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 28, 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 papers on climate change. This dynamic topic model provides the basis for a *topography* of climate change literature. The thematic development of the field is outlined and used to inform an analysis of the topics which are better and less well covered by IPCC reports.

We know that the scientific literature on climate change is growing rapidly [1], 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 [2]. In the light of policymakers' demand for more solutions-oriented knowledge from assessments [3]; calls for more representation of certain disciplines in IPCC reports [4, 5]; and criticism of bias in the knowledge reflected by the IPCC [6, 7], *what*, in the (dwindling) proportion of the literature that gets cited by IPCC reports, is more salient than ever.

Despite this, 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. Understanding these trends, and their reflection in IPCC reports is a crucial task if we are to assess and improve comprehensiveness in global environmental assessments.

	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

Table 1 depicts the scale of the challenge to the IPCC. In the years since the publication of 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, intended nationally determined contributions and mixed matrix membranes 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 the context 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 set of topics which can be used to describe the corpus [8], and 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. In this sense, 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 [9] (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.

To understand the map 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.

Further, the map allows us to examine topics consisting of research from across disciplines. With reference to SI figure [x] 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.

more k

Mappi
system
atic m
knowle
gaps, v
we kno
about
ipcc ar
literatu

Disciplinary bias in IPCC citations?

It was argued after the fifth assessment report that the IPCC needs to do more to incorporate knowledge from the social sciences [4]. Further, a scientometric study from 2011 claimed that the IPCC gave a greater *emphasis* to natural sciences and, within the social sciences, to economics [6]. This claim has been interpreted as a disciplinary bias by the IPCC [7, 10], 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. In each field, the under-representation has been present across assessment reports.

