Discovering Scientific Influence using Cross-Domain Dynamic Topic Modeling

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

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

***Abstract*—We describe an approach using dynamic topic modeling to model influence and predict future trends in a scientific discipline. Our study focuses on climate change and uses assessment reports of the Intergovernmental Panel on Climate Change (IPCC) and the papers they cite. Since 1990, an IPCC report has been published every five years that includes four separate volumes, each of which has many chapters. Each report cites tens of thousands of research papers, which comprise a correlated dataset of temporally grounded documents. We use a custom dynamic topic modeling algorithm to generate topics for both datasets and apply cross-domain analytics to identify the correlations between the IPCC chapters and their cited documents. The approach reveals both the influence of the cited research on the reports and how previous research citations have evolved over time. For the IPCC use case, the report topic model used 410 documents and a vocabulary of 5911 terms while the citations topic model was based on 200K research papers and a vocabulary more than 25K terms. We show that our approach can predict the importance of its extracted topics on future IPCC assessments through the use of cross domain correlations, Jensen-Shannon divergences and cluster analytics.**

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

I. Introduction

Given an interdisciplinary scientific domain evolving over time, the sheer volume of publications involved makes it difficult for policy makers, general public and even scientists to track research and comprehend evolving discoveries, findings and implications. Uncovering relatedness across such scientific domains can provide insights into how concepts and findings interact and predict how topics will change in the future. For example, it is not possible for a climate scientist to know all of the literature in all of the disciplines related to the physical science and how the science will lead to various regional and global potential impacts, no less what are the implications of proposed mitigation responses. Having tools that automatically processes past assessments and citation information could reduce the time spent in composing future assessment reports.

Dynamic topic models can summarize large collections of documents at a given point in time and provide a way to find published research related to specific topics. Researchers are thus spared the time needed to manually discover and read related documents. Our approach can automatically link authors of research fields across multiple year assessment reports and citations, enabling researchers to see social network-based influence among climate change researchers. More importantly, we find which chapters across three decades of assessment reports are directly related and how they connect to current and prior research papers. Our prototype is an example of a powerful tool to understand interdisciplinary climate research and give insights into how the importance of specific concepts evolves.

In our work we use *domain* to refer to a collection of documents pertaining to a given scientific discipline. We describe two domains in this work, one that includes the chapters in the IPCC assessment reports and another that encompasses the collection of papers they cite from journals and conferences. In this example, the two domains are more precisely defined as sub-domains of a larger climate science domain of research. However, the two sub-domains are quite distinct. The first acts as the agreed upon consensus of the current state of climate change and gives a summary of the most important and influential research. The second acts as information retrieved from 'sensors' that is used to support the claims made by the first domain. It tends to be noisy, includes a wider mixture of topics, and can also be less relevant to the concepts described in the report. However, it can also expand upon what is described in the reports.

The concept of domain influence can be extended to identify relationships between two weakly related domains, such as climate change and the economy, or climate change and health care. This concept of identifying domain influence can also be extended to include more than two domains. For example, it could be used to understand how climate change and food scarcity influences events, such as violence in the Middle East [1].

A topic model for each domain is created from the collection of documents of that domain. The topic model uses probabilities and word co-occurrences to establish concept-based summaries of the document collections in terms of 'topics'. Each document is seen as a mixture model given by the topic probabilities. For example, a document pertaining to 'Radiative Forcing' is seldom just about 'Radiative Forcing' but rather may describe other

concepts such as 'greenhouse gases' and 'greenhouse effect'. The topics capture this mixture quite well. In this work the report topic model used 410 documents and a vocabulary of 5911 terms while the citations topic model was based on 200,000 research papers and a vocabulary more than 25,000 terms. Cross-domain analysis was used to understand how the research domain and reports domain are related and how the research influences the reports. The Jensen-Shannon method [2] was used to find topic distributions across the two models that present a low divergence, hence establishing which topics across the two domains had similar word distributions. This was then used to find the documents from the two domains that could be related. Since the topic models are defined in terms of time, by finding related documents across time slices among research papers and reports, the research papers that were published before a related report implies a level of influence. The lower the divergence between the two topics, the more likely the research paper and report relate to each other. Given the research paper was published before the related report, there is a higher likelihood of influence.

Our work uses dynamic topic modeling [3], concept evolution and cross-domain analysis to understand relatedness and influence between two domains. We now describe each of these components in more detail.

Concept Evolution: Given a scientific domain, often there is an evolving set of concepts that naturally emerge from the collective research literature. Understanding how a scientific domain changes over time can be accomplished by understanding how these concepts evolve. Hence concept evolution provides insight into domain evolution. By learning how a domain has evolved in the past, future projections can be inferred as to how the domain may change in the future. These sorts of projections can be used to guide the direction of future research areas.

