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 March 4, 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.

The IPCC sees its role as to "assess on a comprehensive, objective, open and transparent basis the scientific, technical and socio-economic information relevant to [...] climate change" [1]. Climate science is so broad, multi-disciplinary, and laden with uncertainties and values, that the role of the IPCC as assessment maker is vitally important to developing evidence-based international climate policy. In the Pragmatic Enlightened Model of the science-policy interface, the task of assessment making is a cartographic one [2]. Assessment makers map out the problem and the option space for policy-makers.

Further, it has been pointed out that, in the age of "big literature", providing assessments that are comprehensive, objective and transparent has become much more difficult [3]. When the IPCC's citations constitute an ever-decreasing proportion of the totality of science on climate change, questions about the map that the IPCC reports produce become more pressing:

- Is the map up to date?
- Is the map complete?
- Is the map's projection representative? Does it emphasise or obscure certain areas?

The IPCC as an object of scientific investigation

Various researchers have attempted to do empirical research on the assessment reports, and processes of inter. alia. the IPCC [4, 5, 6, 7]. Among the clear conclusions is that policy makers, when asked about their interactions with the IPCC call for a greater focus on solutions [8]. Studies that look directly at the content of IPCC reports, though, are similarly challenged by the the size of the literature that the reports assess. Without engaging with this literature, the conclusions drawn lack a reference point, and the phenomena identified (such as an over-representation of the natural sciences in IPCC citations [5]) are hard to disentangle as specific to the way the IPCC assesses its literature, rather than a feature of the literature itself.

Some studies have attempted to identify and characterise the literature on climate change through bibliometric techniques [9, 10]. Such studies are the jumping off point for this work, and form the basis of the identification strategy for the literature, but they do not engage with the *texts* of the work identified, nor do they make the link to the IPCC. A growing body of scientific work deals with the unsupervised analysis of texts using topic modelling [11, 12, 13, 14]. Applications to the field of climate change though are rare and often limited to closer analysis within sub-topics [15], or calls for their greater use [16].

A problem of scale

The case for a greater application of text-mining approaches in understanding and assisting the IPCC is made by the scale of the challenge it faces, depicted in table 1. Less than a thousand documents relevant to climate change were published before the first assessment report (see Methods for data, exclusions and processing). The abstracts of these documents contained 1,380 unique terms. In the three complete years since the publication of AR5, 128,266 documents have been published, containing 74,196 unique terms. The new words in each assessment period unearthed by a basic machine-reading of the literature already give an indication of the how the field has expanded. They chart the emergence of

	AR1	AR2	AR3	AR4	AR5	AR6
Documents	625	7623	16395	34510	117758	128266
Words	1380	12409	20453	32644	67064	74196
New words	change (296)	loss (552)	downscaling (197)	sres (217)	biochar (1752)	mmms (192)
	climate (262)	efficiency (515)	degreesc (145)	petm (95)	redd (1058)	c3n4 (132)
	model (168)	mol (439)	ncep (130)	amf (87)	cmip5 (656)	cop21 (107)
	effect (160)	ambient (417)	otcs (87)	sf5cf3 (81)	cmip3 (569)	cmip6 (104)
	co2 (156)	coal (404)	nee (87)	cwd (74)	wrf (334)	zika (75)
	atmospheric (152)	photosynthetic (393)	fco (80)	embankment (72)	mofs (288)	brgdgts (71)
	climatic (133)	concern (381)	hadcm2 (78)	aod (69)	sdm (283)	twitter (68)
	global (131)	chamber (353)	dtr (75)	clc (69)	gosat (281)	jing (66)

Table 1: Growth in climate change literature

- new compounds or materials, such as trifluoromethyl sulfur pentafluoride (SF5CF3) in AR4, or mixed matrix membranes (MMMs) in AR6
- new models, or modelling approaches, such as HadCM2 in AR3, or CMIP in AR5 and AR6.
- the involvement of new actors, such as NCEP (National Centers for Environmental Prediction) in AR3
- new approaches, such as biochar and REDD in AR5
- new applications of equipment, such as OTCs (Open-top Chambers) in AR3;
- new phenomena in relation to climate change, such as zika and twitter;
- and new concepts, such as CO2 fugacity (fco) in AR3

To put this growth into context, the 236,634 publications published in AR5 and AR6 are significantly larger than the 178,118 publications recorded in the first volume of the 'Catalogue of Scientific Papers', compiled by the Royal Society to record the entirety of scientific output from 1800 to 1863 [17]. In the last 10 years, climate science has produced significantly more papers than the were published across all scientific disciplines in the first half of the 19th Century.

