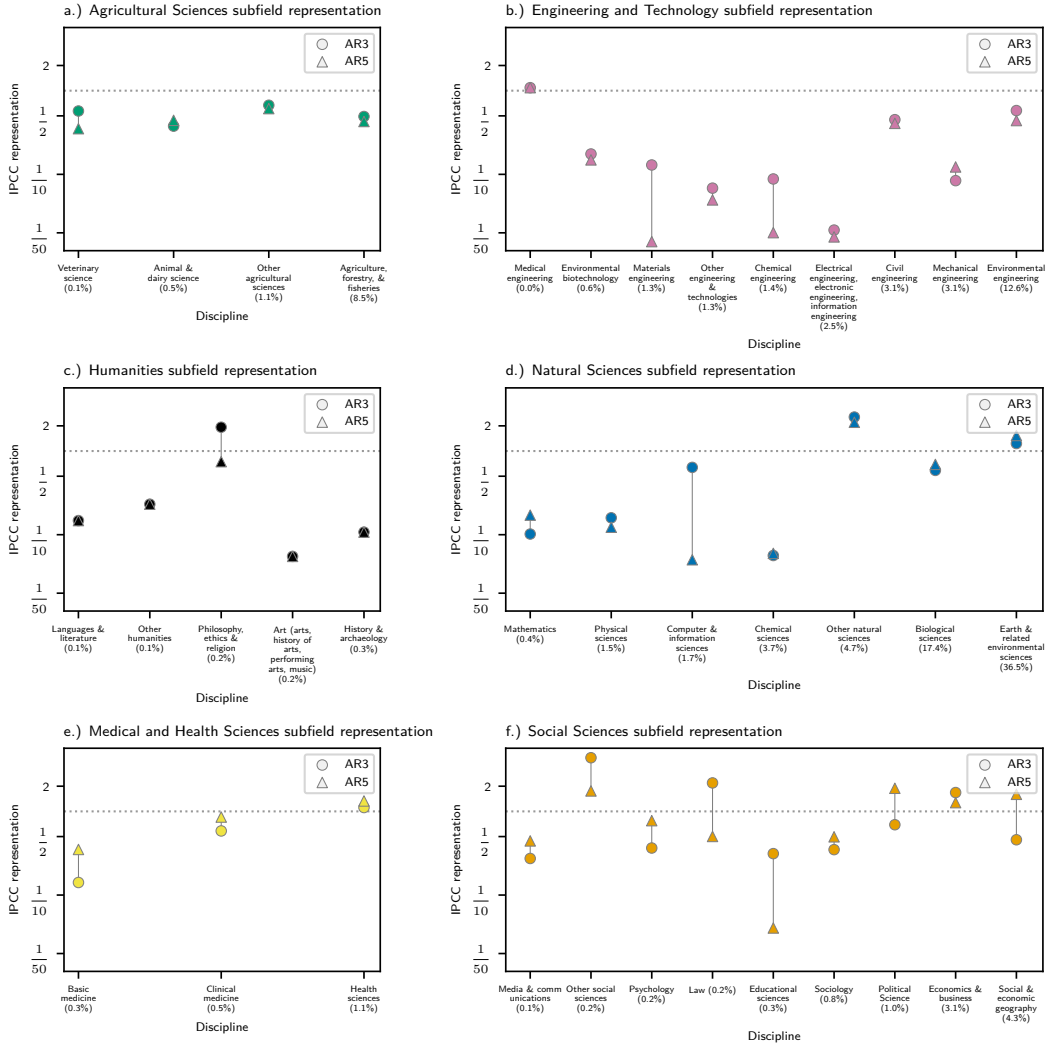
Extended Figures and Tables

AR WG	1	2	3	4	5
1	8%	25%	37%	47%	58%
2	6%	12%	30%	38%	47%
3	3%	9%	15%	22%	35%

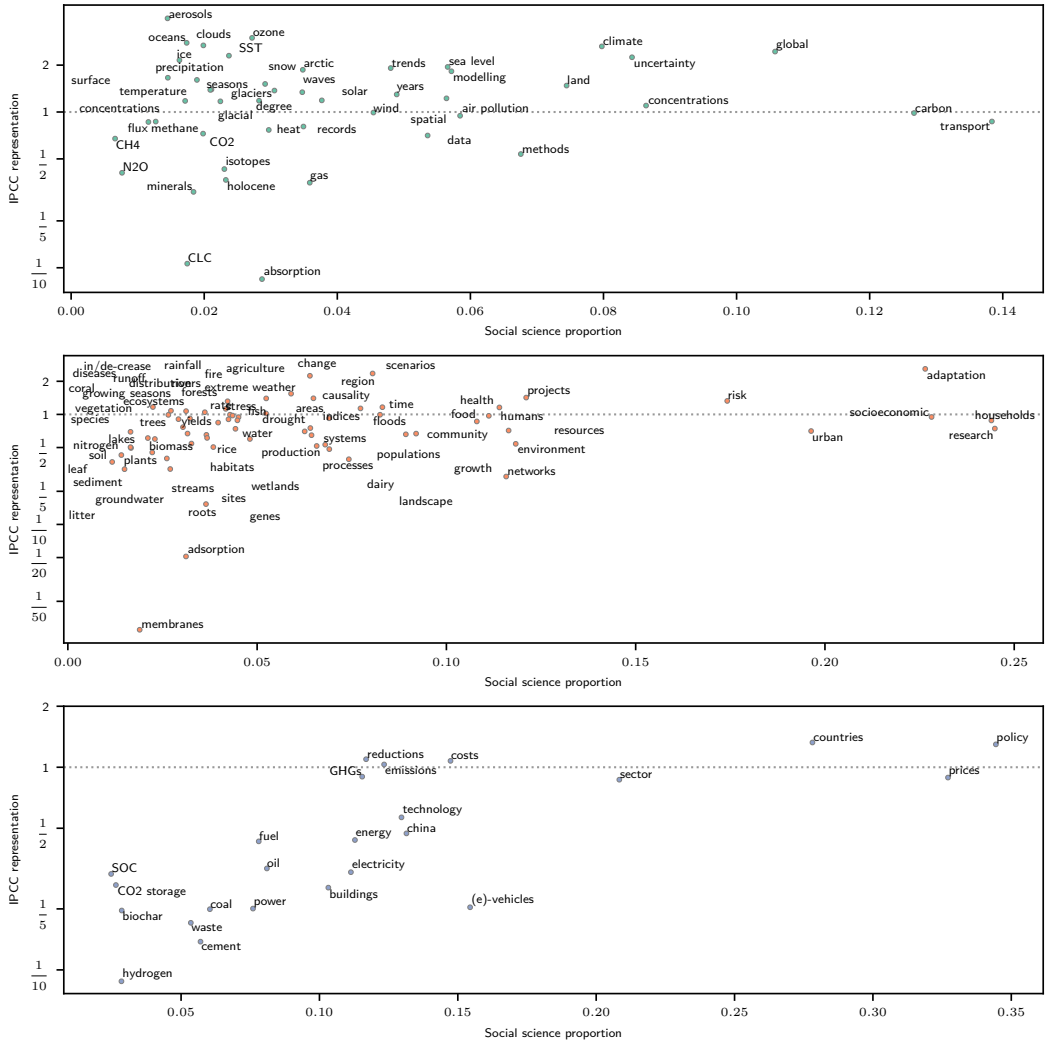
Extended Table 3.2: The proportion of citations in each report that could be matched with a document in our query from the Web of Science



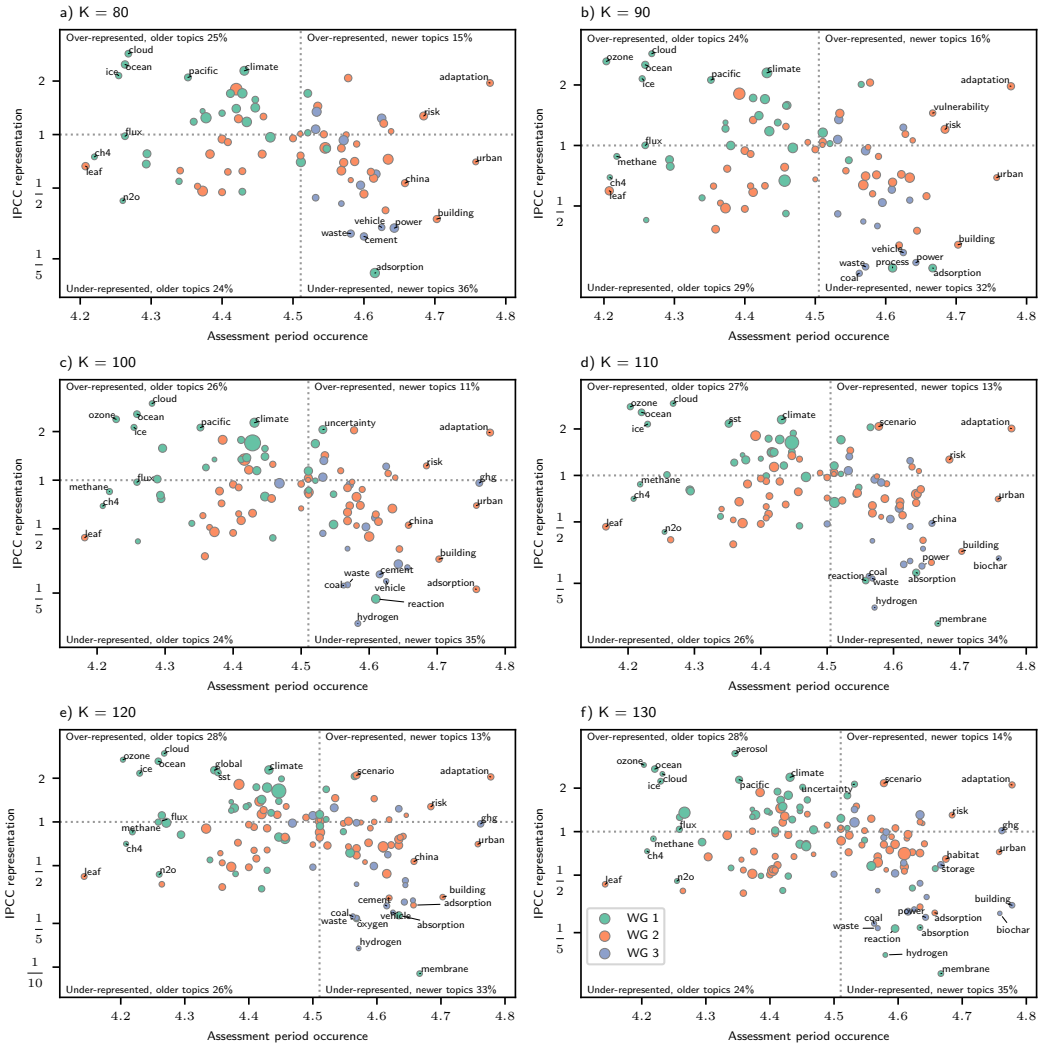
Extended Figure 3.6: Disciplinary Entropy of Topics. Coloured bars show the proportion of each topic made up of papers from each disciplinary category. Crosses show the Disciplinary Entropy of each topic (see methods for details).



Extended Figure 3.7: IPCC Representation by subfield. Representation is the share of the subset of documents being cited by the IPCC divided by the share of the subset in the whole literature. We plot on a log scale so that 0.5 is equally distant to 1 as 2; plot labels show real values.



Extended Figure 3.8: SI Social science & representation in topics across working groups. Representation is the share of the subset of documents being cited by the IPCC divided by the share of the subset in the whole literature. Social science proportion shows the proportion of the total document-topic score coming from documents in the social sciences.



Extended Figure 3.9: Topic representation over different values of K (number of topics). Topics in the upper or lower 6.66th percentile of either dimension are labelled. Representation is the share of the subset of documents being cited by the IPCC divided by the share of the subset in the whole literature. Assessment period occurrence refers to the center of a topic's distribution across assessment periods (see methods for further details).

Bibliography

Beller, E., Clark, J., Tsafnat, G., Adams, C., Diehl, H., Lund, H., Ouzzani, M., Thayer, K., Thomas, J., Turner, T., Xia, J., Robinson, K., Glasziou, P., and On behalf of the founding members of the ICASR group (2018). Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews*, 7(1):1–7.

Bjurström, A. and Polk, M. (2011). Physical and economic bias in climate change research: A scientometric study of IPCC Third Assessment Report. *Climatic Change*, 108(1):1–22.

Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*.

Chalmers, I., Hedges, L. V., and Cooper, H. (2002). A Brief History of Research Synthesis. *Evaluation & The Health Professions*, 25(1):12–37.

Chaney, A. J. B. and Blei, D. M. (2012). Visualizing Topic Models. *Icwsm*, pages 419–422.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. (2009). Reading Tea Leaves : How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems 22*, pages 288–296.

Corbera, E., Calvet-Mir, L., Hughes, H., and Paterson, M. (2016). Patterns of authorship in the IPCC Working Group III report. *Nature Climate Change*, 6(1):94–99.

David G. Victor (2015). Embed the social sciences in climate policy - David Victor. *Nature*, 520:7–9.

Grieneisen, M. and Zhang, M. (2011). The Current Status of Climate Change Research. *Nature Climate Change*, 1:72–73.

Hall, D., Jurafsky, D., and Manning, C. D. (2008). Studying the history of ideas using topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, pages 363–371.

Haunschild, R., Bornmann, L., and Marx, W. (2016). Climate Change Research in View of Bibliometrics. *PLoS ONE*, 11(7):1–19.

Hulme, M. and Mahony, M. (2010). Climate change: What do we know about the IPCC? *Progress in Physical Geography*, 34(5):705–718.

IPCC (2013). Principles governing IPCC work.

IPCC (2014). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland.

Khabsa, M. and Giles, C. L. (2014). The number of scholarly documents on the public web. *PLoS ONE*, 9(5).

Kowarsch, M., Jabbour, J., Flachsland, C., Kok, M. T. J., Watson, R., Haas, P. M., Minx, J. C., Alcamo, J., Garard, J., Riousset, P., Pintér, L., Langford, C., Yamineva, Y., von Stechow, C., O'Reilly, J., and Edenhofer, O. (2017). A road map for global environmental assessments. *Nature Climate Change*, 7(6):379–382.

Le, Q. V. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *ICML*, 32.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91.

Minx, J. C., Callaghan, M., Lamb, W. F., Garard, J., and Edenhofer, O. (2017). Learning about climate change solutions in the IPCC and beyond. *Environmental Science & Policy*.

Moss, R. H., Edmonds, J. A., Hibbard, K. A., Manning, M. R., Rose, S. K., Van Vuuren, D. P., Carter, T. R., Emori, S., Kainuma, M., Kram, T., Meehl, G. A., Mitchell, J. F., Nakicenovic, N., Riahi, K., Smith, S. J., Stouffer, R. J., Thomson, A. M., Weyant, J. P., and Wilbanks, T. J. (2010). The next generation of scenarios for climate change research and assessment. *Nature*, 463(7282):747–756.

Nunez-Mir, G. C., Iannone, B. V., Pijanowski, B. C., Kong, N., Fei, S., and Fitzjohn, R. (2016). Automated content analysis: addressing the big literature challenge in ecology and evolution. *Methods in Ecology and Evolution*, 7(11):1262–1272.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python Fabian. *Journal of Machine Learning Research*, 12:2825–2830.

Rao, V. B., Maneesha, K., Sravya, P., Franchito, S. H., Dasari, H., and Gan, M. A. (2019). Future increase in extreme El Nino events under greenhouse warming increases Zika virus incidence in South America. *npj Climate and Atmospheric Science*, 2(1):2–8.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Chapter 4

AI based evidence and attribution mapping of 100,000 climate impact studies

Abstract

Increasing evidence suggests that climate change impacts are already observed around the world. Global environmental assess- ments face challenges to appraise the growing literature. Here we use the language model BERT to identify and classify studies on observed climate impacts, producing a comprehensive machine-learning-assisted evidence map. We estimate that 102,160 (64,958–164,274) publications document a broad range of observed impacts. By combining our spatially resolved database with grid-cell-level human-attributable changes in temperature and precipitation, we infer that attributable anthropogenic impacts may be occurring across 80% of the world's land area, where 85% of the population reside. Our results reveal a substantial 'attribution gap' as robust levels of evidence for potentially attributable impacts are twice as prevalent in high-income than in low-income countries. While gaps remain on confidently attributing climate impacts at the regional and sectoral level, this database illustrates the potential current impact of anthropogenic climate change across the globe.

4.1 Introduction

There is overwhelming evidence that the impacts of climate change are already being observed in human and natural systems (Cramer et al., 2014). These effects are emerging in a range of different systems and at different scales, covering a broad range of research fields from glaciology to agricultural science and from marine biology to migration and conflict research (IPCC, 2014), The evidence base for observed climate impacts is expanding (Hansen, 2015), and the wider climate literature is growing exponentially (Haunschild et al., 2016; Bornmann and Mutz, 2015). Systematic reviews and systematic maps offer structured ways to collectively identify and describe this evidence while maintaining transparency, attempting to ensure comprehensiveness and reduce bias (Haddaway and Pullin, 2014). However, their scope is confined to very specific questions covering no more than dozens to hundreds of studies.

In climate research, evidence assessments of observed climate change impacts are performed by the Intergovernmental Panel on Climate Change (IPCC) (IPCC, 2014).Since the first assessment report (AR) of the IPCC in 1990, we estimate that the number of studies relevant to observed climate impacts published per year has increased by more than two orders of magnitude (Fig. 4.10a). Since the third AR, published in 2001, the number has increased tenfold. This exponential growth in peer-reviewed scientific publications on climate change (Haunschild et al., 2016; Bornmann and Mutz, 2015) is already pushing manual expert assessments to their limits. To address this issue, recent work has investigated ways to handle big literature in sustainability science by scaling systematic review and map methods to large bodies of published research using technological innovations and machine-learning methods (Callaghan et al., 2020b; Porciello et al., 2020; **?**; Westgate et al., 2018; Lamb et al., 2019). Much of this work builds on a related literature that has applied natural language processing (NLP) techniques to problems of evidence synthesis in the health sciences (Cohen et al., 2006; Marshall et al., 2017).

Fully utilizing the available knowledge on emerging climate change impacts is key to informing global policy processes (Schleussner and Fyson, 2020) as well as local risk assessments and on-the-ground action on climate adaptation (Fankhauser, 2017; Bedsworth and Hanak, 2010). While the global policy process may be served well with literature assessments presenting results aggregated on the level of continents or world regions (IPCC, 2014, 2012), informing climate adaptation typically requires

more highly localized and contextualized information on climate impacts (Hallegatte and Mach, 2016; Conway et al., 2019).