Relatedness: Relatedness refers to the degree that two documents are similar. For example, chapters from an IPCC report book that pertains to 'Radiative Forcing' are related to chapters in another that pertains to 'desertification'. Documents are rarely about a single concept, but rather characterized as being about a mixture of concepts. This idea is the essence of topic modeling and makes it a suitable approach for discovering relatedness. Since this work includes 'dynamic' topic models, a discrete time dimension is included. That dimension enables our models to shine light on how a document might influence other documents. The cross-domain divergence method produces a mapping between topic domains that links documents by means of common topics. Given one topic model for a set of research papers and a second based on the assessment reports based on them, the cross-domain mapping provides a way to understand how a documents from an earlier time slice may have influenced the report chapters that referenced it.

Domain Influence: In this work, the type of influence discovered is indicative of the data itself. Scientific textual data is specifically used, which in some ways can be described by its impact on society. The cross-domain method is used specifically to uncover how the research for a given scientific field influences the assessments put forth by committees and groups. Assessing research and the generation of formalized reports is common when the outcome of the scientific research has a global impact.

To understand global problems such as climate change, pandemics, terrorism, food scarcity or cybersecurity threats, a multi-disciplinary approach is required. This suggests that the data which supports analysis of these issues is also multi-disciplinary, meaning the data originates from multiple domains and sub-domains. This type of data is important to our work because typically there is a synthesis of the research for a given period of time and that synthesis is a concise summary of the research. Treating the research as observations emitted from 'sensors' (e.g., papers in journals, conference and reports) and the topic model as the model for which this information is synthesized, then relating these two or more domains over time provides more than just relatedness. If a research paper was published prior to a summary and that research paper was associated with a highly ranked topic, it is likely there is an influence on the summary which is also related to that topic. This is how influence is defined in our work.

II. Background

Dynamic Topic Modeling (DTM) [3] is an extension of topic modeling that analyzes temporally tagged documents with the goal of capturing how topics are changing over discrete time periods. Documents are split into time slices and a topic model is composed for each slice then linked together where topics and topic proportions are allowed to 'evolve' over the set of time slices. A normal distribution is used over topics and approximate inference is achieved using an expectation maximization algorithm [4].

DTM uses the concept of state space models and maintains the natural parameter for topic term distributions and document topic distributions. Multinomial distributions are described in terms of Gaussian distributions, enabling the parameters to be held in a state space model which is used to evolve the parameters [3]. A topic model is built for each time slice, however the topics at time t are evolved based on what is held in the state space model for time slice $t-1$.

Given there are k topics in time slice t, there is a vector of 'natural parameters' for t and k where $\beta_{t,k}$, Gaussian noise as defined by the following Equation 1 is evolved.

$$\beta_{t,k}|\beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k,\sigma^2 I}) \tag{1}$$

Instead of a Dirichlet distribution for document proportions, Blei et al. uses a logistic normal distribution [3]. The

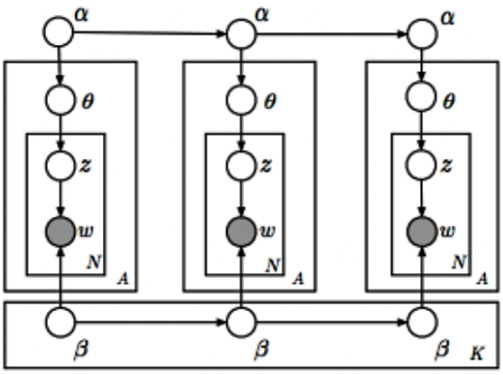

Figure 1: Dynamic topic modeling plate diagram

diagram shown in Figure 1 [3] is typically used to capture the generative process where β represents the parameters of a topic for some time slice t_i and topics along with topic proportions evolve over the time slices [3]. For posterior inference, Kullback-Leibler divergence to the true posterior is used to approximate the Gaussian observations with the true posterior [3]. Variational approximations account for time using Kalman filters [5].

III. Related Work

Early work by Dietz et al.[6] focused on defining influence by means of visualizing how topics flow from research paper to research paper with the goal of finding the influence a citation has on the paper citing it. They showed they are able to improve citation influence identification given their citation influence model which is based on two topic models, a citation model and a model that cites. This work provides a strong foundation as to why our cross-domain mapping method is a reasonable approach, though it is solving a different research problem.

Li et al. [7], developed a topic correlation and Jenson-Shannon divergence measure for cross-domain text classification by treating three separate term vocabularies. In this work they use topic modeling and the Jensen-Shannon divergence method to act as a way to classify unlabeled data based on a model that was trained with labeled data. Their method is less relevant to finding relatedness and influence, which is the focus of our work. However, their use of Jensen-Shannon is similar by nature, in that they are trying to relate different topic models to each other by using topic divergences.