Clearly, if the IPCC is to continue producing comprehensive assessments, it has to engage in machine-reading in order to remain anchored to the wider literature. Without such an approach, it becomes harder to justify which ever-diminishing proportion of the wider literature is included in assessments. The same arguement holds for scholars and critics of the IPCC. Without machine reading the literature at large, it becomes harder to analyse or criticise, with quantitatively evidenced claims, the outcomes of assessment processes.

In order to draw substantive messages from this high-dimensional document-word space, we apply a form of dimensionality reduction called topic-modelling [18]. The non-negative matrix factorisation (NMF) approach to topic modelling is based on generating a "perception of the whole based on perception of its parts" [11]. Here the corpus of documents (whole) is summarised by its parts (topics), which appear in varying proportions in each of the documents. Importantly, in such approaches, the parts are not pre-defined, but learned by the algorithm, leveraging systematic co-occurences of words within documents. This makes the discovery of unlooked-for features possible and can reduce the human bias risked by the pre-definition of categories.

Despite these advantages, machine reading and machine learning approaches can of course not replace the task of human assessment-making. The contribution that could be made, though, is to pre-process the literature, producing a topographical map, used to navigate the literature while producing a more detailed assessment with human judgement. In fact this happens already - when IPCC authors search for literature on a topic, the results which appear on the search engine they use will be subject to algorithms based on the processing of millions of records of article text and metadata. This can be done in a much more systematic way when scientists perform directed analyses of the literature at scale.

In applying a dynamic version of NMF [13], this study demonstrates how topic modelling can be used to gain an overview of an otherwise unmanageably large body of literature. This overview, or topography, describes the thematic development of the climate change literature and, in a novelly systematic way, examines how comprehensively the IPCC has been able to engage with it. In pulling together strands from text-mining, bibliometrics, and the study of science and policy, this study advances our understanding of the literature on climate change and the role of the IPCC in communicating this to policy makers.

Results

Table 2 shows the 10 most prominent topics within the literature on climate change. The topics are characterised by the (stemmed) words that define them, and by the documents which are most well described by them. For example, the topic "energi, renew, consumpt" features the words energi, renew, consumpt, effici, demand, save, sector, sourc, industri, use, and best describes the documents "Energy issues and energy priorities" [19] and "Energy efficiency and CO2 emissions in Swedish manufacturing industries" [20]. The topics learned by dynamic NMF identify a set of interpretable topics and can identify which documents belong to these topics.

Such first results show the promise of topic modelling in the identification of studies in a thematic area, which is of high value to assessment exercises like the IPCC, but in themselves do not give a sense of the broad structure of the corpus. Figure 1 projects the documents on to a two dimensional representation of the their topic composition derived through t-distributed Stochastic Network Embedding (t-SNE) [21]. In this approach, documents are positioned close to other documents with similar topic compositions. This projection enables us to show a topography of the whole literature on climate change.

In the two panels of the figure, the documents are coloured by the IPCC working group which cites them (if any), and the disciplinary classification according to the Web of Science. The topic structure cuts across working group and disciplinary lines, but the disciplinary and working group structure is clearly visible in the topography. For example, in the south-eastern section of the map, we can see a concentration of documents in the social sciences that were cited by working group III (mitigation).

The topography_map_reference.csv [`https://github.com/mcallaghan/cc-topography/blob/master/tables/tsne_topic_index_665.csv`] file in the supporting materials aids more detailed analysis of the map, describing the topic composition of each square. For example the square whose center is [12.5,-7.5] is made up of the topics {energi, renew, consumpt}, {fuel, fossil, engin}, {emiss, reduct, reduc}, {countri, develop, trade} and {cost, optim, price}, typical social science, working group III topics. The square whose center is [-17.5,-2.5], containing mostly working group I (the physical science basis) documents on the other hand, is made of largely by the topic ozon, stratospher, tropospher as well as by increas, concentr, decreas, and climat, chang, impact.

