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 July 24, 2019

The massive expansion of scientific literature on climate change poses challenges for global environmental assessments and how we understand them. 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 on climate change. We update current knowledge on the IPCC, showing that 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 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 were published in the Web of Science (WoS) (see Table 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 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 [6]. 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% [2], 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 [7], 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

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

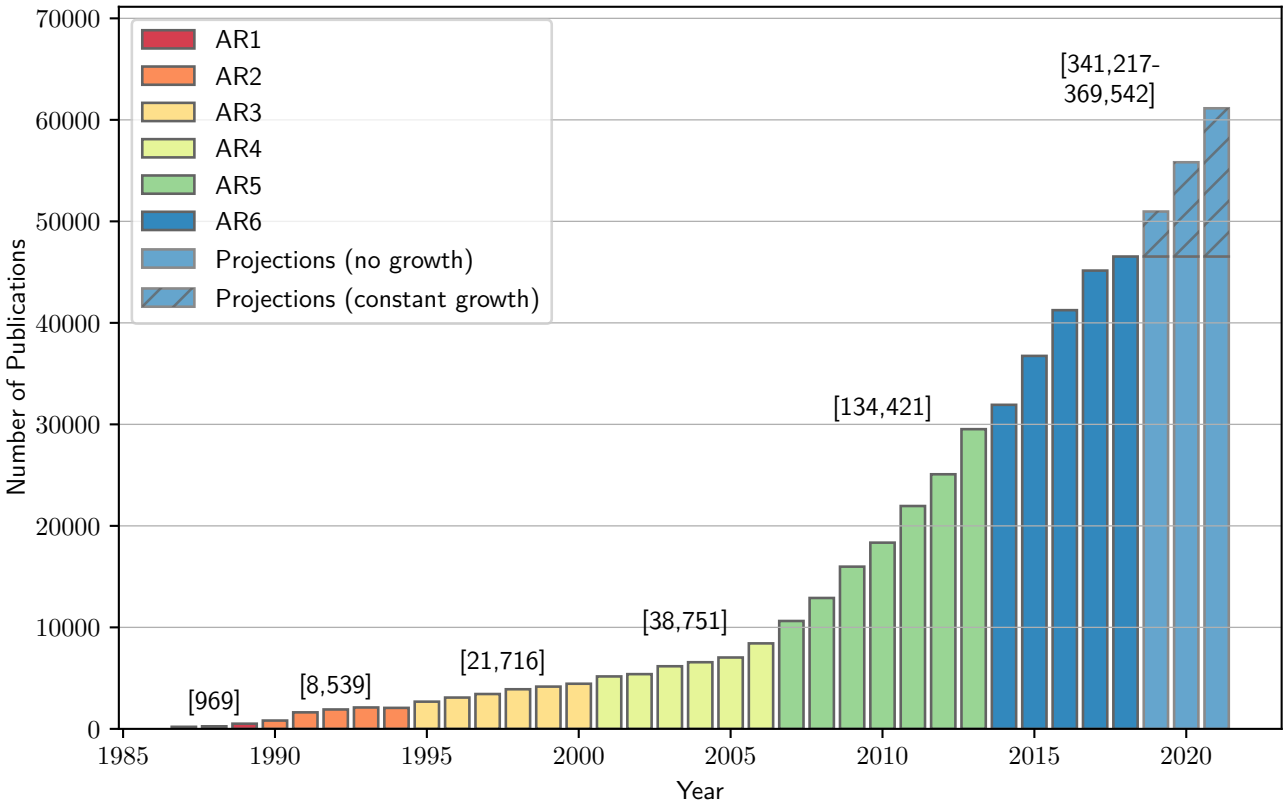

Figure 1: The number of climate change documents in the Web of Science in each year. For 2019-21 we project the number of papers assuming there is no more growth, and assuming that growth continues at the same rate as over the past five years

	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

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 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 [9]. *Finally*, overwhelming amounts of literature may be a major reason why studies of scientific assessments [10, 11, 12] do not offer robust quantifications, for claims about the relationship between report citations and the underlying literature.

