TU WIEN Informatics

Machine Learning-Based Classification of Climate Legislation

A Comparative Analysis of Legislative Coverage and Hazard Intensity

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Emilie Caillerie
Matrikelnummer 12402208

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ. Prof. Dr. Allan Hanbury
Mitwirkung: Dr. Mihaela Juganaru-Mathieu

Wien, 28. August 2026

Emilie Caillerie | Allan Hanbury

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Informatics

Machine Learning-Based Classification of Climate Legislation

A Comparative Analysis of Legislative Coverage and Hazard Intensity

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Business Informatics

by

Emilie Caillerie
Registration Number 12402208

to the Faculty of Informatics

at the TU Wien

Advisor: Univ. Prof. Dr. Allan Hanbury
Assistance: Dr. Mihaela Juganaru-Mathieu

Vienna, August 28, 2026

Emilie Caillerie | Allan Hanbury

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Declaration of Authorship

Emilie Caillerie

I hereby declare that I have written this thesis independently, that I have completely specified the utilized sources and resources and that I have definitely marked all parts of the work – including tables, maps and figures – which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.
I further declare that I have used generative AI tools only as an aid, and that my own intellectual and creative efforts predominate in this work. In the appendix "Overview of Generative AI Tools Used" I have listed all generative AI tools that were used in the creation of this work, and indicated where in the work they were used. If whole passages of text were used without substantial changes, I have indicated the input (prompts) I formulated and the IT application used with its product name and version number/date.

Vienna, August 28, 2026

Emilie Caillerie

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acknowledgements

I want to thank everyone who guided and supported me throughout my studies, as this Master's thesis marks the end of my journey to becoming an engineer.

First and foremost, I would like to express my deepest gratitude to my supervisor, Univ.Prof. Dr. Allan Hanbury, who took on my project at short notice when I needed it most. His guidance and support, through helpful suggestions and constructive criticism, were of vital importance to the completion of this work.

I am also thankful to École Nationale Supérieure des Mines de Saint-Étienne for giving me the opportunity to pursue a double degree at Technische Universität Wien, whom I also thank for welcoming me. Special thanks to Dr. Mihaela Juganaru-Mathieu for always supervising and encouraging my work.

Lastly, I would like to thank my family and friends for their support and encouragement during my studies. I am grateful to my parents for their continuous backing throughout my education, which made my two-year journey in Austria possible. I am also thankful to my friends for their emotional and moral support, even when we were far apart.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Kurzfassung

Klimagesetzgebung prägt globale Bemühungen gegen den Klimawandel. Ob legislative Maßnahmen den physischen Klimarisiken der Länder angemessen sind, wurde jedoch kaum global untersucht. Bestehende Studien stützen sich meist auf manuelle Auswertung und verbinden gesetzgeberische Abdeckung selten mit beobachteter Gefahrenintensität. Diese Arbeit untersucht, wie maschinelles Lernen Klimagesetze in Gefahrenkategorien klassifizieren kann, wie die Abdeckung mit der Gefahrenintensität zusammenhängt, und wie dies über ein Dashboard kommuniziert werden kann.

Regelbasierte und Zero-Shot-LLM-Ansätze klassifizierten Klimagesetze aus zwei Quellen in sechs Gefahrenkategorien, evaluiert anhand von 356 manuell annotierten Dokumenten. Die Abdeckung wurde mit der aus sechs Erdbeobachtungsvariablen abgeleiteten Gefahrenintensität verglichen. Zudem wurde ein Dashboard zur Visualisierung der Abdeckungslücken entwickelt und in einer Nutzerstudie evaluiert.

Zero-Shot-LLMs übertrafen regelbasierte und Mehrheitsklassen-Baselines deutlich, mit Schwächen bei einzelnen Kategorien. Der Zusammenhang zwischen Abdeckung und Gefahrenintensität war gefahrenspezifisch statt einheitlich und variierte teils zwischen Länder- und Kontinentalebene – gesetzgeberische Reaktionen sollten daher nicht als uniform angenommen, sondern gefahrenspezifisch und auf mehreren geografischen Ebenen bewertet werden. Das über drei Nutzerdurchläufe evaluierte Dashboard erhielt hohe Bewertungen für Vollständigkeit; Korrektheit und Klarheit verbesserten sich mit iterativer Verfeinerung.

Die Studie liefert eine reproduzierbare Pipeline für LLM-basierte Klassifizierung sowie ein Design-Artefakt zur Kommunikation von Abdeckungslücken und trägt so zur Politikanalyse und Klimagovernance-Forschung bei. Angesichts wachsender Klimarisiken ergänzen solche Werkzeuge die rein qualitative Politikanalyse sinnvoll.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Abstract

Climate legislation plays a central role in framing global efforts to respond to climate change. Yet, whether legislative efforts are proportionate to physical climate risks countries face has rarely been assessed at a global scale. Despite a growing number of climate laws and climate hazard datasets, existing studies of climate legislation rely largely on manual review and hardly connect legislative coverage to observed hazard intensity. This thesis addresses these challenges by investigating how machine learning-based methods can be used to classify climate legislation into hazard categories, how the resulting legislative coverage relates to observed climate hazard intensity, and how these relationships can be communicated through an interactive dashboard.

Rule-based and zero-shot large language models (LLMs) approaches were used to classify a dataset of climate laws combined from two data sources into six hazard categories, evaluated against a manually labeled benchmark of 356 documents. Legislative coverage was then correlated with hazard intensity derived from six Earth observation variables. Finally, a dashboard visualizing coverage-intensity alignment gaps was developed and evaluated through a preliminary user study.

The findings reveal that zero-shot LLM classification substantially outperformed both rule-based and majority-class baselines, though some difficulty remained for certain hazard categories. The relationship between legislative coverage and hazard intensity was found to be hazard-specific and in some cases differed between country- and continental-level aggregation, suggesting that legislative response to climate risk cannot be assumed to be uniform across hazard types and that assessments of policy-hazard alignment should be hazard-specific and reported across multiple levels of geographic aggregation. The dashboard, evaluated across three iterative user walkthroughs, was rated highly for completeness, with correctness and clarity improving as the artifact was iteratively refined based on user feedback.

The study contributes to policy analysis and climate governance research by providing a reproducible pipeline for LLM-based classification of climate legislation and a design artifact for communicating alignment gaps between legislative coverage and observed hazard intensity. As climate risks intensify and legislative bodies face growing pressure to respond, tools that can indicate where legislative attention is needed offer a valuable complement to purely qualitative policy review.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Contents

Kurzfassung **ix**

Abstract **xi**

Contents **xiii**

1 Introduction **1**
1.1 Aim of the Thesis . . . 2
1.2 Research Questions . . . 2
1.3 Methodology . . . 3
1.4 Structure of the Thesis . . . 4

2 Background **7**
2.1 Artificial Intelligence . . . 7
2.2 Classification Approaches . . . 8
2.3 Hierarchical Time Series . . . 9

3 Related Work **11**
3.1 Classification of Climate-related Documents . . . 12
3.2 Uses of Climate Observation Data . . . 13
3.3 Visualization of Climate Information . . . 14

4 Datasets **15**
4.1 Legislation Data . . . 15
4.2 Sensor Data . . . 24
4.3 Linking Legislation Data and Sensor Data . . . 32

5 Experiment Setup **35**
5.1 Models . . . 35
5.2 Evaluation Metrics . . . 41

6 Results **47**
6.1 Classification Results . . . 47
6.2 Correlation Results . . . 53

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

6.3 Visualization Results 61

7 Discussion **65**
7.1 Research Question 1 65
7.2 Research Question 2 66
7.3 Research Question 3 68

8 Conclusion **71**
8.1 Summary 71
8.2 Limitations 72
8.3 Future Work 72

A Detailed Hazard Label Mapping **75**

B Detailed Classification Results **77**

C Visualization Dashboard **81**

Overview of Generative AI Tools Used **83**

Bibliography **85**

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 1

Introduction

Climate laws and policies are major instruments for climate actions at national and international levels [MM21]. They provide a framework for compliance with treaties, such as the United Nations Framework Convention on Climate Change and the Paris Agreement. In practice, the analysis of laws and policies is often limited by the difficulty to parse long legislative documents, especially when language and terminology vary with the country. This makes the understanding of climate legislation particularly challenging for policymakers seeking to improve their efficacy. In this regard, previous work has been carried out to extract and classify climate policy content into target categories [JMM+24], and to classify Nationally Determined Contributions into targets, actions, and sectors to assess their coherence with treaties [SLT24]. However, the classification of climate laws and policies by labeling them with the hazard they are addressing does not yet exist on a large scale.

The Climate Change Laws of the World database [CG23] contains national climate-related laws and policies from all over the globe. In this database, only 4.4% of the over 12,000 climate-related documents were manually annotated with a label on the hazard they are addressing. This makes it impossible to observe which climate risks are addressed and which are neglected by legislation.

Furthermore, no existing study has compared nor made visible the alignment of legislative coverage against observed risk intensity for multiple hazard categories on a global scale. This is critical, not only for policymakers trying to identify areas where efforts are misaligned with actual risks, but also for citizens aiming to understand the extent to which their country is addressing the climate hazards it faces.

This research is motivated by the resolution of the aforementioned problem. By using the machine learning approaches developed in this work, a wide range of legislative documents relating to climate change are labeled with the hazards they address. Additionally, comparison and visualization of legislative coverage against observed risk intensity is

performed for a given set of hazard categories, revealing gaps where legislative attention does not align with the severity of the risks countries face.

1.1 Aim of the Thesis

This thesis aims at developing a Natural Language Processing (NLP) and Large Language Model (LLM)-based pipeline to classify climate legislation by hazard category, and then quantitatively comparing and visualizing the resulting legislative coverage against observed hazard intensity. The analysis focuses on the time period from 2000 to 2025, covered by all chosen datasets.

Three distinct goals and artifacts are defined, one for each research question (Section 1.2). First, NLP and LLM-based multi-label classification pipelines are developed to annotate climate legislation by hazard category (RQ1). Second, the legislative coverage is computed and compared against observed climate hazard intensity, incorporating a hierarchical trend analysis to account for regional dynamics (RQ2). Third, the identified gaps between coverage and intensity are be communicated through a visualization dashboard (RQ3).

1.2 Research Questions

To successfully and efficiently address the above-mentioned problem, the following research questions are defined.

RQ1: *How do NLP and LLM-based approaches perform and compare for multi-label classification of climate laws into hazard categories in terms of precision, recall, and F1-score?*

RQ1 addresses the absence of hazard classification of climate legislation. A rule-based NLP approach, where rules are used to process textual data, and zero-shot classification approaches by different LLMs, where a model is not trained on the classes and can classify data it has never seen before, are developed and evaluated against a benchmark of labeled documents from the Climate Change Laws of the World database.

RQ2a: *To what extent does climate legislative coverage correlate with the intensity of locally observed climate hazard risks?*

RQ2b: *To what extent do continental-level correlation between legislative coverage and hazard intensity differ from country-level correlation?*

Both RQ2 sub-questions focus on the link between legislative coverage and hazard intensity. The labels generated from RQ1 can be used to compute the cumulative count of laws per country per hazard category per year, referred to as legislative coverage. Hazard intensity is defined as the deviation of the variable associated with each hazard, relative to its historical baseline.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Because a country's risk exposure is not only affected by its own legislation but also by that of its neighbors, the correlation analysis is conducted on two levels of granularity. First, the correlation is computed on a country-level (RQ2a). Then, countries are grouped by continent and a hierarchical trend analysis is conducted, providing a complementary analysis of alignment gaps (RQ2b).

RQ3: *To what extent can a visualization dashboard effectively communicate national gaps between climate hazard risk intensity and legislative coverage in terms of completeness, correctness, and clarity?*

RQ3 uses the variables computed in RQ2 to develop a visualization dashboard, incorporating a world map with country-level gap indicators, a timeline of legislative coverage, hazard intensity, and policy gap evolution per hazard category. Three design criteria are defined as follows: completeness requires all hazard categories and countries with available data to be represented, correctness ensures that the visualization of the gaps matches the underlying data, and clarity guarantees that the gaps between legislative coverage and hazard intensity are unambiguously readable. The extent to which the dashboard fulfills these criteria is assessed through a preliminary user study involving a small group of participants, providing an indication on the dashboard's communicative effectiveness.

1.3 Methodology

The research approach followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) [MPCOF+21] methodology. Within it, three modeling and evaluation phases were performed in a sequential iteration, each phase needing the previous one to be completed. The development of the visualization dashboard as a design artifact followed the Design Science Research Methodology [PTRC07]. Additionally, a Literature Review was carried out, following the guidelines by Okoli [Oko15], to help establish an understanding of the current state of the art. The steps in the context of this thesis can be described as below:

1. **Business Understanding**. This phase was covered by the submission and the presentation of a proposal. Outlining the problem and success criteria, alongside the definition of research questions established the business objectives and data mining goals.
2. **Literature Review**. After narrowing down the literature by screening a large panel of research and defining which studies to include and exclude, information was extracted from chosen papers to conduct the review.
3. **Data Understanding**. An exploratory analysis of the Climate Change Laws of the World database [CG23], the Climate Policy database [NWP25] and chosen datasets from the catalog of the Copernicus Climate Data Store [Cop] was undertaken.
4. **Data Preparation**. The data cleaning and preprocessing was documented, explaining how the datasets used for modeling were constructed.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5. **RQ1 Modeling**. The multi-label classification models were built without training: a rule-based NLP method using a domain-specific glossary of hazard-related keywords, and a zero-shot LLM-based method prompting multiple pre-trained models to assign labels from a predefined set of hazard categories. A majority-class baseline was included as a reference.
6. **RQ1 Evaluation**. The models were assessed and compared by a controlled benchmarking experiment. The chosen metrics are *precision*, *recall*, and *F1-score*, as well as *Jaccard score* and *Hamming loss* for additional insight. The final outcome is that the models are evaluated and compared on the subset of labeled documents from the Climate Change Laws of the World database.
7. **RQ2 Modeling**. Using the model resulting from the previous step, legislative coverage and observed hazard intensity were computed per hazard category and country, in order to conduct a Spearman correlation analysis.
8. **RQ2 Evaluation**. Correlation results were assessed for statistical significance and evaluated by a quantitative data analysis. The chosen metrics are *correlation coefficient* and *p-value*. The final outcome is that the statistical significance of the correlation results is tested at both country- and continental-level.
9. **RQ3 Modeling**. A visualization dashboard was designed and implemented to communicate gaps between legislative coverage and hazard intensity.
10. **RQ3 Evaluation**. The dashboard was evaluated by a user study combined with a manual verification. The three chosen design criteria are *completeness*, *correctness*, and *clarity*. The final outcome is that all three criteria are assessed.
11. **Deployment**. The resulting classification, correlation, and dashboard were documented and made available for traceability and reproducibility, with the source code published in a public repository.[1]

1.4 Structure of the Thesis

The structure of the thesis is as follows:

Chapter 2 provides background information on topics relevant to this thesis. This gives basic knowledge to people unfamiliar with the tools and concepts that are used in the remainder of the work.

Chapter 3 presents the literature review conducted ahead of the research itself. State-of-the-art work on methods for multi-label text classification, on current uses of climate observation data, and on visualization tools for communicating climate information are covered.

Chapter 4 describes the datasets used for the research. An in-depth exploratory analysis is detailed, examining data sources and statistical properties, to justify the data preprocessing steps.

[1] https://github.com/emiliecaillerie/master-thesis

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Chapter 5 introduces the experimental setup of the thesis. The models chosen for each step are described alongside the metrics selected for their evaluation.

Chapter 6 analyzes the results emerging from the experiments.

Chapter 7 addresses each research question individually and provides an answer to them.

Chapter 8 concludes the thesis through a summary. The limitations of this work are discussed, contemplating potential extensions with future work.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 2

Background

This chapter introduces the theoretical foundations behind the methodological choices made throughout this thesis. It begins by situating Large Language Models (LLMs) within the domain of Artificial Intelligence (AI), by exploring the evolution of the field. Then, it examines classification as a predictive modeling task, distinguishing between different label structures and presenting the main text classification techniques that motivate the comparative analysis conducted in this research. Finally, it presents the concept of hierarchical time series, which motivates the multi-level aggregation structure of legislative coverage and hazard intensity data used in the subsequent analysis.

2.1 Artificial Intelligence

Recent years have seen the rise of LLMs, particularly, the mainstream success of OpenAI's ChatGPT has impacted the society in varied sectors and the popularity of AI has risen. The rapid developments in AI justify the need to properly identify and define each underlying concept. The following definitions, provided by the 2026 Guide to Machine Learning [SLBS26], help understand the nested hierarchy of the notions present in this thesis illustrated in Figure 2.1.

AI is a technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity, and autonomy.

Within this vast domain, Machine Learning (ML) covers algorithms that can be trained to grasp patterns from data and to then use this knowledge to make conjectures on unseen data. This pattern recognition ability enables ML models to make decisions or predictions without explicit instructions. Deep Learning (DL) is a subset of ML, with multilayered neural networks in its center. The most prominent application of DL in the context of this thesis are LLMs, a category of DL models trained on large amounts of data, able of understanding and generating natural language or other types of content.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

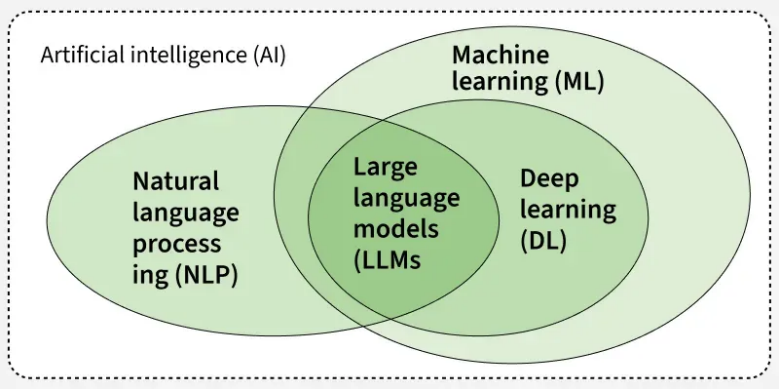


Figure 2.1: Differentiation between LLM and NLP (Source: [Gee25])

LLMs are built on a type of neural network architecture, a transformer, particularly good at handling sequences of words and capturing patterns in text.

While LLMs represent the current state-of-the-art in text generation, they function as a primary tool within the broader field of Natural Language Processing (NLP), a subfield of computer science and AI using ML to enable computers to understand and communicate with human language.

It is interesting to study the evolution of NLP, as characterized by Noguer [NIA24], in order to identify the three main approaches used in the field. In the 1950s, rule-based systems emerged as the initial approach. Relying on syntactic rules and grammar, they were quickly limited by the lack of flexibility and adaptability to the nuances of human language. In the 1980s, a shift towards statistical approaches took place, allowing models to learn patterns and thus to handle the complexities of natural language. Since the 2000s, DL and neural networks further improved the performances of NLP. This latest revolution also saw the emergence of advanced transformers models such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT).

2.2 Classification Approaches

In Machine Learning (ML), classification is a predictive modeling process where algorithms are used to predict the correct label for input data [SLBS26]. The classifying approach varies based on the structure of the labels and the nature of the data to be analyzed.

When it comes to the structure of the labels, there exist five kinds of classification according to Herrera et al. [HCRdJ16]. Binary classification is the simplest case, resulting in a dataset where each instance belongs to exactly one of two distinct classes. Multi-class classification also assigns exactly one attribute to each instance, but from three or more classes. Multi-label classification assigns each instance a binary (present/absent) label for each of several classes, resulting in any number of classes from none to all being assigned to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

a single instance. Multi-dimensional classification generalizes multi-labeling by replacing these binary labels with multiple labels. For example, instead of indicating whether an instance belongs to each of several hazard categories independently (multi-label), it might assign each instance a severity level for each hazard category (multi-dimensional). Multi-instance classification is rather different, it assigns a class label to a bag of instances rather than to individual ones. As for the nature of the data, depending on the context, one can be required to classify numbers, text, or images.

In the context of this thesis, the focus is placed on multi-label text classification. Each document of the dataset will be assigned zero, one, or more hazard labels, indicating which climate risk it addresses.

In Natural Language Processing (NLP), text classification is generally used for tasks such as spam detection, sentiment analysis, topic labeling, or news categorization. Multiple approaches and techniques can be used for text classification.

First, rule-based classification is the most fundamental technique. It uses predefined IF-THEN rules and keyword patterns to label the data without requiring training. Second, traditional ML text classification algorithms such as Naive Bayes Classifier, Support Vector Machines, and Logistic Regression rely on statistical patterns learned from labeled training data. Finally, modern ML text classification methods using neural networks or transformers are increasingly being used with diverse strategies. While zero-shot learning relies on the model's pre-trained knowledge only, fine-tuning relies on training a pre-trained model on a task-specific dataset, adjusting its weights to make it highly specialized [BMR+20].

In this thesis, the chosen datasets are limited and do not allow for training. Indeed, only a handful of the documents is labeled with the target variable, which will be used for evaluating the models. Consequently, while fine-tuned domain-specific models might offer better results, this research focuses on evaluating and comparing a rule-based approach and a zero-shot LLM classification for identifying climate hazards addressed in legislative documents.

2.3 Hierarchical Time Series

A hierarchical time series is a collection of time series organized in a hierarchical aggregation structure, where time series at a lower level of the hierarchy can be summed to produce the time series of the next level [HA18]. It is especially used as a forecasting method in many applications such as energy, tourism, and in the retail industry [MPS21].

As illustrated in Figure 2.2, each circle represents one time series. Each parent node's series is obtained by summing its children's series at each matching time point. This requires all child series being aggregated to share a common time index, which is the case for both legislative coverage and hazard intensity in this thesis, as both are available for the same range of years across all countries.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

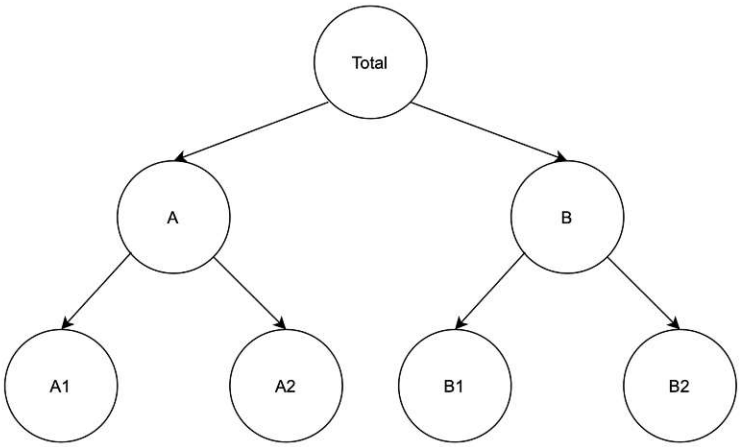


Figure 2.2: An example of hierarchical time series (Source: [Lew21])

In the context of this thesis, both legislative coverage and hazard intensity are computed as time series at the country-level. Aggregating these into continental-level series provides an additional perspective on a country's exposure to climate hazards, which is not only determined by its own legislative efforts but also by that of its neighbors. This is done to observe whether regional dynamics, that the country-level analysis may not reveal, can be captured at the continental-level.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 3

Related Work

This chapter covers the literature review conducted to analyze existing related work in three areas relevant to this thesis.

