

Hidden Stories: Topic Modeling in Hydrology Literature

Mashrekur Rahman¹, Grey Nearing², and Jonathan Frame¹

¹University of Alabama

²NASA Goddard Space Flight Center

November 21, 2022

Abstract

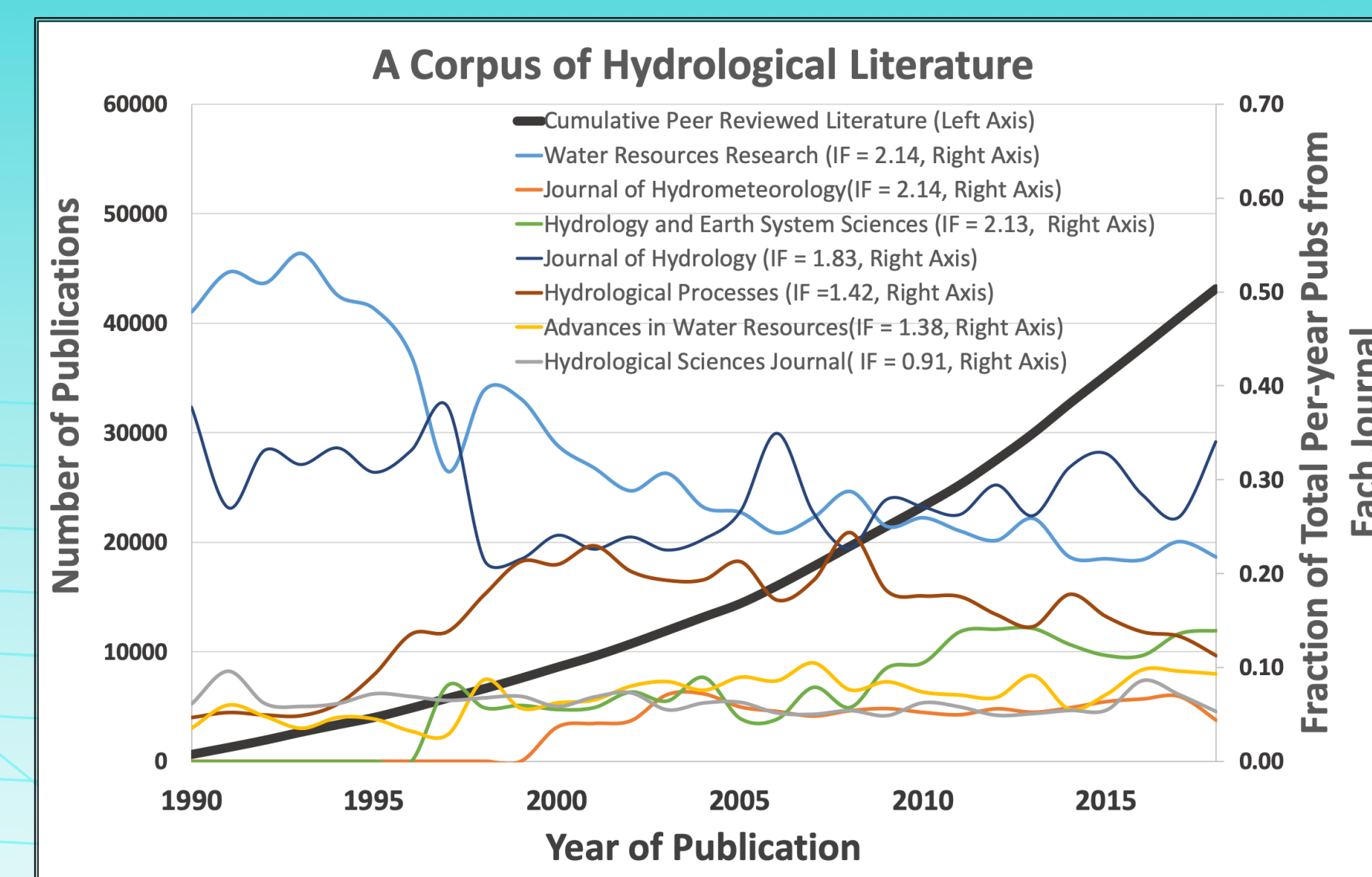
Recent advancement of computational linguistics, machine learning, including a variety of toolboxes for Natural Language Processing (NLP), help facilitate analysis of vast electronic corpuses for a multitude of objectives. Research papers published as electronic text files in different journals offer windows into trending topics and developments, and NLP allows us to extract information and insight about these trends. This project applies Latent Dirichlet Allocation (LDA) Topic Modeling for bibliometric analyses of all abstracts in selected high-impact (Impact Factor > 0.9) journals in hydrology. Topic modeling uses statistical algorithms to extract semantic information from a collection of texts and has become an emerging quantitative method to assess substantial textual data. The resulting generated topics are interpretable based on our prior knowledge of hydrology and related sub-disciplines. Comparative topic trend, term, and document level cluster analyses based on different time periods was performed. These analyses revealed topics such as climate change research gaining popularity in Hydrology over the last decade. An inter-topic correlation analysis also revealed the nature of information exchange and absorption between various communities within the hydrology domain. The primary objective of this work is to allow researchers to explore new branches and connections in the Hydrology literature, and to facilitate comprehensive and inclusive literature reviews. We aim to use these results combined with probability distribution between topics, journals and authors to create an ontology that is useful for scientists and environmental consultants for exploring relevant literature based on topics and topic relationships.