This study uses topic modelling [13] 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 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 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.

Mapping out the landscape of climate change literature

Figure 2 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 [14], 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 [15] [2].

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 [16]. 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 2 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 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** and **oceans** comes almost exclusively from the natural sciences).

[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

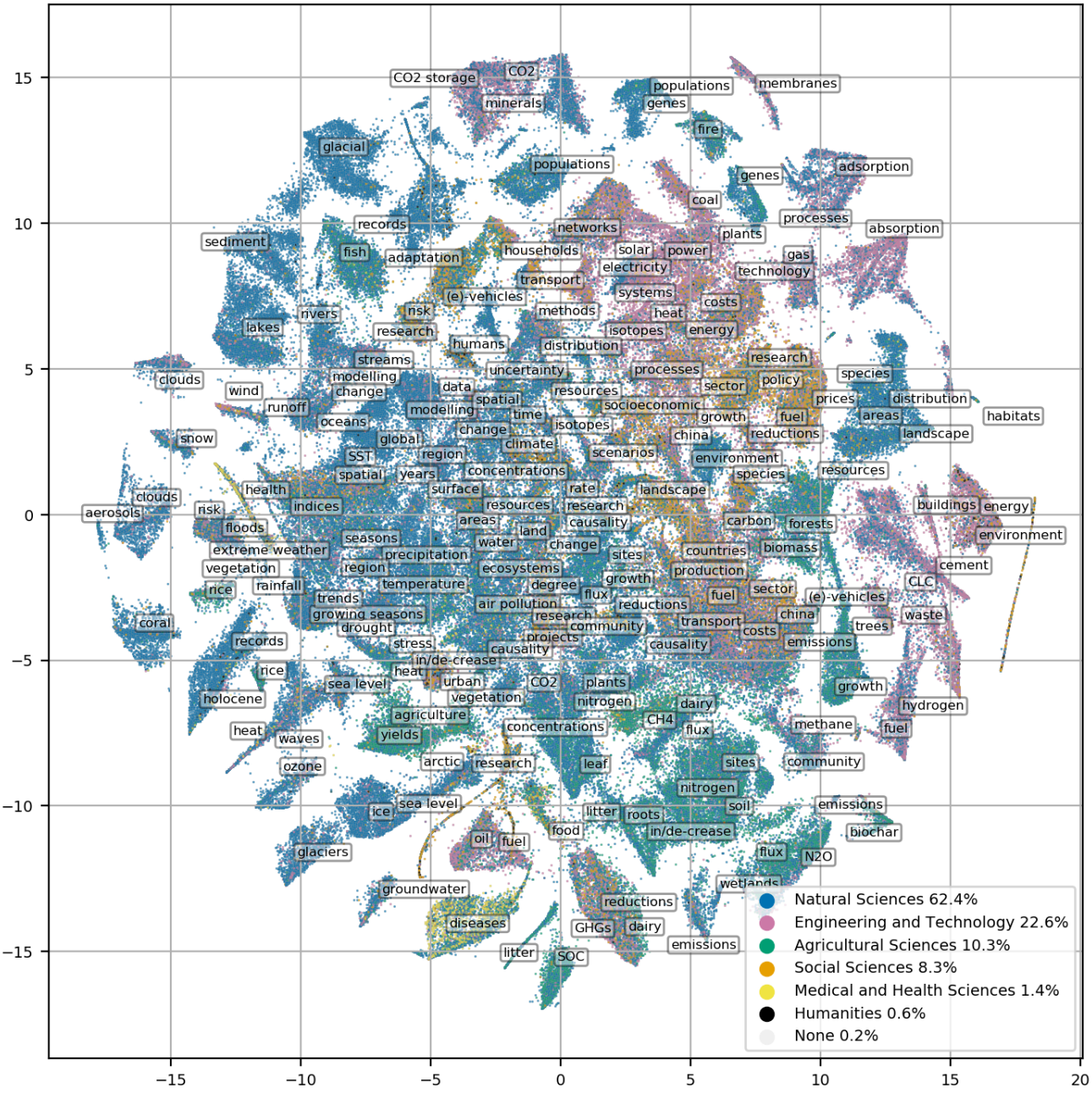

Figure 2: A map of the literature on climate change. Document positions are obtained by reducing the topic scores to two dimensions via t-SNE Documents are coloured by web of science discipline category. Topic labels are placed in the center of each of the large clusters of documents associated with each topic.