First, papers on classification of textual documents were reviewed to identify and compare approaches applicable for this study of climate legislation. Second, current uses of climate observation data in policy-relevant contexts were surveyed to justify processing methods of datasets. Third, existing visualization tools and approaches for communicating climate information were reviewed to understand how to convey a sense of urgency in a dashboard. Key results, gaps, and limitations are identified across these three areas further motivate the research questions.

The initial search was executed by using one set of keywords for each intended purpose of the review to use as search prompts on five different research-oriented engines. To narrow down the research, only papers written in English language and published during the period of 2018 to 2026 were considered. This choice is justified by the need to review recent papers using up-to-date tools and methods, especially in the ever growing field of Machine Learning. They have to be freely accessible, whether it be through open access or through the TU Wien Library. Papers relating directly to climate change or more generally to climate are examined. The only exception is made if a paper, even though not referencing climate, appears to use tools or methods relevant to this research. All papers whose central focus is prediction or forecasting of climate-related issues are ignored, because such tasks are not relevant for classification, which is the purpose of this research. It comes naturally that each paper has to be relevant to one or more research questions.

The literature screening process was conducted in four distinct steps, by gradually increasing the depth of analysis. First, any records found to be irrelevant by their title were filtered out during the search. Any duplicates emerging from different databases were removed, to ensure the uniqueness of the each record. Second, the abstracts of all

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

remaining records were reviewed. Papers were excluded if they did not address a topic relevant to the research, or if they were not available through institutional access. Third, the introduction and conclusion of the reports were assessed for eligibility. Articles were excluded at this point if their scope was too narrow to address the research questions or if they were found unrelated upon closer inspection. Finally, a quality assessment was performed on the remaining articles. This final filter ensured that only studies with sufficient rigor and peer-reviewed credibility were included.

3.1 Classification of Climate-related Documents

Natural Language Processing (NLP) has extensively been applied to climate policy documents in recent years with varied approaches. Following the developments in Machine Learning (ML), five methods for the classification of climate-related documents were noted, evolving from most basic to most advanced.

First, rule-based approaches have been used as baseline models [WKBL22, BBA25], yet they remain competitive for non-complex tasks and have the advantage of being most transparent[CSX+21].

Then, traditional ML algorithms like Support Vector Machine, Random Forest, or Logistic Regression have been used [CBS25] or reviewed [GMM25] but were not viewed as particularly effective, especially compared to newer methods.

Next, general-purpose Large Language Models (LLMs) like OpenAI's ChatGPT, Google's Gemini, or Meta's LLaMA, have recently started being used for climate and sustainability purposes [TLCV23, SCCBK24, KZNM25]. These models have the advantage of being applicable in a zero-shot setting. In spite of that, or maybe because of that, they have demonstrated conflicting results in classification tasks [GL24]. They can showcase competitive results on simple tasks and, on specific tasks, they can be improved by adjusting the prompts without retraining [TLCV23], underlying the importance of prompt engineering. This dependence on human intervention and decision-making limits the reproducibility and scalability of these models [LHC+25].

Finally, transformers are the most state-of-the-art models, especially when domain-specific. Pre-trained BERT-based models, for example DistilBERT or RoBERTa, have been fine-tuned for climate policies' investigation purposes [SLT24, Dut25]. Although in climate-related fields, ClimateBERT [WKBL22], a model based on DistilRoBERTa and pre-trained on a climate-related text corpus, is seen as the reference. It has been used as a primary model [JMM+24] and appears to be the better-performing one [TLCV23, BBA25]. ClimateBERT has also been fine-tuned for specific tasks, such as identifying climate literature that investigates climate policy instruments [CBDH+25] or detecting net zero and reduction targets in corporate disclosures [SBK+23]. Most studies agree that fine-tuned domain-specific models outperform general-purpose LLMs and rule-based approaches on climate-specific tasks [SBK+23, GL24, YSZ24, BBA25]. Still, it is worth noting a caveat: fine-tuned models may overfit rather than properly generalize, because the currently

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

available datasets may be too simple [CBS25].

There exist countless datasets containing climate-related documents, varying in source and scope. Textual information has been extracted from news, reports, and social media posts [WKBL22, OOOO23]. By contrast, Nationally Determined Contributions (NDCs) have served as primary data source [CSX+21, SCCBK24, SLT24], sometimes alongside Sustainable Development Goals (SDGs) [CA25, LHC+25]. Alternatively, Climate Policy Radar, gathering climate laws, policies, and United Nations Framework Convention on Climate Change (UNFCCC) reports, has also been used as a data source [BBA25, JMM+24]. Climate Policy Radar is also behind the initiative of Climate Change Laws of the World, which has been used on its own [Dut25] or alongside the Climate Policy Database by the New Climate Institute [CBDH+25].

The majority of the papers investigated use English documents only, a limitation acknowledged explicitly [CBDH+25, KZNM25, SLT24], because of the language and geography bias it might induce [GMM25]. In the case that larger datasets were chosen, machine translation was performed [BBA25, JMM+24], which itself may lead to parsing or translation errors, potentially affecting the results.

In this thesis, the choice of datasets is limited to English-language and mirrors the one of Callaghan et al. [CBDH+25], by merging the Climate Change Laws of the World [CG23] and the Climate Policy Database [NWP25].

Lastly, documenting the impacts of working with AI tools, especially in a climate-related research, is essential. Out of all the previously cited studies, only two of them monitored the emissions associated with training and running their models [SLT24, WKBL22]. As a result, in this thesis, a calculation of carbon emissions is also carried out.

3.2 Uses of Climate Observation Data

Climate observation data refers to satellite imagery and reanalysis, such as land cover imagery, sea level measurements, and reanalysis datasets that provide continuous, large-scale measurements of Earth's climate.

Machine Learning (ML) techniques applied to this type of data primarily focus on prediction and forecasting of climate variables. By contrast, Santhosh [San25] has demonstrated Artificial Intelligence (AI) tools' utility for habitat monitoring and conservation. AI enables the automation of vital tasks, such as land cover classification or change detection. However, ethical concerns such as privacy violations from surveilling local communities, and unequal access to these technologies must be considered, along with limitations in data availability and spatial resolution.

The utility of satellite and reanalysis data specifically for policy monitoring and risk assessment is also demonstrated. Institutional case studies, on international [KJMB+20]

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

or local [SAL+25] scales, reinforce the role of Earth observation not only for scientific monitoring but also for policy support. This lays the groundwork for linking policy documents with unstructured satellite data [CA25, RDK+23], which is the approach this thesis builds on by combining hazard intensity derived from reanalysis data with legislative coverage extracted from policy documents.

3.3 Visualization of Climate Information

Visual communication strategies, such as data visualizations or interactive infographics, are facilitating accessibility to complex scientific information by simplifying data into easily understandable representations.

The effects of visuals on awareness of climate change have been reviewed [Met20], displaying a positive impact of images on people's understanding of the situation, a rise in their level of concern, or a change in their attitudes. Varied emotions can be elicited from visuals, whether positive or negative depending on the content, and seem to affect people's intention to take action. However, fear is not an emotion one wants to induce, as it is found to be counterproductive by making people feel helpless. Additionally, while not all best practices are recommended or applicable to every case, they enable the comprehension of a key to effective visualization: balancing information density, interactivity, scientific credibility, and user-centered design [TCC22].

Visual formats, especially when interactive, outperform text alone in improving environmental literacy and drive behavioral change in the context of climate change [GHW25]. Custom prototypes of animated [SZC+18], interactive [AL24], and personalized [RFM25] communication tools have been implemented, further emphasizing that visual stimulation and storytelling increases comprehension. However, while personalization and localization of imagery improves engagement, such narrative risks reinforcing existing beliefs or introducing bias [RFM25].

Although research provides the fundamentals of visual communication and strategies to adopt for effective and impactful design, existing tools remain limited in scope. The United Nations Environment Programme (UNEP) Web Intelligence platform [UNE21] represents public sentiment and media coverage trends around environmental topics, drawn from social media and news sources, but it does not address legislative content or hazard-specific risk data. To this date, no existing tool integrates, in an interactive dashboard, the coverage of climate legislation across hazard categories with the observed intensity of those hazards.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

CHAPTER 4

Datasets

This chapter presents the datasets used throughout this thesis. It first introduces the legislation data, covering its sources, features, statistical properties, missing values, ethical concerns, and the preparation steps applied to it. It then examines the sensor data in the same way. Finally, it describes how the legislation and sensor datasets are linked.

4.1 Legislation Data

This section introduces the legislation data, consisting of climate-related laws and policies. It first presents the two source databases and their respective scope and limitations. It then examines the features retained for analysis, their statistical properties, and the missing values and ethical concerns they raise. Finally, it details the data preparation steps applied to harmonize, clean, filter, and merge both databases into a single dataset.

4.1.1 Source

For laws and policies, the Climate Change Laws of the World database [CG23] and the Climate Policy database [NWP25] are chosen. They are both standard databases when it comes to legal climate-related documents. From now on, the Climate Change Laws of the World database will be referred to as CCLW, and the Climate Policy database as CPDB.

CCLW is maintained by the Grantham Research Institute on Climate Change and the Environment at the London School of Economics, in partnership with Climate Policy Radar [CG23]. It covers climate-related laws and policies, which they define as legal documents that are directly relevant to climate change mitigation, adaptation, loss and damage or disaster risk management. The database is updated continuously and the version dated from April 26th 2026 was downloaded for this thesis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

One of their core principles is to be inclusive and, as such, the database covers the 196 countries part of United Nations Framework Convention on Climate Change (UNFCCC) and several other territories not part of the United Nations like Taiwan, Palestine, and Western Sahara. Nonetheless, they currently focus on national level contributions only, thus climate change policy response cannot be captured fully in countries where decision-making at the sub-national level is significant.

They are very transparent with the limitations of their database. Their current data collection process prioritizes most recently passed laws and policies. Language barrier, media coverage, and expertise of collaborators may lead to certain countries being more present than others. Up until January 2023, targets and hazard were manually identified by researchers in laws and policies. They used translation software to understand the content of non-English documents, and as such, the categorization of a given law might differ with interpretation.

CPDB is maintained by NewClimate Institute in collaboration with Wageningen University and PBL Netherlands Environmental Assessment Agency [NWP25]. It covers national mitigation-related policies. The database is updated annually and the version dated from April 21st 2026 was downloaded for this thesis.

The CPDB's central focus is put on tracking national climate policy developments of 42 countries including the European Union, which is treated as one country. Other countries are also included with non-exhaustive information.

It is important to note that their focus is primarily put on mitigation laws and policies, meaning the adaptation coverage is much lighter than CCLW. CPDB will be used as a complementary source to CCLW, which constitutes the primary database for this thesis.

4.1.2 Features

CCLW contains 12,478 rows described by 36 features, and was merged with the 6,727 rows described by 28 features of CPDB, only keeping the common features as listed in Table 4.1.

CCLW column (type)	**CPDB column** (type)	**Description**
`Family Title` (str)	`policy_title` (str)	Title of the document
`Family Summary` (str)	`policy_description` (str)	Summary of the document
`Geography ISOs` (str)	`country_iso` (str)	ISO 3166-1 alpha-3 code
`Geographies` (str)	`country` (str)	Name of the country
`First event in timeline` (str)	`decision_date` (float)	Year of law passing
`Hazard` (str)	N/A	Hazard category addressed

Table 4.1: Key features of the legislation datasets

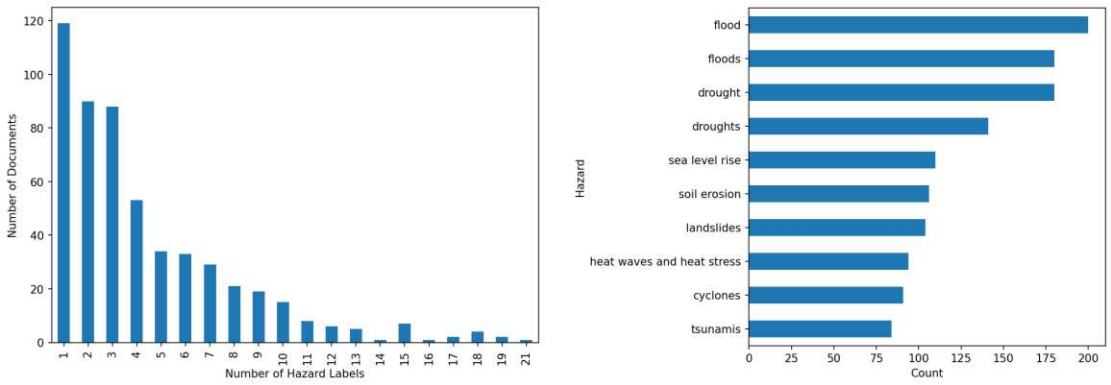
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.1.3 Statistical Properties

This subsection examines the statistical properties of the merged legislation dataset, focusing only on the features used in this research. It first analyzes the target variable used for the classification task, before investigating the geographic and temporal distribution of the documents. Finally, it looks at the distribution of text length, which determines the suitability of each document for the classification pipeline.

Target Variable

The target variable is the `Hazard` column in CCLW. It represents the hazard(s) addressed inside each document. The 538 rows where this column is non-empty are used as a benchmark dataset. The distribution of hazard label counts per document is plotted, in order to validate the relevance of a multi-label classification task. Over half of the labeled documents are annotated with three hazards or less, suggesting that most laws address a limited number of hazards. Nevertheless, around a quarter have between six and twenty-one different labels, and seem to be addressing a wide range simultaneously (Figure 4.1a).



(a) Distribution of hazard label counts

(b) Most frequent hazard labels

Figure 4.1: Distribution and frequency of hazard labels in the benchmark dataset

There exist 81 unique labels in total, with the most frequent ones displayed in Figure 4.1b. A deeper look into them reveals the need for a cleaner hazard mapping which will be handled during data preparation (Section 4.1.6). Some labels exist in both their singular and plural form, for instance `flood` and `floods` or `drought` and `droughts`. Other variations of the same label exist, like `heat waves` and `heatwaves` or `wild fire` and `wildfire`, while some may even contain mistakes such as `earhquakes`. Lastly, diverse regional terms, `cyclones`, `hurricanes`, or `typhoons`, all fall under the same meteorological phenomenon and can be grouped under the same category of `storm`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Geographic Coverage

A total of 203 countries are found in CCLW, compared to the 198 of CPDB. Among them, 195 are common in both datasets.

In CCLW, all documents with International Organization for Standardization (ISO) codes `XAB` and `XAA` indicate respectively „International“ and „No Geography“. After a quick manual search, it looks like these documents refer to UNFCCC reports and are irrelevant in the context of this thesis. This choice does not reduce the size of the benchmark, because none of the documents with either ISO code is labeled with hazard(s).

Laws and policies at the European Union (EU) level are also present, with ISO codes `EUR` in CCLW and `EUE` in CPDB. All of such documents are kept, not only because the ISO codes are among the most represented in both datasets, but also due to the fact that EU legislation affects all 27 member states and are relevant in each of them. Later in the work, EU-level laws are distributed to each relevant member state.

Figure 4.2 illustrates the geographic distribution of documents in both databases. In this map, the special ISO codes `XAB`, `XAA`, `EUR`, and `EUE` described earlier are set aside. It is unambiguously noticeable that the coverage of both CCLW and CPDB is rather focused on a small number of developed countries. In CCLW, Brazil, France, Canada, China, and Australia stand out, while most parts of Africa and the Middle East show very low coverage. Likewise in CPDB, the USA, Canada, Australia, China, and India stand out, while Africa, South America, Eastern Europe, and Central Asia are underrepresented. This second observation is not surprising, considering their focus on 42 countries.

Temporal Coverage

The time period covered by the documents is rather long and ranges from 1957 to 2026 in CCLW, and from 1927 to 2026 in CPDB. Though solely 2.5% in CCLW and 8% in CPDB are dated before 2000, meaning the majority of the documents are recent. This aligns with the scope of this thesis, targeting the period from 2000 to 2025.

Within this relevant time space, the yearly distribution of documents reveals different patterns between the two datasets (Figure 4.3). CCLW displays an almost uniform increase every year, with a steeper slope after 2015, which may be coinciding with the adoption of the Paris Agreement. CPDB is more steady and does not exhibit any particular pattern. In both cases, 2025 and 2026 show lower counts, confirming that recent documents take some time to be added to the databases.

Title and Summary Length Distribution

The documents' title and summary are the features used for classification. It was attempted to retrieve full document text from source URLs, instead of only relying on the titles and summaries that can sometimes be incomplete or lack information. However, this turned out to be unsuccessful for the majority of the tested sample, due to the rendering of pages in JavaScript, scanned PDF formats that would require going through Optical

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

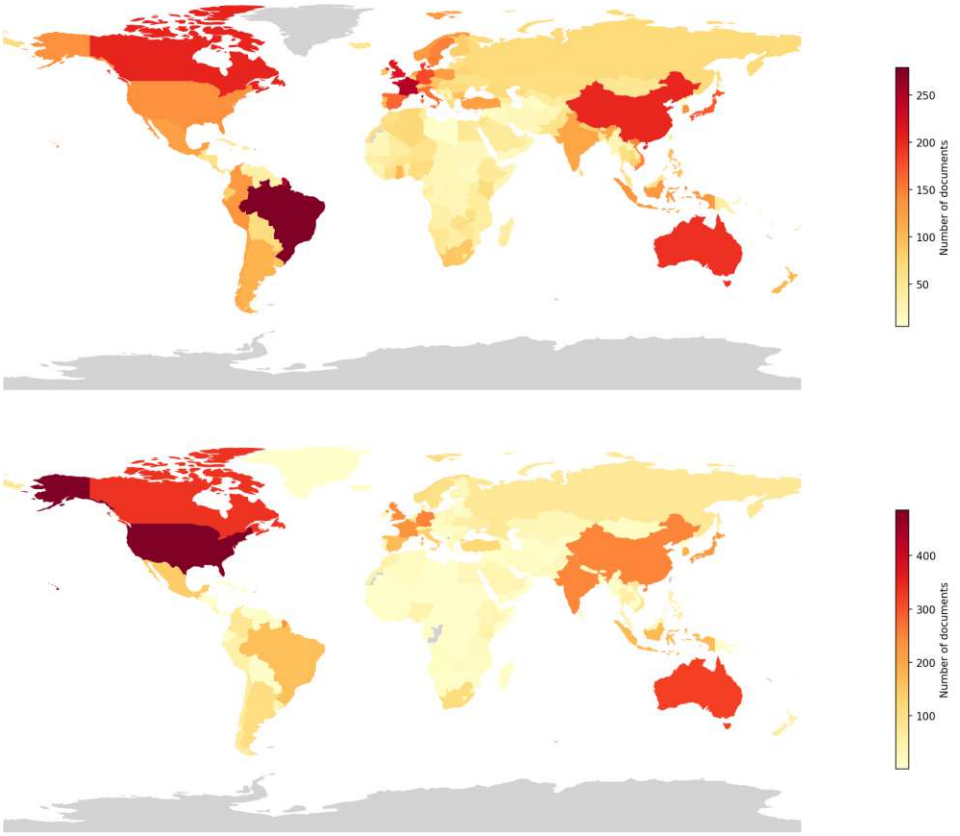


Figure 4.2: Number of documents per country in CCLW (*top*) and CPDB (*bottom*)

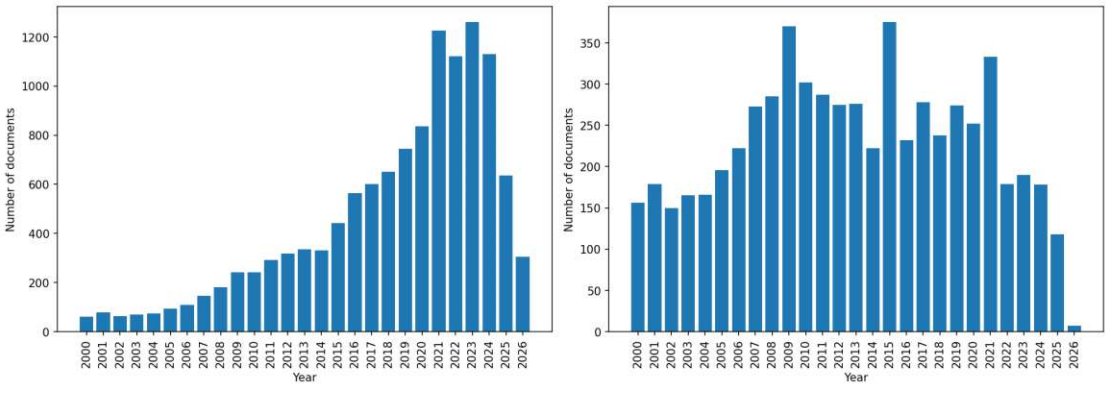


Figure 4.3: Number of documents per year in CCLW (*left*) and CPDB (*right*)

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Character Recognition process, access restrictions, and non-English content. Full-text retrieval was therefore abandoned in favor of classification solely based on available titles and summaries. Table 4.2 reports the length statistics for both titles and summaries in each dataset.

Statistic	CCLW title	CPDB title	CCLW summary	CPDB summary
Mean	9	7	88	47
Median	7	5	56	32
Standard deviation	6	4	109	58
Minimum	1	1	0	0
Maximum	76	36	1,879	706
Under 25 words	97.1%	99.2%	36%	45.1%

Table 4.2: Title and summary length statistics of the datasets, measured in words

The summaries need to be of appropriate length and need to contain enough content to be able to extract some information. On average, the summaries in CCLW are twice as long as the descriptions in CPDB. In CCLW, the mean and median word counts are respectively 88 and 56, indicating a skewed distribution where really long documents pull the mean upwards. In CPDB, 26.3% of the documents have a description length of 0 words, reflecting a large proportion of documents with simply no description.

A notable portion of the summaries contain only very short text which might be insufficient for a classification task. In CCLW, 36% of documents' summary are written with less than 25 words. Manually looking into them reveals that these are principally titles of reports and Nationally Determined Contributions, offering little to no descriptive content on the hazards addressed. In CPDB, 45.1% of documents' summary are written with less than 25 words, which are mostly one-sentence descriptions. Unlike the CCLW case, manual inspection shows that many of these remain informative enough to identify a hazard or climate response despite their short length.

