Dynamic Topic Modeling to Infer the Influence of Research Citations on IPCC Assessment Reports

Jennifer Sleeman, Milton Halem, Tim Finin
Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, MD 21250 USA
{jsleem1,halem,finin}@cs.umbc.edu

Mark Cane
Lamont-Doherty Earth Observatory of Columbia University
Columbia University
New York, NY 10027 USA
mac6@columbia.edu

***Abstract*—A common Big Data problem is the need to integrate large temporal data sets from various data sources into one comprehensive structure. Having the ability to correlate evolving facts between data sources can be especially useful in supporting a number of desired application functions such as inference and influence identification. As a real world application we use climate change publications based on the Intergovernmental Panel on Climate Change, which publishes climate change assessment reports every five years, with currently over 25 years of published content. Often these reports reference thousands of research papers. We use dynamic topic modeling as a basis for combining report and citation domains into one structure. We are able to correlate documents between the two domains to understand how the research has influenced the reports and how this influence has changed over time. In this use case, the topic report model used a total number of 410 documents and 5911 terms in the vocabulary while in the topic citations the vocabulary consisted of 25,154 terms and the number of documents was closer to 200,000 research papers.**

***Keywords*-big data; topic model; cross-domain correlation; data integration; domain influence;**

I. Introduction

Combining large data sets from a combination of sources can be complicated by a number of factors such as contextual differences and semantic differences. However, the benefit is finding facts that provide additional meaning and are supportive of tasks such as classification and inference that could not be acquired from processing a single source of data.

When working with scientific research that has an impact on society, panels are often formed to report on the findings of that research. Understanding how that research influences the recommendations and assertions in the reports also entails combining sources of data. However, the sources combined are constrained since they are mostly related to the same general domain, i.e., they are not vastly different domains.

Understanding how research influences report recommendation or assessments, or in general how one domain influences another, provides a basis for inferring how that research may influence future reports. An important part of understanding how one domain influences another is to understand how these domains are changing over time.

In this work we use Dynamic Topic Models (DTM) [1] to model how topics evolve for different domains. We apply our own methodology to find a common vocabulary between the two domains and build divergence matrices using Jensen-Shannon divergence to discover pairs of cross-domain topics that have low divergences. We use these cross-domain topic pairs to cluster documents from the two domains to discover documents that are semantically closely related.

A. Use Case - Climate Change Assessment Reports

“The Intergovernmental Panel on Climate Change (IPCC) is a scientific body established in 1988 under the auspices of the United Nations” at the request of member governments, to provide “the world with an objective, scientific view of climate change and its political and economic impacts” [2]. Thousands of scientists voluntarily contribute to conducting computational simulated scenarios, writing sections of chapters and formulating the summary recommendations in the Assessment Reports, which are then reviewed by governments. The IPCC issues Assessment Reports approximately every five years beginning in 1990 with AR1, 1995 (AR2), 2001 (AR3), 2007 (AR4) and 2014 (AR5). Each report obeys a formal structure and consists of four books, (Physical Science Basis, Impacts, Adaptations and Vulnerability, Mitigation of Climate Change and Synthesis Reports) [2]. Each book has 12 to 16 chapters with a Summary for Policy Makers. Each chapter contains 800 to 1200 citations.

B. Contribution

Our work makes contributions to help scientists gain insights into how climate change research is evolving and also contributes a general approach for integrating data from a scientific domain and from bibliographic sources which support it. This is the first time to our knowledge that anyone has created a semantic language model of the IPCC literature and its citations. We have modeled over 30 years of climate change research and can show how key concepts have evolved over time. By using our approach to integrating two domains into a common space, we can show how particular aspects and threads of climate change research relate to the various IPCC reports.

C. Application

This work can also be used as a tool for climate change researchers to quickly reference and acquire a coarse understanding of climate change research past and present. It can reduce the human time spent in understanding how climate change research has evolved over the past 30+ years, and can be used, with additional user interfaces, to support a search tool for a climate change researchers.

II. Background

In this section we describe the purpose of the IPCC. We also describe topic modeling since this method of finding thematic structure in documents is foundational to our work.

A. IPCC

Approximately every five years dating back to 1989, thousands of climate scientists, research centers and government labs volunteer to prepare comprehensive Assessment Reports for the Intergovernmental Panel on Climate Change (IPCC) [2]. These are highly curated reports distributed to approximately 200 nation policy makers. There have been five IPCC Assessment Reports to date, the latest leading to a Paris Agreement in December 2016 [3] signed thus far by 172 nations to limit the amount of global Greenhouse gases emitted to producing no more than a 2° C warming of the atmosphere.