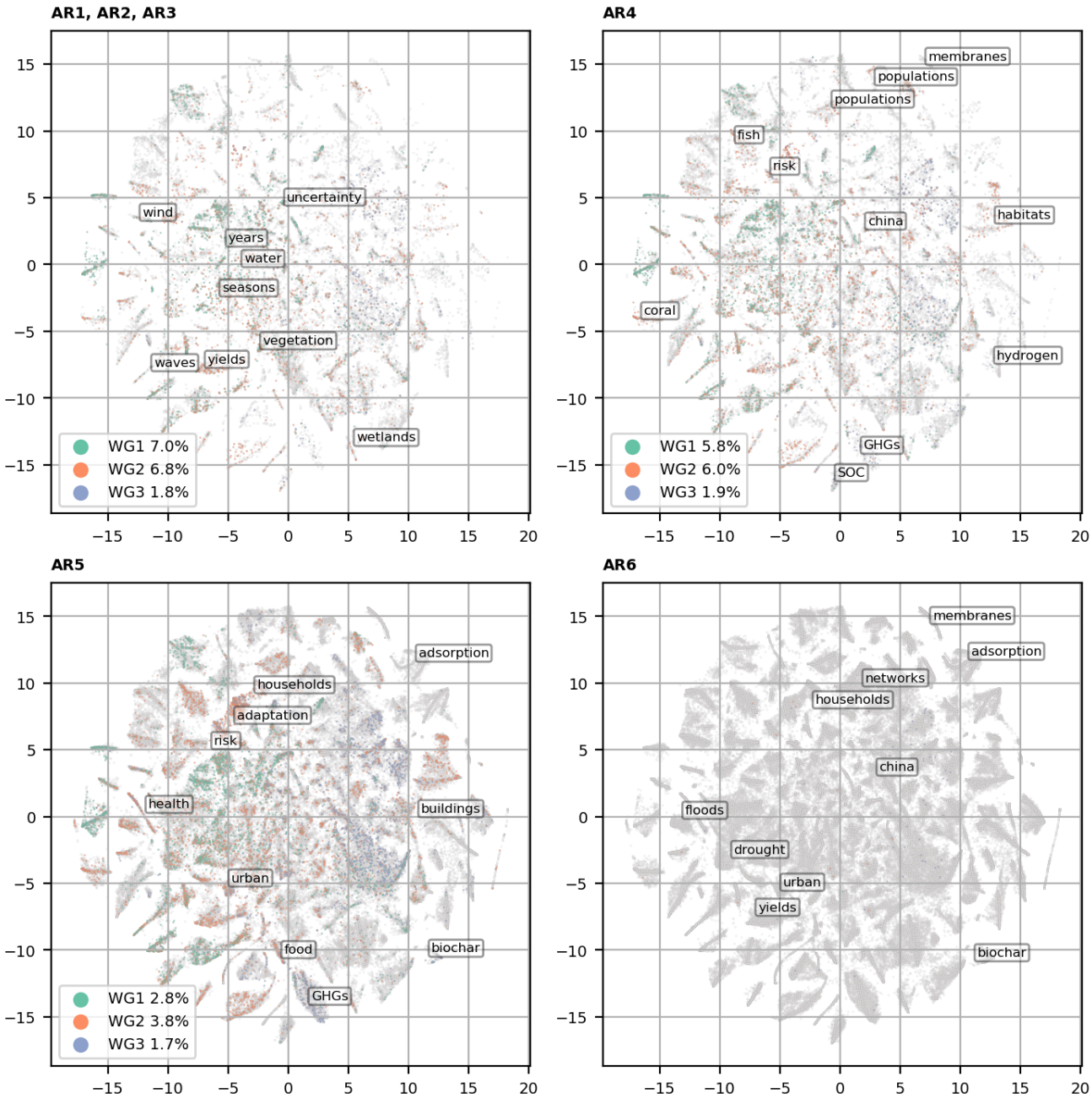

Figure 3: Evolution of the landscape of climate change literature. In each period, the 10 fastest growing topics are labelled. Where documents could be matched to IPCC citations, they are coloured by the working group citing them.

Finally, the topography shows the thematic evolution of the literature (Figure 3), 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 the representation of science in IPCC assesssments and how it manages to respond to 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 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 [7]. Our database of publications allows us to study representation with such a benchmark, and over time rather than for single assessment cycles.

Figure 4.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 SI.3f) 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 4b and 4c plot the representation of the topics shown in the map. Figure 4c 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 [3].

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, 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**,

[3] 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

and **GHGs** contain more. As shown in figure 4c, **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 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**, **cement** and **coal**, are more than 3 times more prevalent in the wider literature 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 (see Figure SI.4), and complicates the perception of the under-representation of the social sciences.

Recalling policymakers' demands for more solution-oriented assessments [17], 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.

Machine-learning for climate change assessments

We have shown that social science literature about climate change is over-represented in the IPCC, while technical literature on solutions is under-represented. A perfectly proportional representation of every part of the literature is of course not optimal, yet these two features represent new knowledge about the interaction between the IPCC and the literature, and have important implications for the community. The