Hidden Stories: Topic Modeling in Hydrology Literature

Mashrekur Rahman, Jonathan M. Frame, Grey S. Nearing; University of Alabama – Department of Geological Sciences

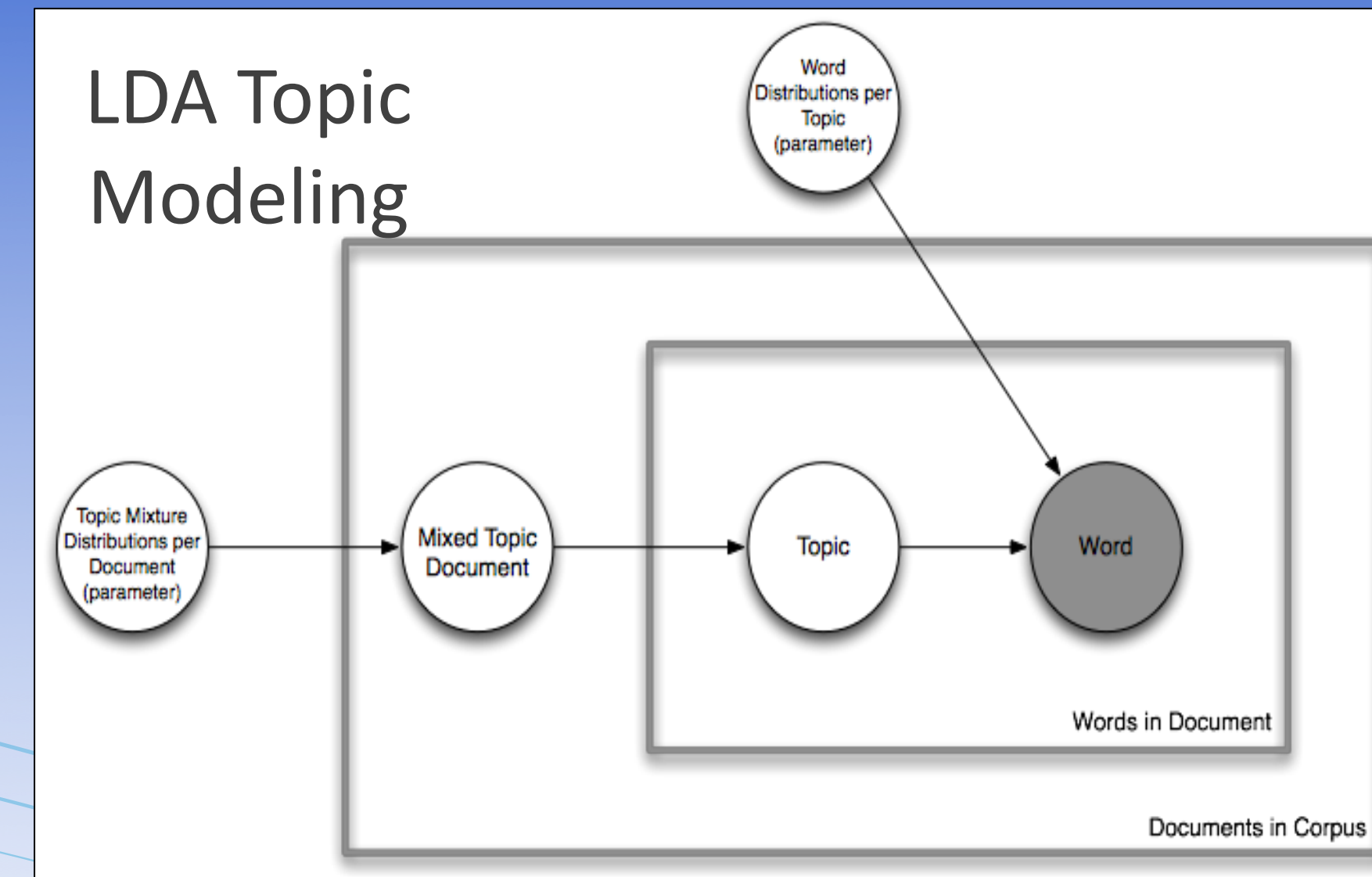


DATA



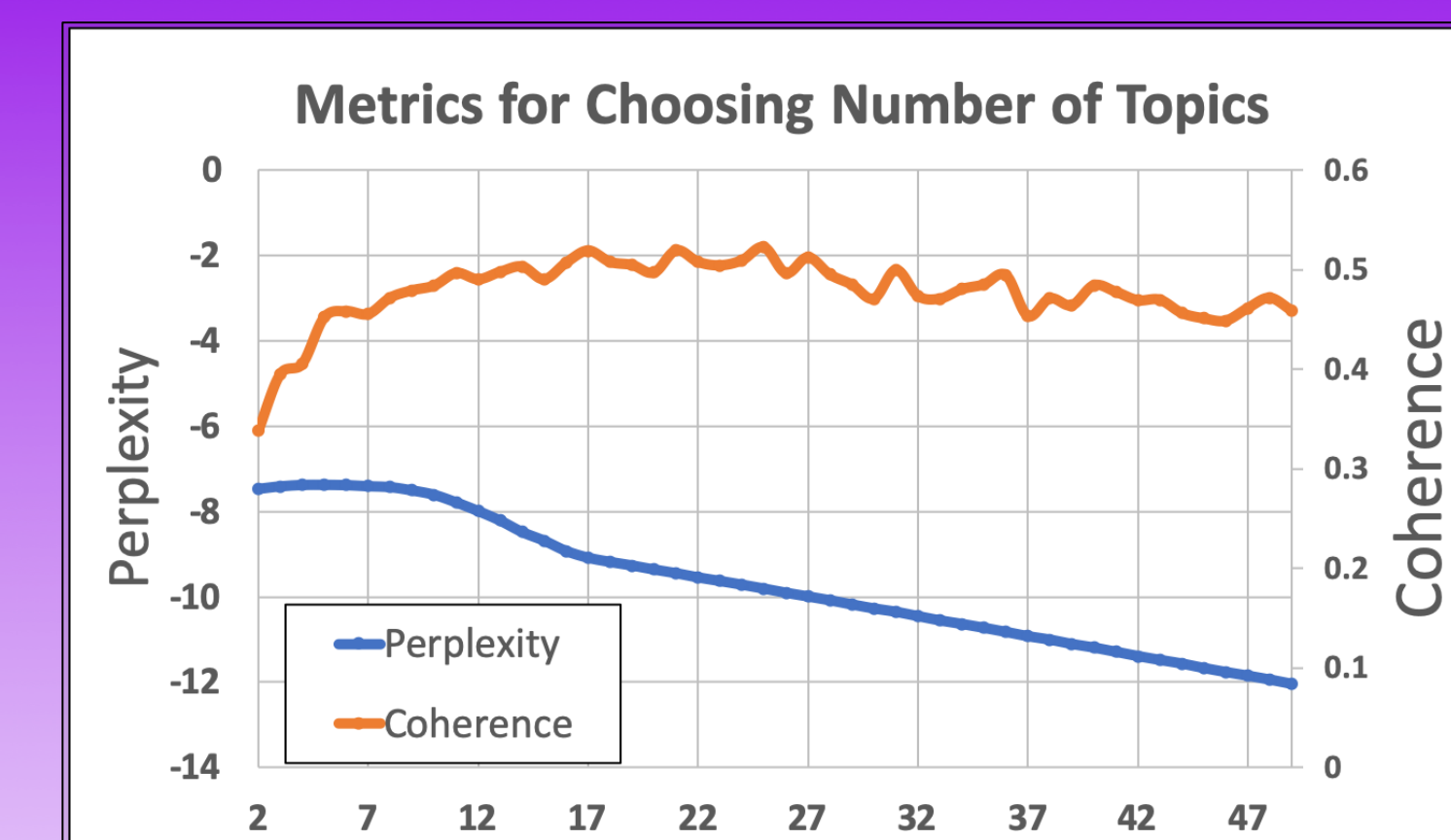
Our corpus consists of abstracts from all articles (43,000+) published in seven major hydrology journals since 1991.