Blei et al. [3] used DTM to model the evolution of a collection of articles from *Science* and showed the evolution of topics for specific terms such as 'Atomic Physics' and 'Neuroscience'. Using topic chaining, it capture known trends among the collection of articles. There was no mention of the citations referenced in each paper and this work does not address the cross-domain nature of our work.

Hall et al. [8] address how scientific ideas have changed over time by modeling temporal changes employing DTM, with probability distributions for the ACL Anthology. Their work proposes extensions to their model by integrating topic modeling with the citations as done in this paper. Work by Shalit et al. [9] 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. However, influence and relatedness is an important component of our work.

More recent work by Hu et al. [10] also highlighted dynamic topic modeling for topic evolution in a software project. The documents for this model were commit messages for a project revision control system. This work did not modify the DTM algorithm itself but instead performed post-processing methods based on the document topic and topic term distributions. Our work similarly applies additional methods to the output of the DTM modeling.

Tang et al. [11] investigate the use of topic modeling to identify extreme events based on numerical atmospheric model simulations. They associate text terms with statistical ranges of numerical variables. This work is most closely related to ours as it has a similar use case. There were other methods that examined topic evolution over time [12] that is less relevant as it keeps topics as constant and uses topic co-occurrence patterns to identify changes over time. Our goal is to understand how topics evolve over time and to use the topics as a means to map between disciplines.

IV. The IPCC Data Set

There are currently five IPCC assessment reports, AR1-AR5, each of which follows a similar structure consisting of four distinct books: Physical Science Basis, Impacts, Adaptations and Vulnerability, Mitigation of Climate Change and Synthesis Reports. Each book has between 11 and 25 chapters and each chapter typically contains between 800 and 1200 citations to external documents. This structure is formally defined as follows: 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 $l(m,n)$ chapters $ch_{n,m,1}, ch_{n,m,2}, ...ch_{n,m,l}$ where $ch_{n,m,l} \subset br_{n,m}$. For each $ch_{n,m,l}$, $k(m,n,l)$ citations $ci_{n,m,l,1}, ...ci_{n,m,l,k}$ found in that document are extracted.

V. Methodology

Given two or more domains, where each domain is described by a set of temporal documents, a dynamic model is built for each domain. Given the use case in this work, a dynamic topic model is built for citations referenced in the IPCC reports and for the IPCC reports. The intention of this work is to find research that is related to the various chapters defined within the IPCC reports and research that may have influenced a particular report. Identifying relatedness

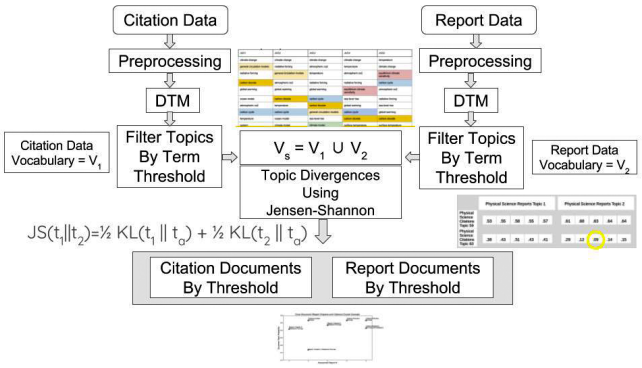

Figure 2: A cross-domain overview of mapping between two dynamic topic models to discover relatedness and influence.

and influence is performed using a novel cross-domain divergence method which will be described further.

A. Preprocessing

In order to build topic models with useful topics, preprocessing is a necessary step. A climate change glossary was used, based on a custom ontology, to guide the preprocessing of text, whereby a 'bag of words' consisted of a 'bag of domain-specific phrases and words'. Lemmatization and stop word removal was performed and functional and numeric words are removed. Words with low frequency and words with a length less than three were also removed to reduce the noise incurred during the PDF to text conversion, as often mistakes in the conversion result in stray characters or incomplete words.

B. Model Generation

The DTM code [13] was used to build dynamic topic models of both of the domains. As with LDA, generating a model involves calculating the frequency of each word found in a document. Model generation also involves changing hyperparmeters and the number of topics. A particularly important parameter in DTM is the parameter which controls how much variance is allowed from one time slice to another.

C. Cross-Domain Mappings

Figure 2 shows a high level view of the methodology described in this work with particular emphasis on discovering relatedness and influence among citations and reports. Given two domains, though this approach could be applied to more than two domains, documents are pre-processed to eliminate stop words and to discover relevant domain-specific phrases. A dynamic topic model is built for each domain. To discover cross-domain relatedness and influence, Jensen-Shannon divergence [2] is calculated for each pair of topics across the domains by first reducing the vocabulary based on the intersection of top n words for each topic, as shown if Figure 3.

