A Topography of Climate Change Research

Max Callaghan[1,2]

[1]Mercator Research Institute on Global Commons and Climate Change, Torgauer Straße, 10829 Berlin, Germany
[2]School of Earth and Environment, University of Leeds, Leeds LS2 9JT, United Kingdom

Draft current May 31, 2019

The massive expansion of scientific literature on climate change challenges the Intergovenmental Panel on Climate Change (IPCC)'s ability to assess the science according to its objectives. Moreover, the number and variety of papers hinders researchers of the science-policy interface from making objective judgements about those IPCC assessments. In this paper, we present a novel application of a machine-reading approach to model the topical content of 400,000 papers on climate change. This topic model provides the basis for a *topography* of climate change literature. By comparing the climate change literature as a whole to that which is cited by IPCC reports, we show that, contrary to previous estimations [1], social science research on climate change is relatively well represented by IPCC reports. Conversely, topics on technical solutions to climate change, and knowledge from Engineering and the Agricultural Sciences in general, are under-represented.

We know that the scientific literature on climate change is growing rapidly [2, 3], and that this growth poses problems for the IPCC. Their task of providing a comprehensive and transparent summary of the literature on climate change is severely challenged by the growth of the literature: the proportion of the literature that gets cited by the IPCC is dwindling [4]. In the light of policymakers' demand for more solutions-oriented knowledge from assessments [5]; calls for more representation of certain disciplines in IPCC reports [6, 7]; and criticism of bias in the knowledge reflected by the IPCC [1, 8], what gets cited and what doesn't is more salient than ever.

However, we know little about thematic trends within the literature, or how the IPCC performs in reflecting different parts of this growing body of knowledge on climate change. We lack an overview of the literature itself, and without this, our understanding of the IPCC [] lacks the context of the literature which it is supposed to assess. Mapping out this literature, and its reflection in IPCC reports is a crucial task if we are to assess and improve the IPCC's comprehensiveness.

Table 1 depicts the scale of the challenge faced by the IPCC. Since the publication of the fifth assessment report (AR5), almost as much literature has been published as in the 30 years previously. Moreover, not only are more articles being published, but the vocabulary of climate knowledge has expanded. While the 8,539 documents published in AR2 contained 12,480 unique words, the 201,606 documents published in AR6 contain a vocabulary of 94,746 unique words.

The zika virus, the Sustainable Development Goals (SDGs), Intended Nationally Determined Contributions (INDCs), and mixed matrix membranes (MMMs) are all significant parts of the literature since 2014 which were simply not discussed in the context of climate change before the last IPCC report. In view of this expanding vocabulary, this study employs topic modelling to draw out patterns in the content of scientific literature. Topic modelling is an unsupervised machine-learning technique, where patterns of word co-occurrences in documents are used to learn a

	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
Words	2000	12480	23346	34637	71867	94746
New words	change (560)	oil (287)	downscaling (217)	sres (234)	biochar (1791)	mmms (313)
	climate (428)	deltac (283)	degreesc (187)	petm (95)	redd (1113)	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

set of topics which can be used to describe the corpus [9]. It is applied here for the first time to the whole scientific literature on climate change.

A Topography of Climate Change Research

Topics, from the greek word “topos” (meaning place), refer here to concepts or themes within the literature. Using the same geographic-thematic metaphor, the topographic map shown in figure 1 *situates* the 400,000 documents about climate change in a topical landscape derived from the 140 topics discovered through topic modelling. Using t-distributed stochastic network embedding [10] (t-SNE) as a dimensionality reduction technique, the 140 dimensional topic space is *projected* onto two dimensions, such that documents with similar topical content are placed next to each other. If scientists, in global environmental assessments, act as mapmakers [11], this topography, as a broad outline of the features of the research landscape, help understand the relationship of the map to the mapped.

To make the map readable, we find clusters of documents relating to each topic, and superimpose the topic label in the center of each cluster (see methods for a more detailed description). Documents are coloured by their categorisation in the Web of Science into broad disciplinary categories, showing us how the topical content of the documents fits into disciplinary structures. Across the map, it is clear that different disciplines focus on different groups of topics, with more natural science research on the left hand side of the map, more engineering in the upper right and right, more agricultural sciences at the bottom of the map, and more social sciences on the inner right.