The IPCC operationally defines climate science in three parts:

1) The study of the physical, chemical and biological mechanisms that determine the workings of the climate system, which is comprised of the atmosphere, the oceans, the cryosphere and the land surface. It investigates past, present and future climate though observational, theoretical and modeling.
2) The assessment of impacts, especially anthropogenic influences on the climate system, on ecosystems and socio-economic systems. Important components include sea level rise, extreme events such as hurricanes, impacts on agriculture.
3) The possibilities for mitigating climate change through such strategies as conservation, renewable energy, carbon capture and sequestration, and geoengineering.

These reports are a living evolving big data collection tracing 30 years of climate science research, observations, and model scenario intercomparisons. They contain more than 200,000 citations over a 30 year period that trace the evolution of the physical basis of climate science, the observed and predicted impact, risk and adaptation to increased greenhouse gases and mitigation approaches, pathways, policies for climate change.

The task of studying these reports and understanding the research that supports each report is laborious, hence as soon as one report is published the work begins for the composition of the next report, a five year process. This sort of effort can be seen in other domains such as law, economics and various scientific domains. Leading to the question, how does one come to understand and infer new information from big data that spans over many decades?

B. Topic Modeling

Foundational to this work is both Latent Dirichlet Allocation (LDA) [4] and Dynamic Topic Models (DTM) [1].

Topic modeling has a long history of research particularly in the Natural Language Processing domain and is used to identify the semantic structure or 'hidden' structure of a collection of documents. LDA is based on the early work of Deerwester et al. [5] who introduced the concept of Latent Semantic Analysis (LSA) based on which uses singular value decomposition and Hofmann [6] who introduced the concept of Probabilistic Latent Semantic Indexing (pLSI) which introduced a probabilistic generative approach. Blei et al. [4], [7] furthered the probabilistic generative approach by incorporating a Dirichlet prior and using a Bayesian estimation.

With LDA each document has a mixture of topics represented as a probability distribution and each topic is represented by a probability distribution over the set of words found in the vocabulary which is composed from the words found in the collection of documents. Topics are drawn from a Dirichlet distribution.

The generative process of LDA is typically shown as a plate diagram, conveyed by Figure 1, where ω represents the words, $\beta_{1..k}$ are the topics, $\theta_{d,k}$ is the topic proportion of topic k in document D, and $Z_{d,k}$ is the topic assignments [4], [7].

From this generative process, a joint probability distribution is obtained over observed and hidden variables [4], [7]. Equations 1 and 2 [4], [7] show the joint distribution of hidden and observed variables and the posterior which is intractable and is estimated using a variational method such as an EM algorithm [8] or a sampling method such as Gibbs sampling [9].

$$\begin{aligned} p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) \\ = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \\ (\textstyle\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d:n})) \end{aligned} \tag{1}$$

$$\begin{aligned} p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \\ \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \end{aligned} \tag{2}$$

Where k is a specified input number of topics across the document collection and n is a specified input number of terms per topic. β_k is the topic proportions θ_k is marginal probability of observed variables (computed by summing the joint distribution over all possible hidden topic structure) z_d is the topics assignments W_d is the observed word for doc d

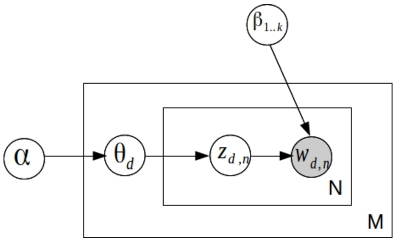


Figure 1: Plate diagram of LDA (Blei et al., 2003)

DTM tries to capture how topics evolve over some specified time period, where documents are split into specified time slices and a topic model is generated for each given time slice. Topics evolve over time and are captured in this model. The Figure 2 is typically used to capture this generative process where β represents the parameters of a topic for some time slice t_i and topics along with topic proportions evolves over the time slices [1]. A normal distribution is used over topics and approximate inference is achieved with a variational method. Further discussion can be found in the work by Blei et al. [1].

Our time slices are discrete, in that we use assessment report periods as time slices. Therefore, DTM is an appropriate way to model how these time slices are changing. It may be less appropriate if we were solely analyzing continuous time slices.

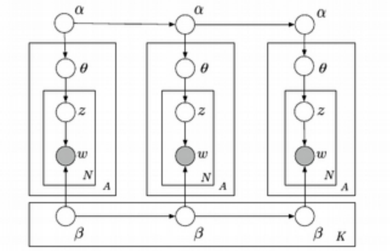


Figure 2: Plate diagram of DTM (Blei et al., 2006)

III. Methodology

In this section we will explain the structure of the IPCC reports, how we acquire the citations, preprocessing, our method for modeling and how we use the document-topic distributions and topic-term distributions. Figure 3 visually depicts the structure of the IPCC reports. This structure is important to this work as it relates to how we build the topic model. The topic model is based on conceptual documents but what is treated as a document can vary based on how we wish to use topic modeling.