For each topic in domain 1 described by d_1 and each topic in domain 2 described by d_2, using the term probabilities in

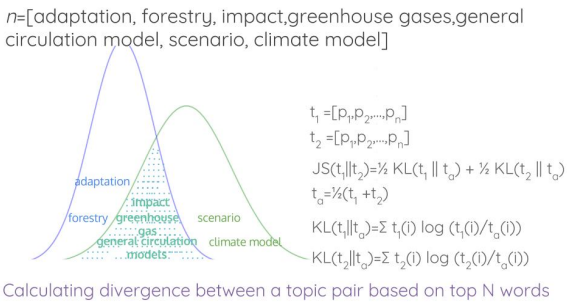

Figure 3: Using Jensen-Shannon divergence to find cross-domain topic pairs with the lowest divergence.

each topic, take n highest probable terms and generate a new vector of terms V that consists of the top terms from t_{d_1} and the top terms from t_{d_2}. For each term in V, if the term exists in t_{d_1}, assign the probability from t_{d_1} to $V_{t_{d_1}}$. If the term exists in t_{d_2}, assign the probability from t_{d_2} to $V_{t_{d_2}}$. Normalize $V_{t_{d_1}}$ and $V_{t_{d_2}}$ such that their probabilities are redistributed and sum to 1. This will result in two new probability distributions for t_{d_1} and t_{d_2}.

After this method was performed for every pair of topics across d_1 and d_2, using Jensen-Shannon divergence [2] pairs with the divergences below a given threshold were found. The Jensen-Shannon divergence between two probability distributions $t1$ and $t2$ is defined as:

$$JS\,(t1\|t2) = \frac{1}{2}KL(t1\|\frac{t1+t2}{2}) + \frac{1}{2}KL(t2\|\frac{t1+t2}{2}) \quad (2)$$

where KL is the Kullback Leibler divergence between two distributions. The Jensen-Shannon divergence is used, rather than the Kullback Leibler divergence, for its property of symmetry.

The smaller the threshold, the few pairs will be used to obtain documents across the two domains. Given a pair of topics below the divergence threshold, documents for each topic were discovered based on a second threshold which defines how much of the document mixture model should pertain to the paired topic. There were two threshold parameters that can be used to control how many documents from the two domains are partitioned together. The first threshold controls how many topic pairs will be used. The second determines how many documents are returned based on how significant the topic is in the document mixture. Both are configurable and could be automatically discovered and, in this work, optimized based on the data set.

Documents from the two domains that met the defined threshold were partitioned. Each partition was defined over time slices, hence if there was some ontological relationship between the documents in one domain and the documents in the other, if the documents were partitioned together, then influence was inferred between the documents.

VI. Experiments and Results

Experiments were conducted that were related to optimizing DTM parameters, such as the number of topics K,

comparing different values for variance, comparing subsections vs. chapters as the documents for the topic models. These experiments are described further.

A. Measurements

Likelihood can be described in terms of a probability model, where a given topic modeling algorithm should try to maximize the probability of observable values given a set of parameter values. When using variational inference, which is the method used in this research, the parameters are estimated and likelihood is approximated.

Perplexity can be described as the inverse probability of a held-out test set and is a measurement of how well a probability distribution or probability model predicts a sample. It may be used to compare probability models. A low perplexity indicates the probability distribution is good at predicting the sample. For example, when perplexity is applied to language models, the models are evaluated based on how well they can predict the next word in a sentence [14]. A better model would assign a higher probability to the word that should occur next in the sentence. Since the perplexity is the inverse probability of the held-out test set, lower perplexity would indicate a better model.

There are variations on how perplexity is measured from per-word perplexity to model perplexity. For LDA-based models, generally perplexity of w, a held-out set can be defined by equation 3, where N_d is the number of words in the d^{th} document and M represents the model. In our work we used a per-word perplexity, as described in Wang et al. [15].

$$Perplexity(testset_w) = \exp(-\frac{\sum_{d=1}^{D} \log p(w_d|M)}{\sum_{d=1}^{D} N_d}) \quad (3)$$

Coherence measures are based on co-occurrences of words. Given a topic that is filtered by the top N words, the words are evaluated to determine how much they likely co-occur. External sources can be used for this measure, as well as the internal data. Higher coherence scores indicate a better topic. In this work coherence was evaluated using the UMASS measure introduced in [16] and described by equation 4.

$$Coherence_{UMASS}(w_m, w_l) = \log(\frac{p(w_m, w_l) + \epsilon}{p(w_l)}) \quad (4)$$

This metric uses the count of occurrences in the documents to measure coherence. Words should be ordered and evaluated in decreasing probability order. This measure is not symmetric and assumes that word order is important where a word is compared with the preceding word. Where the probability of seeing word w_m and word w_l together is divided by the probability of $p(w_m)$ and ϵ acts as a smoothing variable.