Based on this distinction, a minimum threshold of 25 words is applied to summaries originating from CCLW, to exclude documents lacking sufficient text for hazard classification. The exclusion is not applied to summaries from CPDB, for two reasons. First, as explained above, many of its short descriptions remain usable for classification despite their brevity. Second, applying the same threshold to CPDB would exclude nearly half of the database, substantially reducing the size and diversity of the data.

Titles are considerably shorter than summaries in both datasets, with mean word counts of 9 in CCLW and 7 in CPDB. Almost all titles fall under 25 words (97.1% in CCLW and 99.2% in CPDB), meaning a length-based threshold, as applied to summaries, would not be meaningful for titles. Despite their brevity, titles often contain hazard-relevant information not always repeated in the summary, and are therefore included alongside the summary as input for classification.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.1.4 Missing Values

Among the features used in this research, a few have a significant amount of missing values. The most critical in terms of number and importance is the `Hazard` column in CCLW, with a missing values rate of 95.7%. This absence is what constitutes the motivation for the classification task of RQ1. The documents' summary column, which is essential for the classification as it provides the input text, is missing for only 0.3% of documents in CCLW but for 26.3% in CPDB, meaning over a quarter of CPDB documents are excluded because they cannot be used for classification without any text available. Several other features not used in this research also contain substantial missing values but are not discussed further, as their absence has no impact on the analysis.

4.1.5 Ethical Concerns and Biases

As shown in Figure 4.2, the countries of the Global North are better represented than those of the Global South. This unbalanced distribution leads to a geographic bias, where some of the most vulnerable regions, like Sub-Saharan Africa or South Asia, are the least represented. This may affect the reliability of the results of RQ2, which will be more robust for well-represented countries.

Additionally, a moderately positive correlation exists between document count and population size in both databases, with a correlation coefficient of 0.33 in CCLW and of 0.48 in CPDB. Similarly, a slightly higher positive correlation exists between document count and country size, with a correlation coefficient of 0.39 in CCLW and of 0.57 in CPDB. These indicate that larger population and area are associated with bigger document counts but neither suffices in explaining this distribution. The higher correlations in CPDB reflects its focus on 42 countries, including the whole G20, with significant economies and legislation.

Overall, this suggests that such countries are overrepresented because they have more legislative bodies adopting more laws and policies, and because they receive greater attention on the international scene. However, this does not fully explain the geographic distribution, which could also be influenced by data collection methodology or language bias.

4.1.6 Data Preparation

This subsection describes the data preparation pipeline applied to the legislation data. It first covers the harmonization of columns across both databases, then the cleaning steps applied to fix errors and standardize formats. It then examines the filtering criteria used to remove irrelevant or unusable documents, and finally details the process of identifying and merging duplicates across both datasets.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Harmonizing

To simplify the work of cleaning and filtering both datasets at the same time, the names of the features kept for merging were harmonized following Table 4.3. A Source column (CCLW or CPDB) was added to track document origin. The Hazard column was also included in CPDB to prepare for the classification task.

CCLW column (type)	**CPDB column** (type)	**Merged column** (type)
`Family Title` (str)	`policy_title` (str)	`Title` (str)
`Family Summary` (str)	`policy_description` (str)	`Summary` (str)
`Geography ISOs` (str)	`country_iso` (str)	`Country` (str)
`Geographies` (str)	`country` (str)	`Country_name` (str)
`First event in timeline` (str)	`decision_date` (float)	`Year` (int)
`Hazard` (str)	N/A	`Hazard` (str)
N/A	N/A	`Source` (str)

Table 4.3: Harmonization of the datasets' columns

Cleaning

Three cleaning steps were performed either to fix some errors or to standardize formats.

Before anything else, a manual correction of erroneous dates in CCLW was carried out directly in the source file. Three rows contained invalid years, `0021`, `0022`, and `1010`. These were updated respectively to `2021`, `2022`, and `2010`.

The first cleaning process was to fix parsing or encoding errors in the Summary column of both datasets. Most rows still contained HTML tags that needed to be removed to have a proper string format.

Then, all labels in the Hazard column were mapped to new categories (Appendix A). To do so, every unique label was retrieved and manually assigned to a new one. This step gathers, under one label, various words that can be considered as one same category. In the end, ten hazard categories were defined:

- `flood`. Laws addressing river and coastal floodings;
- `drought`. Laws addressing water scarcity, desertification, and prolonged precipitation deficits;
- `temperature_extremes`. Laws addressing heatwaves, heat stress, cold waves, and changes in temperature;
- `sea_level_rise`. Laws addressing coastal erosion and sea level change;
- `storm`. Laws addressing storms, windstorms, cyclones, hurricanes, and typhoons;
- `wildfire`. Laws addressing wildfires and forest fires;
- `melting`. Laws addressing glacial melting, loss of snow cover, and permafrost thawing.
- `erosion`. Laws addressing land and soil erosion.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- `other`. Laws addressing a climate change concern that does not fit into any of the previous categories.
- `none`. Laws addressing no climate hazard.

Finally, the date format of Year was standardized to be uniform across both datasets. Only the year the document was passed was kept, as an integer.

Filtering

For CCLW, two filtering steps were performed. The documents with ISO codes `XAA` and `XAB`, and those with less than 25 words in `Summary` were removed. Starting with 12,478 documents, 7,626 remained after filtering. For the rows with a non empty `Hazard` column, they went down from 538 to 529.

For CPDB, one filtering step was performed. The documents with empty `Summary` were removed. Starting with 6,727 documents, 4,956 remained after filtering.

Merging

Within each dataset, internal duplicates were first identified by matching Title and Summary. This reduced CCLW from 7,626 to 5,802 documents and CPDB from 4,956 to 4,803 documents. As a consequence, the number of documents with a non-empty Hazard label dropped from 529 to 356.

Then, duplicates needed to be identified across both datasets. To do so, they were identified and gathered in pairs by identical Year and Country. Within these candidate pairs, titles were compared after two cleaning procedures. Anything in parentheses was excluded, and words shorter than four characters were removed. This threshold is based on manual observation of the dataset, where titles referring to the same document often differed only by short connecting words such as *the* or *of*, or by the presence of an act number (for example „Tax“ and „Tax Act 15“ or „Decree“ and „Decree No. 449“). A stop-word list would not address these cases, since it would leave the differing act or decree numbers. Removing words shorter than four characters instead addresses both cases at the same time. This approach automatically removed 848 documents. Remaining pairs, with at least four words of at least four characters in common, were manually reviewed. Out of 287 pairs, 230 were confirmed as duplicates and removed. In each case where duplicates were found, the version of CCLW was kept.

One last issue needed to be resolved, as 10 rows of the final set had an empty Year column. These were manually added by verifying the information from the source.

The final merged dataset contains 9,527 unique documents and was saved as `law.csv`. The final benchmark, consisting only of documents with hazard labels, contains 356 documents and was saved as `benchmark.csv`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.2 Sensor Data

This section introduces the sensor data, consisting of Earth observation variables. It first introduces the six datasets selected to represent the relevant environmental variables. It then examines the data preparation pipeline applied to convert raw gridded observations into a country-level format, followed by the resulting features, statistical properties, missing values, and ethical concerns.

4.2.1 Source

For Earth observation variables, the Climate Data Store [Cop] is chosen. From now on, the Climate Data Store will be referred to as CDS.

The CDS supports the implementation of the Copernicus Climate Change Service (C3S), one of six thematic information services provided by the Copernicus Earth Observation Programme of the European Union (EU), and implemented by the European Centre for Medium-Range Weather Forecasts (ECMWF). The CDS makes C3S data observations, reanalyses, and projections available through a web interface and an Application Programming Interface (API). The database's update frequency varies from one dataset to the other.

C3S' mission is to support adaptation and mitigation policies by providing consistent and authoritative information about climate change, offering free and open access to climate data and tools based on the best available science.

The ERA5-Land monthly averaged data from 1950 to present [CC22] dataset is used to acquire the variable `2m temperature`. ERA5-Land is a reanalysis dataset providing a view of the evolution of land variables from 1950 onward. Though the original temporal resolution is of one hour, monthly-mean averages are calculated and directly available for download. The data is gridded globally, to a regular latitude-longitude grid of 0.1°x0.1°. It is updated monthly, with a delay of two to three months, and the version dated from May 6th 2026 was downloaded for this thesis.

The ERA5 monthly averaged data on single levels from 1940 to present [CC23] dataset is used to acquire the variables `total precipitation`, `instantaneous wind gust`, and `snowmelt`. ERA5 is the fifth generation ECMWF reanalysis for the global climate and weather, with data available from 1940 onward. ERA5 provides hourly estimates for a large number of atmospheric, ocean-wave and land-surface quantities. Monthly-mean averages are calculated and directly available for download. The data is gridded globally, to a latitude-longitude grid of 0.25°x0.25°. It is updated monthly and the version dated from May 6th 2026 was downloaded for this thesis.

The Sea level gridded data from satellite observations for the global ocean from 1993 to present [CC18] dataset is used to acquire the variable `sea level anomaly`. This dataset provides gridded daily and monthly mean global estimates of sea level anomaly based on satellite altimetry observations provided by the Copernicus Marine Environment

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Monitoring Service. The data is gridded across oceans globally, to a latitude-longitude grid of 0.25°x0.25°, expecting some spatial gaps on the land. It is updated about three times a year, with a delay of about five months, and the version dated from June 1st 2026 was downloaded for this thesis.

The Monthly drought indices from 1940 to present derived from ERA5 reanalysis [Eur25] dataset is used to acquire the variable Standardized Precipitation Evapotranspiration Index (`SPEI`). ERA5–Drought is a global reconstruction of drought indices from 1940 to present. The two drought indices present in the dataset are calculated using precipitation and potential evapotranspiration from the fifth generation of the ECMWF atmospheric reanalyses. The data is gridded globally, to a latitude-longitude grid of 0.25°x0.25°. It is updated monthly, with a delay of two to three months and the version dated from May 27th 2026 was downloaded for this thesis.

4.2.2 Data Preparation

This subsection describes the data preparation pipeline applied to the sensor data. It first covers the collection and processing of the raw gridded files into country-level observations, then the harmonization of columns across the datasets. It then details the cleaning steps applied to standardize country codes and dates, and finally describes the merging of all six datasets into a single merged dataset.

Collecting

The interactive interface of the CDS allows to directly make requests and download the corresponding data or use the API request code in Python. By doing so, six ZIP folders were obtained. Contrary to the case of laws and policies, the data here cannot be directly interpreted for an exploratory analysis. Thus, the collection step already includes some preprocessing and it was decided to complete the whole data preparation before diving into data understanding.

The nature of the files inside the ZIP folder varies depending on which dataset it originates from, and required processing pipelines following the same logic overall, but with a few distinct aspects when loading the files.

On the one hand, the variables `2m Temperature`, `Instantaneous Wind Gust`, `Snowmelt`, and `Total Precipitation` are each gathered in their respective file of GRIB format, commonly used in meteorology to store historical and forecast weather data. Each of them requires the use of the `cfgrib` library [ECM25] to decode the format into an xarray, much simpler to use for processing. For the `instantaneous wind gust` dataset, loading the file in the same way as the others raised an error, as the GRIB file bundles two different statistical representations of the variable under the key `stepType`. Since the data used in this thesis consists of monthly averaged values, the file was reopened while explicitly filtering for the `avgas` step type.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

On the other hand, the variables `Sea Level Anomaly` and `SPEI` are stored in NetCDF files, one for each month with available data. As such, it was required to load all files simultaneously, concatenating them along the time dimension.

Though the CDS request was narrowed down to the variable of interest only, a dataset may contain more than one variable because it automatically bundles related meteorological quantities. In that case, the code was implemented so that the user is prompted to choose from them and select the relevant variable.

Then, a pre-defined function was applied to convert the grid into countries. By taking the xarray as an input, it outputs a pandas dataframe with one row per country per month, compatible with the `pandas` library [Pan26]. Using the `regionmask` library [Reg24], each cell of the global grid can be assigned to a country. The resolution of the grid depends on the dataset, yet the same Natural Earth boundaries at 1:110m scale were used arbitrarily from available land masks of the previously cited library. Iterating over each country, the spatial mean of the variable was computed in order to have a single value per country per month.

Finally, the data for each variable was stored and saved in a CSV file.

Harmonizing

To simplify the work of cleaning and filtering every dataset at the same time, the columns kept and their names were harmonized following Table 4.4. A Year column was added later on to match with legislation data.

.grib column (type)	**.nc column** (type)	**Merged column** (type)	**Description**
`iso` (str)	`iso` (str)	`Country` (str)	ISO 3166-1 alpha-3 code
`country` (str)	`country` (str)	`Country_name` (str)	Name of the country
N/A	N/A	`Year` (int)	Year of observation
`valid_time` (str)	`time` (str)	`Date` (str)	Year and month of observation
`value` (float)	`value` (float)	`[variable name]` (float)	Variable observed

Table 4.4: Harmonization of the dataset's columns

Cleaning

Two cleaning steps were performed to standardize formats.

First, it was noticed that, contrary to legislation data where the ISO codes are given in ISO 3166-1 alpha-3 codes, in the sensor data they are mostly given in ISO 3166-1 alpha-2 code. There are also isolated issues: countries already in ISO 3166-1 alpha-3 code, countries identified by one letter, countries identified by four letters, or countries identified by three letters but not following the ISO 3166-1 alpha-3 code. Additionally, the letter 'J' is used to refer to three countries at the same time, Jamaica, Jordan, and Japan. These were manually mapped and fixed directly. Subsequently, the rest of the countries have had their alpha-2 codes changed into alpha-3 codes.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Then, the dates were cleaned to only keep the year and the month, the day was then arbitrarily set to the 1st without any consequences for the rest of the work.

Merging

The final merged dataset contains 183,549 rows, and was saved as `sensor.csv`.

4.2.3 Features

In total, our sensor dataset contains 183,549 rows described by 9 features, described in Table 4.5 and each variable detailed hereinafter.

Column (type)	**Description**
`Country` (str)	ISO 3166-1 alpha-3 code of the country
`Country_name` (str)	Name of the country
`Year` (int)	Year of the observation
`2m_temperature`	In Kelvin
`Instantaneous_wind_gust`	In m/s
`Sea_level_anomaly`	In meters
`Snowmelt`	In m of water equivalent)
`SPEI`	Without unit
`Total_precipitation`	In meters

Table 4.5: Features of the sensor dataset

2m Temperature

The temperature of air at two meters above the surface of land, sea or in-land waters, measured in Kelvin [CC22]. It is calculated by interpolating between the lowest model level and the Earth's surface, taking account of the atmospheric conditions. The dataset contains 162,132 rows.

Total precipitation

The accumulated liquid and frozen water, including rain and snow, that falls to the Earth's surface, expressed in meters [CC23]. It is measured as the depth the water would have if it were spread evenly over a grid box. This does include fog, dew or the precipitation that evaporates in the atmosphere before it lands at the surface of the Earth. The dataset contains 183,549 rows.

Instantaneous Wind Gust

The maximum wind gust at the specified time, at a height of ten meters above the surface of the Earth, measured in m/s [CC23]. A wind gust is defined as the maximum of the wind averaged over 3 second intervals. The dataset contains 183,549 rows.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Snowmelt

The accumulated amount of water that has melted from snow in the snow-covered area of a grid box, expressed in meters of water equivalent [CC23]. It is measured as the depth of water there would be if the melted snow (from the snow-covered area of a grid box) were spread evenly over the whole grid box. The dataset contains 183,372 rows.

Sea Level Anomaly

The sea surface height above mean sea surface computed with respect to a 20-year mean reference period (1993-2012), expressed in meters [CC18]. A positive value indicates the sea surface is above the historical mean, while a negative value indicates it is below. The dataset contains 65,844 rows.

Standardized Precipitation Evapotranspiration Index

The deviation from the reference climate, expressed in standard deviations [Eur25]. The Standardized Precipitation Index (SPI) measures the precipitation deficit that accumulated over the preceding months and evaluates the deficit with respect to a reference period. The Standardized Precipitation Evapotranspiration Index (SPEI) is an extension of the SPI and incorporates potential evapotranspiration to capture the impact of temperature on drought. A positive value indicates wet conditions, while a negative value indicates dry conditions with respect to the reference climate. The dataset contains 183,018 rows.

4.2.4 Statistical Properties

This subsection examines the statistical properties of the merged sensor dataset. It first looks at the geographic and temporal coverage of the data, before investigating the distribution and trend of each environmental variable over the observation period.

Geographic Coverage

The same 177 countries are found in every dataset. The geographic distribution of the data is perfectly uniform across all datasets, meaning that each country is represented by an identical number of rows.

Figure 4.4 shows the geographic overlap between the legislation and sensor datasets. There is a total of 169 countries overlapping in both datasets.

Out of the five states only present in sensor data, four are territories depending on legislation of other countries: France for French Southern Territories and New Caledonia, the United Kingdom for Falkland Islands, and the United States of America for Puerto Rico. The other is Antarctica, an uninhabited piece of land.

Out of the 34 countries only present in legislative data, most of them are micro-states like Andorra, Malta, and Singapore, enclaves like Hong Kong, Monaco, and San Marino, or

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

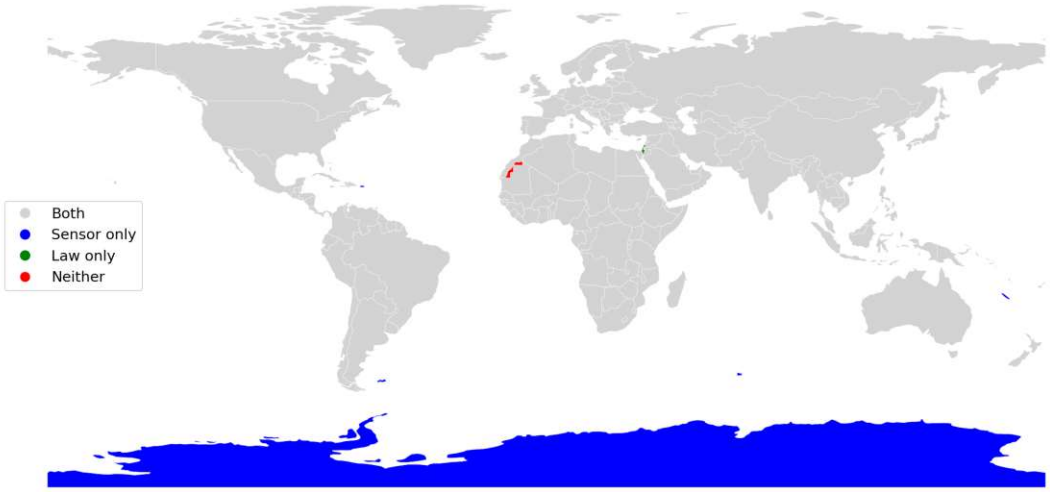


Figure 4.4: Geographic alignment of legislative and sensor data

small island states like Barbados, Comoros, and Dominica. Their land areas are smaller than one cell of a grid and fall below the resolution of the Natural Earth boundaries, so no cell can be assigned to them.

Temporal Coverage

The ERA5-Land variables contain observations ranging from 1950 to 2026, while the ERA5 variables and the drought indices extend further back to 1940. The sea level anomaly dataset has the shortest coverage, from 1993 to 2023, constrained by the availability of satellite altimetry data.

All datasets but sea level anomaly fully cover the 2000-2025 window used in the analysis, the data missing between 2023-2025 only leads to a shorter time period. The extended coverage back to 1940 and 1950 provides a baseline period for computing variables' deviation, as values prior to 2000 are used to establish a historical baseline.

Variables

Different types of visualizations help in understanding the distribution of each variable. Table 4.6 reports descriptive statistics for each variable of the dataset, while figure 4.5 shows the distribution of each variable across all country-month observations, and figure 4.6 displays the global mean of each variable averaged across all countries per year. The spikes or drops visible at the end of some series are due to incomplete yearly coverage in the most recent years of the observation period.

First, `2m temperature` ranges from 221.8K to 312.9K (approximately -51°C to 40°C), explained by the dataset covering polar and tropical regions. The trend observed is an

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Variable (unit)	**Mean**	**SD**	**Min**	**Median**	**Max**
2m Temperature (K)	290.2988	11.2522	221.7932	294.3349	312.9114
Instantaneous Wind Gust (m/s)	6.1272	1.5256	2.9526	5.9969	18.3736
Sea Level Anomaly (m)	0.0361	0.1206	-1.4647	0.0430	0.5918
Snowmelt (mm)	0.1651	0.5316	0.0000	0.0000	15.4151
SPEI (N/A)	0.0436	0.9287	-8.0707	0.0323	7.6594
Total Precipitation (mm)	3.1282	3.2504	0.0000	2.1276	42.1091

Table 4.6: Descriptive dtatistics of environmental target variables

increase of the mean value over time, reflecting global warming.

Second, `instantaneous wind gust` follows a normal distribution centered around 6 m/s, with a long tail in higher values corresponding to windy regions, more exposed to storms. Additionally, like temperature, the trend shows an increase over the observation period, consistent with intensification of extreme weather.

Third, the positive `sea level anomaly` mean and median indicate a slight positive bias across the observation period, consistent with a sea level rise trend since the reference baseline. It shows the clearest trend, rising monotonically, directly reflecting accelerating global sea level rise.

Then, `snowmelt` shows a median of 0mm, which can be expected from a dataset covering the entire globe, including tropical or temperate regions with no snow.

Next, `SPEI` is symmetric and centered around zero, ranging -8.1 to 7.6 standard deviations these extreme values reflect important drought and wet events.

Finally, `total precipitation`'s distribution is strongly skewed towards low values, with most observations near zero and a long tail corresponding to high-rainfall regions.

4.2.5 Missing Values

A few columns have a significant number of missing values. The most critical in terms of number is the `Sea_level_anomaly` columns with a 73% missing values rate. This can be explained by the fact that the original database only covers coastal countries, and all landlocked countries have empty values. One other variable, `2m_temperature` (12% missing values rate), is lacking values. This may be due to some regions being out of the coverage of the observations derived from the ERA5-Land. Additionally, `Snowmelt` and `SPEI` respectively have 177 and 531 missing values, which amounts for less than 1% of the dataset and does not impact the global coverage.