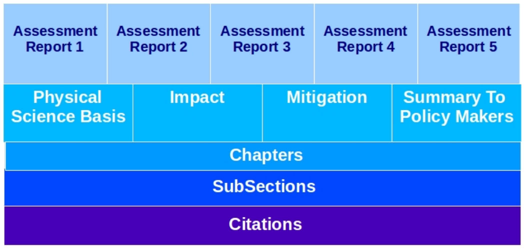


Figure 3: The Structure of IPCC

A. IPCC Hierarchical Structure

The IPCC is naturally hierarchical. Reports are composed of books, books are composed of chapters, chapters are composed of subsections.

There are n reports $ar_1, ar_2, ..., ar_n$, currently $n = 5$.

There are m books $br_{n,1}, br_{n,2}, ...br_{n,m}$ where $br_{n,m} \subset ar_n$, currently $m = 4$ for all ar_n.

There are l chapters $ch_{n,m,1}, ch_{n,m,2}, ...ch_{n,m,l}$ where $ch_{n,m,l} \subset br_{n,m}$.

The citation modeling consist of documents ci that are cited in the chapters of the IPCC books.

For each $ch_{n,m,l}$, we extract k citations $ci_{n,m,l,1}, ...ci_{n,m,l,k}$ found in that document. The citations are stored in a directory structure by chapter, by year.

The main focus of this work is to model reports and to model citations. For citations a document in our topic model is equivalent to the full text of a paper. For reports, we experimented with treating a chapter as a document, a book as a document and even a full report as a document. Treating chapters as documents yields a higher number of documents per time slice.

B. Preprocessing

Report chapters are in the form of PDF files, we used a pdf to text converter to convert PDFs to text using PDFMiner[1]. Unfortunately these conversions can produce noisy text but the essence of the document can be captured. At the end of each chapter is typically a set of citations to support the chapter. We parsed these sections to extract citations and generated a knowledge base of citations linked to each chapter which captures the chapter, the year, the authors, and the title of the citation. We built parsers that were specific to each assessment report so as to maximize the number of extracted citations. Using Microsoft Bing [10], we retrieve approximately 150,000 citations referenced across chapters and convert those citations to text also. As one could imagine, this is a challenging task as the citations are often formatted in non-standard ways or the retrieval is not possible. Errors are expected and noise is

[1] https://pypi.python.org/pypi/pdfminer/

not uncommon. For the Physical Science Basis book, we calculated approximately a 20% error rate between parsing, extraction and retrieval. Meaning we were able to retrieval approximately 80% of the citations.

Using a word n-gram model based on a heterogeneous climate change glossary referenced in Table I, we first look for word n-grams, then singular words. We use lemmatization to get a single form of words. We remove all stopwords, words that are significantly numeric (necessary since many of the reports contain extensive numeric analysis) and functional words. We also apply noise filtering and exclude words with a frequency less than ten or a length less than three. Word frequencies are calculated for chapters and citations.

Table I: Climate Change Glossary Sample

Glossary Sample
anthropogenic climate change
anthropogenic influence
anthropogenic carbon dioxide
atlantic oscillation
atmospheric aerosol
atmospheric dynamic
atmospheric model
atmospheric co2
anthropogenic emission
anthropogenic co2 emission
aerosol cloud

C. Modeling

Using the C code based implementation of DTM[2] we model books, where each chapter in a book is treated as a document. We also build topic models for citations for each chapter. Temporal document sets are built based on the assessment period, where there are five periods. We use this time slicing for both the chapters and citations. Document-topic and topic-term probability distributions are built from the vocabularies of the respective assessment report chapters and citations. We currently use variations in modeling according to Table II where we also vary the number of topics K according to the size of the collection. We highlight our use of K in the Experimentation section of this paper.

D. Micro-Filtering

The output from DTM enables us to generate topic-term probability matrices and document-topic matrices from which we apply our own algorithms to achieve 'micro-filtering', single-domain and cross domain correlation.

From the topic-term probability matrix we calculate n high rank terms per topic. Where n defines how many high rank terms to include. We also build m high rank topics per document. These measures allow us to really understand which terms and topics are most significant in the models we build.

Table II: Variations in Modeling

Model Type	*Information Used*
Report	All Books
Report	Specific Book
Report	Specific Chapter
Citation	All Citations
Citation	Specific Books
Citation	Specific Chapters

E. Cross-Domain Correlations

Based on the high rank terms for each topic, we formulate a new 'micro-model'. We do this for any set of domains we wish to correlate. In our experiments in particular we generated 'micro-models' for reports and citations. In order to perform any sort of correlation between 'micro-models' we generate a common shared 'micro-model' vocabulary which is simply stated as given $V1$ and $V2$, we want to generate $VS = V1 \cup V2$. Using VS, we formulate new word vectors and calculate divergences using Jensen-Shannon divergence and correlations using Pearson correlation for each set of topics.

