From Texts to Pixels: Leveraging Artificial Intelligence to Achieve Novel Insights into Hydrologic Research, Human-Drought Interactions, and Global Drought Prediction

By

MASHREKUR RAHMAN

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Hydrologic Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Thomas Harter, Chair

Samuel Sandoval Solis

Grey S Nearing

Committee in Charge

2024

To
my Mother, Mariam Begum
my Father, AKM Mostafizur Rahman
& my role model, Dr. Janice L. Kirsch

Contents

List of Figures

Overview - Hydrology in a World of Artificial Intelligence

Hydrology has historically lagged in adopting and advancing interdisciplinary knowledge, including Machine Learning (ML) research and applications. However, this trend is gradually changing as the hydrologic community begins to increasingly recognize the immense potential of ML in addressing complex water-related challenges. The increasing availability of big data from remote sensing, climate models, in-situ measurements, and human-generated texts, images, sounds coupled with the growing computational capabilities, has opened up increasingly new avenues for ML-driven hydrologic research.

The advancement of ML in hydrology is not just a matter of academic interest but an essential step towards tackling the pressing water issues facing California, and our precious planet's environment. Climate change coupled with unsustainable, and often misinformed resources management is putting unprecedented pressure on our water resources, necessitating innovative solutions that can help us better understand, predict, and manage hydrologic systems. Traditional hydrologic models, while valuable, often struggle to capture the nonstationarity and nonlinearities inherent in these systems, particularly when it comes to understanding the complex interactions between human behavior and evolving environmental conditions. ML, with its ability to learn from data and uncover hidden patterns, offers a promising complementary approach to address these challenges.

In my PhD research, I have focused on leveraging ML to address three key problems in hydrology: (1) understanding the evolution of research topics and interdisciplinarity in hydrologic sciences using Natural Language Processing (NLP), (2) exploring the complex interactions between droughts and human awareness by applying Computer Vision and NLP to remotely sensed and human-generated data, and (3) developing a global drought prediction model using a novel Computer Vision technique. I demonstrated the capability and versatility of ML in advancing hydrologic knowledge, capturing anthropogenic behaviors, and informing

real-world decision-making. My research not only pushes the boundaries of what is possible with ML in hydrology but also lays the groundwork for future studies in this exciting interdisciplinary field.

However, the impact of my research extends beyond the academic realm. To ensure that our endeavors benefit society most, we must keep fostering strong collaborations between hydrologists, ML experts, and stakeholders like water managers, policymakers, and local communities. This involves developing ML tools that are not only scientifically rigorous but also user-friendly, interpretable, and adaptable to different contexts. Although not contained in this dissertation, during my doctoral pursuit, I have engaged in multiple other collaborative research projects which leveraged ML for tasks such as post-processing the US National Water Model, estimating inputs to the Rangeland Hydrology & Erosion Model, and an educational module for hydrologists for building ML intuition. We need to increasingly invest in ML education and capacity building within the hydrologic community, empowering researchers and practitioners with the skills and knowledge to harness these powerful techniques effectively and ethically.

As the field of ML continues to evolve at a rapid pace, with breakthroughs like ChatGPT, Claude 3, Bard, DALL-E, Sora, etc. showcasing the unprecedented potential of Generative Artificial Intelligence, and the possibility of Artificial General Intelligence (AGI) in the horizon, the hydrology community must also adapt and position itself to leverage these advancements. AGI may trigger unprecedented rates of technological advancements with profound implications for hydrology. It could lead to the development of hyper-intelligent systems capable of autonomously managing water resources, predicting and mitigating natural disasters, and optimizing water infrastructure. However, it also raises important questions about the role of human expertise, the transparency and accountability of AI systems, and the potential for unintended consequences. As we approach this uncharted territory, it is crucial

that hydrologists engage in proactive discussions and collaborations to ensure that the development and deployment of these technologies align with the principles of sustainability, equity, and societal well-being.

Finally, as we navigate the rapidly evolving landscape of ML in hydrology, it is crucial that we prioritize scientific integrity and societal benefit over narrow interests or contractual obligations. The models and techniques we develop and adopt should be chosen based on their reliability, robustness, realism, and generalizability, rather than on biased considerations. We stand at an exciting juncture in the history of hydrologic sciences with unparalleled opportunities for discovery, innovation, and societal impact.

Abstract

This dissertation presents three interconnected studies that leverage advanced computational techniques, including Natural Language Processing (NLP), Computer Vision, Machine Learning, and Big Data analytics to gain insights into various aspects of hydrologic sciences and drought research.

In the first study, we applied NLP to assess topic diversity in approximately 75,000 research articles from eighteen water science and hydrology journals published between 1991 and 2019. We found that individual water science and hydrology research articles are becoming increasingly diverse in the sense that, on average, the number of topics represented in individual articles is increasing, which may be a sign of increasing interdisciplinarity. This is true even though the body of water science and hydrology literature as a whole is *not* becoming more topically diverse. Topics with the largest increases in popularity were *Climate Change Impacts*, *Water Policy & Planning*, and *Pollutant Removal*. Topics with the largest decreases in popularity were *Stochastic Models* and *Numerical Models*. At a journal level, *Water Resources Research*, *Journal of Hydrology*, and *Hydrological Processes* are the three most topically diverse journals among the corpus that we studied.

The second study focused on understanding the relationship between droughts and drought awareness, which is crucial for decision-making, policy development, and socioeconomic outcomes related to water management and conservation strategies. We used computer vision (UNet models) to analyze nonlinear, lagged correlations between Standardized Precipitation Evapotranspiration Index (SPEI) and Google Trends Search Interest within the Continental United States (CONUS). We also used Twitter data to asses people's sentiments about droughts. The most important drivers of this relationship are the variability and ranges of drought trends and severity, as well as climatic extremes. This relationship was the strongest for Western states, followed by Northeastern, Southeastern, and Central regions. Search interest tends to lag droughts by a period of ˜1-3 months. We also found evidence that reductionist linear approaches, such as Principal Component Analysis, might not be

as effective as UNet models in capturing the nuanced relationship between droughts and drought awareness at various dimensions and scales. We subsequently applied sentiment analysis on a set of 2.5 million georeferenced tweets related to droughts and found that people's sentiments towards drought have become increasingly positive with decreasing neutral sentiments since 2014 within the United States.

In the third study, we propose a novel approach for global drought prediction using the Vision Transformer (ViT) model, leveraging its ability to contextually learn spatial and temporal patterns from high-dimensional climate data. Using a sliding window approach, we trained the ViT model on a global dataset spanning from January 1970 to December 2004, using Sea Surface Temperature (SST), 2-meter Air Temperature (T2M), and Total Precipitation (TP) as input variables, and the Standardized Precipitation Evapotranspiration Index (SPEI) (looking ahead 0, 1, and 2 months) as the target variable. The model's performance is evaluated on a test dataset from January 2005 to December 2020 using accuracy, precision, recall, and F1 score metrics. Our results demonstrate the ViT model's effectiveness in predicting drought occurrences, with high accuracy scores ranging from 0.9456 to 0.9475 and precision scores from 0.8747 to 0.8781 for a three-month prediction horizon. The model's relatively lower recall scores (0.6285 to 0.6465) indicate room for improvement in capturing all drought occurrences, particularly in regions with complex or sporadic drought patterns. The findings of this study indicate substantial potential of the ViT model in predicting increasingly complex meteorological drought occurrences on a global scale.

Collectively, these studies contribute to the advancement of hydrologic sciences by providing operational tools and insights for researchers, policymakers, and all stakeholders in the field of water resources science and management.

Acknowledgments

As I pause to reflect on this incredible journey that culminates in these pages, my heart is filled with gratitude towards the wonderful people whose contributions, wisdom, and kindness have been instrumental towards my successes.

I would like to begin with my greatest appreciation for Dr. Grey Nearing, who has advised and supported me through thick and thin since I first arrived in the United States. Dr. Nearing's scientific accomplishments motivated and inspired me to follow his footsteps and delve deeper into the fascinating world of Artificial Intelligence in Hydrology. I am grateful for all that he has done for my career and will forever be a proud student of his.

I would also like to convey my heartfelt appreciation for Prof. Thomas Harter. His compassionate advising and unwavering wisdom were crucial towards navigating the hurdles in my years of pursuit.

I sincerely thank Prof. Samuel Sandoval Solis for mentoring me in my journey. As a fellow Star Wars fan, it has always been a pleasure to talk to him. From a fellow Mandalorian, "This is the way!"

I appreciate Prof. Mark Lubell, Prof. Gregory Pasternack, Prof. Jimmy Lin (University of Waterloo), Prof. Geoffrey Tick (University of Alabama), Prof. Fred Andrus (University of Alabama), and Dr. Craig Pelissier (NASA) for their mentorship and contributions in various stages of my journey.

I am grateful to all the co-authors and contributors of my primary dissertation projects for their time and inputs. Special thanks to Dr. Mahmoud Saeedimoghaddam and Dr. Jonathan Frame for collaborating with me in multiple projects.

I would like to acknowledge the National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), United States Department of Agriculture (USDA), Google Research Credits Program, American Geophysical Union, and the Hydrologic Sciences Graduate Group for funding my research and travel needs at various stages.

On a personal note, I am here today because of the incredible sacrifices and hard work of my parents, who put my well-being and education above everything else. Thank you for loving me unconditionally and being the greatest Ammu (Mom) and Abbu (Dad) any child can ever ask for.

In my toughest of times in this journey, I have had the incredible support of Dr. Janice L Kirsch of Berkeley, California. No amount of gratitude will be enough to express my gratefulness for what she has done for me and my family. She is the most munificent and magnanimous personality I have come across in my life, and I aspire to be like her.

I am thankful to all my friends, family members, acquaintances, and everyone else who I have come across whose kindness and companionship have helped me cope with the challenges of life. Special shout-out to my fellow soccer, badminton, and tennis buddies over the years.

Last, but not least, I would like to express my love and appreciation for Dr. Baheya Jaber. She has been my constant support system through this journey and I am deeply thankful for her unconditional love and support in my life.

As I step into the next phase of my professional life, I feel excited about the potential of AI to revolutionize our understanding and stewardship of the planet's water resources. I now carry with me not only the knowledge and skills gained during my doctoral pursuit, but also a deep commitment to leveraging AI for making meaningful progress toward a sustainable and water-secure future.

CHAPTER 1

Hydrology Research Articles are Becoming More Topically Diverse

1.1. Introduction

In the early stages of hydrology and water resources science, the focus was on bringing together natural scientists, engineers, and social scientists (Harshbarger and Evans, 1967). Freeze (1990) identified a separation between physical and social sciences in water research and encouraged the journal Water Resources Research (WRR) to encourage then-limited partnerships to encourage the mixture of these topics. A report by the National Research Council (1991) focused on the importance of a diverse educational base in hydrology, and encouraged multidisciplinary hydrological research as necessary to understand (and predict) the full global water cycle. Over the next decade, hydrologic sciences became central to new research topics (e.g., hydroclimatology, hydrometeorology, geobiology, hydroecology, hydrogeomorphology, ecogeomorphology, earth system dynamics, etc.) (National Research Council, 2012).

In the modern era, Montanari et al. (2013) argued that the Scientific Decade 2013-2022 would focus on advanced monitoring and data analysis techniques, and that diversity in water science could be sought through connecting economic sciences and geosciences. Montanari et al. (2015) later argued that this branching of sub-topics in hydrologic sciences has given rise to a vibrant interdsiciplinary research culture that focuses on a wide range of spatial and temporal scales, and interactions between water, earth, and biological systems. Ruddell and Wagener (2015) mentioned that hydrology education must expand beyond traditional scopes to address the evolving and unique needs of society (e.g., data and modeling driven cybereducation, developing an international faculty learning community, hydro-economics, etc.).

Vogel et al. (2015) described a modern hydrologic science that develops deeper understanding of human-nature connections. He argued that every theoretical hydrologic model introduced previously is in need of revision to properly capture nonstationarity in nature; proposing knowledge discovery through 'Big Data' to understand the coupled human/hydrologic system. The 21st century saw a sharp rise in demand for more robust, diverse hydrologic models which account for nonstationarity associated with climate change (e.g., Milly et al., 2008; Bayazit, 2015; Galloway, 2011), and leverage large samples of available data (Gupta et al., 2014). Nearing et al. (2021) argued that modern data science has the potential to transform water science given concerted effort to bring together hydrologists with data scientists, computer scientists, and statisticians.

Regardless of how we perceive open challenges in the discipline, it is important for scientists and practitioners to have some idea about whether and how the water science and hydrology science community is changing. In this study, we identify and quantify trends and interactions in and between different subtopics within the discipline. Specifically, we measure trends and diversity of different sub-topics within the discipline, and we use these analyses to provide some insight into the state of topic diversity in the field. Water research articles encompass a wide range of research topics including groundwater, streamflow, climate change, eco-hydrology, biogeochemistry, water quality etc., all of which are consequential to global socioeconomic well-being. McCurley and Jawitz (2017) attempted to assess interdisciplinarity in hydrology by analyzing instances of topic keywords in article titles, however, their corpus consisted of article titles from only one journal - WRR, and used pre-identified keywords and topics. In this paper we look at a broad spectrum of water science and hydrology research publications (our corpus encompasses 18 high-impact journals), and use data science techniques to help (partially) automate the process of identifying distinct sub-topics in the discipline.

One of the major challenges faced by all scientific communities is the increasing volume of peer reviewed literature – Figure 1.1 quantifies this phenomenon in hydrology and water

science. Recent advances in computational linguistics, machine learning, and a variety of application-ready toolboxes for Natural Language Processing (NLP) can help facilitate analyses of vast electronic corpora for a variety of objectives (Cambria and White, 2014). These techniques, which include information retrieval, text categorization, and other text mining techniques based on machine learning have been gaining popularity in information systems since the 1990s (Sebastiani, 2002).

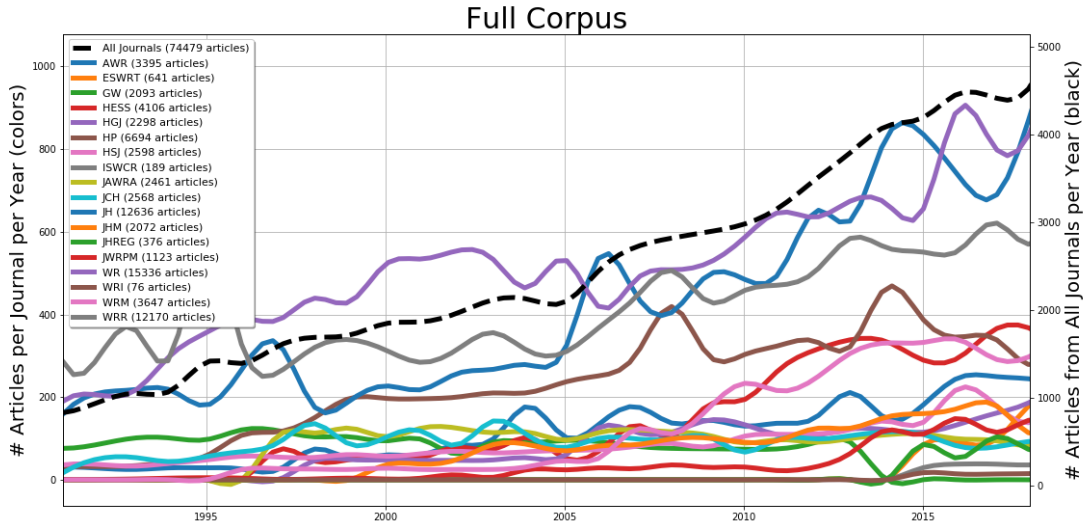


FIGURE 1.1. Number of articles published per year between 1991 and 2019 in 18 major water research journals (Source: Web of Science)

Topic modeling is a particular type of NLP that uses statistical algorithms to extract semantic information from a collection of texts in the form of thematic classes (Jiang et al., 2016). Topic models can be applied to massive collections of documents (Blei, 2012) and have been used to recommend scientific articles based on content and user ratings (Wang and Blei, 2011). Topic modeling has also been used to cluster scientific documents (Yau et al., 2014), improve bibliographic search (Jardine and Teufel, 2014; Pham et al., 2018; Shu et al., 2009; Tang et al., 2008; Paul and Girju, 2009), and for a variety of application-specific objectives such as statistical modeling of the biomedical corpora (Blei et al., 2006), bibliometric exploration of hydropower research(Jiang et al., 2016), in the analysis of research trends in

personal information privacy (Choi et al., 2017), development of meta-review in cloud computing literature (Upreti et al., 2016), literature review of social science articles (Li and Liu, 2018), discovering themes and trends in transportation research (Sun et al., 2017), identifying contribution of authors in knowledge management literature (Jussila et al., 2017), exploring the history of cognition (Priva and Austerweil, 2015), and exploring topic divergence and similarities in scientific conferences (Hall et al., 2008). As opposed to *scientometrics* techniques (Mingers and Leydesdorff, 2015), which have been traditionally used for ranking articles and authors based on citation data, topic modeling allows for a contextual understanding of particular scientific domains and disciplines.

Motivated by the success of topic modeling in a wide range of applications, we explore its potential to aid bibliometric exploration of peer-reviewed water science literature. In particular, we explore the question of whether peer-reviewed water science literature is increasing in diversity with respect to sub-topics in the discipline. The specific hypotheses that we will explore are:

- Individual hydrology research papers are becoming more topically diverse, i.e. it is increasing at the level of individual research projects.
- The hydrology and water science corpus as a whole is becoming more topically-diverse.
- There is a difference in per-paper topic diversity between different water science journals.
- Some topics are negatively correlated to diversity in the community research output.

1.2. Methods

Table 1.1 lists notation used throughout this paper, including variables and indices related to the model and corpus. The corpus that we analyzed is described in Subsection 1.2.1 below. We analyzed this corpus using sequential Latent Dirichlet Allocation (LDASeq) in GenSim (Řehřek and Sojka, 2011), based on Blei and Lafferty (2006)'s Dynamic Topic Model (DTM), to identify dominant topics and to associate topics with individual research articles. LDASeq

is described in 1.6.1 — this NLP method identifies topics by associating a unique set of words that frequently co-appear together in timestamped documents and assigns weights to each of those words based on their likelihood of appearance within a particular topic.

TABLE 1.1. List of notation for indices, parameters and variables

Notation	Meaning
Indices	
d	Documents index
k	Topics index
j	Journals index
t	Years index
Corpus Parameters	
M	Number of documents
N_d	Number of words in document d
t_d	Year of publication of document d
j_d	Journal of document d
A_t	Slice of documents based on year of publication
LDASeq Model Components	
K	Number of topics
K^{opt}	Optimal number of topics
α	Parameters of a Dirichlet prior on on the per-document topic distribution
β	Parameters of a Dirichlet prior on the per-topic word distribution
μ_k	Distribution of topics over document d
μ_{kd}	Weight of topic k assigned to document d
z	list of K topics
z_d	Per-word topic vector for document d
z_k	Particular topic from a list of K topics
w_d	Word collection in document d
Derived Distributions	
μ_{kj}	Weight of topic k over all documents in journal j
μ_{kt}	Average weight for topic k over all documents at time t
$\hat{\mu_k}$	Mean weight of topic k over all documents
μ_{kjt}	Weight of topic k in journal j at time t
μ^m	Topic distribution over entire corpus of M documents
Derived Metrics & Functions	
p	LDA model perplexity score
c	LDA model coherence score
JSD	Jensen-Shannon Divergence
KLD	Kullback-Leibler Divergence
I	Indicator function
H_j	Shannon Diversity of journal j
H_d	Shannon Diversity per document d
$\hat{H}_t$	Mean Shannon Diversity of topics in documents per year
H_{jt}	Shannon Diversity of topics in documents per journal per year
r	Correlation coefficient
$r_{k,j}$	Correlation coefficient between topics k and j
r_{μ,H_d}	Correlation coefficient between document-topic distributions μ and their corresponding article diversity scores

1.2.1. Corpus. Peer-reviewed abstracts offer snapshots of the historical and current trends and developments in both theoretical and applied research. In this study, we use abstracts because they are intended to be concise representations of full-texts and are used often for bibliometric analyses (Griffiths and Steyvers, 2004; Gatti et al., 2015). The corpus that we use consists of abstracts from all peer-reviewed articles published in eighteen water science journals between 1991 and 2019 - this is all water science journals with a 2018 Impact

TABLE 1.2. Repository of article-abstracts

Journal Name	Abbreviation	IF	Years Available	Total Abstracts
Advances in Water Resources	AWR	1.384	1991-2019	3395
Environmental Science: Water Research and Technology	ESWRT	1.104	2015-2019	641
Groundwater	GW	0.911	1991-2013	2093
Hydrology and Earth System Sciences	HESS	2.134	1997-2019	4106
Hydrogeology Journal	HGJ	0.940	1998-2019	2298
Hydrological Processes	HP	1.417	1991-2019	6694
Hydrological Sciences Journal	HSJ	0.913	1991-2019	2598
International Soil and Water Conservation Research	ISWCR	1.134	2015-2019	189
Journal of the American Water Resources Association	JAWRA	1.026	1997-2019	2461
Journal of Contaminant Hydrology	JCH	0.960	1991-2019	2568
Journal of Hydrology	JH	1.830	1991-2019	12636
Journal of Hydrometeorology	JHM	2.410	2000-2019	2072
Journal of Hydrology: Regional Studies	JHREG	1.378	2015-2019	376
Journal of Water Resources Planning and Management	JWRPM	1.418	1991-2019	1123
Water Research	WR	2.721	1991-2019	15336
Water Resources and Industry	WRI	1.255	2015-2019	76
Water Resources Management	WRM	1.097	1996-2019	3647
Water Resources Research	WRR	2.135	1991-2019	12170

Factor (IF) of greater than 0.9 (Scimago Journal and Country Rank). The list of journals and journal abbreviations, along with corresponding IFs, years of available data, and total number of abstracts, are listed in Table 1.2. In total, 74,479 article-abstracts were acquired from the Web of Science core collection in the form of bib files. Methods for pre-processing this corpus are described in 1.6.1.

1.2.2. Analysis Methods. To reiterate from the introduction, the hypotheses that we want to test are about whether hydrology and water science research is becoming more topically diverse over time. We will test these hypotheses by exploring sub-topics within the discipline, and measuring whether individual research articles, individual journals, and the body of water science and hydrology literature as a whole is becoming more topically diverse. The analysis tools that we use to address these research questions are described below. This analysis was applied to the posterior document-topic and topic-word expectations from a trained LDASeq model with 45 topics ($K^{opt} = 45$). We used a combination of objective-subjective method to choose the opitmal number of topics. Details of this process can be found in 1.6.1.

1.2.2.1. *Temporal Trends in Topic Distributions.* There are multiple methods of analyzing temporal trends and distributions of topics. Griffiths and Steyvers (2004) applied a

disjointed time-blind topic model and rearranged documents according to their publication dates. Blei and Lafferty (2006) developed a sequential topic modeling approach that learns time-dynamic parameters for the document-topic and topic-word distributions constrained by linear filtering theory. Wang and McCallum (2006) introduced a non-Markov joint modeling framework where topics are associated with a continuous distribution over document timestamps. We initially tested Griffiths and Steyvers (2004)'s approach of time-unaware topic modeling and post-hoc aggregation of results according to timestamp for benchmarking. Due to the sequential nature of our data, we chose dynamic topic modeling (Blei and Lafferty, 2006) approach for this study because, unlike a time-blind topic model, it provides a qualitative scope into the contents of a large textual dataset in addition to providing us with a quantitative, predictive model for our sequential corpus.

We calculated temporal topic distributions for a given year μ_{kt}, as the proportion of all topic weights over all papers from a given year, t:

$$\mu_{kt} = \frac{\sum_{d=1}^{M} \mu_{kd}\, I(t_d - t)}{\sum_{d=1}^{M} I(t_d - t)}. \tag{1.1}$$

μ_{kd} represents the weight for topic k assigned to document d, t_d is the year in which document d was published, and I is an indicator function such that $I(0) = 1$ and $I(x) = 0$ for $x \neq 0$. Henceforth, I will carry the same meaning.

Statistical significance of these trends were assessed using standard linear regression analysis between variables. In each case, we computed the (i) Pearson correlation coefficient (r) as the strength of association between variables, (ii) the p-value for the t-test of the correlation coefficient against a null hypothesis of zero-trend, and (iii) the Bayes Factor (B10) as a measure of the strength of evidence toward the alternate (nonzero-trend) hypothesis.

1.2.2.2. *Measurement of Topic Diversity.* Shannon entropy (Shannon, 1948) is a classic diversity metric that is used - among many other things - in ecology studies to quantify the diversity of species in a given ecosystem or location (e.g., Harte and Newman, 2014; Sherwin and Prat i Fornells, 2019). Intuitively, we propose that articles can be analogous

to ecological sites and topics are analogous to species. We used the Shannon Entropy based metric applied to topic distributions to measure diversity at corpus and article levels.

1.2.2.3. *Measuring Diversity at the Article Level.* We used Shannon entropy to measure the topic diversity H_d for each article in our corpus as:

$$H_d = -\sum_{k=1}^{K}(\mu_{kd} log(\mu_{kd})), \tag{1.2}$$

Where μ_k is the distribution of topics over document d. We also calculated the mean Shannon diversity in documents per year as $\hat{H}_t$:

$$\hat{H}_t = \frac{\sum_{d=1}^{M} H_d \; I(t_d - t)}{\sum_{d=1}^{M} I(t_d - t)}, \tag{1.3}$$

Finally, we calculated the Shannon diversity of topics in documents per journal per year H_{jt} as:

$$H_{jt} = \frac{\sum_{d=1}^{M} H_d \; I(|j_d - j| + |t_d - t|)}{\sum_{l=1}^{K}\sum_{d=1}^{M} H_d \; I(|j_d - j| + |t_d - t|)}, \tag{1.4}$$

Shannon diversity is represented using the natural unit of information (*nat*), where 1 *nat* represents the information contained in an event when the probability of that event occuring is *1/e*.