Get ri
numbe
replac
letters

The map also allows us to examine topics consisting of research from across disciplines. With reference to figure **??** which shows the disciplinary diversity of each topic, we can see that research on coral comes almost exclusively from the natural sciences, while papers discussing socioeconomic issues come from a variety of disciplines.

The map plays an important role in generating knowledge about what we do and do not know about climate change. Though not strictly systematic, in that all studies have been checked for relevance and categorised by hand, it moves in the direction of evidence maps, or systematic maps [12, 13], by categorising and quantifying the literature and its uptake by the IPCC. The approach is weakly systematic in the sense that the studies are selected according transparent criteria (see methods), and are categorised consistently by the topic model. Further, that the categories are learnt from the data without human input, means that we can find patterns we didn’t know we were looking for.

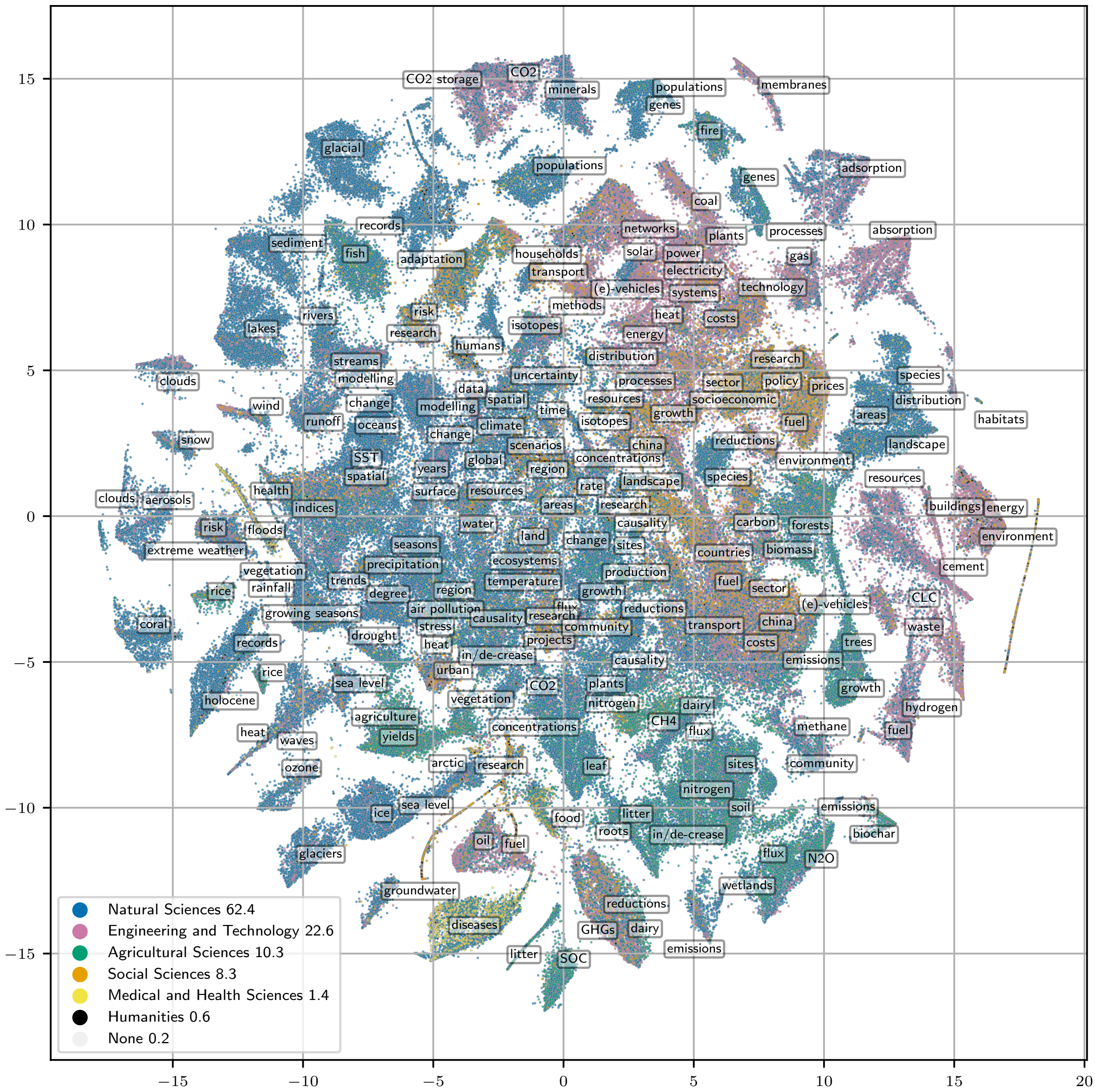

Figure 1: A map of the literature on climate change. Document positions are obtained by reducing the topic scores to two dimensions via t-SNE Documents are coloured by web of science discipline category. Topic labels are placed in the center of each of the large clusters of documents associated with each topic.

Disciplinary bias in IPCC citations?

It was argued after the fifth assessment report that the IPCC needed to do more to incorporate knowledge from the social sciences [6]. Further, a scientometric study from 2011 claimed that the IPCC gave a greater *emphasis* to natural sciences and, within the social sciences, to economics [1]. This claim has been interpreted as a disciplinary bias by the IPCC [8, 14], but the study operationalised disciplinary emphasis as simply the share of citations from each field, and was based on analysis of the Third Assessment report, published in 2001. The share of citations in each discipline does not take into account the distribution of climate change research across disciplines in the wider literature. Here we look at *representativeness*, that is, the share of IPCC citations in each field divided by the share of all climate related documents in that field, and carry out this analysis across assessment periods.