4.2.6 Ethical Concerns and Biases

The sensor datasets introduce a geographic bias, where small island states are not represented though among the most vulnerable countries, by being hidden behind the resolution of the mask because of their size. This mirrors the under-representation of the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

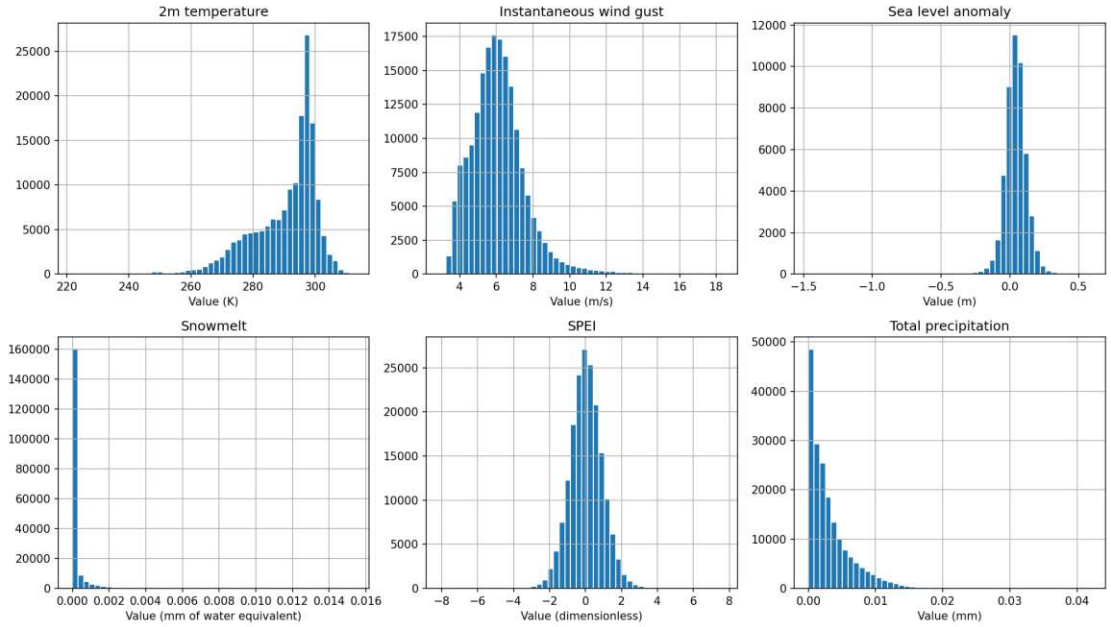


Figure 4.5: Distribution of each climate variable across all country-month observations

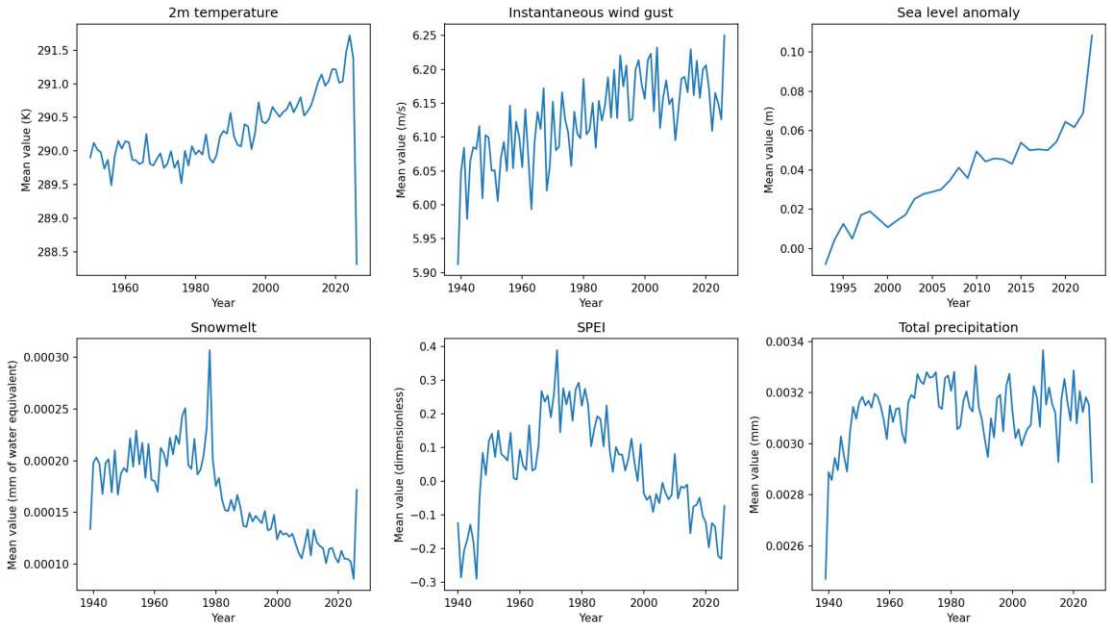


Figure 4.6: Global yearly mean of each climate variable averaged across all countries

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

legislation data as well, leading to these countries being made invisible in the correlation analysis.

Additionally, some systematic bias may be introduced due to the provenance of the data. The ERA5 reanalysis datasets are model estimates and not observations. Even though they are highly reliable and widely used in research, some regions of the world might have less data to input in those models.

4.3 Linking Legislation Data and Sensor Data

Each hazard label from legislation data, established in Section 4.1.6, was associated with a variable from sensor data, described in Section 4.2.3. Table 4.7 matches the hazard labels with the variables chosen to represent them as measurable physical parameters.

Hazard	**Variable** (unit)
`temperature_extremes`	2m Temperature (K)
`storm`	Instantaneous Wind Gust (m/s)
`sea_level_rise`	Sea Level Anomaly (m)
`melting`	Snowmelt (m of water equivalent)
`drought`	Standardized Precipitation Evapotranspiration Index (N/A)
`flood`	Total Precipitation (m)

Table 4.7: Hazard labels and corresponding Earth observation variables

Initially, ten distinct hazard labels were defined. It comes naturally that no physical measure can be associated with `none` and `other`. Additionally, no variable was found to be consistent with `wildfire` and `erosion` on a global scale, and both are thus discarded from the analysis.

For the six remaining hazard categories, multiple relevant variables could have been chosen. However, the scope of this thesis requires selecting only one variable per hazard to maintain consistency and enable cross-hazard comparison. The variables selected represent the most direct measures available on the Climate Data Store (CDS) with consistent global coverage across countries and years.

First, `temperature_extremes` is associated with 2m temperature, the standard near-surface air temperature measure used in climate monitoring. According to the Global Climate Observing System, surface air temperature is a key indicator of climate change, contributing to the global surface temperature record [Glo24]. Rising temperature anomalies are the direct physical driver of heat waves and extreme cold periods, making this variable the most natural representation of extreme temperatures available on the CDS.

Second, `storm` is associated with instantaneous wind gust, which captures the maximum wind intensity at a given time and location. Wind speed measurements are used to identify and track dangerous weather events such as hurricanes and tornadoes [NAS24c],

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

making wind gust, as the measure of peak wind intensity, the most direct indicator available on the CDS to measure storm severity.

Third, `sea_level_rise` is associated with sea level anomaly, which directly measures the deviation of the sea surface height from a historical mean, making it, among the variables available on the CDS, immediately and easily interpretable.

Then, `melting` is associated with snowmelt, which measures the rate of snow and ice melting per unit area. Changes in glacier mass balance, driven by changes in the size and mass of the world's glaciers, have long served as an indicator of climate fluctuation [NAS25], making snowmelt, among the variables available on the CDS, the primary physical manifestation of the melting hazard in the context of climate change.

Next, `drought` is associated with the Standardized Precipitation Evapotranspiration Index (SPEI), which measures the balance between precipitation and evapotranspiration. Unlike the Standardized Precipitation Index (SPI), a commonly used operational drought indicator based on precipitation alone [NAS24a], SPEI accounts for the effect of temperature on water demand, making it the most comprehensive indicator of drought conditions available on the CDS.

Finally, `flood` is associated with total precipitation, as accumulated rainfall is the primary driver of river flooding and surface water flooding. Precipitation is one of the key remote sensing observations used, alongside soil moisture, digital elevation, and land cover, to support flood monitoring and prediction when combined with hydrologic models [NAS24b]. Among the variables available on the CDS, total precipitation is chosen as the most direct proxy for flood risk.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 5

Experiment Setup

This chapter presents the experimental setup used to address the three research questions of this thesis. It first introduces the models developed for each modeling phase and then examines the evaluation metrics used to assess the performance of the classification models and the quality of the visualization dashboard.

5.1 Models

This section presents the models developed for each of the three research questions. It first examines the classification approaches for identifying hazard categories addressed in legislative documents. It then describes the correlation approaches used to relate legislative coverage to hazard intensity. Finally, it introduces the visualization approaches designed to communicate the resulting alignment gaps.

5.1.1 Classification Approaches

The objective of the first modeling phase was to develop and compare approaches for multi-label classification of climate laws and policies into hazard categories. After preprocessing (Section 4.1.6), only 3.7% of documents in the resulting dataset were annotated with one or more labels on the hazards they are addressing. Through classification, a document may be assigned any combination of hazard labels, including none.

To assign hazard categories to all documents, three automated classification pipelines were developed and evaluated against the benchmark of 356 labeled documents. A majority-class baseline, a rule-based NLP classifier, and zero-shot LLM-based classifiers were chosen. All three approaches function without training, because the size of the labeled benchmark is insufficient for reliable fine-tuning across all hazard categories.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Baseline Classifier

The baseline classifier assigns the most frequent hazard label(s) in the benchmark dataset to every document, regardless of the content of their summary. Any approach that does not outperform the baseline is found to have failed to extract any meaningful signal from the text.

Two versions were implemented to account for the multi-label nature of the task. The first baseline assigns only the most frequent label to every document, while the second one assigns the three most frequent labels. They will be respectively referred to as baseline top-1 and baseline top-3 classifiers. In the benchmark dataset, the most frequent label is `flood`, followed by `drought` and `other`. These baselines will both serve as reference points against which other approaches are compared.

Rule-based NLP Classifier

The rule-based classifier assigns hazard labels based on the presence of predefined keywords in the document summary. For each hazard category, a glossary of related keywords was manually established based on the original set of labels, synonyms, lexical field, and domain knowledge. A document is assigned a label if at least one keyword from the corresponding glossary appears in its summary. This approach's main limitation is sensitivity to vocabulary choices and inability to capture related terms not explicitly listed in the glossary.

Zero-shot LLM Classifier

The zero-shot classifiers use pre-trained large language models to assign hazard labels without any training. Eight models were chosen from the Ollama API using the Python library [Oll25]. The models were prompted with the list of hazard categories and the document summary, and asked to assign all applicable labels with the following prompt:

> Classify the following climate law or policy summary into one or more hazard categories.
>
> - Only use labels from this list: ['flood', 'drought', temperature_extremes', 'sea_level_rise', 'storm', 'wildfire', 'melting', 'erosion', 'other', 'none'].
> - Use 'none' only if the document has no relation to any physical climate hazard.
> - Use 'other' if the document addresses climate hazards not covered by the other categories.
> - Assign all applicable labels.
>
> Summary: {summary}

Released in August 2025, gpt-oss-120b and gpt-oss-20b are OpenAI's state-of-the-art open-weight language models trained a mix of reinforcement learning and techniques informed by OpenAI's most advanced internal models, with a focus on reasoning, efficiency, and real-world usability [Ope25]. Both are freely available for download, making AI accessible

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

with models matching the performances of OpenAI o4-mini for gpt-oss-120b and of OpenAI o3-mini for gpt-oss-20b. Both use a Mixture of Experts architecture with a 128k-token context window.

Released in March 2025, Gemma 3 4B and Gemma 3 27B are multi-modal open-weight models from a family of open models from Google, built from the same research and technology used to create the Gemini models [Dee25]. Both support a 128k-token context window and are capable of processing and analyzing text, images, and video. Gemma 3 4B offers a balance between performance and flexibility, while Gemma 3 27B is enhanced for understanding and targets more sophisticated applications.

Released in October 2025, MiniMax M2 is a compact and cost-effective open-source model from MiniMax, built for coding and agentic tasks [Min25]. Released in February 2026, MiniMax M2.5 is MiniMax's state-of-the-art open-source model, improving performance and speed for coding and complicated agentic tasks [Min26]. Both share a 205K-token context window.

Released in July 2025, Qwen3-Coder-480B is a long-context model from Alibaba, for agentic and coding tasks [Qwe25a]. Released in September 2025, Qwen3-Next-80B is the first installment of the Qwen3-Next series from Alibaba, with strong performance in terms of parameter efficiency and inference speed [Qwe25b]. Both models support a 256K-token context window.

5.1.2 Correlation Approaches

The objective of the second modeling phase was to conduct a correlation analysis between legislative coverage and hazard intensity.

Variables Definition

Legislative coverage was defined as the cumulative count of laws and policies per country per hazard category per year, from 2000 to 2025. The legislative coverage LC of hazard h for country c at year y is given by:

$$LC_{c,y}(h) = LC_{c,y-1}(h) + N_{c,y}(h) \tag{5.1}$$

with $N_{c,y}(h)$ the count of laws passed for country c at year y and $LC_{c,1999}(h) = 0$ as the initialization condition.

In the case that a country had no law classified under a given hazard category, the count was set to zero until the first year when a law is found to have been passed. If necessary, the count stays null for all years. In the case that a country's last passed law predated 2025, the count was kept constant at its final value for the remaining years. This ensured a complete time series for every country and hazard combination, reflecting the fact that existing legislation remains in force from the moment they are passed.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Hazard intensity was defined as the deviation of the climate variable associated with each hazard category, relative to a historical baseline period. The intensity I of hazard h for country c at year y is given by:

$$I_{c,y}(h) = \overline{h}_{c,y} - \mu_c \tag{5.2}$$

where $\overline{h}_{c,y}$ is the mean value for country c at year y and μ_c is the mean of the variable for country c over the baseline period. For this computation, the reference period starts in 1950 and ends in 1999.

Two variables did not require this step and were kept in their original form. SPEI is expressed in standard deviations and already uses a reference period (1991-2020), similarly sea level anomaly is expressed as a deviation from a baseline period (1993-2012). Both were directly interpretable as anomaly measures and were used as-is in the correlation analysis.

Spearman Correlation

The relationship between legislative coverage and hazard intensity was evaluated using Spearman's rank correlation coefficient ρ. Spearman correlation was chosen over Pearson correlation, more fitting for non-linear relationships and robust to outliers. Its value directly allows to understand the relationship between the two metrics, and observe how legislative coverage evolves with hazard intensity.

For each pair of country and hazard, Spearman's ρ was computed between the annual cumulative law count and the annual hazard intensity over the period 2000-2025. The correlation between legislative coverage $LC_{c,y}(h)$ and hazard intensity $I_{c,y}(h)$ for country c, hazard h, at year y is given by:

$$\rho = \text{Spearman}[LC_{c,y}(h),\ I_{c,y}(h)] \tag{5.3}$$

Statistical significance can be tested with the p-value and a threshold of $p < 0.05$.

Legislation Lag

The enforcement of laws and policies does not necessarily happen in the same year as events that might trigger them. Drafting and passing legislation takes time. To account for this, a lag analysis was conducted by shifting the legislative coverage time series backwards by k years before computing correlations, testing whether legislative coverage at year y is better explained by hazard intensity at year $y - k$ than at year y. The lagged correlation between legislative coverage $LC_{c,y-k}(h)$ and hazard intensity $I_{c,y}(h)$ for country c, hazard h, and lag k is given by:

$$\rho_k = \text{Spearman}[LC_{c,y-k}(h),\ I_{c,y}(h)] \tag{5.4}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Lags of one to five years were tested. Hazard intensity was computed over an extended period preceding the analysis window, ensuring that legislative coverage observed as early as 2000 can still be paired with hazard intensity values. The lag producing the highest proportion of statistically significant correlations across all country-hazard pairs was selected as the reference lag for subsequent analysis.

Hierarchical Trend Analysis

To examine whether regional dynamics influence the relationship between legislative coverage and hazard intensity, the analysis was conducted at two levels of geographic aggregation: country-level and continental-level. At the country-level, Spearman correlation was computed independently for each country-hazard pair. At the continental-level, legislative coverage and hazard intensity were first computed by continent and correlations are then calculated on these aggregated time series.

This hierarchical structure allowed comparison of the correlation at different scales. Because a country's risk exposure is not only affected by its own legislation but also by that of its neighbors, regional aggregation may reveal other patterns. As detailed in Figure 5.1, each country was mapped to one of the six continents: Africa, Asia, Europe, North America, Oceania, and South America.

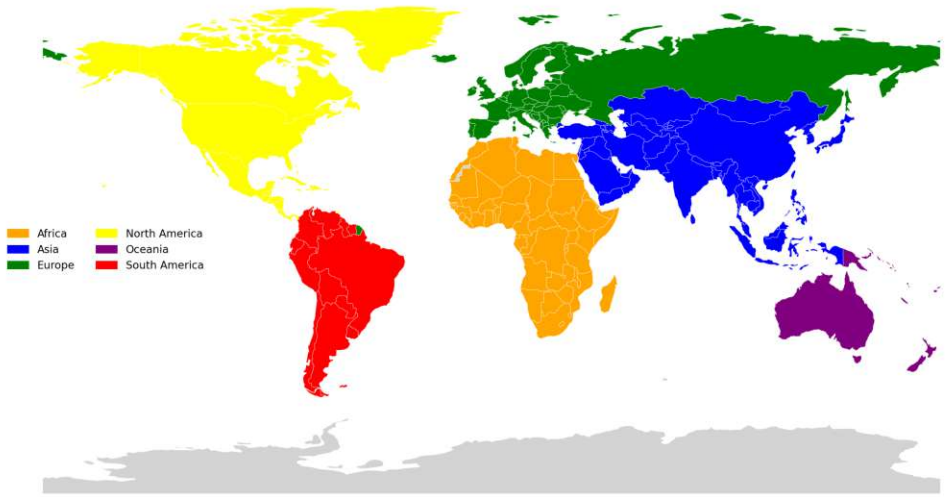


Figure 5.1: Country-to-continent mapping used for continental aggregation

5.1.3 Visualization Approaches

The objective of the third modeling phase was to design and implement a visualization dashboard to communicate the gaps between legislative coverage and hazard intensity.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The visualization dashboard was developed as an interactive web application using Plotly [Plo26], a Python visualization library that supports interactive charts, maps, and multi-panel dashboards. This choice was motivated by the consistency of Python's usage throughout this entire thesis, the flexibility to implement custom designs, and the fact that the dashboard runs as a local web application that can be opened in any browser without requiring software installation. Screenshots of the dashboard are provided in Appendix C.

The aim of the dashboard was to communicate national gaps between climate hazard risk intensity and legislative coverage. To this end, the said gap was defined as the difference between the computed values of coverage and intensity, both normalized to range from 0 to 1, to ensure the comparability of metrics using different units and scales. The gap for country c, hazard h, and year y is given by:

$$\text{gap}_{c,h,y} = NLC_{c,h,y} - NI_{c,h,y} \tag{5.5}$$

where $NLC_{c,h,y}$ and $NI_{c,h,y}$ are the normalized legislative coverage and hazard intensity. A positive gap indicates that a country's legislative coverage exceeds observed intensity, while a negative gap indicates that observed hazard intensity exceeds its legislative coverage. This definition comes with the thought that negative numbers communicate a negative feeling.

Five complementary visualization types were implemented, each designed to communicate a different dimension of the alignment gap between climate hazard intensity and legislative coverage. Hazard and country selection can be made through a slicer drop-down shared across the figures, so that choosing a hazard or a country updates every view that depends on it simultaneously. Year selection, where applicable, is provided through Plotly's built-in timeline slider.

World Map

A color-filled map was chosen to display the gap per country per hazard category per year. A color scale was used to differentiate the scenarios, from red tones indicating countries where hazard intensity exceeds legislative coverage to green tones indicating countries where legislative coverage exceeds hazard intensity. Greener states are considered more proactive as they are properly covering hazards, while redder ones are at risk with very low legislative protection. The shared hazard drop-down allows the user to switch between the hazard categories and the year slider provides the possibility to select the year. This enables an immediate observation of countries' differences.

Timeline

A timeline was adopted to exhibit the evolution of both legislative coverage and hazard intensity over the 2000-2025 period. The shared hazard and country drop-downs control which pair is displayed. The two variables are plotted on separate y-axes to account for

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

their different units and scales. This view allow the user to observe how legislative trend matches the hazard intensity, and whether it accelerates in response to the risks.

Another version of the timeline displays the gap, using normalized values of legislative coverage and hazard intensity. This facilitates the visualization of the gap, whether it is widening or narrowing over the period.

Bar Chart

A comparative bar chart was selected to simultaneously compare the gap for all hazard categories in a country selected by the user for a year selected through the slider. Bars were colored by the direction of the gap, red for positive and green for negative. For a given country and year, this view allow to quickly identify which hazards are least aligned with observed risk and require more legislative efforts.

Scatter Plot

A scatter plot was used to have a global view of each country for a chosen hazard and year, in a two-dimensional space defined by normalized hazard intensity and normalized legislative coverage. One country is represented as a point. Countries above the diagonal are over-legislated relative to their risk level, while countries below are under-legislated. Points are colored by continent, adding the hierarchical geographic dimension of the analysis and allowing visual identification of regional clustering patterns.

5.2 Evaluation Metrics

This section presents the metrics used to evaluate the outputs of the two modeling phases requiring quantitative assessment. It first examines the classification metrics used to compare the performance of the models developed for RQ1. It then introduces the visualization metrics used to assess the dashboard developed for RQ3.

5.2.1 Classification Metrics

There exists a multitude of evaluation metrics for classification tasks. For multi-label classification, accuracy is not an appropriate evaluation metric. A prediction is only considered correct if the entire list of predicted labels exactly matches the true set, which is too strict in the scenario where even one single correct label is meaningful. Instead, metrics that account for partial matches across multiple labels are required.

Primary Metrics

Three metrics were chosen and computed for each of the ten hazard categories and then aggregated in four different ways using the `classification_report` function of the `sklearn.metrics` library [Sl25].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

First, precision, ranging from 0 to 1, is the ratio of true positives (TP) among all samples labeled positively, meaning all true positives and false positives (FP):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.6}$$

It reflects the ability of a classifier not to label as positive a sample that is negative.

Second, recall, also ranging from 0 to 1, is the ratio of true positives among all samples that are positive, meaning all true positives and false negatives (FN):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.7}$$

It reflects the ability of the classifier to find all the positive samples.

Third, F1-score is calculated as the harmonic mean of precision and recall:

$$\text{F1-score} = 2 * \frac{precision * recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{5.8}$$

It helps in evaluating a model by balancing both precision and recall.

Aggregation Methods

Four averaging strategies exist for multi-label classification, in order to obtain a general score of the model by computing the values for each hazard category in various ways.