Figure 2 shows the evolution of the landscape of climate change literature over the 6 (including the in progress AR6) assessment periods of the IPCC. In each assessment period, the grid squares which contain a higher (lower) proportion of documents than in the previous assessment period are coloured green (red). This mapping allows us to identify topic areas as they appear in the literature. For example, the square with the center [-12.5,12.5] is composed largely of the topic {coral, reef, bleach}, and contained no documents before a sudden growth in AR3. Indeed, the topic of coral bleaching was little discussed in the context of climate change before AR3, with just two out of 693

title	top words	top docs	share
climat, chang, impact	[climat, chang, impact, respons, futur, effect, shift, sensit, affect, may]	Climate oscillations and changes over Russia; World Regionalization of Climate Change (1961-2010)	2.73%
soil, moistur, microbi	[soil, moistur, microbi, organ, respir, content, miner, depth, matter, efflux]	PARTITIONING OF SOIL RESPIRATION IN A FIRST ROTATION BEECH PLANTATION; Responses of soil respiration to N fertilization in a loamy soil under maize cultivation	2.73%
emiss, reduct, reduc	[emiss, reduct, reduc, greenhous, factor, total, estim, inventori, nox, measur]	China's CH4 and CO2 emissions: Bottom-up estimation and comparative analysis; Monitoring total emissions from industrial installations	2.21%
carbon, dioxid, sequestr	[carbon, dioxid, sequestr, sink, organ, cycl, storag, stock, terrestri, atmospher]	Interpreting carbon-isotope excursions: carbonates and organic matter; PARTICULATE FLUXES OF CARBONATE AND ORGANIC-CARBON IN THE OCEAN - IS THE MARINE BIOLOGICAL-ACTIVITY WORKING AS A SINK OF THE ATMOSPHERIC CARBON	1.74%
temperatur, air, mean	[temperatur, air, mean, surfac, minimum, maximum, daili, increas, effect, degreesc]	Observed changes in shallow soil temperatures in Northeast China, 1960-2007; Beyond the Mean: Biological Impacts of Cryptic Temperature Change	1.71%
record, dure, glacial	[record, dure, glacial, reconstruct, last, period, holocen, event, late, core]	HIGH-RESOLUTION CLIMATE RECORDS FROM THE NORTH-ATLANTIC DURING THE LAST INTERGLACIAL; HIGH-RESOLUTION CLIMATIC INFORMATION FROM SHORT FIRN CORES, WESTERN DRONNING MAUD LAND, ANTARCTICA	1.7%
speci, distribut, rang	[speci, distribut, rang, rich, invas, nich, predict, extinct, shift, abund]	Northward range extensions of some mesopelagic fishes in the Northeastern Atlantic; Natural occurrence and backwater infection of C-4 plants in the vegetation of the Yangtze hydropower Three Gorges Project region	1.7%
increas, concentr, decreas	[increas, concentr, decreas, effect, atmospher, doc, result, organ, nutrient, may]	TERRESTRIAL HIGHER-PLANT RESPONSE TO INCREASING ATMOSPHERIC [CO2] IN RELATION TO THE GLOBAL CARBON-CYCLE; Hydrological response to climate change in the Black Hills of South Dakota, USA	1.61%
forest, tropic, stand	[forest, tropic, stand, deforest, disturb, stock, boreal, redd, harvest, wood]	Spatially explicit estimates and temporal changes of forest tree biomass in a typical department of forest management, Turkey; Analysis of the changes in forest ecosystem functions, structure and composition in the Black Sea region of Turkey	1.56%
energi, renew, consumpt	[energi, renew, consumpt, effici, demand, save, sector, sourc, industri, use]	Energy issues and energy priorities; Energy efficiency and CO2 emissions in Swedish manufacturing industries	1.56%