Looked at this way (Figure 2.a), we see that the social sciences were indeed under-represented in the third assessment report, but by the fifth assessment report were over-represented. The disciplines under-represented in IPCC reports (with respect to the distribution of studies in the wider literature) are in fact Agricultural Sciences, Engineering & Technology and Humanities (although the humanities make up a very small proportion of the literature). In each field, the under-representation has been present across assessment reports.

A similar story is visible within the social sciences (see figure **??**f). Economics was previously over-represented among social sciences, while other subfields were under-represented. In AR5, though, the share of economics citations in the IPCC was close to that in the wider literature, while social & economic geography (4.3% of the literature), political science (1.0%), and sociology (0.8%) were better represented than in AR3 and above or close to a proportional representation.

With more up to date data and a more comprehensive method, both results contradict previous evidence on disciplinary representation in the IPCC. This new evidence is a compelling reason to rethink our view of the relationship between disciplinary knowledge and the IPCC.

Supply and demand for solutions-oriented knowledge

Beyond the disciplinary categories analysed so far, the topographical map helps us to dig down further into the themes that are more or less proportionately represented. Figures 2b and 2c plot the representation of the topics. Figure 2c shows that topics more commonly cited by IPCC working group I are older and largely better represented in IPCC reports. It makes intuitive sense that these topics, for example on ozone, oceans, clouds, aerosols and sea levels, are older, predominantly cited by WG I, and well cited by IPCC reports, as these are 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. They are the topics that could be seen as potential gaps in the IPCC's coverage of the science. These topics are primarily in working group III, on the mitigation of climate change, with the exception of adsorption, CLC and hydrogen which are primarily of relevance to WG III but are miscategorised due to the low number of citations of relevant documents [1]. and citations of tangentially relevant documents by other working groups

Although it is not surprising that these newer topics are less well represented than the older topics that make up

[1]For example, the word "capacity" is relevant to the adsorption topic, so documents talking about adaptive capacity receive a low relevance score for that topic. Because only very few documents highly relevant to the adsorption 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 WG II documents outweighs the sum of the topic scores of the strongly relevant WG III documents