Another core challenge of literature reviews and assessments of observed climate impacts relates to the question of whether climate impacts can be attributed to anthropogenic forcing (Hansen and Stone, 2016). While anthropogenic climate change signals have been identified in observed trends in a number of variables (Hansen and Stone, 2016), including temperature (Knutson et al., 2013), precipitation (Knutson and Zeng, 2018), sea level rise (Nerem et al., 2018), water resources (Gudmundsson et al., 2019), and selected extreme weather events (Padrón et al., 2020) have been identified, the confidence in these assessments is still subject to substantial regional variations and remains relatively tentative at smaller spatial scales even if very high confidence levels can be reached for larger-scale (for example, global scale) attribution findings. Confidence also strongly depends on the variable being considered and specifically decreases further down the impact chain, that is, for indicators of changes in human and natural systems that are driven by changes in other climate impact variables Hansen and Stone (2016). In addition, methodological approaches and robustness criteria for climate change attribution differ widely among studies and disciplines, requiring expert judgement on a case-by-case basis to compile a comprehensive evidence base.

This points towards the added value of joining the body of evidence documenting regional or local-scale studies about climate impacts linked to common climate drivers such as temperature and precipitation change to a spatially resolved detection/attribution database of those variables.

Using Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art deep-learning language representation model (Devlin et al., 2019a), we develop a machine-learning pipeline to identify, locate and classify studies on observed climate impacts at a scale beyond that which is possible manually (Figure 4.1). We combine this spatially resolved dataset with an approach to attributing observed trends in surface temperature and precipitation at the grid-cell level (5°x 5°and 2.5°x 2.5°cells, respectively) to human influence on the climate. In doing so, we establish a new paradigm for assessing the impacts of climate change across human and natural systems.

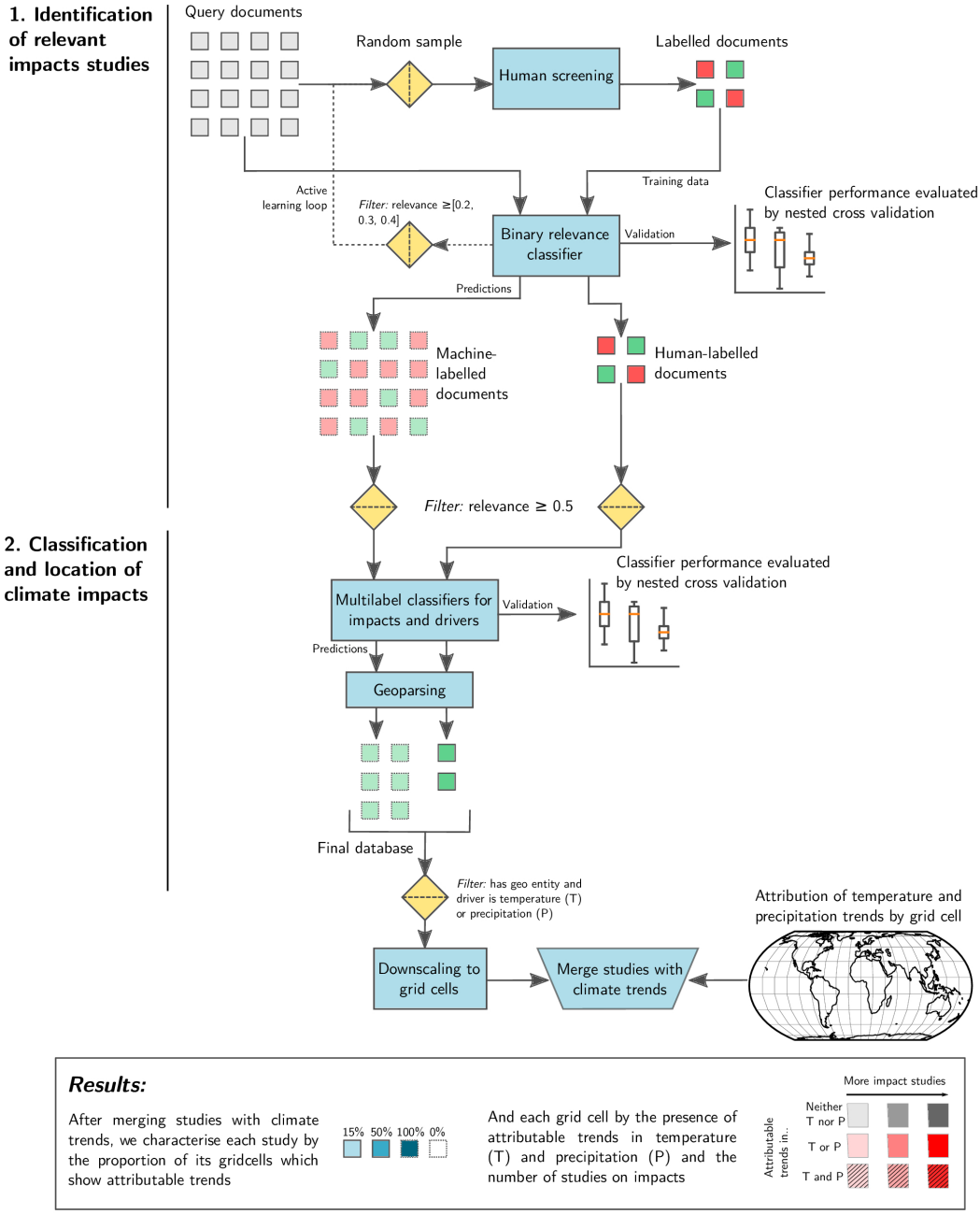


Figure 4.1: A visual representation of the workflow of our machine learning assisted attribution map. Squares represent documents (not to scale), boxes represent the steps taken. Documents are screened by hand, and those labels are used to generate predictions and machine label documents. These machine-labelled documents are matched by location with information from observations and climate models on the detection and attribution of trends in temperature and precipitation.

4.2 Methods

4.2.1 Data Collection

Potentially relevant documents were assembled by developing a query to search bibliographic databases. To validate the query, we tested this against a set of records known to be relevant. Tables 18.5-18.9 in the contribution of Working Group II to the Fifth Assessment Report of the IPCC (AR5 WGII) (IPCC, 2014) contain the studies considered in their assessment of the observed impacts of climate change. After extracting these references, we built a query that would return all of the references in the tables that specifically referred to the role of climate change (rather than of counterfactual explanations for impacts). The query is reproduced in the Supplementary Information (in the format for Web of Science) and is made up of three lists of keywords linked with boolean ANDs. The first set of keywords refer to climate and climate variables, the second to impacts, and the third to observations and attribution.

The query was performed on Scopus and the following citation indices from the Web of Science Core Collection:

- Science Citation Index Expanded (SCI-EXPANDED) –1900-present
- Social Sciences Citation Index (SSCI) –1900-present
- Arts & Humanities Citation Index (A&HCI) –1975-present
- Conference Proceedings Citation Index- Science (CPCI-S) –1990-present
- Conference Proceedings Citation Index- Social Science & Humanities (CPCI-SSH) –1990-present
- Emerging Sources Citation Index (ESCI) –2015-present

The queries were updated on October 19 2020: Web of Science returned 411,194 documents, and Scopus returned 476,778 documents. The total number of records after deduplication through fuzzy title and publication-year matching using trigram similarity was 601,667. The queries were imported into a database and deduplicated using the NACSOS review platform (Callaghan et al., 2020a).

4.2.2 Inclusion and exclusion criteria

We take a broad definition of climate impacts to include all studies relevant to understanding the observed impacts of climate change. This includes the following (with some documents belonging to more than one category):

- Studies that explicitly link impacts to climate change (8% of coded studies)
- Studies that link impacts to trends in climate drivers such as temperature or precipitation (42% of coded studies)
- Studies that link impacts to extreme climate events (6% of coded studies)
- Studies that link impacts to variation in climate drivers (39% of coded studies)
- Studies that document regional or local climate trends (11% of coded studies)

Documents that provide only evidence of likely future impacts of climate change were excluded.

With this broad definition of climate impacts evidence, we do not claim that each study alone is evidence of the impacts of climate change. Rather, taken together, and in the context of observations and climate models, this collection of included studies constitutes the evidence base necessary for understanding climate impacts.

4.2.3 Coding impacts and drivers

Where documents were selected for inclusion, reviewers coded the attribution category, the climate impacts and the drivers (where appropriate) for each paper. Impacts and their drivers were chosen from a selection of 75 specific categories, which were aggregated according to the hierarchy of categories included in the supplementary file category_aggregation.csv. Ninety-three percent of included studies coded impacts in one or more of the five broad impact categories used by IPCC AR5:

- Mountains, snow and ice (11.42% of included studies)
- Rivers, lakes and soil moisture (21.27% of included studies)
- Terrestrial ecosystems (33.13% of included studies)
- Coastal and marine ecosystems (13.21% of included studies)

- Human and managed systems (21.42% of included studies)

Remaining studies documented only trends in climate variables without reference to any of these systems.

4.2.4 Screening and coding

A total of 2,373 documents were screened by members of the author team using the NACSOS platform48, of which 1,125 were included as relevant and coded for impacts and drivers. The median number of documents coded per user was 133, and the mean was 173.

In addition, documents extracted from the tables 18.5–18.9 in AR5 WGII were automatically labelled as relevant and tagged with the broad impact categories corresponding to the table in which they were found.

To mitigate a highly unbalanced sample (few relevant documents among many irrelevant documents), and to make best use of reviewing resources, some documents were selected for screening using an adapted active learning pipeline. With active learning, a classifier (see following section for details) is trained using existing screening decisions to predict the relevance of documents yet to be reviewed. Usually, reviewers screen subsequent documents in decreasing order of predicted relevance, and the classifier is periodically updated with the new data that have been generated. Given that our goal was not to screen all relevant documents but to generate useful labels efficiently, we created samples with relevance predictions greater than 0.2, 0.3 and 0.4 to exclude documents with a low likelihood of being relevant. Documents were first screened by a small group of reviewers who developed the categorization scheme for impacts and drivers. A subsequent set of documents was screened by all reviewers, and differences in coding were discussed and alterations recorded. Reviewers were then split into teams corresponding with the AR5 impact categories according to expertise and screened documents predicted to be rather relevant (¿0.33) to the given category. Each team screened a sample of documents and discussed differences in screening and coding decisions. Teams reached average Cohen's Kappa scores between 0.66, indicating substantial agreement, and 1.0, indicating full agreement (McHugh, 2012). After this initial round of double coding, reviewers proceeded to screen documents individually. Additional documents were selected for screening using keyword searches (https://github.com/mcallaghan/regional-impacts-map/blob/

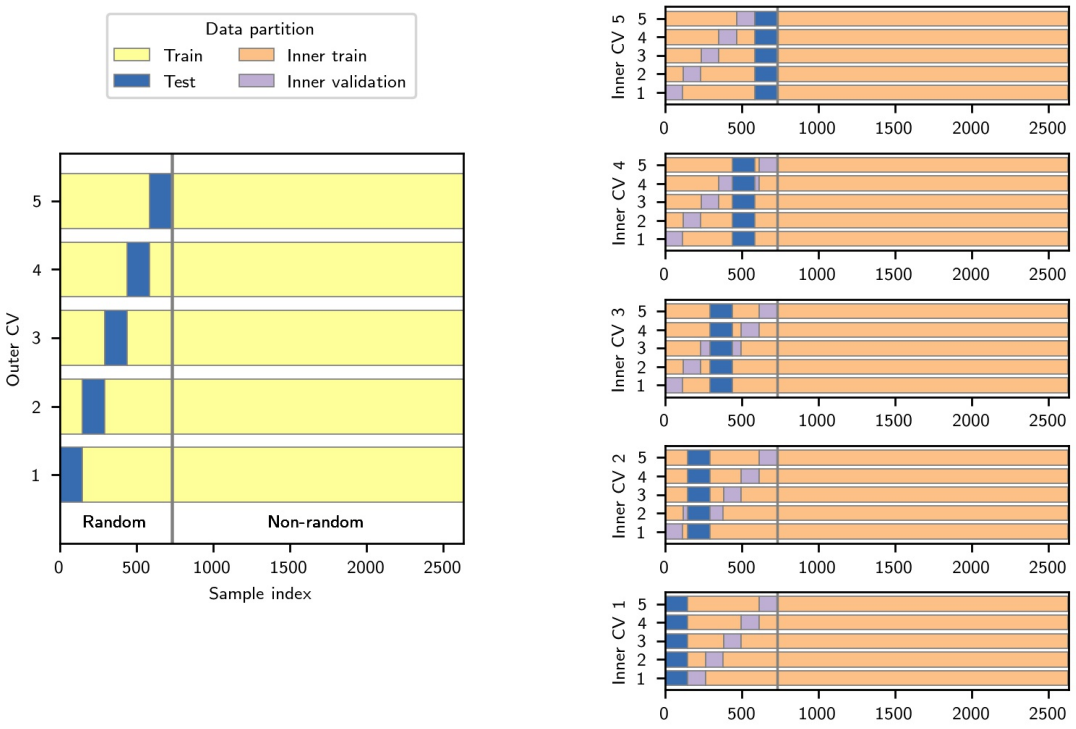


Figure 4.2: Nested cross validation (CV) procedure for the binary relevance classifier. Models are fit using training documents and evaluated on validation/test documents. The inner CV loop is used to search for optimal hyperparameter settings, which are then evaluated on the outer test sets.

master/literature_identification/category_keywords.ipynb) to identify documents from infrequently appearing subcategories.

Because the documents selected using the methods described are unlikely to be representative of the full set of documents returned by the query, we also screened 732 documents drawn at random, which we used for validation.

4.2.5 Machine-learning classifiers for inclusion, impact type and drivers

We first trained a binary classifier to predict the inclusion/exclusion decision given by reviewers. We use a nested cross-validation (CV) procedure (Figure 4.2) to optimize parameter settings and evaluate the performance of a support vector machine (SVM) classifier (Chang and Lin, 2011) as well as a pretrained DistilBERT model fine-tuned

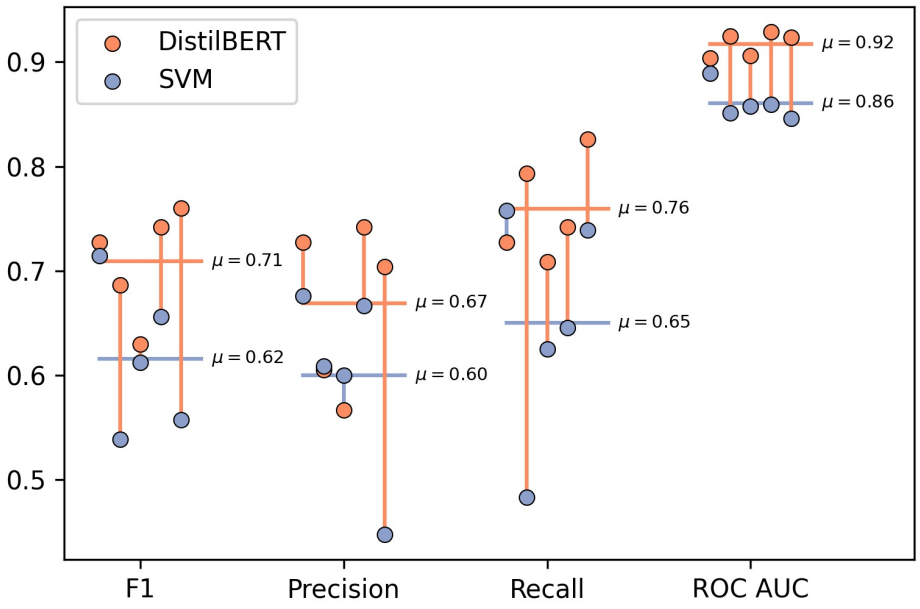


Figure 4.3: Performance metrics for the binary inclusion/exclusion classifier. Each pair of dots represents the scores for a distinct cross-validation fold. Horizontal lines show the mean score across folds.

with out labelled dataset Sanh et al. (2020). SVMs have a long history of applications in evidence synthesis (Cohen, 2006), while the BERT (Devlin et al., 2019b) model recently achieved state-of-the-art results in a variety of NLP challenges and has begun to be used in evidence synthesis pipelines (Porciello et al., 2020). However, large language models such as BERT can have non-trivial climate impacts (Bender et al., 2021), motivating our decision to use the lighter and faster DistilBERT, which retains "97% of its language understanding" (Sanh et al., 2020) with greatly reduced computational resource usage.

In our nested cross-validation procedure, we first separate those documents that were drawn at random from the population of documents identified by the query from the remaining unrepresentative documents. Only randomly selected documents are used in validation and test sets to ensure that the estimation of the performance of the classifier on the whole dataset is not biased. In the outer fold of the cross-validation loop, a separate test set is drawn from the randomly selected documents for each fold,

k, and all other documents are assigned to the test set. The inner cross-validation loop draws k inner validation sets from the remaining random documents in the training set and allocates all other documents in the training set to an inner training set. The inner loop is used to optimize hyperparameters for each model using grid search: a model is initialized with each combination of hyperparameters and fit on each inner training set and evaluated on each inner validation set. The combination of hyperparameters with the best mean F1 score across inner folds is selected as the best model. This model is fit with the training data from the outer cross-validation and evaluated with the test data. The outer cross-validation thus returns k scores for each metric, which we report in the following. We note that our cross-validation approach, while transparent, robust and thorough, is computationally expensive and that alternative procedures such as random search may provide similar results at lower computational cost, or minor improvements at the same cost (Bergstra and Bengio, 2012). In principle, additional improvements to the model may also be generated through additional pre-training (Gururangan et al., 2020) using the unlabelled corpus of climate-relevant abstracts. Pre-training BERT-like models on climate science corpora remains an area for future investigation.

We evaluated our binary inclusion/exclusion classifiers with five inner and outer folds. DistilBERT clearly outperformed SVM across all metrics, achieving an average F1 score of 0.71 and an average ROC AUC score of 0.92 (Figure 4.3). A final DistilBERT model configuration was chosen using the same procedure on the outer folds. Each combination of parameter settings was tested on each outer fold, and the combination of parameter settings with the highest mean F1 score was selected.

This final model was used to predict the relevance of all remaining documents. To create a confidence interval for each prediction, five versions of the final model were trained on five folds of the data. Upper and lower estimates for each document are given by the mean plus or minus one standard deviation. All documents where the lower estimate was below 0.5 were excluded from the study.

We then trained multilabel classifiers to predict the impact category and the driver category of included documents. Classifiers parameters were optimized and classifiers evaluated with the same nested cross-validation method using only those labelled documents that were included. Because documents selected for screening using the active learning process are broadly representative of the documents to which the multilabel classifiers are applied, all documents selected in this manner are also used for validation.