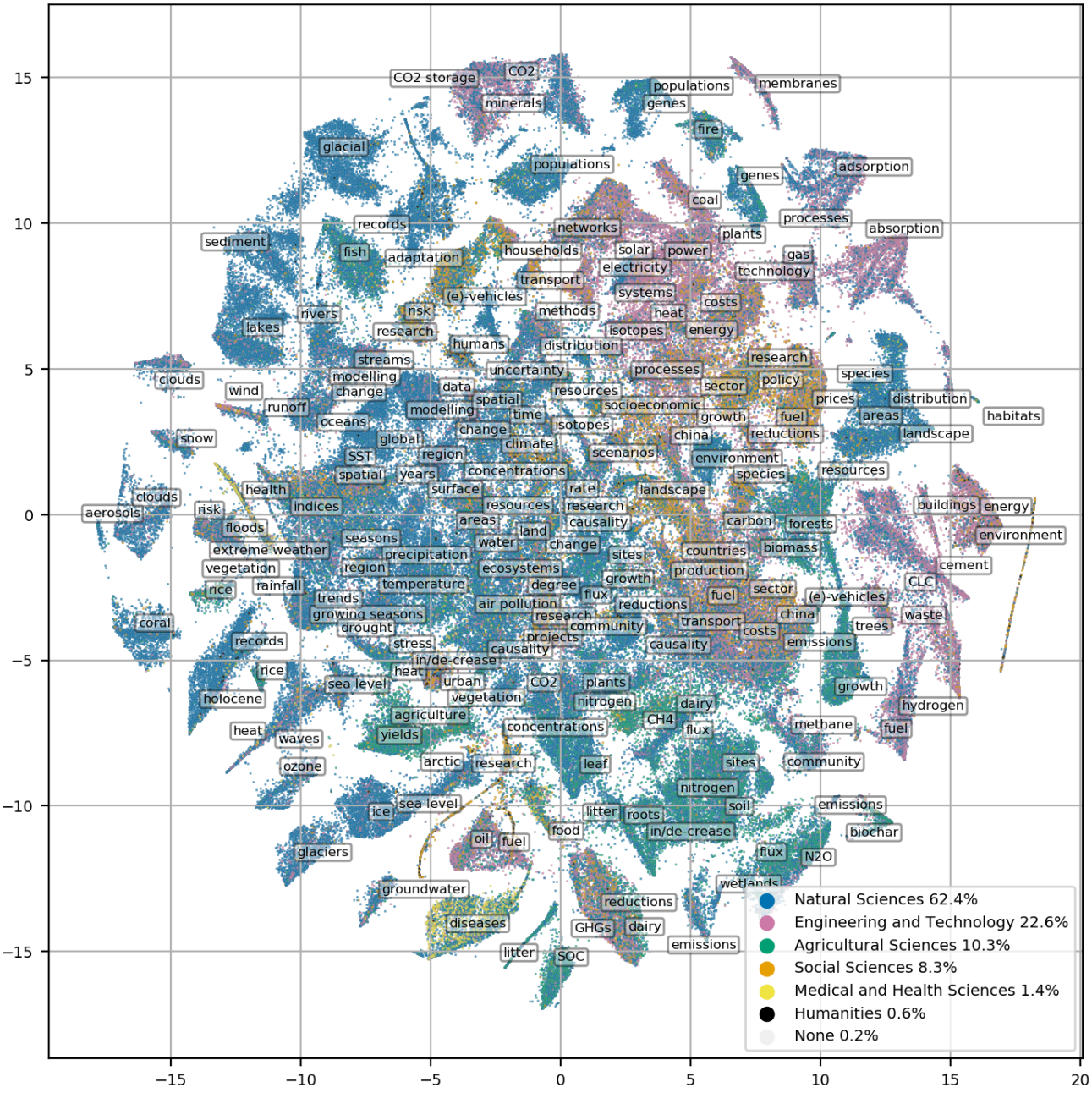

Figure 1: A map of the literature on climate change. Document positions are obtained by reducing the topic scores to two dimensions via t-SNE Documents are coloured by web of science discipline category. Topic labels are placed in the center of each of the large clusters of documents associated with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