For report chapter topics and citation topics, for each (ch_tp, ci_tp) pair exists a set of correlation measures $CORR_{0..n} = corr1, corr2, ...corr_m$ where n is the number of (ch_tp, ci_tp) pairs. In this case we have two measures for each pair, a divergence measure and p-correlation measure. We use these sets of correlation measures to establish points in space where each $(ch\_tp, ci\_tp)_n$ pair is a point. Given that we have established m high rank topics per document, our $(ch\_tp, ci\_tp)_n$ points give way to a set of chapter,citation points which we use to establish relatedness and influential relationships.

The Jensen-Shannon divergence between two probability distributions $P1$ and $P2$ is defined as:

$$JSD\,[P1, P2] = \frac{1}{2}(KL[P1, \frac{P1+P2}{2}] + KL[P2, \frac{P1+P2}{2}]) \tag{3}$$

Where KL is the KullbackLeibler divergence.

The Pearson correlation is defined as:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}. \tag{4}$$

IV. Related Work

Blei et al. [1] used DTM to model the evolution of a collection of Science articles and showed promising results related to the evolution of topics for specific terms such as 'Atomic Physics' and 'Neuroscience'. This work through the topic chaining was able to capture known trends among the collection of Science articles. Their document collection was from JSTOR and they used an OCR engine to obtain

[2] https://github.com/blei-lab/dtm

a text version of these papers. It was not specifically stated but implied that each paper represented a document in this model. There was no mention of the citations referenced in each paper. It is likely the citations found in each document were not parsed, retrieved and used in their model.

Work by Shalit et al. [11] used DTM for modeling the musical influence. They applied this work to a large data set of songs for a continuous time period from 1922 to 2010. Their problem is similar from a hierarchical perspective, i.e. sound segments - songs - album structure is similar to our data - chapter - book - report structure. Influence is an important component of our work but we also examine how the science has changed over time, which is less of a focus in this work.

Recent work by Li et al. [12] described a number of experiments they conducted with dynamic topic models. They modified the original algorithm by using a hierarchical Pitman-Yor process. Document topic proportions and distributions are evolved using this method. They also used Gibb sampling rather than a variational method. This method was also built for multi-core use. Their experiments consisted of the ABC data set and used a time period with three months of content per time slice over the course of a ten year period. They showed similar term evolution results via the topics. This work primarily focused on building a dynamic scalable model and rather than the content, which is really the essence of our work.

More recent work by Hu et al. [13] also highlighted using dynamic topic modeling for topic evolution. Their work was focused on the evolution of a software project. The documents for this model were commit messages for a project's revision control system. This work did not modify the DTM algorithm itself but instead performed post-processng methods based on the document topic and topic term distributions. We similarly apply additional methods to the output of the DTM modeling.

There were other methods that examined topic evolution over time [14] that were less relevant to this work.

V. Experiments and Results

Our preliminary experiments focused primarily on one book, the Physical Science Basis.

Our first experiment involved a specific chapter from each assessment report with a significant representation of the concept 'Radiative Forcing', i.e., the main theme of the chapter. We built two models for this experiment, a radiative forcing chapter report model with five documents (each chapter document is decomposed into several subsections) and a vocabulary of 790 terms, and a radiative forcing citation model with 1451 documents as shown in Table III and 5744 terms.

Table III: Physical Science - Radiative Forcing Citation Counts

Assessment Report	Citation Count
AR1	12
AR2	164
AR3	345
AR4	435
AR5	490

Table IV: Physical Science Citation Counts

Assessment Report	Citation Count
AR1	1051
AR2	3393
AR3	4527
AR4	7096
AR5	9545

We built models for reports and citations with topic size varying as $K = [5, 10, 20, 30, 40]$. We found for reports the most useful model based on topic term rank was five topics with really the core topics clustering in two specific areas: 'radiative forcing' and 'anthropogenic' topics. For citations we found ten topics was sufficient.

Our second experiment involved all chapters within the Physical Science book. We limited the number of chapters for the Physical Science book to 11 chapters for each assessment period, where the assessment periods represent the five time slices. We built two models for this experiment, a physical science chapter report model with 55 documents and a vocabulary of 1953 terms, and a physical science citation model with 25,612 documents as shown in Table IV and 15,485 terms.

We experimented with changing a number of variables. For example, we experimented with changing the number of topics K. For the Physical Science reports we experimented with [5, 10, 20, 30, 40] topics. For Physical Science citations, we experimented with 10-number intervals from 10 to 100 topics. We also experimented with term rank, we tested our approach with a top 10 rank, top 15, and top 20 rank.

In the below results, we used 20 topics for the reports and 60 topics for the citations. We used low topic repetition as a measure for determining how large K should be. An example of a report topic evolution among high ranked terms is shown in Table V for the reports.

Among 'black carbon' research there was a publication peak around 2001 which is about the time the third assessment report was published. So we would expect to see a peak around AR3 or AR4 in Figure 4. We saw a slight peak at AR4 however it is not statistically significant. Overall the probabilities for 'black carbon' are low. We compare this with Figure 5 which shows the evolution of 'black carbon' for citations and do not see this peak at all. The probabilities are slightly higher in the citations model.