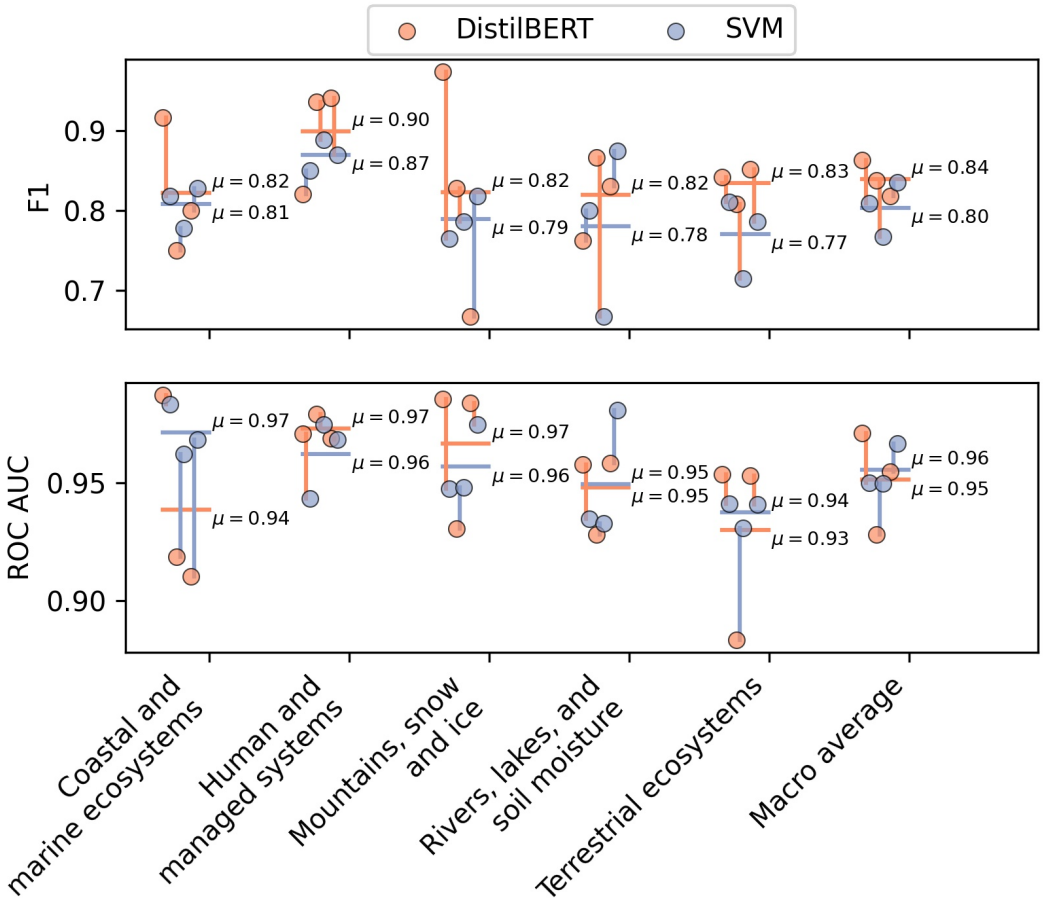


Figure 4.4: Receiver operating curve area under the curve scores (ROC AUC) and F1 scores for the classification of impact categories. Each pair of dots represents the scores for a distinct cross-validation fold. Horizontal lines show the mean score across folds.

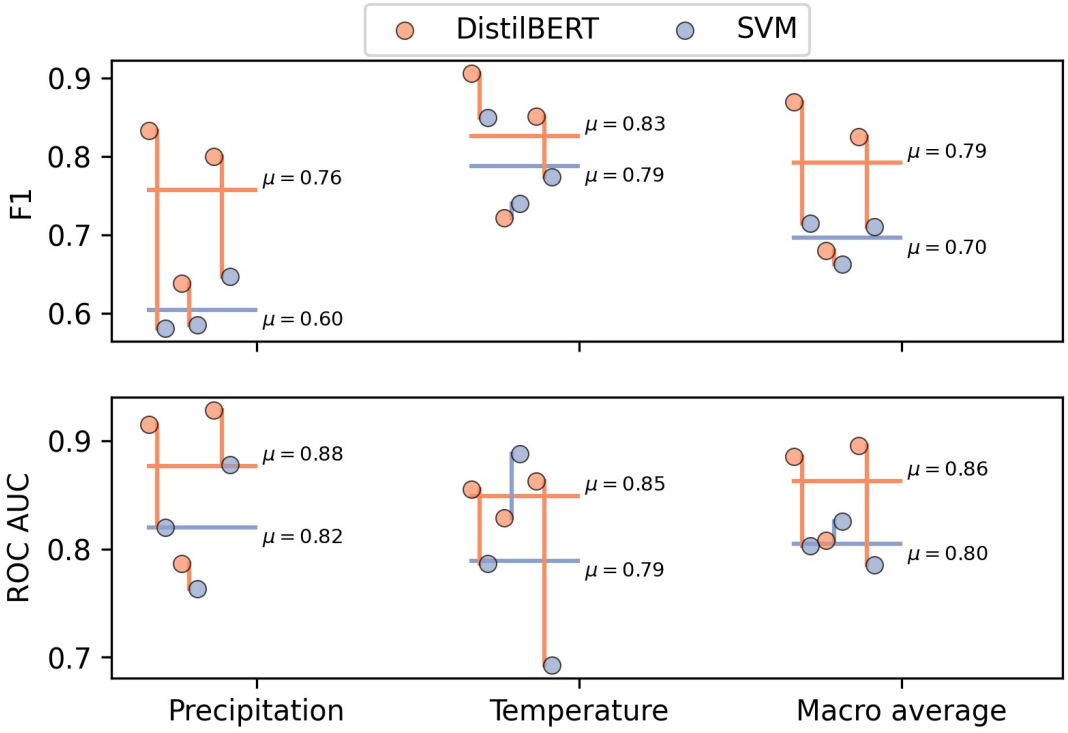


Figure 4.5: Receiver operating curves area under the curve scores (ROC AUC)(ROC) and F1 scores for the classification of drivers. Each pair of dots represents the scores for a distinct cross-validation fold. Horizontal lines show the mean score across folds.

Due to the lower number of documents, and lower number of documents drawn from a random sample in this set, we used a smaller k value of 3 for cross-validation. We treat each class equally and optimize using the macro F1 score. For the prediction of impact categories, DistilBERT outperforms SVM, achieving a macro-averaged F1 score of 0.84 and a macro-averaged ROC AUC score of 0.95 (Extended Data Fig. 4.). For classification of climate drivers, we optimize for the macro-averaged F1 score for the categories temperature and precipitation. DistilBERT outperforms SVM, achieving an average F1 score of 0.79 and an average ROC AUC score of 0.86. Where no individual class has a prediction larger than 0.5, documents are classed as 'Other systems'.

4.2.6 Detection and attribution

To put our database of impact studies in context, we match studies with grid-cell-level detection and attribution of temperature and precipitation trends to human influence on the climate.

Updating attribution of temperature and precipitation trends

We followed a previously published methodology (Knutson et al., 2013; Knutson and Zeng, 2018) used to attribute observed temperature and precipitation trends to human influence around the globe, at the level of typical climate model grid cells (5 °grid boxes for temperature and 2.5 °grid boxes for precipitation. It relies on a comparison of local trends in observational datasets for temperature (HadCRUT4 version 4.6 (Morice et al., 2012)) and precipitation (GPCC v2018, [1]), with those produced in climate model runs from Coupled Model Intercomparison Project Phase 6 (CMIP6) (Eyring et al., 2016) simulating climate over the historical period under the influence of all forcings (i.e., both natural and anthropogenic, referred to as "ALL") or natural forcings only (referred to as "NAT").

We analysed the outputs of these simulations from ten CMIP6 models: MIROC6, IPSL-CM6A-LR, CanESM5, HadGEM3-GC31-LL, CNRM-CM6-1, GFDL-ESM4, CCESS-ESM1-5, BCC-CSM2-MR, NorESM2-LM and CESM2. The model selection was based on the availability of ALL and NAT as well as 'piControl' runs (simulating internal climate variations in the absence of external forcings, apart from a constant solar forcing). The analysis provides a test of the ability of the corresponding ALL simulations

[1] obtainable from https://psl.noaa.gov/data/gridded/data.gpcc.html

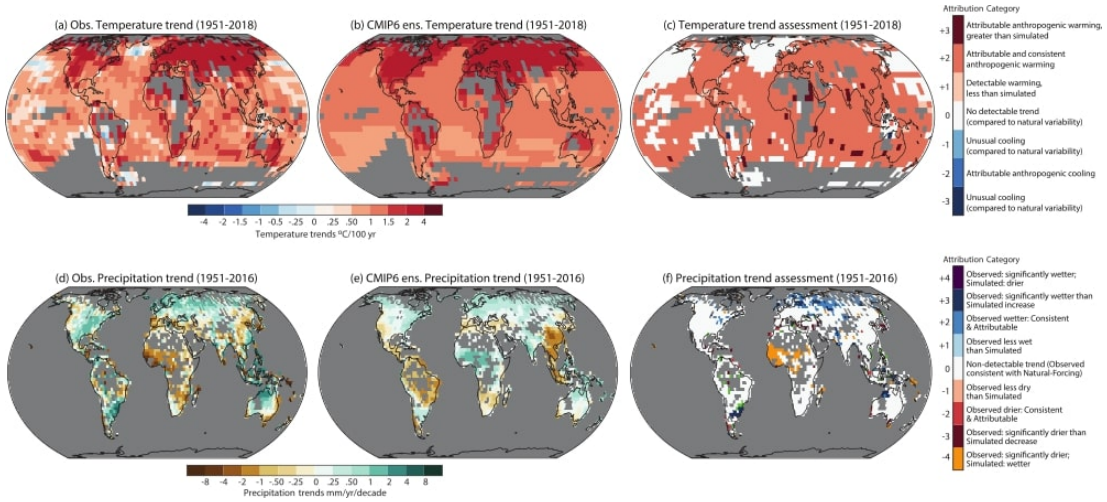


Figure 4.6: Geographical distribution of surface trends. Temperature from 1951 to 2018 (left) and precipitation trends from 1951 to 2016 (right) in (a),(b) observations and (c),(d) CMIP6 10-model ensemble mean all-forcing runs. Bottom panels (e),(f) show observations categorised into attribution categories, following Knutson et al. (2013); Knutson and Zeng (2018), respectively. Observed cooling/warming or drying/wetting trends that–after accounting for internal climate variability–are inconsistent with the simulated response to natural forcings but consistent with the simulated response to both natural and anthropogenic forcings are indicated by categories -/+2. This is clearest case of changes that are at least partially attributable to anthropogenic forcing, according to the CMIP6 ensemble. Categories -/+1 have detectable observed changes, but are not assessed as attributable to anthropogenic forcing because the observed changes are significantly less than those simulated in the average all-forcing runs. Categories -/+3 have detectable changes and are assessed as at least partly attributable anthropogenic forcing, although the observed changes are inconsistent with the all-forcing runs. That is, they are in the same direction as, but are significantly stronger than, the mean of the all-forcing runs. Categories -/+4 represents cooling/warming or drying/wetting trends that are inconsistent with the simulated response to natural forcings but whose sign is opposite to that of the average simulated all-forcing response; category 0 represents trends that are not distinguishable from natural variability alone. Categories -/+4 and 0 are considered to be examples of non-detectable trends).

to reproduce the regional trends in annual mean temperature and precipitation against observational data (Beusch et al., 2020). For some models, the ALL simulations were not available after 2014, in which case we combined them with the first few years of the ssp585 simulations of future climate conditions to match the length of the observational data.

Linear trends over the 1951–2018 (for temperature) and 1951–2016 (for precipitation) periods were computed over each grid cell with adequate data for each observational dataset, following the criteria of Knutson et al. (2013); Knutson and Zeng (2018) (Figures 4.6a&b). For temperature, we computed a linear trend for each ensemble member of the HadCRUT4 dataset, from which observed trend distributions were derived. Precipitation trends were not computed over grid cells where less than 20% of data was available for the first or last 10% of the observed time series or where the entire time series had less than 70% of data available. For temperature, we divide the trend period into five roughly equal periods and require that each period has at least 20% temporal coverage for annual means. We consider an annual mean as available if at least 40% of the months are available for the year.

To be compared with the observational data, for each model the data from both the ALL and NAT runs were first re-gridded onto the observational grids (5°x 5°for temperature and 2.5° x 2.5° for precipitation), excluding times and grid locations where observed data were missing, before linear trends were computed over each grid cell in which adequate temporal coverage was available (Figures 4.7c,d). For each model, we then assessed the potential effect of internal variability by computing trends of the length being investigated in 50 random samples of the corresponding piControl runs from each model. The model control runs had beforehand been corrected for any long-term drift and the anomaly series adjusted by a factor to ensure consistency of low-frequency variability between model control runs and estimated internal variability from observations (further discussed in the following). We then combined the resulting trend distributions from the piControl runs with the trends computed in the ensemble mean of ALL and NAT runs. Following previous studies (Knutson et al., 2013; Knutson and Zeng, 2018), the final distribution for temperature was based on an aggregate distribution of all constructed model trend distributions (and thus included the spread of different model ensemble means) whereas for precipitation, an average distribution of model trends across the ensemble was used (that is, the distribution had the average

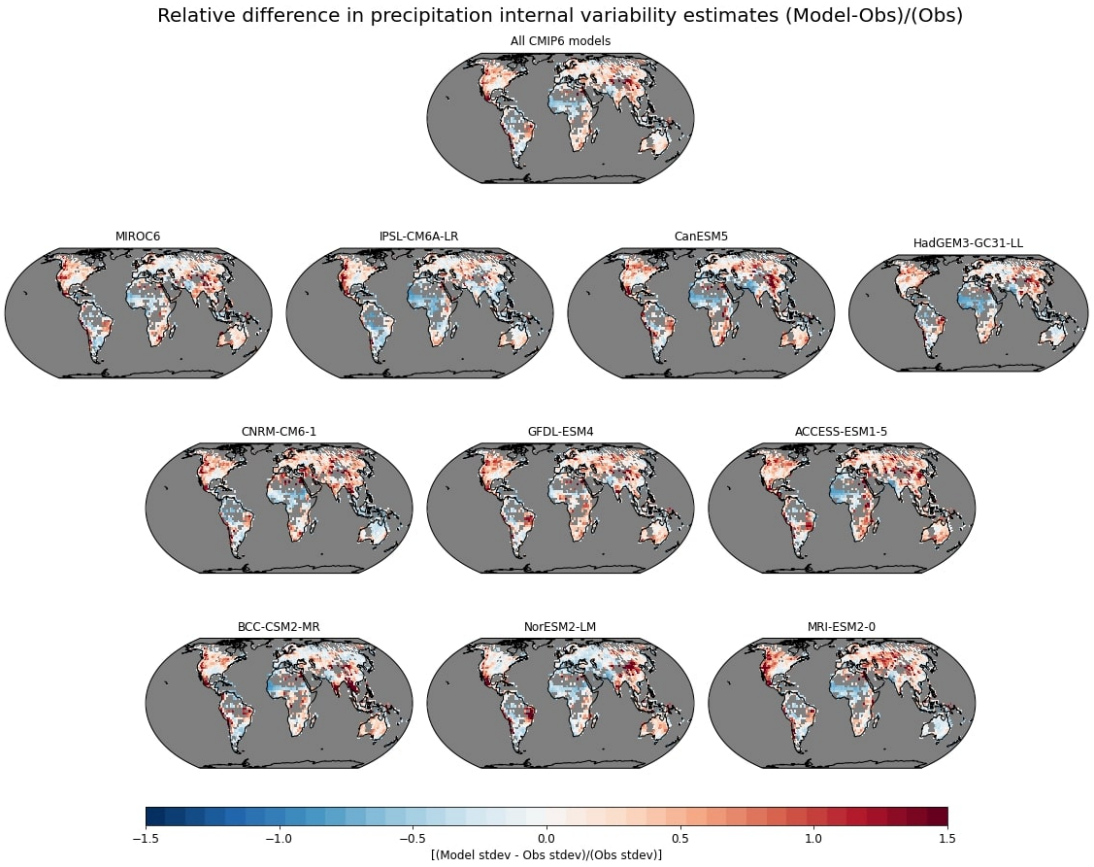


Figure 4.7: Fractional difference between average CMIP6 modeled low-frequency standard deviation of annual mean precipitation vs observed precipitation. To estimate the internal low-frequency variability for both models and observations, the observed time series were detrended and low-pass filtered with a 7-year running mean filter prior to computing the standard deviations while for the models we used the full available control runs (7-yr running mean filtered) to estimate the internal low-frequency variability for each model. The top panel shows the multi-model ensemble standard deviation comparison while the ten individual panels below it show the comparison for each individual CMIP6 model used in the study. The fraction difference was computed as: [(Model st. dev. - Observed st. dev.) / (Observed st. dev.)].

characteristics of the ten CMIP6 models).

Attribution categories were assigned to grid cells (Figure 4.6e,f) on the basis of where their observed trend (or trend distribution in the case of temperature) lay relative to the final trend distributions derived from the ALL and NAT runs. Over the grid cells where an observed trend was in the same direction (sign) as the mean of the ALL trend distribution and was outside the trend distribution 5th–95th percentile range for the NAT simulations, the observed trend was categorized as -3 (+3), -2 (+2) or -1 (+1) depending on whether it was significantly stronger, the same, or weaker than the simulated decrease (increase). Categories -3 (+3) and -2 (+2) are defined as decreases (increases) that are detectable and at least partially attributable to anthropogenic forcing, according to our methodology. Categories -1 (+1) are detectable but not attributable. If the observed trend was significantly different from the NAT distribution, but was in the opposite direction to the mean of the All-Forcing distribution, it was categorized as -4 (observed decrease, modelled increase) or +4 (observed increase, modelled decrease). All observed trends (or trend distributions, in the case of temperature) that intersected with the 5th–95th percentile range of the corresponding trend distributions derived from the NAT runs were categorized as non-detectable, or indistinguishable from natural variability (category 0). Note that for cases where observed trends or trend distributions had a different sign of the mean trend from that of the trend distribution derived from the ALL runs, but were within the range of the Nat run distribution, the corresponding grid cells were also categorized as non-detectable (category 0).

Once the grid cells were categorized, in the case of temperature the results were re-gridded to a 2.5° x 2.5° grid to allow superposition with the categories obtained for precipitation.

