

# A Topography of Climate Change Research - Methods

Max Callaghan<sup>1,2</sup>

<sup>1</sup>Mercator Research Institute on Global Commons and Climate Change, Torgauer Straße, 10829 Berlin, Germany

<sup>2</sup>School of Earth and Environment, University of Leeds, Leeds LS2 9JT, United Kingdom

Draft current May 31, 2019

## 1 Data

This study reproduces the query developed by [1], which is carried out on the Web of Science core collection. Though not exhaustive, the Web of Science 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.

## 9 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 occurring words, or “stopwords” are removed, as are all words shorter than 3 characters, and all words containing only punctuation or numbers.

The documents are transformed into a document-term matrix, where 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 occurrence in the corpus. This gives more weight to terms which appear in few documents, and less weight to those which appear in many.

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

## 24 Topic Model

We use non-negative Matrix Factorisation (NMF) [3], 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  $V$ :

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

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 [4], and are saved in a database and topic visualisation system based on [5]<sup>1</sup>.

## Model selection

Topic models are calculated for 70, 80, 90, 100, 110, 120, 130 and 140 topics. The relative usefulness of each model was assessed subjectively by the authors, based on inspection of the online visualisation tool, and the spreadsheet **topic\_comparison.xlsx** accompanying the supporting information. The spreadsheet shows each set of topics in adjacent columns. Topics from each model are placed next to the topics with the largest number of each topic's 10 highest scoring words in common. This helps authors to find an appropriate level of granularity for the analysis.

## 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

Disciplinary Entropy inverts the measurement of a conference's topical diversity suggested in [6], by measuring a topic  $z$ 's entropy  $H$ , where

$$H(f|z) = - \sum_{i=1}^K \hat{p}(f|z) \log \hat{p}(f|z) \quad (3)$$

based on the empirical distribution of a field  $f$  in the documents  $d$  in each topic:

$$\hat{p}(f|z) = \sum_{d:z_d=z} \hat{p}(f|d) \hat{p}(d|z) \quad (4)$$

---

<sup>1</sup>The system adds new functionality to [5] and combines it with a system for managing sets of documents and queries. The code and additional information is published online at <https://github.com/mcallaghan/tmv>