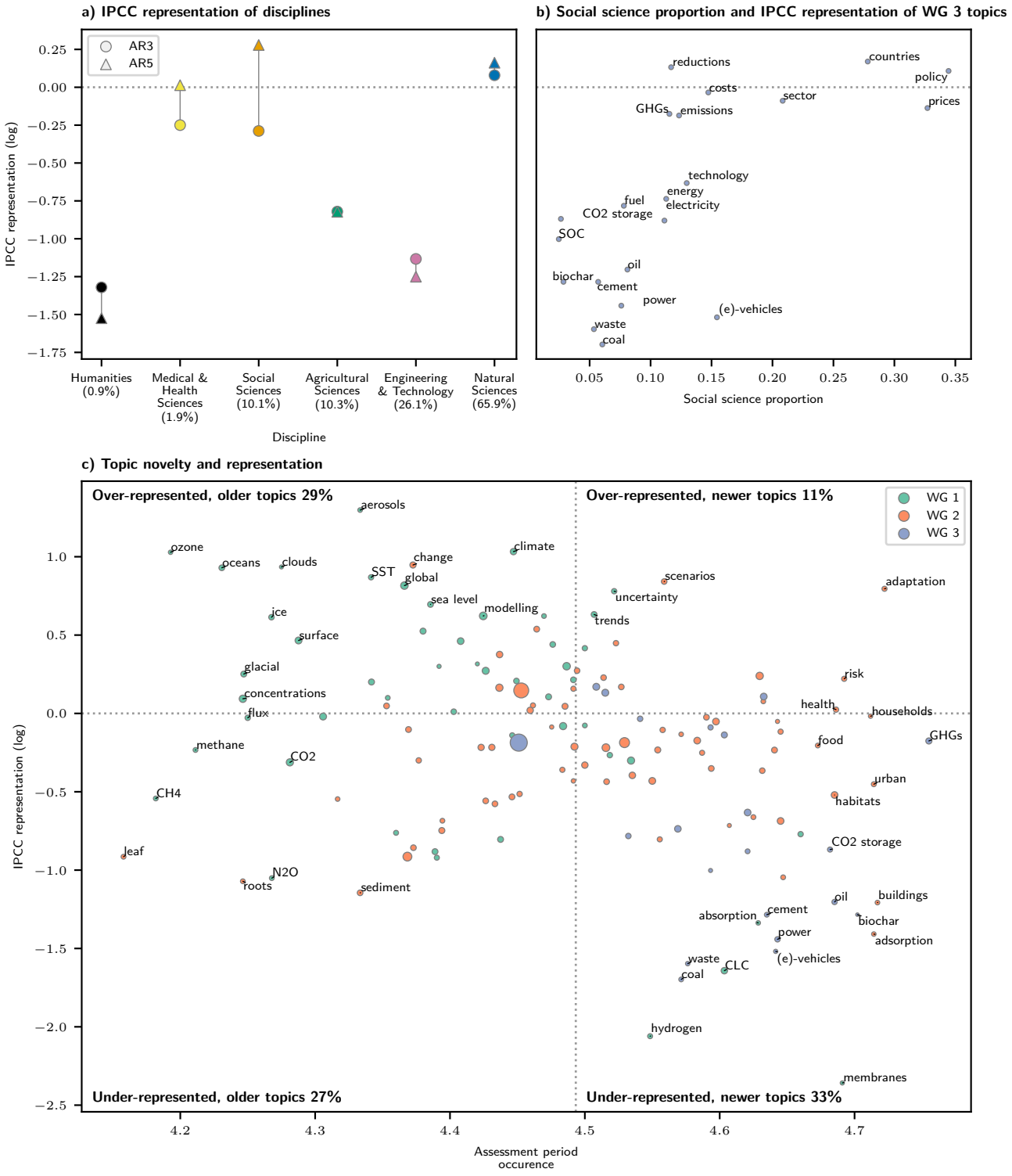

Figure 2: Representation in IPCC reports: **a)** by discipline, **b)** by social science proportion of WG 3 topics, **c)** and novelty of all topics, where topics in the highest and lowest 10% of either axis are labelled. Topics are coloured according to the working group from which they receive the 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.

A similar story is visible within the social sciences (see figure SI.3f). 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 which locate the documents on the map, showing us, in figure 2c that topics more commonly cited by IPCC working group I, are older, and are largely better represented in IPCC reports. That these topics, on ozone, oceans, clouds, aerosols and sea levels, for example, are older, cited by WG I and well cited by IPCC reports makes intuitive sense, as these are some of the core topics of the physical science of climate change.

The topics in the lower right hand part of the graph are the most pertinent to the question of whether the IPCC is well representing the existing knowledge on climate change. These topics are relatively newer and until now 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 and citations of tangentially relevant documents by other working groups [1].

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

One reason seems to be Not only about social sciences, or disciplines, but what topic+social science

Given policymakers' demands for more solution-oriented assessments, we could also interpret these topics as solutions relevant. Many deal with negative emissions, or with mitigation options in the transport, buildings and power sectors. However, while policymakers' demands for solutions-oriented knowledge was rather about policy options, these under-represented new topics deal rather with technical solutions.