MODEL



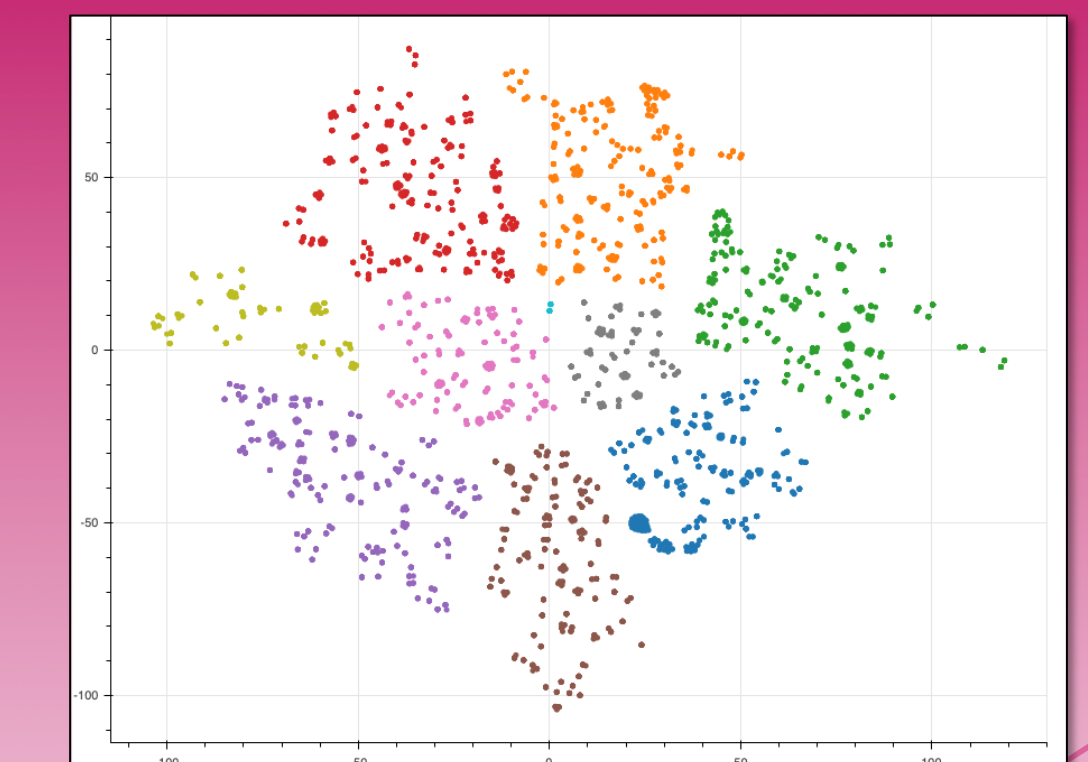
Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collection of discrete data. Unsupervised learning with LDA was used to identify dominant topics in the corpus and to allocate papers to each topic.