First, the macro-averaged value is calculated using the arithmetic mean of the corresponding metric over all the classes, for example:

$$\text{Macro-averaged precision} = \frac{1}{|classes|} \sum_{c \in classes} precision_c \tag{5.9}$$

It weights each label equally regardless of frequency, making it sensitive to performance on less frequent categories.

Second, the micro-averaged value is computed using the formula of the corresponding metric over the whole dataset, without accounting for the class, for example:

$$\text{Micro-averaged precision} = \frac{\sum_{c \in classes} TP_c}{\sum_{c \in classes} TP_c + \sum_{c \in classes} FP_c} \tag{5.10}$$

It aggregates contributions from all labels proportionally to their frequency.

Third, the weighted-averaged value is calculated taking the mean of the corresponding metric over all the classes, weighted by each class's support, for example:

$$\text{Weighted-averaged precision} = \frac{\sum_{c \in classes} precision_c * support_c}{\sum_{c \in classes} support_c} \tag{5.11}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The support is the count of all occurrences of the class in the dataset. It accounts for label imbalance by weighting each label by its support.

Finally, the sample-averaged value is computed by calculating the corresponding metric for each row and averaging them over the whole dataset, for example:

$$\text{Sample-averaged precision} = \frac{1}{|rows|} \sum_{r \in rows} precision_r \tag{5.12}$$

Macro F1 was used as the primary comparison metric, as it treats all hazard categories equally and is most informative in case of class imbalance.

Additional Metrics

Two additional metrics specific to multi-label classification complemented the analysis, respectively using the `hamming` and `jaccard_score` functions of the `sklearn.metrics` library [Sl25].

Hamming loss, ranging from 0 to 1, measures the fraction of incorrectly predicted labels to the total number of labels, across all rows and classes:

$$\text{Hamming loss} = \frac{1}{|rows| \cdot |classes|} \sum_{r \in rows} \sum_{c \in classes} \text{xor}[y(r,c), y_{pred}(r,c)] \tag{5.13}$$

where xor is the operator that returns zero when the target and prediction are identical and one otherwise.

Jaccard score, ranging from 0 to 1, is the ratio of correctly predicted labels to the union of predicted and true labels per document, averaged across all documents:

$$\text{Jaccard score} = \frac{|y \cap y_pred|}{|y \cup y_pred|} = \frac{|y \cap y_pred|}{|y| + |y_pred| - |y \cap y_pred|} \tag{5.14}$$

It measures the similarity between the predicted labels and the true labels for each document.

To be able to compute these metrics, the predicted and true label sets were first encoded into binary vectors, transforming each document's label list into a fixed-length binary vector with one position per hazard category. For example, the label set ['drought', 'temperature_extremes', 'wildfire', 'other'] became the array (0,1,1,0,0,1,0,0,1,0), following the ordering of `HAZARDS` = ['flood', 'drought', 'temperature_extremes', 'sea_level_rise', 'storm', 'wildfire', 'melting', 'erosion', 'other', 'none']. This encoding enabled the use of the standard classification metrics at the label level.

5.2.2 Visualization Metrics

The dashboard was evaluated against three design criteria: completeness, correctness, and clarity. These criteria are adapted from the seven Cs of effective communication,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

formulated by Scott Cutlip and Allen Center [Cut52] and widely used in communication studies, from which the three criteria most relevant to the communication of visual data are selected. Each criterion was defined as a set of conditions that can be assessed following a heuristic evaluation approach and through a user study.

Preliminary User Study

A preliminary user study was conducted to assess the clarity criterion, and provides a secondary check for completeness. The study relied on a sample of three participants, recruited from this author's personal network. All three participants were in their early twenties, enrolled in scientific engineering programs (logistics, data science, and architecture, respectively), and had no prior knowledge of the topic of this thesis. While this sample is not representative of the dashboard's intended user base, it provides an initial indication of the dashboard's usability by individuals unfamiliar with both the underlying data and the research context.

Each session took the form of a 30-minute screen-shared video call. The dashboard was run locally and shared via a temporary public URL using ngrok [ngr26], a tool that exposes a locally running application to the internet without requiring deployment, allowing participants to interact with the artifact remotely.

Each session opened with a brief contextualization of the thesis, and explanation of legislative coverage and hazard intensity. Participants were then asked to complete five tasks for which completion, time taken, and any spontaneous comments were recorded:

1. Using the map, find the policy gap for sea level rise in the Netherlands in 2022;
2. Using the first timeline, determine whether Canada's legislative coverage for flood is increasing or decreasing over time;
3. Using the second timeline, determine how the policy gap for drought in Brazil is evolving;
4. Using the bar chart, identify which hazard has the largest policy gap for for Japan in 2010;
5. Using the scatter plot, identify one country-year pair where coverage is higher than temperature extremes' intensity would suggest is necessary.

Following the tasks, participants answered open-ended questions targeting completeness, correctness, and clarity:

- Was there any information you expected to find but couldn't locate?
- Did you ever feel like you needed to look somewhere outside the dashboard to answer a question?
- Was the meaning of 'gap' clear?
- Which chart was easiest to read? Which was hardest? Why?
- Was the color scale intuitive, or did you have to check the legend repeatedly?
- Did the first timeline make sense at a glance, or did you need a second look?

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Finally, participants were asked to rate three statements on a scale from one (strongly disagree) to five (strongly agree):

- Statement 1: The dashboard contained all the information I needed to answer the tasks;
- Statement 2: I trusted that the values shown were accurate and clearly defined;
- Statement 3: The figures were easy to interpret without extra explanation.

These questions and ratings allowed to understand how the dashboard was perceived by an external viewer, and the feedback helped in identifying areas of improvement.

Completeness

Completeness is defined as the condition that all hazard categories and all countries with available data are represented in the dashboard.

It was evaluated by verifying that all six hazards were present and all countries were illustrated in each visualization type (map, timeline, bar chart, and scatter plot). Countries absent from the satellite dataset and those with certain missing values, like landlocked countries for sea level, are shown in grey. As a secondary check, participants were asked whether they encountered any information they expected to find but could not locate, and whether they ever felt the need to consult a source outside the dashboard.

Correctness

Correctness is defined as the condition that the values displayed in the dashboard accurately reflect the underlying computed data.

It was evaluated by spot-checking a sample of ten country-hazard combinations. Given the size of the dataset, ten combinations were selected as a reasonable sample size that could be manually checked while still covering all hazard categories, all continents, varied years, and a range of gap magnitudes, ensuring the sample is representative despite its size. For each selected combination, the value displayed in the dashboard was compared against the corresponding value in the source file. The dashboard was considered correct if all ten spot-checked values match. Additionally, the direction of the gap (positive or negative) was verified for each case to ensure that the color coding was not inverted.

Clarity

Clarity is defined as the condition that the direction and magnitude of the gap can be read unambiguously by a first-time viewer without domain expertise.

It was assessed through the preliminary user study described previously, specifically through participants' completion of the five tasks and their ratings of Statements 1 to 3. The dashboard was considered to satisfy the clarity criterion if participants were able to complete the five tasks in reasonable time and correctly answer the associated questions without requiring access to the underlying data.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 6

Results

This chapter presents the results obtained for each of the three research questions. It first examines the classification results, comparing the performance of the rule-based and zero-shot approaches used to assign hazard labels. It then presents the correlation results, analyzing the relationship between legislative coverage and hazard intensity at country- and continental-level. Finally, it introduces the visualization results, reporting on a preliminary user study and the evaluation of the dashboard against the completeness, correctness, and clarity criteria.

6.1 Classification Results

This section presents the results of the classification approaches introduced in Section 5.1.1. It first compares the performance of the baseline, rule-based, and zero-shot LLM models across all evaluation metrics, before examining the computational cost and emissions associated with each. It then reports on the effect of prompt engineering on model performance, and concludes with the selection of the model retained for classifying the full dataset.

For efficiency in the evaluation process, all classification approaches were applied to the 356 documents of the benchmark dataset. By comparing each model's predictions against the actual labels, the best-performing model was then used to classify the full dataset of 9,527 documents.

Table 6.1 summarizes the macro-, micro-, weighted-, and sample-averaged precision, recall, and F1-score for all approaches. Table 6.2 compiles the Jaccard score and Hamming loss for each model. To facilitate the observation and analysis of the results, a color grading is used. The classification report function also allows to look at the models' performance more in detail (Appendix B), by having each metric computed per hazard category.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

	Macro			Micro			Weight			Sample		
	pre	rec	F1	pre	rec	F1	pre	rec	F1	pre	rec	F1
Baseline top-1	0.07	0.10	0.09	**0.74**	0.23	0.35	0.17	0.23	0.19	**0.74**	0.27	0.37
Baseline top-3	0.19	0.30	0.23	0.63	0.58	0.60	0.37	0.58	0.45	0.63	0.64	0.58
Rule-based	0.54	0.21	0.28	0.49	0.25	0.33	0.74	0.25	0.35	0.41	0.33	0.32
gpt-oss-20b	0.65	0.22	0.32	0.62	0.28	0.39	**0.80**	0.28	0.40	0.56	0.36	0.39
gpt-oss-120b	**0.69**	0.22	0.33	0.58	0.26	0.36	0.79	0.26	0.39	0.49	0.34	0.36
Gemma 3 4B	0.49	0.47	0.47	0.62	0.61	0.61	0.65	0.61	0.61	0.61	0.64	0.57
Gemma 3 27B	0.47	**0.63**	**0.52**	0.54	**0.74**	**0.63**	0.60	**0.74**	**0.65**	0.57	**0.76**	**0.60**
MiniMax M2.5	0.64	0.18	0.27	0.51	0.22	0.30	0.77	0.22	0.33	0.43	0.29	0.31
MiniMax M2	0.65	0.18	0.28	0.53	0.23	0.32	0.78	0.23	0.35	0.44	0.30	0.32
Qwen3-Coder-480B	0.46	0.61	0.47	0.51	0.64	0.57	0.64	0.64	0.58	0.52	0.64	0.54
Qwen3-Next-80B	0.60	0.33	0.42	0.64	0.37	0.47	0.72	0.37	0.49	0.61	0.44	0.45

Table 6.1: Classification performance per averaging strategy for each model
(*Three best values in green and three worst in red, overall best value in* ***bold***)

	Jaccard	Hamming
Baseline top-1	0.07	0.28
Baseline top-3	0.19	**0.25**
Rule-based	0.17	0.33
gpt-oss-20b	0.20	0.29
gpt-oss-120b	0.20	0.30
Gemma 3 4B	0.33	**0.25**
Gemma 3 27B	**0.38**	0.29
MiniMax M2.5	0.16	0.32
MiniMax M2	0.17	0.32
Qwen3-Coder-480B	0.34	0.32
Qwen3-Next-80B	0.27	0.27

Table 6.2: Jaccard score and Hamming loss for each model
(*Three best values in green and three worst in red, overall best value in* ***bold***)

The baseline models showcase high micro- and sample-averaged precision, solely because it predicts either the most frequent label or the three most frequent labels. Looking at other values, especially the macro-averaged metrics reflect their incapacity at predicting the other categories. It comes naturally that these are not be used for the classification task but rather serve as baselines for all other approaches, any model performing worse than these is set aside.

The rule-based approach does not perform well. It achieves the worst Hamming loss and one of the lowest F1-score no matter the averaging strategy. It achieves near perfect precision on `flood` and `drought` but very low recall across all hazard categories. The main issue with this model is that it assigns no label to the majority of documents, though when predicting a label it tends to be correct. It is mostly limited by the non-exhaustive list of keywords in the glossary and different terminologies it may not contain. Outperformed in some areas by the baseline, this rule-based approach is not used for the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

classification task on the whole set of documents.

The two gpt-oss versions show highest macro- and weighted-average precisions among all models, but also very low recall, suggesting the models are selective when making their predictions. Having high precision and low recall is a sign that when the model predicts a label, it is often correct but it rarely takes the risk of predicting a label incorrectly. Neither model reliably identifies the less frequent hazard types. They are both consistently outperformed by the Gemma 3 variants on all F1 and Jaccard scores. As such, these models are not used for the classification task on the whole set of documents.

The two Gemma 3 variants stand out as the best-performing models. Highlighted in green across almost all metrics, both achieve great balance between precision and recall. Gemma 3 27B realizes the best performance overall, with highest recall and F1-scores in all four averaging strategies, but also highest Jaccard score. Compared to Gemma 3 4B, it is only slightly outperformed by it in terms of precision. Gemma 3 4B achieves the most balanced performance across all categories, also being one of the few reaching over 0.5 in F1-score on `storm` and `temperature_extremes`. The strong performance of both Gemma 3 variants across diverse hazard categories suggests that their general-purpose aspect transfers well to this domain-specific multi-label classification task.

The two MiniMax models perform among the worst, not only of LLM approaches, but out of all of the considered methods. It is more often than not worse performing than the baseline models. Despite their 230B parameter size, their low recall and high Hamming loss indicate that their specialization for coding and agentic tasks does not transfer well to this task of climate policy classification. They are both easily dropped from the list of considered models for the classification task on the whole set of documents.

The two Qwen3 variations are strong contenders to the best-performing Gemma 3 models. Qwen3-Coder-480B performs comparably to Gemma 3 4B on several metrics, but shows lower precision and a higher Hamming loss, suggesting more false positive predictions. It performs rather poorly on `other`, suggesting it tends to assign specific hazard labels which inflates the rate of false positives for specific classes. Qwen3-Next-80B achieves intermediate performance but does not outperform either Gemma model.

Looking into the overall results per hazard category (Appendix B), consistent patterns are observed. The best-classified categories are `flood` and `wildfire`, benefiting from explicit and precise vocabulary in summaries. On the other hand, `melting` and `sea_level_rise` are the categories that are consistently less well-predicted, with most models achieving F1-scores below 0.50.

All in all, a clear trade-off between precision and recall divides the models into two groups. MiniMax and gpt-oss models favor higher precision with lower recall, by predicting few labels per document but those predictions tend to be correct. Gemma and Qwen-Coder models achieve better recall at the cost of some precision, resulting in substantially higher F1 and Jaccard scores.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

6.1.1 Monitoring Emissions

The environmental cost of running each model was estimated using CodeCarbon [CSGK+26], an open-source tool that tracks the energy consumption of code execution and estimates the corresponding carbon dioxide emissions based on the hardware used and the carbon intensity of the local electricity grid. It was used when running each model for the predictions on the benchmark. The three most important metrics reported by CodeCarbon are summarized in Table 6.3.

Model	**Duration**	**Emissions** (kgCO2eq)	**Energy** (kWh)
Baseline	1″	4.87e-7	4.39e-6
Rule-based	2″	5.19e-7	4.69e-6
gpt-oss-20b	13′20″	3.65e-4	3.29e-3
gpt-oss-120b	11′27″	3.16e-4	2.85e-3
Gemma 3 4B	3′41″	1.03e-4	9.27e-4
Gemma 3 27B	4′33″	1.31e-4	1.18e-3
MiniMax M2.5	47′10″	1.28e-3	1.15e-2
MiniMax M2	32′22″	8.61e-4	7.77e-3
Qwen3-Coder-480B	10′41″	2.91e-4	2.62e-3
Qwen3-Next-80B	149′2″	4.48e-3	4.04e-2

Table 6.3: CodeCarbon emissions' calculation for each model

It is no surprise that the baselines and rule-based models are the fastest to be executed, completing classification in under two seconds. Additionally, they are by far the less emitting and energy-consuming.

The interesting comparison happens at the level of the LLM models. Both Gemma 3 versions complete classification of the benchmark in under five minutes, over twice as fast as the gpt-oss models and consuming roughly three times less energy. This efficiency advantage is particularly notable given that the two outperform them on all classification metrics, demonstrating that smaller but well-trained models can simultaneously be more accurate and more sustainable. The MiniMax models represent the worst overall trade-off, taking over thirty minutes while achieving the lowest classification performance among all LLM approaches. Qwen3-Next-80B is clearly the most resource-intensive model, requiring nearly 2.5 hours and consuming 0.040 kWh, about 34 times more energy than Gemma 3 27B.

These results highlight that model size is not a proof of performance nor efficiency in this task. The two Gemma models consistently outperform much larger models while consuming a fraction of the energy.

6.1.2 Prompt Engineering

Prompt engineering is the process of adapting and fine-tuning prompts to improve the quality and accuracy of a generative AI's responses. This technique was implemented to observe the influence of different prompts on the outputs of the two gpt-oss and the two

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Gemma 3 models, and examine any improvements in the results, summarized in Table 6.4.

The original prompt, used to produce the initial results, was:

> Classify the following climate law or policy summary into one or more hazard categories.
>
> - Only use labels from this list: ['flood', 'drought', temperature_extremes', 'sea_level_rise', 'storm', 'wildfire', 'melting', 'erosion', 'other', 'none'].
> - Use 'none' only if the document has no relation to any physical climate hazard.
> - Use 'other' if the document addresses climate hazards not covered by the other categories.
> - Assign all applicable labels.
>
> Summary: {summary}

This was modified by adding the documents' title alongside the summary, and the model was instructed to identify the hazards mentioned or implied before classifying. This change was motivated by the observation that document titles can contain explicit hazard-related keywords not always repeated in the summary. In addition, this technique of chain-of-thought prompting, encouraging the model to reason about hazards before assigning labels, separates hazard identification from label assignment and gives the model room to reason about implicitly stated hazards [WMD+23]. The second version of the prompt became:

> First, identify which physical climate hazards are mentioned or implied in the document. Then, classify the following climate law or policy summary into one or more hazard categories.
>
> - Only use labels from this list: ['flood', 'drought', 'temperature_extremes', 'sea_level_rise', 'storm', 'wildfire', 'melting', 'erosion', 'other', 'none'].
> - Use 'none' only if the document has no relation to any physical climate hazard.
> - Use 'other' if the document addresses climate hazards not covered by the other categories.
> - Assign all applicable labels.
>
> Title: {title}
>
> Summary: {summary}

Then, the accent was put on confidence and prioritizing precision over recall by adding two instructions. The model was told to only assign a label when confident that the document explicitly or clearly addresses the corresponding hazard, and that assigning fewer labels was preferable to assigning uncertain ones. This change was motivated by the observation that the previous versions tended to over-assign labels based on loosely related or ambiguous mentions, and explicitly discouraging low-confidence assignments was expected to reduce false positives at the cost of potentially missing some weaker or implicit hazard mentions. The third version of the prompt became:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

First, identify which physical climate hazards are mentioned or implied in the document. Then, classify the following climate law or policy summary into one or more hazard categories.

- Only use labels from this list: ['flood', 'drought', 'temperature_extremes', 'sea_level_rise', 'storm', 'wildfire', 'melting', 'erosion', 'other', 'none'].
- Only assign a label if you are confident the document explicitly or clearly addresses that hazard.
- It is better to assign fewer labels than to assign uncertain ones.
- Use 'none' only if the document has no relation to any physical climate hazard.
- Use 'other' if the document addresses climate hazards not covered by the other categories.

Title: {title}

Summary: {summary}

	Macro			**Micro**			**Weight**			**Sample**				
	pre	rec	F1	pre	rec	F1	pre	rec	F1	pre	rec	F1	jac	ham
gpt-oss-20b (v1)	0.65	0.22	0.32	**0.62**	0.28	0.39	0.80	0.28	0.40	0.56	0.36	0.39	0.20	0.29
gpt-oss-20b (v2)	0.58	0.29	0.38	**0.62**	0.36	0.45	0.75	0.36	0.48	0.57	0.41	0.43	0.25	0.28
gpt-oss-20b (v3)	0.70	0.15	0.25	0.46	0.18	0.26	0.82	0.18	0.30	0.37	0.27	0.29	0.15	0.34
gpt-oss-120b (v1)	0.69	0.22	0.33	0.58	0.26	0.36	0.79	0.26	0.39	0.49	0.34	0.36	0.20	0.30
gpt-oss-120b (v2)	0.61	0.28	0.38	0.59	0.32	0.42	0.74	0.32	0.44	0.52	0.39	0.40	0.24	0.29
gpt-oss-120b (v3)	**0.71**	0.15	0.25	0.46	0.18	0.26	**0.84**	0.18	0.29	0.36	0.27	0.29	0.14	0.34
Gemma 3 4B (v1)	0.49	0.47	0.47	**0.62**	0.61	0.61	0.65	0.61	0.61	**0.61**	0.64	0.57	0.33	**0.25**
Gemma 3 4B (v2)	0.49	0.46	0.46	0.60	0.61	0.60	0.64	0.61	0.61	0.58	0.66	0.57	0.33	0.26
Gemma 3 4B (v3)	0.49	0.46	0.46	0.61	0.59	0.60	0.64	0.59	0.60	0.59	0.65	0.57	0.32	0.26
Gemma 3 27B (v1)	0.47	**0.63**	**0.52**	0.54	**0.74**	**0.63**	0.60	**0.74**	**0.65**	0.57	**0.76**	**0.60**	**0.38**	0.29
Gemma 3 27B (v2)	0.49	0.55	0.49	0.57	0.67	0.61	0.63	0.67	0.63	0.58	0.70	0.59	0.35	0.27
Gemma 3 27B (v3)	0.57	0.48	0.47	0.61	0.59	0.60	0.68	0.59	0.59	**0.61**	0.64	0.59	0.34	0.26

Table 6.4: Classification performance of each model with prompt engineering (*v1 refers to original prompt, v2 to second version, v3 to third version, three best values in green and three worst in red, overall best value in* ***bold***)

For the gpt-oss models, the first prompt variation notably improves recall and F1 score, with smaller gains for the 120B model. On the contrary, the second prompt variation focused on precision, does increase this metric but reduces the performances in recall and F1 score, but also those of Jaccard score and Hamming loss. This confirms that forcing the model to predict fewer labels produces more accurate predictions on the individual level but misses too many labels to be useful. Despite the improvements from the title-inclusive prompt, both gpt-oss models remain outperformed by both Gemma variants across all metrics.

For the Gemma models, the impact is less significant. Gemma 3 4B shows barely no variation across all three prompts and for all metrics. Gemma 3 27B displays more fluctuations, but improvements in a given metric comes at the cost of the others. It suggests that the original prompt was already extracting sufficient information from summaries.

Overall, the second prompt variant, including the title, provides meaningful gains for the gpt-oss models but minimal effects for the Gemma models. The third variant, focused

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

on precision and confidence, worsens Jaccard score and Hamming loss across all models, confirming this strategy is inappropriate for this multi-label classification task.

6.1.3 Model Selection

Based on the classification performance and computational cost results, Gemma 3 27B with the second prompt version, is selected to classify the full dataset of 9,527 documents. It is the most reliable model for assigning hazard labels, achieving the best classification performance across all metrics, with highest recall and F1-score, but also best Jaccard score.

