International Journal of Recent Advances in Multidisciplinary Topics
Volume 7, Issue 6, June 2026
www.ijramt.com | E-ISSN: 2582-7839 | RESAIM Publishing | www.resaim.com

A Topic Tone Framework for Tracking Linguistic Framing in Climate Science Abstracts: Evidence from arXiv (1995–2026)

Kajol Bala[1*], Sonia Akter[2]
[1]Department of Medical Physics and Biomedical Engineering, Gono University, Savar, Bangladesh
[2]Department of Electrical and Electronic Engineering, Gono University, Savar, Bangladesh

***Abstract*: Climate change has become central to global policy and public debate, but it's less clear whether this shift shows up in the more guarded language of scientific writing itself. Because standard sentiment analysis tools don't work well on formal academic prose, we instead build a topic tone framework that pairs Latent Dirichlet Allocation with custom lexicon-based scores for uncertainty, certainty, and risk intensity capturing how confidently and how urgently claims are made, rather than emotional tone. Using climate-related abstracts from the arXiv metadata corpus (1995–2026), we model topics, score tone, and run regression and change point analyses to track trends over time. We finds a gradual but statistically meaningful rise in risk-oriented language, apparently accelerating after 2015, while uncertainty fluctuates rather than steadily declining possibly reflecting the growing complexity of climate modeling. Topics also carry distinct linguistic signatures: policy related work shows more risk language, while modeling heavy work hedges more. None of this points to scientific writing becoming "alarmist," but it does suggest a real, incremental shift in how climate risk is discussed one this reproducible framework could help track in other fields too.**

***Keywords*: arXiv corpus, change point detection, climate science communication, epistemic stance, hedging and risk language, Latent Dirichlet Allocation, topic modelling.**

1. Introduction

Climate change has moved from a specialist scientific concern to a central issue in global policy and public life, driven by international reports, extreme-weather events, and sustained media coverage. Whether this growing urgency is reflected in scientific writing itself is less clear: academic prose is bound by norms of caution, qualification, and restraint, so even a more pressing subject matter may not translate into visibly different language. This paper asks whether the linguistic framing of climate science has shifted in any measurable way over time, or whether it remains within its traditional, hedge-heavy register.

Standard sentiment-analysis tools are not well suited to this question. Lexicon-based methods such as VADER work well on informal, emotionally explicit text [1], [2], but reviews of sentiment analysis in specialised domains consistently find that general-purpose lexicons misfire on technical prose, where everyday words carry domain-specific meanings [3]. Scientific abstracts rarely read as overtly positive or negative; instead, meaning is carried by modal verbs, hedges, and risk-related vocabulary that signal epistemic stance rather than emotion the difference between “this model suggests an increase” and “this model demonstrates an increase” is one of commitment, not sentiment. Hyland's work on hedging established that such devices are a core rhetorical resource scientists use to calibrate claims [5], with later studies cataloguing how hedges and boosters vary across disciplines and genres [6], [7]. Separately, topic modelling particularly LDA [4] has been widely used to trace thematic change in climate-related literature, from cities and adaptation research [8], [9] to broader sustainability and climate-finance corpora [10], [12], and even social-media climate discourse [11]. These two literatures rarely meet: hedging research has the right linguistic categories but is seldom applied at corpus scale over long time spans, while topic-modelling studies track thematic change but not how language within those themes is framed.

This paper combines both strands into a topic-tone framework. We define three interpretable tone dimensions uncertainty (hedges such as “may,” “might,” “suggests”), certainty (assertive terms such as “demonstrates,” “confirms”), and risk intensity (urgency or severity terms such as “critical,” “extreme”), following established hedging and stance categories [5]-[7] and pair them with an LDA topic model [4] to ask not just how language changes over time, but whether that change is driven by shifts within topics or by shifts in which topics dominate the literature.

The framework is applied to roughly three decades of climate-related abstracts from the arXiv metadata corpus [13]. Section 2 describes the dataset, preprocessing, topic model, and tone metrics. Section 3 presents result on topic evolution and temporal tone trends. Section 4 discusses implications and limitations, and Section 5 concludes.

2. Materials and Methods

A. Dataset and Filtering

The data come from the arXiv metadata corpus, a public collection of bibliographic records titles, abstracts, subject

*Corresponding author: kajolbala.eee@gmail.com

categories, and submission dates for roughly 1.7 million papers [13]. Climate-related abstracts were extracted via keyword filtering ("climate change," "global warming," "climate risk"), yielding a corpus spanning 1995–2026 (Table 1).