Figure SI.1: SI Disciplinary Entropy

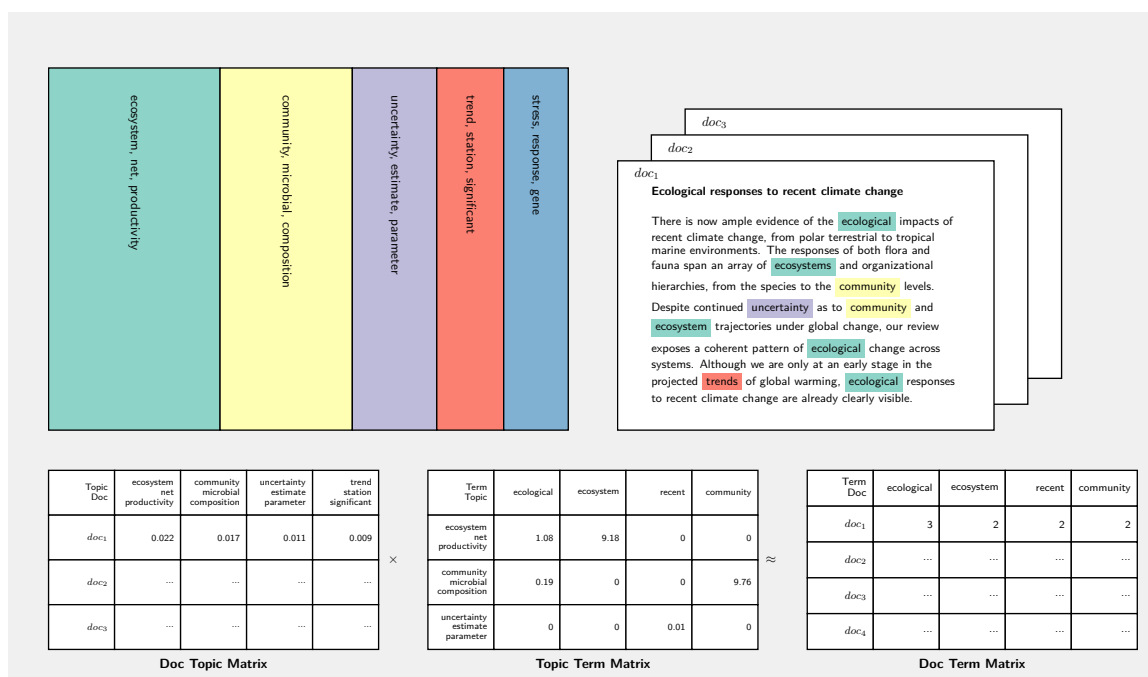


Figure SI.2: SI Topic make up of a single document

## Glossary

- ncep:** National Centers for Environmental Protection
- fco:** Fugacity of Carbon Dioxide
- pfc:** Perfluorocompound
- otcs:** Open Top Chambers
- dtr:** Diurnal Temperature Range
- sres:** Special Report on Emissions Scenarios (200)
- petm:** Paleocene Eocene Thermal Maximum
- amf:** Arbuscular Mycorrhizal Fungal
- sf5cf3:** trifluoromethyl sulfur pentafluoride (A Potent Greenhouse Gas Identified in the Atmosphere, 2000)
- clc:** Chemical Looping Combustion
- cwd:** Coarse woody debris
- etm:** Enhanced Thematic Mapper (NASA satellite sensor)
- cmip5:** Coupled Model Intercomparison Project 5 (Starting 2008)
- cmip3:** Coupled Model Intercomparison Project phase 3 (first published 2007 [7])
- mofs:** metal-organic frameworks (for CO<sub>2</sub> storage)
- sdm:** statistical-dynamical model
- mmms:** Mixed Matrix Membranes (for CO<sub>2</sub> capture)
- cop21:** 21st Conference of Parties (Paris 2015)
- c3n4:** Carbon nitride (a synthetic nanomaterial used for hydrogen production)
- sdg:** Sustainable Development Goals