[1] 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

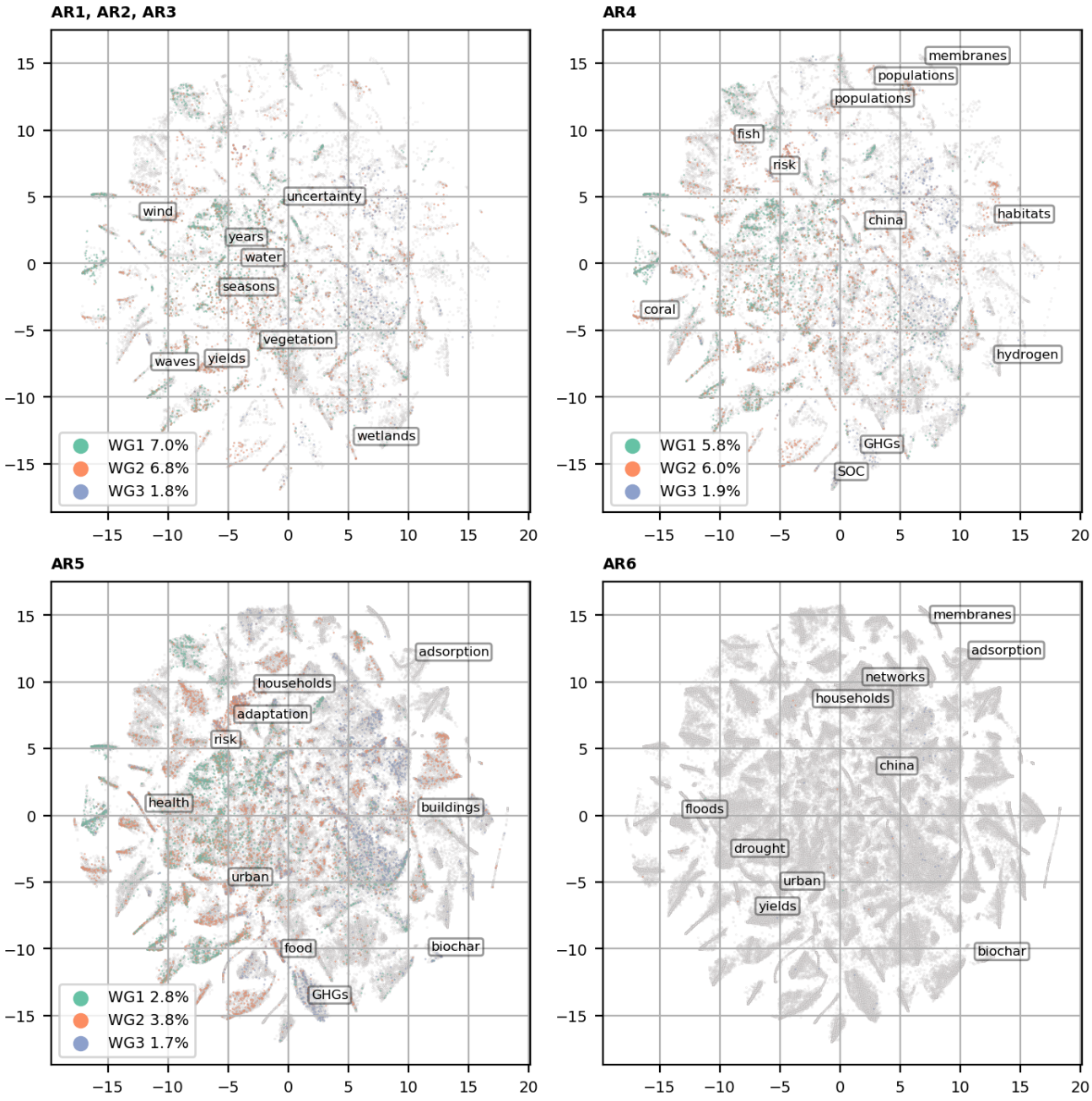

Figure 3: Evolution of the landscape of climate change literature

The IPCC as an informed decision-maker on topical representation

IPCC best to decide but informed technical vs general knowledge. Technical questions about the effectiveness of specific technologies are perhaps just as disputed as the workings of climate change

information helps to set priorities

if we feel we need more social science knowledge, the information here tells us that we may need to change the social sciences

once again, this is to be answered best by the IPCC,

limitation, not all literature, doesn't fundamentally change conclusions, means the challenges are even larger than portrayed here. More potentially relevant literature, other potentially important and under-cited topics. Other types of knowledge.

References

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

[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] 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.

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

[5] 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.

[6] 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.

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

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

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

[10] 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.

[11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Mattheiu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python Fabian. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A Topography of Climate Change Research - Methods

Max Callaghan[1,2] and 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
[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 28, 2019

Data

This study reproduces the query developed by [1], which is carried out on the Web of Science core collection. Though not exhaustive, it gives a good coverage of the literature in major peer-reviewed journals. Each document is assigned to an assessment period according to the timeline shown in table 1.

We use the references scraped from IPCC assessment reports from [2], and attempt to match these with the results from the web of science. Table [x] shows the percentage of IPCC citations matched in each working group for each assessment report.

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 [], titles are substituted for abstracts where they are not available. The words of the documents are lemmatized/stemmed, replacing different forms of the same word (i.e. word/words) with a single instance. Commonly occuring words, or "stopwords" are removed, as are all words shorter than 3 characters, and all words containing only punctuation or numbers.

For each period, the documents are transformed into a document-term matrix, 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 occurence in the corpus. This gives more weight to terms which appear in few documents, and less weight to those which appear in many.

$$tf(t,d) = f_{t,d}, \quad idf(t,D) = \log \frac{N}{|\{d \in D : t \in d|\}} \tag{1}$$

Topic Model

We use non-negative Matrix Factorisation (NMF), 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:

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu} \tag{2}$$

As demonstrated in Figure SI.2, each topic is represented as a set of word scores, and each document a set of topic scores. The combination of the two give 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 [11]

Model selection

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.