1.2.2.4. *Measuring Diversity at the Journal and Corpus Level.* We calculated Shannon diversity at the corpus level and then computed these corpus indexes for both the entire corpus and for each journal. To do this, we began by calculating the K-nomial distribution over topics μ_j in a particular set of articles j (either a journal or the whole corpus, although we will hereafter refer to subscript j as referring to a specific journal):

$$\mu_{kj} = \frac{\sum_{d=1}^{M} \mu_{kd} \; I(j_d - j)}{\sum_{l=1}^{K}\sum_{d=1}^{M} \mu_{ld} \; I(j_d - j)}, \tag{1.5}$$

where μ_{kj} is the relative popularity of a particular topic in a particular journal as a fraction of popularity of all topics in the journal. We then calculated the total entropy of each μ_j topic distribution as H_j, which is a measure of the Shannon diversity of the per-journal topic

distributions:

$$H_j = -\sum_{k=1}^{K}(\mu_{kj} log(\mu_{kj})), \tag{1.6}$$

The popularity of a particular topic in a particular journal for a particular year, μ_{kjt} is a fraction of the popularity of all topics in that journal and year:

$$\mu_{kjt} = \frac{\sum_{d=1}^{M} \mu_{kd}\, I(|j_d - j| + |t_d - t|)}{\sum_{l=1}^{K}\sum_{d=1}^{M} \mu_{ld}\, I(|j_d - j| + |t_d - t|)}, \tag{1.7}$$

We used these per-year, per-journal topic distributions to construct timeseries of individual topic popularity in each journal, μ_{kjt}, which allowed us to quantify the evolving diversity of topic distributions in individual journals over time.

1.2.3. Correlation Between Pairs of Topics. We calculate the correlation coefficient between pairs of topics. This allows us to broadly separate frequently co-occuring (i.e., exist within the same article) topics from the ones which do not frequently co-occur.
The correlation coefficient between topic weights over the whole corpus M for each pair of topics, $r_{k,j}$, was calculated as:

$$r_{k,j} = \frac{\sum_{d=1}^{M} (\mu_{kd} - \hat{\mu_{kd}})(\mu_{jd} - \hat{\mu_{jd}})}{\sqrt{\sum_{d=1}^{M} (\mu_{kd} - \hat{\mu_{kd}})^2}\sqrt{\sum_{d=1}^{M} (\mu_{jd} - \hat{\mu_{jd}})^2}}, \tag{1.8}$$

where μ_{kd} is the weight for topic k assigned to document d, and $\hat{\mu_{kd}}$ is the mean weight for a topic k assigned over all documents in the corpus, and μ_{jd} is the weight for a topic j assigned to document d, and $\hat{\mu_{jd}}$ is the mean weight for topic j assigned over all documents in the corpus. We only report correlations greater than 0.2.

1.2.4. Correlation between Topics and Per-article Diversity. We observe the correlation between topics and per-article diversity in water science articles by observing the statistical relationship between topic distribution weights and article diversity. This allows us to identify which topics participate more or less often in articles with greater or lesser topic diversity. Intuitively, a negative statistical relationship between topic distribution

weights and article diversity indicates decreasing article diversity when certain topics are more present within an article.

We denote the correlation coefficient between document-topic distributions and their corresponding article diversity scores (entropy metrics) as r_{μ,H_d}.

1.3. Results and Analysis

1.3.1. Naming the Topics. The first step towards using the posterior expectations of the LDASeq model is naming the topics. We identified and named $K = 45$ topics by first looking at the topic-word distributions (the set of words most likely to appear within a particular topic), and the per-document topic distributions (from the titles of 100 articles most closely associated with each topic). We reinforced our choices of topic names with an informal survey sent to four qualified hydrologists outside of our research group. Figure 1.2 illustrates the topic-word distributions of $K = 45$ topics in the form of wordclouds, along with our chosen topic names.

This topic naming analysis was similar to what was done by McCurley and Jawitz (2017), who looked at topic diversity in WRR papers as described in the introduction. Those authors assigned seven topics in hydrology prior to their analysis: catchment-hydrology, hydrogeology, hydro-meteorology, contaminant hydrology, socio-hydrology, and hydro-climatology. Our post-hoc identified topics extracted using LDASeq were conceptually similar to these, however LDASeq was able to extract a larger and more nuanced set of topics through unsupervised learning.

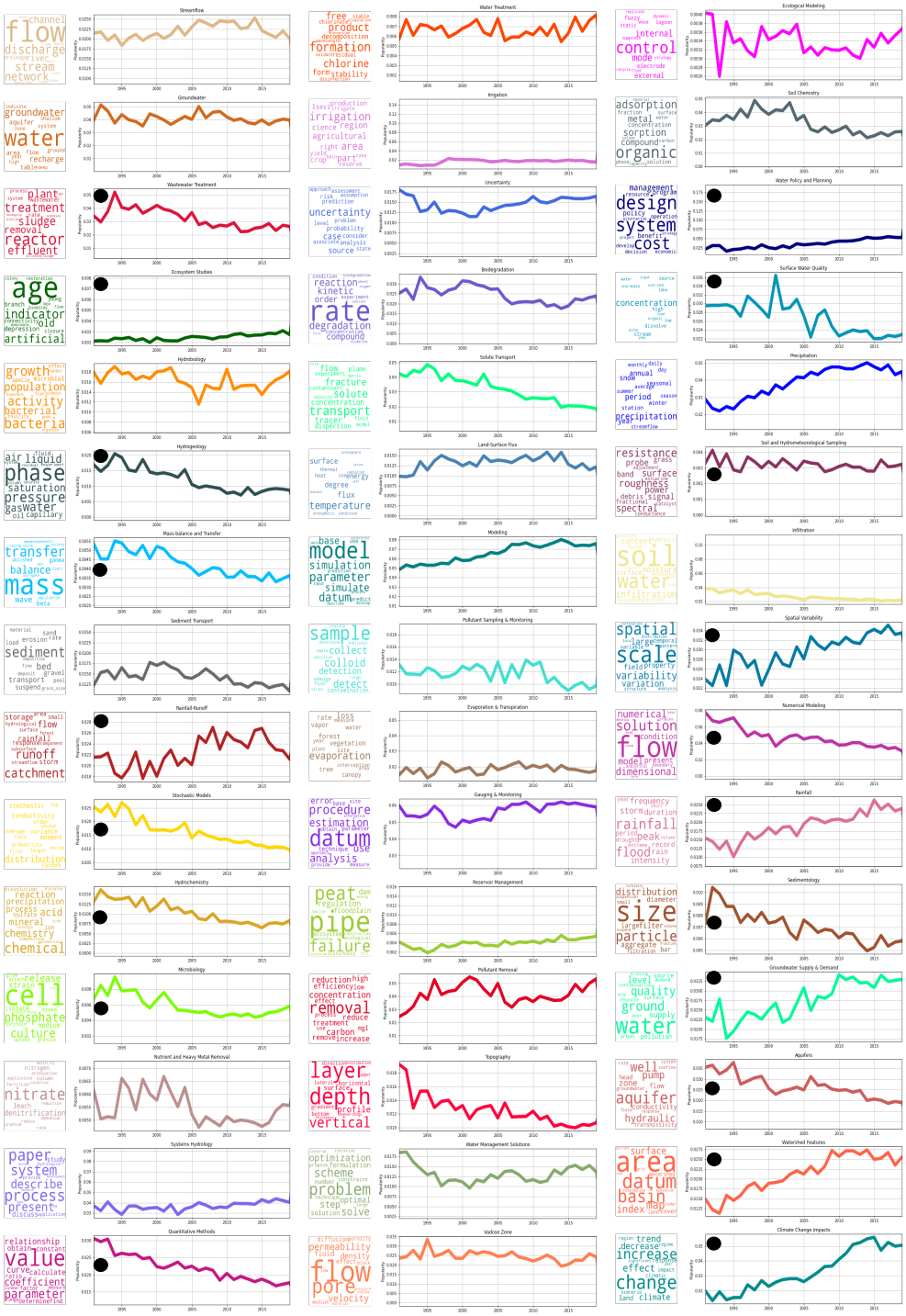


FIGURE 1.2. Wordclouds show the words most strongly associated with each topic, and the sizes of words within the wordclouds are proportional to their likelihood of appearance within individual topics. Statistically significant trends are indicated by black dots on the left axes(20/45 are significant) their respective plots. Topic trends are independent and not depicted relative to each other (see Figure 1.3).

1.3.2. Temporal Trends of Topics in the Full Corpus. The popularity of each topic changes with time, and these trends are also shown in Figure 1.2. Some topics demonstrated statistically significant rising trends in popularity (table 1.3). Some of these rising topic trends (e.g. 'Rainfall-Runoff', Precipitation', 'Rainfall', 'Spatial Variability') might be attributed to researchers increasingly leveraging the availability and accessibility of hydrology related data, both in terms of breadth and depth. Other topics demonstrated statistically significant downward trends (table 1.3). The remainder of topics do not demonstrate any significant trend within our corpus.

Figure 1.3 shows the relative popularity of topics over time plotted on the same scale (Figure 1.2 shows the same topic trends but not normalized). Considering the relative popularity of topics in 1991 vs. 2019, topics that lost the most popularity within our corpus (over -50%) are "Stochastic Models" (-62%), "Numerical Modeling" (-61%), "Solute Transport" (-56%). Conversely, the topics that gained the most (over +50%) are "Climate Change Impacts" (+155%), "Water Policy & Planning" (+143%), "Pollutant Removal" (+117%), "Watershed Features" (+72%), "Irrigation" (+60%), "Modeling" (+57%), "Precipitation"(+57%), and "Rainfall"(+55%). These changes in the popularity of topics can be, perhaps, interpreted as shifting focus of researchers who publish their works within the journals in our corpus. These changes reflect the increasing effects of climate change on water availability and rainfall, increasing water stress and an ever more pressing need for sustainable and efficient water management (e.g., Mehran et al., 2017; Tabari, 2020; Padrón et al., 2020; Gudmundsson et al., 2021). In addition to leveraging the availability of data, water researchers are responding to the needs of the time.

TABLE 1.3. Rising and falling temporal trends of topics (only statistically significant trends are reported)

Rising Trends

Topic	p-val	BF10
Rainfall-Runoff	1.24E-04	253.82
Water Policy and Planning	2.42E-04	139.38
Precipitation	1.92E-04	171.40
Spatial Variability	8.20E-05	367.25
Rainfall	1.30E-04	242.14
Groundwater Supply & Demand	5.12E-09	2.50E+06
Watershed Features	5.61E-13	1.13E+10
Climate Change Impacts	1.06E-14	4.47E+11
Ecosystem Studies	4.46E-03	10.74

Falling Trends

Topic	p-val	BF10
Wastewater Treatment	4.86E-07	3.85E+04
Hydrogeology	1.41E-10	6.86E+07
Mass-balance and Transfer	1.94E-10	5.11E+07
Stochastic Models	1.34E-14	3.58E+11
Hydrochemistry	2.21E-11	3.79E+08
Microbiology	1.52E-07	1.11E+05
Quantitative Methods	5.38E-16	7.05E+12
Surface Water Quality	2.35E-06	9.13E+03
Numerical Modeling	3.54E-10	2.93E+07
Sedimentology	5.51E-08	2.83E+05
Aquifers	6.43E-10	1.69E+07

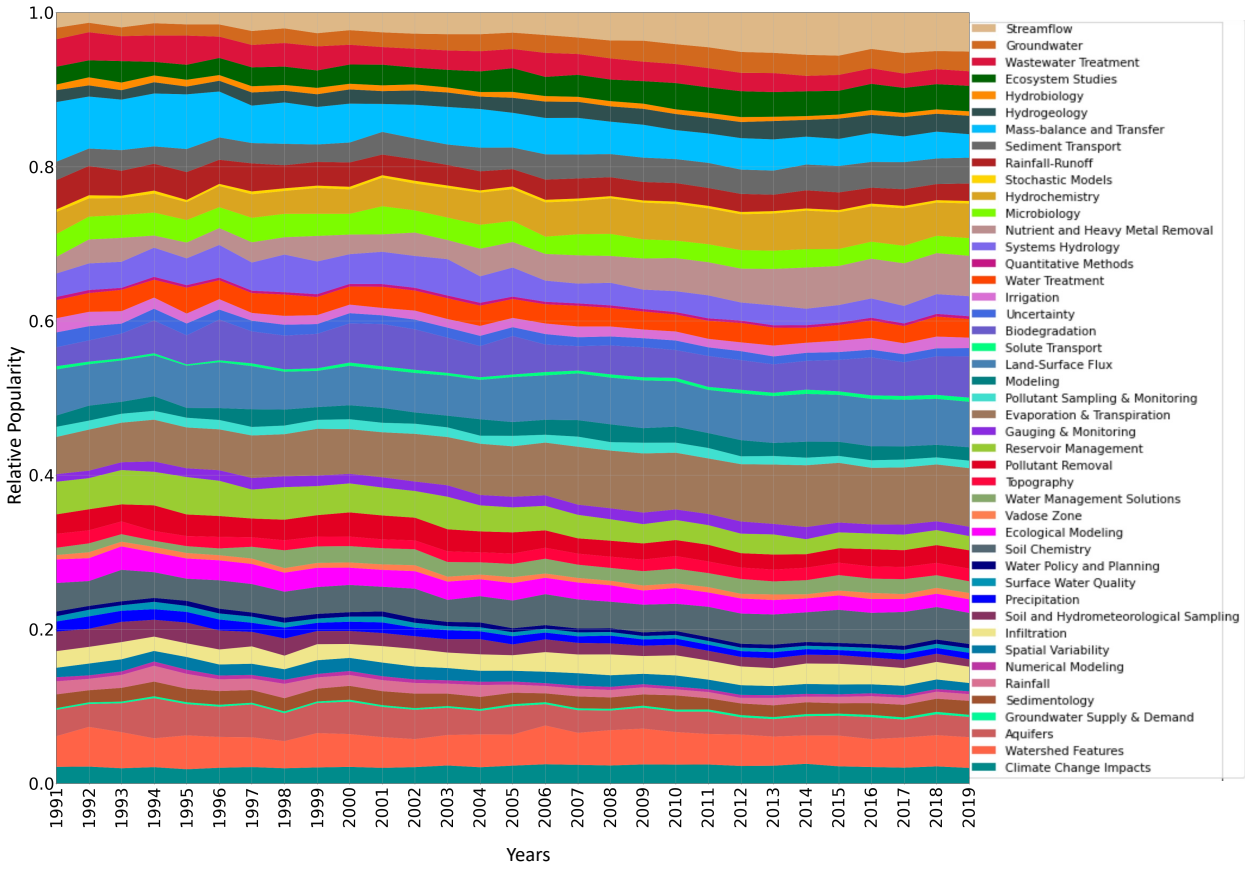


FIGURE 1.3. Temporal variation of topic popularity relative to each other.

1.3.3. Are Articles becoming More Topically Diverse? The corpus-wide mean per-article diversity metric is shown in Figure 1.4. Our findings indicate the average diversity of topics within individual water science articles is increasing overall. Regression-based trend analysis for the Shannon diversity metric time from the entire corpus are: $r = 0.95$, p-value = 1.36e-14, B10 = 3.39e+11, indicating a statistically significant trend at any reasonable significance threshold.

To gain an intuitive interpretation of this change in diversity, we applied another metric from ecological/biological sciences - ENS(Effective Number of Species). In our case, we will call it ENT(Effective Number of Topics), where $ENT = e(H_d)$. As an example, if $ENT = x$

for mean per-article diversity $\hat{H}_t$ for year(t), $\hat{H}_t$ is equivalent to articles containing x count of equally-common topics. In our corpus, the mean effective number of topics (ENT) per article steadily rose from 13.62 in 1991 to 15.29 in 2019. This means a 4.44% rise in mean per article topic diversity translates to 12.26% rise in the number of equally-common topics per article between 1991 and 2019. This rising ENT can also be interpreted intuitively as an indicator of water researchers absorbing knowledge from topics within other disciplines through interdisciplinary collaborations and education.

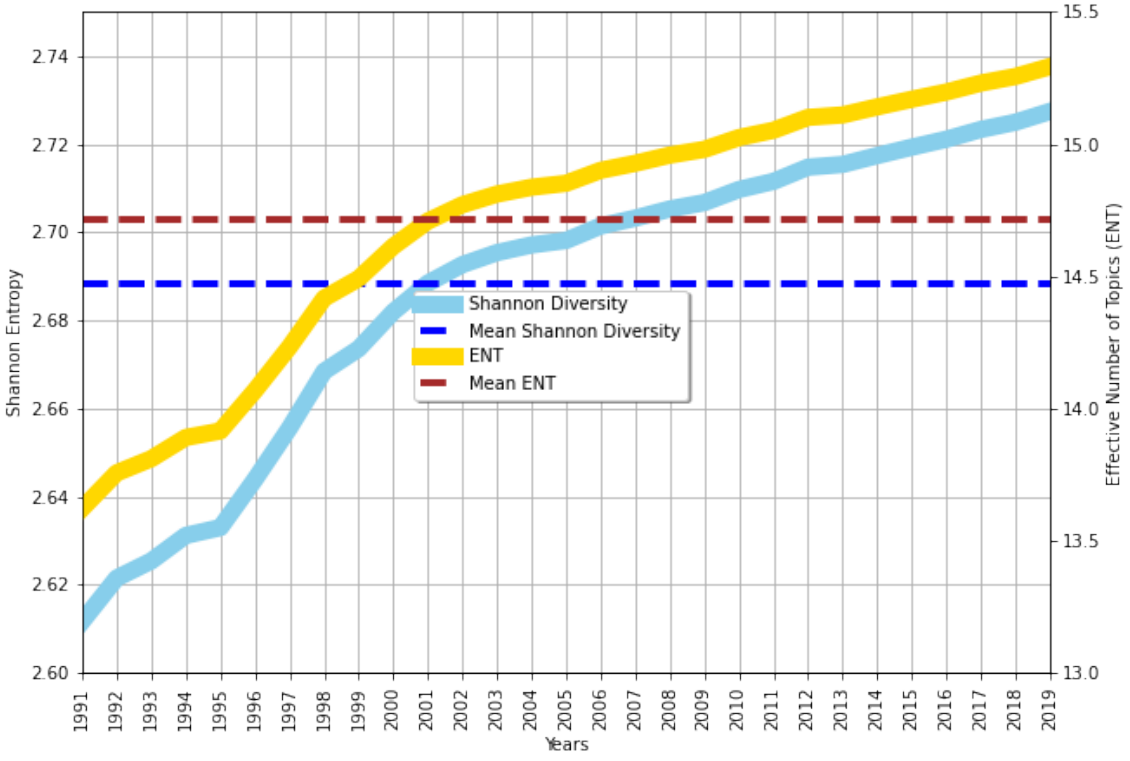


FIGURE 1.4. Mean per-article diversity (left axis) and ENT per year (right axis). The dashed lines represent the mean per-article diversity and ENT over the entire corpus.

1.3.4. Which Journals Are Contributing to Per-Article Diversity? To understand which journals are contributing to the trend of increasing diversity of topics in individual research articles, we calculated the mean diversity of articles per year for each of the

eighteen journals as shown in Figure 1.5. As before, we used linear regression to assess the significance of temporal trends in these per-journal time series.

Water Resources Research WRR demonstrates the strongest rise (as an individual journal) in the mean diversity of topics per article published between 1991 and 2019 (R = 0.92, p-value = 2.39e-12, BF10 = 2.77e+09). Other journals with overall rise in per-article diversity within our corpus are Advances in Water Resources AWR (R = 0.69, p-value = 5.69e-05, BF10 = 513.33), Water Research WR (R = 0.67, p-value = 9.14e-05, BF10 = 336.08), Journal of Contaminant Hydrology JCH (R = 0.67, p-value = 1.05e-05, BF10 = 297.751), and Journal of Hydrology JH (R = 0.57, p-value = 1.57e-03, BF10 = 27.06). While these results do not directly translate to a rise of interdisciplinarity within these journals, they most certainly indicate increasing diversification of topics. This increasing diversification can be driven by multiple factors, which again includes researchers creating new and absorbing knowledge from other disciplines.

Journals which demonstrate moderate rises in per-article diversities are Water Resources Management WRM (R = 0.46, p-value = 0.026, BF10 = 2.68), and Hydrogeology Journal HGJ (R = 0.43, p-value = 0.05, BF10 = 1.59). Journal of Water Resource Planning & Management $JWRPM$ (R = 0.28, p-value = 0.15, BF10 = 0.62), Journal of the American Water Resources Association $JAWRA$ (R = 0.11, p-value = 0.64, BF10 = 0.29), and Hydrological Processes HP (R = 0.02, p-value = 0.94, BF10 = 0.24) do not demonstrate any significant trend at a significance level of $\alpha = 0.01$. Average diversity of articles published in Hydrologic Sciences Journal HSJ (R = -0.46, p-value = 0.01, BF10 = 4.11), Hydrology & Earth System Sciences $HESS$ (R = -0.36, p-value = 0.09, BF10 = 1.00), and Journal of Hydrometeorology JHM(R = -0.30, p-value = 0.21, BF10 = 0.59) decreased. The rest of the journals do not have publication records long enough for trend analysis. The declining per-article diversity trends could mean that these journals are increasingly favoring a particular set of topics or that researchers working on certain topics are favoring these journals.

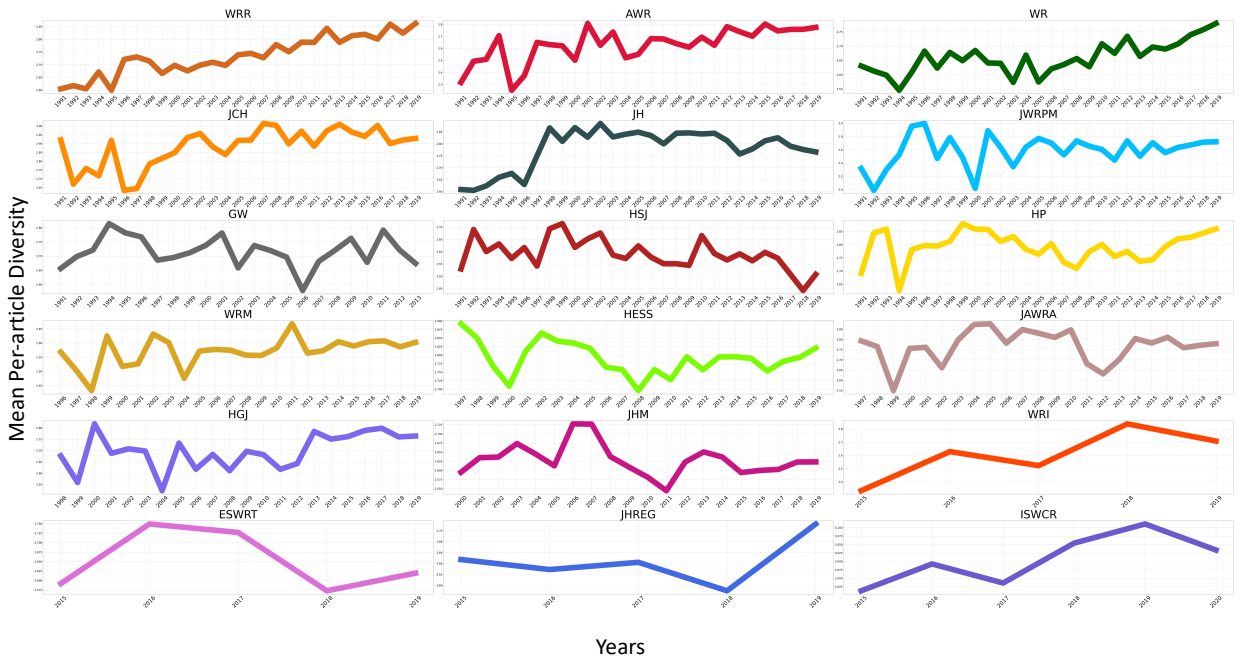


FIGURE 1.5. Mean per-article diversity (Shannon entropy) per-journal over time

1.3.5. Is the Whole Corpus becoming More Topically Diverse? Figure 1.6 shows the temporal variability of topic entropy (diversity) over time for the entire corpus (dashed black line) and for each individual journal (solid colored lines). This differs from the average per-article diversity metrics reported in the previous subsection in that these metrics are calculated over the topic distributions averaged over all papers in the corpus (journal). Whereas the per-article diversity metrics diversity of (presumably) individual research projects, the corpus metrics measure the diversity of topics overall in a journal or corpus and measure the mixture of topics at community level rather than at the level of individual research projects. The diversity for the entire corpus rose very slightly in the late 1990s and, since then, the entropy of the entire corpus has remained steady or slightly decreased. However, no definite trend exists overall (R = -0.43, p-value = 0.02, BF10 = 0.69). This emphasized the disentanglement of per-article diversity from corpus diversity - showing that, increasing article-level diversity does not necessarily translate to overall corpus diversity.

We used Figure 1.6 to also visualize the per-journal topic diversity trends. Statistically significant upward diversity trends can be seen for Advances in Water Resources AWR (R = 0.79, p-value = 2.68e-07, BF10 = 6.59e+04), Water Resources Research WRR (R = 0.713, p-value = 1.39e-05, BF10 = 1824.36), Journal of Water Resources Planning & Management $JWRPM$ (R = 0.69, p-value = 3.73e-05, BF10 = 745.97), and Hydrogeology Journal HGJ (R = 0.52, p-value = 0.01, BF10 = 5.13). Journals which demonstrated statistically significant downward trends were Water Research WR (R = -0.64, p-value = 1.70e-04, BF10 = 191.81) and Hydrological Sciences Journal HSJ (R = -0.59, p-value =8.04e-04, BF10 = 48.40). Other journals did not demonstrate any significant trend in entropy over time. Here again, evidences of disentanglement between per-article diversity and overall corpus diversity can be seen at a journal level.

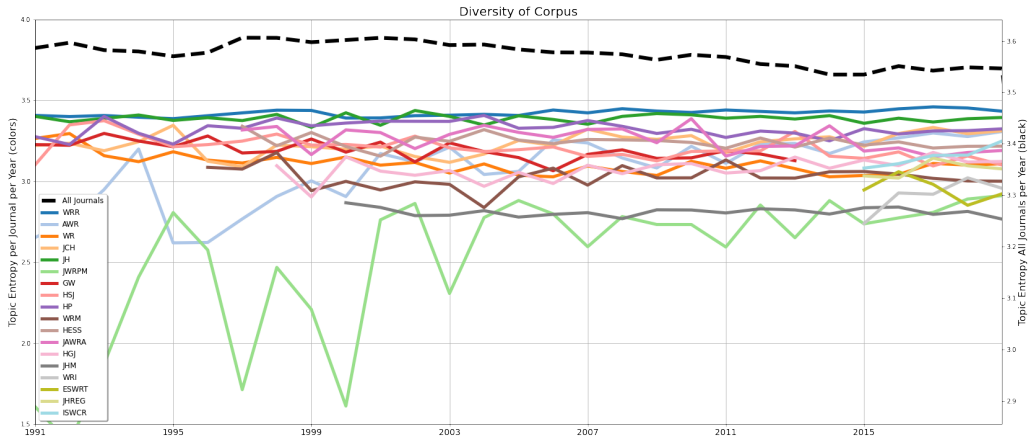


FIGURE 1.6. Temporal variation of the diversity of each journal, as measured by the entropy of that journal's topic distribution in a particular year.