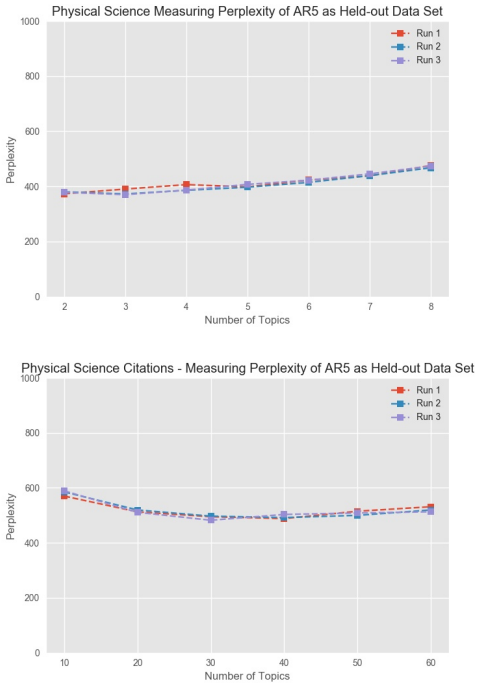

Figure 4: Understanding Effects of Changing the Number of Topics on the Physical Science Topics (top) and Citations (bottom).

B. Comparing Different Sizes for K

In the following experiments, perplexity (lower is better) is shown as a function of the number of topics K for both reports and citations. By visually inspecting topics and using perplexity, the best K for reports and citations was obtained. The perplexity scores were generally consistent with visual inspections. The perplexity had a tendency to reflect data collection size differences, as a larger collection might reach a higher K before perplexity increases. An example of these difference is shown when comparing the perplexity of reports and citations as shown in Figure 4. Reports typically had better perplexity with a slightly lower K value, as there are fewer documents to represent. However, the citation topic model had a tendency to reach a higher K before perplexity increased.

These experiments exposed an interesting weakness in the DTM approach for scientific research. Since the models in time slices $t + \{1...n\}$ were constrained by the topics that are generated in time slice t, if the document set tends to grow over time slices, that growth will be hard to model. If the number of latent topics grow over time, adjusting the model to support this growth is not possible.

C. Comparing Different Values for Variance

In this experiment, the behavior of changing the variance parameter in DTM was explored. In Table I, the effects of changing variance from .05 to .15 to .5, for the *Physical Science* assessment reports are shown. The top 10 most probable words change from assessment to assessment in

Table I: The Effect of Variance on Topic Evolution - Physical Science.

vari-ance	*AR1*	*AR2*	*AR3*	*AR4*	*AR5*
0.05	carbon dioxide, ocean, concentration, atmosphere, water, ecosystem, soil, surface, plant, effect	ocean, carbon dioxide, atmosphere, effect, surface, atmospheric, response, water, concentration, greenhouse gases	carbon dioxide, anthropogenic, atmospheric co2, aerosol, emission, ocean, carbon cycle, greenhouse gases, atmosphere, concentration	carbon dioxide, atmospheric co2, carbon cycle, anthropogenic, ocean, emission, temperature, effect, atmosphere, concentration	carbon cycle, atmospheric co2, carbon dioxide, anthropogenic, land use, methane, emission, global, fossil fuel, atmosphere,
0.15	carbon dioxide, ocean, concentration, ecosystem, temperature, soil, plant, atmosphere, global, water	sea level rise, ocean, effect, global, carbon dioxide, response, atmosphere, surface, atmospheric, water	carbon dioxide, sea level rise, anthropogenic, atmospheric co2, aerosol, emission, ocean, carbon cycle, greenhouse gases, effect	carbon dioxide, atmospheric co2, carbon cycle, temperature, ocean, anthropogenic, emission, sea level rise, concentration, effect	carbon cycle, atmospheric co2, carbon dioxide, sea level rise, anthropogenic, land use, methane, global, emission, scenario
0.5	carbon dioxide, concentration, ocean, plant, soil, ecosystem, temperature, atmosphere, methane, emission	ocean, effect, carbon dioxide, atmosphere, surface, response, atmospheric, ecosystem, system, marine	carbon dioxide, anthropogenic, atmospheric co2, aerosol, emission, carbon cycle, ocean, greenhouse gases, ppb, concentration	carbon dioxide, carbon cycle, temperature, atmospheric co2, anthropogenic, emission, ocean, effect, aerosol, flux	carbon cycle, atmospheric co2, carbon dioxide, anthropogenic, land use, methane, emission, global, fossil fuel, nitrogen

Table II: Measuring Coherence Given Different Variance Averaged Over Assessments - Physical Science.