NUMBER OF TOPICS



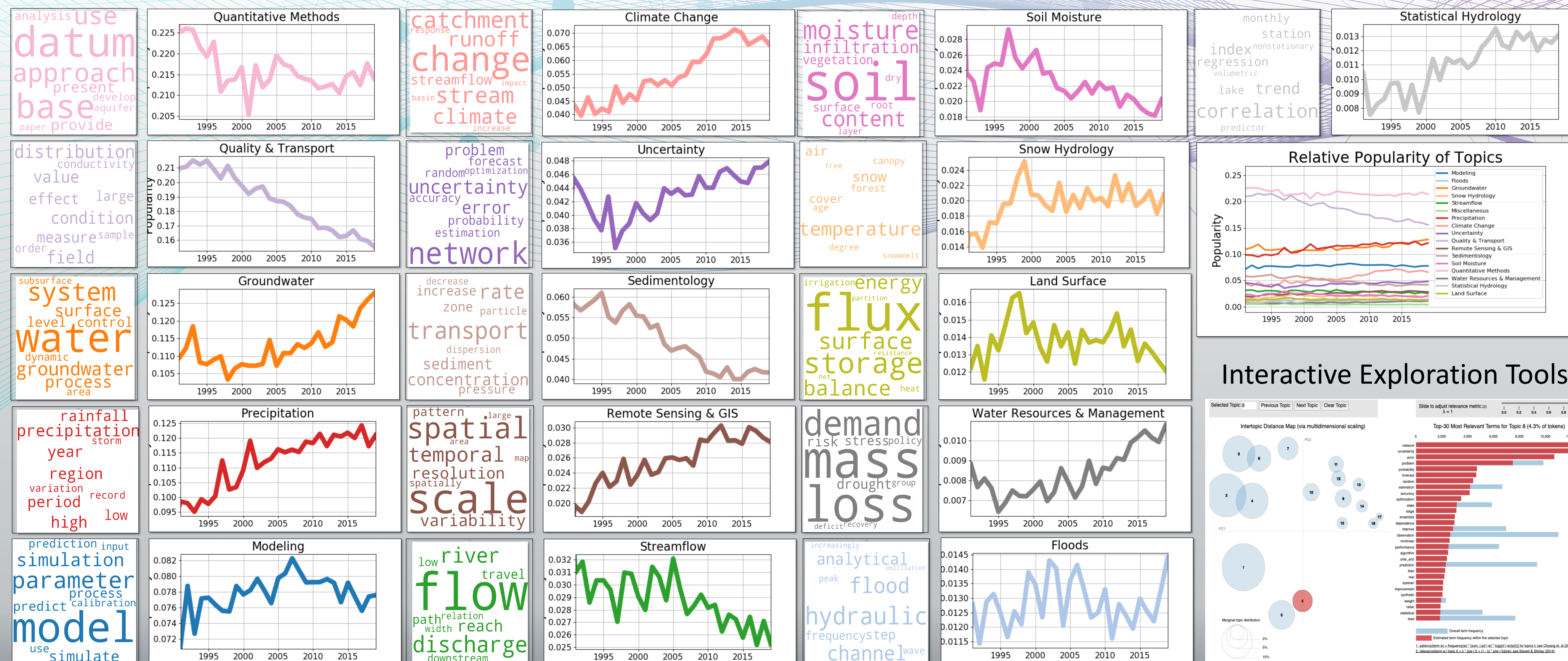
Perplexity and coherence scores were used to identify an optimal number of topics. Perplexity is related to the classification likelihood of a holdout set, and coherence is related to how often words in the same topic appear together. We chose 17 topics, based on inflection points in the two measures.