Table 1
Dataset summary

Metric	Value
Total/filtered climate papers	~4,704
Time range covered	1995–2026
Average abstract length	~189 words

B. Preprocessing

Abstracts were lower-cased, stripped of punctuation and numeric tokens, filtered against a stop-word list, lemmatised, and cleaned of citation artefacts (e.g., "et al."). Domain relevant terms such as "risk," "model," and "uncertainty" were retained despite high frequency. This reduced average tokens per abstract from 188.6 to 113.7, with a cleaned vocabulary of 18,259 terms (Table 2).

Table 2
Preprocessing statistics

Metric	Value
Average tokens (raw → cleaned)	188.64 → 113.67
Vocabulary size	18,259

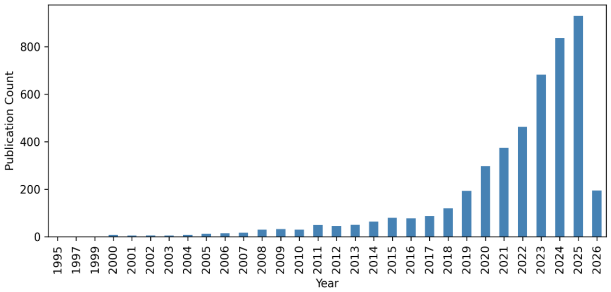

Fig. 1. Top 50 domain-vocabulary terms by frequency in the cleaned corpus, following a Zipfian distribution dominated by terms such as climate, model, and emission

C. Temporal Structuring

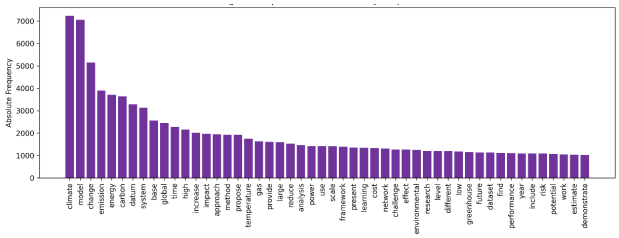

Fig. 2. Climate-related abstracts published per year, 1995-2026, showing a marked increase from the mid-2010s onward

Each abstract was tagged by submission year; records with missing dates were dropped. The corpus grows substantially over time (Figure 2), a compositional shift that is revisited in Section 3.

D. Topic Modelling

Topic modelling used Latent Dirichlet Allocation [4], with K ∈ {5, 10, 15} compared by coherence score:

$$P(w|d) = \sum_{k=1}^{k} P(w|z_k)P(z_k|d)$$

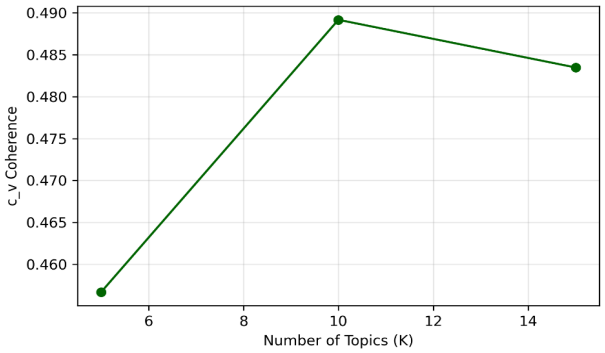

Fig. 3. Coherence versus number of topics; coherence peaks near K = 10

K = 10 was selected as the best balance of interpretability and granularity. Table 3 lists each topic's top keywords and assigned label.

E. Tone Metrics

Three lexicon-based tone metrics were defined, following established hedging and stance categories [5–7]: an Uncertainty Score (hedges: "may," "might," "suggests"), a Certainty Score (assertives: "demonstrates," "confirms"), and a Risk Intensity Score (urgency/severity terms: "critical," "extreme"), each computed as

$$Score = \frac{(count\ of\ lexicon\ terms)}{total\ tokens} \times 100$$

i.e., occurrences per 100 words. Document-level distributions of Uncertainty and Risk Intensity are heavily right-skewed (Figure 4), motivating the use of annual means for temporal analysis.