Although the original prompt achieves a higher Jaccard score, the second prompt is preferred for the reason that the dataset contains a substantial number of rows with short summaries, for which titles may contain significant information when the summaries alone do not suffice.

Moreover, completing the classification of the benchmark in a little under five minutes, with an estimated energy consumption of 0.00118 kWh and emission of 0.000131 $kgCO_2eq$, the model offers a favorable trade-off between performance and computational cost. The classification of the entire dataset takes three hours and thirty-two minutes, with emissions of 0.005773 $kgCO_2eq$, a rather negligible environmental cost for a one-time classification task.

6.2 Correlation Results

This section presents the results of the correlation analysis introduced in Section 5.1.2. It first examines the relationship between legislative coverage and hazard intensity at country-level, before turning to the continental-level, and comparing the two levels of granularity. It then investigates the effect of legislation lag on the strength of the observed correlations, and concludes with a case study contrasting Australia and Europe to illustrate the hierarchical effect in a concrete setting.

6.2.1 Country-level Correlation

At the country-level, the initial step was to distribute EU-level laws to each relevant member state. Each corresponding law was duplicated for every member country, provided that the law's year falls on or after that country's entry year. In the case of the United Kingdom, the law's year must also precede its withdrawal year. This distribution increased the total number of rows from 9,527 to 16,787.

Figure 6.1 shows the distribution of Spearman ρ values across all country-hazard pairs. The graph on the left presents two clusters, one centered around $\rho = -0.1$ and another centered around $\rho = 0.6$, displaying the variety of negative and positive correlations in the dataset. When distinguishing significant and non-significant correlations with the p-value, it is observed that the positive tail ($\rho > 0.5$) is made of most significant

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

correlations, indicating that when a statistically significant relationship exists it tends to be strongly positive.

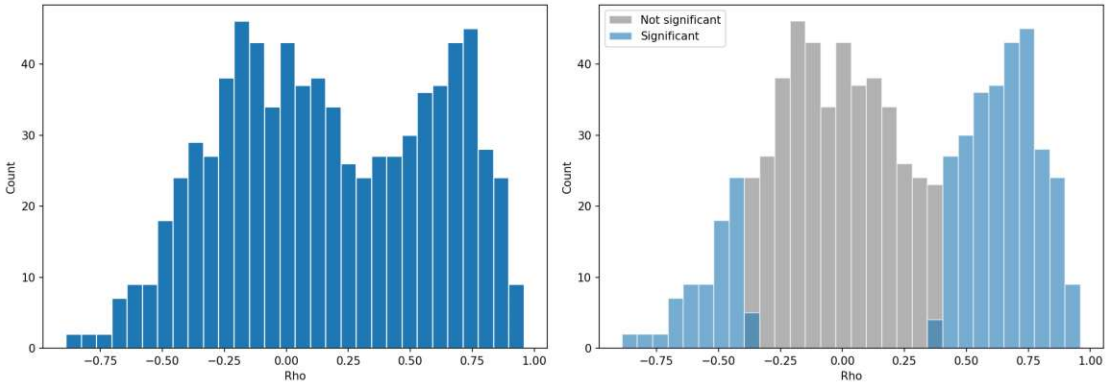


Figure 6.1: Distribution of country-level correlation coefficients (*left*) distinguished by significance (*right*)

Table 6.5 details the correlation results per hazard category and per significance. Non-significance refers to $p \geq 0.05$, a significant result is strongly positive when $\rho > 0.3$, strongly negative when $\rho < -0.3$, and weak otherwise. Alongside Figure 6.2, three distinct results across hazard categories are revealed. On the one hand, `sea_level_rise` and `temperature_extremes` achieve a positive Spearman ρ average value, on the other hand `drought` and `melting` are all negative, while `flood` and `storm` have near-zero mean.

	Not significant	Strong (-)	Strong (+)	Weak
Total	437	283	78	44
`drought`	120	36	7	0
`flood`	137	6	15	0
`melting`	38	19	3	0
`sea_level_rise`	7	2	96	44
`storm`	124	15	9	0
`temperature_extremes`	11	0	153	0

Table 6.5: Country-level correlation per hazard per significance

Both `sea_level_rise` and `temperature_extremes` show positive and statistically significant correlations, suggesting that legislative coverage of these hazards generally increases with observed hazard intensity over the period. This can be explained by the fact that these two hazards might be the most known. Even though sea level rise is only experienced in coastal country, its danger is acknowledged. Both variables' have been showing increasing trends for years, making them visible to policymakers and governments, explaining a high legislative coverage.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

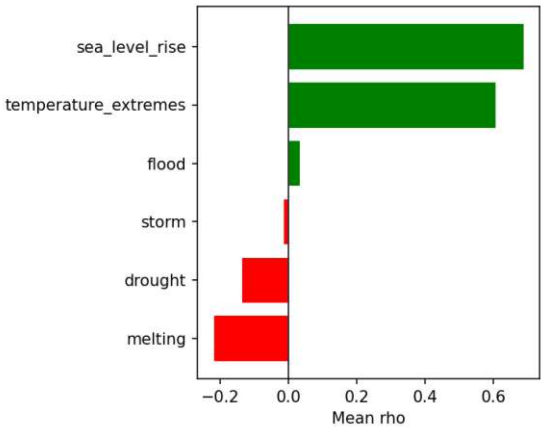


Figure 6.2: Average country-level correlation by hazard category

`flood` and `storm` show near-zero mean correlations, indicating no systematic relationship between the hazard variable and related legislation. This possibly is mostly due to the chosen variables not fully explaining the hazards.

`drought` and `melting` show stronger negative mean correlations. The negative correlation for `melting` suggests that while snowmelt anomalies decrease over time, legislative coverage increases. This is actually consistent global snowmelt trends that have been declining, with glaciers retreat producing less snowmelt, and simultaneously, the policy response to this long-term loss increases. For `drought`, the negative correlation is explained by the definition of the SPEI variable. With positive values indicating wet conditions and negative values indicating drought, a positive SPEI trend correlates negatively with drought legislation even when both are increasing.

6.2.2 Continental-level Correlation

At the continental-level, the total number of observations reduced to 36 pairs, with six hazard categories in six geographic regions, and 50% of them show statistically significant correlations, a higher proportion than at country-level.

Figure 6.3 shows the distribution of Spearman ρ values across all continent-hazard pairs. Similarly to the country-level distribution, two clusters are observed: one of negative values centered around $\rho = -0.4$ and another of positive values around $\rho = 0.6$. This polarization reflects the divergence between hazard categories, previously discussed in the country-level analysis. Moreover, the significant results are located in the two tails of the distribution. The central region around zero is predominantly non-significant, indicating that continental aggregation tends to either reinforce or filter relationships rather than producing weak intermediate signals.

Table 6.6 details the correlation results per hazard category and per significance and Figure 6.4 displays the average correlation per hazard category. Here, two distinct results

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

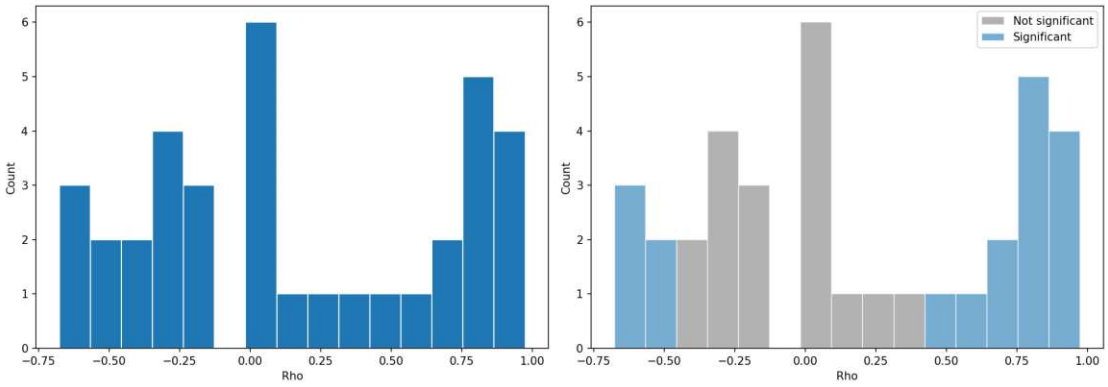


Figure 6.3: Distribution of continental-level correlation coefficients (*left*) distinguished by significance (*right*)

across hazard categories are observed. On the one hand, `temperature_extremes`, `sea_level_rise`, and `flood` achieve a positive Spearman ρ average value, on the other hand `storm`, `drought`, and `melting` are all negative.

	Not significant	Strong (-)	Strong (+)	Weak
Total	18	13	5	0
`drought`	4	1	1	0
`flood`	5	0	1	0
`melting`	4	2	0	0
`sea_level_rise`	0	1	5	0
`storm`	5	1	0	0
`temperature_extremes`	0	0	6	0

Table 6.6: Continental-level correlation per hazard per significance

Though all have positive correlation coefficients on average, each variable has its individual pattern. `temperature_extremes` shows the most consistent and strongest signal at continental-level, with all six regions' correlation coefficient ranging from 0.737 (South America) to 0.809 (North America), with a mean of 0.774. This consistency across confirms that the relationship between rising temperatures and legislative response is a global phenomenon. `sea_level_rise` shows the highest individual correlation in the entire analysis, with Africa achieving $\rho = 0.974$. However, the standard deviation is large (0.560) compared to the mean (0.659), reflecting variation across continents. Oceania (0.708) shows a strong significant positive correlation, consistent with the vulnerability of isolated Pacific island nations to sea level rise driving legislative activity. The high variance suggests that related legislation is geographically concentrated in the most exposed regions. `flood` improves substantially from country to continental-level, the mean ρ rising from 0.04 to 0.205. This improvement suggests that, while individual country flood responses are inconsistent, regional aggregate patterns are more coherent,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

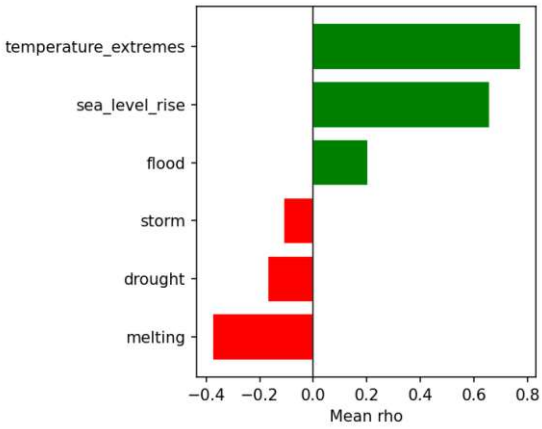


Figure 6.4: Average continental-level correlation by hazard category

possibly reflecting shared regional precipitation trends driving collective responses.

The same goes for the negative correlation coefficient. `storm` shows near-zero negative mean correlation, with no continent achieving a significant positive correlation. Wind gust anomalies appear to have no systematic relationship with storm legislation at either geographic scale, which may reflect that storm legislation is reactive to specific extreme events rather than to long-term trend changes in average wind intensity. `drought` remains negative at continental-level, with Africa showing the strongest negative signal. This significant negative correlation in Africa may indicate that drought legislation in Africa has increased over time while SPEI values have become positive, or that drought legislation precedes rather than follows observed drought intensity. Oceania is the only continent showing a significant positive drought correlation. `melting` shows consistently negative correlations across all continents with no significant positive values. This reinforces the country-level finding that declining snowmelt trends coincide with increasing legislative attention to melting hazards.

6.2.3 Granularity-level Comparison

Table 6.7 summarizes and compares the country-level and continental-level correlation results across all six hazard categories. The sign (positive or negative) of the correlation is preserved across both levels for all six categories, but the magnitude and the direction of the change between levels reveals different patterns emerging from the aggregation.

First, `flood` and `temperature_extremes` both show a difference of +0.17 after continental aggregation. For `temperature_extremes`, the mean increases from country-level to continental-level, suggesting that the relationship between rising temperatures and legislative coverage is more coherent when observed at a regional scale. For `flood`, the near-zero country-level mean becomes a slightly positive continental mean, indicating that a positive regional signal emerges when countries are aggregated.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Hazard	Mean rho (country)	Mean rho (continent)	Difference
flood	0.036	0.205	0.169
temperature_extremes	0.608	0.774	0.166
sea_level_rise	0.690	0.659	-0.031
drought	-0.138	-0.171	-0.033
storm	-0.015	-0.111	-0.095
melting	-0.220	-0.377	-0.157

Table 6.7: Comparison of correlation between country- and continental-level aggregation

By contrast, `sea_level_rise` and `drought` show barely any change between levels, suggesting that aggregation neither adds nor removes signal for these categories.

Finally, `storm` and `melting` both become more negative at continental-level. This is the opposite of the noise-reduction effect that could be expected from aggregation. For `melting` in particular, the continental mean is substantially stronger than the country-level mean, indicating that the negative relationship between declining snowmelt and increasing legislative coverage is a regional phenomenon, perhaps most visible in continents such as Europe and Asia where glacier retreat is a shared, continent-scale concern driving coordinated legislative responses.

This demonstrates that hierarchical aggregation is not only a noise-reduction technique. Depending on the hazard, it can either reveal a hidden positive signal (flood), amplify an existing positive one (temperature), leave the signal unchanged (sea level rise, drought), or deepen a negative one (storm, melting).

6.2.4 Legislation Lag

Table 6.8a and Table 6.8b report the proportion of statistically significant correlations at country- and continental-level respectively, for legislation lags of 0 to 5 years. A lag of k years tested whether legislative coverage at year y could be better explained by hazard intensity from k years prior than by coincident intensity.

Lag	Significant count	% Significant
0	361	42.9
1	344	40.9
2	320	38.0
3	310	36.8
4	323	38.3
5	330	39.1

(a) Country-level results

Lag	Significant count	% Significant
0	18	50.0
1	18	50.0
2	17	47.2
3	19	52.8
4	19	52.8
5	17	47.2

(b) Continental-level results

Table 6.8: Legislative lag

At country-level, the highest proportion of significant correlations is observed at lag 0 (42.9%), with a decline through lags 1-3 before a slight increase at lags 4 and 5. This

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

pattern is counterintuitive to the hypothesis that legislative response lags observed hazard intensity. If legislation reacted to prior-year hazard signals, one would expect lag 1 or lag 2 to outperform lag 0. The observed decline instead suggests that for the majority of country-hazard pairs, legislative coverage and hazard intensity move together rather than with a delay. This is consistent with the interpretation that both variables are driven by the same underlying long-term trend, particularly for `temperature_extremes` and `sea_level_rise`, where both the physical signal and the legislative response have been increasing steadily since 2000, rather than reflecting a reactive legislative mechanism responding to specific past hazard events.

At continental-level, the pattern is markedly different. The proportion of significant correlations remains virtually stable across all lag values, ranging only between 47.2% and 52.8%, with no clear directional trend. This stability indicates that at the continental scale, the relationship between hazard intensity and legislative coverage is equally strong regardless of temporal offset, further supporting the interpretation that the continental-level signal is driven by shared long-term trends rather than by reactive policy mechanisms with identifiable delays. The slightly higher significance at lags 3 and 4 could suggest that continental legislative cycles operate with a multi-year policy response lag, though the difference is too small to draw firm conclusions given the limited sample size of 36 continent-hazard pairs.

Overall, the lag analysis does not support the hypothesis that legislative coverage systematically lags hazard intensity. The reference lag of 0 years yields the strongest country-level results, and the continental pattern shows no meaningful lag sensitivity. The concurrent relationship identified at lag 0 is therefore retained as the primary analytical framework for the remainder of the analysis.

6.2.5 Case Study: Australia versus Europe

To illustrate the hierarchical effect in a concrete environment, this case study contrasts Australia, a single and geographically isolated country, against Europe, a bloc of countries subject to shared hazards. Russia, Turkey, and Iceland are excluded from the European group due to their positioning relative to the European continent. To avoid inflating legislative coverage counts for European countries through the EU law distribution step, EU-level laws are weighted by $\frac{1}{27}$ per member state. This is chosen to consider each EU policy as one unit of legislative coverage across the union rather than 27 times.

Table 6.9 reports the Spearman correlation between hazard intensity and legislative coverage for both entities. The results reveal a consistent pattern where for hazard categories where a signal exists, Europe achieves a stronger aggregate correlation than Australia, while for categories with no signal, both entities converge toward zero.

The most notable result is `temperature_extremes`, where both Australia and Europe show statistically significant positive correlations. At country-level in Europe, 37 out of 42 countries show significant positive correlations, confirming that this is a robust signal across the continent that cannot be attributed to supranational law weighting

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Hazard	Rho (AUS)	p-value (AUS)	Rho (EUR)	p-value (EUR)
flood	0.094	0.648	0.070	0.507
drought	-0.185	0.365	-0.056	0.496
temperature_extremes	0.468	0.016	0.662	0.002
sea_level_rise	0.533	0.007	0.671	0.035
storm	0.265	0.191	-0.011	0.511
melting	-0.250	0.218	-0.243	0.165

Table 6.9: Correlation by hazard category for Australia and Europe

alone. Europe's higher aggregate coefficient suggests that national legislative coverage across European countries moves more coherently with observed temperature trends than Australia does on its own, plausibly reflecting shared European climate frameworks and cross-border policy coordination that amplify individual national responses.

Similarly for sea_level_rise, both show significant positive correlations with comparable magnitudes, and 22 out of 42 European countries are individually significant. The similarity between the two entities here suggests that sea level rise legislation is driven primarily by national coastal exposure rather than by supranational coordination.

For melting, neither Australia nor Europe show statistically significant aggregate correlations, though the direction and magnitude are similar between the two. With 11 out of 42 European countries individually significant, the negative relationship is present but heterogeneous across the continent. The near-identical values between Australia and Europe for this category suggest that the melting signal is driven by geography and glacier exposure rather than by institutional factors.

For flood, drought, and storm, neither entity shows a statistically significant aggregate correlation. Only 2 European countries are significant for flood, 3 for drought, and 2 for storm. The structural absence of signal for these three categories is therefore robust to geographic scale, institutional context, and the treatment of supranational legislation.

Overall, this case study reveals that the hierarchical effect is not uniform. It manifests only where an underlying hazard-coverage relationship already exists at the national level. For temperature_extremes and sea_level_rise, aggregation to the European scale strengthens or preserves a signal that is independently visible in the majority of individual member states, indicating that supranational grouping amplifies coherence. For melting, the aggregate effect is negligible, while for flood, drought, and storm the absence of correlation is stable across both cases. This pattern supports the idea that hierarchical structures can reinforce legislative-hazard alignment when some response already exists, but they cannot substitute for it where national-level coverage is structurally disconnected from hazard intensity.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

6.3 Visualization Results

This section presents the results of the visualization approaches introduced in Section 5.1.3. It first reports on the preliminary user study conducted with three participants, before examining the evaluation results against the completeness, correctness, and clarity criteria defined in Section 5.2.2.

6.3.1 Preliminary User Study

The initial walkthrough, with User 1, was completed with no errors, the participant providing correct answers to the five tasks in 38 seconds per task on average. The second walkthrough, with User 2, was also completed with no errors, the participant providing correct answers to the five tasks in 22 seconds per task on average. The third walkthrough, with User 3, was also completed with no errors, the participant providing correct answers to the five tasks in 55 seconds per task on average. The ratings given by the participants to the three statements are summarized in Table 6.10.

Statements	User 1	User 2	User 3
The dashboard contained all the information I needed to answer the tasks	5	5	5
I trusted that the values shown were accurate and clearly defined	2.5	4	5
The figures were easy to interpret without extra explanation	4	4.5	3

Table 6.10: Ratings on a scale from 1 (strongly disagree) to 5 (strongly agree)

User 1

The map and the timelines were deemed easier to understand. This participant noted that maps and timelines are among the most common formats for communicating data to a general audience, so they could draw on prior experience with this type of visualization. In contrast, the scatter plot was evaluated as harder to read, since it took them longer to understand how the variables were plotted and how they evolved.

The coloring scheme, not only for visualizing the gap differences in the map and the bar chart but also for differentiating intensity and coverage in the timeline were noted as helpful visualization tools.

Some arrangements are implemented to meet the suggestions of User 1. First, the participant made explicit the need to see the data sources in order to have more confidence in what is shown in the dashboard. This scarcity is the reason behind the score given to Statement 2 addressing correctness. In this regard, hyperlinks referencing the data sources' websites are added to the header of the dashboard. Then, the explanation needed to be given by this author to grasp the idea behind policy gap led to the participant removing one point on the rating of Statement 3. For this reason, a simpler definition of the gap in the way of a mathematical subtraction is preferred. Additional discussion with User 1 revealed a design flaw on the first timeline, where the two y-axes respectively

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

of hazard intensity and legislative coverage were aligned on different grids, making the reading ambiguous.

Further comments were made but not addressed, to observe whether they are pointed out by different users. For example, a recurring issue this user had was with the year slider claiming that the selected year is not visible enough when looking for a specific date. Also, the positioning of the country slicer, on top of the map even though the latter is not linked to it, might be counter intuitive. It was discussed whether the place of the slicer should be moved to below the map, or the choice of a country should zoom in to said country when selecting it.

User 2

This user, with some background in mathematics and statistics, had a deeper understanding of the purpose of the dashboard and who it is meant for. Their familiarity with dashboards prompted them in recommending the addition of key figures at the top of the page, so that the context can be better understood and can encourage a user to look into the details of a specific country or hazard.

Similarly to User 1, this participant judged the map as one of the easiest figures to read and the scatter plot as hardest. This participant pointed out that the color palette from red to green is properly conveying a sense of urgency. Though this user's background helped in quickly understanding how to read the scatter plot, a struggle could be found in understanding that each dot represents a country as they are all stacked by default.

Some arrangements are enforced to meet the suggestions of User 2. First, the color scheme is standardized for uniformity across all the figures. This change is rather simple on the timeline of the policy gap by adjusting the plot color, and more advanced on the scatter plot, by removing the differentiation by continent to replace it with the color gradient of the gap value. Then, key figures were added at the top of the page to catch the attention of users with what can be learned through the dashboard. Finally, though the presence of the sources helped in increasing the score given to Statement 2, one point was removed for the absence of the date, hence its addition.

Furthermore, the lack of effect of the country slicer on the map was also observed by this participant, with an additional comment on the problems the map may cause to people with less general knowledge on geography. In this regard, the implementation of an automatic zoom on the region when selecting a country was attempted but unsuccessful, due to lack of experience on the Plotly library.

User 3

This user had more difficulty in understanding the general layout of the dashboard, though the setting and purpose of it was well interpreted. The addition of key figures prompted by User 2 as a headline, especially the hazard with the largest average policy gap being

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

highlighted, it first confused the participant in thinking that the whole dashboard was about said hazard.