Table 2: Top 10 topics in climate change literature

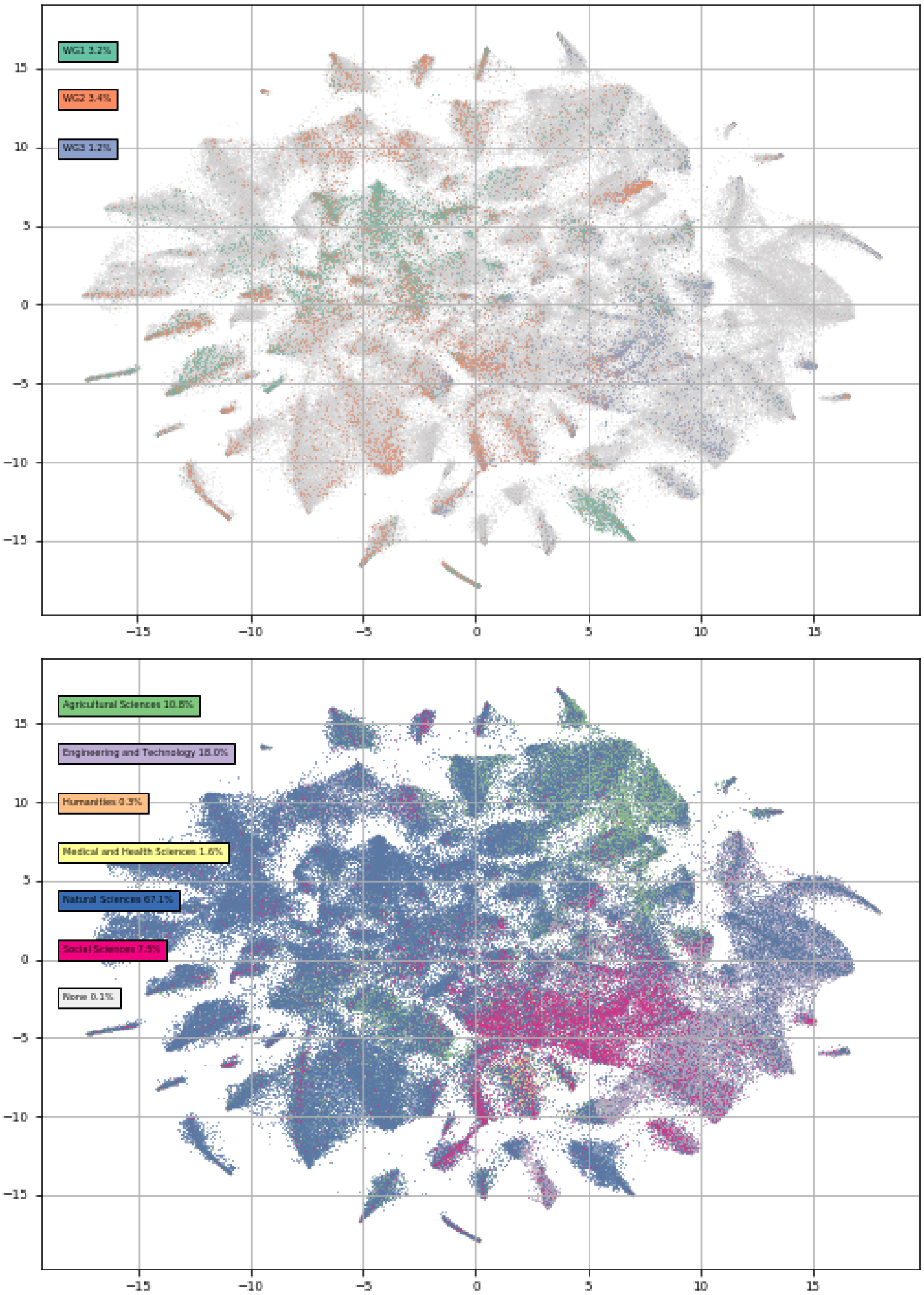


Figure 1: A map of the literature on climate change. Document positions are obtained by reducing the topic scores to two dimensions via t-SNE Documents are coloured by working group citations (top) and web of science discipline category (bottom). See SI table for topic composition of each grid square