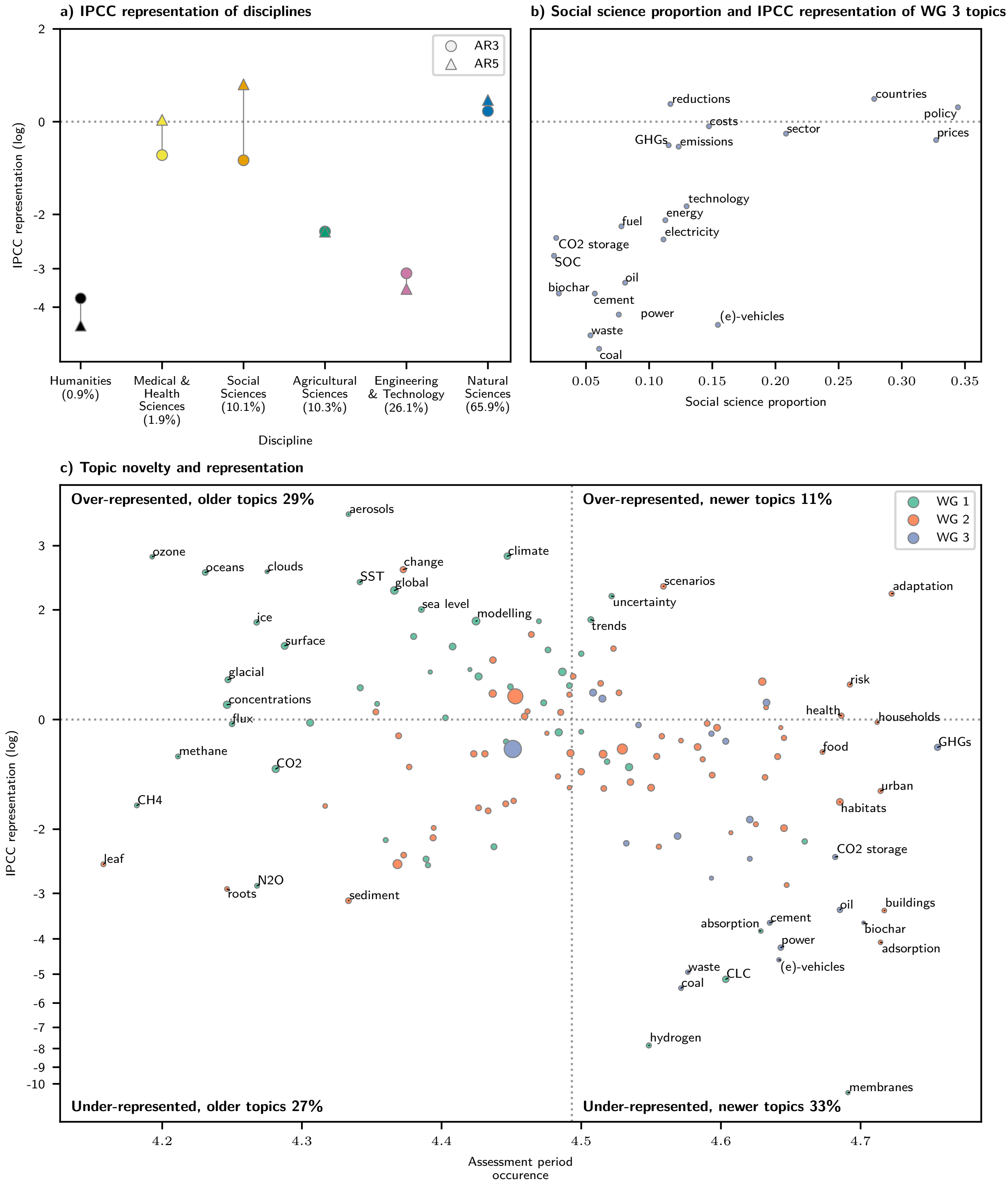

Figure 2: Representation in IPCC reports: **a)** by discipline, **b)** by social science proportion of WG 3 topics, **c)** and novelty of all topics, where topics in the highest and lowest 10% of either axis are labelled. Topics are coloured according to the working group from which they receive the 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. The log is taken so that 0 is equal to perfectly proportional representation, and -1 and 1 are equally under and over represented.