1.3.6. Overall Journal Diversity. The stacked bar plots in Figure 1.7 show the relative fraction of topic representation in each journal, with the total height of each bar representing the journal's topic entropy. Water Resources Research WRR (3.45 nats), Journal of Hydrology JH (3.40 nats), Hydrological Processes HP (3.35 nats), and Journal of

the American Water Resources Association $JAWRA$ (3.25 nats) are the most topically diverse journals in our corpus. We can again intuitively interpret these values in terms of ENT, meaning that these journals have published the highest numbers of equally-common topics within the entire dataset. The overall Shannon Diversity per journal decreases for more specialty journals – i.e., journals which focus atmospheric science topics - Journal of Hydrometeorology JHM and water management topics - Water Resources Management WRM, Journal of Water Resources Planning & Management $JWRPM$. Journals with a fairly recent publication history – i.e., Environmental Science: Water Research and Technology $ESWRT$, International Soil and Water Conservation Research $ISWCR$, Journal of Hydrology: Regional Studies $JHREG$, and Water Resources and Industry WRI had lower overall diversity compared to the rest of the corpus, which is expected.

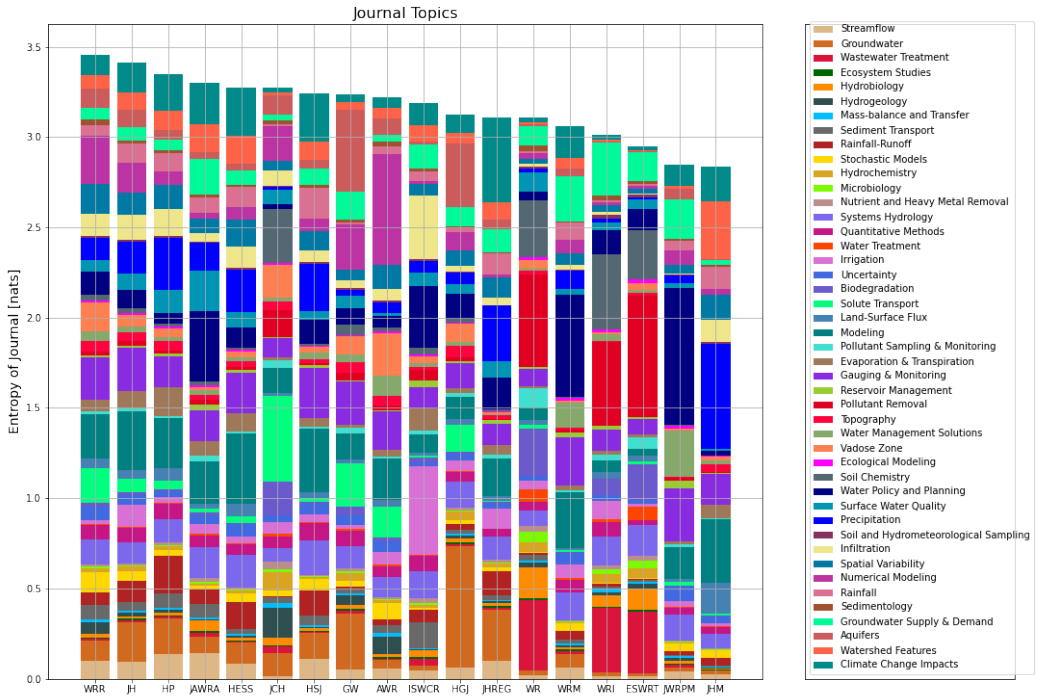
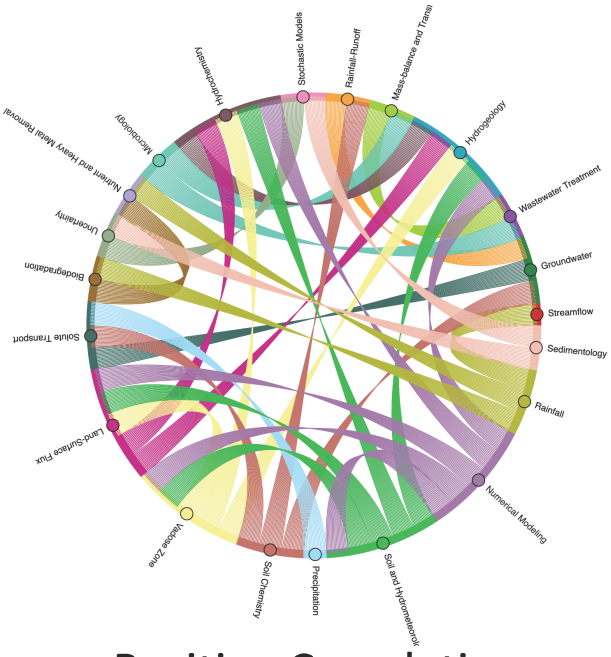


FIGURE 1.7. Total bar height represents the overall diversity of topic distributions of each journal for the whole study period. The stacked color bars represent the fraction of papers representing each individual topic in that journal.

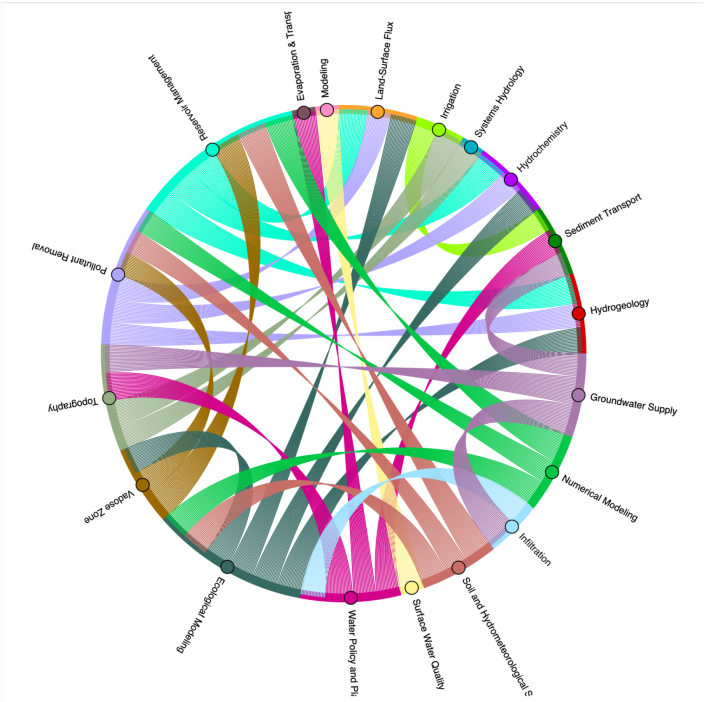
1.3.7. Correlations Between Topic Pairs. To reiterate from Section 1.2.3, we observe the correlations (both positive and negative) between pairs of topics to understand which topics co-appear frequently in our corpus.

1.3.7.1. *Co-appearing Topics.* An intuitive way to depict inter-topic correlations $r_{k,j}$ are chord-diagrams. $r_{k,j}$ correlation coefficients measure relationships between per-paper topic weights, meaning that a higher $r_{k,j}$ value indicates papers that contain word groups associated with topic k also tend to contain word groups associated with topic j. Positive correlation coefficients between pairs of topics indicate some degree of co-appearance of these topics in research articles, and vice-versa. Positive and negative inter-topic correlations are shown

in Figure 1.8, where the width of each chord represents the overall correlation between a pair of topics. For ease of viewing, positive correlations are only plotted for $r_{k,j} > 0.20$ and negative correlations $r_{k,j} <$ -0.20. Inter-topic correlation plots for the entire corpus lends us a snapshot of co-appearing topics.



Positive Correlations



Negative Correlations

FIGURE 1.8. Inter-topic correlations: positive correlations in the upper subplot and negative correlations in the lower subplot. Only correlations $|r_{k,j}| > 0.20$ are shown.

1.3.7.2. *Positive and Negative Inter-Topic Correlations.* Positive Correlations or likelihood of co-occurrence can be observed for a range of topics, e.g. between "Rainfall" and "Streamflow", "Rainfall" and "Spatial Variability", "Uncertainty" and "Stochastic Models", "Land Surface Flux" and "Hydrogeology", "Groundwater" and "Solute Transport", and "Microbiology" and "Wastewater Treatment". Anti-correlations indicate that there are set of vocabulary in the water science literature that are largely not shared between sub-communities. For example, "Pollutant Removal" and "Land Surface Flux", "Pollutant Removal" and "Vadose Zone", "Water Policy and Planning" and "Uncertainty", "Numerical Modeling" and "Reservoir Management", and "Irrigation" and "Sediment Transport" are less likely to co-appear within our corpus. These negative correlations between topics indicate potential for expanding avenues of collaborative research.

1.3.8. Correlation between Individual Topics and Per-article Diversity. Reiterating from Section 1.2.4, we quantify relationships between per-article diversity and corresponding topic weights. Some topics in our corpus tend to reduce the paper-wise diversity when they appear in an article (meaning they are less likely to appear alongside a wide variety of other topics). Statistical relationship between mean per-article Shannon Diversities H_d and their corresponding topic distribution weights μ are shown in Figure 1.9. Topics which demonstrate statistically significant relationships with per-article diversity are indicated with black dots on their corresponding bars.

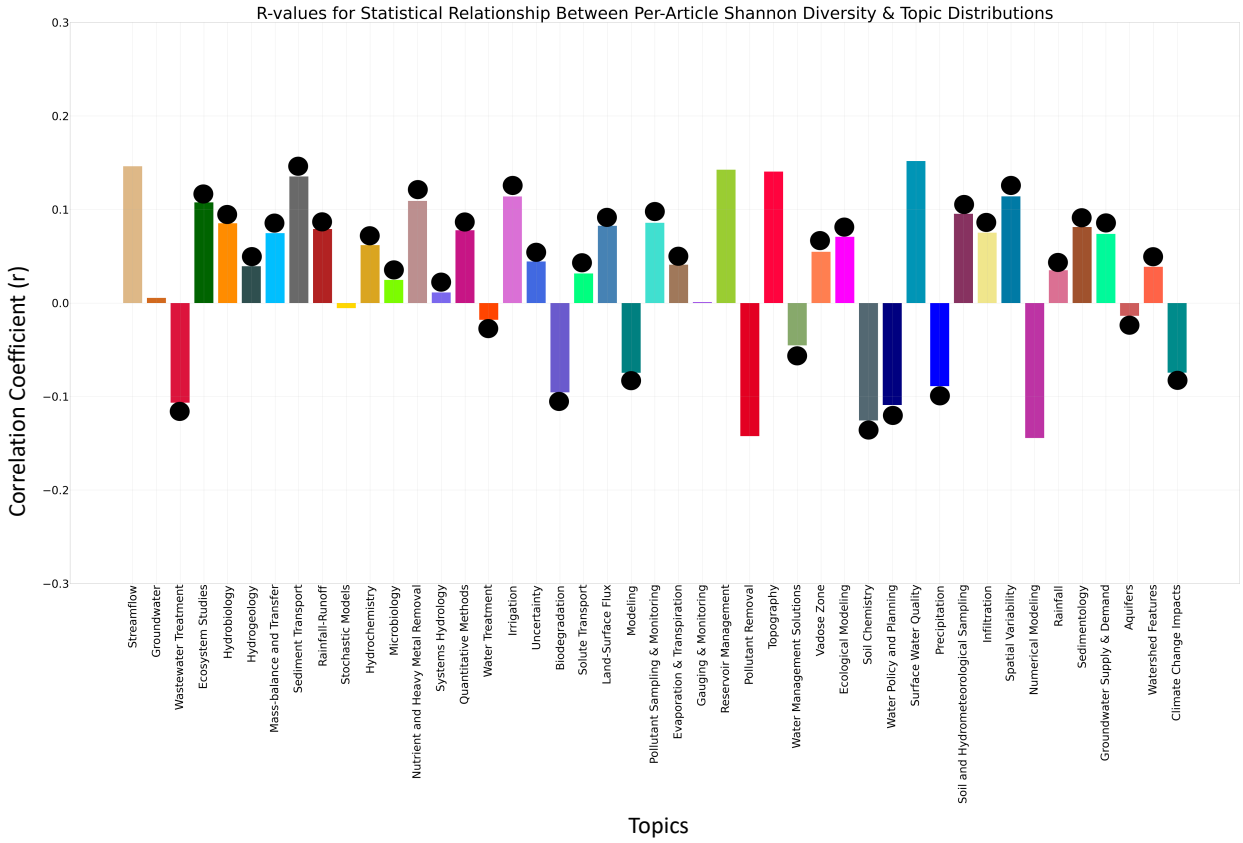


FIGURE 1.9. Pearson correlation coefficients for statistical relationships between per-article Shannon diversity metrics and per-topic distribution weights. Statistically significant relationships are indicated with black dots on their corresponding bars.

Figure 1.10 shows the temporal behavior of these negatively correlated ($r_{\mu,H_d} < -0.05$) topics. "Wastewater Treatment" ($r = 0.74$, p-value = 6.54e-06, BF10 = 3602.16) was the only topic becoming less negatively correlated with per-article diversity, indicating an increasing co-appearance with a wider variety of other topics in individual articles. Opposite trend was observed for "Modeling" (r = -0.88, p-value = 4.93e-10, BF10 = 2.12e+07),"Water Policy and Planning" (r = -0.63, p-value = 2.28e-4, BF10 = 147.57), "Precipitation" (r = -0.87, p-value = 2.53e-09, BF10 = 4.71e+06), and "Climate Change Impacts" (r = -0.87, p-value

= 1.30e-09, BF10 = 8.69e+06). "Soil Chemistry" ($r = 0.17$, p-value = 0.37, BF10 = 0.34) does not demonstrate any significant trend.

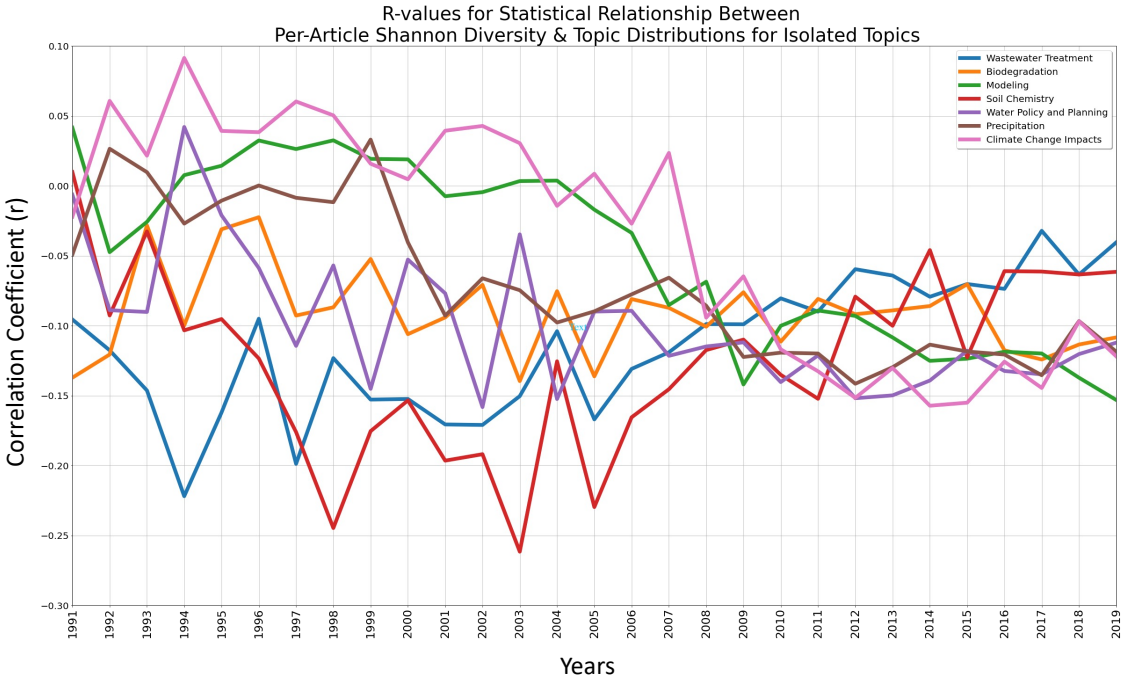


FIGURE 1.10. Trends of Pearson correlations between per-article Shannon diversity and topic distributions for topics which negatively correlate ($r_{\mu,H_d} < -0.05$) with per-article diversity.

1.4. Conclusions & Discussion

We use semantic-based topic diversity to quantify two types of topic diversities in hydrology and water science articles: (i) within individual articles and (ii) across corpora (both within individual journals and within a corpus of all water science journals with a 2018 IF greater than 0.9). We tested the hypotheses that diversity was increasing in both respects and found evidence to support one of those hypotheses but not the other. Individual researchers appear to be broadening their scope across different subtopics in the discipline (i.e., per-paper topic diversity is increasing – Figure 1.4), and while individual topics are changing in popularity

over time (Figure 1.3), the water science and hydrology corpus as a whole is not increasing, nor decreasing, in diversity (Figure 1.6).

The primary findings of this study are (see the four hypotheses outlined in Section 3.1):

(1) At an article level, the average (Shannon) diversity of topics in individual research papers is increasing over the entire corpus (r = 0.95, p-value =1.36e-14, B10 = 3.39e+11). There was a 4.44% rise in mean per-article topic diversity, translating to a 12.26% rise in the number of equally-common topics per article between 1991 and 2019.

(2) At a corpus level, the average (Shannon) diversity of topics in the whole corpus is neither increasing nor decreasing (r = -0.43, p-value = 0.02, BF10 = 0.69).

(3) At a journal level, the most topically-diverse water science journals are Water Resources Research WRR (3.45 nats), Journal of Hydrology JH (3.40 nats), Hydrological Processes HP (3.35 nats), and Journal of the American Water Resources Association $JAWRA$ (3.25 nats). Certain journals are increasing in their average per-article topic diversity (Water Resources Research WRR, Advances in Water Resources AWR, Water Research WR, Journal of Contaminant Hydrology JCH, and Journal of Hydrology JH), and three journals are decreasing in their average per-article topic diversity (Hydrological Sciences Journal HSJ, Hydrology and Earth System Sciences $HESS$, and Journal of Hydrometeorology JHM).

(4) At a topic level, certain topics are more semantically different from the others in a sense that their appearance in an article tend to reduce the article diversity. These topics are: "Wastewater Treatment", "Modeling", "Biodegradation", "Soil Chemistry", "Water Policy and Planning", "Precipitation", and "Climate Change Impacts".

Our interpretation of these findings is that water science research articles are becoming more topically diverse. The increasing mixture of research topics in articles is most likely a bottom-up effect driven by changing efforts, attitudes, and vision by individual researchers

and - perhaps - of increasingly multidisciplinary education, as called for by National Research Council (1991). However, diversity of the overall corpus is not increasing. If it were the case that both per-paper diversity and the overall corpus diversity were increasing, it would have been difficult to disentangle these effects. The hydrology community could benefit from top-down policies and actions which encourage more topically diverse and cross-disciplinary research, which we think will raise overall diversity.

The ability to automatically detect distinct sets of vocabularies (as topics) is a strength of unsupervised dynamic topic modeling, however it is important to remember that any results from an analysis of topic model outputs is related to the words that define the topics. As more topics emerge within our discipline through new knowledge, increasing collaborations, and conducive policies, we expect topic modeling to continue being helpful towards tracking the evolution of hydrological sciences.

1.4.1. Future Outlook. The volume of scientific research in general is growing rapidly. This makes it difficult for researchers to be confident about fully understanding the state of the science, and also makes it challenging to expand into new research topics since so much background information is available for synthesis. We expect that in the future machine learning methods like topic modeling will be an integral part of the tool set available to help scientists synthesize scientific literature. While this paper provides multi-level (per-paper, per-journal, and whole-corpus) contextual insights into the current state of topic diversity in water research, we envision that similar NLP-based efforts might help us address problems related to semantically synthesizing diverse bodies of water science and hydrological literature. There have been several biobliometric analyses of hydrology literature (e.g., Clark and Hanson, 2017; Zare et al., 2017; Rajaram et al., 2015; Koutsoyiannis and Kundzewicz, 2007; McCurley and Jawitz, 2017), however NLP has the potential to allow for faster, and more contextual analyses of larger corpora. LDASeq also allows us to look at the evolution of topics in terms of their probabilistic distance and also their varying word-topic distributions. This paper serves as a preface to a currently undergoing hydrology topic evolution study.

Interdisciplinary research has been identified as one of the ways to solve the world's biggest problems (Ledford, 2015). However, academia continues to be strangled by traditional stereotypes: interdisciplinary proposals are less likely to receive funding (Bromham et al., 2016) and institutions continue to enable this discrimination (Ledford, 2015). While we cannot definitively say that interdisciplinarity is increasing in hydrological sciences, the increasing per-article diversity is an indicator that it may be. This article lays the groundwork for further, and much needed, focus on interdisciplinarity in hydrological sciences.

1.5. Acknowledgements (Project Specific)

Mashrekur Rahman was partially supported by the NASA Advanced Information Systems Technology program (award ID 80NSSC17K0541).

The code and data to reproduce all results and figures are available at https://doi.org/10.5281/zenodo.6976752.

The authors appreciate the help of Dr. Kevin Walker and Mangala Krishnamurthy from the University of Alabama Libraries for their assistance in acquiring large quantities of full-text journal articles that we used for benchmarking. The authors are also thankful to Dr. Hoshin V. Gupta and Dr. Ty Ferre from the University of Arizona, Dr. Bart Nijssen from the University of Washington, and Dr. Cris Prieto Sierra from Universidad de Cantabria for their help identifying topic names.

1.6. Appendix - Chapter 1

1.6.1. Preprocessing the Corpus. Performance of dynamic topic modeling is influenced by the quality of input training data. Article-abstracts were preprocessed into a canonical format for efficacious feature extraction (Feldman et al., 2007). To prepare the data, we used separate temporally-segregated dataframes of abstracts and metadata from each journal. All sets of data were processed through identical multi-layered cleaning routines. We used Spacy and NLTK Python libraries to filter non-semantic elements such as stopwords, punctuation, and symbols, and in addition we manually identified and removed

unwanted elements that were common in our article abstracts (the cleaned abstracts are available in the repository linked in the Data and Code Availability statement at the end of this article).

In the next step, we formed bi-grams and segmented texts by tokenizing with whitespaces as word boundaries. This was followed by lemmatization, to extract semantic roots from conjugations, etc. Using this corpus, we created a map between words and integer identifiers. We then converted this dictionary into a bag-of-words format, making the corpus ready for ingestion by an LDASeq model implemented in *Gensim* - a Python library for NLP (Řehřek and Sojka, 2011).

1.6.2. Dynamic Topic modeling. To understand dynamic topic modeling, we must start with Latent Dirichlet Allocation (LDA). LDA builds on another more traditional topic modeling approach (Latent Semantic Analysis) (Landauer et al., 1998), and captures the intuition that text documents exhibit multiple topics in different proportions. Documents are represented as mixtures of topics (per-document topic distributions) and each topic is characterized by a distribution over words (per-topic word distributions).