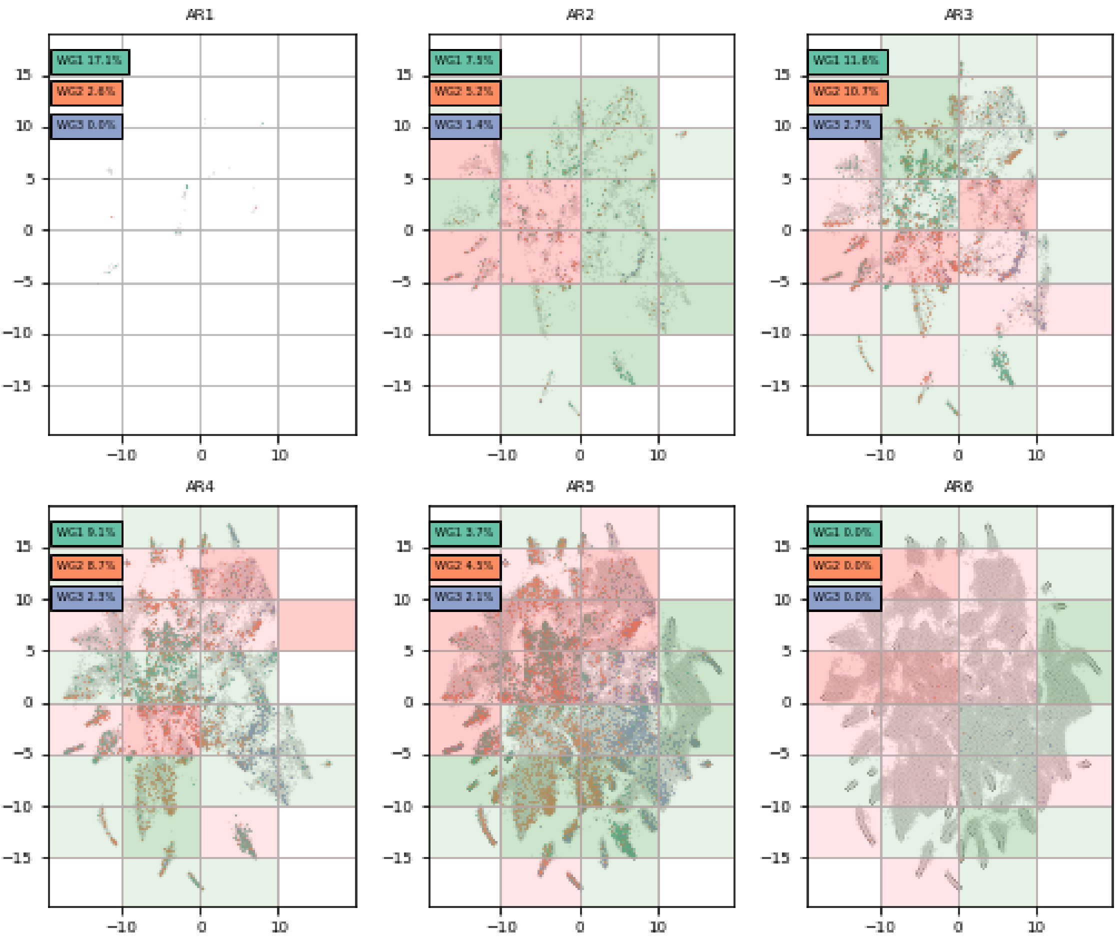


Figure 2: The evolution of the landscape of climate change literature

results from Web of Science for the terms "Coral bleaching" AND "climate change" published before 1996. More recently, the square [12.5,12.5] identifies the topic {biochar, amend, applic} as a quickly emerging field within climate change in present times.

The dynamic nature of the topic model also allows us to identify changes within topics. SI figure 5 shows the evolution of the topic {research, social, issu}. In this overarching topic on research priorities we can detect an increasing prominence of the word "sustainability".

Finally, by comparing the topic distribution across all documents (published before the end of AR5) with the topic distribution across documents that were cited by the IPCC, we can identify those topics which were over- or under- represented, with respect to the underlying literature, in