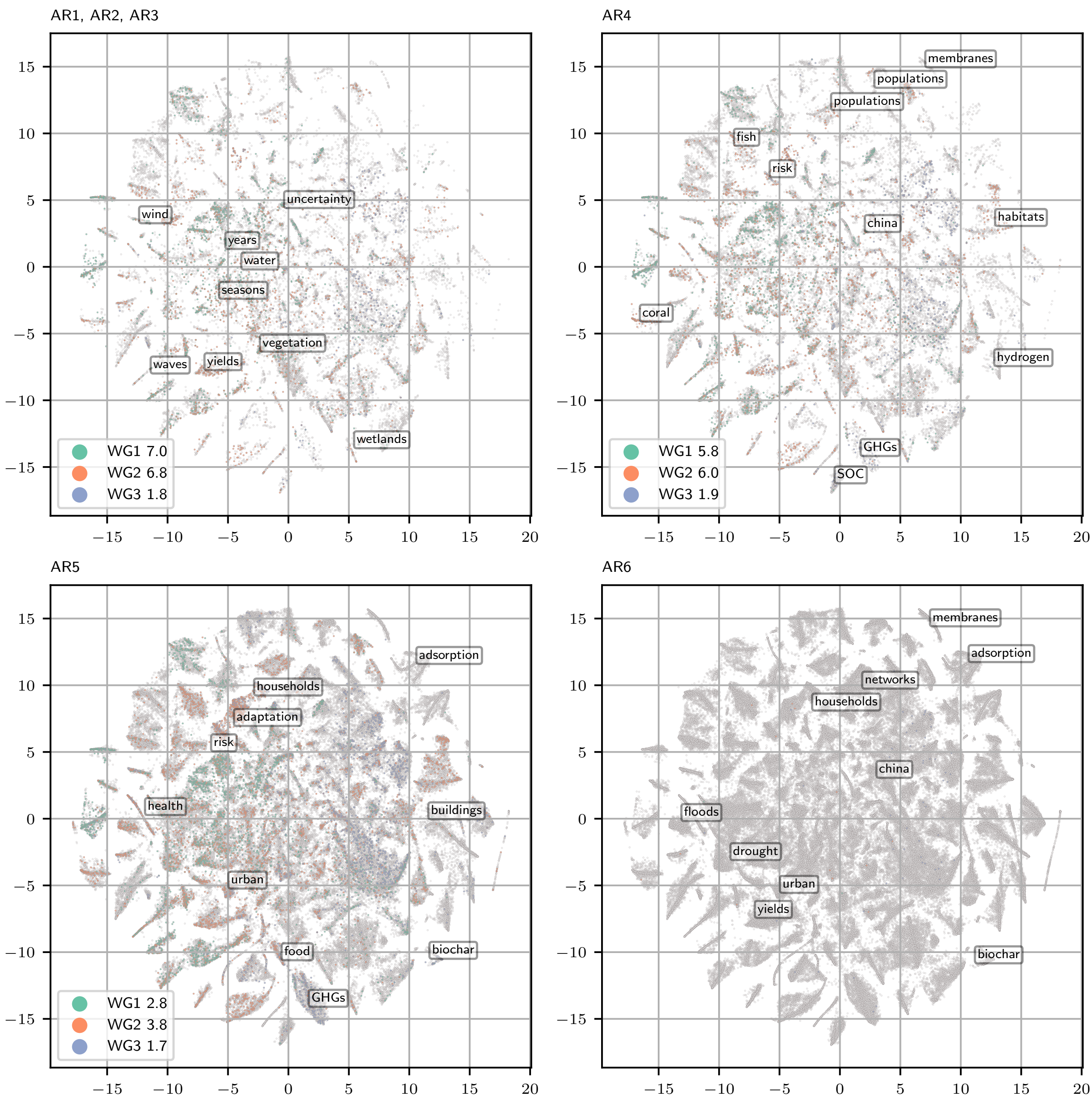

Figure 3: Evolution of the landscape of climate change literature

the core of the physical science research on climate change, the difference between these new topics and other new topics that are better represented is intriguing. This difference is visible in figure 3, where the fastest growing topics in each period are labelled, and the documents are coloured according to the working group, if any, which cites them. 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. This is also evident in figure 2c, where these topics display corresponding representation values. The IPCC, then, 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.