	Variance	*5 Topics*	*10 Topics*	*20 Topics*
Top 5 Words	.05	-0.33	-0.36	-0.43
	.15	-0.42	-0.50	-0.44
	.50	-0.39	-0.41	-0.47
Top 10 Words	.05	-0.38	-0.38	-0.40
	.15	-0.37	-0.48	-0.45
	.50	-0.40	-0.44	-0.48
Top 20 Words	.05	-0.42	-0.45	-0.49
	.15	-0.48	-0.50	-0.49
	.50	-0.45	-0.46	-0.47

a gradual way, given a variance of .05. However, when the variance is changed to .50, there are more abrupt changes among the top 10 words.

Figure 5 shows that a .05 variance on average had a higher UMASS coherence score given changes in the number of topics and the number of top words used. The raw numbers are shown in Table II. In this particular example, the number of topics at 5-10 topics was most useful for the cross-domain method. Though the UMASS coherence measure was slightly better when variance was at .05, the gradual change given 5 discrete time slices did not provide enough effect to understand how the reports were evolving. In general, for smaller data sets (the reports), we found topic evolution benefited from the variance value being increased, in some cases we increased to .5 variance. For example, with the topic models for the Impact reports we used a .5 variance. For the citation data sets, when the variance was increased to more than .05, the topics were visually harder to interpret and we also saw the lower coherence scores. Hence a smaller variance was preferred.

D. Comparing Subsections and Chapter

In the above described experiments, chapters were treated as documents to construct the topic models. In this experiment, chapters were further segmented into subsections based on subsection headings. A topic model was built treating the subsections as documents. This was compared to a topic model built using chapters as documents. Coherence and perplexity were measured to understand the behavior between these two approaches. Perplexity in both models was calculated by using a DTM that included AR1-AR4 for which AR4 for training the model and AR5 as the held-out data set. These experiment were performed using the *Physical Science* book.

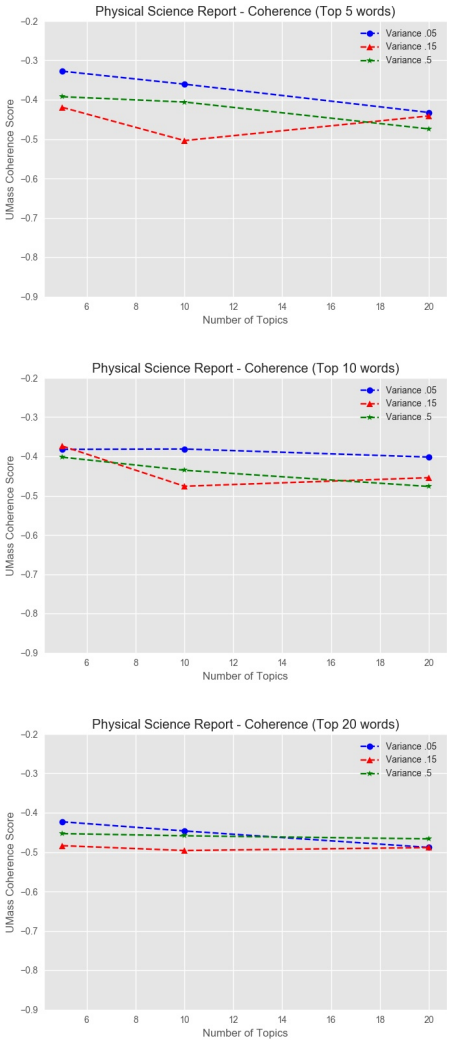

Figure 5: Understanding the Effects of Changing Variance Using Top 5, 10 and 20 Words Averaged Over Assessments - Physical Science.

When observing perplexity as a function of the number of topics, as shown in Figure 6, perplexity had an upward trend almost immediately. Subsection extraction is challenging to perform without error. Figure 7 shows that coherence appeared to improve with the documents as chapters rather than the subsections. This was consistent for top n where n

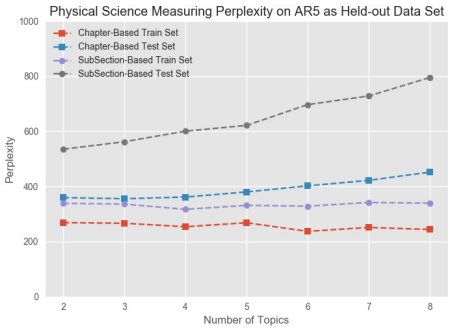

Figure 6: Comparing Physical Science Chapters and Physical Science Subsections as Documents Perplexity

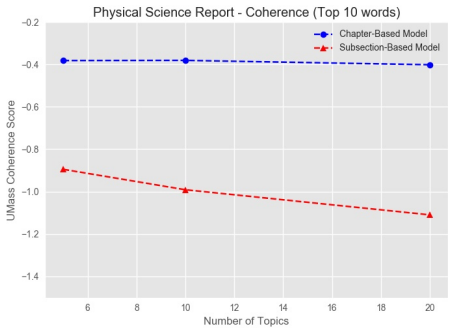

Figure 7: Comparing Physical Science Chapters and Physical Science Subsections as Documents - UMASS Coherence - Top 10 Words

was varied from five to 20.

