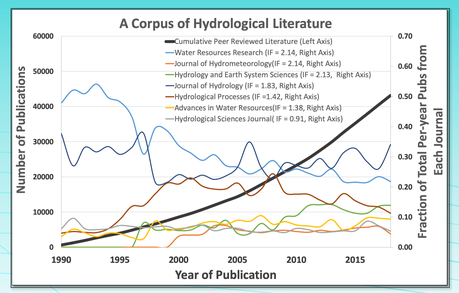
Hidden Stories: Topic Modeling in Hydrology Literature

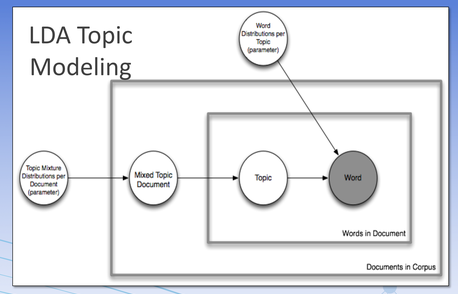
Mashrekur Rahman, Jonathan M. Frame, Grey S. Nearing; University of Alabama – Department of Geological Sciences

DATA



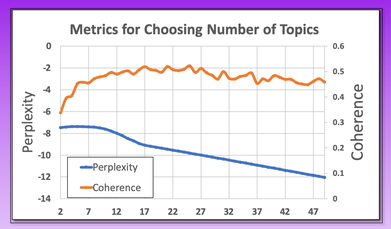
Our corpus consists of abstracts from all articles (43,000+) published in seven major hydrology journals since 1991.

MODEL



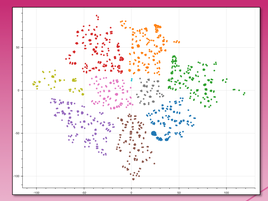
Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collection of discrete data. Unsupervised learning with LDA was used to identify dominant topics in the corpus and to allocate papers to each topic.

NUMBER OF TOPICS



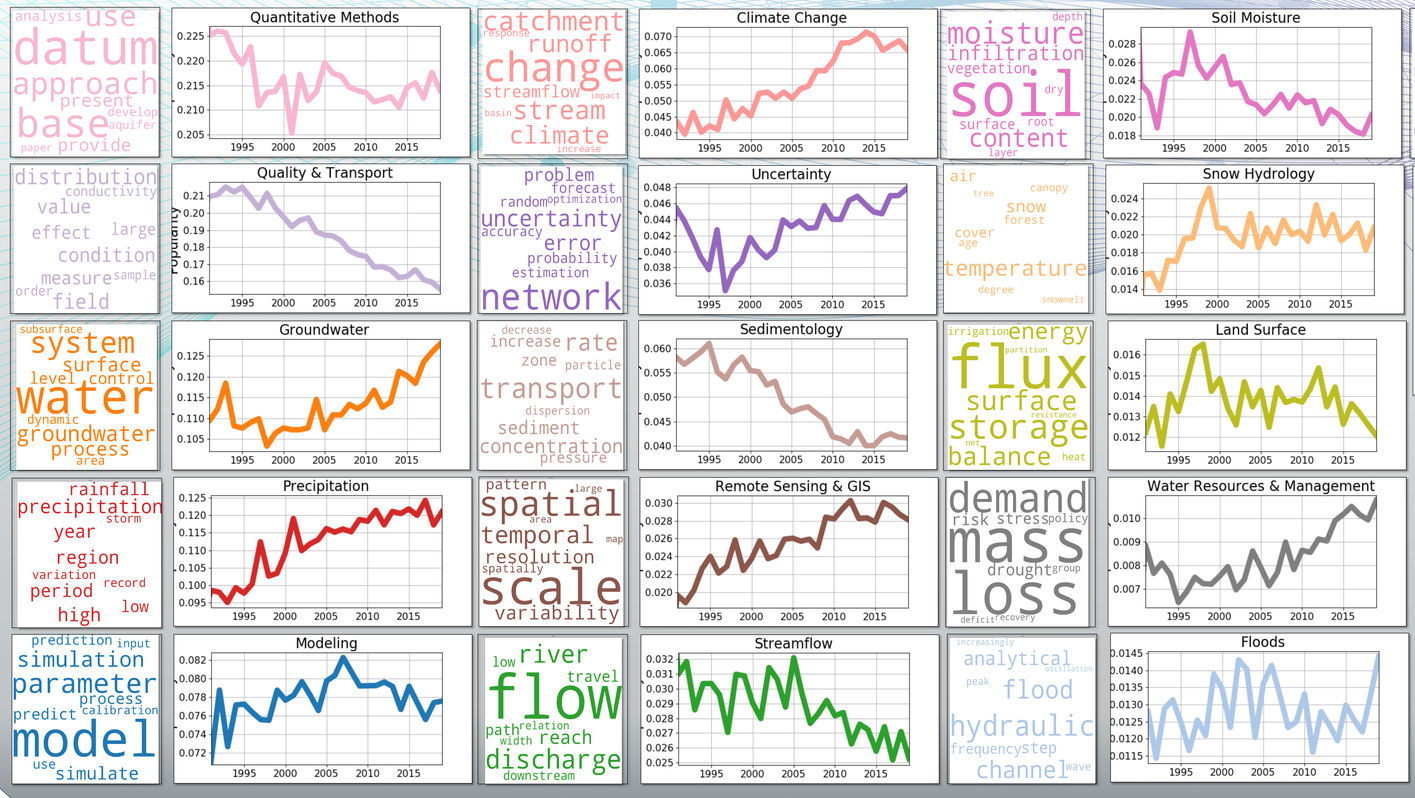
Perplexity and coherence scores were used to identify an optimal number of topics. Perplexity is related to the classification likelihood of a holdout set, and coherence is related to how often words in the same topic appear together. We chose 17 topics, based on inflection points in the two measures.

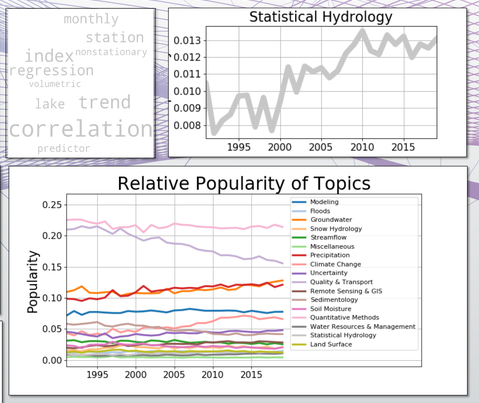
FUTURE DIRECTIONS



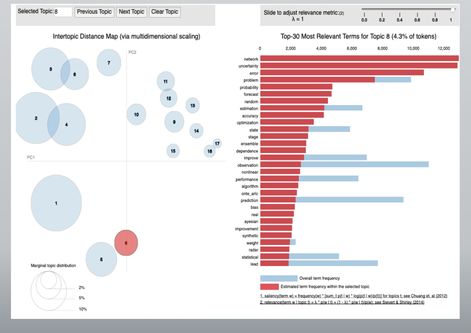
We are working toward a web-based tool to allow for topic-guided exploration and hydrology literature. If you read a paper you like, then it should be easy to find other papers that are related along different topical dimensions. We would also like to analyze trends in research topics, crossovers, and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

RESULTS: TOPICS & TRENDS





Interactive Exploration Tools



Negative Inter-Topic Correlations

Positive Inter-Topic Correlations