Our analysis requires the internal variability for each grid location and variable to be estimated via model control runs. To compare observed estimated internal variability and trends with those generated by the model control runs, Figures 4.7 and 4.8 show fractional difference maps for estimated internal low-frequency variability (model versus observed) for each model individually and for the ensemble mean of the modelled variability (the latter being most relevant for our analysis, which is based on combined estimated variability across the models). The observed low-frequency internal variability is estimated by subtracting the multimodel ensemble All-Forcing change

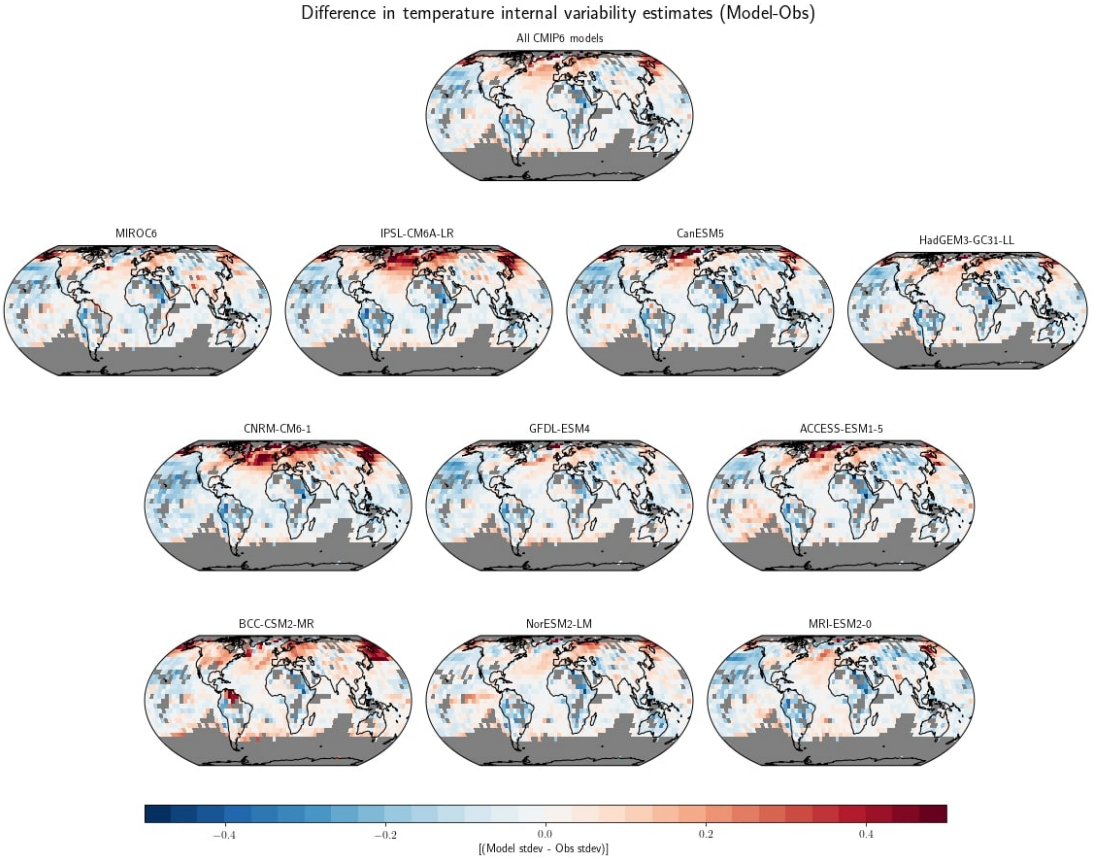


Figure 4.8: Difference between average CMIP6 modeled low-frequency standard deviation (°C) of annual mean surface air temperature vs observed surface temperature. To estimate the internal low-frequency variability for both models and observations, the observed time series were detrended and low-pass filtered with a 7-year running mean filter prior to computing the standard deviations while for the models we used the full available control runs (7-year running mean filtered) to estimate the internal low-frequency variability for each model. The top panel shows the multi-model ensemble standard deviation comparison while the ten individual panels below it show the comparison for each individual CMIP6 model used in the study.

from the observations and computing the standard deviation of the annual residuals, after application of a seven-year running mean filter. For models, we use the simulated variability from the various control runs, again smoothed with the seven-year running mean smoother. The averaged internal low-frequency variability comparison plot for precipitation (Figure 4.7, top panel) shows reds in most regions, indicating that by this measure of internal low-frequency variability, the CMIP6 models tend to overestimate observed variability levels. So our detection results for precipitation will tend to be conservative while, conversely, the ability of All-Forcing to be consistent with observations will tend to be liberal because the modelled spread is relatively wide. However, blue regions are evident in Figure 4.7 in some tropical regions, including over Africa and South America, indicating an undersimulation of internal low-frequency variability there. We took the internal variability comparisons versus observed estimated internal variability in Figure 4.7 and adjusted the control-run variability and trends by the ratio of observed s.d./model s.d. before computing our assessment categories. Results without this variability adjustment (not shown) are broadly similar but show more category -4 (unexplained trends of incorrect sign) over Africa, where internal low-frequency variability appears to be underestimated in models according to this analysis; unadjusted results show slightly less detectable human influence in middle and high latitudes, where internal variability is apparently overestimated in models.

For surface temperature (Figure 4.8) the internal variability comparison results versus observed estimates are similar to those of Knutson et al. (2013) for CMIP3 and CMIP5 with a mixture of results: models tend to simulate more internal variability than the observed estimate in northern mid- to high latitudes, typically less than observed over most other ocean regions at lower latitudes and mixed results over land regions. Whether we include the grid-point-scale adjustment of simulated internal variability in our detection/attribution analysis or not, the results are similar (unadjusted control-run-based assessment not shown). For the assessment of 1951–2018 observed trends (Figure 4.6), there are some additional regions with detectable anthropogenic warming compared with Knutson et al. (2013), but that is as expected since the Knutson et al. analysis examined trends only through 2010. With the termination of the "global warming hiatus" around 2014, the additional recent years have been adding to an ongoing strengthening warming signal and leading to even greater assessed area with detectable anthropogenic warming. In Figure 4.6 and elsewhere in the study, we use the

adjusted control-run results for our assessments for both temperature and precipitation.

4.2.7 Spatial resolution of studies

To match these data with the finest-scale resolution of our database, we resolved each study to the set of 2.5 °grid cells contained by the smallest geographical entity extracted from each paper's title and abstract using the geoparser Mordecai (Halterman, 2017). For each study, we calculated the proportion of the grid cells that this entity corresponds to in which an attributable trend for each variable can be found. For example, Figure 4.9a,b shows that 20 out of Sudan's 27 grid cells show an attributable anthropogenic warming trend, so each study referring to Sudan and documenting impacts predicted to be driven by temperature receives a precipitation trend proportion value of 20/27. Such a study would therefore add towards the dark red bars in Fig. 3, which count studies where an attributable temperature trend can be demonstrated for more than 50% of the grid cells the study refers to.

We also calculate a weighted number of studies for each grid cell by adding 1 divided by the number of grid cells a study refers to to each of those grid cells, and repeating this procedure for all identified relevant studies. Figures 4.9c. and d. show 16 studies which refer to impacts driven by precipitation trends in Sudan. For each of these studies we add 1/27 to each gridcell. Given that some geographical entities were too small to hold one 2.5 degree grid cell, their longitude-latitude values were interpolated to the nearest grid cell instead and the grouped studies apportioned to that one grid cell. Because 3 additional studies refer to Khartoum, for each of them we added 1/1 to the weighted studies value in the grid cell containing Khartoum.

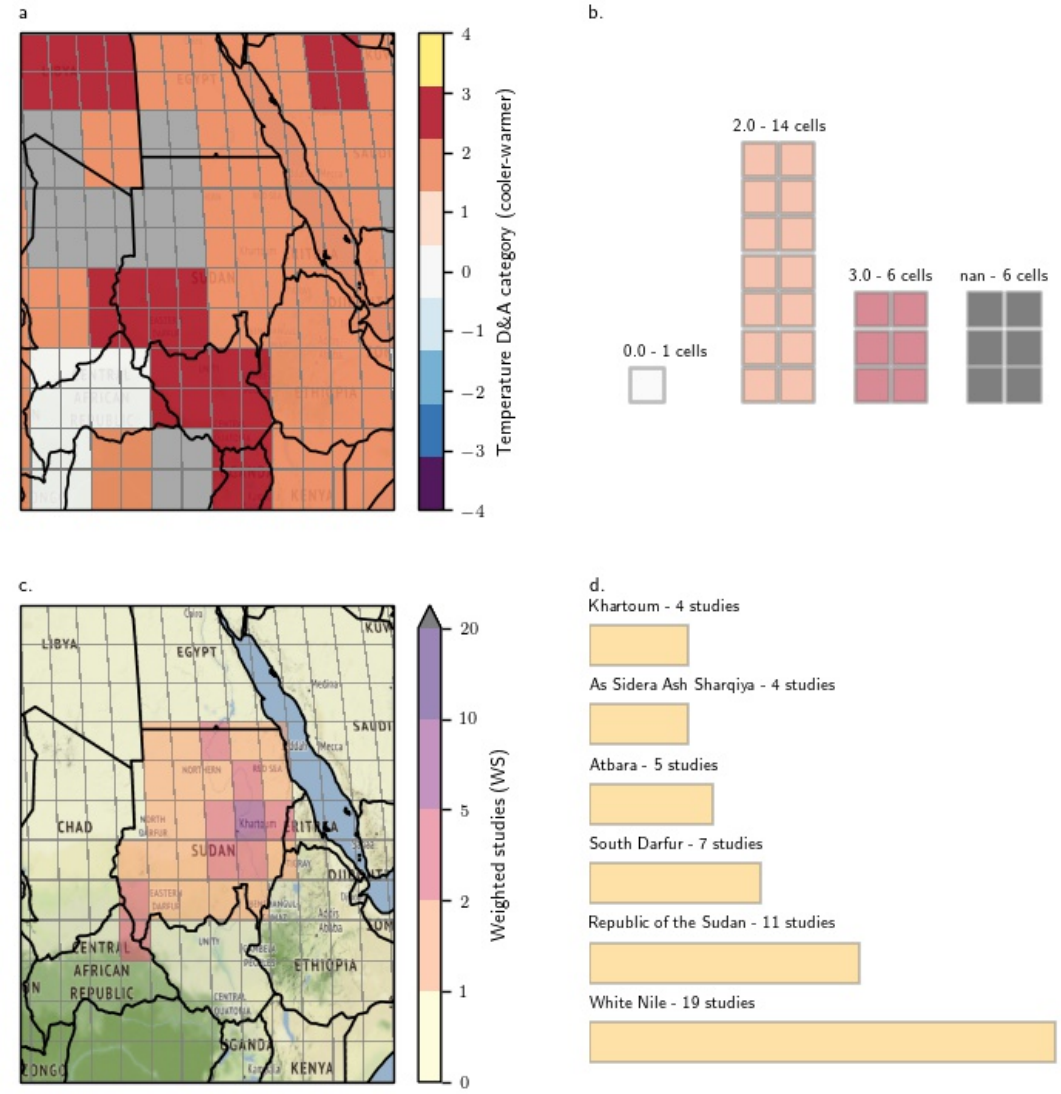


Figure 4.9: An illustration of the spatial resolution and weighting methodology. Detection and attribution categories for temperature in East Africa; b. the number of grid cells of each type in Sudan; c. weighted studies for each grid cell in Sudan; d. The number of studies referring to each extracted geographical location in Sudan.

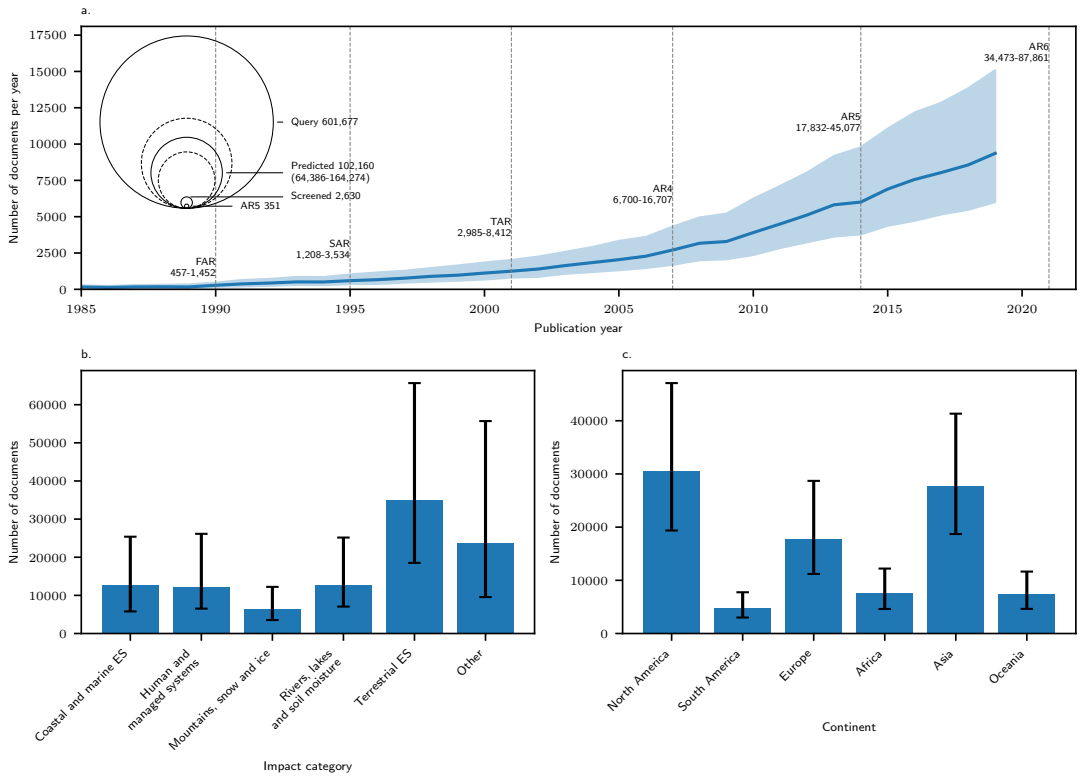


Figure 4.10: All results shown are based on our search queries and subsequent classification by the machine-learning pipeline. Uncertainty ranges denote the number of studies whereby the mean ±1 s.d. for the range of predictions for relevance and category membership obtained via bootstrapping is greater than 0.5 a, Growth in the scientific literature relevant to observed climate impacts over the past 30 years (cumulative totals for IPCC assessment periods are highlighted for reference). Inset: numbers of documents considered in the total query and in the IPCC AR5 WGII Tables 18.5–18.9. b,c, The estimated number of studies for each impact category (b) and continent (c) in our database (note that uncertainty bars consider uncertainty over relevance as well as impact category). ES, ecosystem; FAR, First Assessment Report; SAR, Second Assessment Report; TAR, Third Assessment Report.

4.3 Tens of thousands of impact studies

We searched two large bibliographic databases (Web of Science and Scopus) using an inclusive and transparent search method to systematically identify the literature on climate impacts. We assessed comprehensiveness by ensuring that our search string returned all references from tables 18.5–18.9 in the Fifth Assessment Report (AR5) Working Group II (WGII), which deal with the detection and attribution of climate impacts. Recent breakthroughs in NLP have extended the capabilities of text classification. BERT is a deep-learning language model trained using semi-supervised learning on massive corpora to represent text where word representations depend on context. Such models are able, to some extent, to capture the context-dependent meanings of texts. The pretrained model can be fine tuned on downstream tasks and has achieved state-of-the-art results across a range of NLP tasks. Using training data assembled by collaboratively screening and coding 2,373 abstracts, we use supervised machine learning, fine tuning the smaller and faster BERT variant DistilBERT (Sanh et al., 2020), to classify (also on the basis of the abstract text) documents relevant to understanding the observed impacts of climate change in general and to predict the human or natural systems for which they document impacts (the impact categories), as well as the climate variable(s) driving the documented impacts. Uncertainty estimates for the predictions are derived from bootstrapping. We employ a nested cross-validation approach to hyperparameter tuning, model selection and classifier evaluation and find that our binary inclusion classifier achieves an average F1 score (the harmonic mean of precision and recall) of 0.71 and receiver operating curve area under the curve (ROC AUC) score of 0.92. The prediction of impact type is achieved with an average macro F1 score of 0.84 while the prediction of climate driver is achieved with an average F1 score of 0.79 (Figures 4.1-4.5).

Our query returned 601,677 unique documents (Figure 4.10a), many more than would have been possible to screen by hand. We estimate that 102,160 (64,958–164,274) of these documents are relevant to understanding the observed impacts of climate change in general, judging from the spread of inclusion/exclusion predictions obtained from our model via bootstrapping (Figure 4.10a). This base of relevant publications has grown substantially through the IPCC assessment cycles; 46,426 (34,464–87,824) articles have been published in the sixth assessment cycle so far. This represents more than twice the number of studies published during the AR5 period.

We used a geoparser pretrained using neural networks (Halterman, 2017) to extract structured geographic information from the titles and abstracts of the studies in our database. Although the number of relevant studies in North America, Asia and Europe is much higher than in South America, Africa and Oceania, there is a large body of relevant studies available on all continents (Figure 4.10c). Adjusted for population, the number of papers focusing on Oceania far exceeds the size of the literature devoted to other continents, with Africa and Asia receiving the least attention per million inhabitants. The relevant publications are also unevenly distributed across impact categories, with by far the largest number of studies, 34,974 (18,516–65,631), documenting impacts on terrestrial and freshwater ecosystems (Figure 4.10b). However, the category with the comparably smallest coverage—mountains, snow and ice—still has 6,306 (3,526–12,225) studies.

In contrast to the map of observed impacts produced by the IPCC, we do not include only papers that formally attribute impacts to observed trends in climate. Instead, we take a more comprehensive approach reflecting that our objective is to map all possibly relevant studies on climate-related changes, rather than a list of studies where the relationship between an observed climate trend and specific impacts has been demonstrated with high confidence, or even linked to human influence on the climate. This includes studies attributing impacts to observed trends in climate variables, even where the authors do not attribute these trends to human influence, such as, for example, a study documenting the influence of the date of snowmelt on the phenology and population growth of mammals (Lane et al., 2012). In addition, we include studies that provide evidence on the sensitivity of human or natural systems to climate metrics, such as how heart disease mortality responds to variations in temperature (Zhang et al., 2016). Finally, we include documents describing the impacts of extreme events and studies that detect significant trends in climate variables or climate extremes (Barry et al., 2018), regardless of whether these trends are in line with the expected effects of anthropogenic climate change. We exclude all studies that describe only potential or modelled impacts of future climate change.

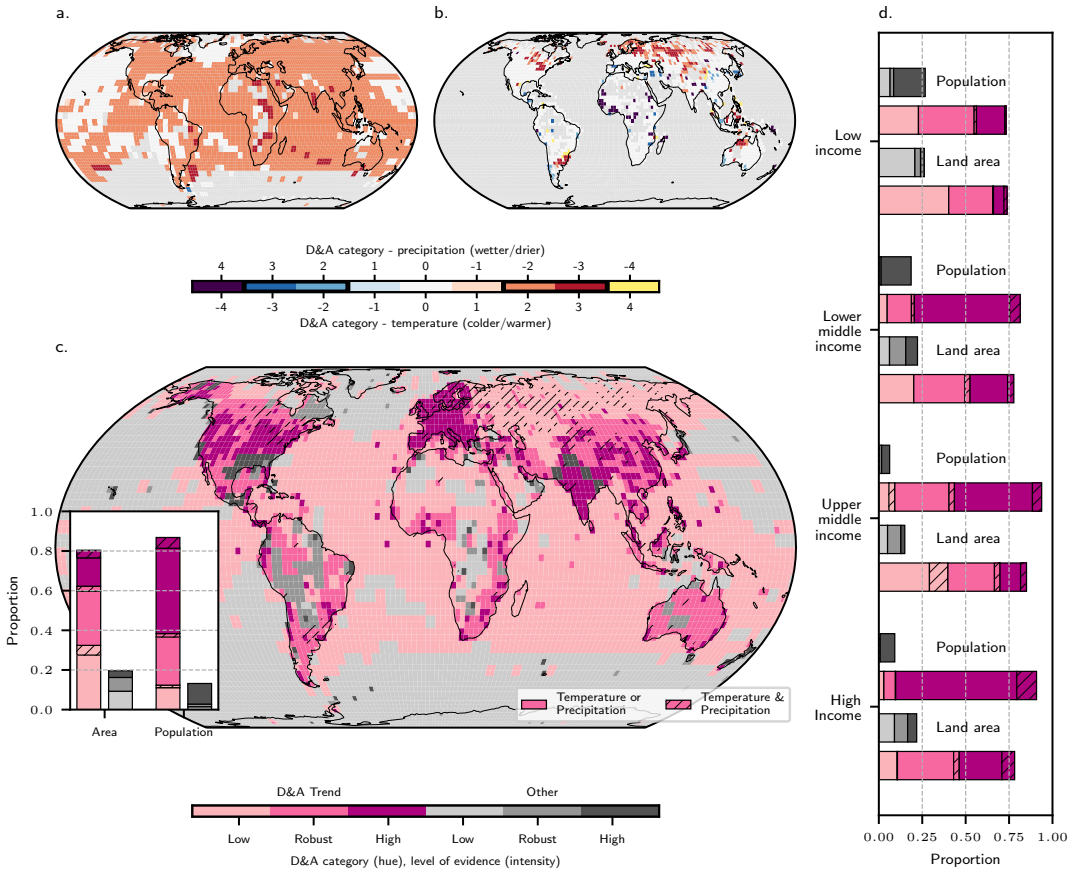


Figure 4.11: Potential attribution of impact studies to regional anthropogenic temperature and precipitation trends. a,b, Model-based assessment of the attribution of regional temperature for the time span 1951–2018 (a) and precipitation trends for the time span 1951–2016 (b) to human influence. Cooling/warming or drying/wetting trends in the regions marked as categories +-2 and +-3 are assessed as attributable in part to human influence (Methods). c, Global map of area-weighted studies coloured by the existence of detectable and attributable (D&A) trends (purple for attributable trends in at least one variable, cross-hatched for attributable trends in both variables, grey for no attributable trends) and indicating the localized evidence density (Low: ¡5 weighted studies; Robust: 5–20 weighted studies; High: ¿20 weighted studies). d,e, The proportion of land area (d) and population (e) with each grid-cell type, grouped by country income category.