When evaluating a sample of the subsections it was found that only 90% of the subsections were correctly captured. This would have an effect on the coherence and perplexity. This may explain the results related to perplexity and coherence. Let the reader take note that chapters were used as the topic model documents for the rest of the work described.

E. Cross-Domain Relatedness and Influence

Cross-domain mappings were discovered between the Physical Science reports and Physical Science citations. A low divergence mapping between topics across domains is shown in Table III.

From this mapping of topic pairs, the following related chapters across assessments were partitioned together:

- AR1 Ch. 1 (Greenhouse Gases and Aerosols)
- AR1 Ch. 10 (Effects on Ecosystems)
- AR2 Ch. 1 (The Climate System: an Overview)
- AR2 Ch. 9 (Terrestrial Biotic Responses to Environmental Change and Feedbacks to Climate)
- AR2 Ch. 10 (Marine Biotic Responses to Environmental Change and Feedbacks to Climate)
- AR2 Ch. 11 (Advancing our Understanding)

Table III: A Physical Science Report and Citations Topic Pair after Divergence.

	AR1	*AR2*	*AR3*	*AR4*	*AR5*
Citations Model	radiative, radiative forcing, band, ipcc, aerosol, anthropogenic, temperature, effect, carbon dioxide, estimate	radiative, radiative forcing, ipcc, aerosol, anthropogenic, effect, temperature, estimate, greenhouse gases, carbon dioxide	radiative, radiative forcing, aerosol, ipcc, anthropogenic, effect, estimate, greenhouse gases, temperature, carbon dioxide	radiative, radiative forcing, aerosol, anthropogenic, estimate, ipcc, effect, temperature, greenhouse gases, carbon dioxide	radiative, radiative forcing, aerosol, ipcc, anthropogenic, effect, estimate, temperature, greenhouse gases, carbon dioxide
Report Model	radiative, radiative forcing, effect, greenhouse gases, carbon dioxide, emission, concentration, ozone, aerosol, solar	radiative, radiative forcing, ipcc, aerosol, ozone, emission, methane, concentration, carbon dioxide, anthropogenic	radiative, radiative forcing, aerosol, ipcc, effect, temperature, estimate, carbon dioxide, greenhouse gases, cloud	radiative, radiative forcing, anthropogenic, aerosol, cloud, carbon dioxide, emission, effect, estimate, temperature	radiative, anthropogenic, radiative forcing, aerosol, cloud, temperature, effect, carbon dioxide, emission, estimate

Table IV: An Impact Report and Citations Topic Pair after Divergence.

	AR1	*AR2*	*AR3*	*AR4*	*AR5*
Citations Model	adaptation, risk, impact, vulnerability, strategy, coastal, action, adaptive, development, capacity	adaptation, vulnerability, the united nations framework convention on climate change, impact, management, resource, risk, strategy, action, ecosystem	adaptation, management, sea level rise, ecosystem, national, biodiversity, resource, specie, option, coral reef	adaptation, biodiversity, forestry, management, ecosystem, ipcc, specie, national, sea level rise, impact	adaptation, biodiversity, management, national, resource, sea level rise, ecosystem, food security, impact, the united nations framework convention on climate change
Report Model	adaptation, impact, method, mitigation, option, development, policy, economic, measure, information	adaptation, impact, method, option, mitigation, development, policy, economic, measure, cost	adaptation, development, capacity, adaptive, cost, impact, vulnerability, policy, method, economic	adaptation, mitigation, development, policy, capacity, cost, adaptive, decision, economic, sustainable	adaptation, decision, policy, planning, action, option, local, making, cost, capacity

- AR3 Ch. 3 (The Carbon Cycle and Atmospheric Carbon Dioxide)
- AR3 Ch. 4 (Atmospheric Chemistry and Greenhouse Gases)
- AR3 Ch. 5 (Aerosols, their Direct and Indirect Effects)
- AR3 Ch. 14 (Advancing Our Understanding)
- AR4 Ch. 1 (Historical Overview of Climate Change Science)
- AR4 Ch. 7 (Coupling Between Changes in the Climate System and Biogeochemistry)
- AR5 Ch. 6 (Carbon and Other Biogeochemical Cycles)

Over 140 research papers across assessments were found, the majority of which come directly from the cited chapters.