Table 3
Top keywords per LDA topic

Topic	Top keywords (label)
T1	temperature, global, warming, surface, earth, solar, ocean (physical climate science)
T2	datum, learning, base, method, framework, approach, performance (ML methodology)
T3	datum, forest, image, satellite, monitoring, wildfire, ecosystem (remote sensing)
T4	system, power, vehicle, network, emission, urban, grid (energy/engineering systems)
T5	extreme, event, risk, weather, impact, uncertainty, precipitation (extreme events & risk)
T6	system, dynamic, state, transition, equation, theory, tipping (dynamical systems)
T7	social, research, policy, human, impact, science, public (social/policy)
T8	emission, gas, methane, star, galaxy, infrared, greenhouse (astrophysical/GHG)
T9	gas, ice, material, flow, amoc, simulation, condition (physical climate / cryosphere)
T10	energy, carbon, emission, system, cost, consumption, power (energy/carbon systems)

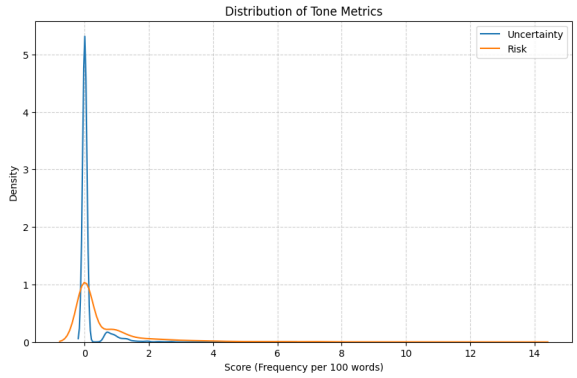

Fig. 4. Kernel density estimates of document-level Uncertainty and Risk Intensity scores

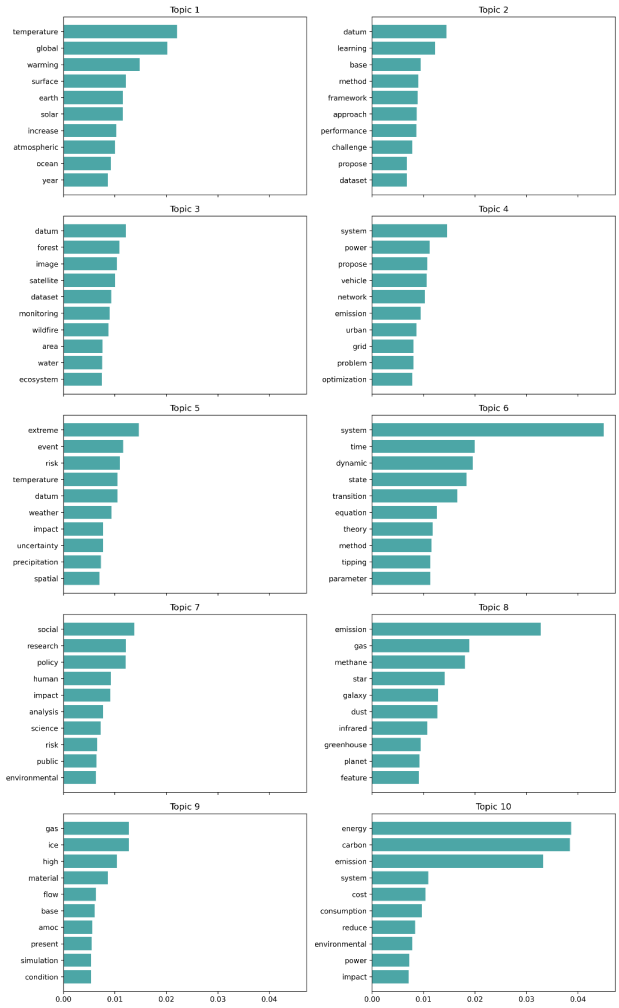

Fig. 5. Relative weight of the top ten keywords within each LDA topic (T1–T10)

Each panel corresponds to one topic and lists its highest-probability terms in descending order of weight. The length of each bar indicates how strongly that word characterizes the topic relative to the others in its top-10 list. Topics with one or two dominant, high-weight terms (e.g., T6 with "system," or T10 with "energy" and "carbon") have a sharply skewed distribution a single keyword carries much of the topic's identity. Topics with more evenly distributed weights (e.g., T1 or T7) suggest a more diffuse, multi-term thematic signature, where no single word dominates the topic's interpretation. This figure essentially validates the keyword-based labels assigned in Table 3 it shows the underlying weight structure that justifies calling T5 "extreme events and risk" or T10 "energy and carbon systems," by visualizing how concentrated or spread out each topic's defining vocabulary is.