4.4 Combining geolocated literature with climate information

To add context on the role of anthropogenic climate change in driving impacts, or more precisely the role of historical changes in anthropogenic climate forcing agents such as greenhouse gases and aerosols, we combine our literature database of studies selected using machine learning with spatially explicit analysis of detectable and attributable trends in two key climate variables. Combining evidence from climate model simulations and observational datasets allows identification of trends probably attributable in part to anthropogenic climate change for near-surface temperature and precipitation at the level of 5°(temperature) or 2.5°(precipitation) grid cells Knutson et al. (2013); Knutson and Zeng (2018). In this article, we apply this methodology to analyse trends from 1951 to updated observational data until 2018 for temperature (Figure 4.11a) and until 2016 for precipitation (Figure 4.11b). Grid cells in categories +-2 or +-3 show where trends cannot be explained by internal variability and are either consistent with or greater than the expected change in climate model simulations that include anthropogenic forcing agents. We infer that these cells display detectable and at least partly attributable trends.

We next resolve the structured geographic information extracted from our studies, which ranges from continental scale to individual watersheds or communities, to sets of grid cells (Figure 4.9 and Methods). We can then derive the weighted number of studies per grid cell according to the number of grid cells to which each study relates. By combining studies related to temperature or precipitation with the gridded information on attributable trends in temperature and precipitation, this provides a necessary (though not necessarily sufficient) condition for a systematic two-step attribution to anthropogenic activities of the impacts predicted by the classifier (Hegerl et al., 2010). Where studies documenting impacts associated with changes in temperature or precipitation co-occur with attributable trends in those variables, we claim that there is at least preliminary evidence for attributable impacts in these areas. This approach is similar in nature to the 'joint attribution' applied in IPCC AR4 (Rosenzweig et al., 2007, 2008).

In general, we note that this type of automated assessment procedure is no substitute for careful assessment by experts but can identify large numbers of studies for a

region that may point towards attributable human influence on impacts. Confidence in multi-step attribution claims depends on confidence in the attribution of the individual components (steps) along with the confidence or limitation in linking the different steps in the proposed causal chain (Rosenzweig et al., 2008). One limitation of the partially automated two-step attribution approach is that we cannot verify that every temperature or precipitation trend cited in impact studies matches, in sign, magnitude or period, those attributed to human influence by the regional detection and attribution studies for temperature (Knutson et al., 2013) and precipitation (Knutson and Zeng, 2018). This is a greater problem for studies driven by precipitation, where both wetting and drying trends occur with greater temporal variation, although these make up the minority of partially attributed studies and grid cells. We also note that not all studies in our database document impacts in response to trends in climate variables. Where impacts are attributed to extreme events or variation in temperature or precipitation, the fact that recent trends in temperature or precipitation can be attributed to human influence provides important context but does not allow robust attribution of those impacts. These factors limit confidence in our cases of potential attribution of impacts to anthropogenic forcing. Our approach could be extended with more fine-grained analysis of studies or with attribution of additional signals in climate variables to make more robust attribution statements.

For 80% of global land area (excluding Antarctica), trends in temperature and/or precipitation can be attributed at least in part to human influence on the climate (purple cells, Figure 4.11c). Using gridded population density data (Center for International Earth Science Information Network - CIESIN - Columbia University, 2018), we calculate that this covers 85% of the world's population. The majority of land grid cells show attributable warming trends, with exceptions where trends cannot be robustly distinguished from internal variability (white cells, category 0) or where there is insufficient data to establish trends (grey cells). For precipitation, attributable wetting and drying trends are found with greater geographical variation. There are also more grid cells where a trend in precipitation cannot be established, or where the observed trend is opposite in sign to that simulated by climate model historical simulations (green and yellow cells, ±4).

Although most of the world's population resides in areas where trends in temperature and or precipitation can be at least partially attributed to human influence, there

is substantial geographical variation in the degree to which the impacts of temperature and precipitation on human and natural systems have been studied. We characterize areas with fewer than 5 weighted studies per grid cell as displaying low levels of evidence, areas with 5-20 weighted studies as robust levels of evidence and areas with more than 20 weighted studies as high levels of evidence.

For 48% of global land area (hosting 74% of global population), we find robust or high levels of evidence of impacts on human and natural systems colocated with attributable temperature or precipitation trends (Figure 4.11c). Areas with this combination of evidence are indicated by the darker purple cells. These constitute almost all grid cells in western Europe, North America, and South and East Asia, and there are parts of all continents that have similar pockets of substantial preliminary evidence.

However, for 33% of global land area (hosting 11% of global population), although there is evidence that long-term trends in precipitation and temperature are attributable at least in part to human influence, there is relatively little evidence in the existing literature about how these trends impact human and natural systems (Figure 4.11c lightest purple shading). This imbalance suggests, in line with research measuring climate impacts using remote sensing (Frank et al., 2015), that the lack of evidence in individual studies is because these locations are less intensively studied, rather than because there is an absence of impacts in these areas. Parts of western Africa and southeastern, western and northern Asia contain several light purple grid cells where there is evidence to suggest that the climate (temperature and/or precipitation) has changed because of human influence, but there is little evidence on how this may be impacting human and natural systems. These demonstrable evidence gaps suggest a lack of impacts research commensurate with current knowledge of how the local climate (temperature and/or precipitation) is changing.

Some of the spatial features can be explained by the geographical characteristics. Among the regions with limited evidence are vast, sparsely populated and difficult-to-reach areas with a comparable uniform biosphere and climate such as Siberia or the Saharan desert. But beyond these features, our results clearly reveal a substantial 'attribution gap'. We find that 23% of the population of low-income countries live in areas with low impact evidence despite at least partially attributable trends in temperature and/or precipitation (Fig. 2d). In high-income countries, this figure is only 3%. A density of 5 or more studies per grid cell with attributable impacts is 1.76 times as

prevalent by population for high-income countries (88%) as for low-income countries (50%), while a density of 20 or more studies with attributable impacts is more than 4 times as prevalent (81% compared with 17%).

In the remaining grey grid cells (Figure 4.11c), trends in precipitation and temperature have not been attributed to human influence on the climate according to the methodology in refs. (Knutson et al., 2013; Knutson and Zeng, 2018), as applied to CMIP6 models. This does not rule out the possibility that some trends in precipitation or temperature have occurred in these regions that have been driven, at least in part, by human influence on the climate. However, due to various factors, such as lack of adequate observational data, high levels of natural variability compared with the climate change signal or limitations in modelling or estimated climate forcings, some observed changes that include anthropogenic contributions may not yet be attributable at the grid-cell level. This categorization of individual grid points may well change as new observational data are collected, as models improve, as the global climate continues to warm or as detection/attribution methodologies improve. Darker grey grid cells (10% of analysed land area) indicate where there are no detectable trends in temperature or precipitation that can be attributed to human influence at a grid-cell level but where there nevertheless appears to be substantial evidence that local trends in some climate variables lead to impacts on human and natural systems. For example, many studies refer to the impacts of temperature in the state of Western Australia, but of the 40 grid cells in the state, an attributable temperature trend can be demonstrated for 22 cells. For 16 of the remaining cells, a lack of data means that a detectable trend cannot be established, and for the remaining 2 cells, no attributable trend can be established.

The lightest grey cells (17% of land area) describe areas where we do not detect anthropogenic influence on regional temperature or precipitation and find few publications about the impacts of temperature or precipitation on human and natural systems. Apart from high latitudes and over the ocean, these cells are primarily in Africa. For example, in the light grey patch over the central part of sub-Saharan Africa, limitations of observed data, models or low signal-to-noise imply that we are unable to attribute temperature or precipitation trends to human influence on the climate using the methodologies employed here (Figure 4.6); further, we have identified few studies analysing the impacts of climate change on human and natural systems in those regions. These evidence gaps constitute substantial blind spots in understanding of climate im-

pacts and, in some cases, understanding of attributable anthropogenic influence on regional precipitation and/or temperature.

In total, 57,366 studies discuss impacts related to a driver that our analysis suggests can be attributed in part to human influence on the climate in at least one grid cell to which the study refers. We find hundreds of partially or mostly attributable studies (where there are attributable trends in the relevant climate variable for at least 1% or more than 50% of grid cells, respectively) in each impact category across all continents (Figure 4.12, indicated by the darker green and purple bars). This figure ranges from 268 (143–514) studies of impacts on mountains, snow and ice in Africa to 7,835 (4,308–13,552) studies of impacts on terrestrial ecosystems in North America. Wide confidence intervals here reflect the compound uncertainty deriving from classification of relevance, impact and driver.

Our analysis also allows quantification of how the share of research on each impact category varies from continent to continent. For example, research on human and managed systems makes up 12% of all research globally, but only 10% of research in Europe, compared with 19% in Africa. This focus on human and managed systems in Africa is remarkable given that the absolute numbers of studies in Africa (1,466) is similar to that in Europe (1,799) despite the vast difference in total numbers of studies between the two continents. This greater share of research in Africa documenting impacts in human and managed systems may reflect the high vulnerability of particularly sub-Saharan Africa to climate impacts (Schleussner et al., 2018).

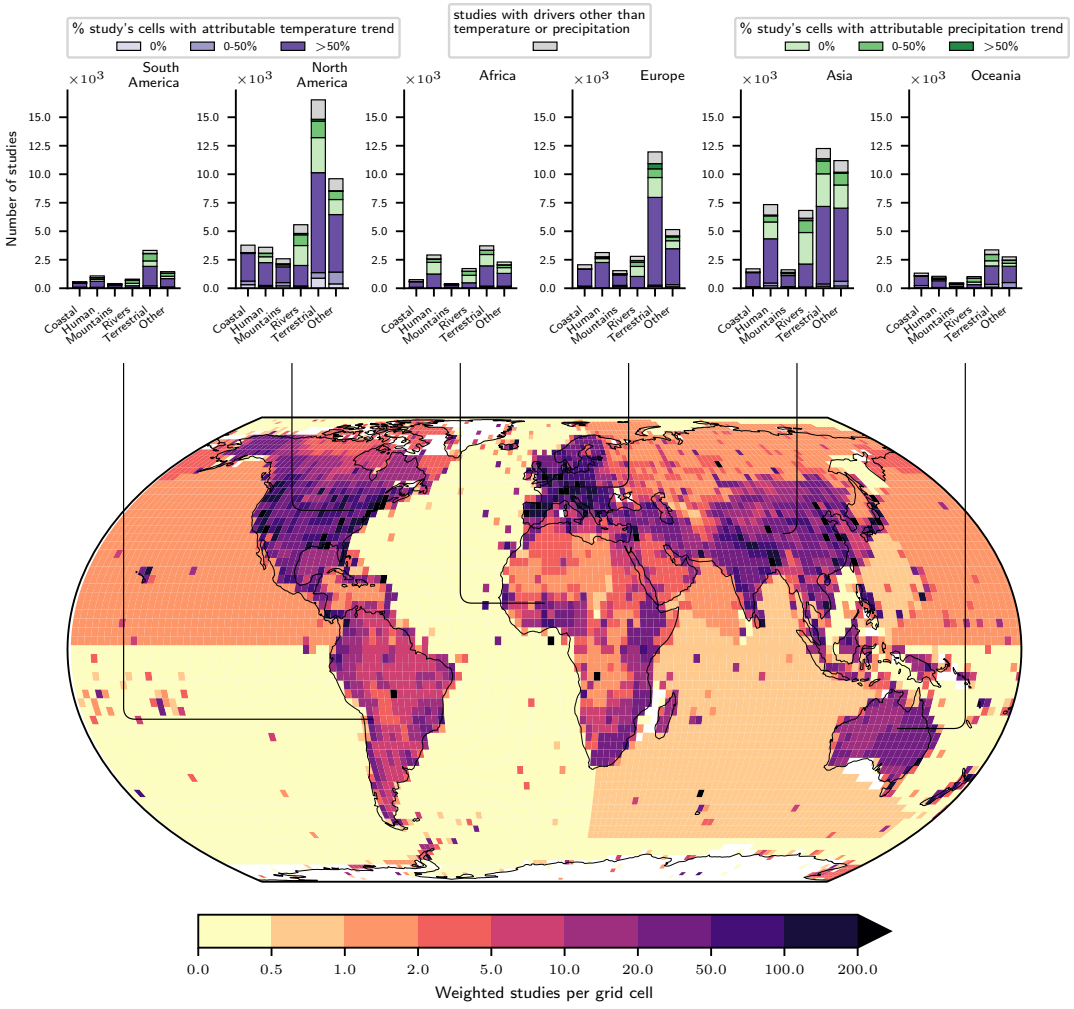


Figure 4.12: A global density map of climate impact evidence. Map colouring denotes the number of weighted studies per grid cell for all evidence on climate impacts (N = 77,785). Bar charts show the number of studies per continent and impact category. Bars are coloured by the climate variable predicted to drive impacts. Colour intensity indicates the percentage of cells a study refers to where a trend in the climate variable can be attributed (partially attributable: $> 0\%$ of grid cells, mostly attributable: $> 50\%$ of grid cells).

4.5 Discussion and Conclusion

We develop a two-step attribution process that combines a transparent and reproducible (Peng, 2011; Müller-Hansen et al., 2020) machine-learning approach to identifying studies on observed climate impacts with model-based assessments of detectable anthropogenic contributions to historical temperature and precipitation trends. Using machine learning to scale up evidence synthesis allows us to map 100,000 studies of climate impacts, providing a comprehensive picture of the evidence base. Bringing together these two lines of evidence on climate change and climate impacts provides a new bridge between the climate science community and the impacts, adaptation, and vulnerabilities communities, and highlights the synergistic nature of their approaches.

Our spatially resolved approach allows for a systematic provision of regional to local, sector-specific climate impact information to local or regional experts and adaptation practitioners. This offers perspectives for a new climate service supporting the uptake of scientific information in local contexts and providing relevant information for adaptation action. Second, the quantification of an 'attribution gap' highlights the need for more research on climate impacts in low-income countries. Furthermore, the automated nature of the assessment allows for continuous updating of the database, creating a 'living' evidence map that can also be improved and extended by incorporating additional sources of relevant publications (for example, non-English-speaking evidence or improved/expanded regional detection/attribution studies) and targeted assisted learning in regional or topical areas of interest.

The compiled database is vast but neither complete nor perfect. Our systematic query-based literature search is extensive but will also exclude some relevant studies. The selection and categorization of studies was achieved using machine learning, meaning that results are subject to additional uncertainties, which compound for each level of classification. Further, documents were coded only at the abstract level, and only the abstracts were used as inputs to our classifiers. Given the relative simplicity of the types of information we extract (focusing on the impact area studied and the documented driver), we expect them to be covered in the abstract, which provides the condensed summary of the study's findings. Applying classifiers to noisy full texts that contain contextual information and related research as well as the results and topic of a study would greatly increase the risk of false positives. We thus find our approach well justified for such high-level syntheses.

The database we assemble will also incorrectly exclude some relevant documents and contain some documents that have been incorrectly included or incorrectly coded, but the approach enables us to report both classifier performance and associated uncertainties. In addition, some included studies may be of low quality as no process for critical appraisal (a key component of formal systematic reviews) was followed either by human reviewers or in the machine-learning pipeline. In the case of systems subject to other anthropogenic interference such as the global biosphere, managed systems such as agriculture or human systems themselves, identifying a robust climate change driver requires careful assessment of other socioeconomic factors (Shepherd, 2019; Rosenzweig and Neofotis, 2013), adding additional levels of complexity (Mengel et al., 2020).

The two-step attribution process is also applied only for the subset of papers that provide evidence on impacts driven by temperature and precipitation. Exploring the role of human influence for studies analysing the effects of factors other than trends in mean temperature or precipitation as the main driver would require additional attribution strategies, but these could, in principle, be combined with individual studies in similar ways. There is a growing literature on attributable human influence on a number of climate metrics at the regional scale as well as extreme events (Gudmundsson et al., 2021; Diffenbaugh, 2020; Herring et al., 2021) and, therefore, much scope for expansion of this approach. Finally, we note that plausible causal chains of cascading impacts are not covered by our attribution approach (such as temperature driving an increase in drought, leading to reduced agricultural yields) except where studies address each part of the causal chain.

These caveats highlight that the type of machine-learning-assisted evidence map we present here is no substitute for careful assessment by experts, either in the context of a gold-standard systematic review (Higgins et al., 2019) or in IPCC assessments. However, in an age of 'big literature' (Nunez-Mir et al., 2015; Callaghan et al., 2020b), it is an invaluable complement. The use of machine learning means we consider more evidence than would otherwise be feasible, showing where evidence appears to be more prevalent and where important gaps can be observed. While traditional assessments can offer relatively precise but incomplete pictures of the evidence, our machine-learning-assisted approach generates an expansive preliminary but quantifiably uncertain map. Further, it enables us to provide an automated, living systematic map of climate impacts that can be readily updated. Ultimately, we hope that our global, living, automated and

multi-scale database will help to jump start a host of reviews of climate impacts on particular topics or particular geographic regions.

Machine-learning pipelines as developed here will be useful to prepare the IPCC for the age of big literature by scaling systematic evidence mapping approaches. However, our results also show how synthesis and transparency can be lifted to new levels by combining hitherto disparate lines of evidence and reporting classifier performance as well as associated uncertainties. If science advances by standing on the shoulders of giants, in times of ever-expanding scientific literature, giants' shoulders become harder to reach. Our computer-assisted evidence mapping approach can offer a leg up.