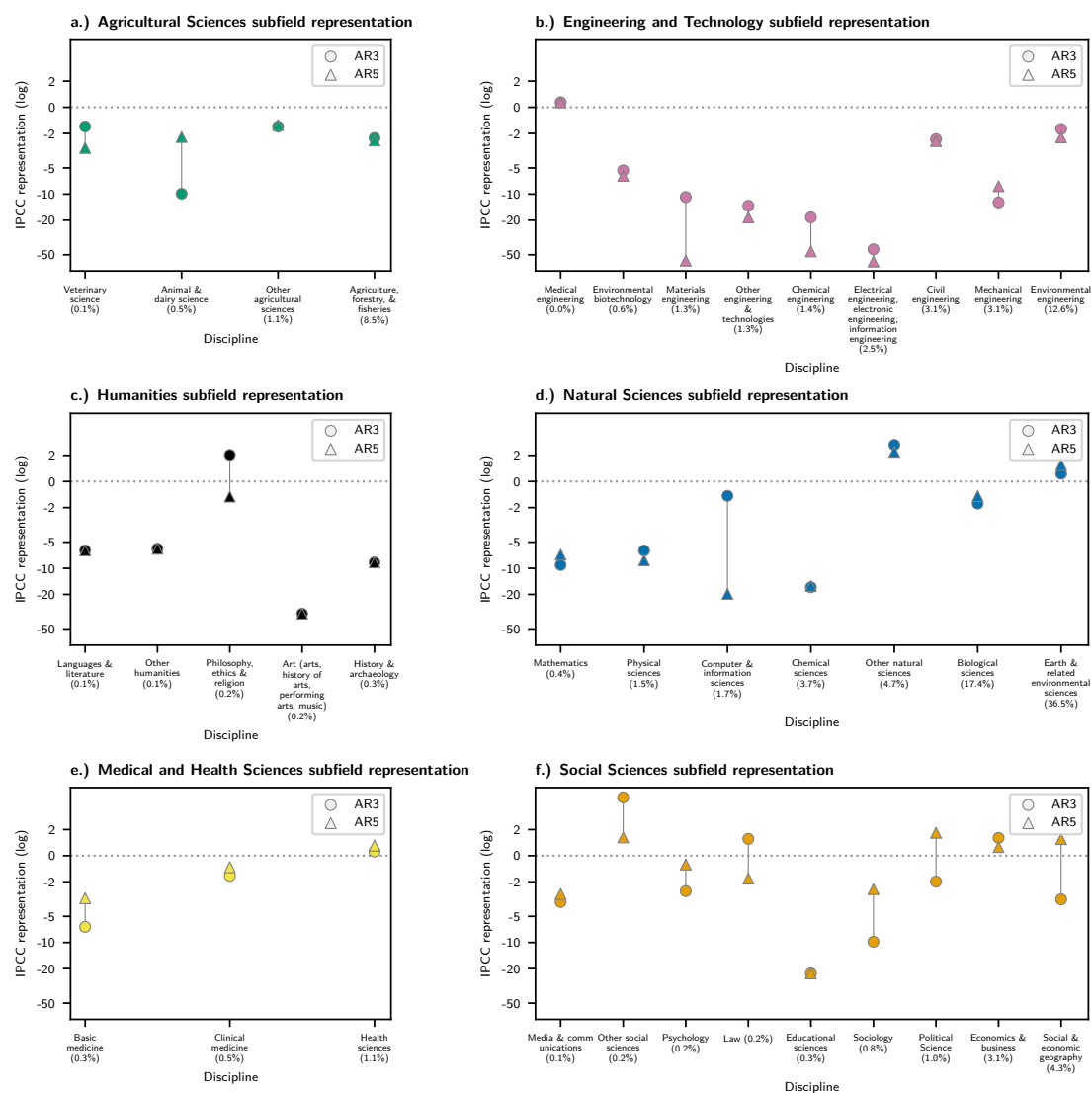


Figure SI.3: SI Representation by subfield

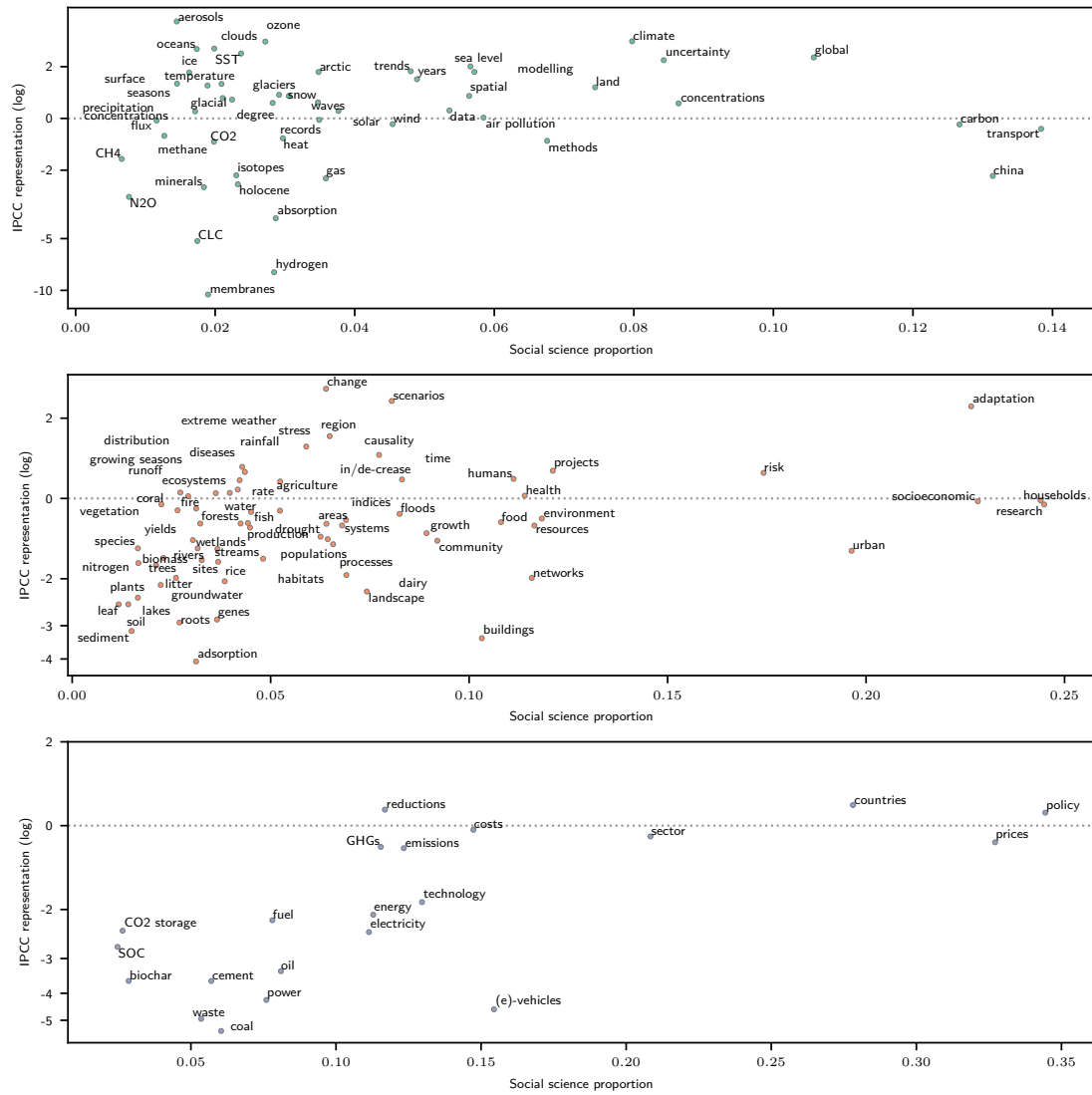


Figure SI.4: SI Social science & representation in topics across working groups

## 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] D D Lee and H S Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91, 1999.
- [4] 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.
- [5] Allison J B Chaney and David M. Blei. Visualizing Topic Models. *Icwsn*, pages 419–422, 2012.
- [6] David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the history of ideas using topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, pages 363–371, 2008.
- [7] Gerald A. Meehl, Curt Covey, Thomas Delworth, Mojib Latif, Bryant McAvaney, John F.B. Mitchell, Ronald J. Stouffer, and Karl E. Taylor. The WCRP CMIP3 multimodel dataset: A new era in climatic change research. *Bulletin of the American Meteorological Society*, 88(9):1383–1394, 2007.