A Topography of Climate Change Research

Max Callaghan[1,2], Jan Minx[1,2], and Piers M. Forster[2]

[1]Mercator Research Institute on Global Commons and Climate Change, Torgauer Straße, 10829 Berlin, Germany

[2]Priestley International Centre for Climate, University of Leeds, Leeds LS2 9JT, United Kingdom

Draft current October 2, 2019

The massive expansion of scientific literature on climate change poses challenges for global environmental assessments and our understanding of how these assessments work. Big data and machine learning can help us deal with the large collections of text represented by scientific fields. Such methods help make the production of assessments more tractable, and give us better insights about how past assessments have engaged with the literature as it has evolved. We use topic modelling to identify the thematic structure and draw a comprehensive topic map, or topography, of over 400,000 scientific publications from the Web of Science (WoS) on climate change. We update current knowledge on the Intergovernmental Panel on Climate Change (IPCC), showing that, at least when compared to the baseline of the literature identified in the WoS, the social sciences are in fact over-represented in recent assessment reports, and that technical, solutions-relevant knowledge - especially in the agricultural and engineering sciences - are under-represented. We point to 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. We highlight fast-growing topics on solutions that could be better integrated into future IPCC reports. The perceived lack of social science knowledge in solutions-relevant IPCC reports does not necessarily imply a bias towards the natural sciences. It rather suggests a need for more social science research with a focus on "technical" topics related to climate solutions.

We live in an age of "Big Literature" [1, 2], where the science of climate change is expanding exponentially [3, 4]. In the five years since the publication of the last IPCC assessment report [5], 202,000 papers on climate change were published in the Web of Science (WoS) (see Table 1). This is almost as much as the 205,000 papers identified in the same query [3] during the first five assessment periods; a period of nearly 30 years. Around 350,000 new publications can be expected for before the sixth assessment report (AR6) of the Intergovernmental Panel on Climate Change (IPCC), based on current growth patterns (Figure 1). Moreover, from the expansion of the literature's vocabulary (see methods) - from 2,000 unique words in the first assessment period to 95,000 words so far in the sixth - we can observe the literature's increasing diversity of content. For example, the zika virus, 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, whose outbreak in Brazil in 2016 was declared a public health emergency by the World Health Organization, is set to increase under rising global temperatures [6]. Similar rapid emergence patterns can be seen for Intended Nationally Determined Contributions (INDCs) in AR6, and Biochar in AR5, among others[1].