FUTURE DIRECTIONS

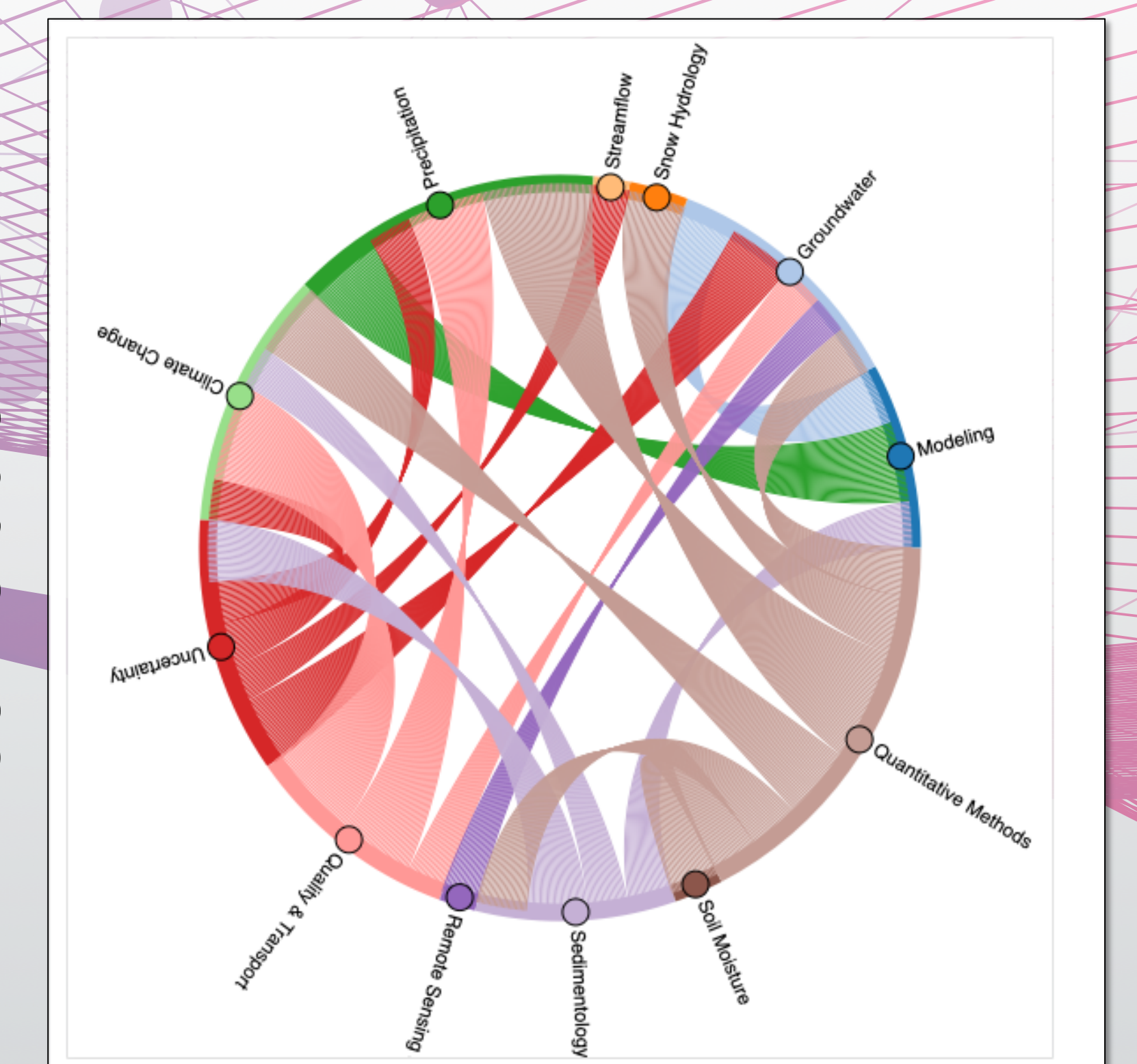


We are working toward a web-based tool to allow for topic-guided exploration and hydrology literature. If you read a paper you like, then it should be easy to find other papers that are related along different topical dimensions. We would also like to analyze trends in research topics, crossovers, and collaborations.