Bibliography

Barry, A. A., Caesar, J., Tank, A. M. G. K., Aguilar, E., McSweeney, C., Cyrille, A. M., Nikiema, M. P., Narcisse, K. B., Sima, F., Stafford, G., Touray, L. M., Ayilari-Naa, J. A., Mendes, C. L., Tounkara, M., Gar-Glahn, E. V. S., Coulibaly, M. S., Dieh, M. F., Mouhaimouni, M., Oyegade, J. A., Sambou, E., and Laogbessi, E. T. (2018). West Africa climate extremes and climate change indices. *International Journal of Climatology*, 38(S1):e921–e938. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.5420.

Bedsworth, L. W. and Hanak, E. (2010). Adaptation to Climate Change. *Journal of the American Planning Association*, 76(4):477–495. Publisher: Routledge _eprint: https://doi.org/10.1080/01944363.2010.502047.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Bergstra, J. and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(10):281–305.

Beusch, L., Gudmundsson, L., and Seneviratne, S. I. (2020). Crossbreeding CMIP6 Earth System Models With an Emulator for Regionally Optimized Land Temper-

ature Projections. *Geophysical Research Letters*, 47(15):e2019GL086812. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019GL086812.

Bornmann, L. and Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222. _eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23329.

Callaghan, M., Müller-Hansen, F., Hilaire, J., and Lee, Y. T. (2020a). NACSOS: NLP Assisted Classification, Synthesis and Online Screening.

Callaghan, M. W., Minx, J. C., and Forster, P. M. (2020b). A topography of climate change research. *Nature Climate Change*, 10(2):118–123. Number: 2 Publisher: Nature Publishing Group.

Center for International Earth Science Information Network - CIESIN - Columbia University (2018). Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11. Place: Palisades, NY.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

Cohen, A. M. (2006). An Effective General Purpose Approach for Automated Biomedical Document Classification. *AMIA Annual Symposium Proceedings*, 2006:161–165.

Cohen, A. M., Hersh, W. R., Peterson, K., and Yen, P. Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219.

Conway, D., Nicholls, R. J., Brown, S., Tebboth, M. G. L., Adger, W. N., Ahmad, B., Biemans, H., Crick, F., Lutz, A. F., De Campos, R. S., Said, M., Singh, C., Zaroug, M. A. H., Ludi, E., New, M., and Wester, P. (2019). The need for bottom-up assessments of climate risks and adaptation in climate-sensitive regions. *Nature Climate Change*, 9(7):503–511.

Cramer, W., Yohe, G. W., Auffhammer, M., Huggel, C., Molau, U., Dias, M. A. F. S., Solow, A., Stone, D. A., and Tibig, L. (2014). Detection and attribution of observed impacts. In Field, C. B., Barros, V. R., Dokken, D. J., Mach, K. J., Mastrandrea,

M. D., Bilir, T. E., Chatterjee, M., Ebi, K. L., Estrada, Y. O., Genova, R. C., Girma, B., Kissel, E. S., Levy, A. N., MacCracken, S., Mastrandrea, P. R., and White, L. L., editors, *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel of Climate Change*, pages 979–1037. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. Section: 18.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019a). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 4171–4186.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. arXiv: 1810.04805.

Diffenbaugh, N. S. (2020). Verification of extreme event attribution: Using out-of-sample observations to assess changes in probabilities of unprecedented events. *Science Advances*, 6(12):eaay2368. Publisher: American Association for the Advancement of Science Section: Research Article.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958. Publisher: Copernicus GmbH.

Fankhauser, S. (2017). Adaptation to Climate Change. *Annual Review of Resource Economics*, 9(1):209–230. _eprint: https://doi.org/10.1146/annurev-resource-100516-033554.

Frank, D., Reichstein, M., Bahn, M., Thonicke, K., Frank, D., Mahecha, M. D., Smith, P., Velde, M. v. d., Vicca, S., Babst, F., Beer, C., Buchmann, N., Canadell, J. G., Ciais, P., Cramer, W., Ibrom, A., Miglietta, F., Poulter, B., Rammig, A., Seneviratne, S. I., Walz, A., Wattenbach, M., Zavala, M. A., and Zscheischler, J.

(2015). Effects of climate extremes on the terrestrial carbon cycle: concepts, processes and potential future impacts. *Global Change Biology*, 21(8):2861–2880. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.12916.

Gudmundsson, L., Boulange, J., Do, H. X., Gosling, S. N., Grillakis, M. G., Koutroulis, A. G., Leonard, M., Liu, J., Schmied, H. M., Papadimitriou, L., Pokhrel, Y., Seneviratne, S. I., Satoh, Y., Thiery, W., Westra, S., Zhang, X., and Zhao, F. (2021). Globally observed trends in mean and extreme river flow attributed to climate change. *Science*, 371(6534):1159–1162. Publisher: American Association for the Advancement of Science Section: Report.

Gudmundsson, L., Leonard, M., Do, H. X., Westra, S., and Seneviratne, S. I. (2019). Observed Trends in Global Indicators of Mean and Extreme Streamflow. *Geophysical Research Letters*, 46(2):756–766. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018GL079725.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv:2004.10964 [cs]*. arXiv: 2004.10964.

Haddaway, N. R. and Pullin, A. S. (2014). The Policy Role of Systematic Reviews: Past, Present and Future. *Springer Science Reviews*, 2(1):179–183.

Hallegatte, S. and Mach, K. J. (2016). Make climate-change assessments more relevant. *Nature News*, 534(7609):613. Section: Comment.

Halterman, A. (2017). Mordecai: Full Text Geoparsing and Event Geocoding. *Journal of Open Source Software*, 2(9):91.

Hansen, G. (2015). The evolution of the evidence base for observed impacts of climate change.

Hansen, G. and Stone, D. (2016). Assessing the observed impact of anthropogenic climate change. *Nature Climate Change*, 6(5):532–537. Number: 5 Publisher: Nature Publishing Group.

Haunschild, R., Bornmann, L., and Marx, W. (2016). Climate Change Research in View of Bibliometrics. *PLoS ONE*, 11(7):1–19.

Hegerl, G. C., Hoegh-Guldberg, O., Casassa, G., Hoerling, M., Kovats, S., Parmesan, C., Pierce, D., and Stott, P. (2010). Good Practice Guidance Paper on Detection and Attribution Related to Anthropogenic Climate Change. In Stocker, T., Field, C. B., Qin, D., Barros, V., Plattner, G.-K., Tignor, M., Midgley, P., and Ebi, K. L., editors, *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Detection and Attribution of Anthropogenic Climate Change*. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland.

Herring, S. C., Christidis, N., Hoell, A., Hoerling, M. P., and Stott, P. A. (2021). Explaining Extreme Events of 2019 from a Climate Perspective. *Bulletin of the American Meteorological Society*, 102(1):S1–S116. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A., editors (2019). *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Chichester, 2nd edition edition.

IPCC (2012). *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. Cambridge Univ ersity Press, Cambridge.

IPCC (2014). *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Knutson, T. R. and Zeng, F. (2018). Model Assessment of Observed Precipitation Trends over Land Regions: Detectable Human Influences and Possible Low Bias in Model Trends. *Journal of Climate*, 31(12):4617–4637. Publisher: American Meteorological Society.

Knutson, T. R., Zeng, F., and Wittenberg, A. T. (2013). Multimodel Assessment of Regional Surface Temperature Trends: CMIP3 and CMIP5 Twentieth-Century Simulations. *Journal of Climate*, 26(22):8709–8743. Publisher: American Meteorological Society.

Lamb, W. F., Creutzig, F., Callaghan, M. W., and Minx, J. C. (2019). Learning about urban climate solutions from case studies. *Nature Climate Change*, 9(4):279–287.

Lane, J. E., Kruuk, L. E. B., Charmantier, A., Murie, J. O., and Dobson, F. S. (2012). Delayed phenology and reduced fitness associated with climate change in a wild hibernator. *Nature*, 489(7417):554–557. Number: 7417 Publisher: Nature Publishing Group.

Marshall, I. J., Kuiper, J., Banner, E., and Wallace, B. C. (2017). Automating Biomedical Evidence Synthesis: RobotReviewer.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Mengel, M., Treu, S., Lange, S., and Frieler, K. (2020). ATTRICI 1.0 - counterfactual climate for impact attribution. *Geoscientific Model Development Discussions*, pages 1–26. Publisher: Copernicus GmbH.

Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Atmospheres*.

Müller-Hansen, F., Callaghan, M. W., and Minx, J. C. (2020). Text as big data: Develop codes of practice for rigorous computational text analysis in energy social science. *Energy Research & Social Science*, 70:101691.

Nerem, R. S., Beckley, B. D., Fasullo, J. T., Hamlington, B. D., Masters, D., and Mitchum, G. T. (2018). Climate-change-driven accelerated sea-level rise detected in the altimeter era. *Proceedings of the National Academy of Sciences of the United States of America*, 115(9):2022–2025.

Nunez-Mir, G. C., Iannone, B. V., Curtis, K., and Fei, S. (2015). Evaluating the evolution of forest restoration research in a changing world: a "big literature" review. *New Forests*, 46(5):669–682.

Padrón, R. S., Gudmundsson, L., Decharme, B., Ducharne, A., Lawrence, D. M., Mao, J., Peano, D., Krinner, G., Kim, H., and Seneviratne, S. I. (2020). Observed changes in dry-season water availability attributed to human-induced climate change. *Nature Geoscience*, 13(7):477–481. Number: 7 Publisher: Nature Publishing Group.

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, 334(6060):1226–1227. Publisher: American Association for the Advancement of Science Section: Perspective.

Porciello, J., Ivanina, M., Islam, M., Einarson, S., and Hirsh, H. (2020). Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning. *Nature Machine Intelligence*, 2(10):559–565. Number: 10 Publisher: Nature Publishing Group.

Rosenzweig, C., Casassa, G., Karoly, D. J., Imeson, A., Liu, C., Menzel, A., Rawlins, S., Root, T., Seguin, B., and Tryjanowski, P. (2007). Assessment of observed changes and responses in natural and managed systems. In *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 79–131. Cambridge University Press, Cambridge, United Kingdom.

Rosenzweig, C., Karoly, D., Vicarelli, M., Neofotis, P., Wu, Q., Casassa, G., Menzel, A., Root, T. L., Estrella, N., Seguin, B., Tryjanowski, P., Liu, C., Rawlins, S., and Imeson, A. (2008). Attributing physical and biological impacts to anthropogenic climate change. *Nature*, 453(7193):353–357. Number: 7193 Publisher: Nature Publishing Group.

Rosenzweig, C. and Neofotis, P. (2013). Detection and attribution of anthropogenic climate change impacts. *WIREs Climate Change*, 4(2):121–150. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcc.209.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*. arXiv: 1910.01108.

Schleussner, C.-F., Deryng, D., D'haen, S., Hare, W., Lissner, T., Ly, M., Nauels, A., Noblet, M., Pfleiderer, P., Pringle, P., Rokitzki, M., Saeed, F., Schaeffer, M., Serdeczny, O., and Thomas, A. (2018). 1.5°C Hotspots: Climate Hazards, Vulnerabilities, and Impacts. *Annual Review of Environment and Resources*, 43(1):135–163. _eprint: https://doi.org/10.1146/annurev-environ-102017-025835.

Schleussner, C.-F. and Fyson, C. L. (2020). Scenarios science needed in UNFCCC periodic review | Nature Climate Change. *Nature Climate Change*, 10(272(2020)).

Shepherd, T. G. (2019). Storyline approach to the construction of regional climate change information. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2225):20190013. Publisher: Royal Society.

Westgate, M. J., Haddaway, N. R., Cheng, S. H., McIntosh, E. J., Marshall, C., and Lindenmayer, D. B. (2018). Software support for environmental evidence synthesis. *Nature Ecology & Evolution*, 2(4):588–590. Number: 4 Publisher: Nature Publishing Group.

Zhang, Y. Q., Yu, C. H., and Bao, J. Z. (2016). Acute effect of daily mean temperature on ischemic heart disease mortality: a multivariable meta-analysis from 12 counties across Hubei Province, China. *Zhonghua Yu Fang Yi Xue Za Zhi [Chinese Journal of Preventive Medicine]*, 50(11):990–995.

Chapter 5

Discussion and Conclusion

This thesis demonstrates three main results which show clear methodological advances and contributions to our understanding of climate science and policy. These are discussed below and put into context. The following section sketches out a conceptual framework for machine-learning-assisted evidence synthesis, before critically reflecting on the approach used in this thesis, and outlining directions for future work.

5.1 Summary of results

1. Machine-learning-assisted screening in systematic reviews can save work *and* achieve reliable levels of recall. This is the first time this has been demonstrated. Chapter 2 also shows that methods suggested in the existing literature do not achieve this, and can have serious negative consequences.

2. Contrary to what we thought we knew about the IPCC, the social sciences do not seem to be under-represented in IPCC reports. In fact, technical disciplines in engineering and agricultural sciences are under-represented. Chapter 3 shows how technical, solutions-oriented topics are both under-represented in IPCC reports and not well covered by social science literature.

3. Chapter 4 shows that climate impacts with a plausible attribution to anthropogenic causes can be identified in locations covering 80% of the world's land area where 85% of the world's population reside. These impacts are demonstrated in all continents and in all sectors.

5.1.1 Saving work in systematic reviews using machine learning

The first result identifies a clear research gap within a novel but developed field. Automated study identification is a well-defined task, which can be easily evaluated using datasets from previous systematic reviews. Several papers have demonstrated ways in which different machine learning approaches can perform this task, and have been able to show marginal improvements in potential work savings (Cohen et al., 2006; Miwa et al., 2014; Bannach-Brown et al., 2019; Przybyła et al., 2018). A systematic review has even been conducted on how automated study identification can help systematic reviews (O'Mara-Eves et al., 2015).

However, all the work savings demonstrated in these studies are predicated on the *a priori* knowledge of how many studies are relevant. Figure 5.1 shows different ways to draw 100 relevant documents from a sample of 1000 documents. If documents are chosen at random, the number of relevant documents identified will roughly follow the diagonal line from the bottom left hand corner to the top right hand corner (Figure 5.1 **a**). This means that if you want to identify 95% of relevant documents, you have to screen around 95% of all documents. However, if machine learning can increase the likelihood that relevant documents are drawn (orange line), 95% of relevant documents (dotted grey line) will be identified long before 95% of all documents have been seen. The proportion of documents not yet seen by the time 95% of relevant documents have been seen is traditionally counted as work saved.

Banking these savings is dependent on deciding to stop at exactly the right time. Panel **b** and **c** show that the researcher does not have enough information to decide when to stop. We do not know how many relevant documents there are, so we do not know where the 95% line will be. We may guess based on the shape of the curve, but even with this information, researchers may be tempted to stop too early. Figure 5.1 b shows a case where the curve has appeared to plateau, but panel c shows that this turns out to be too early. This means that automated screening has until now remained a promising experimental approach, using which some *potential* work savings can be

demonstrated. The stopping criteria developed in chapter 2 solve this problem by indicating when to stop if a target recall level is to be achieved with a given confidence level. This transforms automated study identification into an approach that is ready for live reviews. Further discussion now needs to be had about how this can be incorporated into systematic review guidelines.

In the social sciences - particularly in the field of climate solutions - the lack of a culture of research synthesis slows down learning (Minx et al., 2017). Technology-assisted reviews may offer a route to systematising knowledge synthesis where resourcing requirements for traditional systematic reviews prove too large. Incentive structures in the social sciences do not often reward time-consuming and rigorous systematic reviews. By lowering the human input required for screening, machine-learning assisted evidence synthesis can provide for a more systematic and transparent process for study identification than would have been possible otherwise. Alternatively, the use of machine-learning in evidence synthesis can make projects possible that would have been too ambitious before, because the literature to be surveyed was too vast. Indeed, the application of such methods has begun to bear fruit (Ivanova et al., 2020; Khanna et al., 2021; Berrang-Ford et al., 2021). To further enable this cultural change, future research should work towards providing estimates of potential time requirements for machine-learning assisted reviews (Haddaway and Westgate, 2019), on developing more accurate methodologies for estimating safe stopping points, and on continuing to make the process of machine-learning-assisted screening interpretable to non-specialists.

Any potential time savings are dependent on various known, estimable, and unknown parameters of the dataset and approach. Theoretically we know how these will behave. The total number of documents, the performance of the machine learning algorithm, and the proportion of documents which are relevant, will all be positively correlated with the potential work savings to be gained in technology assisted systematic review. This is confirmed by the results of chapter 2, where higher potential work savings (between 40% and 60% of the total dataset size are identified for datasets with more than 2,000 documents). However, no dataset had more than 10,000 documents and we can expect greater work savings for very large systematic reviews that may have tens of thousands of potentially relevant documents. Technology-assisted systematic review is therefore particularly suited to such large-scale projects where larger proportional work savings also translate to large absolute reductions in human effort. Further

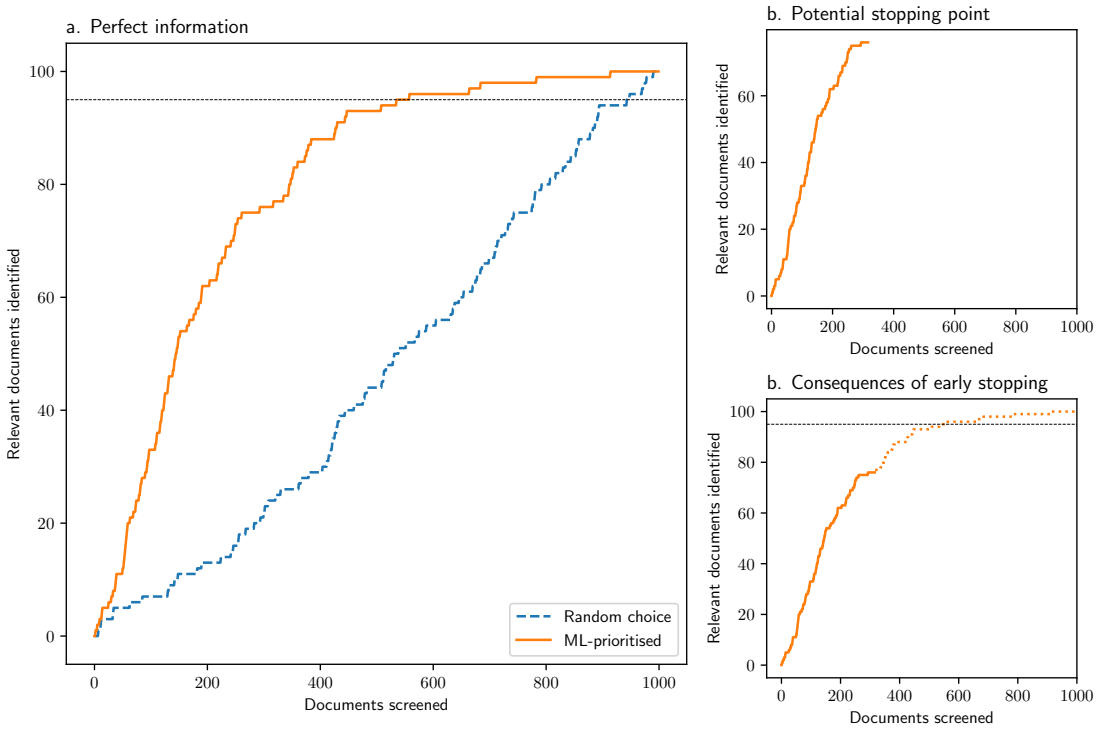


Figure 5.1: Automated study identification for systematic reviews: 3 views of the number of relevant documents included and the number of documents seen in a toy dataset with 1000 documents of which 100 are relevant. Orange lines show values for a hypothetical ML-prioritised ordering, where relevant documents are more likely to be identified first.

theoretical and empirical work is necessary to estimate, with uncertainty, potential work savings given a set of the parameters enumerated above.