F. Topic-Tone Integration and Temporal Analysis

Each document was represented jointly by its topic distribution $P(z_k|d)$ and its three tone scores, allowing per-topic mean tone profiles to be computed (Figure 6). Annual mean tone scores were regressed on year via OLS to estimate linear trends, dominant-topic (arg-max) proportions were tracked over time, and a first-difference change-point procedure in the spirit of PELT [14,15] was applied to the annual Risk Intensity series to detect structural breaks.

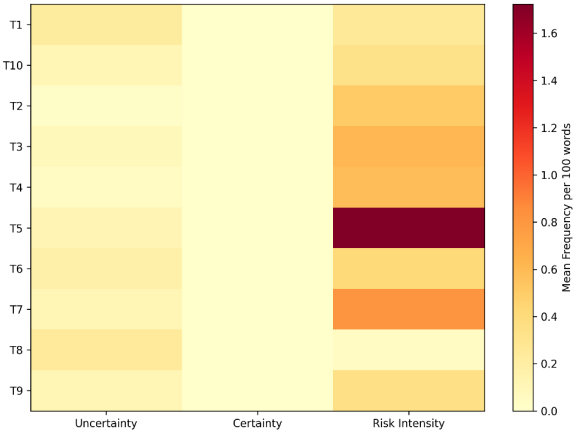

Fig. 6. Topic tone heatmap: mean Uncertainty, Certainty, and Risk Intensity per topic. Policy-related topics (T5, T7) show elevated risk intensity; modelling topics show higher uncertainty; certainty is near-flat across topics

The heatmap in Figure 6 shows that policy-oriented topics T5 and T7 in particular stand out with noticeably higher risk-intensity scores than the rest, while the certainty dimension is essentially flat across all ten topics. Uncertainty is somewhat higher for the more modelling- and simulation-heavy topics. We come back to these patterns in Section 3.

3. Results

A. Topic Evolution

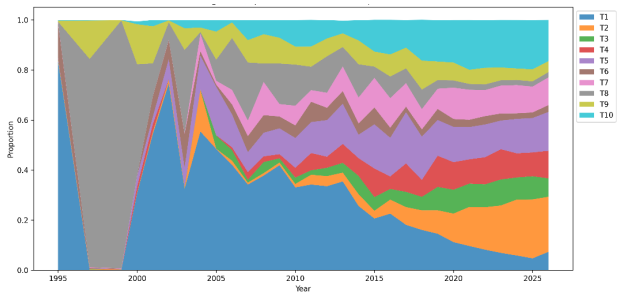

Fig. 7. Stacked relative proportions of the ten LDA topics (T1–T10) over time, 1995–2026

Figure 7 tells a fairly clear story about how the thematic

make-up of the corpus has shifted. In the earliest years pre-2000, where sample sizes are tiny Topic 1 (physical climate science) and Topic 8 (the astrophysical/greenhouse-gas cluster) account for most of the documents. From around 2005-2010 onward, T1's share steadily declines, while Topics 2 and 10 machine-learning methodology and energy/carbon systems, respectively grow substantially and become two of the most prominent topics in recent years. In other words, the corpus has gradually shifted away from "pure" physical climate science toward more applied, methods-driven, and engineering-oriented work.

Figure 8 backs this up at the level of individual words. The terms "impact" and "policy" barely appear before roughly 2005 and 2015, respectively, but become noticeably more common after those points right around when the topic-proportion shifts in Figure 7 start to show up. The two pictures line up well enough that we're reasonably confident the topic-level trends reflect real changes in vocabulary, not just an artefact of how the topic model happens to carve things up.