We can build an intuition of this model as follows. It is assumed that the per-document topic distributions of all documents in a corpus share a common Dirichlet prior (parameterized by parameters α), and that the per-topic word distributions also share a (different) common Dirichlet prior (parameterized by parameters β). The distribution over a particular word w in a document d with topic distribution μ_d can be understood as (Blei et al., 2003):

$$p(w|\mu_d, \beta) = \sum_{k=1}^{K} p(z_k|\mu_d)p(w|z_k, \beta), \tag{1.9}$$

where z_k is a particular topic from K total topics. Treating the per-document topic distribution as latent and integrating over all N_d words in each document d and over all M documents in corpus D gives:

$$p(D|\alpha, \beta) = \sum_{d=1}^{M} \int_{\mu_d} p(\mu_d|\alpha) \left(\prod_{n=1}^{N_d} p(w_{dn}|\mu_d, \beta) \right) d\mu_d \tag{1.10}$$

The above is an intuition only. In actuality, LDA assumes a generating model (i.e., a model of how the corpus was produced) that samples each μ_d once for each word in a corpus, which means that each document contains a mixture of topics, which is why each document has its own topic distribution (called a per-document topic distribution). This means that each document d can be associated with an N_d vector of topics, z_d, - one topic assignment (out of K total topics) for each word in the document. This generating model is described in more detail by

In a static topic model (LDA), it is implicitly assumed that the documents are drawn from a fixed set of topics in an exchangeable sense. However, for many collections of documents, the order of the documents reflect an evolving set of topics. In the Dynamic Topic Model (DTM) or LDASeq (Figure 1.11), we divide the data by timestamps and then model each slice of documents with a number of topics where topics in time slice t_d evolve from the topics associated with slice t_d-1. Unlike the static LDA model, the uncertainty about the distributions over words cannot be modeled by a Dirichlet prior β. We instead chain the natural parameters of each topic in a state space model which evolves with statistical noise. In the same way, the uncertainty over the per-document topic distribution in each time slice is modeled using a logistic normal distribution with a mean α. In this way topics and topic proportion distributions are chained together, sequentially tying a collection of topic models. Here, we use an LDASeq implementation in the Python *Gensim* package. We trained our models with the number of passes set to 5000 and chunksize (number of documents in a batch) set to 100. For finding the optimal number of topics, we used a parallelized implementation of LDA in *Gensim* to train individual models with topic sizes ranging from $K = 10$ to $K = 80$; each model trained using 40 shared-memory cores on a single node of a high performance cluster. Using these settings it takes on the order of a few hours to train a single model: between 3-15 hours per model on our particular machine, depending on K. Given that the LDASeq models take in the order of weeks to run, we used the parallelized static LDA for

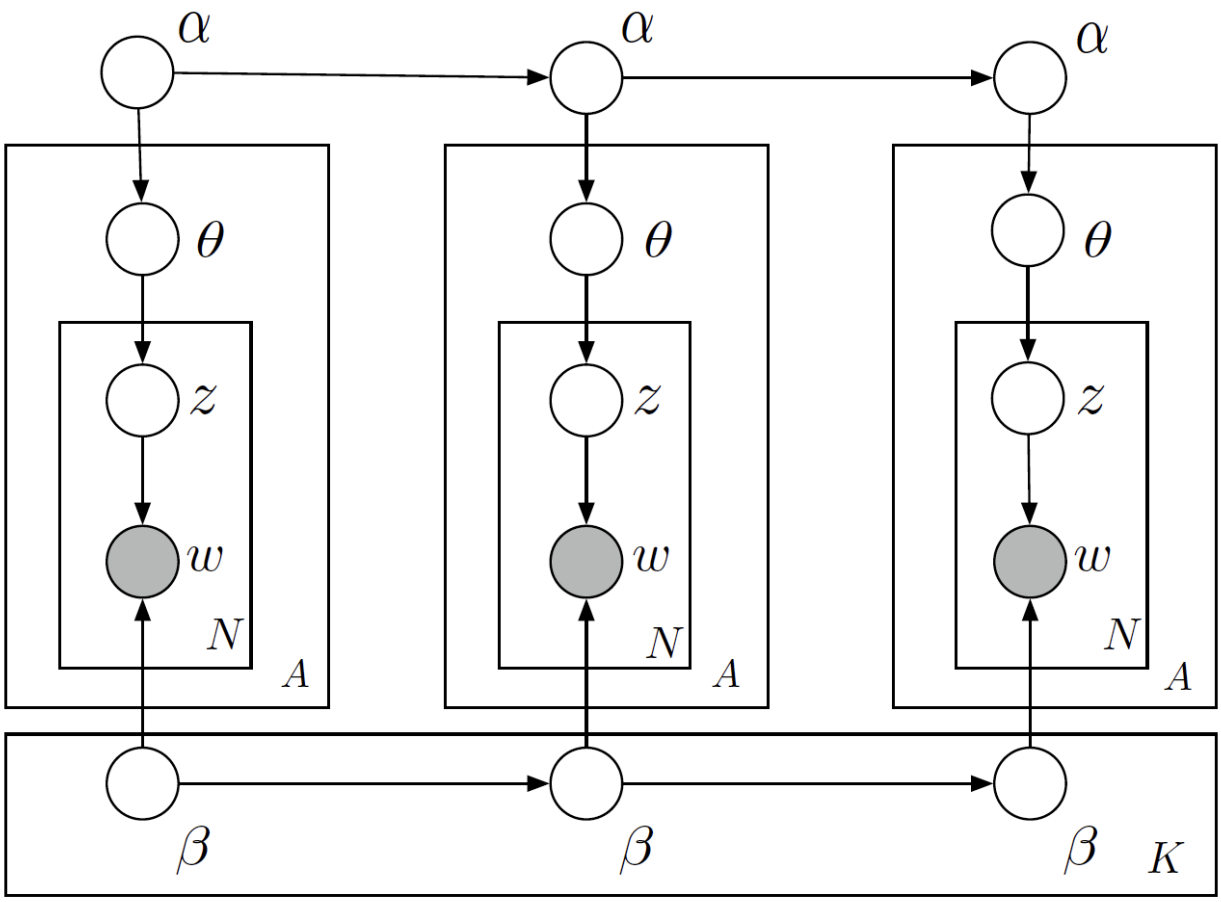


FIGURE 1.11. Graphical model of a DTM with three time slices. Natural parameters $\beta_{t,k}$ and the mean parameters α_t of the logistic normal distribution for topic proportions of each topic evolve together. A represents the slices of documents.

this analysis, because our objective was to estimate a range of topics which might be optimal for both these classes of models.

1.6.3. Choosing an Optimal Number of Topics. Ideally it is desirable to maximize the number of topics identified by LDASeq to increase variety and "depth" in terms of how the model partitions subtopics in the discipline. In practice, a number of topics, K, above some (unknown) optimal number of topics, K^{opt}, increases the occurrence of common words among different topics, resulting in compromised quality of topics (Lu et al., 2011).

We therefore adopted a hybrid quantitative/qualitative approach for deciding the optimal number of topics, K^{opt}.

1.6.3.1. *Data-Driven Approach to Choose an Optimal Number of Topics.* We used a combination of perplexity p and coherence c scores to evaluate model performance over a range of different numbers of topics. Details on how coherence and perplexity are calculated, and their underlying algorithms are given in 1.6.4.

We trained LDA models using identical hyperparameters for different numbers of topics from $K = 10$ to $K = 80$, logging the coherence c and perplexity p scores for each value of K. The goal of this multi-model training routine was to acquire a range of values of K within which K^{opt} was likely. The resulting scores are plotted in Figure 1.12. Coherence (higher is better) peaked at around $K = 25$ with substantial noise around that value, and there was no clear optimum in perplexity (lower is better). Therefore, to determine K^{opt} we additionally qualitatively considered a range of $K = 25$ to $K = 50$ (see next subsection).

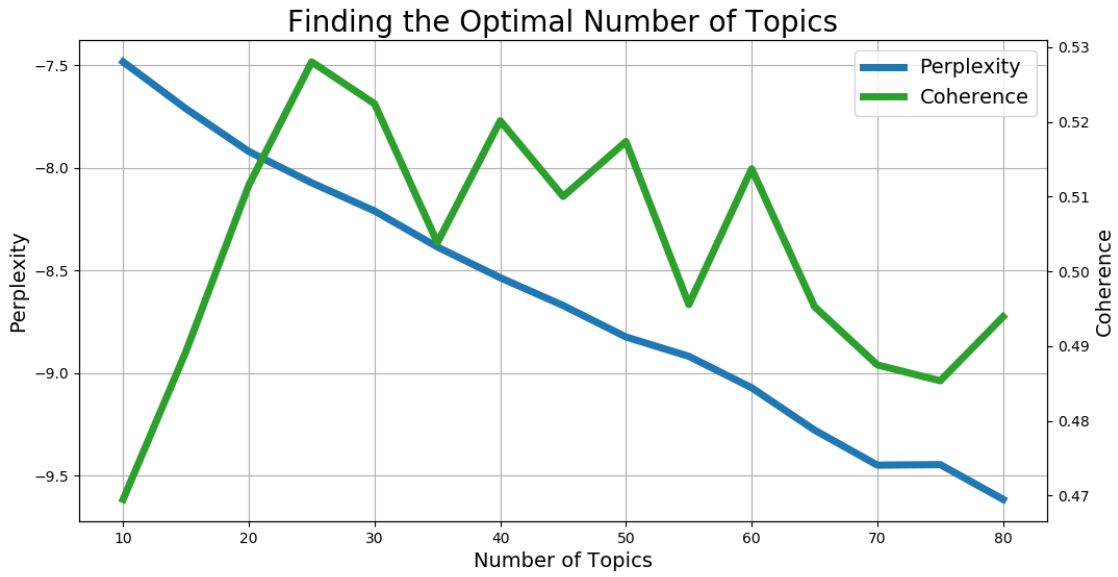


FIGURE 1.12. Variation of topic coherence c and perplexity p based on LDA models trained for a range of topic numbers ($K = 10$ to $K = 80$). Lower perplexity and higher coherence indicate a better model. These values guide our subjective analysis for choosing K^{opt}

1.6.3.2. *Qualitative Approach to Choosing Optimal Number of Topics.* Qualitative perception of topics is a common step in essentially all topic modeling research (e.g., Sun et al., 2017; Paul and Dredze, 2014; Jiang et al., 2016) and allows for data-driven evaluation metrics to be supported by manual validation. We assessed the quality of topics for various values of K, looking for increasing or decreasing occurrence of similar words within certain topics and backtracking into the dataframe to observe the titles of documents associated with each topic. We drew on our prior experience in hydrology to make these assessments, and also solicited input from several other professional hydrologists. We used the aforementioned range of values of K, and this subjective assessment to choose $K^{opt} = 45$.

1.6.4. Perplexity and Coherence. Perplexity is a popular metric for evaluating language models (Chen et al., 1998). Perplexity is an information theory metric that measures something like how surprised the model might be on the introduction of new data (Zhao et al., 2015). Formally defined by (Blei et al., 2003), perplexity for a collection of M documents is:

$$p = exp\left\{-\frac{\sum_{d=1}^{M} log\, p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{1.11}$$

Perplexity is a decreasing function of the probability assigned to each per-document word distribution. Lower perplexity indicates a better model.

Topic coherence c is a measure of similarity in semantics between the high probability words in a certain topic. We use *Gensim's* built-in topic coherence model, which is an implementation of the method described by (Röder et al., 2015). Calculating topic coherence is a four-stage process involving segmentation of word subsets, probability calculation, confirmation measure, and aggregation.

Figure 1.13 Röder et al. (adapted from 2015) illustrates these four steps. t represents an input collection of words, and the first stage creates a set of different kinds of segmentation of words S from t, since coherence measures the fitting together of words or a set of words. Secondly, probabilities of occurrence of words P are calculated based on reference corpus.

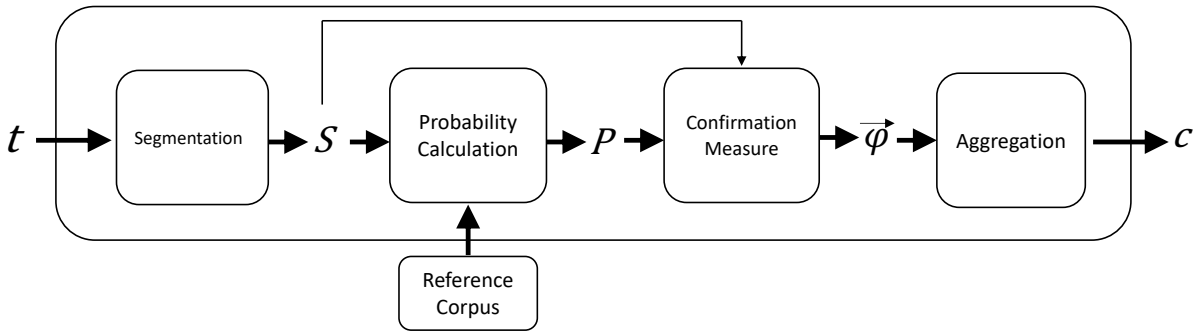


FIGURE 1.13. Illustration of the four stages of the unified topic coherence framework. In stage 1, input words t are segmented into smaller sets S. Probabilities of occurrence P of words are calculated based on the reference corpus in the second stage. In the third stage, P and S are ingested to measure φ between pairs of words S. Coherence c is calculated in the final step.

Confirmation measure ingests both P and S to yield the agreements φ of pairs of S. In the final step, the aforementioned scores are aggregated to compute coherence c.

CHAPTER 2

Drought Awareness over Continental United States

2.1. Introduction

Economic damage caused by droughts in the United States is estimated to be in the tens of billions of dollars annually (Smith and Matthews, 2015). Public perceptions and attitudes towards droughts – both pre- and post-drought – are indicators of public reception of water management and conservation measures (Campbell et al., 2004; Clarke and Brown, 2006). Adams et al. (2013) assessed the influence of attitudes and perceptions regarding multiple factors on water conservation use in nine U.S. states and found that public perception of the importance of water resources management significantly influenced water conservation outcomes. A study on the sociological impacts of drought perception in South-Central Nebraska revealed that crop and livestock producers were becoming increasingly concerned about water scarcity resulting from droughts (Woudenberg et al., 2008). Similar concerns were shared by farmers in a study conducted in the Jucar River Basin in Spain (Urquijo and De Stefano, 2016) and in South Africa (Bahta et al., 2016). After a record breaking drought in Texas in 2011, residents were significantly more concerned with water availability and water conservation (Gholson et al., 2019). A recent study on two cities in Alabama found that public awareness of drought is significantly dependent on geographic, physical, and social contexts (Shao et al., 2022). That study used Google Trends data in addition to survey data, and suggested that future studies consider Twitter data to gain a more complete understanding of social responses to drought hazards.

Understanding the complex relationship between droughts and people's perception about droughts, offers key insights for decision makers who can leverage these insights for early

warnings, public service announcements, or targeted water conservation initiatives. Recognizing shifts in sentiments has the potential to aid in tailoring effective messaging to communities. Such insights directly align with societal objectives, notably, reducing water consumption.

Evaluating public responses to drought through traditional surveys can be costly and limited in scale and interpretability due to biases (McLafferty, 2016; Vaske, 2011; Fisher, 1993; Groves, 2006). The rise of internet and social media use, with 92% of U.S. households having computer access and 85% broadband subscriptions in 2018 (Martin et al., 2021), and the availability of these data open avenues for large-scale web-based studies. Previous research has effectively used large-scale online data, including Google Trends Search Interest, as a proxy for surveys in diverse fields such as epidemiology, mental health, political science, telemedicine, population health, etc. (Carneiro et al., 2018; Yang et al., 2010; Stephens-Davidowitz, 2014; Hong et al., ????; Arora et al., 2019; Mellon, 2013).

Previous studies, differing significantly in approach and scope of this study have also explored questions surrounding public awareness of droughts. Gonzales and Ajami (2017) used the relative number of Google searches for the term 'California drought' as an indicator of public interest to suggest that multiple socio-political factors have contributed to growth in public awareness during and following droughts. Kam et al. (2019) leveraged google trends data and standard precipitation indices and found that there was no detectable social response during the onset of the 2011-2017 California droughts. Kim et al. (2019) they found that drought severity (as indicated US Drought Monitor (USDM) data) in most Northeastern and Southeastern U.S. states showed significant temporal correlations with corresponding Google search interest in droughts, while most Western and Central U.S. states displayed weaker or insignificant correlations.

In this paper, we use computer vision models trained on lagged data to investigate the non-linear, heterogeneous relationship (and their drivers) between meteorological droughts and people's search interest on the topic of "drought" - allowing for a more holistic view of

the spatial and temporal variability (at state and regional levels) of drought awareness over Continental United States (CONUS). We also complement this investigation by leveraging social media data to understand how people's sentiments towards drought in the U.S. has varied.

2.1.1. Research Questions. In this paper, we address the following specific research questions:

- Is there a relationship between droughts and drought awareness, as measured by Search Interest, and what factors drive this relationship?
- What are people's sentiments about drought, as indicated from social media posts, and have these sentiments changed over time?

2.2. Methods

In this section, we discuss the data-driven experimental approaches and methods that we used to address the research questions outlined above.

To address the first research question, we broke it down into a set of sub-queries. We first looked at the averages, trends, and variances of the state-wise search interests and SPEI across CONUS within our study period (2004-2020). We also investigated any existence of statistically significant linear correlation among the corresponding variables. We then applied six computer vision (UNet) models trained on lagged drought (input) data and corresponding Search Interest (target) data to investigate the existence of nonlinear dynamics between these variables. We then conducted a feature importance analysis to find out the most important drivers of this nonlinear relationship. We further compared these results with a Principal Component Analysis applied on the same dataset. We then explored the variability of these relationship aggregated at the state and regional levels and looked at the best lagged models per state and computed the sum of the R-squared (cumulative and population-weighted) per model predictions.

TABLE 2.1. Overview of data sources and availability

Data	Abbreviation	Source	Years Available	Years Used
Standardized Precipitation and Evapotranspiration Index	SPEI	spei.csic.es	1900-2020	2004-2020
Google Trends Search Interest	SI	Google Trends API (trends.google.com)	2004-Present	2004-2020
Tweets	NA	Twitter API (developer.twitter.com/en/docs/twitter-api)	2006-Present	2008-2020

To address the second research question, we performed sentiment analysis on a subset of 2.5 million georeferenced tweets containing drought-related terms, with the aim of tracking changing people's sentiments towards drought over time in the United States.

2.2.1. Data. Table 3.1 summarizes the three data types used in this study, including where the data was sourced and the temporal periods that we acquired and used.

2.2.1.1. *Standardized Precipitation and Evapotranspiration Index (SPEI).* We use the Standardized Precipitation and Evapotranspiration Index (SPEI) (Beguería, 2022) as our meteorological drought dataset. Intuitively, SPEI serves as a quantitative measure of droughts by providing a standardized index to assess moisture deficit over time and space. It is calculated by taking the difference between total precipitation and total potential evapotranspiration (PET) over a given period of time (e.g., monthly). SPEI is is expressed in units of standard deviations calculated over local (per pixel) climatologies, allowing us to assess whether it is drier or wetter than expected conditions. Calculating SPEI involves the following steps:

(1) Calculate the difference between precipitation (P) and reference evapotranspiration ($ET0$) for each month or time step:

$$D_i = P_i - ET0_i \tag{2.1}$$

where D_i is the difference between precipitation and reference evapotranspiration for the i-th month or time step, P_i is the precipitation for the i-th month or time step, and $ET0_i$ is the reference evapotranspiration for the i-th month or time step.

(2) Calculate the climatic water balance for each month or time step:

$$WB_i = \sum_{j=1}^{n} D_{i-j+1} \tag{2.2}$$

where WB_i is the climatic water balance for the i-th month or time step, and n is the time scale (e.g., 3, 6, or 12 months).

(3) Fit a probability distribution, such as the three-parameter log-logistic distribution with the Maximum Likelihood Estimation (MLE) method, to the climatic water balance values:

$$F(WB) = \frac{1}{1 + \left(\frac{WB-\alpha}{\beta}\right)^{-\gamma}} \tag{2.3}$$

where $F(WB)$ is the cumulative probability distribution of the climatic water balance, and α, β, and γ are the distribution parameters that need to be estimated.

(4) Calculate SPEI by transforming the fitted probability distribution to a standard normal distribution:

$$SPEI = \Phi^{-1}(F(WB)) \tag{2.4}$$

where SPEI is the Standardized Precipitation Evapotranspiration Index, Φ^{-1} is the inverse standard normal cumulative distribution function, and $F(WB)$ is the cumulative probability distribution of the climatic water balance.

SPEI can be calculated for a variety of time periods, ranging from 1 month to 48 months. We use monthly data in our analysis because Google Search Interest (target) data has the same temporal resolution. Positive SPEI values indicate wetter conditions, while negative values indicate drier conditions. We created images of monthly averaged SPEI over CONUS for each month from January 2004 to December 2020 to be used as input data for training and prediction for our analysis. Fig 2.1 shows an example of one of these SPEI maps.

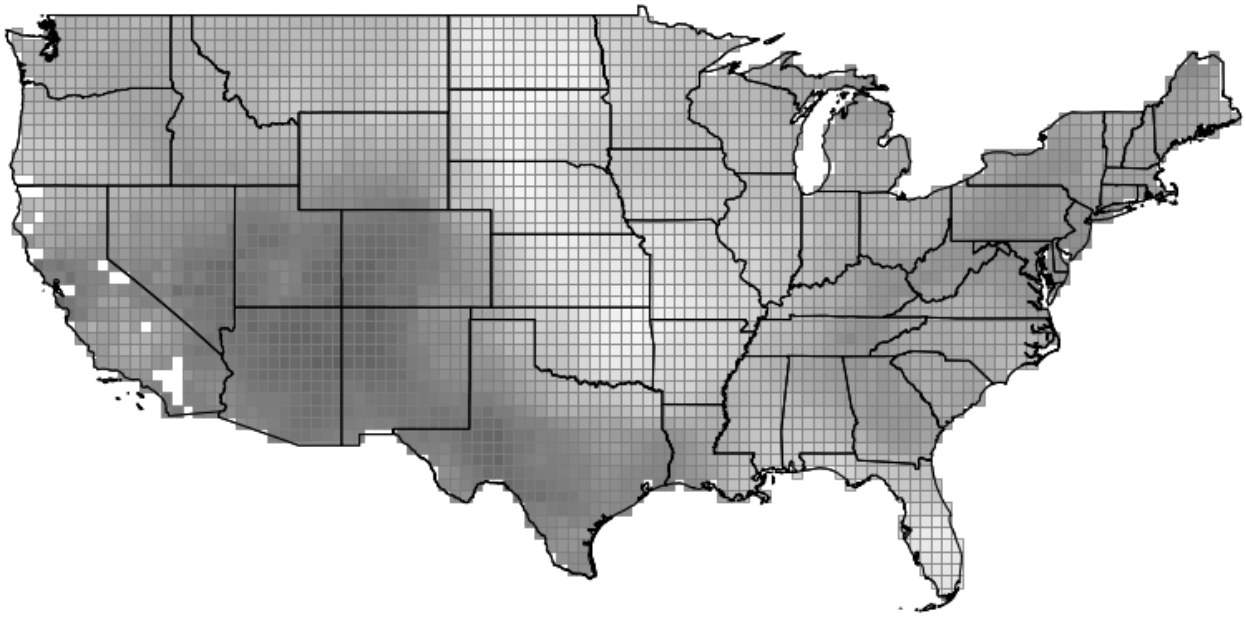


FIGURE 2.1. Example of a SPEI map fed into our UNet models(Date: 08/2019). The image is in grayscale (darker shades represent lower SPEI - drier conditions), without coordinates, grids, or legend consistent with our actual input images.

2.2.1.2. *Google Trends Search Interest (SI).* We acquired Google Search Interest (SI) data from Google Trends using the Trends API. The Trends API allows programmatic access to Google trends data and track the popularity of different topics over time and place. We selected "Drought" as a topic. When a topic is selected, Google Trends aggregates all searches related to the concept of drought, regardless of the specific terms or phrases used. This includes searches in different languages and searches using related terms (e.g., "water shortage," "dry weather"). The topic approach provides a broader view of public

interest or concern about drought as a general subject. It captures the overall awareness and information-seeking behavior related to drought without being limited to a specific keyword. Given a specific topic, T, and a time range from t_1 to t_n, the search interest for T at each time point, t_i, is calculated as:

$$SI(T, t_i) = \frac{S(T, t_i)}{S_{\max}(T)} \times 100 \tag{2.5}$$

where $SI(T, t_i)$ is the search interest for the topic T at time point t_i, $S(T, t_i)$ is the search volume for T at t_i, and $S_{\max}(T)$ is the maximum search volume for T within the specified time range. We use monthly state-wise SI data on "Drought" topic from 2004 to 2020 to create maps over the CONUS. Fig 2.2 shows an example of a Google SI target map.

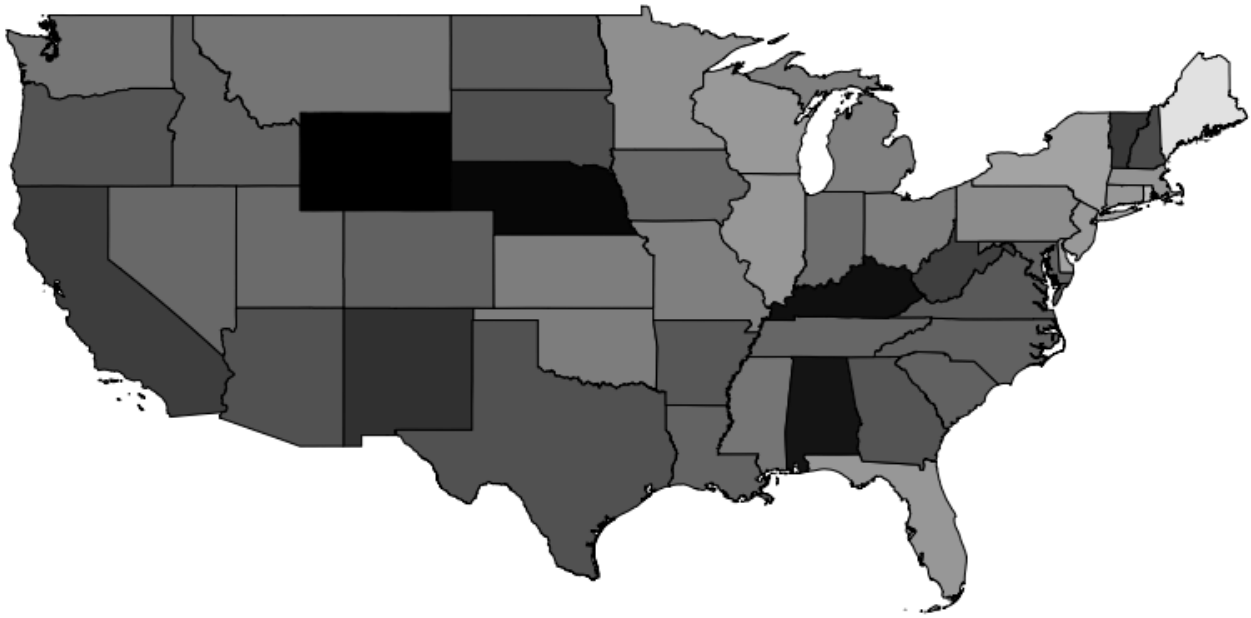


FIGURE 2.2. Example of a Google SI map (Date: 08/2019) used as a target image during model training and testing. The image is in grayscale(darker shades represent higher search interest), without coordinates, grids, or legend) consistent with our actual target images used for model training and testing.

2.2.1.3. *Tweets.* We acquired Twitter data (Tweets) using the Twitter API. Our data set consists of 2.5 million (restricted to English language) tweets related to the "Drought" topic from 2008 to 2020. We use the Twitter data primarily for sentiment analysis. Since

the Twitter data was only partially geotagged with user supplied locations, we leveraged the GeoNames gazetteer database (Ahlers, 2013)to tag tweets which reference a location within the United States. We then applied the Sentiment Analysis to those tweets.

2.2.2. Analysis Methods.

2.2.2.1. *Search Interest and SPEI Averages, Trends, and Variances.* Both droughts and people's search interest in droughts varies spatially and temporally in a multi-directional (complex and nonlinear) manner rather than follow a steady trend. We first calculated the averages, trends (using linear regression), and variances aggregated by states for both search interest and SPEI. We then investigated for any statistically significant linear correlation between these corresponding variables.

2.2.2.2. *Relationship between droughts and search interest.* We used machine learning models to investigate the non-linear relationship(s) between meteorological droughts and search interest over CONUS. Specifically, we trained U-Net models (Ronneberger et al., 2015) to predict SI from SPEI maps. Details of our model architecture, training, and evaluation are in Appendix 2.6.1). In summary, we trained 6 models on 6 sets of lagged SPEI maps as inputs (from 0 months lag to 5 months lag) and the target data were their corresponding search interest maps. Correlation between (out-of-sample) search interest predictions made by these trained models and real search interest data is an estimate of the non-linear correlation between SPEI and SI at a given lag time. We evaluated the models on time periods not used for training (training period: 01/01/2004 - 07/31/2017 and test period: 08/01/2017 - 12/31/2020). This approach, aligning with the standard 80/20 training/testing data split, ensured a robust assessment of the models' ability to generalize across different time frames and drought scenarios. We report the models' performances based on the per pixel coefficient of determination (R^2) between the model predicted and observed search interest during the test period:

$$O_{a,b}(t) = U(T_{a,b}(t)) \tag{2.6}$$

where $O_{a,b}(t)$ is the U-Net output for pixel (a, b) at time t, $T_{a,b}(t)$ is the input data for pixel (a, b) at time t, and U is the U-Net model function.