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

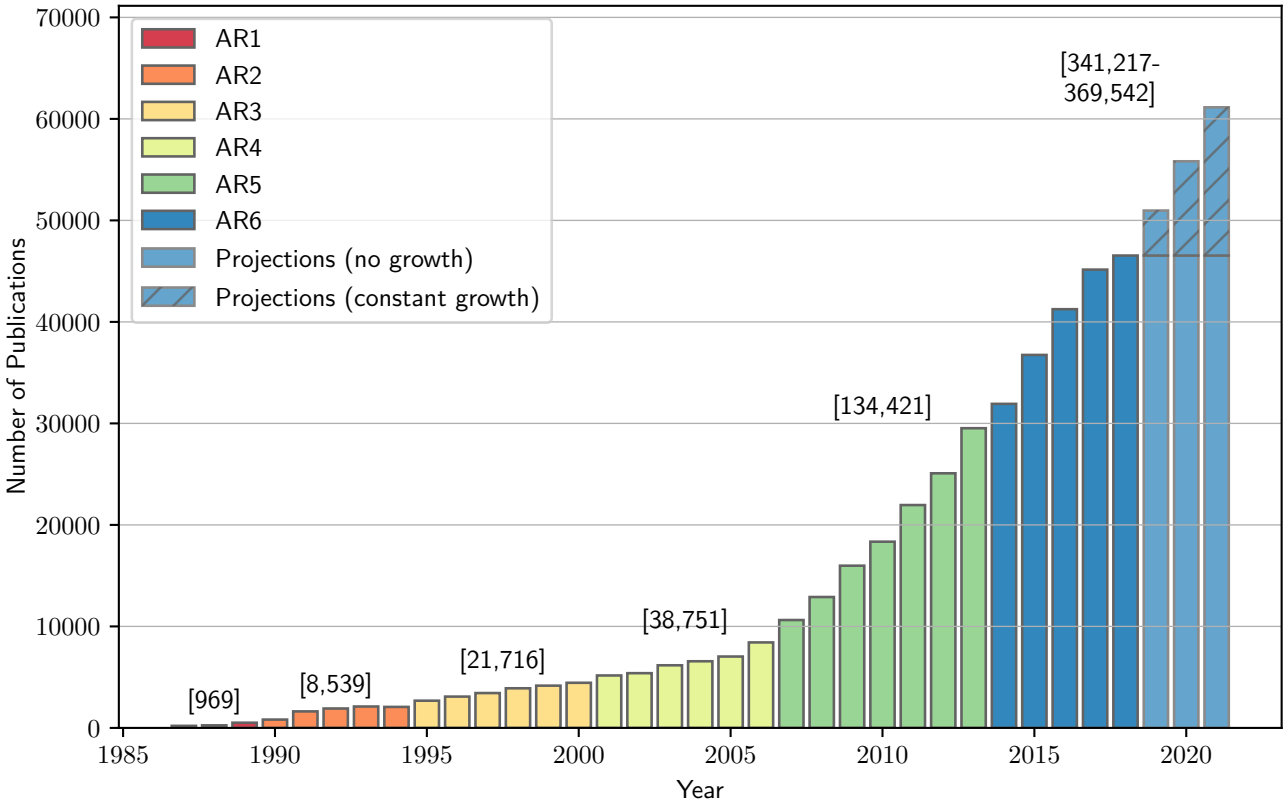

Figure 1: The number of climate change documents in the Web of Science in each year. A total of 406,191 documents were published until the end of 2018. The number of publications in each assessment period is shown in square brackets. For 2019-21 we project the number of papers assuming there is no more 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 1: Growth of Literature on Climate Change. A glossary of acronyms is provided in SI

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 struggle to address the exploding literature base. For example, the ratio of studies cited in IPCC reports to the number of studies on climate change in the WoS has declined from 60% to 20% [2], posing a rapidly growing risk of selection bias. The exponentially increasing volume of literature means that the provision of "comprehensive, objective, open and transparent" assessments of the available scientific literature, as defined in the principles governing IPCC work [7], is no longer possible by traditional means. Machine reading and learning methods, among other data science applications, are required to enable an understanding of the field of climate change research at scale. *Second*, evidence synthesis - the enterprise of reviewing the literature based on a formal and systematic set of methods [8] - becomes increasingly important for aggregating and consolidating 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 [9]. *Finally*, overwhelming amounts of literature may be a major reason why studies of scientific assessments [10] do not offer robust quantification for their claims about the relationship between report citations and the underlying literature.

This study uses topic modelling [13] to map the vast body of evidence on climate change. Topic modelling is an unsupervised machine-learning technique, where patterns of word co-occurrences are used to learn a set of topics, groups of words, which 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 (Figure 2), 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 then use this map to understand how 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 on representation by discussing topics as well as disciplines.

Mapping the landscape of climate change literature

Figure 2 shows a *thematic* or *topographic map* of the 378,000 publications on climate change in our dataset with abstracts. Using non-negative matrix factorization [14], the 140 topics are machine-learned from the papers' abstracts (see methods for details). The topic scores of each document are reduced to the two dimensions shown through t-distributed stochastic neighbour embedding (t-SNE) [15]. [2]. The two dimensions represent a projection of the 140-dimensional topic scores of each document that seeks to preserve small distances between topically similar documents.

Our map covers a broad range of topics, with related topics in clusters. Generally, topics related to climate science and impacts are on the left, while solution-oriented topics are on the right. More fine-grained research areas can also be distinguished. For example, publications related to urban infrastructure (**buildings**, **cement**, **waste**) are located on the right, physical climate impacts (**sea-level**, **droughts** or [crop] **yield**) are in the lower left and energy systems are in upper right. Larger groups of documents at the fringes of the map relate mainly to one or two specific topics like **biochar** or **coral**. Interestingly, scenarios feature centrally in the map, at the interface between different scientific communities. This corresponds to their integrative nature in IPCC reports [16].