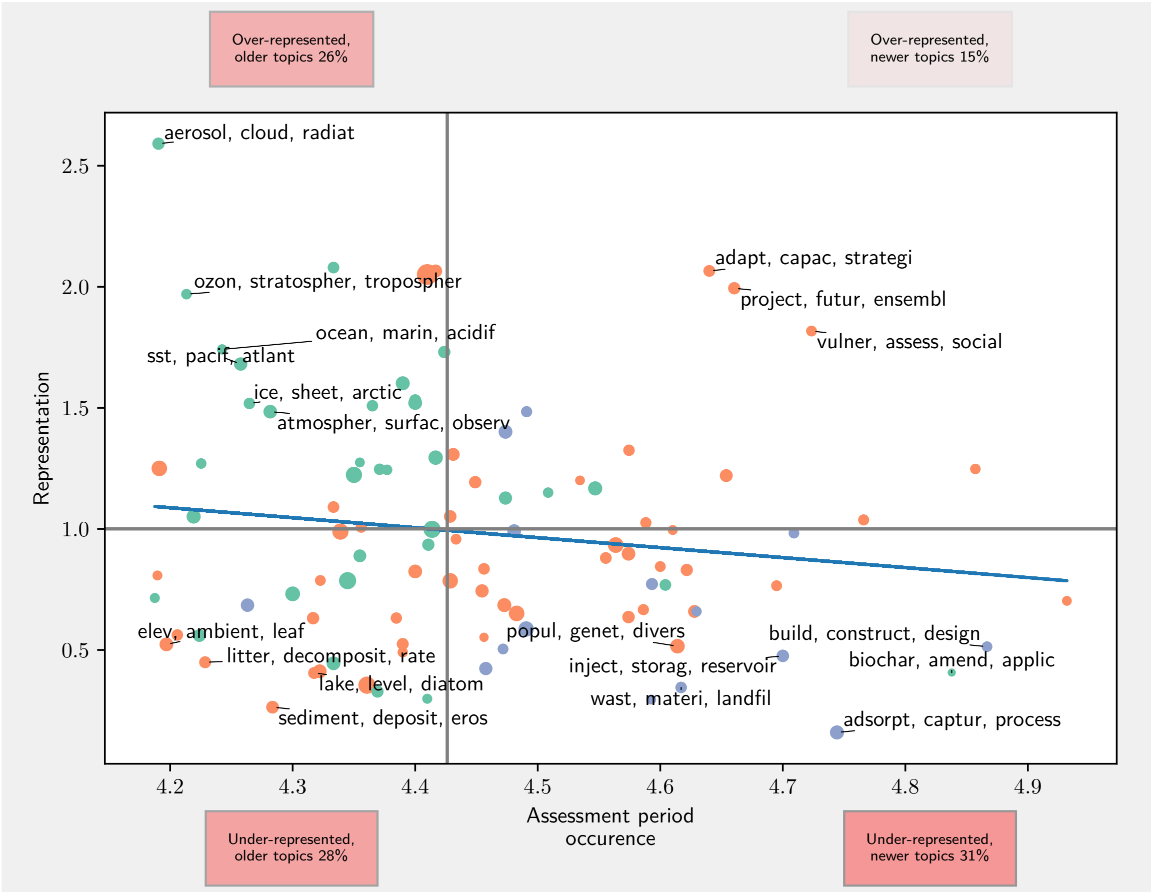


Figure 3: Representation and newness of dynamic topics

IPCC reports. Figure 3 plots the representation and novelty of all topics in the model, labelling those topics most under- or over-represented. The topics are coloured according to the working group from which most IPCC citations to its documents came.

The negative slope of the plot shows that the IPCC better represents those topics which have a longer history in the literature. The majority of well represented older topics come from working group I, such as {aerosol, cloud, radiat}. New topics in working group II on adaptation and vulnerability are well represented in IPCC reports. However, a cluster of new topics on mitigation issues appear to be under-represented by the IPCC. Moreover, these topics, such as those on buildings, waste, biochar, carbon capture and carbon storage are those that deal with climate solutions.

Discussion

The results from a dynamic topic model of climate change literature delineate the structure and evolution of the field, telling its history and identifying areas of rapid growth that are of high relevance for future assessments. Further, they suggest that the demand by policy-makers from the IPCC for a greater focus on solutions is well-founded. Topics related to climate solutions are the least well represented in the IPCC's assessments.

There may well be excellent reasons for over- or under-representation of topics in IPCC reports, and indeed it is the institution of the IPCC, not machines, who can best make judgements about what to include and what not to include. However, it is unclear whether the decisions which resulted in such unequal distributions were made consciously, on the basis of an understanding of the breadth of the extant literature. For future assessments to be relevant, transparent and comprehensive in the age of big literature, these decisions should be supported by machine-supported engagement with the literature as a whole.