In this case, we only found a single topic that had 'black carbon' in the top terms however, often concepts can

Table V: Dynamic Topic Generation - 20 Topics Physical Science Reports

	Topic 14 - Top 10 Terms	Topic 15 - Top 10 Terms
Assessment 1	climate change	cloud
	general circulation models	anthropogenic
	radiative forcing	radiative forcing
	carbon dioxide	black carbon
	global warming	climate change
	ocean model	effect
	atmospheric co2	temperature
	carbon cycle	boundary layer
	temperature	fossil fuel
	system	sulphate
Assessment 2	climate change	cloud
	radiative forcing	anthropogenic
	general circulation models	radiative forcing
	atmospheric co2	black carbon
	global warming	climate change
	carbon dioxide	effect
	temperature	fossil fuel
	carbon cycle	temperature
	ocean model	boundary layer
	climate model	sulphate
Assessment 3	climate change	anthropogenic
	atmospheric co2	radiative forcing
	temperature	black carbon
	radiative forcing	cloud
	global warming	climate change
	carbon cycle	temperature
	carbon dioxide	fossil fuel
	general circulation models	boundary layer
	sea level rise	atmospheric aerosol
	climate model	sulphate
Assessment 4	climate change	anthropogenic
	temperature	radiative forcing
	atmospheric co2	black carbon
	radiative forcing	surface temperature
	equilibrium climate sensitivity	temperature
	sea level rise	climate change
	global warming	radiation budget
	carbon cycle	boundary layer
	carbon dioxide	tropospheric
	surface temperature	climate model
Assessment 5	temperature	radiative forcing
	climate change	anthropogenic
	equilibrium climate sensitivity	temperature
	carbon cycle	surface temperature
	atmospheric co2	climate change
	radiative forcing	black carbon
	sea level rise	boundary layer
	global warming	global warming
	carbon dioxide	climate model
	surface temperature	atmospheric aerosol

be found among a number of topics. Changes in concept probabilities for a given topic could imply a trend. However because these concepts may be found among the high ranked terms of multiple topics, it is important to observe how the concept trends among those topics. For example, the concept 'climate model' remains steady over assessments, shown in Figure 6. This was consistent with the other four topics that had 'climate model' present in the top ten terms. The concept 'general circulation model' showed a decrease in probability as seen in Figure 8. By observing another topic for 'general circulation model' as shown in Figure 9, we see though the trend isn't exactly the same, there is still a decrease in probability from the top ten. In climate change research the use of 'general circulation model' has been found in publications since the 1980s. More recently, 'climate models' are used more frequently than 'general circulation model'. Our models do tend to show these attributes. We compare 'climate model' topics among citations in Figure 7 with 'general circulation model' topics among citations in Figure 10 and there does also seem to be some agreement with citations.

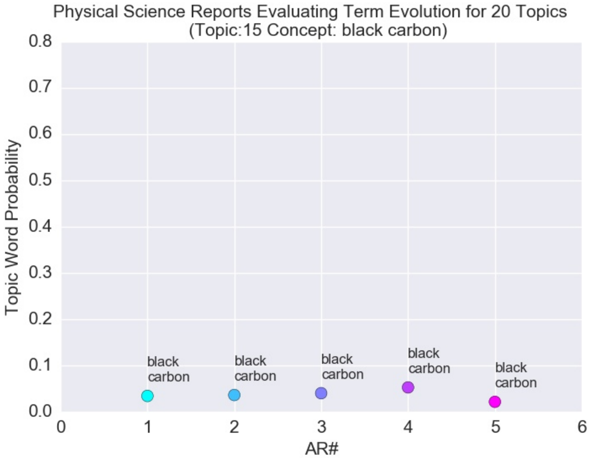


Figure 4: Physical Science Report Chapter Topic Evolution of 'Black Carbon'

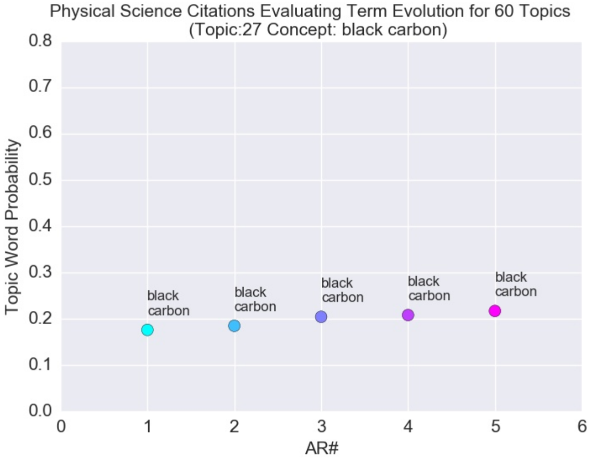


Figure 5: Physical Science Citation Topic Evolution of 'Black Carbon'

There were also results that seem to be in disagreement from reports to citations, as seen with the concept 'sea level rise' in Figure 11. The report topic trend tends to decline, whereas with the citations there tends to be a slight increase in probability as seen in Figure 12. This was also consistent among other topics which included 'sea level rise' as a term.