The disciplinary composition of this research topography indicated by the different colours in Figure 2 highlights the dominance of natural sciences in climate change research. More than 60% of the literature is published in natural science journals. Similarly, 115 of 140 topics contain a greater share of publications from natural science journals than

[2] 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

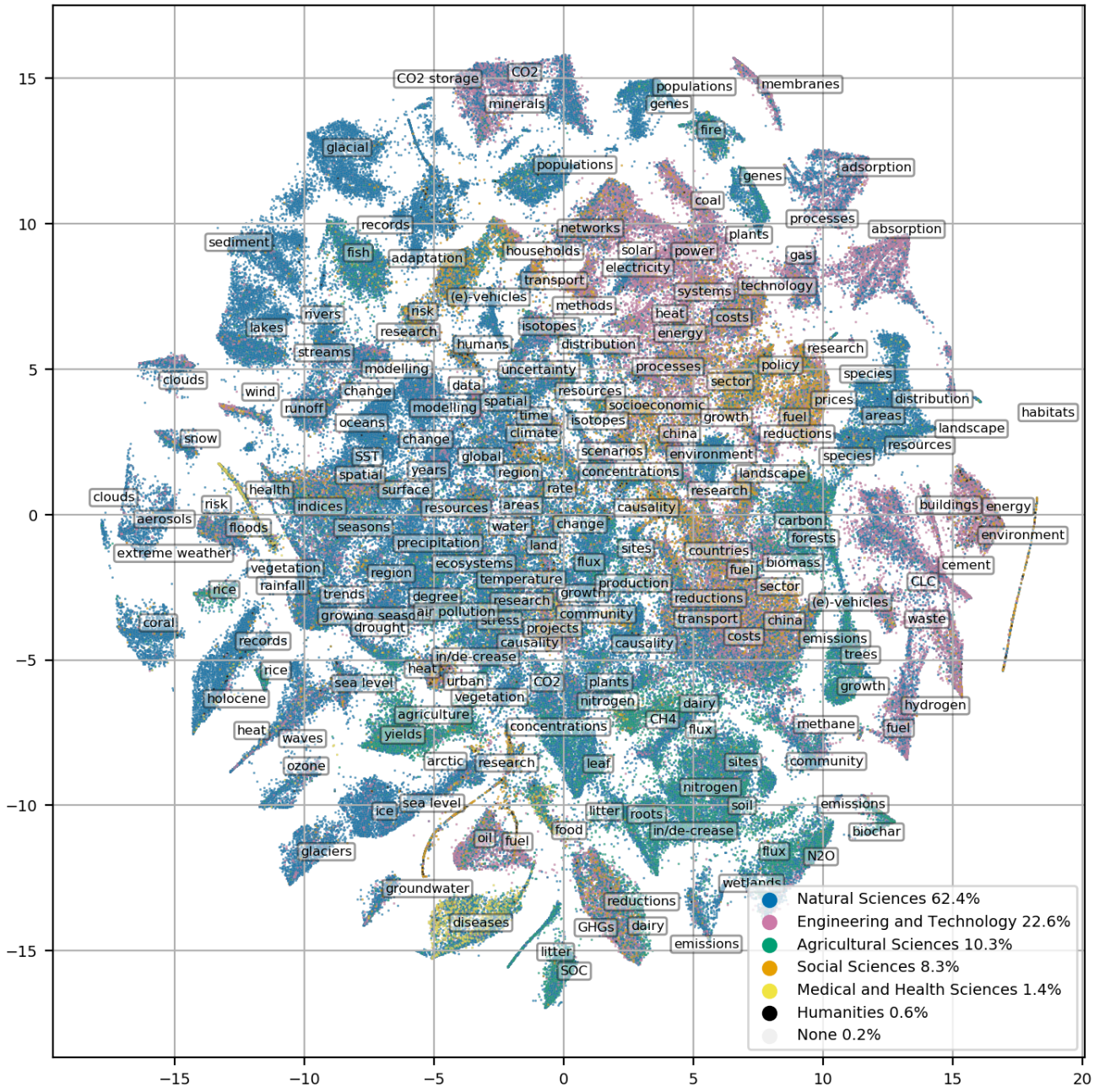

Figure 2: A map of the literature on climate change. Document positions are obtained by reducing the topic scores to two dimensions via t-SNE (see methods for further details). The two axes therefore have no direct interpretation, but represent a reduced version of similarities between documents across 140 topics. 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.

any other discipline. We calculate disciplinary entropy of topics as a measure of their degree of interdisciplinarity (Figure SI.1 and methods for details). This shows how research on **health**, **food**, or **policy** comes from a range of disciplines, while research on **ice** or **oceans** comes almost exclusively from the natural sciences).

Finally, the topography shows the thematic evolution of the literature (Figure 3), 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.