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

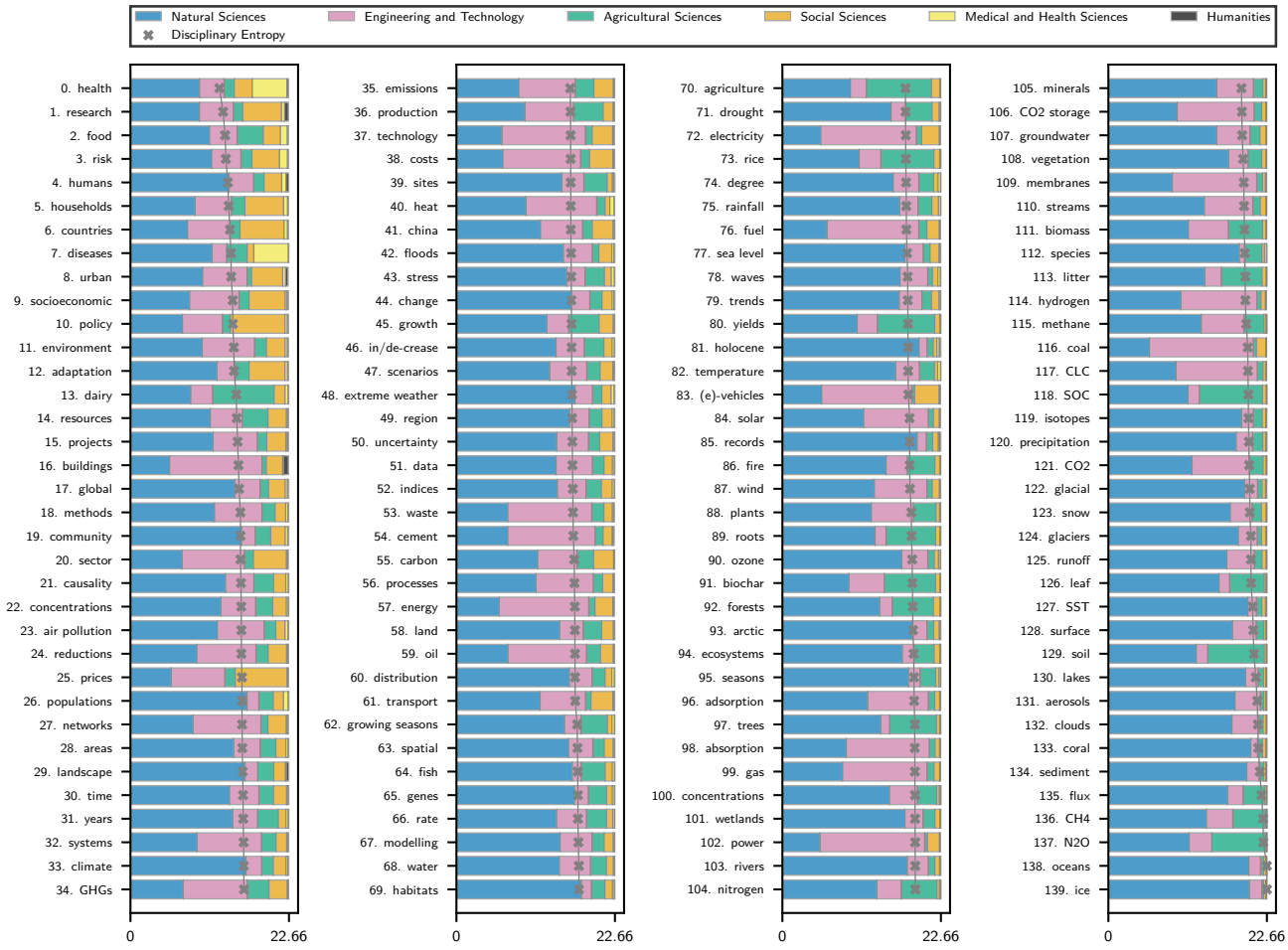

Figure SI.1: SI Disciplinary Entropy

Topic Doc	ecosystem net productivity	community microbial composition	uncertainty estimate parameter	trend station significant
doc_1	0.022	0.017	0.011	0.009
doc_2	...	...	...	...
doc_3	...	...	...	...