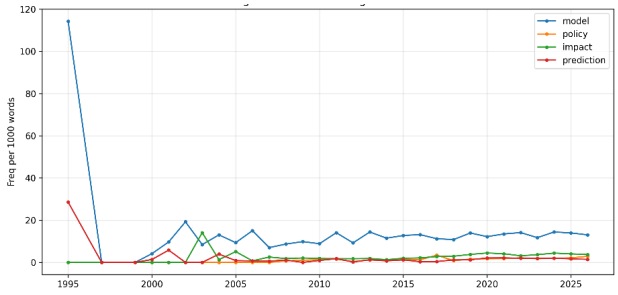

Fig. 8. Frequency (per 1,000 words) of selected lexical items "model," "policy," "impact," and "prediction" over time

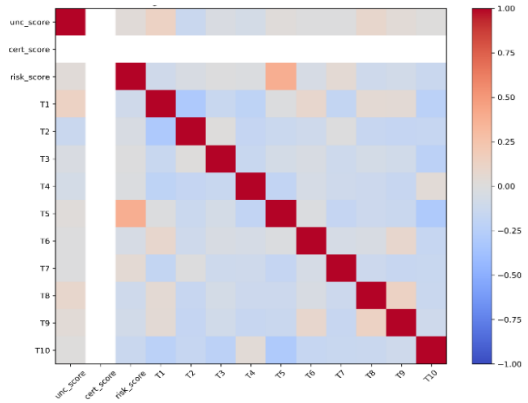

Fig. 9. Document-level correlation matrix between the Uncertainty score, Risk Intensity score, and the ten LDA topic proportions (T1–T10). The Certainty score row or column is shown blank owing to near-zero variance across documents

The correlation matrix in Figure 9 shows a clear positive relationship between the Risk Intensity score and Topic 5 (extreme events and risk) documents that load heavily on T5 tend to use noticeably more risk-related vocabulary, which is more or less what you'd expect given how that topic is defined. The Uncertainty score, by contrast, doesn't correlate strongly with any single topic; its relationship with the topic structure is weaker and more diffuse. The Certainty score essentially drops out of this picture entirely, since it shows almost no variation across documents something we discuss further in Section 4.

B. Dominant-Topic Distribution

Figure 10 shows that the corpus is far from evenly distributed across the ten topics. T10 (energy and carbon systems) and T2 (machine-learning methodology) are by far the most common dominant topics, with roughly 900 and 790 documents respectively, followed by T5 (extreme events and risk) at around 680. At the other end, T6, T8, and T9 are comparatively rare. This lines up with the topic-evolution pattern in Figure 7 energy/engineering, methodology, and risk-related research make up the bulk of the corpus, while the remaining topics represent smaller, more specialised pockets of the literature.

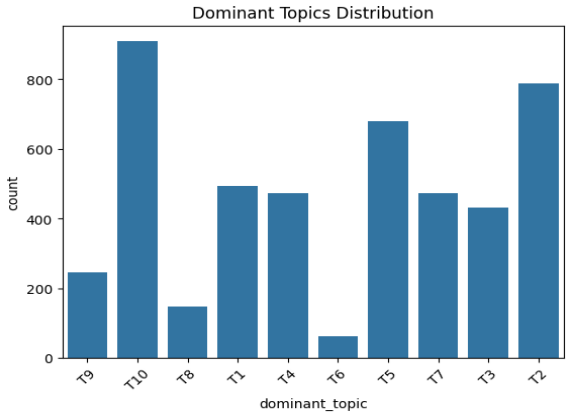

Fig. 10. Distribution of documents by dominant (arg-max) topic assignment

C. Temporal Trends in Tone

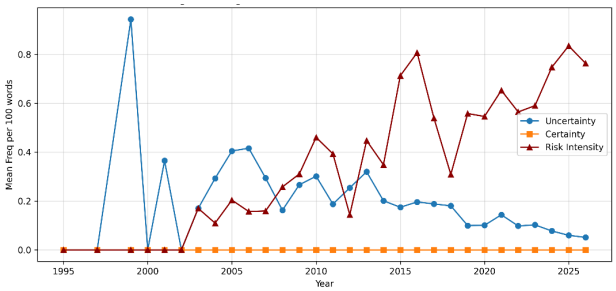

Fig. 11. Annual mean Uncertainty, Certainty, and Risk Intensity scores, 1995–2026

This is arguably the central result of the paper. Risk Intensity (Figure 11) climbs gradually over the study period, with the increase becoming much more visible after about 2010 and the highest sustained values appearing after 2015. Uncertainty tells a messier story: after an early spike around 1999-2001 almost certainly driven by the very small number of documents in those years it rises through the mid-2000s, peaks around 2006-2008, and then drifts down with continued ups and downs through to 2026. Certainty, meanwhile, stays close to zero and essentially flat across the entire period.