Research representation in IPCC reports

We apply our topic map to understand how IPCC assessments represent the science and respond to policymakers' and consulted experts' demands for more solution-oriented knowledge [17]. 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 [10, 12, 11, 18]. Where these claims were based on an analysis of IPCC citations [10], they assess this without measurable baseline. In view of the organisation's mandate to provide "comprehensive, objective, open and transparent" assessments of the available science [7], our dataset of publications allows us - albeit imperfectly, as discussed in the concluding section - to study representation with a meaningful baseline. Further we provide an update to the last quantitative assessment of IPCC citations [10], which looked only at AR3. This baseline forms a starting point for informed discussion about how to represent the literature according to the IPCC's priorities.

By matching the documents in our dataset to a set of references scraped from all published IPCC reports [2], we assess the representation of a group of studies by comparing its share in IPCC citations with its share in the dataset of WoS studies on climate change (see methods). Figure 4.a shows that social science documents (as identified by WoS) were indeed under-represented in AR3, but by AR5 were the most over-represented discipline, with a share in the literature cited by IPCC reports 1.32 times higher than their share in our WoS dataset. Likewise, social & economic geography, political science, and "Other social sciences" were better represented in AR5 than economics. This challenges what we think we know about the IPCC. Instead of under-representing the social sciences, the IPCC has been under-representing the Agricultural Sciences and Engineering & Technology.

The topography allows us to delve deeper into subjects that receive more or less attention in the IPCC. Figure 4c shows that topics more commonly cited by IPCC working group I (WGI) are older and largely better represented in IPCC reports. These topics, for example **ozone**, **oceans**, and **aerosols**, are core topics for WGI, which addresses 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. They are newer and until now have been under-represented in IPCC reports. Their novelty may be highly salient in a periodic assessment process. These topics are primarily in working group III, on mitigation and are "solutions-relevant". But while policymakers' demands for solutions-oriented IPCC assessments were often focussed on policy options, these under-represented new topics deal with more technical solutions and are found in technical disciplines within engineering & technology and the agricultural sciences.

Further, WGIII topics that are well represented contain a greater proportion of social science research (figure 4b). The topics **countries**, **policy**, and **prices** are close to a proportional representation and are made up of around 30% social science research. **Waste**, **biochar**, and **cement**, are more than 3 times more prevalent in the wider literature