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 March 4, 2019

Data

This study reproduces the query developed by [10], 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 3.

We use the references scraped from IPCC assessment reports from [3], 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.

name	years
AR1	1988-1989
AR2	1990-1994
AR3	1996-2000
AR4	2001-2006
AR5	2007-2013
AR6	2014-

Table 3: Assessment period time windows

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.

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

Dynamic Non-negative Matrix Factorisation

Non-negative Matrix Factorisation (NMF) is 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 4, 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.

In Dynamic Non-negative Matrix Factorisation, proposed by Greene [13], a separate topic model is run for each period. These are then joined through another topic model, which takes the topic-term matrices of the all periods as V, and produces dynamic topics, which describe the window topics according to the words which occur in them. Similar topics across and within time periods are thereby grouped together.

While Greene uses an automatic approach to deciding on topic numbers within time periods, we found the number of topics derived from topic coherence scores (as used by Greene) to be noisy, and instead opt for identifying (subjectively) an optimal number of topics for each window. We do this by comparing topic lists with increasing numbers of topics, where similar topics are automatically placed next to each other. We similarly compare different numbers of Dynamic topics.

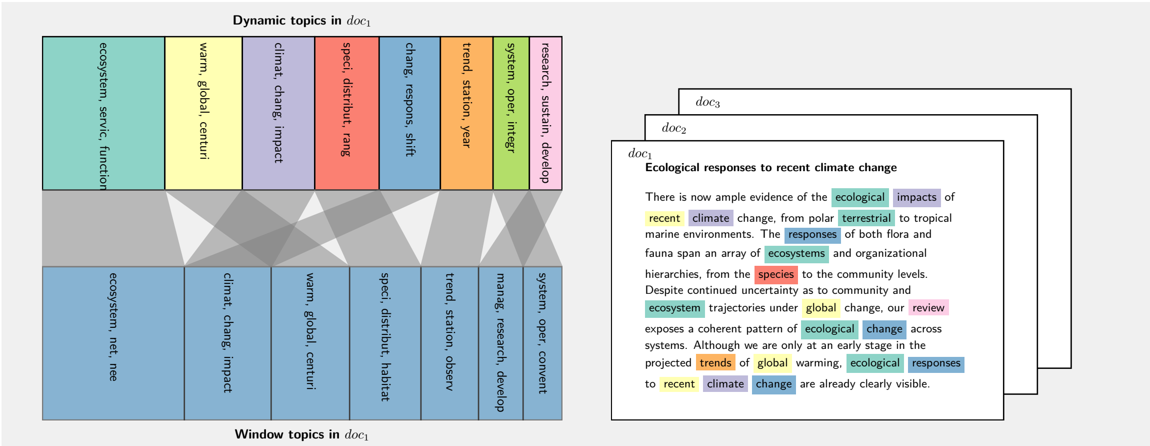
We settle on [x,x,x,x,x] topics for ARs 1 to 6, and [x] dynamic topics.

Topics are calculated using the scikitlearn library [22]

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 dynamic topic's total score as the sum of document-dynamic topic scores (which in turn are made of the product of all document-window topic and window topic-dynamic topic scores). A dynamic topic's window score is the sum of document-dynamic scores considering only documents in the given time window. To represent a dynamic 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.



Topic / Doc	ecosystem net nee	climat chang impact	warm global centuri	speci distribut habitat
doc_1	0.034	0.021	0.019	0.017
doc_2	...	...	...	...
doc_3	...	...	...	...

Doc Topic Matrix

×

Term / Topic	chang	ecolog	global	recent
ecosystem net nee	0.18	0.39	0.01	0.03
climat chang impact	6.33	0	0.22	0.19
warm global centuri	0	0	3.72	0.31

Topic Term Matrix

≈

Term / Doc	chang	ecolog	global	recent
doc_1	4	3	2	2
doc_2	...	...	...	...
doc_3	...	...	...	...
doc_4	...	...	...	...