We then tested the model performance over time (using the coefficient of determination or R^2) for each individual pixel.

$$R^2_{a,b} = R^2(O_{a,b}(t), G_{a,b}(t)) \tag{2.7}$$

where $R^2_{a,b}$ is the coefficient of determination for pixel (a, b), $G_{a,b}(t)$ is the ground truth (SI) data for pixel (a, b) at time t, and R^2 is the coefficient of determination function. We then created a binary mask array based on the geometry of CONUS and constructed heatmaps of the R^2 values averaged over states and then aggregated to regions (as defined by the US Census Bureau) (U.S. Census Bureau, 2021). This approach allows us to observe overall model prediction performances over time across CONUS, and also by individual states and regions. We further evaluated the strength of correlation and significance of the statistical relationship between the model R^2 and corresponding averaged, trends, and variance of search interest and SPEI.

For contrast, we applied Principal Component Analysis (PCA) on the same dataset. The PCA served as a linear, dimensionality reduction technique to identify the major patterns of variance in droughts and corresponding people's search interest state-wise and regionally. Our objective was to evaluate the strengths and limitations of nonlinear versus linear methods in capturing the dynamics of drought awareness.

2.2.2.3. *Drivers of the relationship between droughts and drought awareness.* We investigated the most important drivers of the relationship between droughts and drought awareness. We first generated the histograms per state to visualize the distribution of (i) SPEI, (ii) trends of SPEI, (iii) variance of SPEI, (iv)length of droughts, and (v)severity of droughts.

To calculate the length of droughts, we first selected a threshold of SPEI value of -1.0 and below to classify a drought occurrence based on commonly accepted standards in drought assessment. We recorded the number of consecutive months a pixel was in drought condition considering it as a continuous drought period and then aggregated the results by state. For drought severity, we identified all time points per pixel in drought conditions within the study period and computed the mean SPEI value.

For each metric, we fitted an appropriate statistical distribution to the data. We assumed a Weibull distribution for drought lengths and normal distribution for rest of the metrics. We then extracted the parameters (for normal distributions: mean, standard deviation, skewness, kurtosis, interquartile range, and for Weibull distribution: form and scale) of these distributions. We then We trained a Random Forest Regressor model using the parameters derived from the distribution fits as input features, with the $\mathbf{R^2}$ values from our Unet models as the target variable. We then conducted a feature importance analysis to identify the drivers that most influence the relationship between droughts and drought awareness.

2.2.2.4. *Best Lagged Model Per State.* In our analysis, the term 'lag' refers to the time lag between variation in SPEI and corresponding changes in people's search interest. We mapped the models (between 0 month lagged and 5 months lagged) which demonstrate the best overall predictive value per state.

To perform this analysis, we followed these steps:

First, we calculated the highest R^2 value for each model:

$$max\_r\_squared_{a,b} = \max_{k=1}^{6} R^2_{a,b,k} \tag{2.8}$$

where $max\_r\_squared_{a,b}$ is the highest R^2 value for pixel (a, b), and $R^2_{a,b,k}$ is the R^2 value for pixel (a, b) in model k.

Then we assigned each pixel with the highest R^2 value:

$$best\_model\_index_{a,b} = \arg\max_{k=1}^{6} R^2_{a,b,k} \tag{2.9}$$

where $best\_model\_index_{a,b}$ is the index of the model with the highest R^2 value for pixel (a, b). We then aggregated these values to a state level.

We also computed the sum of per model R^2 values and compared it with the population-weighted sum per model. The population weights were imparted by multiplying the averaged per-state per-model R^2 with the per-state population of 2020 according to the US Census Bureau.

2.2.2.5. *Sentiment analysis on Twitter data.* We performed sentiment analysis on 2.5 million tweets related to drought topics. Sentiment analysis allows us to understand the tone of human generated texts. We used VADER (Valence Aware Dictionary and Sentiment Reasoner) – a lexicon and rule-based sentiment analysis tool that is specifically designed to work with social media data sets (Hutto and Gilbert, 2014). The overall sentiment score (compound score) assigned to a tweet is a number between -1 and 1. A score of -1 points towards a very negative sentiment while a score of 1 indicates a very positive sentiment. A score of 0 indicates a neutral sentiment. Fig. 2.3 demonstrates some examples of how some tweets about drought can receive a positive compound score (positive sentiment).

To analyze people's sentiment about droughts within the United States, we first randomly sampled a fraction (20%) of the tweets per year from our corpus. We then leveraged the US GeoNames data from the GeoNames database which contains detailed geographic information about locations within the United States, including place names (in various forms), latitude and longitude coordinates, and other geographic identifiers (Ahlers, 2013). We then defined a function to check if the text from tweets (either from a location-related field or the tweet text itself) contains any place names listed in the geonames dataframe. We iterated through unique place names in the GeoNames data and checked for their presence in the tweet text, marking the tweet as related to the US if any matches were found.

We then applied a time-series analysis of the percentages of different sentiments on tweets pertaining to the United States. We also calculated the overall percentages of these sentiments.

tweet	sentiment	Month
@RadHertz But a wetter climate - 9 days straig...	pos	2008-01
Govt tight-fisted on drought relief: Nelson: F...	pos	2008-01
Nelson cautious on drought relief plans: Feder...	pos	2008-01
NFF welcomes talks with Govt on drought assist...	pos	2008-01
listening to the rain. After a 7 year drought,...	pos	2008-01

FIGURE 2.3. Examples of tweets about droughts which were assigned with positive sentiments

2.3. Results and Analysis

In this section we address the research questions laid out previously and discuss the results of our analyses.

2.3.1. Is there a relationship between droughts and drought awareness as measured by search interest, and what factors drive this relationship? In order to gain a comprehensive understanding of the dynamics between droughts and drought awareness, we break down this complex question into a subset of smaller questions.

2.3.1.1. *How have drought conditions and drought awareness varied across CONUS from 2004 to 2020?* We first look at the averages, trends, and variances of search interest on droughts and SPEI across CONUS (Figure 2.4). Our objective with this analysis is to provide an overview of droughts and drought awareness and investigate any simple linear correlation between them within our study period of 2004-2020.

The average search interest map (Figure 2.4(A)) demonstrates the spatial variations in public awareness of droughts. Overall drought awareness were the highest for California, New Mexico, South Dakota while the lowest awareness were seen for Mississippi, Tennessee, Ohio, Pennsylvania, New York, and Florida. California and New Hampshire had the strongest increasing trend of drought awareness between 2004 and 2020, while the rest of the CONUS states demonstrated slightly increasing to stable trends (Figure 2.4(B)). We observed the highest degrees of variability in drought awareness (Figure 2.4(C)) for Wyoming, California, North and South Dakotas, Montana, New Mexico, Vermont and New Hampshire.

The average SPEI map (Figure 2.4(D)) provides an overview of climatic tendencies in moisture balance (droughts) across CONUS states. The overall driest conditions were seen in Arizona, Georgia, Oregon, New Mexico, Utah whilst most Midwestern and Northeastern had wetter conditions relative to the rest of the CONUS. The strongest increasingly dry to drought conditions were observed over Arizona, New Mexico, California, Utah, Nevada, Colorado, Maine, Vermont and New Hampshire (Figure 2.4(E)). Conversely, strongest increasingly wet conditions were observed for North Carolina, Virginia, Tennessee, Kentucky, Massachusetts, and Wisconsin. The highest fluctuations in drought conditions were seen over Texas, Florida, and Michigan (Figure 2.4(F)), indicating the least predictable moisture conditions over these states. These variances were also notable for California, Idaho, Colorado, Montana, and North Carolina while the least variances were seen for majority of the New England states.

We found statistically significant inverse relationship (R = -0.64, p = 7.61e-07) between the trend in SPEI and search interest (Figure 2.4(H)), indicating that drought awareness has generally increased in states with increasingly dry to drought conditions. We did not find any statistically significant correlation between averages of and variances in SPEI and search interest.

2.3.1.2. *Are there nonlinear relationships between droughts and drought awareness?* As described in the methods section, we trained six UNet models (of the same architecture and

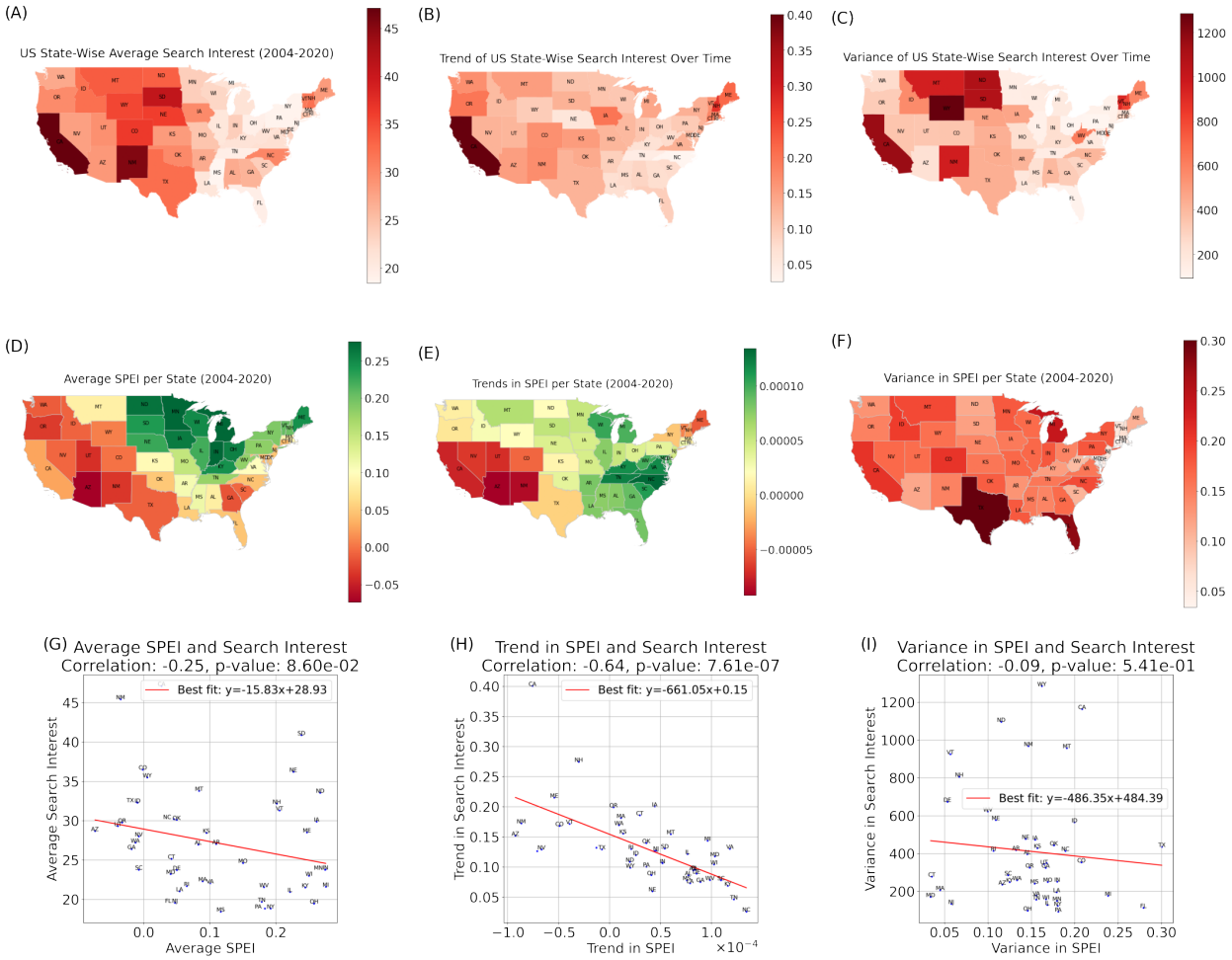


FIGURE 2.4. This panel shows the per-state (over CONUS) search interest and SPEI averages, trends, variances, and their (linear) correlations. (A) Average search interest for drought-related terms per state - darker shades indicate increased public awareness or concern about drought conditions. (B) Trend of drought search interest over time - showing the magnitude of the slope from a linear regression analysis of search interest against time. (C) Variance in search interest per state, with color intensity denoting the degree of variability in public search behavior. A higher variance (darker shade) suggests fluctuating public interest, possibly in response to episodic drought events or media coverage. (D) Average SPEI per state - greener shades indicate wetter average conditions, and red shades depict drier conditions. (E) Trends in SPEI per state - gradient reflects the slope of the trend, where green represents an increasing wetness trend, and red indicates an increasing dryness trend. (F) Variance in SPEI per state, capturing the fluctuations in drought conditions - darker shades of red states have experienced more significant variability in SPEI. (G) Correlation between average SPEI and search interest across states. (H) Correlation between the trend in SPEI and search interest. (I) Correlation between the variance in SPEI and search interest.

hyperparameter settings but trained on 6 sets of lagged data) and tested their predictions for the period of June 2017 to December 2020. We measured the association (R^2) between the per-pixel values of target images and predicted pixel values by the U-Net models, and then aggregated and averaged them by individual states (Fig. 2.5). The U-Net's capacity for capturing nonlinear relationships is a key in this analysis. Results highlight significant nonlinear associations between variability in SPEI and corresponding changes in drought awareness. Although there is some degree of variability between the models, some of the common strongest nonlinear associations were seen for Colorado, Nebraska, and Arizona, suggesting heightened awareness in response to fluctuating drought conditions, while the least associations were observed over North Dakota, Mississippi, and New York.

2.3.1.3. *What factors drive this relationship?* As mentioned in section 2.2.2, we extracted a diverse range of parameters from the distributions fitted to SPEI, SPEI trends, SPEI variances, drought length, and drought severity data. We then trained a Random Forest Regressor using these parameters as input features and our UNet model R^2 as targets and carried out a feature importance analysis to identify the most important factors driving the nonlinear dynamics between droughts and drought awareness.

Figure 2.6 shows the results of the feature importance analysis. The interquartile range of the trend data was the most influential feature, showing that middle 50% spread of SPEI changes over time is highly predictive of the UNet model R^2 - indicating that periods of moderate variability in drought conditions are significantly important to the shifts in public search interest on droughts. SPEI median was the second most important feature. Variability of SPEI trends was another important feature indicating that both average fluctuations and mean fluctuations of drought conditions were key factors driving the dynamics between droughts and drought awareness. Drought severity range, drought mean, and drought trends mean were also important drivers. SPEI variance kurtosis and trend kurtosis were both key drivers in this relationship, suggesting that the 'tailedness' or the presence of extreme

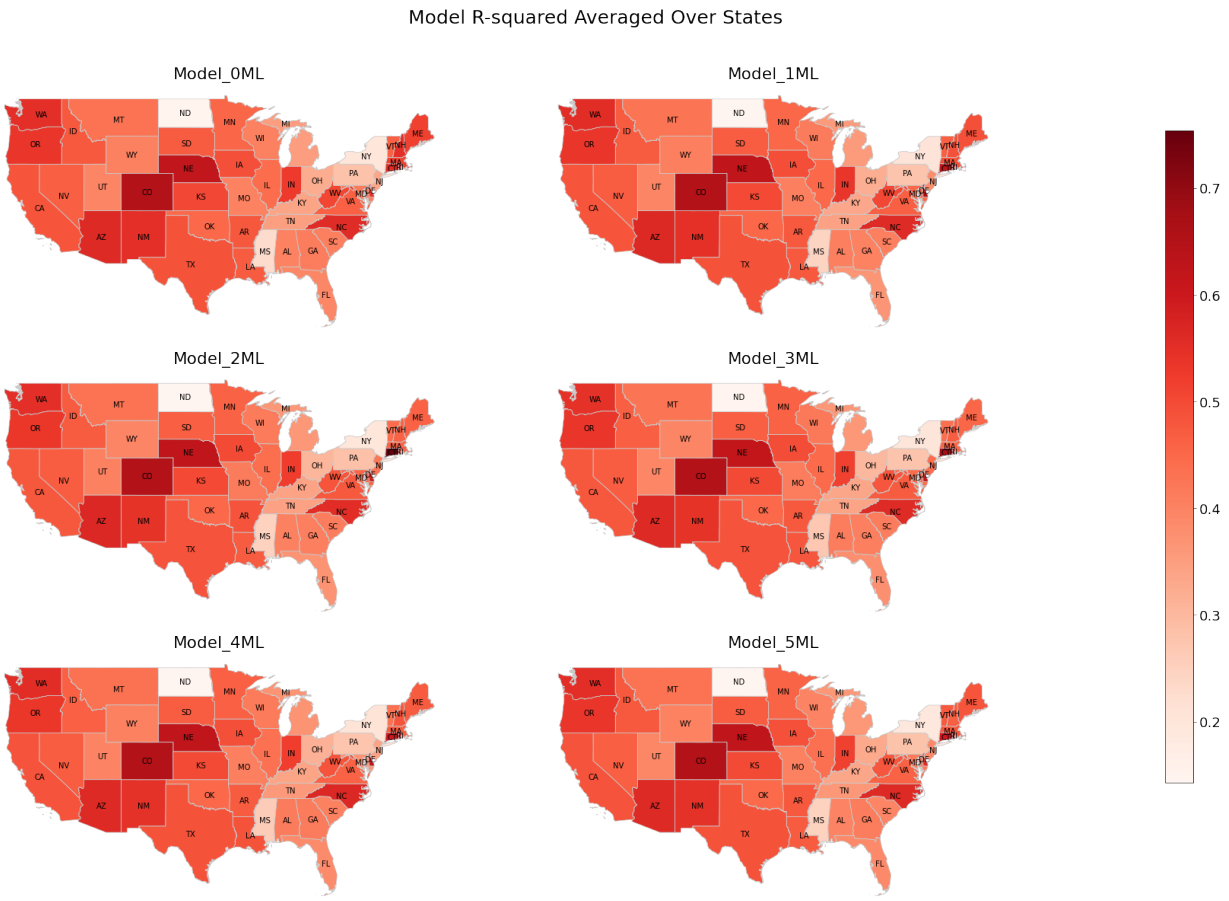


FIGURE 2.5. R^2 values per-pixel averaged over per-state for the six models trained on lagged SPEI input data. Here 'ML' indicates 'Months Lagged'(0-5 months). Significant nonlinear correlations are observed for all the six model predictions.

variability in drought conditions (outliers) significantly influenced the nonlinear dynamics between droughts and drought awareness.

2.3.1.4. *How do the nonlinearity captured by the UNets and the PCA's dimensionality reduction compare at the state and regional levels?* We aggregated the UNet models' average R^2 over states and regions (Fig. 2.7). We observed spatial variability of the average R^2 in both state and regional levels, with some degree of association found all over CONUS. The strongest relationships were observed for the Western region, followed by the Northeastern, Southeastern, and the Central regions.

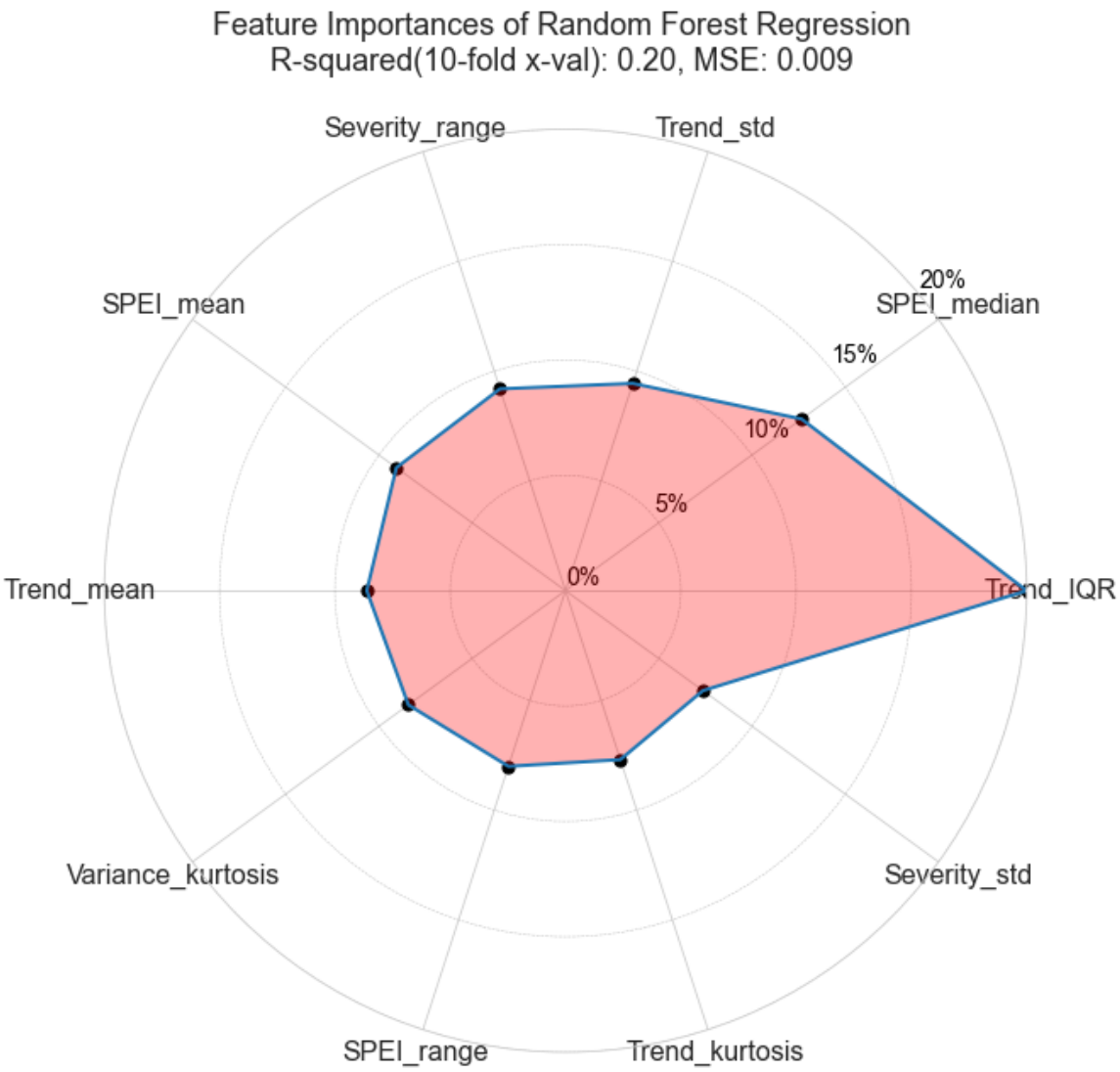


FIGURE 2.6. Ranked feature importances derived from the Random Forest Regressor model predicting UNet R^2 values. R^2 score from 10-fold cross-validation is 0.20, and the mean squared error (MSE) is 0.009

We additionally compared these results with the results of a PCA (a reductionist approach with two components) performed over our dataset (SPEI and Search Interest). The objective of the PCA was to identify the principal components (PCs) that captured the largest variances within our dataset. We found that PC1 had an explained variance of 25%, and

PC2 had roughly 17%. We represent the spatial variability (by states and then aggregated by regions) of PC1 scores (indicating how regions relate to the patterns of variance captured by the principal component) in Fig. 2.7 and found the highest PC1 scores for California, South Dakota, and New Mexico while there was no to negative relationship for the other states. Regionally, we also saw some positive PC1 scores for the Western states with none or negative scores for the other regions. These findings further point towards significantly nuanced interactions between droughts and drought awareness - likely better captured by our UNet models (non-linear approach) which can account for the various dimensions and scales at which these interactions occur.

Kim et al. (2019) previously applied PCA to understand the spatiotemporal patterns of drought awareness over CONUS. Our findings are different to Kim et al. (2019)'s. We designed our Computer Vision models to capture the complex nonlinear dynamics between droughts and drought awareness which linear and reductionist methods such as PCA might miss. Our method also allows pixel-wise prediction capability which is not possible with PCA. There are differences between their dataset and ours in terms of the periods of analysis as well as the drought indicators used, however, it makes sense that the complex interplay between any indicator of droughts and drought awareness would be better captured by nonlinear methods rather than linear ones.