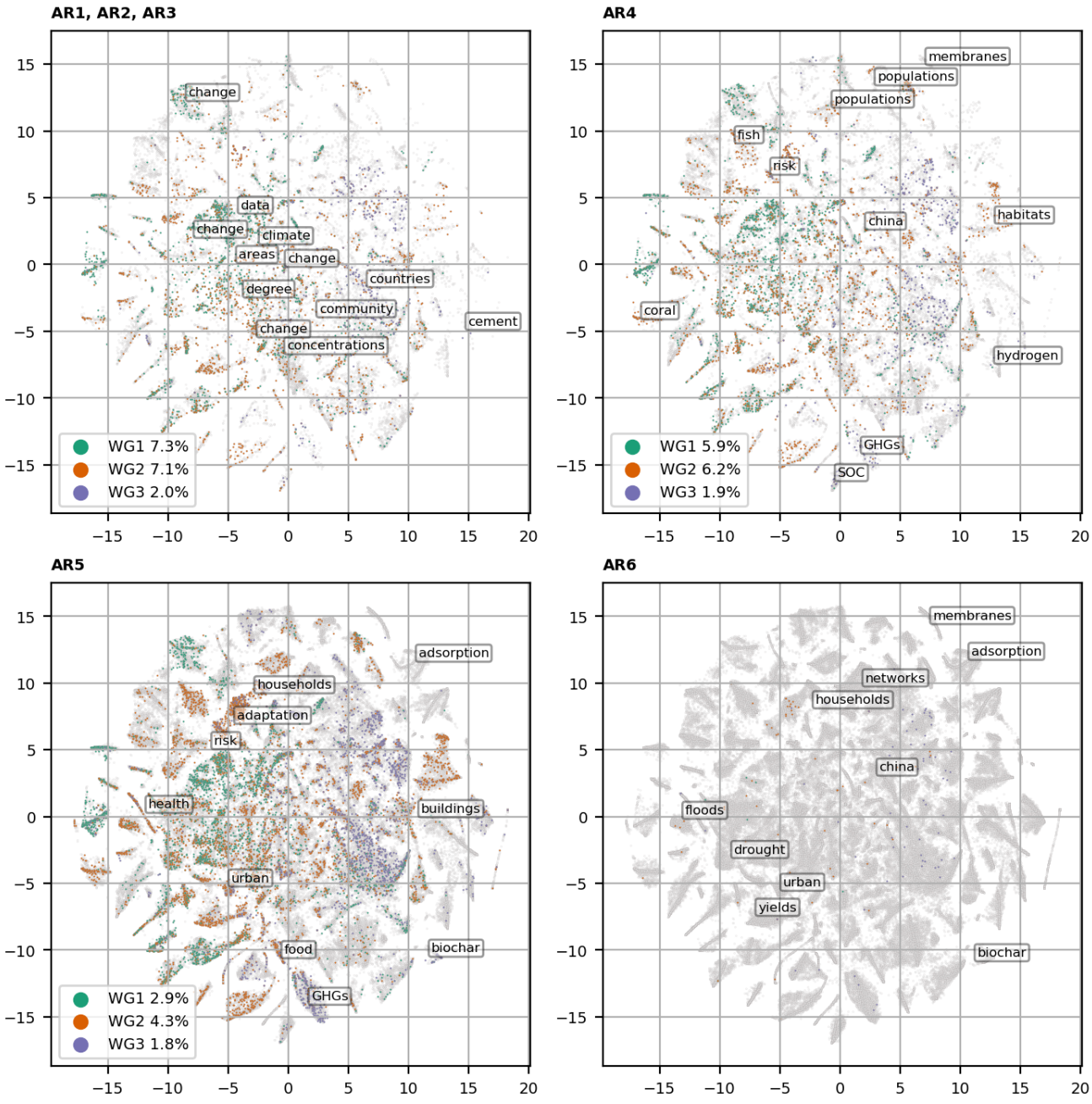

Figure 3: Evolution of the landscape of climate change literature. In each period, the 10 fastest growing topics are labelled. Where documents could be matched to IPCC citations, they are coloured by the working group citing them.

than in the literature cited by the IPCC, and are made up of around 5% social science research. This pattern is not visible in other working groups (Figure SI.4).

The difference between under-represented new topics and new topics that are better represented is intriguing. This is visible in figure 3, where in AR5, the clusters of documents around the, **buildings** and **biochar** topics contain few IPCC citations, whereas the clusters around, **adaptation** and **food** contain more. As shown in figure 4c, **buildings** and **biochar** are 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 **adaptation** is 2.22 times more prevalent in IPCC citations respectively.

Machine-learning for climate change assessments

Notwithstanding the over-representation of social science and under-representation of technical solutions in the IPCC with respect to the WoS, a perfectly proportional representation of the literature is of course not optimal. A recommendation that the IPCC cite more or less of any part of the literature is by no means the goal of such an analysis. The IPCC, as a community of scientific experts, is vastly better placed to decide what is relevant than any algorithm. As with many machine learning applications, we should be mindful of David Hume's is-ought problem. Machine learning can help us to more efficiently understand and describe the landscape of climate change literature, but cannot tell us how things should be. The results represent new knowledge about the interaction between the IPCC and the literature, which can have a variety of implications. If the IPCC needs to include more social science knowledge [12], our analysis suggests that this is a result of insufficient production or funding of social science research on climate, rather than IPCC bias. The under-representation of solutions-relevant topics (despite calls for solutions-oriented assessments), and the small proportion of social science research within these topics, suggests areas for future highly relevant social science research, as well as opportunities for particularly fruitful interdisciplinary collaboration.

As a guide for future assessments, the map could facilitate well informed decisions about the representation of different areas of climate literature, from the early scoping process, through to selection by authors of individual studies. One advantage of topic modelling is that outcomes are not determined by any categorisation scheme imposed by the modeller, facilitating the discovery of "unsearched" for topics. Highlighting recent research on, for example, membranes, biochar or e-vehicles, could prompt discussion in the scoping process about their inclusion in chapter outlines. This mode of discovery can act as a complement to human expertise, which may be better at identifying under-researched niches, existing biases or knowledge requirements. The methods shown here could also aid other processes in the production of IPCC reports, such as the identification of potential authors to achieve a better balance across sectors, regions and genders [18]. The possible benefits or risks of using data science methods for IPCC processes constitutes an important area for future research. Outside of the IPCC, this approach is part of ongoing attempts to make use of machine learning within evidence synthesis. This topographic map is a new approach to rapidly mapping very large literatures.