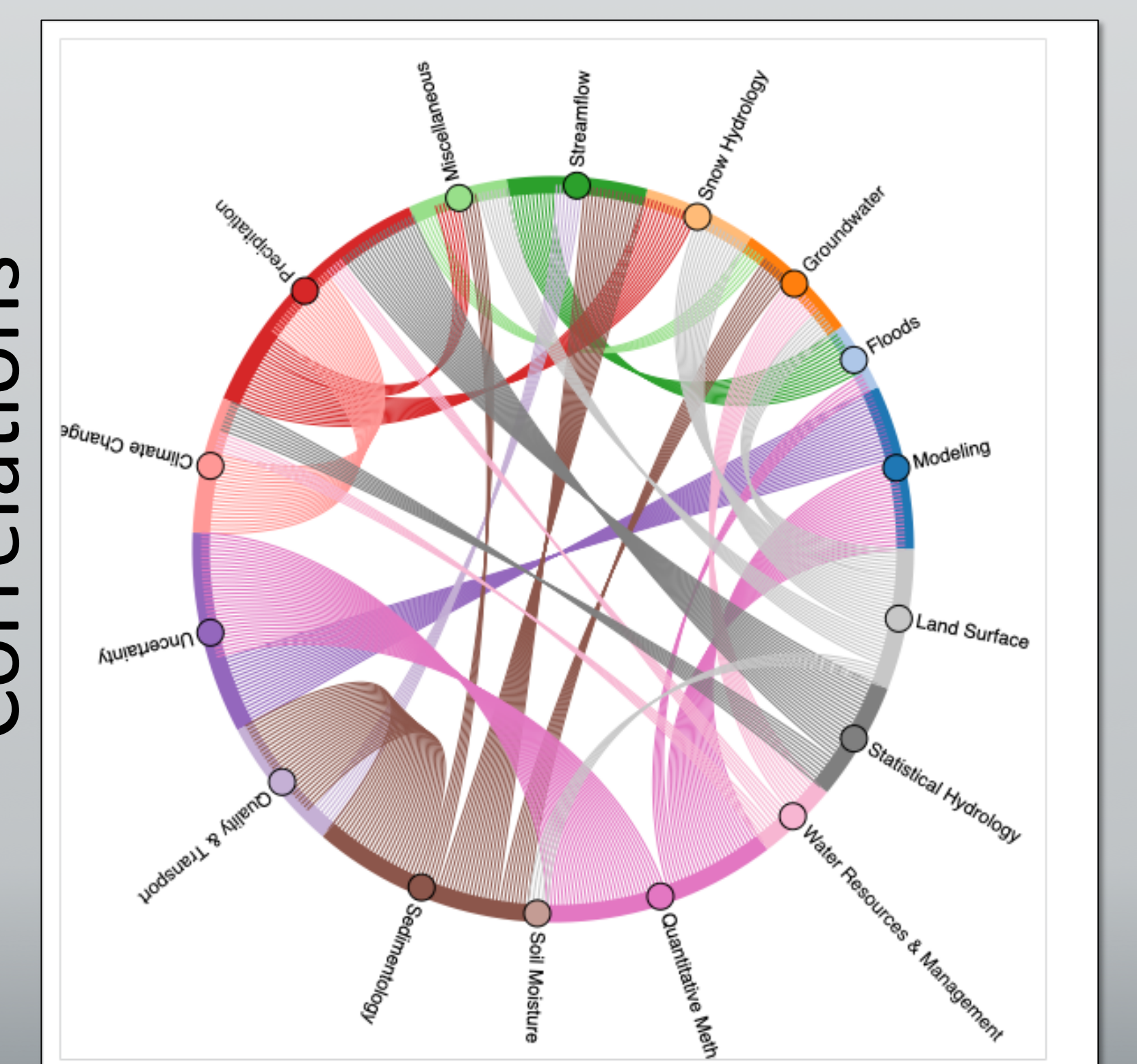
RESULTS: TOPICS & TRENDS



Negative Inter-Topic Correlations



Positive Inter-Topic Correlations



Interactive Exploration Tools