It is important to understand what is truly a trend and what is a side-effect of the model itself and how the documents are composed. For example, a concept in one assessment

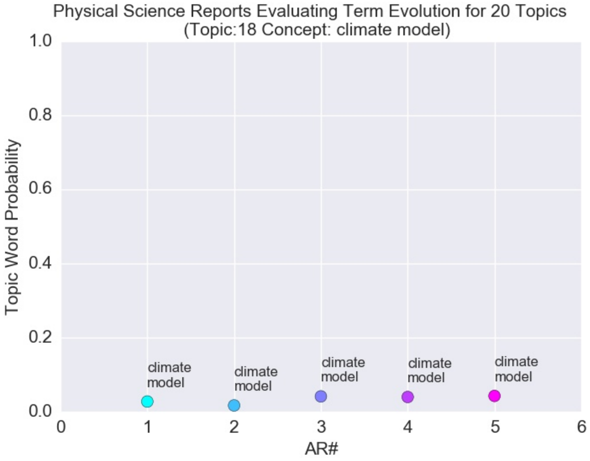


Figure 6: Physical Science Chapter Report Evolution of Climate Model

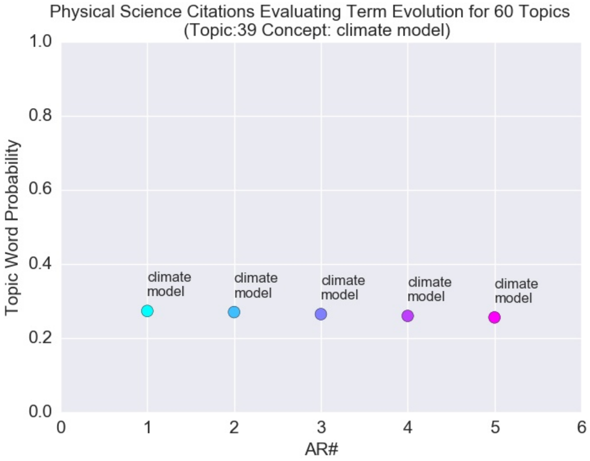


Figure 7: Physical Science Citation Evolution of Climate Model

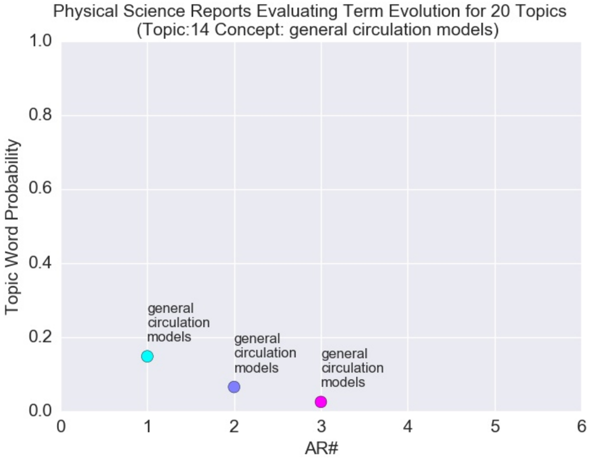


Figure 8: Physical Science Chapter Report Evolution of ’General Circulation Model’

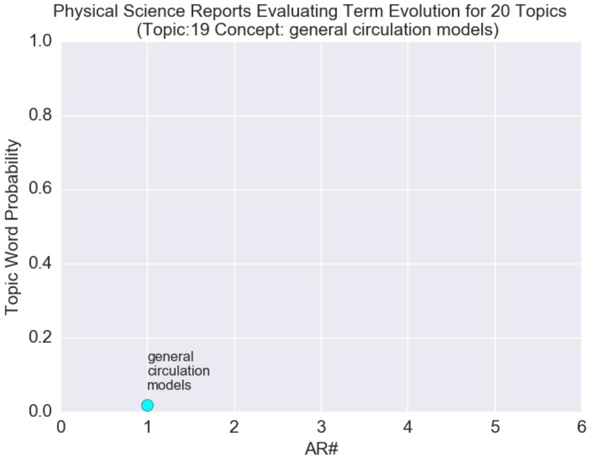


Figure 9: Physical Science Chapter Report Evolution of ’General Circulation Model’