2.3.2. Is there a time lag between droughts and changes in drought awareness?

We mapped the models with the best overall R^2 values per state (color-coded to indicate the lag time from zero to five months) and found a very nuanced spatial distribution across CONUS (Fig. 2.8). We also quantified the overall and population-weighted sums of the models' R^2 values to comparatively assess each model's explanatory power across CONUS. We found that search interest tends to lag changes in drought conditions mostly between 1-3 months. There are multiple factors which could drive such observations, including for example, whether a region primarily relies on rain or water storage for agriculture. Several studies have discussed the diverse impact of droughts and ensuing water scarcity on public

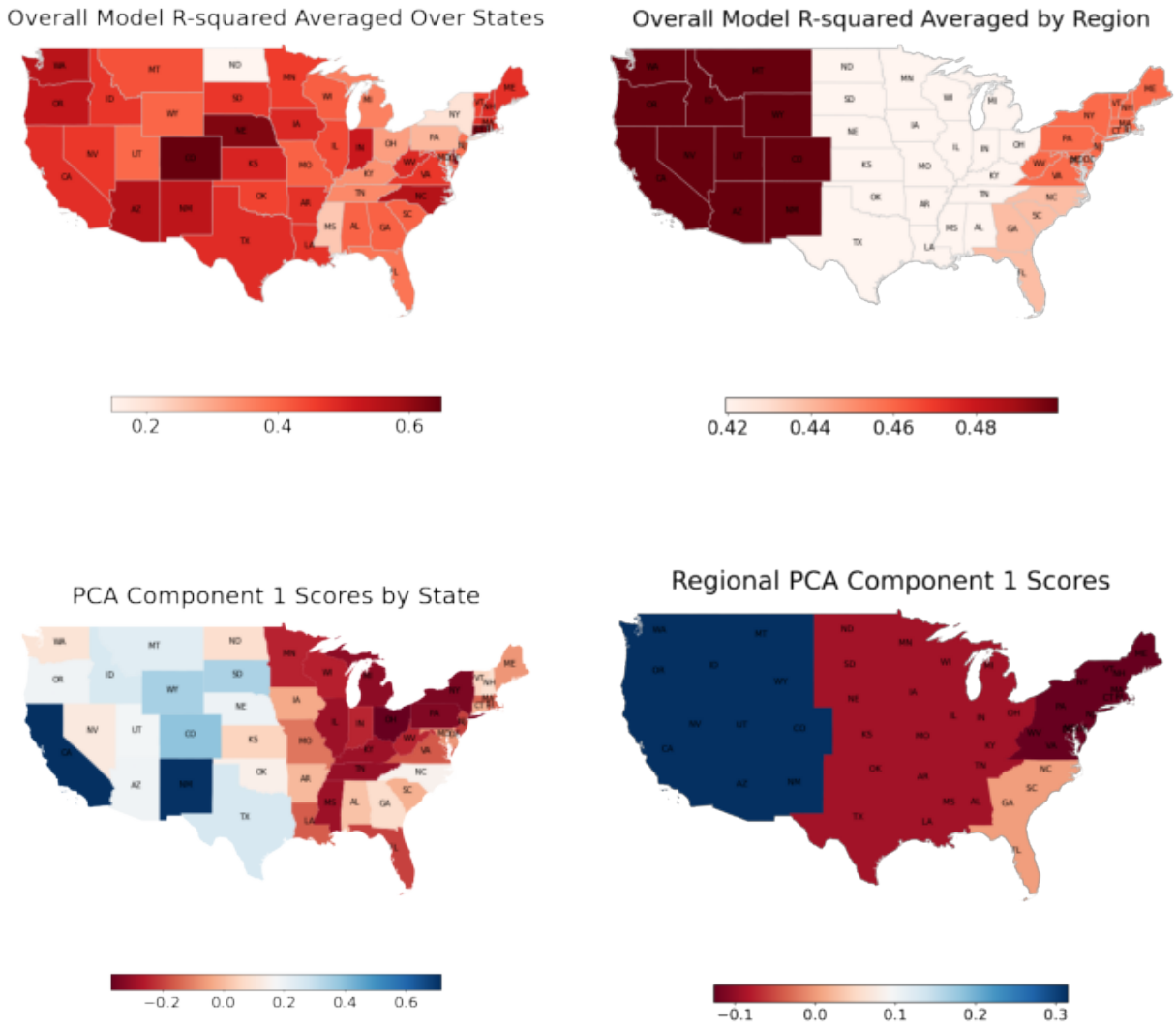


FIGURE 2.7. Comparative analysis of UNet model R^2 and PCA results at the state and regional levels. The top left map illustrates the overall UNet models' R^2 values averaged over states. The top right map shows the average values aggregated by region. The bottom left map presents the PCA Component 1 scores by state. The bottom right map portrays regional PCA Component 1 scores.

reactions and opinions. For example, AghaKouchak et al. (2014) argued that concurrent droughts and extreme heatwaves from climate change have significant social implications, and they could influence public opinions and reactions. Drier conditions are increasingly linked to public health issues (Barrera et al., 2023; Moyo et al., 2023) and vegetation health as well (Chilakala et al., 2023; Bao et al., 2023), which impact people's interest in droughts.

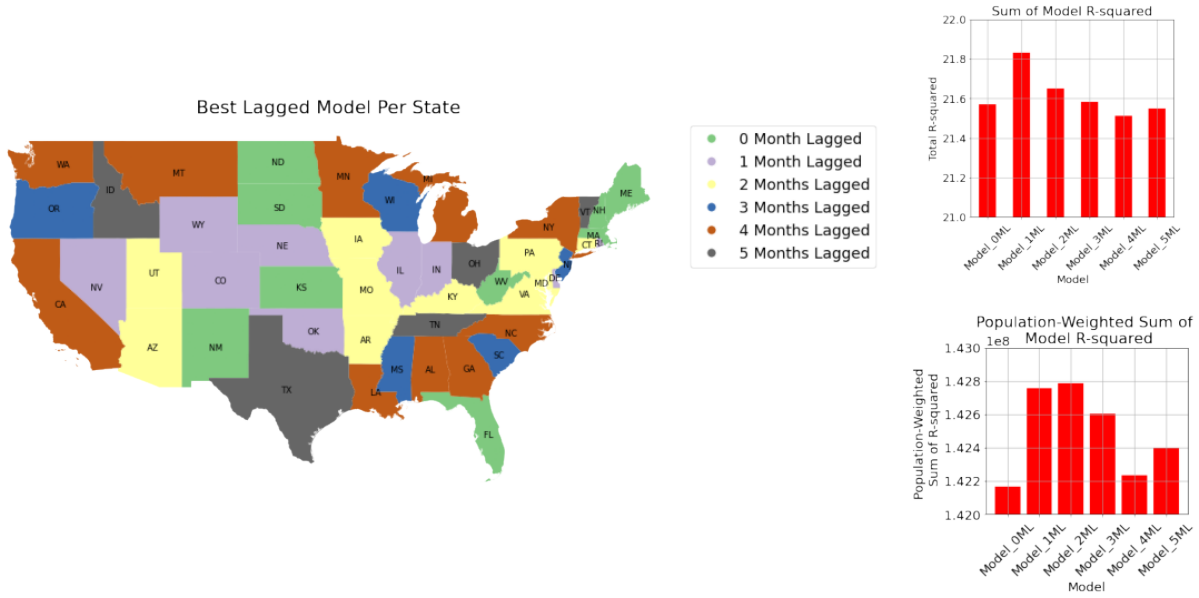


FIGURE 2.8. Panel showing the best lagged UNet model per state (left), overall and population-weighted sums of models' R^2 (right). Here 'ML' should be understood as 'months lagged', indicating models trained on lagged data from 0 to 5 months.

2.3.3. How have people's sentiments about droughts changed over time? As previously discussed in section 2.2.2, we applied sentiment analysis on 2.5 million tweets containing drought terms and analyzed a subset of these tweets which point to a location inside the United States. Figure 2.9 shows the timeseries of the percentages of these sentiments. We observed a clear increasing trend (pre-2014: R= -0.919, p-value = 0.009462 and post-2014: $R = 0.9378$, p-value = 0.00179) of positive sentiments in tweets post-2014 when it started rising sharply, contrary to it's clear decreasing trend from 2008 to 2014. Conversely, we saw a clear positive trend of neutral sentiments from 2008 to 2014 and a clear decreasing trend post-2014 (pre-2014 neutral sentiment: $R = 0.898$, p-value = 0.015, post-2014 neutral sentiment: R = -0.9687, p-value = 0.0003). There were no clear trends seen for percentages of negative sentiments.

The inflection point around 2014 possibly point towards some significant environmental policy shifts at the national and/or global levels at that time. California's groundbreaking Sustainable Groundwater Management Act (SGMA) was passed by the California Legislature

in 2014 and became effective in January 2015 (Leahy, 2015). This act was a response to the severe drought conditions and the largely unregulated use of groundwater resources in California. In January 2014, then US President Barack Obama's State of the Union address focused on environmental policies. Another major policy implementation around that time include the Clean Power Plan of 2015, developed under the Clean Air Act which aimed to set statewise targets for carbon emmissions reductions (Carbonell, 2014). Broad public discourse on climate change related topics (which include droughts) also surrounded a Nebraska judge's ruling in February 2014 that the governor's approval of the Keystone XL pipeline was unconstitutional (Hestres, 2014). The IPCC reports from 2014 and 2015, which were part of the IPCC Fifth Assessment Report (AR5), also drew global attention towards increased frequency and intensity of extreme weather events, including droughts (Mach et al., 2016).

Our findings also suggest that the public sentiments towards droughts may be becoming more polarized (decreasing percentage of neutral sentiments). This is likely due to multiple factors which could include increasing severity of droughts, increasing politicization of climate change (McCright and Dunlap, 2011; Chinn et al., 2020; Bolsen and Druckman, 2018), increased regulations over groundwater to prevent groundwater overdraft (Gage and Milman, 2021), and spread of misinformation over the social media sphere (Treen et al., 2020; Freiling and Matthes, 2023). One possible reason for these trends could be that as drought conditions worsen or become more prevalent, public awareness and concern grow, leading to more online searches and discussions about the topic (Brulle, 2010). Positive sentiments could arise from discussions around successful drought management strategies, water conservation efforts, or community resilience. Negative sentiments might stem from the adverse impacts of droughts, such as crop loss, water shortages, and the associated socio-economic hardships. It can also be said that the rise in sentiment polarity is influenced by increasing public engagement with environmental issues more broadly. As discussions around specific environmental phenomena

(like droughts) become increasingly popular, it is reasonable to think that they will become sentimentally charged and polarized.

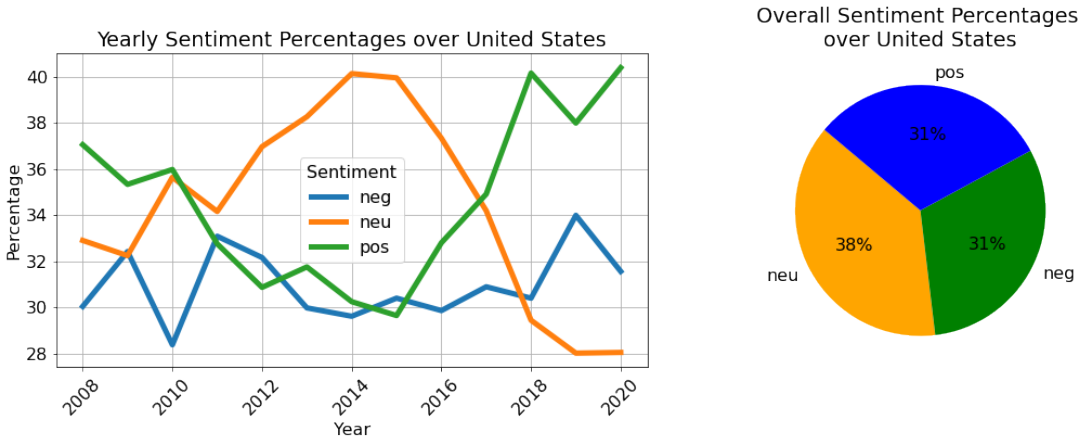


FIGURE 2.9. Sentiment analysis of tweets related to droughts inside the United States. Each tweet is assigned with a positive, neutral, or negative sentiment. The time series of the percentages of the sentiments are represented in the left figure and overall the percentages of the sentiments are presented in the pie chart to the right. Trends and significance of the time series - pre-2014 negative sentiment: $R = 0.106$, p-value = 0.842, post-2014 negative sentiment: $R = 0.7052$, p-value = 0.076, pre-2014 neutral sentiment: $R = 0.898$, p-value = 0.015, post-2014 neutral sentiment: R = -0.9687, p-value = 0.0003, pre-2014 positive sentiment: R= -0.919, p-value = 0.009462, post-2014 positive sentiment: $R = 0.9378$, p-value = 0.00179.

2.4. Conclusions & Discussion

We applied U-Net models and compared them with linear and reductionist approaches to understand the complex dynamics between droughts and drought awareness within CONUS. To do this, we leveraged $SPEI$, Google Trends SI, and Twitter data. The primary findings of this study are (see the research questions outlined in Section 3.1.1):

- We found significant nonlinear relationships between droughts and drought awareness as evidenced by the spatial variability of the UNet models' R^2 values between predicted and observed drought awareness.
- We discovered that the most important drivers of this nonlinear relationship are the variability and extent of drought trends and severity, as well as climatic extremes.

- We also found that reductionist linear approaches, such as a PCA, might not be effective in capturing the nuanced relationship between droughts and drought awareness at various dimensions and scales.
- The strongest nonlinear associations were observed over the Western CONUS, followed by the Northeastern, Southeastern, and lastly the Central regions.
- We found that changes in drought awareness tends to lag changing drought conditions by around 1 to 3 months.
- Upon applying sentiment analysis to a subset of 2.5 million tweets about droughts which reference a location within the United States, we found that post-2014, the percentage of tweets with positive sentiments about droughts have increased notably, while the percentage of tweets with neutral sentiments have reduced significantly. We also found that people's reactions to droughts may be becoming more polarzied.

Our findings strongly underscore the nonlinearity of the dynamics between droughts and drought awareness within the CONUS. We also unraveled the complexity of this relationship by analyzing the driving factors behind it. While dimensionality reduction with linear methods such as the PCA might provide some insights, our custom UNet models (with pixel-level predictions) were significantly better at capturing these interactions and explaining the variances of the variables involved. We also saw that the temporal variability of people's sentiments about droughts coincide with major policy implementations and public discourses about droughts.

Our observations provide significant insights into changes in human behavior with varying climatic conditions across the CONUS. We expect these findings to better inform the stakeholders at various levels making decisions on water resources management, conservation, policies, legislation, communication, and education. Future studies can build on our research by exploring (and modeling) the relationships between various natural and anthropogenic driven phenomena and anthropogenic responses. We particularly encourage future studies

to understand meteorological droughts' impact on agricultural, hydrological, and socioeconomic droughts and quantifying the multidimensional non-stationary interplay between these phenomena and their corresponding human responses at various scales. The results of our study boosts the feasibility of training large-domain drought prediction models in the sense that we provided the evidence that we can now better predict changes in drought awareness based on predicted drought conditions. As the impacts of anthropogenic-driven climate change become increasingly felt and realized by humans across the world, we expect future experiments to discover and decode more complex relationships between extreme weather events (such as droughts) with people's engagement on different platforms across the global web. Given this scenario, computer vision models, such as our custom U-Nets, will continue to be significantly useful towards understanding (evolving) human-drought interactions.

2.5. Acknowledgements(Project Specific)

This study was supported by the Google Cloud Research Credits program with the award GCP19980904.

The authors appreciate the cooperation from the Google Trends team for providing access to Google Trends data through the Google Trends API.

The authors also appreciate the Twitter development team for providing access to Twitter data through the Twitter API.

The code to reproduce all results and figures are available at www.doi.org/10.5281/zenodo.8212808. Due to data and privacy restrictions, SPEI, Google, and Twitter datasets are not made publicly available.

2.6. Appendix - Chapter 2

2.6.1. The U-Net Model. Computer vision techniques such as image segmentation has surged in popularity in recent years, with applications in scene understanding, medical image analysis, robotic perception, video surveillance, augmented reality, and image compression among others (Minaee et al., 2021). Deep learning models such as Convolutional

Neural Networks (CNN) are being increasingly applied in environmental sciences, e.g., air quality modeling (Qin et al., 2019), image classification (Hasan et al., 2019; Zeng et al., 2019) etc. In hydrological sciences, CNNs have been used for lake water level forecasting (Barzegar et al., 2021), prediction of groundwater potential mapping (Panahi et al., 2020), hydrological time series forecasting modeling (Guo-yan et al., 2019; Hussain et al., 2020), rainfall forecasting (Tu et al., 2021; Adaryani et al., 2022), flood susceptibility mapping (Wang et al., 2020), daily runoff prediction (Song, 2022), evapotranspiration estimation (Li et al., 2022), and rapid production of fluvial flood inundation (Kabir et al., 2020) among other applications. In this study, we approach understanding human-meteorological drought(MD)interactions using large-scale (CONUS scale) maps/images, making CNNs ideal candidates for our objectives. We used U-Net, which is a deep learning model used that was developed for image segmentation to learn relationships between meteorological droughts and search interest. Inputs to the model are SPEI maps over the CONUS and targets are Google SI maps over the CONUS.

2.6.2. Model Architecture. The U-Net is a convolutional neural network with an encoder-decoder architecture. Our model architecture (Fig. 2.10) has two encoder blocks. Each encoder block has two convolutional layers followed by ReLU activation functions and one Max-pooling layer. The first encoder block captures low-level features of the input images, e.g. edges, corners and textures. The second encoder block builds on the first one by capturing higher level features and patterns. A middle block consisting of two convolutional layers and ReLU activation functions processes the high-level features captured by the encoder blocks. There are two decoder blocks - the first one starts the upsampling process to generate the output image. It consists of a transposed convolutional layer (also called deconvolutional layer) followed by two convolutional layers and corresponding ReLU activation layers. The first decoder block combines the features from the middle block with the high level features of the second encoder block. The second decoder block continues the upsampling process by combining the features from the first decoder block with the low-level

features from the first encoder block. It has one transposed convolutional layer followed by two 3x3 convolutional layers and one 1x1 convolutional layer to produce the final output image. In a broad sense, this allows the model to skip connections, allowing information to flow directly from the encoder to the decoder blocks, helping the network preserve finer features in the output image.

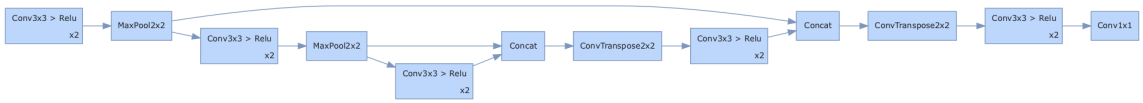


FIGURE 2.10. Our custom U-Net Model Architecture (image generated with Hiddenlayer library).

2.6.3. Model Training. We split our training and testing data using an 80/20 ratio, so that 163 months of SPEI data and corresponding SI data were used for training and 41 months of data were used for testing. We trained six U-Net models on the six different sets of SPEI images (as defined earlier). Starting from zero months lag (0ml) to 5 months lag (5ml). These lags allow us to statistically explore any temporal trends between meteorological drought events and anthropogenic response to these events (in terms of when people within our study domain become interested in topics related to droughts). The training process involved feeding input SPEI images and their corresponding SI labels to the model and loss function, respectively. We used an adaptive moment (ADAM) optimizer with a Mean Squared Error (MSE) loss, a batch size of 1 image, and 20 training epochs.

2.6.4. Model Evaluation. We evaluated models on the test set of 41 images. The output of the U-Net models are probability maps that have the same shape as the input image (i.e., the same spatial resolution as the SPEI images). These vectors represent the probability of each pixel belonging to the target class (SI maps). We then calculated the average R-squared scores for each model.

CHAPTER 3

DroughtVision - Predicting Global Meteorological Droughts with Vision Transformers

3.1. Introduction

Drought is a complex natural hazard that can have severe and widespread impacts on the environment, economy, and society. Accurate and timely prediction of drought occurrences is crucial for effective water resource management, agricultural planning, and disaster risk reduction (Wilhite, 2000; Mishra and Singh, 2010). Droughts are often characterized by their slow onset, long duration, and wide spatial extent, making their prediction a challenging task Traditionally, drought prediction models have relied on statistical methods, such as regression analysis, time series analysis, and probability-based approaches (Mishra and Singh, 2011). These models often utilize various drought indices, such as the Standardized Precipitation Index (SPI) (McKee et al., 1993), the Palmer Drought Severity Index (PDSI) (Palmer, 1965), and the Standardized Precipitation Evapotranspiration Index (SPEI) (Vicente-Serrano et al., 2010), to quantify and predict drought conditions based on meteorological variables. However, these statistical models have limitations in capturing the complex and nonlinear relationships between the input variables and drought occurrences, especially in the context of changing climate conditions (Mishra and Singh, 2011; Hao et al., 2018).

In recent years, machine learning (ML) techniques have emerged as promising tools for drought prediction, offering the ability to learn intricate patterns and relationships from large datasets (Hao et al., 2018). Various ML algorithms, such as artificial neural networks (ANNs) (Morid et al., 2007), support vector machines (SVMs)

In this chapter, we aim to address the gap in global-scale drought prediction by proposing a novel approach based on the Vision Transformer (ViT) model (Dosovitskiy et al., 2020).

The ViT model, originally developed for computer vision tasks, has demonstrated strong performance in capturing long-range dependencies and learning hierarchical representations from complex data (Dosovitskiy et al., 2020). By adapting the ViT model to the task of drought prediction, we leverage its ability to learn spatial and temporal patterns from high-dimensional climate data. Our work builds upon recent advancements in applying transformer-based models to Earth system modeling (Rasp et al., 2021; Pathak et al., 2022) by applying them to the specific problem of global drought prediction.

3.1.1. Research Question. In this chapter, we address the following specific research questions:

- Can the Vision Transformer model predict meteorological drought occurrences on a global scale from Sea Surface Temperature, 2-m Temperature, and Total Precipitation?

3.2. Methods

In this section, we discuss the data-driven experimental approaches and methods that we used to address the research questions outlined above.

3.2.1. Data. Table 3.1 summarizes the data types used in this study, including where the data was sourced and the temporal periods that we acquired and used. We used the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 dataset (Hersbach et al., 2020) to obtain pre-calculated monthly averaged variable data for model inputs. ERA5 provides a comprehensive and consistent record of the Earth's atmosphere, land surface, and ocean waves from 1950 to the present. The dataset is generated using the ECMWF Integrated Forecast System (IFS) and assimilates a vast array of observations from satellites, weather stations, and other sources. All data are provided on a regular grid with a spatial resolution of approximately 31 km and a temporal resolution of 1 hour.

3.2.1.1. *Sea Surface Temperature (SST).* SST is the temperature of the ocean surface. In ERA5, SST is derived from a combination of satellite observations Operational Sea Surface

TABLE 3.1. Overview of data sources and availability

Data	**Abbreviation**	**Source**	**Years Available**	**Years Used**
Sea Surface Temperature	SST	ERA5 dataset (ECMWF)	1950-Present	1970-2020
2-meter Temperature	T2M	ERA5 dataset (ECMWF)	1950-Present	1970-2020
Total Precipitation	TP	ERA5 dataset (ECMWF)	1950-Present	1970-2020
Standardized Precipitation and Evapotranspiration Index	SPEI	spei.csic.es	1900-2020	1970-2020

Temperature and Sea Ice Analysis) and in-situ measurements from buoys and ships. The analysis uses the NEMOVAR data assimilation system, which combines the observations with a prior estimate of the ocean state from the NEMO ocean model (Madec et al., 2017).

3.2.1.2. *2-meter Temperature (T2M).* T2M in ERA5 is calculated using the IFS (Integrated Forecast System) model's land surface scheme, which is based on the HTESSEL (Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land) model (Balsamo et al., 2009). The land surface scheme simulates the energy and water balance at the Earth's surface, taking into account factors such as solar radiation, surface albedo, soil moisture, and vegetation cover. The IFS model assimilates a variety of observations, including surface synoptic observations (SYNOP), ship and buoy measurements, and satellite-derived data, to constrain the model's surface temperature estimates.

3.2.1.3. *Total Precipitation (TP).* TP in ERA5 is derived from a combination of satellite observations, rain gauge measurements, and the IFS model's atmospheric moisture budget. The primary satellite data used for precipitation comes from the GPM (Global Precipitation Measurement) mission, which provides estimates of rainfall and snowfall based on microwave and radar measurements (Huffman et al., 2015). Rain gauge measurements from the GPCC (Global Precipitation Climatology Centre) and other national and regional networks are used

to calibrate and validate the satellite data. The IFS model's atmospheric moisture budget is used to provide a first guess of precipitation, which is then adjusted based on the available observations using the 4D-Var data assimilation system.

3.2.1.4. *Standardized Precipitation and Evapotranspiration Index (SPEI).* We use the Standardized Precipitation and Evapotranspiration Index (SPEI) (Beguería, 2022) as our meteorological drought dataset. SPEI serves as a quantitative measure of droughts by providing a standardized index to assess moisture deficit over time and space. It is calculated by taking the difference between total precipitation and total potential evapotranspiration (PET) over a given period of time (e.g., monthly). SPEI is is expressed in units of standard deviations calculated over local (per pixel) climatologies, allowing us to assess whether it is drier or wetter than expected conditions. Calculating SPEI involves the following steps:

(1) Calculate the difference between precipitation (P) and reference evapotranspiration ($ET0$) for each month or time step:

$$D_i = P_i - ET0_i \tag{3.1}$$

where D_i is the difference between precipitation and reference evapotranspiration for the i-th month or time step, P_i is the precipitation for the i-th month or time step, and $ET0_i$ is the reference evapotranspiration for the i-th month or time step.

(2) Calculate the climatic water balance for each month or time step:

$$WB_i = \sum_{j=1}^{n} D_{i-j+1} \tag{3.2}$$

where WB_i is the climatic water balance for the i-th month or time step, and n is the time scale (e.g., 3, 6, or 12 months).

(3) Fit a probability distribution, such as the three-parameter log-logistic distribution with the Maximum Likelihood Estimation (MLE) method, to the climatic water

balance values:

$$F(WB) = \frac{1}{1 + \left(\frac{WB-\alpha}{\beta}\right)^{-\gamma}} \tag{3.3}$$

where $F(WB)$ is the cumulative probability distribution of the climatic water balance, and α, β, and γ are the distribution parameters that need to be estimated.

(4) Calculate SPEI by transforming the fitted probability distribution to a standard normal distribution:

$$SPEI = \Phi^{-1}(F(WB)) \tag{3.4}$$

where SPEI is the Standardized Precipitation Evapotranspiration Index, Φ^{-1} is the inverse standard normal cumulative distribution function, and $F(WB)$ is the cumulative probability distribution of the climatic water balance.

SPEI can be calculated for a variety of time periods, ranging from 1 month to 48 months. We use monthly data in our analysis consistent with our ERA5 data. Positive SPEI values indicate wetter conditions, while negative values indicate drier conditions. SPEI value $<$ 0 is generally accepted as dry to drought conditions (Vicente-Serrano et al., 2010; Potop et al., 2014; Stagge et al., 2015; Spinoni et al., 2015; Beguería et al., 2014) and we used this threshold to count drought occurrences.

3.2.2. Data Preprocessing and Data Loading. We preprocessed the input and target datasets to be compatible with our state-of-the-art machine learning modeling pipeline. We adjusted the latitude and longitude of the input dataset to ensure consistency and then resampled them to a 1-degree spatial resolution using bilinear interpolation. This step involved defining new latitude and longitude ranges with 1-degree intervals and interpolating the data to match the new grid. We applied a similar approach for our target (SPEI) dataset as well.

We designed a custom dataloader for our pipeline which takes parameters such as the sequence length (number of months of input data), the predict length (number of months to predict), and a boolean flag indicating whether the dataset is used for training or testing. We used 12 months as the input sequence length and 3 months as the predict length (predicting on 12th(Lead Time 0), 13th(Lead Time 1), and 14th(Lead Time 2) month).

We applied Min-Max scaling to the input data to normalize them so that all the input variables per grid point and time step have a value between 0 and 1. We then created a mask to identify the NaN values (land values) in the SST data and replaced these NaN values with corresponding T2M values. We also created a similar mask for NaN SPEI values (Ocean values) and replaced them with a placeholder value. We then converted the SPEI into a binary dataset by setting all SPEI values < 0 to 1 (indicating "dry to drought" conditions) and setting SPEI values > 0 to 0 (indicating "no droughts"). Figure shows examples of SST, T2M, TP, and SPEI images resulting from the preprocessing and used for training and testing our Vision Transformer Model.