×

Term Topic	ecological	ecosystem	recent	community
ecosystem net productivity	1.08	9.18	0	0
community microbial composition	0.19	0	0	9.76
uncertainty estimate parameter	0	0	0.01	0

≈

Term Doc	ecological	ecosystem	recent	community
doc_1	3	2	2	2
doc_2	...	...	...	...
doc_3	...	...	...	...
doc_4	...	...	...	...

Figure SI.2: SI Topic make up of a single document

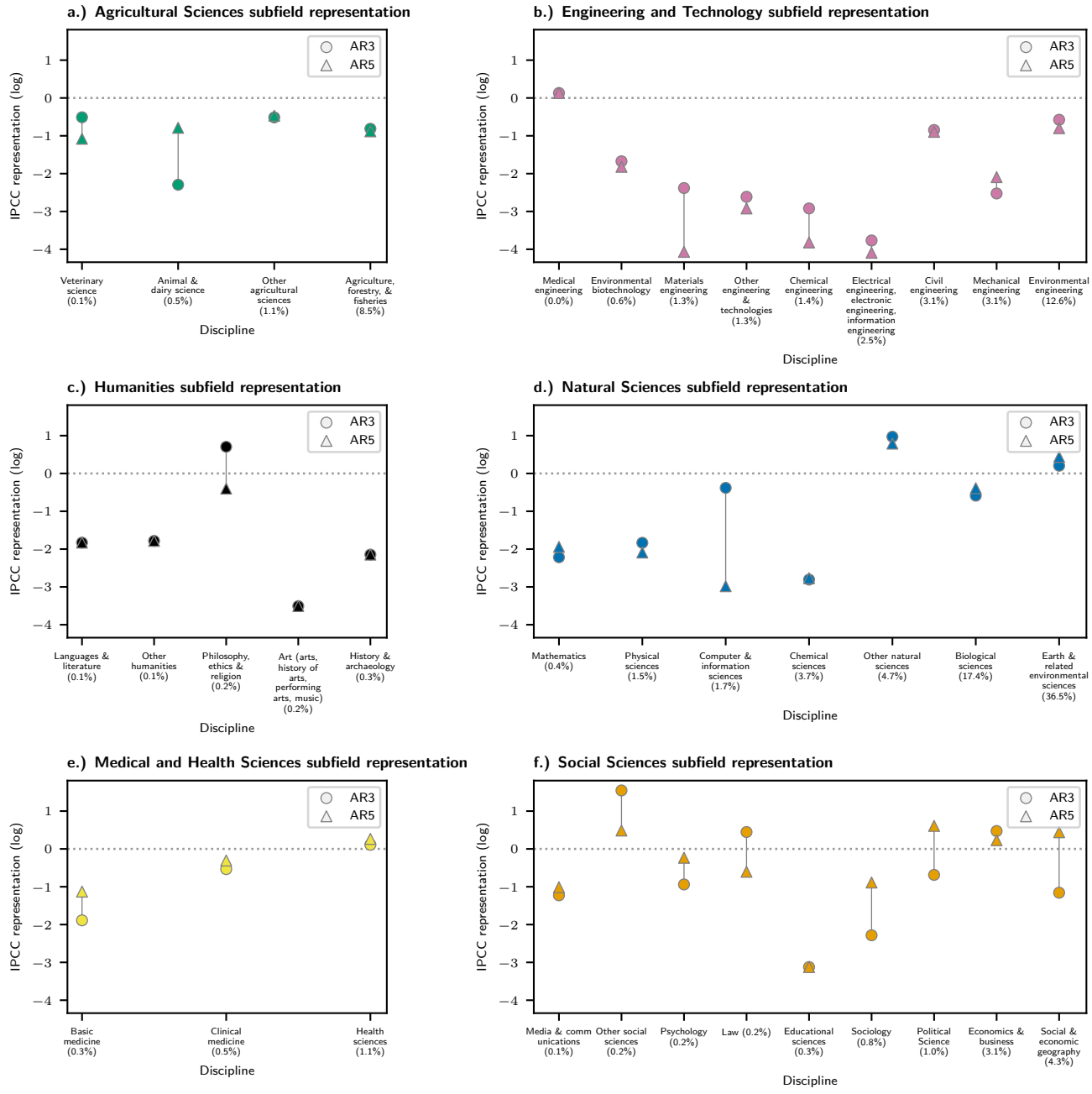

Figure SI.3: SI Representation by subfield

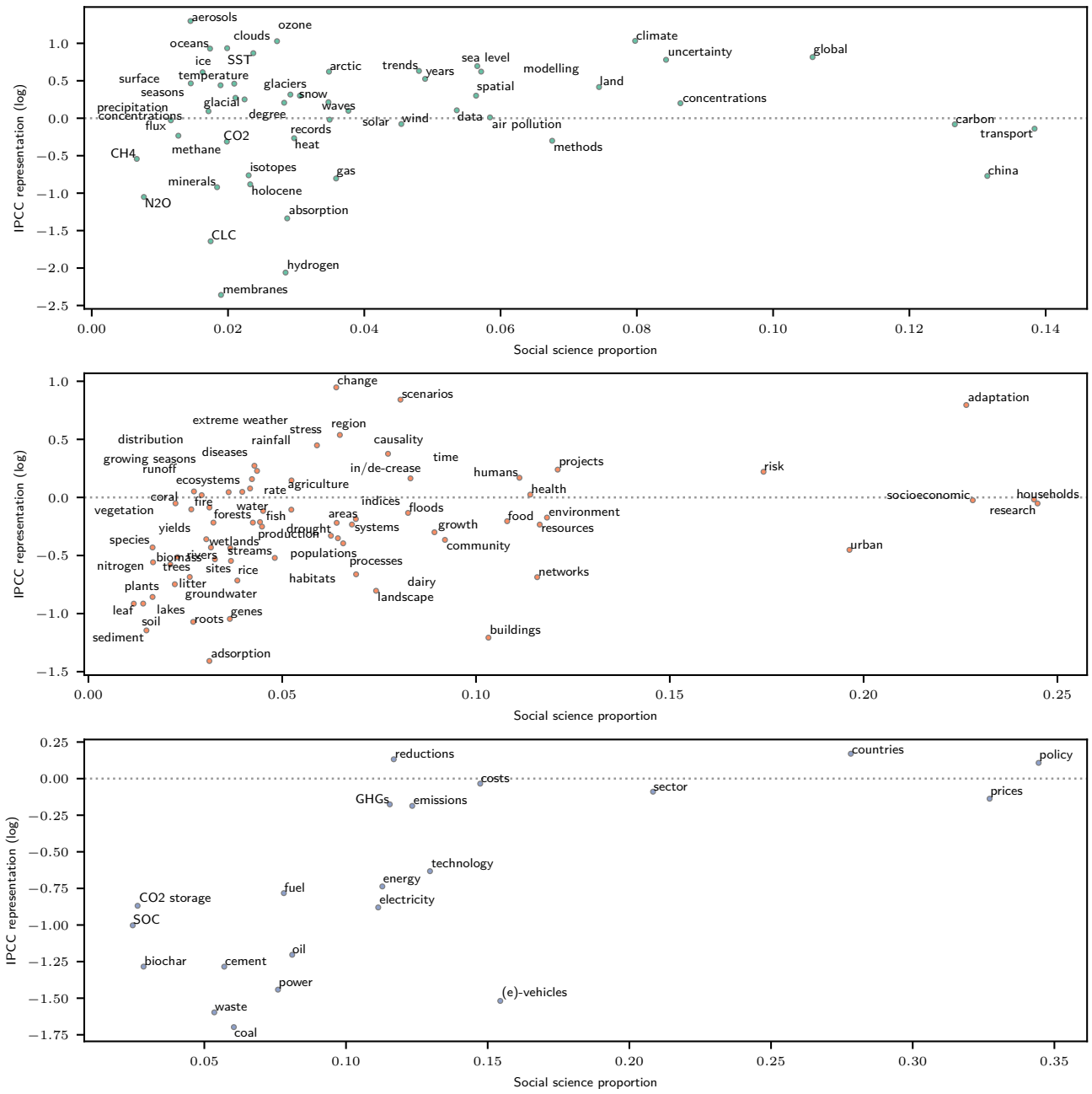

Figure SI.4: SI Social science & representation in topics across working groups

References

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

[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] 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.

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

[5] 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.

[6] 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.

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

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

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

[10] 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.

[11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Mattheiu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python Fabian. *Journal of Machine Learning Research*, 12:2825–2830, 2011.