Like both previous participants, this user evaluated the scatter plot as hardest to read. This was explained by the need to comprehend its layout and take into account both axes. In contrast, the bar chart was judged as the easiest figures to read, thanks to the size and color of each bar which intuitively give away the positive or negative direction of the gap.

Some arrangements are enforced to meet the suggestions of User 3. First, for simpler user experience, a scroll down arrow is added to encourage people in viewing more content straight away. Then, the headline containing key figures is improved to cause less confusion. Finally, a sentence synthesizing the key takeaways from the selected country-hazard pair is added at the bottom of the page. Indeed, this user underlined that there was a lot of information that could be grasped from each figure but with interpretation left to each individual. For people seeking information from such a dashboard but without the will or time to understand the figures and the data, a synthesis is a way to have a conclusion directly available.

6.3.2 Evaluation Results

Completeness

The completeness check confirms that all six hazard categories (drought, flood, melting, sea level rise, storm temperature extremes) can be chosen to be represented across all five visualizations. All 169 countries in the dataset can be selected to appear in each view.

Israel and Antarctica are shown in grey on the map due to absence from either or both legislative and sensor datasets. Other disputed territories are automatically shown in grey, notably Western Sahara, Arunachal Pradesh, Jammu and Kashmir, Ladakh, and Wadi Halfa Salient. Other pieces of land part of other countries are also shown in grey rather than the color of their countries, like Puerto Rico or French Guiana. In the case when a country has missing values for some hazards only, in the likes of landlocked countries for sea level rise, are also greyed out.

No unexpected gaps in coverage were identified beyond these cases. This is validated in the user study where 3 of 3 participants rated the statement „The dashboard contained all the information I needed to answer the tasks“ a 5, and no participant reported needing to consult information outside the artifact during task completion.

Correctness

The ten country-hazard combinations used for the spot-check were generated using Claude Sonnet 5, prompted as follows after input of the CSV file used in making the dashboard:

> Give me 10 country-hazard combinations, balanced across time, continent, and gap magnitude.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 6.11 summarizes the results of the correctness spot-check across the ten selected country-hazard combinations. All ten values on the dashboard match their corresponding source-file values exactly, and the sign of the gap (positive/negative) is correctly reflected by the color coding in every case, confirming that the visualization pipeline introduces no computational or display errors for the sampled combinations.

Country	Hazard	Year	Coverage	Intensity	NCoverage	NIntensity	Gap
Palestine	Drought	2004	1	-0.4124	0.0147	0.4676	-0.4529
Burkina Faso	Flood	2024	10	0.0993	0.1754	0.4104	-0.2349
Haiti	Melting	2006	0	0	0	0	0
Montenegro	Sea Level Rise	2018	3	0.0996	0.0682	0.9206	-0.8524
Uruguay	Storm	2010	5	0.2358	0.0877	0.6369	-0.5492
Australia	Temperature Extremes	2021	68	0.1602	0.5965	0.2656	+0.3309
Ethiopia	Melting	2023	1	0	0.0769	0	+0.0769
United States	Flood	2002	0	-0.0556	0	0.3879	-0.3879
Turkey	Temperature Extremes	2004	3	0.1354	0.0263	0.2604	-0.2341
Egypt	Sea Level Rise	2013	2	0.0612	0.0455	0.8977	-0.8522

Table 6.11: Summary of manually spot-checked country-hazard combinations

In the user study, 3 of 3 participants correctly understood the meaning of the gap after brief exploration of the dashboard, changing its definition from a sentence to an equation made it all the clearer for participants. When the meaning was not clear immediately, the color scale also helped in grasping the definition. Furthermore, more sense of correctness was perceived with the addition of the source and date.

Clarity

As described in Section 6.3.1, all participants were able to complete the five tasks in under 40 seconds on average. In no case did a participant request access to the underlying data to answer a task, with all responses directly accessible by interpreting the visualizations.

However, this clarity is not distributed uniformly across the five visuals. Two participants identified the map and the timelines as the most easily interpretable figures, attributing this to prior familiarity with these chart types from everyday contexts. The other participant identified the bar chart as most easily interpretable, connecting this to the color grading and bar size. In contrast, all three participants identified the scatter plot as the hardest to read: User 1 and User 3 reported needing more time to grasp how the plotted variables related to one another, while User 2 noted difficulty in distinguishing individual countries when points were stacked by default rather than visually differentiated.

The color coding used to represent the policy gap was highlighted as a particularly effective aid to clarity. User 2 specifically noted that the red-to-green scale intuitively communicates the urgency associated with the gap, without requiring to the legend.

These qualitative findings are confirmed by the ratings of the statement „The visualizations were easy to interpret without extra explanation“, which averaged at 3.8 across 3 participants.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

CHAPTER 7

Discussion

This chapter provides answers to the three research questions structuring this thesis by discussing the results of the research.

7.1 Research Question 1

How do NLP and LLM-based approaches perform and compare for multi-label classification of climate laws into hazard categories in terms of precision, recall, and F1-score?

The comparative evaluation on the benchmark set of 356 labeled documents shows two distinct behaviors between the rule-based NLP model and the best-performing LLM (Gemma 3 27B). The NLP approach achieves higher precision (0.54) than recall (0.21), producing a macro F1-score of 0.28 and a Jaccard index of 0.17. The best LLM inverts this pattern, trading precision (0.47) for higher recall (0.63). It outperforms the NLP approach on every other aggregate metric, with the macro F1-score nearly doubling to 0.52, the Jaccard index rising to 0.38, and the Hamming loss improving from 0.33 to 0.29. The inclusion of the document title in the prompt improved classification performance across models, suggesting that titles carry an additional signal that help in identifying the hazard. Nevertheless, this performance was not uniform across hazard categories. `flood` and `wildfire` were classified comparatively well by all models tested, while `sea level rise` and `melting` proved more difficult.

Each approach's inverted tendency to favor precision or recall reflects their underlying mechanisms. The rule-based NLP classifier only assigns a label when hazard-specific keywords are present, which rarely mislabels a document, but fails to recognize hazards described with words absent from the glossary, hence the low recall. The LLM classifier instead infers hazard relevance from larger signals brought by context and semantic, which allows it to correctly label many of the cases the other approach misses, but at the cost of more false positives. The consistently stronger classification of `flood` and `wildfire`

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

suggests these hazards are described with unambiguous and hazard-specific vocabulary, whereas `sea level rise` and `melting` are more often addressed implicitly, making them harder to be identified by both approaches.

The choice between the two approaches is not only a matter of maximizing the metrics, but also of the cost of the errors in the context of this classification. Because RQ2's correlation analysis is built on legislative coverage counts per country and hazard, missed labels (false negatives) understate this coverage, pulling the coverage-intensity relationship towards zero, and wrong labels (false positives) introduce noise. This asymmetry supports the choice of the LLM classifier, favoring recall, despite its lower precision, since under-counting legislative response would have been more damaging for the analysis. More generally, the results suggest that LLM classification is preferable to NLP classification when correct identification of the hazards addressed in legislation matters more than minimizing false positives, while rule-based methods may still be adequate when favoring precision and considerably cheaper to run both in terms of time and cost.

The results indicate that some difficulty remains in the classification process, and neither model flawlessly completed the task of labeling the legislation datasets. Part of this can be likely attributed to constraints on the classification approach itself, including the reliance on titles and summaries rather than full text, the restriction to free and open-source models, and the infeasibility of fine-tuning on a benchmark of this size. Each of these limitations possibly cap the scores of the models, independent of the relative comparison between the rule-based and LLM approaches. Particularly, this raises a question on whether the NLP model could perform competitively next to the LLMs with the access to full documents and an expanded glossary.

Future work should try to extract information directly from the full text of legislative documents, with embedded translation tool to handle documents not originally published in English. Working from full text would allow the classifiers to draw on more details and context clues than a summary can provide and could improve classification, particularly for the hazard categories that were more difficult to disambiguate. Pairing full-text extraction alongside a glossary built with an expert should be tested for the rule-based classifier, to observe whether LLMs' advantage comes from their broader vocabulary only, or also thanks to their reasoning capability. Additionally, future iterations of the classification pipeline could also explore a recall-oriented tuning or approaches that combine both approaches to gain from their respective precision and recall favoring abilities.

7.2 Research Question 2

To what extent does climate legislative coverage correlate with the intensity of locally observed climate hazard risks?

To what extent do continental-level correlation between legislative coverage and hazard intensity differ from country-level correlation?

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The Spearman correlation analysis between legislative coverage and hazard intensity reveals varied results depending on the hazard. At the country-level, `sea level rise` and `temperature extremes` show a positive average correlation between coverage and intensity, `drought` and `melting` show a negative average correlation, and `flood` and `storm` are close to zero. At the continental-level, the pattern shifts for `flood`, now showing a positive correlation. The lag analysis, testing whether legislative coverage responds to hazard intensity with a delay, did not find a systematic lag effect to exist, at either level of aggregation.

The positive relationship for `temperature extremes` and `sea level rise`, consistent across both levels of aggregation, suggests that for these two hazards, countries experiencing more risks tend to produce more climate legislation, a pattern plausibly explained by the high visibility of these hazards, most often mentioned in the public discourse. The negative correlation for `drought` and `melting` is harder to explain. It may reflect that legislative coverage for these hazards is driven less by the country's observed intensity and more by external factors, such as international agreements. The shift for `flood` from near-zero at the country-level to positive at the continental-level is notable. It suggests that country-level particularities dilute the relationship when analyzed individually, but that the relationship becomes visible once intensity and coverage are aggregated per continent. This is consistent with the case study comparing Europe and Australia, where regional governance structure was shown to affect how legislative coverage accumulates independently of hazard exposure. The absence of any lag effect, at either level, indicates that legislative coverage does not follow hazard intensity with a delay. Instead, coverage and intensity vary together, which may be because hazard intensity measures long-term conditions than single events, so by the time a law is passed, the risk has already been present for some time, rather than legislation being a fast reaction to one specific recent event.

These findings imply that legislative response to climate risks is not uniform and needs to be be assessed hazard by hazard. Policy alignment gaps, as visualized in the RQ3 dashboard, are therefore not a failure of climate governance to track risks, but a hazard-specific pattern. This has a direct implication for how the policy gap metric should be read and communicated. A large gap for `drought` should not necessarily be interpreted the same way as a large gap for `temperature extremes`. The scale-dependence observed for `flood` also implies that correlation analysis on a single level only risks under- or over-stating the true relationship, and that both levels should be reported together. The absence of a lag effect implies that using current hazard intensity to anticipate future legislative response, is not supported by this data, and that coverage and intensity should be treated as simultaneous data.

The correlation analysis is Spearman-based and therefore captures monotonic association, a positive or negative correlation for a given hazard should not be interpreted as evidence that hazard intensity does or does not cause legislative coverage, only that the two are or are not, associated. Since coverage counts come from the RQ1 classification, any residual classification error are brought along, particularly for hazard categories more difficult to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

disambiguate, and propagate into the coverage count used here and may explain weak or unexpected correlations for those hazards. The lag analysis only tested a fixed range of 0 to 5 years uniform across all hazards. This restriction might have missed some signals from a longer lag or a hazard-specific one. Beyond these, the correlation analysis is further constrained by how both variables were computed, legislative coverage was measured as a count of laws addressing a hazard, without regard to the depth or scope of each law, and hazard intensity was represented by a single variable per category rather than a combination, a simplification necessary given the scope of this thesis but that does not capture the physical complexity of any given hazard.

Future work could try and test which external drivers explain the correlations observed. Incorporating additional explanatory variables such as population or gross domestic product, but also motivation behind each legislation like international agreements, would provide an assessment of how much of the coverage-intensity relationship is explained by risk exposure itself versus other factors. Additionally, coverage computation could weigh the laws by scope rather than treating all laws addressing a hazard as equivalent. Confidence scores of the classification could also be inputted into the correlation analysis to quantify how much of the observed pattern is attributable to classification noise versus a genuine underlying relationship.

7.3 Research Question 3

To what extent can a visualization dashboard effectively communicate national gaps between climate hazard risk intensity and legislative coverage in terms of completeness, correctness, and clarity?

The preliminary user study, conducted with three participants, found that all three completed the five assigned tasks with no errors, at an average of 38, 22, and 55 seconds per task respectively. Across the completeness, correctness, and clarity criteria, participants rated completeness uniformly high (5/5 for all three), while the other criteria ratings varied more, with correctness rising from 2.5 to 4 to 5 and clarity fluctuating between 4, 4.5, and 3. Across all three users, the map and timeline views were consistently judged the easiest figures to interpret, while the scatter plot was consistently judged the hardest. Each walkthrough revealed concrete design issues, several of which were addressed in the design before the next walkthrough, to observe subsequent users' response to them.

Task completion with no errors by all three users, combined with uniformly high completeness ratings, indicates that the dashboard succeeds at effectively communicating the information needed to answer concrete questions about policy alignment gaps. The variation in correctness and clarity ratings, and their improvement, can be read as evidence that building an artifact in a cycle of building, evaluating, refining functions well. The consistent ranking of the map and timeline as easiest, and the scatter plot as hardest, across all three independent users suggests this might be underlying to each visual. While maps and timelines draw on visualization conventions participants already

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

have prior experience with, a two-axis scatter plot with one point per country requires the viewer to consider both axes and the color grading simultaneously.

These findings imply that an interactive dashboard can communicate policy alignment gaps to users without domain-specific training, provided the visuals draw on familiar conventions (maps, timelines, color-coded bar charts) rather than requiring users to handle less known visuals, as was necessary for the scatter plot. The correctness evaluation process also implies that trust in a dashboard's figures is not established by the figures alone. Explicit provenance through source links and publication dates affected how much participants trusted the same underlying values, independent of any change to the charts themselves. This suggests that source transparency is a prime factor in whether the artifact is judged trustworthy.

The user study evaluates a small sample of only three participants, tested sequentially with the dashboard revised between each session. Its value lies in the qualitative design issues raised and addressed rather than in the quantitative scores. Later participants encountered a different dashboard than earlier ones, and though the rating improvements are consistent with the design addressing real issues, this cannot be viewed as an evolution but may instead reflect the three participants' own standards. More fundamentally, the evaluation relies on a heuristic evaluation approach and a preliminary user study than a validation involving domain experts, which limits the extent to which claims about the dashboard's real-world decision-making relevance can be generalized.

Future evaluation should extend the study to a larger and more diverse participant pool, evaluated against a fixed version of the dashboard rather than one that evolves between sessions. The scatter plot's difficulty to be read by all three users, despite its analytical value, suggests an alternative way of showcasing for the coverage-intensity relationship could be considered. Most importantly, extending the dashboard evaluation with policymakers or other relevant stakeholders, rather than a convenience sample, would bring far more valuable insight into the dashboard's practical value and relevance in a real-world policy context, and should be prioritized.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 8

Conclusion

This chapter summarizes the key findings of the research, discusses limitations, and suggests directions for future research.

8.1 Summary

In this thesis, we explored the task of classifying climate legislation into hazard categories using NLP and LLM-based approaches, and examined how the resulting legislative coverage relates to observed climate hazard intensity across countries and continents.

The core contribution of this work is a pipeline covering multi-label hazard classification of climate laws, a correlation analysis linking legislative coverage to hazard intensity, and an interactive dashboard to communicate these findings through a policy alignment gap metric.

The main conclusions of this work were that zero-shot LLM classification outperforms rule-based and majority-class baselines for hazard classification, though substantial classification difficulty remains for several hazard categories. Moreover, the relationship between legislative coverage and hazard intensity is hazard-specific and partly dependent on scale (country or continent), with no evidence of a lag between hazard intensity and legislative response. Finally, an interactive dashboard using familiar visuals (maps, timelines, color-coded charts) can effectively communicate these findings to non-expert users.

This research contributes an open and reproducible pipeline for combining NLP-based legislative classification with climate hazard data and an artifact for communicating policy-hazard alignment gaps. The complete codebase developed for this thesis is publicly available on GitHub[1].

[1] https://github.com/emiliecaillerie/master-thesis

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

8.2 Limitations

While the work presented in this thesis addresses all three research questions, several limitations should be acknowledged when interpreting the results.

The classification approach was restricted by several factors. First, relying exclusively on titles and summaries rather than the full text of the underlying documents was necessary given the scope of this thesis and thanks to the availability of summary fields, but it means that details and nuances from the full legislative texts were not captured by the classifier. Then, the large language models used were limited to free and open-source options. While this allowed for a reproducible and cost-free process, it also means that potentially stronger models, including domain-specific models fine-tuned on legal or climate policy text, were not included. It is possible that such models would achieve higher classification performance. Additionally, fine-tuning a model on the climate legislation domain was not feasible in this thesis, as the subset of 356 documents annotated with hazards was too small to be reliably fine-tuned without a risk of overfitting. As a result, the classification relied on zero-shot approaches rather than a fine-tuned model.

The measure of legislative coverage used in the correlation analysis is solely based on the number of laws addressing a given hazard, without accounting for the content or scope of each law. This means that one law addressing a hazard in depth is treated equally to a narrower one, which may understate or overstate the true extent of legislative attention a hazard receives. Similarly, hazard intensity was computed using a single representative variable per hazard category, rather than with multiple relevant variables. This simplification was necessary given the scope of this thesis, but it does not fully capture the multi-dimensional nature of hazard intensity.

Finally, the evaluation of the visualization dashboard relied on a heuristic evaluation approach and a preliminary user study with a convenience sample, rather than a validation involving domain experts such as policymakers. This limits the extent to which claims about the dashboard's real-world decision-making relevance can be generalized.

8.3 Future Work

Building on the limitations outlined above, potential directions for future research include the following:

1. Extracting information directly from the full text of legislative documents rather than relying on summaries, potentially using an embedded translation step to handle documents not originally published in English. This would allow the classifier to draw on greater contextual detail and could improve classification granularity.
2. Integrating a pipeline for the automatic integration of newly passed legislation, allowing the classification and correlation analysis to remain up to date without requiring manual reprocessing of the entire dataset.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3. Developing a weighting process to reflect the relative importance and scope of each law, rather than treating them all as equally impactful. This would allow legislative coverage to be measured not just by count, but by significance.
4. Incorporating additional explanatory variables, such as population or gross domestic product, to add context in the correlation findings. This would allow an assessment of how much of the observed relationship between legislative coverage and hazard intensity is explained by risk exposure or other socioeconomic factors.
5. Extending the dashboard evaluation through usability testing with policymakers or other relevant stakeholders. This would bring valuable insight to better assess the dashboard's practical value and relevance in real-world context.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

APPENDIX A

Detailed Hazard Label Mapping

Hazard category	Raw text values
flood	flood, floods, flooding
drought	drought, droughts
temperature_extremes	heat waves and heat stress, heat waves, heatwaves, heat wave, heat stress, changes in average temperature, change in average temperature, cold waves, cold wave, increases in heat waves, heat
sea_level_rise	sea level rise, coastal erosion, sea level change, sea level rise soil erosion
storm	storms, storm, windstorms, windstorm, cyclones, hurricanes, hurricane, typhoons, tropical cyclones, storms/hurricanes/tsunamis/cyclones
wildfire	wildfires, fires, wildfire, forest wildfires, wild fire, fire
melting	glacial melting, loss of snow cover, snow melt, loss of ice mass, loss of sea ice, glacial melt, melting glacial ice, thawing permafrost, permafrost thawing
erosion	soil erosion, erosion, land erosion, soil soil erosion
other	tsunamis, earthquakes, earhquakes, earthquake, changes in average precipitation, changes in precipitation, changes in soil quality, change in soil quality, changes in air quality, change in air quality, changes in surface water, change in surface water, surface water change, changes in groundwater, groundwater change, desertification, biodiversity loss, ocean acidification, high tides, hail/frost, hail, epidemic, diseases, volcano eruptions, forest, lightning, avalanches, landslides, mudslides, landslide, glacial, sea water intrusion

Table A.1: Mapping of raw hazard-related text values to hazard categories

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

APPENDIX B

Detailed Classification Results

	Baseline top-1			Baseline top-3		
	precision	recall	f1-score	precision	recall	f1-score
drought	0.00	0.00	0.00	0.60	1.00	0.75
erosion	0.00	0.00	0.00	0.00	0.00	0.00
flood	0.74	1.00	0.85	0.74	1.00	0.85
melting	0.00	0.00	0.00	0.00	0.00	0.00
none	0.00	0.00	0.00	0.00	0.00	0.00
other	0.00	0.00	0.00	0.54	1.00	0.70
sea_level_rise	0.00	0.00	0.00	0.00	0.00	0.00
storm	0.00	0.00	0.00	0.00	0.00	0.00
temperature_extremes	0.00	0.00	0.00	0.00	0.00	0.00
wildfire	0.00	0.00	0.00	0.00	0.00	0.00

Table B.1: Performance of the baseline models detailed for each hazard

	precision	**recall**	**f1-score**
drought	1.00	0.17	0.29
erosion	0.42	0.12	0.19
flood	0.98	0.23	0.37
melting	0.05	0.22	0.09
none	0.00	0.00	0.00
other	0.69	0.49	0.58
sea_level_rise	0.64	0.38	0.48
storm	0.36	0.14	0.20
temperature_extremes	0.51	0.20	0.29
wildfire	0.75	0.17	0.28

Table B.2: Performance of the rule-based model detailed for each hazard

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

	gpt-oss-20b			gpt-oss-120b		
	precision	recall	f1-score	precision	recall	f1-score
drought	0.94	0.24	0.38	0.91	0.24	0.37
erosion	0.59	0.19	0.28	0.57	0.21	0.31
flood	0.92	0.30	0.45	0.92	0.30	0.45
melting	0.25	0.06	0.09	0.75	0.17	0.27
none	0.00	0.00	0.00	0.00	0.00	0.00
other	0.68	0.42	0.52	0.65	0.32	0.43
sea_level_rise	0.91	0.35	0.51	0.89	0.35	0.50
storm	0.54	0.22	0.31	0.56	0.19	0.28
temperature_extremes	0.85	0.26	0.40	0.88	0.25	0.39
wildfire	0.81	0.19	0.31	0.73	0.18	0.29

Table B.3: Performance of the gpt-oss models detailed for each hazard

	Gemma 3 4B			Gemma 3 27B		
	precision	recall	f1-score	precision	recall	f1-score
drought	0.78	0.61	0.69	0.73	0.85	0.79
erosion	0.47	0.33	0.39	0.41	0.64	0.50
flood	0.86	0.65	0.74	0.85	0.81	0.83
melting	0.14	0.22	0.17	0.46	0.33	0.39
none	0.00	0.00	0.00	0.00	0.00	0.00
other	0.56	0.98	0.72	0.60	0.54	0.57
sea_level_rise	0.71	0.37	0.49	0.49	0.76	0.59
storm	0.54	0.59	0.57	0.37	0.82	0.51
temperature_extremes	0.49	0.65	0.56	0.33	0.92	0.49
wildfire	0.38	0.29	0.33	0.46	0.61	0.52