3.2.3. The Transformer Network. The Transformer network (Vaswani et al., 2017) has revolutionized the field of natural language processing and has been successfully applied to various tasks, including computer vision. The core idea behind the Transformer architecture is to capture long-range dependencies and contextual information through self-attention mechanisms, eliminating the need for recurrent or convolutional layers.

The Transformer consists of an encoder and a decoder, but in the Vision Transformer (ViT) model, we only use the encoder part. The encoder is composed of multiple identical layers, each containing two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network.

The self-attention mechanism allows each element in the input sequence to attend to and compute a weighted sum of the other elements, based on their relevance or importance. The attention scores are calculated by computing the dot product between the query, key, and value vectors, which are derived from the input embeddings. The dot products are then

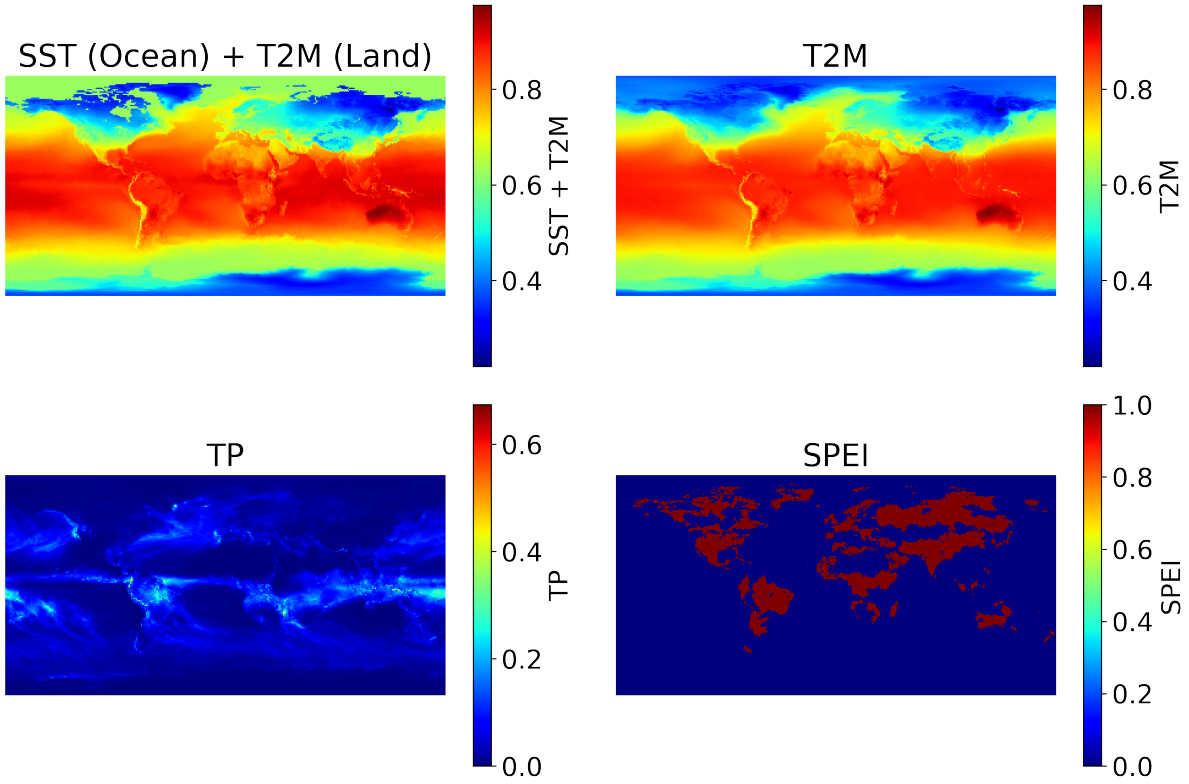


FIGURE 3.1. Examples of the input variables and target variable used for ViT model training and testing. The top-left image shows a SST+T2M image, top right shows a T2M image, bottom left shows a TP image. The bottom right shows a SPEI binary image indicating the presence or absence of drought conditions. The data is from a sample batch of the test dataset, representing a single time step. The color scale represents the normalized values of each variable, with higher values in red and lower values in blue.

scaled and passed through a softmax function to obtain the attention weights. These weights are used to compute a weighted sum of the value vectors, resulting in the attended features for each element. The attention equation can be represented as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.5}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. The softmax function is applied to the scaled dot product of

Q and K^T to obtain the attention weights, which are then used to compute a weighted sum of the value vectors V.

To capture different aspects of the input, the self-attention mechanism is applied multiple times in parallel, known as multi-head attention. Each attention head computes its own set of query, key, and value vectors and performs the attention computation independently. The outputs from all the heads are then concatenated and linearly transformed to obtain the final attended features.

The position-wise feed-forward network consists of two linear transformations with a ReLU activation in between. It is applied independently to each position in the sequence, allowing the model to learn non-linear interactions between the attended features.

The Transformer architecture also incorporates positional embeddings to capture the positional information of the input elements. These embeddings are added to the input embeddings to provide the model with a sense of order and relative position.

3.2.4. DroughtVision - A Vision Transformer Model. In our drought prediction task, we utilized the Vision Transformer (ViT) model, which adapts the Transformer architecture to process visual data.

Patch Embedding: We concatenated the SST, TP, and T2M data along the channel dimension to form a multi-channel input tensor. This tensor was then split into patches of a fixed size (16x16) to serve as the input sequence to the Transformer. Each patch was linearly projected into a hidden dimension using a convolutional layer followed by a flattening operation. This step converted the patches into a sequence of embeddings.

Class Token: We introduced a learnable class token embedding, which was prepended to the sequence of patch embeddings. The class token served as a global representation of the entire input and was used for the final prediction.

Position Embeddings: Learnable position embeddings were added to the patch embeddings to capture positional information. These embeddings helped the model understand the spatial relationships between the patches.

Transformer Encoder: Transformer encoder layers (64 layers) were stacked to learn the contextual representations of the input. Each layer consisted of multi-head self-attention (128 heads) and a feed-forward network.

Decoder: The class token representation from the final Transformer encoder layer was passed through a decoder, which consisted of a linear layer followed by a sigmoid activation function. The sigmoid function was applied element-wise to the output of the linear layer, squashing the values to the range [0, 1]. This produced a probability-like output, indicating the likelihood of drought conditions for each spatial location and time step.

Output Reshaping: The output from the decoder was reshaped to match and align with the spatial dimensions (height and width) of the target SPEI data and the number of predicted time steps. Figure 3.2 is a high-level representation of our custom ViT model training pipeline and its various components.

3.2.4.1. *Model Training.* We trained our ViT model using a variety of hyperparameter settings. For every training routine, we first initialized the model with specified hyperparameters (e.g. hidden dimension, size of patches, number of transformer layers, number of attention heads, etc.). Since our task was binary classification (drought vs. no drought), the binary cross-entropy loss (BCELoss) was chosen as the loss function. BCELoss measures the dissimilarity between the predicted probabilities and the true binary labels, penalizing the model for incorrect predictions.

We applied the Adam optimizer to update the model's parameters during training. Adam is an adaptive learning rate optimization algorithm that combines the advantages of AdaGrad and RMSProp, adapting the learning rates for each parameter based on their historical gradients. The learning rate and other hyperparameters of the optimizer were set based on default values.

In the training loop, the model iterated over the training data for a specified number (20) of epochs. In each epoch, the model was trained on batches of input data (SST, T2M, TP) and corresponding target (SPEI) labels. For each batch, the ViT model performed a

forward pass, generating predicted probabilities for each spatial location and time step. The binary cross-entropy loss was computed between the predicted probabilities and the true (target) binary labels. The gradients of the loss with respect to the model's parameters were computed using backpropagation and the optimizer updated the model's parameters based on the computed gradients to minimize the loss. We monitored the training loss for each epoch to monitor the training progress.

3.2.4.2. *Model Evaluation.* We tested the model performance on 20% of our dataset. For each batch in the testing data, the model performed a forward pass, generating predicted probabilities for each spatial location and time step. The predicted probabilities were stored for each month (Lead Times 0, 1, 2) separately and concatenated along with true labels across all batches. We converted the predicted probabilities to binary predictions using a tunable threshold and then computed the model skills using various evaluation metrics: Accuracy, Precision, Recall, and F1 scores.

Accuracy measures the proportion of correct predictions (both drought and no drought) out of the total predictions. Precision measures the proportion of true positive predictions (correctly predicted droughts) out of the total positive predictions. Recall (Sensitivity) measures the proportion of true positive predictions out of the total actual drought occurrences, and F1 scores measures the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

3.3. Results and Analysis

In this section we address the research question laid out previously and discuss the results of our analyses.

3.3.1. Can the Vision Transformer model predict meteorological drought occurrences on a global scale from Sea Surface Temperature, 2-m Temperature, and Total Precipitation? To investigate if the relationship between SST, T2M, and TP, and drought occurrences was learned by the ViT model, we look at the overall model skill

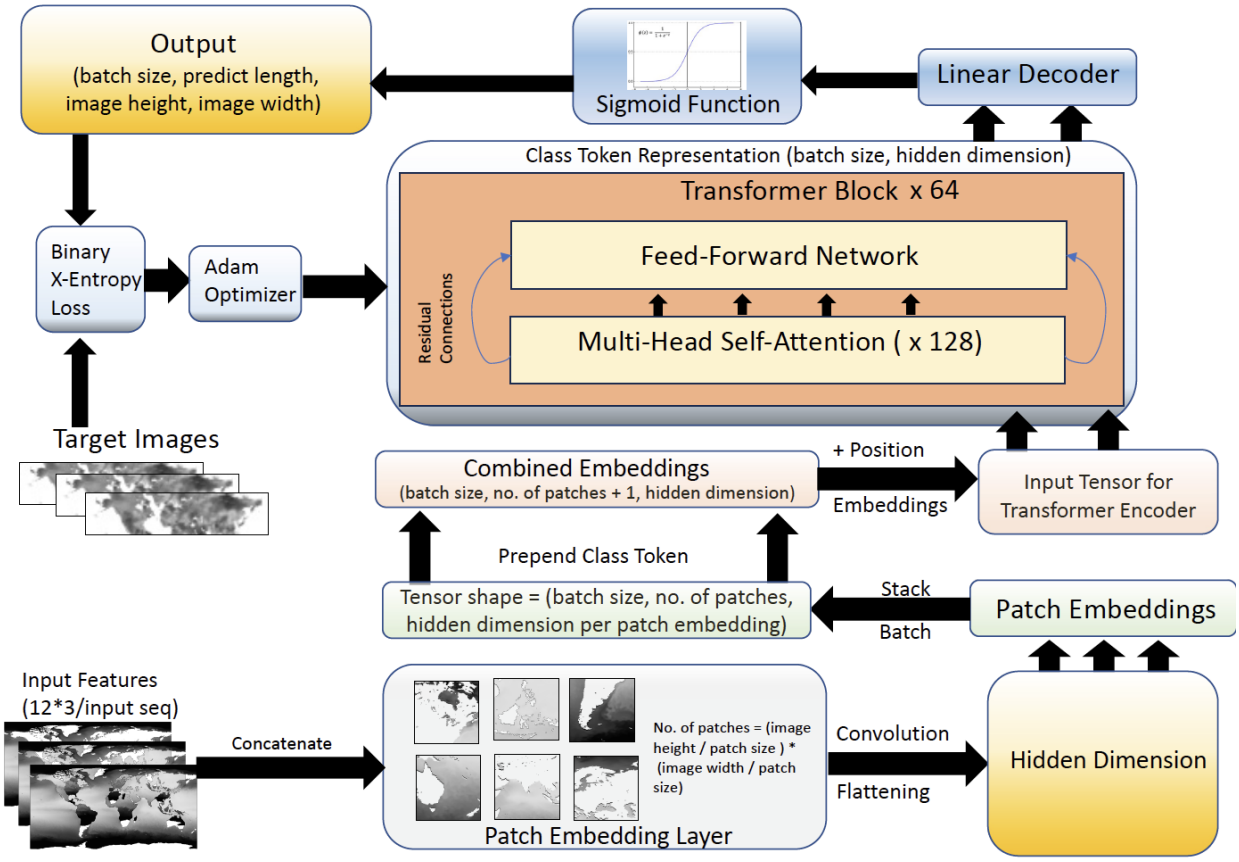


FIGURE 3.2. High-level representation of the ViT model training flow. The input images are broken down into equal-sized patches which are then embedded into tensors. Class token are prepended, position embeddings are learned, and the combined embeddings are fed to the transformer encoder. During model training, class token representations are learned and afterwards decoded. The decoder outputs are passed through a sigmoid function in the final layer and the outputs are generated. Binary cross-entropy loss function is used per epoch to inform the optimizer which also adjusts the learning rate of the network.

scores during our evaluation period (01/01/2005-12/31/2020) for the individual 3 months per input sequence.

3.2 shows the accuracy, precision, recall, and F1 scores for the three months per input sequence.

The accuracy scores for all three months were consistently high, ranging from 0.9456 to 0.9475. This indicates that the model correctly predicts the drought status (drought or no

drought) for a large majority of the locations. The high accuracy scores demonstrate the model's overall effectiveness in distinguishing between drought and non-drought conditions. The precision scores, which measure the proportion of true positive predictions among all positive predictions, were also quite significant. The precision scores range from 0.8747 to 0.8781, suggesting that when the model predicts a drought, it is correct in a significant proportion of cases. These high precision scores indicate that the model has a low false positive rate and is mostly reliable in identifying actual drought occurrences. The precision scores decrease very slightly steadily from 0 months to 2 months lead times, as shown in Figure 3.3.

The recall scores, which measure the proportion of actual drought occurrences that are correctly predicted, are relatively lower compared to the precision scores. The recall scores range from 0.6285 to 0.6465, indicating that the model misses some actual drought occurrences. Recall scores slightly improve for lead time of 1 month compared to lead time of 0 month. This might indicate the model is capturing the global dependencies within input images and input and target images for correctly predicting actual drought occurrences with one month lead time.

The F1 scores, which provide a balanced measure of precision and recall, ranged from 0.7323 to 0.7447. These scores indicate significant overall performance of the ViT model, considering both its ability to correctly identify droughts and its effectiveness in minimizing false positives and false negatives.

Month	Accuracy	Precision	Recall	F1 Score
0	0.9475	0.8781	0.6465	0.7447
1	0.9456	0.8770	0.6285	0.7323
2	0.9471	0.8747	0.6462	0.7433

TABLE 3.2. Evaluation metrics for each month.

We further investigate whether our ViT model captures meaningful relationships beyond the inherent temporal autocorrelation present in the SPEI data. We first calculate the temporal extent over which the SPEI values at a particular pixel remain significantly correlated with

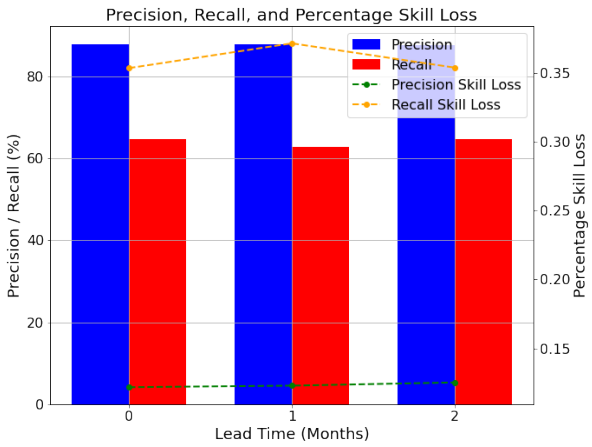


FIGURE 3.3. Precision and recall scores and skill loss percentages for the predicted months. Recall scores slightly improve for lead time of 1 month compared to lead time of 0 month.

each other (correlation length) and generate an exponential decay correlogram. Figure 3.4 shows that the autocorrelation of SPEI values decreases as the lag (in months) increases. The threshold line represents a specific autocorrelation value of exp(-1). The lag at which the autocorrelation crosses this threshold is considered the correlation length. We can observe that the majority of the pixels have a correlation length of around 2-3 months, as the autocorrelation values fall below the threshold within that time range. This suggests that the SPEI values have a relatively short-term memory or persistence, meaning that the SPEI values are more strongly correlated with their recent past values, and the influence of past values diminishes quickly over time. Our ViT model is designed to for near-future predictions (2-3 months), allowing it to contextually learn short-term dependencies between the input variables and SPEI data while also being aware if the long-term dependencies.

Figure 3.5 presents a visual comparison of the actual and predicted drought maps and counts for each of the three months. It is evident that the model captures the overall spatial patterns of drought reasonably well. The predicted drought maps mostly resemble the actual drought maps, indicating that the model has learned to identify the key features and patterns associated with global drought occurrence. There are some significant discrepancies

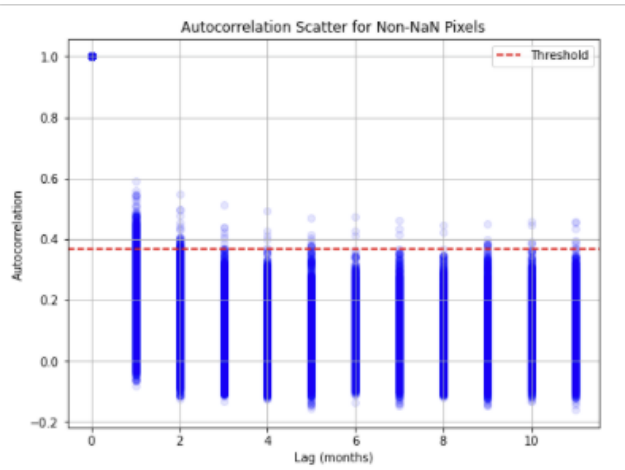


FIGURE 3.4. Autocorrelation decay of SPEI values (blue dots) with increasing lag in months. The rate of decay in the autocorrelation provides information about the persistence of SPEI values over time. The threshold line (red dashed line) represents a specific autocorrelation value (exp(-1) or 0.37). The lag at which the autocorrelation crosses this threshold is considered the correlation length. Majority of the pixels have a correlation length of around 2-3 months.

between the actual and predicted maps as well, particularly in regions where droughts are less prevalent or have a more sporadic nature. We also observe that the ViT model is slightly under-predicting the drought occurrences and slightly over-predicting the no-drought counts. We also looked at the spatial representation of the true positives and false positives (Figure 3.6). There are some distinct regions, such as over the Sahara, Mongolia, Siberia, parts of the Middle East, Eastern Europe, Greenland and Canada where the model is incorrectly predicting drought occurrences. It makes sense physically that most of these regions are prone to complex and persistently dry to drought conditions. Further investigation is required to understand the underlying reasons behind these false-positive results.

Our results also indicate that there are inherent patterns and dependencies between SST, T2M, TP and meteorological drought occurrences. Multiple previous studies have linked temperature and precipitation with drought occurrences, frequencies, and intensity (Vicente-Serrano et al., 2010; Dai, 2011; Trenberth et al., 2014; Mukherjee et al., 2018; Schubert et al., 2016). However, we acknowledge that while these variables exhibit a strong relationship with drought occurrences, there may be other factors influencing global drought dynamics.

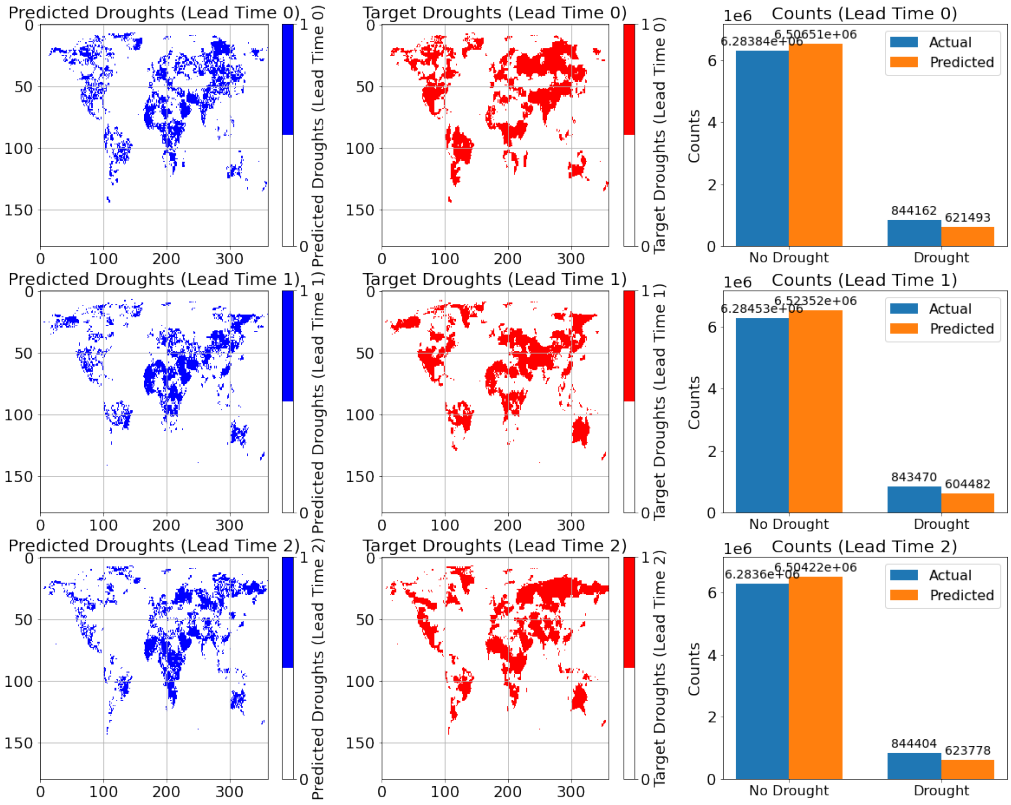


FIGURE 3.5. Comparison of the actual and predicted drought maps for the three individual months per prediction. The maps on the left show the actual drought conditions, while the maps on the right display the model's predictions. The bar plots to the right indicate the counts of predicted and true droughts or no droughts. Based on the drought counts, our ViT model underpredicts droughts by about 26% and overpredicts drought conditions by about 3.5%.

3.4. Conclusions & Discussion

We developed a Vision Transformer (ViT) based model to predict meteorological drought occurrences on a global scale using sea surface temperature, 2-meter air temperature, and total precipitation (TP) as input variables. We trained the model using a sliding window approach on data from January 1970 to December 2004 and evaluated its performance on

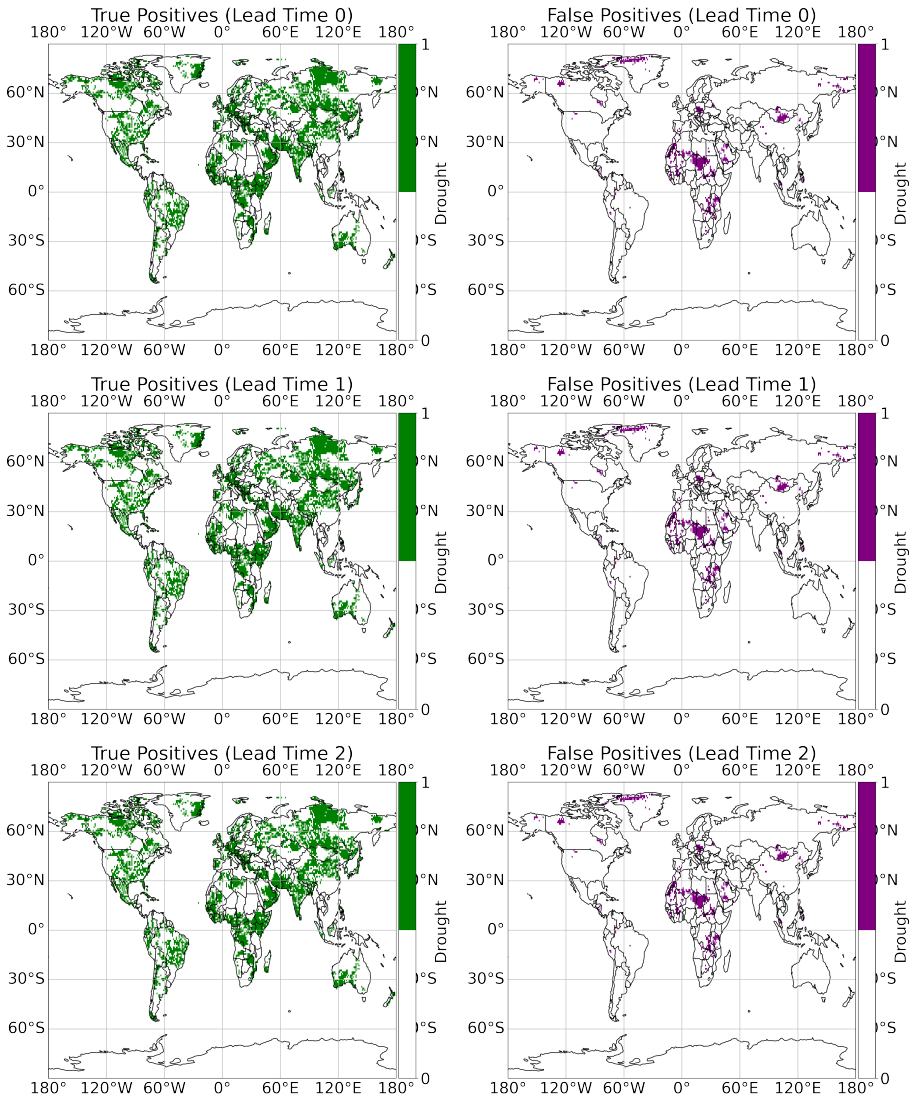


FIGURE 3.6. The true positive and false positive maps for each of the three predicted months. The true positive maps (left column) highlight the regions where the model correctly predicted drought occurrences, while the false positive maps (right column) indicate the areas where the model incorrectly predicted droughts when there were none.

data from January 2005 to December 2020. The model's skill was assessed using a diverse set of metrics. The primary findings of this study are (see the research question outlined in Section 3.1.1):

- The ViT model was able to distinguish between drought and non-drought conditions.
- The model was mostly reliable in identifying actual drought occurrences.

- The model missed some actual drought occurrences, particularly in regions with less frequent or more complex drought patterns possibly driven by other factors
- The model captured global associations between SST, T2M, TP and drought occurrences.

The findings of this study indicate substantial potential of the ViT model in predicting meteorological drought occurrences on a global scale.