5.1.2 What we think we know about the IPCC

The scale of the literature on climate change, means that the IPCC cannot cite all relevant publications. It must make choices about what to reference and what to leave out in order to ensure that assessments remain comprehensive and policy-relevant. Previous assessments of IPCC priorities have been based on comparing the relative proportions of citations from different disciplines (Bjurström and Polk, 2011), or the relative proportions of contributors from different disciplines (Corbera et al., 2016). That comparatively few citations or authors come from the social sciences, as opposed

to the natural sciences, has been taken to indicate bias within the IPCC towards the natural sciences. This has become a truism: "This powerful bias to the natural sciences in the construction of 'IPCC knowledge' about climate change has been remarked on for many years" (Hulme and Mahony, 2010)[1].

In Chapter 3, I argue that differing relative proportions of references or authors are insufficient grounds to argue that the IPCC displays bias against the social sciences. To make this claim, one needs to posit an unbiased distribution of references by discipline, against which the actual distribution can be compared. Figure 5.2 shows three disciplinary distributions of IPCC citations and compares these with the disciplinary distribution across all publications on climate change identified in the study. Panel **a.** shows the actual distribution, making clear that the proportion of social science studies that are cited by the IPCC is greater than the proportion of social science studies in the literature at large. In other words, the social sciences are over-represented in IPCC reports, while engineering and technology, and the agricultural sciences are under-represented. Were the proportions of IPCC references from each discipline to be equal (Figure 5.2b), this would much more severely over-represent some disciplines and under-represent others. To talk about bias, it is more appropriate to discuss the relationship of an actual distribution to a representative distribution (Figure 5.2).

Though this may be enough to debunk claims of IPCC bias, this is not to say that there is not a greater need for social science research in IPCC assessments (David G. Victor, 2015). While the science of climate change, its causes, and its potential impacts has been clear enough for world governments to make commitments to limit global warming to 2℃, and pursue efforts to limit warming to 1.5℃, global emissions have continued to rise. Arguably, solutions to this problem should be rooted in the social sciences, and the IPCC may be well advised to make a particular effort to foreground social science knowledge. We may want to rephrase claims of IPCC bias against the social sciences into calls for the IPCC to introduce a bias to social science research and cite this at a higher rate than other research on climate change. The proportion of IPCC references in each discipline should certainly not be perfectly proportional to the wider literature (Figure 5.2c), but demands for any distribution should be rooted in data about the landscape of publications. The are good reasons for citing one part of the literature more than others, but these decisions should be justified.

[1] Hulme and Mahony (2010) cite an early version of Bjurström and Polk (2011).

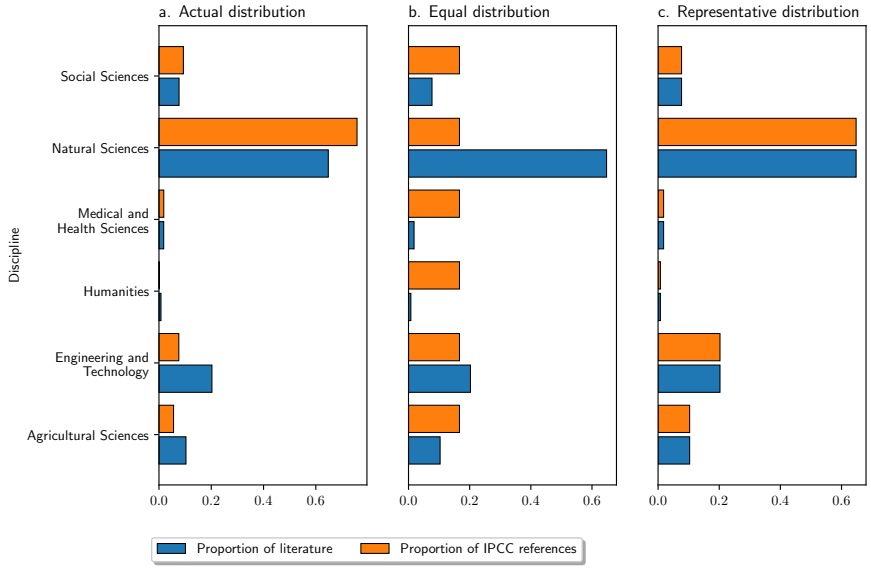


Figure 5.2: Three disciplinary distributions of publications on climate change (blue bars), and the subset of publications cited by the IPCC (orange bars)

Finally, the actual proportion of publications in each discipline can tell us that if we want to have more social science research in the IPCC, we may need to fund and produce more social science research on climate change (Overland and Sovacool, 2020). The innovation of chapter 3 is that analysis of the disciplinary distributions of literature is combined with information about the thematic content of papers. This tells us that topics on technical solutions, such as on negative emissions, e-vehicles, or buildings, are both less cited by the IPCC *and* less written about in social science journals. This sets out clear priorities for policy relevant research on climate solutions in the social sciences, and shows IPCC authors a relatively untapped potential source of knowledge of climate solutions outside the social sciences.

5.1.3 Synthesising local studies of climate impacts with global climate models

Where chapter 3 provides an exploratory and general map of the literature on climate change, chapter 4 gives a specific and directed synthesis. The thematic categories into which the evidence needs to be sorted are already predefined. In this way, the study offers a method to scale up existing synthesis efforts - like that conducted by the IPCC

in AR5 (Cramer et al., 2014). The map of climate impacts literature produced in AR5 was carried out without a systematic search strategy, instead using expert elicitation (Hansen and Stone, 2016). The use of machine learning in chapter 4 meant that more studies could be assessed, classifying documents into the same categories as the IPCC. This then meant it was possible to identify tens of thousands of documents relevant to understanding climate impacts.

The map of impacts in AR5 aimed to assess the role of anthropogenic climate change in driving the impacts in each paper considered. It relied on papers to establish both whether trends in climate variables were driving impacts, and whether these trends were in turn driven by greenhouse gas emissions. Chapter 4 developed an innovative approach to understanding the role of anthropogenic climate change in driving impacts, that leveraged complementary knowledge from observational evidence and climate models.

We extracted the locations discussed in each study, and learned whether a long-term trend was described as driving the impact, and if so, whether that trend was temperature, precipitation, or another variable. We resolved the extracted locations to the set of 2.5°x2.5° grid cells they overlap with. Finally, we merged the data with updated work using observations and climate models to attribute trends in precipitation and temperature at a grid cell level to anthropogenic climate change. In this way we can point to the proportion of grid cells showing attributable trends in impacts for each study. Similarly, for every grid cell, we can show whether it displays attributable trends in temperature or precipitation, and how many studies document impacts driven by these trends.

We find over 100,000 studies related to climate impacts and show that, for either temperature or precipitation, a trend in observational data can be attributed to human influence on the climate for grid cells covering 80% of land area (excluding Antarctica), and 85% of the world's population. For many of these grid cells we find high levels of observational evidence on the impacts of these attributable changes in temperature and precipitation, but this evidence not distributed evenly.

In high income countries, 90% of people live in an area where temperature or precipitation trends can be attributed to human influence on the climate, and for 89% of these people, there is a high level of evidence on how these trends impact human and natural systems, meaning at least 20 weighted studies per grid 2.5 degree grid cell

(where a study covering 2 grid cells contributes half a weighted study to each grid cell). A further 8% live in area with robust levels of evidence (5-20 weighted studies), and 3% live where there are low levels of evidence (fewer than 5 weighted studies).

In low income countries, on the other hand, 73% of people live in area where trends in temperature or precipitation can be attributed to human influence on the climate. That this number is lower is to a large part due to gaps in the observational record which means that trends cannot be robustly assessed. Beyond this, though, only 23% of those 73% of people live in area where there are high levels of evidence on the impacts of climate change in human and natural systems. 45% live in areas where there are robust levels of evidence, and 32% live in areas where there are low levels of evidence. In other words, those who live in an area affected by warming, wetting or drying trends caused by climate change are less likely to have the consequences of that trended documented in large amounts of scientific literature if they live in a low-income country compared to those who live in a high-income country. Thus there is a risk that climate impacts go undocumented in areas where vulnerability to climate change is highest.

Low levels of evidence of climate change have been observed before in low-income countries, for example in the IPCC's fifth assessment report (Cramer et al., 2014). But by bringing together the database of studies with observational and model data on temperature and precipitation trends we can show that the levels of evidence are lower even where temperature or precipitation trends can be observed. Thus we refer to this unbalance as an "attribution gap", arguing that low income countries are understudied in the available literature.

We also observe varying proportions of studies in each impacted system across continents, with large differences particularly prominent for human and managed systems. Such studies make up 12% of research globally, 10% of research in Europe, and 19% of research in Africa. The extent to which this reflects differences in research priorities or variance in actual impacts is unclear, but may reflect the high vulnerability of populations on the African continent, particularly in sub-Saharan Africa, to climate change.

5.2 Towards a typology of machine-learning-assisted evidence maps

This thesis investigates how machine learning affects the practice of evidence synthesis at and between the levels of systematic maps and global environmental assessments. Systematic maps offer a structured methodology for rigorously assessing the quantity and quality of evidence on a specific topic (Haddaway et al., 2016). On the other hand, global environmental assessments attempt to summarise vast fields of research and draw policy-relevant conclusions. Machine learning and natural language processing can be used to stretch what is possible in systematic maps in various ways (Haddaway et al., 2020). A typology of machine-learning-assisted evidence maps is presented below. In each case, distortions and uncertainty are unavoidable, meaning that none are strict systematic maps. However, each of these types of maps can provide relevant evidence for global environmental assessments, going beyond what can be produced in traditional systematic maps.

First, machine-learning-assisted evidence maps can be exploratory (Chapter 3) or directed (chapter 4), with each type of map serving a different purpose. Exploratory maps provide an overview of content, independently of what reviewers may intend to search for. This reduces subjectivity, meaning that topics can be discovered which were previously unknown. This is crucial where a corpus is so large that researchers can no longer hold an overview of all topics, such as in scientific literature on climate change. Few researchers' knowledge could encompass topics as broad as membranes for CO_2 capture, the effects of climate change on fisheries, CO_2 taxation, cloud radiative feedback, e-vehicles, and soil carbon. Yet exploratory evidence maps using topic modelling can uncover each of these topics and more without human input.

This type of map can describe to researchers or policymakers a broad range of topics in a literature, and show how they are related. Further investigation can reveal more about the distribution and dynamics of these topics which can be of practical relevance for environmental assessments. Where scientists are tasked with providing a comprehensive overview of a corpus, but where the relative size and importance of constituent parts may be contested, this type of map can inform the process of drawing up an outline with evidence about the state of the literature.

Directed evidence maps, on the other hand, offer the chance to more efficiently find

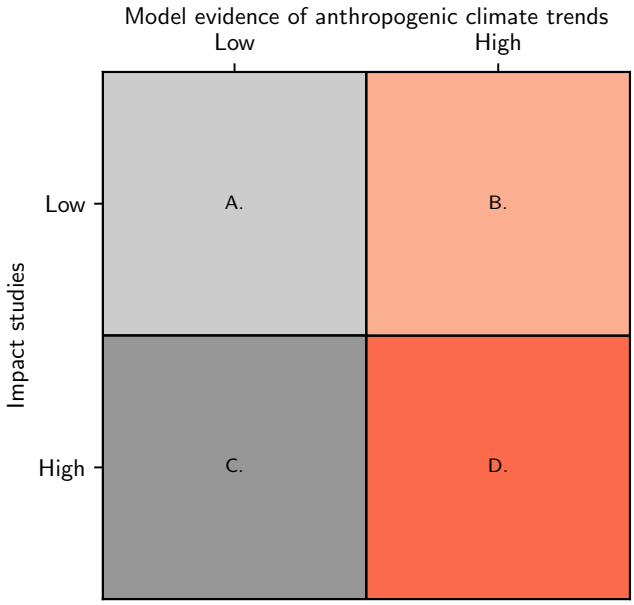


Figure 5.3: (Simplified) evidence gaps and gluts

and classify evidence where the topical categories are already known. These are helpful for characterising research within more specific fields or for specific research questions. In a smaller field, expert knowledge may be more effective at generating meaningful or policy-relevant typologies. Additionally, where an existing categorisation scheme exists, such as the IPCC impact categories referred to in chapter 4, directed evidence maps can use supervised learning to efficiently find new evidence in those categories.

Beyond the question of whether evidence maps are exploratory or directed, a distinguishing feature is the extent, and the way, in which other information is layered onto the thematic categories. This additional information is often used to describe where we have evidence gaps or gluts. For example, we may have more or less evidence about a particular theme in a particular region, and less in another. Describing this uneven distribution as composed of gaps or gluts is not trivial though, as we do not always know what the right distribution would look like.

The layered evidence map in chapter 4 displays the distribution of evidence in different categories across the globe, and combines this with observational and modelling evidence to characterise different types of evidence gaps and gluts. Figure 5.3 shows a simplified model of these. For example, box A describes where there is little or no

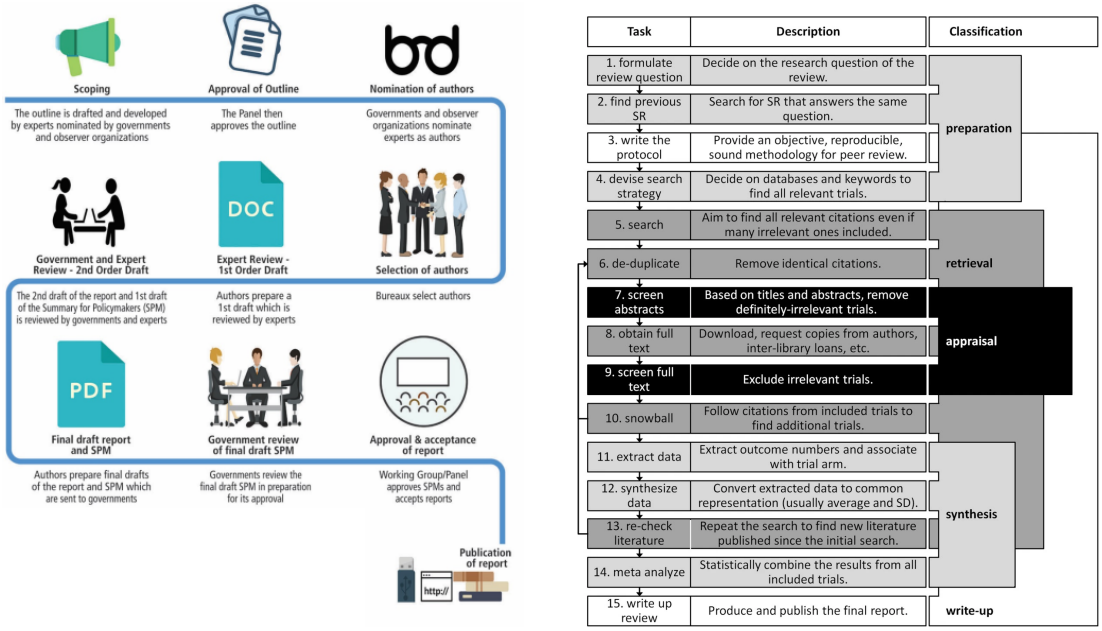
evidence from climate models and observations on whether climate trends are driven by human influence (x axis), and where there are few studies documenting the impacts of climate trends on human and natural systems (y axis). This constitutes a specific type of evidence gap: showing that overall we know little about climate impacts in areas of type A[1]. On the other hand, areas of type B indicate that, given the knowledge we have from climate models and observations on how human influence is driving trends, we know little about how these trends are impacting human and natural systems. The implications of this type of evidence gap are that the area may be understudied, considering the likely impacts of climate trends. In other words, given that we know that temperature or precipitation is changing, we have relatively few studies about the impacts of those changes.

Finally, inquisitive evidence maps can be used to explicitly test hypotheses about the factors which drive differences in the distribution of evidence. In Sietsma et al. (2021), for example, we show that papers on climate adaptation where the first author is from an Annex I country (a grouping of signatories to the UN Framework Convention of Climate Change representing the OECD and other richer countries) are significantly more likely to discuss topics on governance and conceptual issues. Inquisitive evidence maps can ask *why* certain gaps or gluts may occur.

5.3 Further opportunites for natural language processing in evidence synthesis and global environmental assessments

This thesis presents several ways in which natural language processing can aid evidence synthesis and global environmental assessments, but these remain a small subset of possible computer-assisted interventions. Figure 5.4 shows outlines of the processes required for each activity. Possibilities for additional entry points have been highlighted in Minx et al. (2021) and Tsafnat et al. (2014). As already discussed in this thesis, topic maps could be used to inform the IPCC's scoping process to help draft an outline which reflects the latest developments in climate-relevant science. Beyond this, a map of climate publications could provide input to a system that aids the IPCC in selecting

[1] reasons for low model evidence may be to do with lack of data, poor model in predicting trends or because trends were simply not observed



(a) IPCC process (Minx et al., 2021) (b) Evidence synthesis (Tsafnat et al., 2014)

Figure 5.4: Workflows for the IPCC (a) and evidence synthesis (b)

authors. This system could optimise for the combination of objectives of topic expertise, geographic representation, and disciplinary diversity which previous IPCC reports have been criticised for failing to meet (Corbera et al., 2016; Ford et al., 2012).

In evidence synthesis, chapter 2 addresses task 7, screening abstracts, of the 15 tasks described in Tsafnat et al. (2014). It builds on existing work to solve the problem of uncertain recall in automated screening. Beyond this work, computer-assisted interventions have been suggested to assist in many of the other tasks, and there is a wide open space for new approaches, and improvements to existing approaches, for automating these tasks.

In addition to what is documented in this thesis, many of the practical parts of this pipeline are incorporated into a software system, built alongside the thesis, to produce and reproduce the analysis in these chapters (Callaghan et al., 2020). The software includes a web-based interface to search for and import sets of documents through querying; to de-duplicate papers from different databases; to screen abstracts, with the option of using machine learning assistance; and to create and visualise topic models. It has enabled the replication of many of the techniques in this thesis in several other

projects, some of which are listed in chapter 6 of this thesis.