Table 4
Summary of OLS regression results for the three-tone metrics regressed on year (1995–2026)

Tone metric	Estimated trend direction	Statistical significance	Interpretation
Risk Intensity	Positive slope	Statistically significant	Gradual, moderate increase in risk-oriented language over time
Uncertainty	Variable / non-monotonic	Mixed	No steady decline; fluctuates across sub-periods
Certainty	Approximately zero slope	Not significant	No detectable temporal trend; series remains near zero throughout

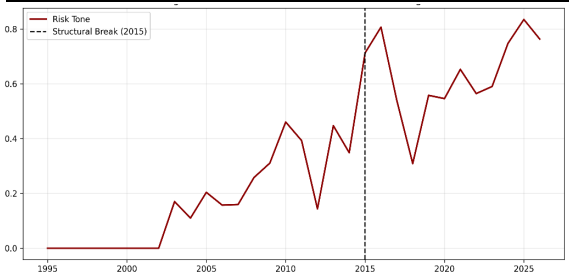

Fig. 12. Trend in the document-level Uncertainty score (unc_score) by year, with OLS regression line and 95% confidence band

The regression results in Figure 12 and Table 4 suggest a negative overall slope for Uncertainty across the full 1995–2026 window, but the confidence band is wide mostly because of how few documents and how much variance there is in the earliest years. Read alongside Figure 11, the fairest description of the uncertainty trajectory is "variable" rather than "steadily declining"; it doesn't move in one direction consistently, which is itself an interesting finding given that a simple narrative of "science getting more confident over time" would predict a steady downward trend.

D. Structural Break in Risk Framing

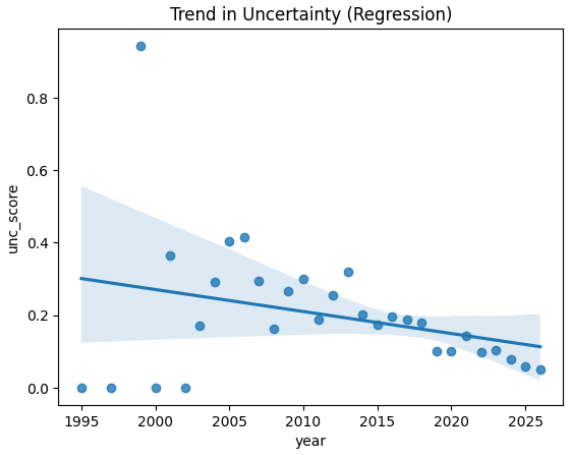

Fig. 13. First-difference-based structural break detection applied to the annual Risk Intensity series, with the identified break point in 2015

The change-point analysis in Figure 13 places an inflection point at roughly 2015 in the annual Risk Intensity series. From 1995 to about 2003, risk-tone scores sit close to zero. From around 2004, the series starts to creep upward; after 2010 it becomes noticeably more volatile and the upward movement steepens, with the highest values appearing after 2015. From 2015 onward, the series settles into a consistently higher range roughly 0.5 to 0.85 occurrences per 100 words suggesting that risk-related framing has become both more common and more pronounced over the past decade of the corpus.

4. Discussion

This study set out to determine whether the linguistic framing of climate-science abstracts has shifted measurably over three decades, and the evidence points to a modest but genuine evolution rather than a dramatic rhetorical turn. Risk-intensity language rose gradually across the study period, with an apparent acceleration after 2015 (Table 4, Figure 13), suggesting that climate research has come to emphasise severity and urgency somewhat more than in earlier decades. At the same time, uncertainty did not decline in the steady, monotonic fashion that a simple "growing scientific consensus" narrative would predict; instead, it fluctuated across the study period (Figures 11-12), while certainty remained close to zero and essentially flat throughout. Topic-level analysis added a further dimension to this picture: policy- and risk-oriented research areas (T5, T7) carried systematically higher risk intensity, while modelling-oriented areas carried more hedging (Figure 6), indicating that linguistic framing is shaped jointly by temporal context and subject matter rather than by time alone.

Part of this aggregate trend likely reflects compositional change rather than within-topic shift: publication volume has grown substantially (Figure 2), and topic prevalence has moved toward policy and risk-oriented themes (Figure 7) that already carry higher baseline risk intensity (Figure 6). The framework's topic-conditional tone estimates partially address this, but a fully topic-stratified longitudinal model would better separate temporal drift from thematic redistribution.