Our study has certain limitations in the sense that the ViT model misses some actual drought occurrences, indicating the need for understanding the underlying reasons and further refinement and tuning of the model. Our method also does not account for other types of droughts, such as agricultural or hydrological droughts. We also acknowledge the need for introducing a more diverse set of input variables (such as soil moisture and vegetation indices) and using other drought indices and datasets to increase the generalizability of this model. Future direction of this study may also involve (i) performing a robust feature importance analysis to understand the importance of different variables for predictions, (ii) exploring the use of long short-term memory networks for better capturing temporal dependencies, (iii) experimenting with various timescales, and (iv) integrating predictions with socioeconomic and environmental data for a more holistic assessment.

Our ViT model leveraged the Attention mechanism of Transformer networks to gain a broad context of global climatic variables and learns their complex and nonlinear relationships with drought occurrences. The non-stationarity of global climate (made increasingly complex by anthropogenic-driven climate change) poses challenges to traditional drought forecasting approaches, emphasizing the need for increasingly broader contextual understanding capability provided by our model. By continuing to leverage deep learning and satellite-derived climate variables, we can develop more reliable and effective tools for drought monitoring and early warning systems. The findings of this study contribute to the ongoing efforts in understanding and mitigating the impacts of droughts under a changing climate, and pave the way for further research and development in this critical area.

3.5. Acknowledgements(Project Specific)

This study was supported by the Google Cloud Research Credits program with the award GCP19980904.

Bibliography

J. Harshbarger, D. Evans, Educational progress in water resources present and future, JAWRA Journal of the American Water Resources Association 3 (1967) 29–44.

R. A. Freeze, Water resources research and interdisciplinary hydrology, 1990.

National Research Council, Opportunities in the hydrologic sciences, National Academies Press, 1991.

National Research Council, Challenges and opportunities in the hydrologic sciences, National Academies Press, 2012.

A. Montanari, G. Young, H. Savenije, D. Hughes, T. Wagener, L. Ren, D. Koutsoyiannis, C. Cudennec, E. Toth, S. Grimaldi, et al., "panta rhei—everything flows": change in hydrology and society—the iahs scientific decade 2013–2022, Hydrological Sciences Journal 58 (2013) 1256–1275.

A. Montanari, J. Bahr, G. Blöschl, X. Cai, D. S. Mackay, A. M. Michalak, H. Rajaram, G. Sander, Fifty years of water resources research: Legacy and perspectives for the science of hydrology, Water Resources Research 51 (2015) 6797–6803.

B. L. Ruddell, T. Wagener, Grand challenges for hydrology education in the 21st century, Journal of Hydrologic Engineering 20 (2015) A4014001.

R. M. Vogel, U. Lall, X. Cai, B. Rajagopalan, P. K. Weiskel, R. P. Hooper, N. C. Matalas, Hydrology: The interdisciplinary science of water, Water Resources Research 51 (2015) 4409–4430.

P. Milly, J. Betancourt, M. Falkenmark, R. M. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, R. J. Stouffer, Stationarity is dead: Whither water management?, Earth 4 (2008) 20.

M. Bayazit, Nonstationarity of hydrological records and recent trends in trend analysis: a state-of-the-art review, Environmental Processes 2 (2015) 527–542.

G. E. Galloway, If stationarity is dead, what do we do now? 1, JAWRA Journal of the American Water Resources Association 47 (2011) 563–570.

H. V. Gupta, C. Perrin, G. Blöschl, A. Montanari, R. Kumar, M. Clark, V. Andréassian, Large-sample hydrology: a need to balance depth with breadth, Hydrology and Earth System Sciences 18 (2014) 463–477.

G. S. Nearing, F. Kratzert, A. K. Sampson, C. S. Pelissier, D. Klotz, J. M. Frame, C. Prieto, H. V. Gupta, What role does hydrological science play in the age of machine learning?, Water Resources Research 57 (2021) e2020WR028091.

K. L. McCurley, J. W. Jawitz, Hyphenated hydrology: Interdisciplinary evolution of water resource science, Water Resources Research 53 (2017) 2972–2982.

E. Cambria, B. White, Jumping NLP curves: A review of natural language processing research, IEEE Computational intelligence magazine 9 (2014) 48–57.

F. Sebastiani, Machine learning in automated text categorization, ACM computing surveys (CSUR) 34 (2002) 1–47.

H. Jiang, M. Qiang, P. Lin, A topic modeling based bibliometric exploration of hydropower research, Renewable and Sustainable Energy Reviews 57 (2016) 226–237.

D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (2012) 77–84.

C. Wang, D. M. Blei, Collaborative topic modeling for recommending scientific articles, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2011, pp. 448–456.

C.-K. Yau, A. Porter, N. Newman, A. Suominen, Clustering scientific documents with topic modeling, Scientometrics 100 (2014) 767–786.

J. Jardine, S. Teufel, Topical PageRank: A model of scientific expertise for bibliographic search, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 501–510.

P. Pham, P. Do, C. D. C. Ta, W-PathSim: Novel approach of weighted similarity measure in content-based heterogeneous information networks by applying LDA topic modeling, in: Asian conference on intelligent information and database systems, Springer, 2018, pp. 539–549.

L. Shu, B. Long, W. Meng, A latent topic model for complete entity resolution, in: Proceedings of the 2009 IEEE International Conference on Data Engineering, IEEE Computer Society, 2009, pp. 880–891.

J. Tang, R. Jin, J. Zhang, A topic modeling approach and its integration into the random walk framework for academic search, in: 2008 Eighth IEEE International Conference on Data Mining, IEEE, 2008, pp. 1055–1060.

M. Paul, R. Girju, Topic modeling of research fields: An interdisciplinary perspective, in: Proceedings of the International Conference RANLP-2009, 2009, pp. 337–342.

D. M. Blei, K. Franks, M. I. Jordan, I. S. Mian, Statistical modeling of biomedical corpora: mining the Caenorhabditis Genetic Center Bibliography for genes related to life span, Bmc Bioinformatics 7 (2006) 250.

H. S. Choi, W. S. Lee, S. Y. Sohn, Analyzing research trends in personal information privacy using topic modeling, Computers & Security 67 (2017) 244–253. URL: http://www.sciencedirect.com/science/article/pii/S0167404817300603. doi:https://doi.org/10.1016/j.cose.2017.03.007.

B. R. Upreti, A. Asatiani, P. Malo, To Reach the Clouds: Application of Topic Models to the Meta-Review on Cloud Computing Literature, in: 2016 49th Hawaii International Conference on System Sciences (HICSS), 2016, pp. 3979–3988. doi:10.1109/HICSS.2016.493.

C.-y. Li, F.-f. Liu, A Topic Modeling Method for Social Science Literature Based on LDA, Computer Technology and Development 2 (2018) 182–187.

S. Sun, C. Luo, J. Chen, A review of natural language processing techniques for opinion mining systems, Information Fusion 36 (2017) 10–25. URL: http:

//www.sciencedirect.com/science/article/pii/S1566253516301117. doi:https://doi.org/10.1016/j.inffus.2016.10.004.

J. Jussila, N. Mustafee, H. Aramo-Immonen, K. Menon, A. Hajikhani, N. Helander, A Bibliometric Study on Authorship Trends and Research Themes in Knowledge Management Literature, International Forum on Knowledge Asset Dynamics, 2017.

U. C. Priva, J. L. Austerweil, Analyzing the history of cognition using topic models, Cognition 135 (2015) 4–9.

D. Hall, D. Jurafsky, C. D. Manning, Studying the history of ideas using topic models, in: Proceedings of the 2008 conference on empirical methods in natural language processing, 2008, pp. 363–371.

J. Mingers, L. Leydesdorff, A review of theory and practice in scientometrics, European journal of operational research 246 (2015) 1–19.

R. Řehřek, P. Sojka, Gensim—statistical semantics in python, Retrieved from genism. org (2011).

D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 113–120.

T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National academy of Sciences 101 (2004) 5228–5235.

C. J. Gatti, J. D. Brooks, S. G. Nurre, A historical analysis of the field of or/ms using topic models, arXiv preprint arXiv:1510.05154 (2015).

X. Wang, 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, 2006, pp. 424–433.

C. E. Shannon, A mathematical theory of communication, Bell system technical journal 27 (1948) 379–423.

J. Harte, E. A. Newman, Maximum information entropy: a foundation for ecological theory, Trends in ecology & evolution 29 (2014) 384–389.

W. B. Sherwin, N. Prat i Fornells, The introduction of entropy and information methods to ecology by ramon margalef, Entropy 21 (2019) 794.

A. Mehran, A. AghaKouchak, N. Nakhjiri, M. J. Stewardson, M. C. Peel, T. J. Phillips, Y. Wada, J. K. Ravalico, Compounding impacts of human-induced water stress and climate change on water availability, Scientific Reports 7 (2017) 1–9.

H. Tabari, Climate change impact on flood and extreme precipitation increases with water availability, Scientific reports 10 (2020) 1–10.

R. S. Padrón, L. Gudmundsson, B. Decharme, A. Ducharne, D. M. Lawrence, J. Mao, D. Peano, G. Krinner, H. Kim, S. I. Seneviratne, Observed changes in dry-season water availability attributed to human-induced climate change, Nature Geoscience 13 (2020) 477–481.

L. Gudmundsson, J. Boulange, H. X. Do, S. N. Gosling, M. G. Grillakis, A. G. Koutroulis, M. Leonard, J. Liu, H. Müller Schmied, L. Papadimitriou, et al., Globally observed trends in mean and extreme river flow attributed to climate change, Science 371 (2021) 1159–1162.

M. P. Clark, R. B. Hanson, The citation impact of hydrology journals, Water Resources Research 53 (2017) 4533–4541.

F. Zare, S. Elsawah, T. Iwanaga, A. J. Jakeman, S. A. Pierce, Integrated water assessment and modelling: A bibliometric analysis of trends in the water resource sector, Journal of Hydrology 552 (2017) 765–778.

H. Rajaram, J. M. Bahr, G. Blöschl, X. Cai, D. Scott Mackay, A. M. Michalak, A. Montanari, X. Sanchez-Villa, G. Sander, A reflection on the first 50 years of water resources research, Water Resources Research 51 (2015) 7829–7837.

D. Koutsoyiannis, Z. W. Kundzewicz, Quantifying the impact of hydrological studies, 2007.

H. Ledford, How to solve the world's biggest problems, Nature News 525 (2015) 308.

L. Bromham, R. Dinnage, X. Hua, Interdisciplinary research has consistently lower funding success, Nature 534 (2016) 684–687.

R. Feldman, J. Sanger, et al., The text mining handbook: advanced approaches in analyzing unstructured data, Cambridge university press, 2007.

T. K. Landauer, P. W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse processes 25 (1998) 259–284.

D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

Y. Lu, Q. Mei, C. Zhai, Investigating task performance of probabilistic topic models: an empirical study of plsa and lda, Information Retrieval 14 (2011) 178–203.

M. J. Paul, M. Dredze, Discovering health topics in social media using topic models, PloS one 9 (2014) e103408.

S. F. Chen, D. Beeferman, R. Rosenfeld, Evaluation metrics for language models (1998).

W. Zhao, J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, W. Zou, A heuristic approach to determine an appropriate number of topics in topic modeling, in: BMC bioinformatics, volume 16, Springer, 2015, p. S8.

M. Röder, A. Both, A. Hinneburg, Exploring the Space of Topic Coherence Measures, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15, ACM Press, New York, New York, USA, 2015, pp. 399–408. URL: http://dl.acm.org/citation.cfm?doid=2684822.2685324. doi:10.1145/2684822.2685324.

A. B. Smith, J. L. Matthews, Quantifying uncertainty and variable sensitivity within the us billion-dollar weather and climate disaster cost estimates, Natural Hazards 77 (2015) 1829–1851.

H. E. Campbell, R. M. Johnson, E. H. Larson, Prices, devices, people, or rules: the relative effectiveness of policy instruments in water conservation 1, Review of policy research 21 (2004) 637–662.

J. M. Clarke, R. R. Brown, Understanding the factors that influence domestic water consumption within melbourne, Australasian Journal of Water Resources 10 (2006) 261–268.

D. C. Adams, D. Allen, T. Borisova, D. E. Boellstorff, M. D. Smolen, R. L. Mahler, The influence of water attitudes, perceptions, and learning preferences on water-conserving actions, Natural Sciences Education 42 (2013) 114–122.

D. L. Woudenberg, D. A. Wilhite, M. J. Hayes, Perception of drought hazard and its sociological impacts in south-central nebraska, Great Plains Research (2008) 93–102.

J. Urquijo, L. De Stefano, Perception of drought and local responses by farmers: a perspective from the jucar river basin, spain, Water Resources Management 30 (2016) 577–591.

Y. Bahta, A. Jordaan, F. Muyambo, Communal farmers' perception of drought in south africa: Policy implication for drought risk reduction, International Journal of Disaster Risk Reduction 20 (2016) 39–50.

D. M. Gholson, D. E. Boellstorff, S. R. Cummings, K. L. Wagner, M. C. Dozier, A survey of public perceptions and attitudes about water availability following exceptional drought in texas, Journal of Contemporary Water Research & Education 166 (2019) 1–11.

W. Shao, J. Kam, E. Cass, Public awareness and perceptions of drought: A case study of two cities of alabama, Risk, Hazards & Crisis in Public Policy (2022).

S. McLafferty, Conducting questionnaire surveys, Key methods in geography 3 (2016) 129–142.

J. J. Vaske, Advantages and disadvantages of internet surveys: Introduction to the special issue, Human Dimensions of Wildlife 16 (2011) 149–153.

R. J. Fisher, Social desirability bias and the validity of indirect questioning, Journal of consumer research 20 (1993) 303–315.

R. M. Groves, Nonresponse rates and nonresponse bias in household surveys, Public opinion quarterly 70 (2006) 646–675.

M. Martin, et al., Computer and internet use in the united states: 2018, American Community Survey Reports. Recuperado de https://www. census. gov/library/publications/2021/acs/acs-49. html (2021).

T. Carneiro, R. V. M. Da Nóbrega, T. Nepomuceno, G.-B. Bian, V. H. C. De Albuquerque, P. P. Reboucas Filho, Performance analysis of google colaboratory as a tool for accelerating deep learning applications, IEEE Access 6 (2018) 61677–61685.

A. C. Yang, N. E. Huang, C.-K. Peng, S.-J. Tsai, Do seasons have an influence on the incidence of depression? the use of an internet search engine query data as a proxy of human affect, PloS one 5 (2010) e13728.

S. Stephens-Davidowitz, The cost of racial animus on a black candidate: Evidence using google search data, Journal of Public Economics 118 (2014) 26–40. URL: https://www.sciencedirect.com/science/article/pii/S0047272714000929. doi:https://doi.org/10.1016/j.jpubeco.2014.04.010.

Y. Hong, J. Lawrence, D. Williams, I. Mainous, Population-level interest and telehealth capacity of us hospitals in response to covid-19: cross-sectional analysis of google search and national hospital survey data. jmir public health surveill. 2020 apr 07; 6 (2): e18961. doi: 10.2196/18961, ????

V. S. Arora, M. McKee, D. Stuckler, Google trends: Opportunities and limitations in health and health policy research, Health Policy 123 (2019) 338–341.

J. Mellon, Where and when can we use google trends to measure issue salience?, PS: Political Science & Politics 46 (2013) 280–290.

P. Gonzales, N. Ajami, Social and structural patterns of drought-related water conservation and rebound, Water Resources Research 53 (2017) 10619–10634.

J. Kam, K. Stowers, S. Kim, Monitoring of drought awareness from google trends: a case study of the 2011–17 california drought, Weather, Climate, and Society 11 (2019) 419–429.

S. Kim, W. Shao, J. Kam, Spatiotemporal patterns of us drought awareness, Palgrave Communications 5 (2019) 1–9.

S. Beguería, sbegueria/speibase: Version 2.7, 2022. URL: https://doi.org/10.5281/zenodo.5864391. doi:10.5281/zenodo.5864391.

D. Ahlers, Assessment of the accuracy of geonames gazetteer data, in: Proceedings of the 7th workshop on geographic information retrieval, 2013, pp. 74–81.

O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

U.S. Census Bureau, Census regions and divisions of the united states, U.S. Census Bureau, 2021. URL: www2.census.gov, accessed: January 20, 2024.

C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the international AAAI conference on web and social media, volume 8, 2014, pp. 216–225.

A. AghaKouchak, L. Cheng, O. Mazdiyasni, A. Farahmand, Global warming and changes in risk of concurrent climate extremes: Insights from the 2014 california drought, Geophysical Research Letters 41 (2014) 8847–8852.

R. Barrera, V. Acevedo, M. Amador, M. Marzan, L. E. Adams, G. Paz-Bailey, El niño southern oscillation (enso) effects on local weather, arboviral diseases, and dynamics of managed and unmanaged populations of aedes aegypti (diptera: Culicidae) in puerto rico, Journal of Medical Entomology (2023) tjad053.

E. Moyo, L. G. Nhari, P. Moyo, G. Murewanhema, T. Dzinamarira, Health effects of climate change in africa: A call for an improved implementation of prevention measures, Eco-Environment & Health (2023).

A. R. Chilakala, P. Pandey, A. Durgadevi, M. Kandpal, B. S. Patil, K. Rangappa, P. C. O. Reddy, V. Ramegowda, M. Senthil-Kumar, Drought attenuates plant responses to multiple rhizospheric pathogens: A study on a dry root rot-associated disease complex in chickpea fields, Field Crops Research 298 (2023) 108965.

L. Bao, L. Yu, Y. Li, F. Yan, V. Lyne, C. Ren, Climate change impacts on agroecosystems in china: Processes, mechanisms and prospects, Chinese Geographical Science (2023) 1–18.

T. C. Leahy, Desperate times call for sensible measures: The making of the california sustainable groundwater management act, Golden Gate U. Envtl. LJ 9 (2015) 5.

T. Carbonell, Epa's proposed clean power plan: Protecting climate and public health by reducing carbon pollution from the us power sector, Yale L. & Pol'y Rev. 33 (2014) 403.

L. E. Hestres, Preaching to the choir: Internet-mediated advocacy, issue public mobilization, and climate change, New Media & Society 16 (2014) 323–339.

K. J. Mach, M. D. Mastrandrea, T. E. Bilir, C. B. Field, Understanding and responding to danger from climate change: the role of key risks in the ipcc ar5, Climatic Change 136 (2016) 427–444.

A. M. McCright, R. E. Dunlap, The politicization of climate change and polarization in the american public's views of global warming, 2001–2010, The Sociological Quarterly 52 (2011) 155–194.

S. Chinn, P. S. Hart, S. Soroka, Politicization and polarization in climate change news content, 1985-2017, Science Communication 42 (2020) 112–129.

T. Bolsen, J. N. Druckman, Do partisanship and politicization undermine the impact of a scientific consensus message about climate change?, Group Processes & Intergroup Relations 21 (2018) 389–402.

A. Gage, A. Milman, Groundwater plans in the united states: Regulatory frameworks and management goals, Groundwater 59 (2021) 175–189.

K. M. d. Treen, H. T. Williams, S. J. O'Neill, Online misinformation about climate change, Wiley Interdisciplinary Reviews: Climate Change 11 (2020) e665.

I. Freiling, J. Matthes, Correcting climate change misinformation on social media: Reciprocal relationships between correcting others, anger, and environmental activism, Computers in Human Behavior 145 (2023) 107769.

R. J. Brulle, From environmental campaigns to advancing the public dialog: Environmental communication for civic engagement, Environmental Communication 4 (2010) 82–98.

S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: A survey, IEEE transactions on pattern analysis and machine intelligence 44 (2021) 3523–3542.

D. Qin, J. Yu, G. Zou, R. Yong, Q. Zhao, B. Zhang, A novel combined prediction scheme based on cnn and lstm for urban pm 2.5 concentration, IEEE Access 7 (2019) 20050–20059.

H. Hasan, H. Z. Shafri, M. Habshi, A comparison between support vector machine (svm) and convolutional neural network (cnn) models for hyperspectral image classification, in: IOP Conference Series: Earth and Environmental Science, volume 357, IOP Publishing, 2019, p. 012035.

D. Zeng, S. Zhang, F. Chen, Y. Wang, Multi-scale cnn based garbage detection of airborne hyperspectral data, IEEE Access 7 (2019) 104514–104527.

R. Barzegar, M. T. Aalami, J. Adamowski, Coupling a hybrid cnn-lstm deep learning model with a boundary corrected maximal overlap discrete wavelet transform for multiscale lake water level forecasting, Journal of Hydrology 598 (2021) 126196.

M. Panahi, N. Sadhasivam, H. R. Pourghasemi, F. Rezaie, S. Lee, Spatial prediction of groundwater potential mapping based on convolutional neural network (cnn) and support vector regression (svr), Journal of Hydrology 588 (2020) 125033.

X. Guo-yan, Z. Jin, S. Cun-you, H. Wen-bin, L. Fan, Combined hydrological time series forecasting model based on cnn and mc, Computer and Modernization (2019) 23.

D. Hussain, T. Hussain, A. A. Khan, S. A. A. Naqvi, A. Jamil, A deep learning approach for hydrological time-series prediction: A case study of gilgit river basin, Earth Science Informatics 13 (2020) 915–927.

T. Tu, K. Ishida, A. Ercan, M. Kiyama, M. Amagasaki, T. Zhao, Hybrid precipitation downscaling over coastal watersheds in japan using wrf and cnn, Journal of Hydrology: Regional Studies 37 (2021) 100921.

F. R. Adaryani, S. J. Mousavi, F. Jafari, Short-term rainfall forecasting using machine learning-based approaches of pso-svr, lstm and cnn, Journal of Hydrology 614 (2022)

128463.

Y. Wang, Z. Fang, H. Hong, L. Peng, Flood susceptibility mapping using convolutional neural network frameworks, Journal of Hydrology 582 (2020) 124482.

C. M. Song, Data construction methodology for convolution neural network based daily runoff prediction and assessment of its applicability, Journal of Hydrology 605 (2022) 127324.

Y. Li, W. Wang, G. Wang, Q. Tan, Actual evapotranspiration estimation over the tuojiang river basin based on a hybrid cnn-rf model, Journal of Hydrology 610 (2022) 127788.

S. Kabir, S. Patidar, X. Xia, Q. Liang, J. Neal, G. Pender, A deep convolutional neural network model for rapid prediction of fluvial flood inundation, Journal of Hydrology 590 (2020) 125481.

D. A. Wilhite, Drought as a natural hazard: concepts and definitions, Drought: A Global Assessment 1 (2000) 3–18.

A. K. Mishra, V. P. Singh, A review of drought concepts, Journal of Hydrology 391 (2010) 202–216.

A. K. Mishra, V. P. Singh, Drought modeling–a review, Journal of Hydrology 403 (2011) 157–175.

T. B. McKee, N. J. Doesken, J. Kleist, et al., The relationship of drought frequency and duration to time scales, in: Proceedings of the 8th Conference on Applied Climatology, volume 17, Boston, MA, USA, 1993, pp. 179–183.

W. C. Palmer, Meteorological drought, Technical Report, US Department of Commerce, Weather Bureau, 1965.

S. M. Vicente-Serrano, S. Beguer'ıa, J. I. L'opez-Moreno, A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index, Journal of Climate 23 (2010) 1696–1718.

Z. Hao, V. P. Singh, Y. Xia, Seasonal drought prediction: advances, challenges, and future prospects, Reviews of Geophysics 56 (2018) 108–141.

S. Morid, V. Smakhtin, K. Bagherzadeh, Drought forecasting using artificial neural networks and time series of drought indices, International Journal of Climatology 27 (2007) 2103–2111.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, N. Thuerey, Weatherbench: A benchmark dataset for data-driven weather forecasting, Journal of Advances in Modeling Earth Systems 13 (2021) e2020MS002203.

J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, et al., Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, arXiv preprint arXiv:2202.11214 (2022).

H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, et al., The era5 global reanalysis, Quarterly Journal of the Royal Meteorological Society 146 (2020) 1999–2049.

G. Madec, R. Bourdallé-Badie, P.-A. Bouttier, C. Bricaud, D. Bruciaferri, D. Calvert, J. Chanut, E. Clementi, A. Coward, D. Delrosso, et al., Nemo ocean engine (2017).

G. Balsamo, A. Beljaars, K. Scipal, P. Viterbo, B. van den Hurk, M. Hirschi, A. K. Betts, A revised hydrology for the ecmwf model: Verification from field site to terrestrial water storage and impact in the integrated forecast system, Journal of hydrometeorology 10 (2009) 623–643.

G. J. Huffman, D. T. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, P. Xie, S.-H. Yoo, Nasa global precipitation measurement (gpm) integrated multi-satellite retrievals for gpm (imerg), Algorithm theoretical basis document (ATBD) version 4 (2015) 30.

S. M. Vicente-Serrano, S. Beguería, J. I. López-Moreno, A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index, Journal of

climate 23 (2010) 1696–1718.

V. Potop, C. Boroneanţ, M. Možnỳ, P. Štěpánek, P. Skalák, Observed spatiotemporal characteristics of drought on various time scales over the czech republic, Theoretical and applied climatology 115 (2014) 563–581.

J. H. Stagge, L. M. Tallaksen, L. Gudmundsson, A. F. Van Loon, K. Stahl, Candidate distributions for climatological drought indices (spi and spei), International Journal of Climatology 35 (2015) 4027–4040.

J. Spinoni, G. Naumann, J. V. Vogt, P. Barbosa, The biggest drought events in europe from 1950 to 2012, Journal of Hydrology: Regional Studies 3 (2015) 509–524.

S. Beguería, S. M. Vicente-Serrano, F. Reig, B. Latorre, Standardized precipitation evapotranspiration index (spei) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring, International journal of climatology 34 (2014) 3001–3023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

A. Dai, Drought under global warming: a review, Wiley Interdisciplinary Reviews: Climate Change 2 (2011) 45–65.

K. E. Trenberth, A. Dai, G. Van Der Schrier, P. D. Jones, J. Barichivich, K. R. Briffa, J. Sheffield, Global warming and changes in drought, Nature Climate Change 4 (2014) 17–22.

S. Mukherjee, A. Mishra, K. E. Trenberth, Climate change and drought: a perspective on drought indices, Current climate change reports 4 (2018) 145–163.

S. D. Schubert, R. E. Stewart, H. Wang, M. Barlow, E. H. Berbery, W. Cai, M. P. Hoerling, K. K. Kanikicharla, R. D. Koster, B. Lyon, et al., Global meteorological drought: a synthesis of current understanding with a focus on sst drivers of precipitation deficits, Journal of Climate 29 (2016) 3989–4019.