Cross-domain mappings were discovered between the Impact reports and Physical Science citations. A low divergence mapping between topics across domains is shown in Table IV. Additional results are shown in [17], [18], [19]. For example, in Table V research papers based on the Physical Science citation domain, were correlated with certain Impact report chapters.

VII. Observations and Conclusions

Identifying relatedness across domains across time slices was consistent with chapter descriptions from our use case

Table V: Cross Domain Document Cluster Results: Physical Science Citations Domain and Impact Report Domain.

Domain 1 Physical Science Citations		
AR		*Title*
3		Timing and duration of the Last Interglacial: evidence for a restricted interval of widespread coral reef growth. [20]
4		Timing and duration of the Last Interglacial: evidence for a restricted interval of widespread coral reef growth. [20]
Domain 2 Impact Reports		
AR	*Chapter*	*Title*
1	6	World oceans and coastal zones
2	9	Coastal Zones and Small Islands
3	6	Coastal Zones and Marine Ecosystems
3	17	Small Island States
4	6	Coastal systems and low-lying areas
4	16	Small islands
5	5	Coastal Systems and Low-Lying Areas
5	29	Small islands

study. Identifying influence was measured by identifying citations from the citation model that were cited in the report chapters for Physical Science. Cross-domain influence was more challenging when identifying influence of Physical Science research on Impact assessment reports. This is due to the fact that chapters are not as correlated since the books have different objectives. Constraining a topic model by using domain specific concepts reduces the likelihood of having topics that are noisy and not relevant. However, duplication is more common is a concept-constrained model as a side-effect of constraining the topic model.

The variance parameter in DTM has a measurable impact on the quality of the topics and understanding the evolution of a domain. From our experiments, increasing variance for smaller document collection topic models and decreasing variance for larger document collection topic models yielded the best models for cross-domain analysis. In some cases, this might be less obvious when measuring coherence using the UMASS coherence measure. However, we found from visual inspection of topics, for smaller data sets (the reports), topic evolution benefited from the variance value being increased. For larger data sets, such as the citations, smaller variance was preferred.

VIII. Acknowledgments

This work was partially supported by NSF award #1439663 and a gift from IBM.

References

[1] C. P. Kelley, S. Mohtadi, M. A. Cane, R. Seager, and Y. Kushnir, "Climate change in the fertile crescent and implications of the recent syrian drought," *Proceedings of the National Academy of Sciences*, vol. 112, no. 11, pp. 3241–3246, 2015.

[2] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991.

[3] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *ICML*. ACM, 2006, pp. 113–120.

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

[5] G. Welch and G. Bishop, "An introduction to the Kalman filter," U. of North Carolina at Chapel Hill, Tech. Rep., 1995.

[6] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised prediction of citation influences," in *Proc. 24th Int. Conf. on Machine learning*. ACM, 2007, pp. 233–240.

[7] L. Li, X. Jin, and M. Long, "Topic correlation analysis for cross-domain text classification." in *AAAI*, 2012.

[8] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *EMNLP*. ACL, 2008, pp. 363–371.

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

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

[11] C. Tang and C. Monteleoni, "Can topic modeling shed light on climate extremes?" *Computing in Science & Engineering*, vol. 17, no. 6, pp. 43–52, 2015.

[12] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *SIGKDD*. ACM, 2006, pp. 424–433.

[13] D. M. Blei and J. D. Lafferty, "Dynamic topic models," http://www.cs.princeton.edu/ blei/lda-c/.

[14] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[15] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," in *Proc. 24th Conf. Uncertainty in Artificial Intelligence*, 2008.

[16] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *EMNLP*. ACL, 2011, pp. 262–272.

[17] J. Sleeman, M. Halem, T. Finin, and M. Cane, "Dynamic topic modeling to infer the influence of research citations on ipcc assessment reports," in *Big Data Challenges, Research, and Technologies in the Earth and Planetary Sciences Workshop, IEEE Int. Conf. on Big Data*. IEEE, 2016.

[18] ——, "Modeling the evolution of climate change assessment research using dynamic topic models and cross-domain divergence maps," in *AAAI Spring Symposium on AI for Social Good*. AAAI Press, 2017.

[19] J. A. Sleeman, "Dynamic data assimilation for topic modeling (ddatm)," Ph.D. dissertation, UMBC, 2016.

[20] C. Stirling, T. Esat, K. Lambeck, and M. McCulloch, "Timing and duration of the last interglacial: evidence for a restricted interval of widespread coral reef growth," *Earth and Planetary Science Letters*, vol. 160, no. 3, pp. 745–762, 1998.