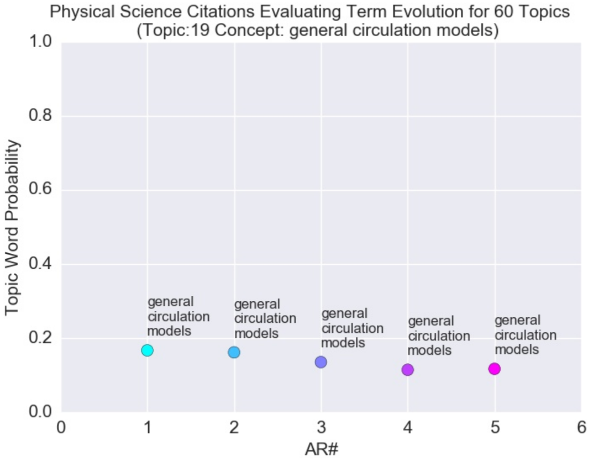


Figure 10: Physical Science Citation Evolution of ’General Circulation Model’

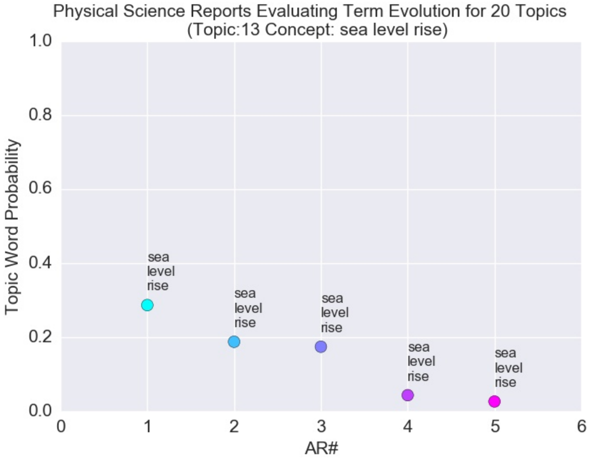


Figure 11: Physical Science Chapter Report Evolution of ’Sea Level Rise’

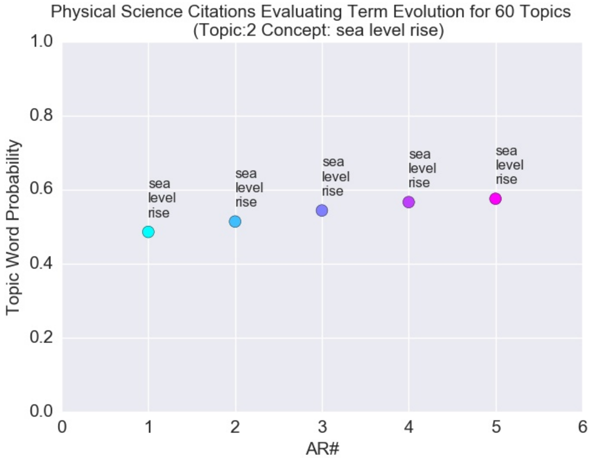
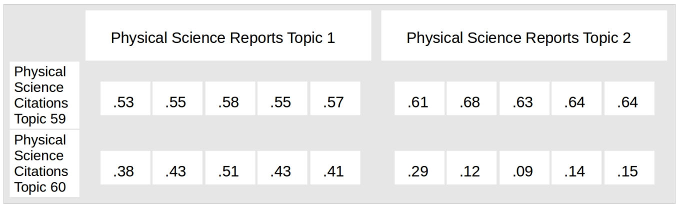


Figure 12: Physical Science Citation Evolution of 'Sea Level Rise'



	Physical Science Reports Topic 1					Physical Science Reports Topic 2				
Physical Science Citations Topic 59	.53	.55	.58	.55	.57	.61	.68	.63	.64	.64
Physical Science Citations Topic 60	.38	.43	.51	.43	.41	.29	.12	.09	.14	.15

Figure 13: Physical Science Report 20-Topic and Citation 60-Topic Divergence Matrix

may be coupled with a number of other concepts in a given chapter. However, in the next assessment, that same concept could appear with other concepts in a different chapter. The concept trend may change for a given topic but it will also change in other topics that capture this different representation. By observing how the probabilities change across all topics which contain a concept in the top ten terms we are able to gain a better understanding of the trend.

A. *Clustering Cross-domain Documents*

We used the output from the dynamic topic models to correlate the citations and chapters. In Figure 13, we show an example of a chapter report 20 topic by a citation 20 topic divergence matrix. The lower the divergence, the more likely the topics contain similar terms. In the experiment which includes chapter report 20 topics and citation 60 topics cross-domain divergences, we set a threshold of 10% and find topic pairs between the two domains that have a divergence of this score or less.

The divergence matrix can also be plotted to show (report topic, citation topic) points in space, where each point represents when the divergence was below the 10% threshold indicating a potential match between the cross-domain topics. For example, in Figure 14 we show report, citation points for Assessment Report 3.

We use the topic divergences to cluster report chapters as documents and full citation papers from the two domains.