Beyond this, though, advancements in the field of NLP present further opportunities for computer-assisted evidence synthesis. A major advancement in NLP occurred in 2019 with the advent of BERT (Devlin et al., 2019). BERT is a language model trained on a large corpus of unlabelled text to represent the meanings of words and sentences in contexts. It can be fine-tuned to out-perform state of the art approaches to many existing NLP tasks. Progress since BERT has profound implications for the work in this thesis. First, new techniques offer the opportunity to perform the tasks described here more accurately. With more advanced NLP pipelines, relevant documents could be more quickly separated from irrelevant documents, and more accurately classified according to their content.

More accurate content classification also means that the classification of more ambitious or abstract content types could be feasible. For example, in chapter 4, a small dataset and relatively simple machine learning model are used to predict which impacts a paper discusses, and whether these are driven by trends or variation. A sophisticated deep learning pipeline using BERT, fine-tuned on a large dataset of annotated full texts, could potentially go further and predict each study's methodology, assess the quality of the study, predict whether trends were significant, and learn to parse complex chains of causal evidence. In evidence synthesis technology, work has started to automate the extraction and synthesis of the effects of interventions (Lehman et al., 2019).

Beyond classification, deep learning could be used to automatically summarise articles (Zhang et al., 2020), potentially offering opportunities in evidence synthesis or in assessments. Advances in AI offer the possibility of generating entirely computer-generated scientific books (Writer, 2019). While the prospect of an automatic IPCC report will rightly be regarded as far-fetched, there may be parts of the report producing process that can be improved or made more efficient with more automation. For example, the increasing capability of AI to answer questions (Yang et al., 2019; Wang et al., 2020) could be used as part of a pipeline to address the tens of thousands of review comments which IPCC authors have to answer. A potential system could be trained with human-generated answers to comments to suggest appropriate answers to similar comments, saving time and enhancing consistency.

5.4 The division of labour between humans and machines

The scale of scientific literature means that we are often unable to meet the task of synthesising and assessing the science of climate change, or indeed many other fields, without computer assistance. Computer assistance makes larger and more ambitious evidence synthesis projects possible, and means that we can draw data-driven conclusions about the content of hundreds of thousands of papers. However, it is important to recognise the trade-offs involved in computer assistance, and reflect accordingly on the appropriate division of labour between experts and machine intelligence.

Human experts will likely always outperform AI systems at the level of individual documents. It is therefore unlikely that systematic reviews or environmental assessments will ever be able to be fully automated. The use of computer assistance increases uncertainty about the internal validity of conclusions drawn from any set of scientific papers – or whether the conclusions accurately represent the evidence considered. However, by widening the net of literature that is considered, computer assistance can make the uncertainty around external validity – whether the conclusions drawn accurately represent all potentially relevant results – more transparent.

Communicating this uncertainty is an important message for the IPCC. In chapter 18 of working group II's contribution to the fifth assessment report, the observed impacts of climate change across the globe are mapped (Cramer et al., 2014). The map quantifies uncertainty about the extent to which these impacts can be attributed to anthropogenic climate change, and the extent to which there is agreement in the literature. However, it fails to address uncertainty over whether all literature is included, giving the misleading appearance that all relevant evidence is considered.

Although computer-assistance can help projects constrained by limited human resources, we should also acknowledge that computational resources are not unlimited. Large language models are computationally intensive, which has impacts in the real world. Training BERT from scratch consumes thousands of dollars worth of computational resources and the equivalent of more than a ton of CO2 in energy (Strubell et al., 2020). We should carefully weigh up the value of potentially marginal gains when considering ever larger models (Bender et al., 2021), and consider whether computer-assistance can always deliver better outcomes with constrained resources. The benefits of computer assistance will need to be assessed on a case by case basis.

Concerns about the use of machine learning in evidence synthesis and environmen-

tal assessments go beyond questions of accuracy and uncertainty. Even a perfectly accurate system can produce unwelcome outcomes. As discussed in chapter 3, accurate representations of data are not always desirable representations of data. This echoes David Hume's warning of the dangers of inferring what *ought* to be from what *is* (Hume, 2014). Machine learning applications are only as good as the data which train them, and there are several ways in which the data of scientific publications can be problematic for large evidence synthesis and assessment projects.

In fact, the data generation process is subject to various distortions. Science does not organically produce literature balanced according to what is relevant and what is useful. It is subject to imbalances of funding which reflect political priorities, the interests of industry, and the agendas of other donors (Overland and Sovacool, 2020). Scientific publishing privileges certain kinds of knowledge over others (Ford et al., 2016), and access to publishing in, and reading, scholarly journals is uneven, notably between richer and poorer institutions, and richer and poorer countries. Further, publication bias (Rothstein et al., 2005) and concerns about a reproducibility crisis (Baker, 2016) mean that we need to be careful about the conclusions drawn in efforts to synthesise the literature. Lastly, we should be aware that scientific literature is often collected in databases with uneven coverage (Mongeon and Paul-Hus, 2016) or limited access to text as data (Haddaway et al., 2020).

Machine learning pipelines fed with biased data will produce biased outcomes (Buolamwini and Gebru, 2018). The application of machine learning to evidence synthesis and environmental assessments needs to proceed with an awareness of potential biases. Without this caution, we are in danger of eliding these biases by presenting a misleadingly positivist representation of research. However, if we proceed cautiously, machine learning can also be used to generate new knowledge about existing biases. Future work is needed to address whether the use of machine learning may effect the outcomes of research synthesis. For example, if machine-learning assisted systematic review misses 5% of studies, it is an important question whether those 5% of relevant studies are representative of the total population (in which case estimates in a meta-analysis could be slightly under-powered), or are more likely to present certain types of evidence (for example null results, in which case estimates in a meta-analysis would be inaccurate). Further analysis could investigate whether the error rate of machine-learning predictions is greater for important but under-represented subgroups of papers, for example

science from the global south, or science on indigenous knowledge.

Scientists can not avoid using machine learning in synthetic research, as it is already embedded in the tools we commonly use to search for literature, be that search engines, bibliographic databases, or recommender systems. It is therefore the task of researchers to engage critically and reflectively on how these can be used to complement human expertise, and how they can be used transparently such that the implications of their use can be interrogated critically.

Bibliography

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452.

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S. C., Ananiadou, S., Liao, J., and Macleod, M. R. (2019). Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews*, 8(1):1–12.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Berrang-Ford, L., Siders, A. R., Lesnikowski, A., Fischer, A. P., Callaghan, M. W., Haddaway, N. R., Mach, K. J., Araos, M., Shah, M. A. R., Wannewitz, M., Doshi, D., Leiter, T., Matavel, C., Musah-Surugu, J. I., Wong-Parodi, G., Antwi-Agyei, P., Ajibade, I., Chauhan, N., Kakenmaster, W., Grady, C., Chalastani, V. I., Jagannathan, K., Galappaththi, E. K., Sitati, A., Scarpa, G., Totin, E., Davis, K., Hamilton, N. C., Kirchhoff, C. J., Kumar, P., Pentz, B., Simpson, N. P., Theokritoff, E., Deryng, D., Reckien, D., Zavaleta-Cortijo, C., Ulibarri, N., Segnon, A. C., Khavhagali, V., Shang, Y., Zvobgo, L., Zommers, Z., Xu, J., Williams, P. A., Canosa, I. V., van Maanen, N., van Bavel, B., van Aalst, M., Turek-Hankins, L. L., Trivedi, H., Trisos, C. H., Thomas, A., Thakur, S., Templeman, S., Stringer, L. C., Sotnik, G., Sjostrom, K. D., Singh, C., Siña, M. Z., Shukla, R., Sardans, J., Salubi, E. A., Safaee Chalkasra, L. S., Ruiz-Díaz, R., Richards, C., Pokharel, P., Petzold, J., Penuelas,

J., Pelaez Avila, J., Murillo, J. B. P., Ouni, S., Niemann, J., Nielsen, M., New, M., Nayna Schwerdtle, P., Nagle Alverio, G., Mullin, C. A., Mullenite, J., Mosurska, A., Morecroft, M. D., Minx, J. C., Maskell, G., Nunbogu, A. M., Magnan, A. K., Lwasa, S., Lukas-Sithole, M., Lissner, T., Lilford, O., Koller, S. F., Jurjonas, M., Joe, E. T., Huynh, L. T. M., Hill, A., Hernandez, R. R., Hegde, G., Hawxwell, T., Harper, S., Harden, A., Haasnoot, M., Gilmore, E. A., Gichuki, L., Gatt, A., Garschagen, M., Ford, J. D., Forbes, A., Farrell, A. D., Enquist, C. A. F., Elliott, S., Duncan, E., Coughlan de Perez, E., Coggins, S., Chen, T., Campbell, D., Browne, K. E., Bowen, K. J., Biesbroek, R., Bhatt, I. D., Bezner Kerr, R., Barr, S. L., Baker, E., Austin, S. E., Arotoma-Rojas, I., Anderson, C., Ajaz, W., Agrawal, T., and Abu, T. Z. (2021). A systematic global stocktake of evidence on human adaptation to climate change. *Nature Climate Change*, 11(11):989–1000.

Bjurström, A. and Polk, M. (2011). Physical and economic bias in climate change research: A scientometric study of IPCC Third Assessment Report. *Climatic Change*, 108(1):1–22.

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA. PMLR.

Callaghan, M., Müller-Hansen, F., Hilaire, J., and Lee, Y. T. (2020). NACSOS: NLP Assisted Classification, Synthesis and Online Screening.

Cohen, A. M., Hersh, W. R., Peterson, K., and Yen, P. Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219.

Corbera, E., Calvet-Mir, L., Hughes, H., and Paterson, M. (2016). Patterns of authorship in the IPCC Working Group III report. *Nature Climate Change*, 6(1):94–99.

Cramer, W., Yohe, G. W., Auffhammer, M., Huggel, C., Molau, U., Dias, M. A. F. S., Solow, A., Stone, D. A., and Tibig, L. (2014). Detection and attribution of observed impacts. In Field, C. B., Barros, V. R., Dokken, D. J., Mach, K. J., Mastrandrea, M. D., Bilir, T. E., Chatterjee, M., Ebi, K. L., Estrada, Y. O., Genova, R. C.,

Girma, B., Kissel, E. S., Levy, A. N., MacCracken, S., Mastrandrea, P. R., and White, L. L., editors, *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel of Climate Change*, pages 979–1037. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

David G. Victor (2015). Embed the social sciences in climate policy. *Nature*, 520:7–9.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 4171–4186.

Ford, J. D., Cameron, L., Rubis, J., Maillet, M., Nakashima, D., Willox, A. C., and Pearce, T. (2016). Including indigenous knowledge and experience in IPCC assessment reports. *Nature Climate Change*, 6(4):349–353.

Ford, J. D., Vanderbilt, W., and Berrang-Ford, L. (2012). Authorship in IPCC AR5 and its implications for content: Climate change and Indigenous populations in WGII. *Climatic Change*, 113(2):201–213.

Haddaway, N. R., Bernes, C., Jonsson, B. G., and Hedlund, K. (2016). The benefits of systematic mapping to evidence-based environmental management. *Ambio*, 45(5):613–620.

Haddaway, N. R., Callaghan, M. W., Collins, A. M., Lamb, W. F., Minx, J. C., Thomas, J., and John, D. (2020). On the use of computer-assistance to facilitate systematic mapping. *Campbell Systematic Reviews*, 16(4):1–9.

Haddaway, N. R. and Westgate, M. J. (2019). Predicting the time needed for environmental systematic reviews and systematic maps. *Conservation Biology*, 33(2):434–443.

Hansen, G. and Stone, D. (2016). Assessing the observed impact of anthropogenic climate change. *Nature Climate Change*, 6(5):532–537.

Hulme, M. and Mahony, M. (2010). Climate change: What do we know about the IPCC? *Progress in Physical Geography*, 34(5):705–718.

Hume, D. (2014). A Treatise of Human Nature. In Norton, D. F. and Norton, M. J., editors, *The Clarendon Edition of the Works of David Hume: A Treatise of Human Nature, Vol. 1: TextsTexts*. Oxford University Press, Oxford.

Ivanova, D., Barrett, J., Wiedenhofer, D., Macura, B., Callaghan, M., and Creutzig, F. (2020). Quantifying the potential for climate change mitigation of consumption options. *Environmental Research Letters*, 15(9).

Khanna, T. M., Baiocchi, G., Callaghan, M., Creutzig, F., Guias, H., Haddaway, N. R., Hirth, L., Javaid, A., Koch, N., Laukemper, S., Löschel, A., Zamora Dominguez, M. d. M., and Minx, J. C. (2021). A multi-country meta-analysis on the role of behavioural change in reducing energy consumption and CO2 emissions in residential buildings. *Nature Energy 2021*, pages 1–8.

Lehman, E., DeYoung, J. B., Barzilay, R., and Wallace, B. C. (2019). Inferring which medical treatments work from reports of clinical trials. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Figure 1):3705–3717.

Minx, J. C., Callaghan, M., Lamb, W. F., Garard, J., and Edenhofer, O. (2017). Learning about climate change solutions in the IPCC and beyond. *Environmental Science & Policy*.

Minx, J. C., Callaghan, M. W., Lamb, W. F., Atwell, E., Biaocchi, G., Berrang-Ford, L., Birkin, M., Biesbroek, R., Bornmann, L., Creutzig, F., Edenhofer, O., Ford, J., Forster, P. M., Haddaway, N. R., Hilaire, J., Kiachopoulus, Y., Krestel, R., Le Quéré, C., Müller-Hansen, F., Saudabayev, A., and Victor, D. (2021). A data science revolution for global environmental assessments. *In preparation*.

Miwa, M., Thomas, J., O'Mara-Eves, A., and Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics*, 51:242–253.

Mongeon, P. and Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106(1):213–228.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., and Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, 4(1):1–22.

Overland, I. and Sovacool, B. K. (2020). The misallocation of climate research funding. *Energy Research and Social Science*, 62(September 2019):101349.

Przybyła, P., Brockmeier, A. J., Kontonatsios, G., Le Pogam, M. A., McNaught, J., von Elm, E., Nolan, K., and Ananiadou, S. (2018). Prioritising references for systematic reviews with RobotAnalyst: A user study. *Research Synthesis Methods*, 9(3):470–488.

Rothstein, H. R., Sutton, A. J., and Borenstein, M. (2005). Publication bias in meta-analysis. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, pages 1–7.

Sietsma, A. J., Ford, J. D., Callaghan, M. W., and Minx, J. C. (2021). Progress in Climate Change Adaptation Research. *Environmental Research Letters*, Submitted.

Strubell, E., Ganesh, A., and McCallum, A. (2020). Energy and policy considerations for deep learning in NLP. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, (1):3645–3650.

Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., and Coiera, E. (2014). Systematic review automation technologies. *Systematic Reviews*, 3(1):1–15.

Wang, Z., Ng, P., Ma, X., Nallapati, R., and Xiang, B. (2020). Multi-passage BERT: A globally normalized BERT model for open-domain question answering. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, (1):5878–5882.

Writer, B. (2019). *Lithium-Ion Batteries A Machine-Generated Summary of Current Research*. Springer, Heidelberg, Germany, ebook edition.

Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., Li, M., and Lin, J. (2019). End-to-end open-domain question answering with BERTserini. *NAACL HLT 2019 - 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Demonstrations Session, pages 72–77.

Zhang, X., Wei, F., and Zhou, M. (2020). Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 5059–5069.

Chapter 6

Other publications

The work in this thesis has led to several related projects which have resulted in jointly authored publications where I am not the first author. A full list of my publications is given below, organised by methodological approach.

Using topic modelling

1. Jan C. Minx, William F. Lamb, **Max W. Callaghan**, Lutz Bornmann, and Sabine Fuss. Fast growing research on negative emissions. *Environmental Research Letters*, 12(3):035007, March 2017. Publisher: IOP Publishing.

2. William F Lamb, **Max W Callaghan**, Felix Creutzig, Radhika Khosla, and Jan C Minx. The literature landscape on 1.5[deg]C climate change and cities. *Current Opinion in Environmental Sustainability*, 30:26–34, February 2018.

3. William F. Lamb, Felix Creutzig, **Max W. Callaghan**, and Jan C. Minx. Learning about urban climate solutions from case studies. *Nature Climate Change*, 9(4):279–287, April 2019. Number: 4 Publisher: Nature Publishing Group.

Using machine learning assisted screening

4. Vivien Fisch-Romito, Celine Guivarch, Felix Creutzig, Jan C. Minx, and **Max W. Callaghan**. Systematic map of the literature on carbon lock-in induced by long-lived capital. *Environmental Research Letters*, 2020.

5. Diana Ivanova, John Barrett, Dominik Wiedenhofer, Biljana Macura, **Max Callaghan**, and Felix Creutzig. Quantifying the potential for climate change mitigation of consumption options. *Environmental Research Letters*, 15(9):093001, August 2020. Publisher: IOP Publishing.

Other publications

6. Finn Müller-Hansen, **Max W. Callaghan**, and Jan C. Minx. Text as big data: Develop codes of practice for rigorous computational text analysis in energy social science. *Energy Research & Social Science*, 70:101691, December 2020.

7. Neal R. Haddaway, **Max W. Callaghan**, Alexandra M. Collins, William F. Lamb, Jan C. Minx, James Thomas, and Denny John. On the use of computer-assistance to facilitate systematic mapping. *Campbell Systematic Reviews*, 16(4):e1129, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cl2.1129.

8. Jan C. Minx, William F. Lamb, **Max W. Callaghan**, Sabine Fuss, Jérôme Hilaire, Felix Creutzig, Thorben Amann, Tim Beringer, Wagner de Oliveira Garcia, Jens Hartmann, Tarun Khanna, Dominic Lenzi, Gunnar Luderer, Gregory F. Nemet, Joeri Rogelj, Pete Smith, Jose Luis Vicente Vicente, Jennifer Wilcox, and Maria del Mar Zamora Dominguez. Negative emissions – Part 1: Research landscape and synthesis. *Environmental Research Letters*, 13(6):063001, May 2018. Publisher: IOP Publishing.

9. Sabine Fuss, William F. Lamb, **Max W. Callaghan**, Jérôme Hilaire, Felix Creutzig, Thorben Amann, Tim Beringer, Wagner de Oliveira Garcia, Jens Hartmann, Tarun Khanna, Gunnar Luderer, Gregory F. Nemet, Joeri Rogelj, Pete Smith, José Luis Vicente Vicente, Jennifer Wilcox, Maria del Mar Zamora Dominguez, and Jan C. Minx. Negative emissions – Part 2: Costs, potentials and side effects. *Environmental Research Letters*, 13(6):063002, May 2018. Publisher: IOP

Publishing.

10. Gregory F. Nemet, **Max W. Callaghan**, Felix Creutzig, Sabine Fuss, Jens Hartmann, Jérôme Hilaire, William F. Lamb, Jan C. Minx, Sophia Rogers, and Pete Smith. Negative emissions – Part 3: Innovation and upscaling. *Environmental Research Letters*, 13(6):063003, May 2018. Publisher: IOP Publishing.

11. Jérôme Hilaire, Jan C. Minx, **Max W. Callaghan**, Jae Edmonds, Gunnar Luderer, Gregory F. Nemet, Joeri Rogelj, and Maria del Mar Zamora. Negative emissions and international climate goals – learning from and about mitigation scenarios. *Climatic Change*, 157(2):189–219, November 2019.