Our dataset of more than 400,000 publications represents a wealth of knowledge on climate change and climate solutions, but is by no means exhaustive. We repeat an established query [4], granting that it may have imperfections. Furthermore, we miss publications not in WoS (some small journals, some books, and most grey literature, not to mention indigenous knowledge [19]); and studies relevant for the work of the IPCC, that do not directly mention climate change (for example on energy policy). We argue that this remains a reasonable system boundary given data availability, and stress that documents not included in our study alter our findings only if they have systematically different patterns of citation by the IPCC. A future topography could be improved by making use of more sources of climate change knowledge, extracting and classifying information from full texts, or exploring author networks and

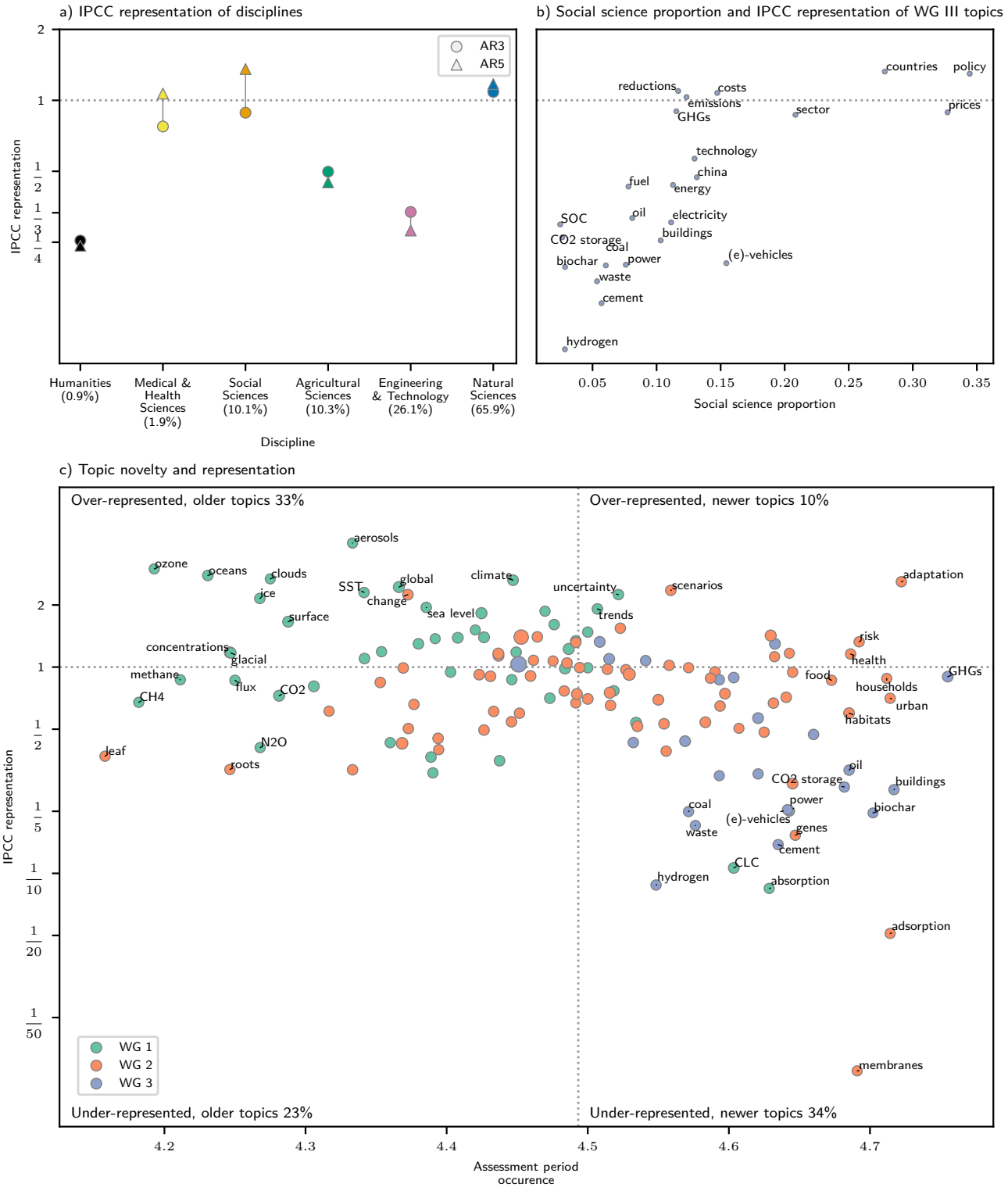

Figure 4: Representation in IPCC reports: **a)** by discipline, **b)** by social science proportion of WGIII topics, **c)** and novelty of all topics, where topics in the highest and lowest 10% of either axis are labelled. Topics are coloured according to the working group from which they receive the most citations, although infrequently cited topics may not correspond to the relevant working group (see methods). 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. Assessment period occurrence refers to the center of a topic's distribution across assessment periods (see methods for further details).

interdisciplinarity. Most importantly, exploring machine learning applications that support IPCC authors in their assessments would prepare the IPCC for the age of big literature.

References

[1] Gabriela C Nunez-Mir, Basil V Iannone, Bryan C Pijanowski, Ningning Kong, Songlin Fei, and Richard Fitzjohn. Automated content analysis: addressing the big literature challenge in ecology and evolution. *Methods in Ecology and Evolution*, 7(11):1262–1272, 2016.

[2] Jan C. Minx, Max Callaghan, William F. Lamb, Jennifer Garard, and Ottmar Edenhofer. Learning about climate change solutions in the IPCC and beyond. *Environmental Science & Policy*, 2017.

[3] Michael Grieneisen and Minghua Zhang. The Current Status of Climate Change Research. *Nature Climate Change*, 1:72–73, 2011.

[4] Robin Haunschild, Lutz Bornmann, and Werner Marx. Climate Change Research in View of Bibliometrics. *PLoS ONE*, 11(7):1–19, 2016.