Table B.4: Performance of the Gemma 3 models detailed for each hazard

	MiniMax 2.5			MiniMax 2		
	precision	recall	f1-score	precision	recall	f1-score
drought	0.90	0.21	0.34	0.89	0.23	0.36
erosion	0.48	0.13	0.20	0.60	0.17	0.26
flood	0.93	0.24	0.38	0.99	0.26	0.41
melting	0.50	0.11	0.18	0.50	0.11	0.18
none	0.00	0.00	0.00	0.00	0.00	0.00
other	0.70	0.31	0.43	0.61	0.29	0.39
sea_level_rise	0.84	0.29	0.43	0.85	0.25	0.39
storm	0.61	0.18	0.28	0.52	0.18	0.27
temperature_extremes	0.78	0.16	0.26	0.82	0.20	0.32
wildfire	0.71	0.13	0.23	0.72	0.15	0.24

Table B.5: Performance of the MiniMax models detailed for each hazard

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

	Qwen3-Coder-480B			Qwen3-Next-80B		
	precision	recall	f1-score	precision	recall	f1-score
drought	0.79	0.76	0.78	0.83	0.33	0.47
erosion	0.42	0.63	0.50	0.50	0.34	0.41
flood	0.92	0.70	0.80	0.90	0.36	0.52
melting	0.11	0.67	0.18	0.56	0.28	0.37
none	0.00	0.00	0.00	0.00	0.00	0.00
other	0.63	0.10	0.17	0.68	0.48	0.56
sea_level_rise	0.49	0.79	0.61	0.76	0.48	0.59
storm	0.39	0.83	0.53	0.59	0.35	0.44
temperature_extremes	0.39	0.87	0.54	0.63	0.42	0.50
wildfire	0.47	0.78	0.58	0.58	0.21	0.31

Table B.6: Performance of the Qwen3 models detailed for each hazard

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

APPENDIX C

Visualization Dashboard

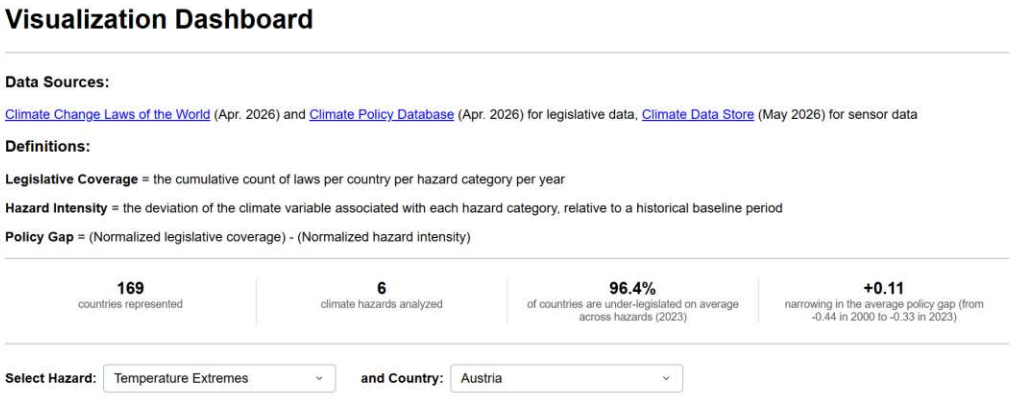


Figure C.1: Header of the dashboard, including sources, definition, and key figures

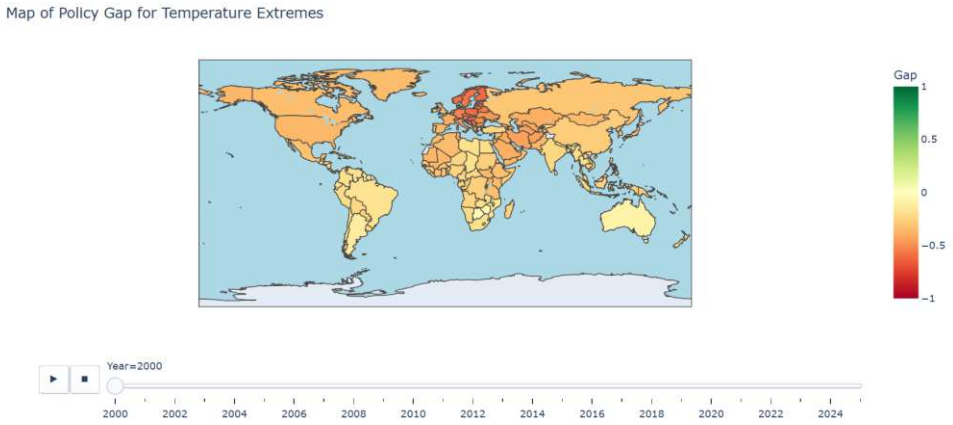


Figure C.2: World Map

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

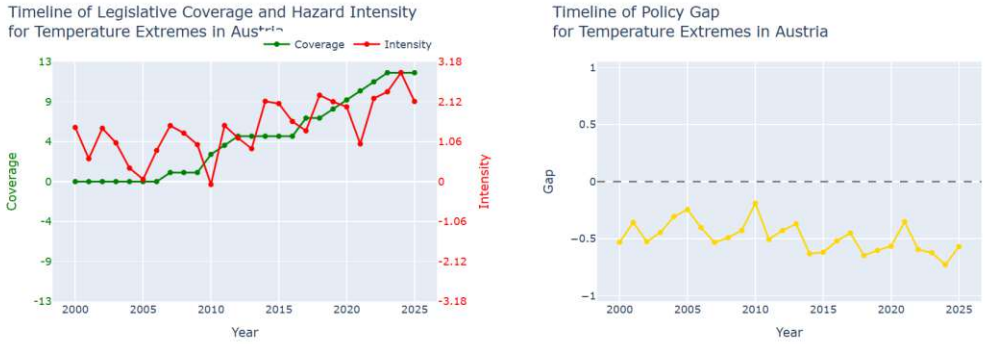


Figure C.3: Timelines

Figure C.4: Bar Chart (*left*) and Scatter Plot (*right*)

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Overview of Generative AI Tools Used

In this thesis, the following AI tools were used to support the writing process:

- Anthropic Claude: Claude Sonnet 4.6 and Claude Sonnet 5;
- Google Gemini: Gemini 3.0 Flash and Gemini 3.5 Flash.

These tools were used for translating between English and German, correcting grammar, enhancing sentence clarity and academic writing style. All AI-generated suggestions were carefully reviewed and manually edited to ensure accuracy and clarity, with no whole passages directly included without substantial changes.

They did not contribute to the development of the methodology, experiments, or results. All ideas, analysis, and findings are the original work of this author.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Bibliography

[AL24] A. Akshara and B. Latha. Precipitation Climate Data Analysis using Power BI. In *2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, pages 1–4, 2024.

[BBA25] J. Bonenkamp, R. Biesbroek, and I. N. Athanasiadis. Transforming adaptation tracking: benchmarking Transformer-based NLP approaches to retrieve adaptation-relevant information from climate policy text. In K. Dutia, P. Henderson, M. Leippold, C. Manning, G. Morio, V. Muccione, J. Ni, T. Schimanski, D. Stammbach, A. Singh, A. Su, and S. A. Vaghefi, editors, *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 266–277. Association for Computational Linguistics, July 2025.

[BMR+20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[CA25] H. Cho and E. Ackom. Artificial Intelligence (AI)-driven approach to climate action and sustainable development. *Nature Communications*, 16(1):1228, 2025.

[CBDH+25] M. Callaghan, L. Banisch, N. Doebbeling-Hildebrandt, D. Edmondson, C. Flachsland, W. F. Lamb, S. Levi, F. Müller-Hansen, E. Posada, S. Vasudevan, and J. C. Minx. Machine learning map of climate policy literature reveals disparities between scientific attention, policy density, and emissions. *npj Climate Action*, 4(1):7, 2025.

[CBS25] T. Calamai, O. Balalau, and F. M. Suchanek. Benchmarking the Benchmarks: Reproducing Climate-Related NLP Tasks. In W. Che, J. Nabende,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17967–18009. Association for Computational Linguistics, July 2025.

[CC18] Copernicus Climate Change Service and Climate Data Store. Sea level gridded data from satellite observations for the global ocean from 1993 to present. https://cds.climate.copernicus.eu/datasets/satellite-sea-level-global?tab=overview, 2018. Last Accessed: June 1st 2026.

[CC22] Copernicus Climate Change Service and Climate Data Store. ERA5-Land monthly averaged data from 1950 to present. https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land-monthly-means?tab=overview, 2022. Last Accessed: June 1st 2026.

[CC23] Copernicus Climate Change Service and Climate Data Store. ERA5 monthly averaged data on single levels from 1940 to present. https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels-monthly-means?tab=overview, 2023. Last Accessed: June 1st 2026.

[CG23] Climate Policy Radar and Grantham Research Institute at the London School of Economics. Climate Change Laws of the World. https://climate-laws.org, 2023. Last Accessed: May 4th 2026.

[Cop] Copernicus Climate Change Service. Climate Data Store. https://cds.climate.copernicus.eu/. Last Accessed: June 1st 2026.

[CSGK+26] B. Courty, V. Schmidt, Goyal-Kamal, inimaz, M. Coutarel, L. Blanche, B. Feld, J. Lecourt, L. Connell, A. Saboni, P. Lloret, supatomic, M. Léval, cianc, A. Cruveiller, ouminasara, F. Zhao, C. Bauer, J.L. Festus, A. Joshi, A. Bogroff, H. de Lavoreille, N. Laskaris, A. Phiev, E. Abati, D. Blank, and rosekelly6400. mlco2/codecarbon: v3.2.6, March 2026.

[CSX+21] T. Corringham, D. Spokoyny, E. Xiao, C. Cha, C. Lemarchand, M. Syal, E. Olson, and A. Gershunov. BERT Classification of Paris Agreement Climate Action Plans. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021.

[Cut52] Cutlip, S. M. and Center, A. H. *Effective Public Relations*. Prentice-Hall, 1952.

[Dee25] Google DeepMind. Gemma 3. https://deepmind.google/models/gemma/gemma-3/, March 2025. Last Accessed: May 26th 2026.

[Dut25] A. Dutta. Quantifying Climate Policy Action and Its Links to Development Outcomes: A Cross-National Data-Driven Analysis. In *NeurIPS 2025 Workshop on Tackling Climate Change with Machine Learning*, 2025.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[ECM25] ECMWF. cfgrib Python library v0.9.15.1, September 2025.

[Eur25] European Centre for Medium-range Weather Forecast Cross Data Store. Monthly drought indices from 1940 to present derived from ERA5 reanalysis. https://cds.climate.copernicus.eu/datasets/derived-drought-historical-monthly?tab=overview, 2025. Last Accessed: June 1st 2026.

[Gee25] GeeksforGeeks. NLP vs LLM: Understanding Key Differences. https://www.geeksforgeeks.org/nlp/nlp-vs-llm/, 2025. Last Accessed: April 27th 2026.

[GHW25] J. Gyeol, S. Ho, and S. Woo. Sustainable Storytelling in Digital Climate Campaigns: A Comparative Analysis of Visual Communication Strategies. *International Journal of Graphic Design*, 3(1):66–80, 2025.

[GL24] F. Grasso and S. Locci. Assessing Generative Language Models in Classification Tasks: Performance and Self-evaluation Capabilities in the Environmental and Climate Change Domain. In A. Rapp, L. Di Caro, F. Meziane, and V. Sugumaran, editors, *Natural Language Processing and Information Systems*, volume 14763, pages 302–313. Springer Nature Switzerland, 2024.

[Glo24] Global Climate Observing System. Essential Climate Variables: Surface Temperature. https://gcos.wmo.int/site/global-climate-observing-system-gcos/essential-climate-variables/surface-temperature, 2024. Last Accessed: August 10th 2026.

[GMM25] H. Ghiloufi, N. Merveille, and S. Mellouli. Exploring Emerging NLP and Machine Learning Methods in Climate Change Discourse Analysis on Social Media: A Systematic Literature Review. In M. Cheriet, L. Boucher, J.-F.and Gondim de Almeida Guimarães, and J.-M. Frayret, editors, *Accelerating the Socio-Ecological Transition: Strategies and Innovations for Sustainable Development*, pages 137–173. Springer Nature Switzerland, 2025.

[HA18] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[HCRdJ16] F. Herrera, F. Charte, A.J. Rivera, and M.J. del Jesus. Multilabel classification. *Springer*, 2016.

[JMM+24] M. Juhasz, T. Marchand, R. Melwani, K. Dutia, S. Goodenough, H. Pim, and H. Franks. Identifying Climate Targets in National Laws and Policies using Machine Learning. *arXiv*, 2024.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[KJMB+20] J. Kucera, G. Janssens-Maenhout, A. Brink, H. Greidanus, P. Roggeri, P. Strobl, G. Tartaglia, A. Belward, and M. Dowell. *Copernicus and Earth observation in support of EU policies - Part I: Copernicus uptake in the European Commission*, volume EUR 30030 EN. Publications Office of the European Union, 2020.

[KZNM25] M. Kurfali, S. Zahra, J. Nivre, and G. Messori. ClimateEval: A Comprehensive Benchmark for NLP Tasks Related to Climate Change. In K. Dutia, P. Henderson, M. Leippold, C. Manning, G. Morio, V. Muccione, J. Ni, T. Schimanski, D. Stammbach, A. Singh, A. Su, and S. A. Vaghefi, editors, *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 194–207. Association for Computational Linguistics, July 2025.

[Lew21] E. Lewinson. Introduction to hierarchical time series forecasting. `https://towardsdatascience.com/introduction-to-hierarchical-time-series-forecasting-part-i-88a116f2e2/`, 2021. Last Accessed: April 27th 2026.

[LHC+25] F. Larosa, S. Hoyas, J. A. Conejero, J. Garcia-Martinez, F. Fuso-Nerini, and R. Vinuesa. Large language models in climate and sustainability policy: limits and opportunities. *Environmental Research Letters*, 20(7):074032, 2025.

[Met20] J. Metag. Climate change visuals: a review of their effects on cognition, emotion and behaviour. In *Research Handbook on Communicating Climate Change*, pages 153–160. Edward Elgar Publishing, 2020.

[Min25] MiniMax. MiniMax M2 & Agent: Ingenious in Simplicity. `https://www.minimax.io/news/minimax-m2`, October 2025. Last Accessed: May 26th 2026.

[Min26] MiniMax. MiniMax M2.5: Built for Real-World Productivity. `https://www.minimax.io/news/minimax-m25`, February 2026. Last Accessed: May 26th 2026.

[MM21] Jan McDonald and Phillipa C. McCormack. Rethinking the role of law in adapting to climate change. *WIREs Climate Change*, 12(5):e726, 2021.

[MPCOF+21] Fernando Martínez-Plumed, Lidia Contreras-Ochando, Cèsar Ferri, José Hernández-Orallo, Meelis Kull, Nicolas Lachiche, María José Ramírez-Quintana, and Peter Flach. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3048–3061, 2021.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[MPS21] P. Mancuso, V. Piccialli, and A. M. Sudoso. A machine learning approach for forecasting hierarchical time series. *Expert Systems with Applications*, 182:115102, 2021.

[NAS24a] NASA Earthdata. Drought Monitoring, Prediction, and Projection using NASA Earth System Data. https://www.earthdata.nasa.gov/learn/trainings/drought-monitoring-prediction-projection-using-nasa-earth-system-data, 2024. Last Accessed: August 10th 2026.

[NAS24b] NASA Earthdata. Monitoring and Predicting Floods Using Earth Observations. https://www.earthdata.nasa.gov/learn/trainings/monitoring-predicting-floods-using-earth-observations-planning-preparedness, 2024. Last Accessed: August 10th 2026.

[NAS24c] NASA Earthdata. Wind Speed. https://www.earthdata.nasa.gov/topics/atmosphere/wind-speed, 2024. Last Accessed: August 10th 2026.

[NAS25] NASA Earthdata. GLIMS Database Advances Glacier Monitoring from Space. https://www.earthdata.nasa.gov/news/feature-articles/glims-database-advances-glacier-monitoring-from-space, 2025. Last Accessed: August 10th 2026.

[ngr26] ngrok. ngrok Agent 3.39.8, June 2026.

[NIA24] M. Noguer I Alonso. Key Milestones in Natural Language Processing (NLP) 1950 - 2024. *SSRN Electronic Journal*, 2024.

[NWP25] NewClimate Institute, Wageningen University and Research, and PBL Netherlands Environmental Assessment Agency. Climate Policy Database. https://climatepolicydatabase.org, 2025. Last Accessed: May 4th 2026.

[Oko15] Chitu Okoli. A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems*, 37:879–910, 2015.

[Oll25] Ollama. Ollama Python library v0.6.2, March 2025.

[OOOO23] J. O. Ojadi, E. C. Onukwulu, C. S. Odionu, and O. A. Owulade. Natural Language Processing for Climate Change Policy Analysis and Public Sentiment Prediction: A Data-Driven Approach to Sustainable Decision-Making. *Iconic Research And Engineering Journals*, 7(3), 2023.

[Ope25] OpenAI. Introducing gpt-oss. https://openai.com/index/introducing-gpt-oss/, August 2025. Last Accessed: May 26th 2026.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[Pan26] Pandas. Pandas Python library v3.0.1, February 2026.

[Plo26] Plotly. Plotly Python library v6.9.0, July 2026.

[PTRC07] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3):45–77, 2007.

[Qwe25a] QwenTeam. Qwen3-Coder: Agentic Coding in the World. `https://qwen.ai/blog?id=qwen3-coder`, July 2025. Last Accessed: May 26th 2026.

[Qwe25b] QwenTeam. Qwen3-Next: Towards Ultimate Training & Inference Efficiency. `https://qwen.ai/blog?id=4074cca80393150c248e508aa62983f9cb7d27cd&from=research.latest-advancements-list`, September 2025. Last Accessed: May 26th 2026.

[RDK+23] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, A. S. Luccioni, T. Maharaj, E. D. Sherwin, S. K. Mukkavilli, K. P. Kording, C. P. Gomes, A. Y. Ng, D. Hassabis, J. C. Platt, F. Creutzig, J. Chayes, and Y. Bengio. Tackling Climate Change with Machine Learning. *ACM Computing Surveys*, 55(2):1–96, February 2023.

[Reg24] Regionmask. Regionmask Python library v0.13.0, December 2024.

[RFM25] M. Rashik, J.-D. Fekete, and N. Mahyar. CLAImate: AI-Enabled Climate Change Communication through Personalized and Localized Narrative Visualizations. In *2025 IEEE Visualization and Visual Analytics (VIS)*, pages 36–40, 2025.

[SAL+25] R. Soden, T. S. Agrawaal, A. Lord, C. Chanen, L. Flawn, Z. Seaifan, M. Classens, and S. Easterbrook. Climate Data Practices: A Research Approach for HCI and Climate Justice. *ACM Trans. Comput.-Hum. Interact.*, 32(4):34:1–34:27, 2025.

[San25] B. Santhosh. Integration of Satellite Data and AI for Habitat Monitoring and Conservation. In Y. Y. Raghav, A. Chauhan, P. Pandey, and S. B. Khan, editors, *AI and Machine Learning Techniques for Wildlife Conservation*, pages 267–295. IGI Global Scientific Publishing, 2025.

[SBK+23] T. Schimanski, J. Bingler, M. Kraus, C. Hyslop, and M. Leippold. ClimateBERT-NetZero: Detecting and Assessing Net Zero and Reduction Targets. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15745–15756. Association for Computational Linguistics, 2023.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[SCCBK24] D. Spokoyny, J. Cai, T. Corringham, and T. Berg-Kirkpatrick. Aligning Unstructured Paris Agreement Climate Plans with Sustainable Development Goals. In D. Stammbach, J. Ni, T. Schimanski, K. Dutia, A. Singh, J. Bingler, C. Christiaen, N. Kushwaha, V. Muccione, S. A. Vaghefi, and M. Leippold, editors, *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 223–232. Association for Computational Linguistics, 2024.

[Sl25] Scikit-learn. Scikit-learn Python library v1.6.1, January 2025.

[SLBS26] C. Stryker, F. Lee, D. Bergmann, and M. Scapicchio. The 2026 Guide to Machine Learning. `https://www.ibm.com/think/machine-learning#605511093`, 2026. Last Accessed: April 27th 2026.

[SLT24] P. Singh, E. Lehmann, and M. Tyrrell. Climate Policy Transformer: Utilizing NLP to track the coherence of Climate Policy Documents in the Context of the Paris Agreement. In D. Stammbach, J. Ni, T. Schimanski, K. Dutia, A. Singh, J. Bingler, C. Christiaen, N. Kushwaha, V. Muccione, S. A. Vaghefi, and M. Leippold, editors, *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 1–11. Association for Computational Linguistics, August 2024.

[SZC+18] A. Sharma, S. M. A. Zaidi, V. Chandola, M. R. Allen, and B. L. Bhaduri. WebGIobe - A cloud-based geospatial analysis framework for interacting with climate data. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, BigSpatial '18, pages 42–46. Association for Computing Machinery, 2018.

[TCC22] M. Terrado, L. Calvo, and I. Christel. Towards more effective visualisations in climate services: good practices and recommendations. *Climatic Change*, 172(1):18, 2022.

[TLCV23] D. Trajanov, G. Lazarev, L. Chitkushev, and I. Vodenska. Comparing the performance of ChatGPT and state-of-the-art climate NLP models on climate-related text classification tasks. *E3S Web of Conferences*, 436:02004, 2023.

[UNE21] UNEP. Live Web Intelligence Platform. `https://unep.ecoresearch.net/weblyzard/`, January 2021.

[WKBL22] N. Webersinke, M. Kraus, J. A. Bingler, and M. Leippold. ClimateBert: A Pretrained Language Model for Climate-Related Text. *arXiv*, 2022.

[WMD+23] B. Wang, S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer, and H. Sun. Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2717–2739. Association for Computational Linguistics, July 2023.

[YSZ24] F. Yudanto, Y. Sari, and M. Z. A. C. Zaki. Climate-NLI: A Model for Natural Language Inference and Zero-Shot Classification on Climate-Related Text. In N. Oco, S. N. Dita, A. M. Borlongan, and J.-B. Kim, editors, *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 600–608. Tokyo University of Foreign Studies, December 2024.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.