Within WG III topics, those that are well represented contain a greater proportion of social science research (figure 2b). Topics on 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 **??**), and complicates the perception of the under-representation of the social sciences.

Recalling policymakers' demands for more solution-oriented assessments [5], we could also interpret the topics that are newer and under-represented as "solutions-relevant". Many deal with negative emissions, or with mitigation (or to a lesser extent adaptation) options in the transport, buildings and power sectors. All are arguably related to climate solutions. However, while policymakers' demands for solutions-oriented knowledge were rather about policy options, these under-represented new topics deal with more technical solutions.

The IPCC as an informed decision-maker on topical representation

Deciding on the optimal proportion of each topic's literature to be cited by the IPCC is not a trivial task. A perfectly proportional representation of each topic is almost certainly not optimal, and as such, decisions about which topics to over-, or under- represent are best made by experts, not by a computer. Indeed, the IPCC itself is the best-placed institution to make these decisions. But in the age of "Big Literature" [15], these decision should at best be underpinned by systematic knowledge of the landscape of the literature, and made in discussion with the wider research community, funders, policymakers and other stakeholders.

The analysis in this study provide a useful input to this discussion. It could be argued that the under-represented technical literature on specific technological or sectoral solutions is not relevant to policymakers, and that working group III should give more weight to evidence about general policy instruments such as carbon taxes. Conversely, one could argue that establishing a technical understanding of specific climate solutions is as important a task as establishing a technical understanding of the physical science of climate change. The desire for more social science research in IPCC reports may imply, given the evidence here, that more social science research on climate change needs to be produced and funded, rather than that the IPCC should cite more of what already exists. Or, one could argue that, in light of the evidence that there is little social science research on solutions-relevant, under-represented topics, that the social sciences could do more to address more technical or specific solutions topics such as negative emissions, green concrete, or waste and recycling.

An understanding of the topography of the literature helps bring these arguments to the foreground, and allows them to be made in an informed way. Machine reading the literature in this way is not without its limitations though, and indeed many are apparent in this analysis. The corpus of potentially relevant literature is of course much wider than what is included in this study. The analysis excludes studies about climate change that do not match our query, studies relevant to tackling climate change that do not mention it specifically, studies written in languages other than

English (unless the abstract has been translated), studies not indexed in the Web of Science, and reports published outside peer reviewed journals, not to mention Indigenous knowledge [16]. Despite this, unless the social science literature in the above categories is systematically less often cited by the IPCC than the social science literature that was analysed in this study, or the solutions-relevant literature in the above categories is systematically more cited by the IPCC, then the two main conclusions of this study hold.

Ultimately, computers can only *assist* humans in understanding a climate change literature which is produced by human individuals and institutions. As in other areas of machine learning [17], computational assessments of the literature run the risk of reproducing biases or gaps in the training data. However, just as with other applications of machine learning, machine-reading the science of climate change can provide information that helps us to understand gaps in what we have knowledge about, and potential biases, or options for emphasis, in how we represent that knowledge in the IPCC.

References

[1] 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.

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

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

[4] 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.

[5] 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.

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

[7] Jessica Barnes, Michael Dove, Myanna Lahsen, Andrew Mathews, Pamela McElwee, Roderick McIntosh, Frances Moore, Jessica O'Reilly, Ben Orlove, Rajindra Puri, Harvey Weiss, and Karina Yager. Contribution of anthropology to the study of climate change. *Nature Climate Change*, 3(6):541–544, 2013.

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

[9] David Blei, Lawrence Carin, and David Dunson. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

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

[11] Ottmar Edenhofer and Martin Kowarsch. Cartography of pathways: A new model for environmental policy assessments. *Environmental Science and Policy*, 51:56–64, 2015.

[12] Madeleine C McKinnon, Samantha H Cheng, Ruth Garside, Yuta J Masuda, and Daniel C Miller. Sustainability: Map the evidence. *Nature*, 528(7581):185–187, 2015.

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

[14] 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.

[15] 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.

[16] 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.

[17] Solon Barocas and Andrew D Selbst. Big Data's Disparate Impact. *California Law Review*, 104:671–732, 2016.