[5] IPCC. *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, 2014.

[6] V. Brahmananda Rao, K. Maneesha, Panangipalli Sravya, Sergio H. Franchito, Hariprasad Dasari, and Manoel A. Gan. 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, 2019.

[7] IPCC. Principles governing IPCC work, 2013.

[8] Iain Chalmers, Larry V Hedges, and Harris Cooper. A Brief History of Research Synthesis. *Evaluation & The Health Professions*, 25(1):12–37, 2002.

[9] Elaine Beller, Justin Clark, Guy Tsafnat, Clive Adams, Heinz Diehl, Hans Lund, Mourad Ouzzani, Kristina Thayer, James Thomas, Tari Turner, Jun Xia, Karen Robinson, Paul Glasziou, and On behalf of the founding members of the ICASR group. 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, 2018.

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

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

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

[13] David Blei, Lawrence Carin, and David Dunson. Probabilistic topic models. *IEEE Signal Processing Magazine*, 2010.

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

[15] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[16] Richard H. Moss, Jae A. Edmonds, Kathy A. Hibbard, Martin R. Manning, Steven K. Rose, Detlef P. Van Vuuren, Timothy R. Carter, Seita Emori, Mikiko Kainuma, Tom Kram, Gerald A. Meehl, John F.B. Mitchell, Nebojsa Nakicenovic, Keywan Riahi, Steven J. Smith, Ronald J. Stouffer, Allison M. Thomson, John P. Weyant, and Thomas J. Wilbanks. The next generation of scenarios for climate change research and assessment. *Nature*, 463(7282):747–756, 2010.

[17] Martin Kowarsch, Jason Jabbour, Christian Flachsland, Marcel T. J. Kok, Robert Watson, Peter M. Haas, Jan C. Minx, Joseph Alcamo, Jennifer Garard, Pauline Riousset, László Pintér, Cameron Langford, Yulia Yamineva, Christoph von Stechow, Jessica O'Reilly, and Ottmar Edenhofer. A road map for global environmental assessments. *Nature Climate Change*, 7(6):379–382, 2017.

[18] Esteve Corbera, Laura Calvet-Mir, Hannah Hughes, and Matthew Paterson. Patterns of authorship in the IPCC Working Group III report. *Nature Climate Change*, 6(1):94–99, 2016.

[19] James D. Ford, Laura Cameron, Jennifer Rubis, Michelle Maillet, Douglas Nakashima, Ashlee Cunsolo Willox, and Tristan Pearce. Including indigenous knowledge and experience in IPCC assessment reports. *Nature Climate Change*, 6(4):349–353, 2016.