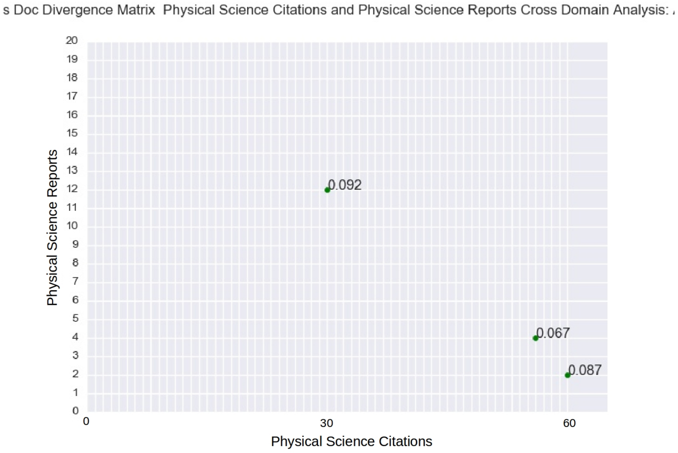


Figure 14: Physical Science Report Topic,Citation Topic Points in Space

With a chapter document topic probability threshold set at 50% and a citation topic probability threshold set at 80%, we were able to find clusters of chapter documents and citation documents from the two domains that are related. For example, in Figure 15 we found a cluster which included three chapters related to 'Radiative Forcing' from AR1, AR2, and AR3. To confirm this result, we mapped specific chapters from each assessment that relate to 'Radiative Forcing' using their table of contents and cosine similarity measure. What we found was that AR1, AR2 and AR3 were significantly more similar and, AR4 and AR5 were significantly more similar. This is consistent with the cluster results shown in Figure 15. The citations in this cluster can be mapped back to either these chapters from AR1, AR2, and AR3, or a chapter from one of the other assessment reports related to 'Radiative Forcing'. Most of these citations can be mapped to chapters which we did not designate as a 'Radiative Forcing' chapter but indeed have a discussion related to 'Radiative Forcing'. The thresholds are important in that they influence our precision and recall. When we relax the thresholds our recall does increase. More experimentation will need to be performed to find the optimal thresholds for each.

VI. Conclusion

This paper highlights our early work and a methodology for assessing relatedness and influence between two domains by topic modeling each domain and using a subset of the vocabulary to find cross-domain topic pairs which could be used to cluster documents from the two domains. We performed topic modeling of five IPCC Assessment Reports and we performed topic modeling of the citations over time based on their occurrence in each chapter of the assessment reports. This unique citation document collection as created by thousands of climate scientists enables one to discover the influence that the referenced publications exert on the

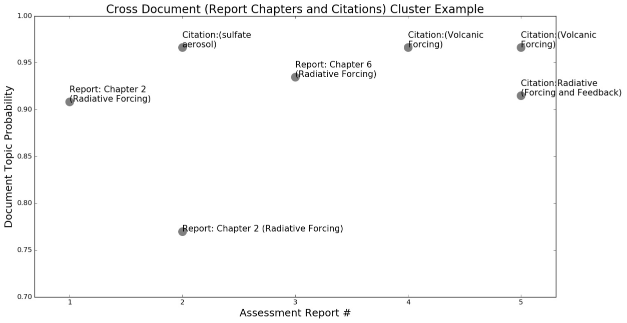


Figure 15: Physical Science Report and Citation Cluster Example

report as the citation research led to new topic emphasis over time.

The continuation of this work will include a service-oriented computing interface for the cross domain of topic report documents and topic citation documents that will enable climate change researchers to mine, correlate and display the influence of selected citations not even appearing at some specified ranked probability measure in that chapter or report. The value of such a service would be to significantly aid climate researchers by enhancing the ability of climate researchers, as well as the general community, to detect changes over time on the importance of portions of assessment reports based on probability measures of the relevant topic citations. This methodology is applicable to document collections from other domains for studying time evolving reports for business, government as well as science and engineering. Finally, we plan to include a Watson interface to be used to optimize client user content in exploring the complex interactions of climate processes and to recognize shifts in document topic terms owing to influences of temporal research contributions.

References

[1] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 113–120.

[2] I. P. on Climate Change, "Intergovernmental panel on climate change," https://www.ipcc.ch/.

[3] C. on Foreign Relations, "Pair agreement," http://www.cfr.org/climate-change/paris-agreement/p37361.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[5] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *JAsIs*, vol. 41, no. 6, pp. 391–407, 1990.

[6] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.

[7] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[8] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.

[9] T. Griffiths, "Gibbs sampling in the generative model of latent dirichlet allocation," 2002.

[10] M. Bing, "Microsoft bing," http://www.bing.com/.

[11] U. Shalit, D. Weinshall, and G. Chechik, "Modeling musical influence with topic models." in *ICML (2)*, 2013, pp. 244–252.

[12] J. Li and W. Buntine, "Experiments with dynamic topic models."

[13] J. Hu, X. Sun, D. Lo, and B. Li, "Modeling the evolution of development topics using dynamic topic models," in *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*. IEEE, 2015, pp. 3–12.

[14] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 424–433.