The non-monotonic uncertainty trend (Figures 11-12) runs counter to the expectation that a maturing field hedges progressively less. A plausible reading is that advances in modelling and data integration continually introduce new uncertainty even as older sources resolve, consistent with hedging as a dynamic, context-dependent resource. However, the lexicon-based metric captures only explicit hedges and may understate implicit caution conveyed through syntax; contextual language models could address this in future work.

The near-flat certainty profile (Figure 11) likely reflects a sparse assertive-term lexicon relative to hedging and risk vocabulary, consistent with the cautious register long documented in academic discourse. The 2015 structural break (Figure 13) is suggestive of external influences such as the IPCC AR5 cycle or the Paris Agreement, but this remains speculative, and first-difference change-point methods are sensitive to short-term noise [14], [15].

5. Conclusion

This study introduced a topic-tone analytical framework that combines Latent Dirichlet Allocation with custom lexicon-based metrics of uncertainty, certainty, and risk intensity to examine how the linguistic framing of climate-science abstracts

on arXiv has evolved between 1995 and 2026. The analysis revealed a gradual but statistically detectable increase in risk-oriented language, with an apparent acceleration after 2015, alongside a fluctuating rather than monotonically declining uncertainty signal and a persistently flat certainty profile. Topic-level analysis further showed that policy and risk oriented research areas carry systematically higher risk intensity, while modelling-oriented areas carry more hedging, indicating that linguistic framing depends jointly on when and what is written.

These findings do not support a narrative of growing "alarmism" in scientific writing; rather, they describe a field whose language is adjusting incrementally alongside its evolving scientific and societal context, while retaining its characteristic epistemic caution. The proposed framework is reproducible, interpretable, and domain-agnostic, offering a template for studying linguistic change in other scientific corpora. Future work should replicate this analysis on curated bibliographic databases, incorporate contextual language models to capture implicit hedging, apply topic-stratified longitudinal modelling to separate temporal and compositional effects, and test alternative change-point algorithms to validate the 2015 structural break.

6. Data and Code Availability

The dataset underlying this study is the arXiv metadata corpus, available at

https://www.kaggle.com/api/v1/datasets/download/Cornell-University/arxiv.

The analysis code is available at

https://colab.research.google.com/drive/1UShQzIJ5OEku6LhJ1MNqmHgrrdvSpMYA?usp=sharing.

References

[1] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, 2014.

[2] S. M. R. Naqvi *et al*., "Public health discussions on social media: Evaluating automated sentiment analysis methods," *JMIR Formative Research*, vol. 9, no. 1, Art. no. e57395, 2025.

[3] E. M. G. Younis *et al*., "Sentiment analysis in public health: A systematic review of the current state, challenges, and future directions," *PMC*, Art. no. PMC12226299, 2025.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] K. Hyland, "Writing without conviction? Hedging in science research articles," *Applied Linguistics*, vol. 17, no. 4, pp. 433–454, 1996.

[6] K. Hyland, *Hedging in Scientific Research Articles*. Amsterdam, The Netherlands: John Benjamins Publishing Company, 1998.

[7] K. Hyland, Metadiscourse: Exploring Interaction in Writing. London, U.K.: Continuum, 2005.

[8] H. Wang *et al*., "Exploring the global research trends of cities and climate change based on a bibliometric analysis," *Sustainability*, vol. 14, no. 19, Art. no. 12302, 2022.

[9] D. Zhang *et al*., "Bibliometric analysis and global research trends of climate change and cities studies for 30 years (1990–2021)," *Environment, Development and Sustainability*, 2023.

[10] S. Sarkar *et al*., "Exploring the research trends in climate change and sustainable development: A bibliometric study," *Heliyon*, 2024.

[11] K. E. Falahy, "Exploring climate change discourse on social media and blogs using a topic modeling analysis," *PMC*, Art. no. PMC11214360, 2024.

[12] S. Battiston *et al*., "Where and how machine learning plays a role in climate finance research," *Journal of Sustainable Finance & Investment*, 2024.

[13] Cornell University, *arXiv Metadata Dataset* [Data set]. Kaggle, 2025.

[14] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[15] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and Information Systems*, vol. 51, no. 2, pp. 339–367, 2017.