over-representation of the social sciences, combined with the perception that the IPCC needs to include more social science knowledge [12], suggests that doing so must involve the funding and production of more social science literature on climate change, not just greater efforts by the IPCC to include it. Moreover, the fact that solutions-relevant topics are under-represented - while policymakers demand more solutions-oriented assessments - and that these topics contain little social science research, suggests areas where social science research may be usefully conducted. Further, this fact can open a debate about the extent to which the IPCC should include more technical knowledge on solutions from disciplines within engineering and the agricultural sciences.

The map can also serve as a guide to future assessments, to ensure that decisions about how different areas of the literature are represented are well-informed. The map can contribute to these decisions as they are made from the scoping process, through to the selection by authors of individual studies. Beyond this, the methods shown here could aid other processes in the production of IPCC reports, such as the identification of potential authors to acheive a better balance across sectors, regions and genders [18]. Outside of the IPCC, this approach is part of ongoing attempts to make use of machine-learning within evidence synthesis. The topographic map presented is a new approach to rapidly mapping very large literatures.

The more than 400,000 publications we deal with here represent a wealth of knowledge on climate change and climate solutions. However, we acknowledge that our dataset is by no means exhaustive. We repeat an established query [4], granting that it may have imperfections. Beyond this we miss publications not in the Web of Science (some small journals, some books, and most grey literature, not to mention indigenous knowledge [19]); and studies relevant for the work of the IPCC, but that do not directly mention climate change (for example on energy policy). We argue that this remains a reasonable system boundary given data availabilty, and stress that the documents not included in