Doc Term Matrix

Figure 4: SI Topic make up of a single document

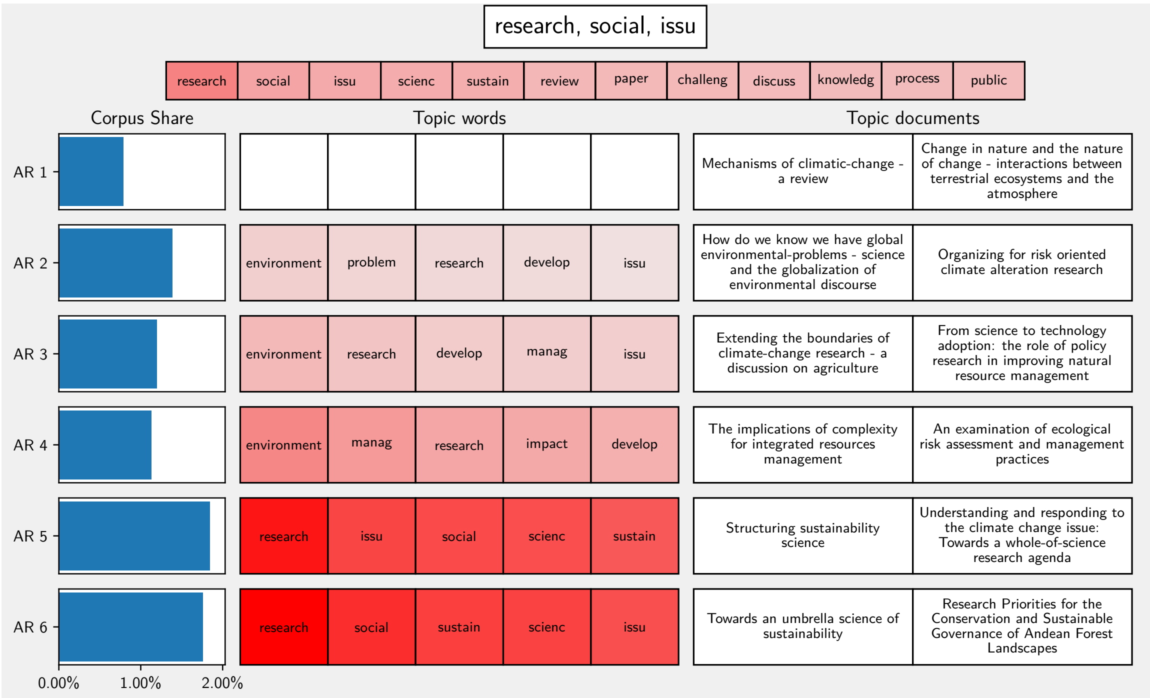


Figure 5: Word and document development of the “Research” dynamic topic

References

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

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

[3] 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.

[4] Jason Jabbour and Christian Flachsland. 40 years of global environmental assessments: A retrospective analysis. *Environmental Science and Policy*, 77(May):193–202, 2017.

[5] 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.

[6] 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.

[7] Mike Hulme. 1.5°C and climate research after the Paris Agreement. *Nature Clim. Change*, 6:222–224, 2016.

[8] 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.

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

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

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

[12] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[13] Derek Greene and James P Cross. Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. pages 1–47, 2016.

[14] Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, and Eduardo M. Airoldi. The structural topic model and applied social science. *NIPS 2013 Workshop on Topic Models*, pages 2–5, 2013.

[15] Emily Grubert. Implicit prioritization in life cycle assessment: text mining and detecting metapatterns in the literature. *International Journal of Life Cycle Assessment*, 22(2):148–158, 2017.

[16] Emily Grubert and Anne Siders. Benefits and applications of interdisciplinary digital tools for environmental meta-reviews and analyses. *Environmental Research Letters*, 11(9):093001, 2016.

[17] Alex Csiszar. How lives became lists and scientific papers became data: Cataloguing authorship during the nineteenth century. *British Journal for the History of Science*, 50(1):23–60, 2017.

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

[19] A. Demirbas. Energy issues and energy priorities. *Energy Sources, Part B: Economics, Planning and Policy*, 3(1):41–49, 2008.

[20] Clara Inés Pardo Martínez and Semida Silveira. Energy efficiency and CO2emissions in Swedish manufacturing industries. *Energy Efficiency*, 6(1):117–133, 2013.

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

[22] 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.