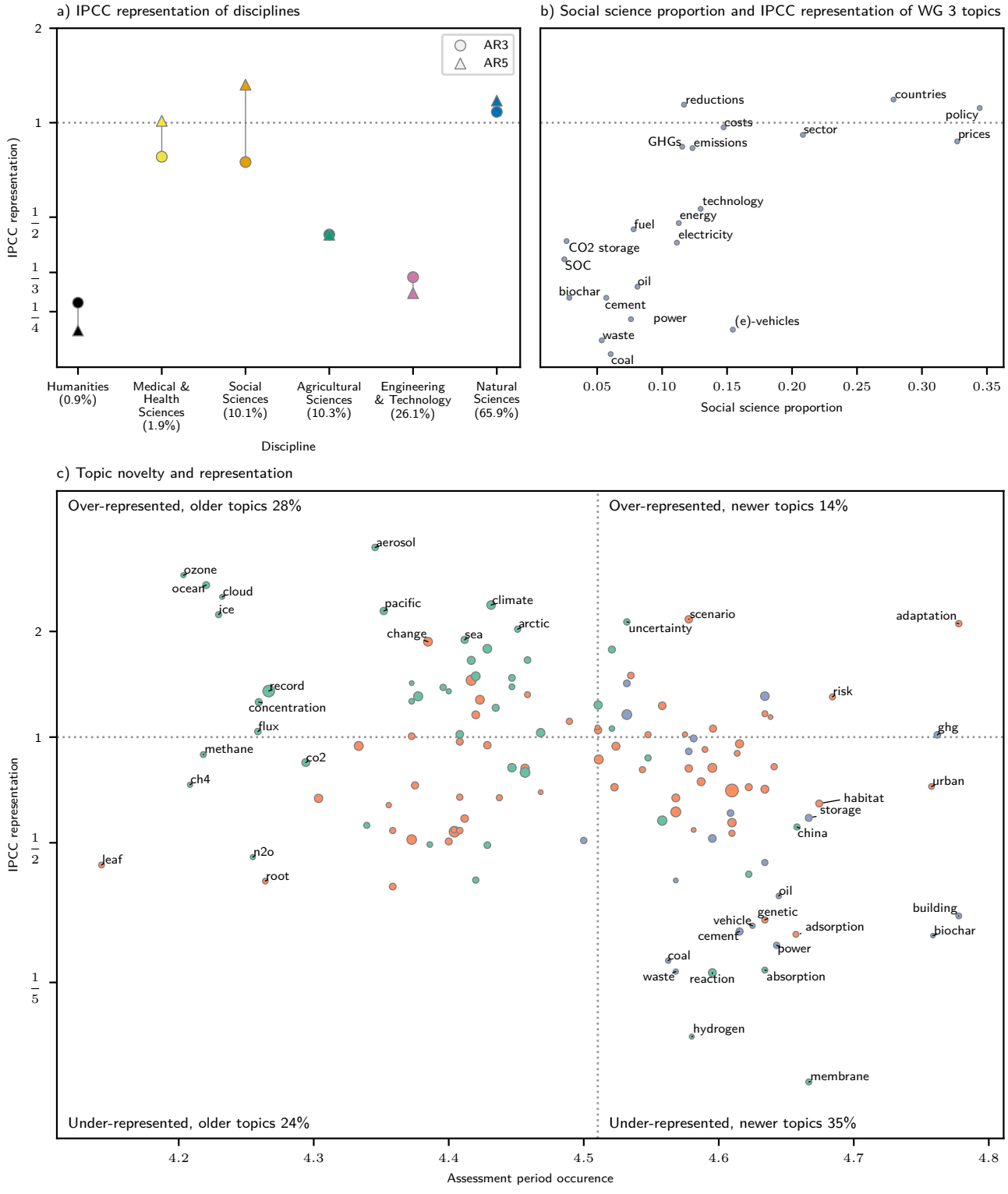

Figure 4: Representation in IPCC reports: **a)** by discipline, **b)** by social science proportion of WG 3 topics, **c)** and novelty of all topics, where topics in the highest and lowest 10% of either axis are labelled. Topics are coloured according to the working group from which they receive the 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.

our study alter our findings only if they have systematically different patterns of citation by the IPCC. In the future, making use of more sources climate change knowledge, and extracting and classifying information from full texts, could improve the granularity of this topography. Most importantly, machine learning applications should be explored that support IPCC authors in assessing the literature. This would prepare IPCC assessments 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.