# GC Insights: Breaking the silos – leveraging NLP to encourage interdisciplinary interaction at the EGU

Jan Sodoge[1], Taís Maria Nunes Carvalho[1,2], and Mariana Madruga de Brito[1]

[1]Department of Urban and Environmental Sociology, UFZ-Helmholtz Centre for Environmental Research, 04318, Leipzig, Germany
[2]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Universität Leipzig, Leipzig, Germany

**Correspondence:** Jan Sodoge (jan.sodoge@ufz.de)

**Abstract.** Thousands of abstracts from various geoscience sub-fields are presented annually at the EGU General Assembly (GA), offering a rich resource for tracking scientific progress. However, rigid session groupings can limit cross-disciplinary exploration. Here, we show that participants focusing only on their broad disciplinary ~~session~~ divisions miss an average of 44 % of the 10 most relevant contributions. To break this compartmentalization, we propose using natural language processing (NLP), enabling the geoscience community to explore the full breadth of knowledge beyond traditional disciplinary boundaries.

## 1 Introduction

Each year, the European Geosciences Union (EGU) General Assembly (GA) gathers over ~~15~~20,000 geoscientists worldwide, with participation steadily increasing since its inception in 2004 (EGU, 2024). To organize this vast array of research, the ~~EGU GA is structured~~ abstracts are organized into 22 disciplinary ~~and a varying number of inter- and transdisciplinary sessions~~programme groups (PGs), alongside Inter- and Transdisciplinary Sessions (ITS), Education and Outreach Sessions (EOS), and other Union-wide Sessions. We hypothesize that this ~~compartmentalization~~ structure may inadvertently create knowledge silos, as EGU GA ~~attendants~~attendees might tend to focus on their own scientific ~~divisions, potentially~~ programme groups —also referred to as divisions[1] potentially missing relevant developments from other disciplines. As a result, they may ~~be only~~ only be exposed to ideas within their peer group, thereby reinforcing existing perspectives. This phenomenon mirrors the well-documented effects of filter bubbles and selective exposure to information observed on social media platforms (Robertson et al., 2023; Spohr, 2017). In scientific settings, such bubbles can hinder interaction between fields that could catalyze creativity and innovation (Kittur et al., 2019; Burt, 2004). While the idea of bursting such potential bubbles is thought to foster research progress (Portenoy et al., 2022), this phenomenon remains underexplored in the context of large scientific conferences~~like~~, such as the EGU GA.

Recent advancements in natural language processing (NLP) offer a promising solution to address this compartmentalization. Language models can extract structured insights from vast amounts of scientific documents, thereby revealing knowledge that would otherwise remain hidden within the sheer volume of scientific output (Yau et al., 2014; De Battisti et al., 2015; Chen

---

[1]We use the terms "divisions" throughout the text for clarity

et al., 2019; Sodoge and de Brito, 2024). For example, language models have been successfully used to map climate-change-related publications (Callaghan et al., 2020), moving towards understanding the meaning and context of language.

25    In this work, we investigate how the ~~current compartmentalization at EGU GA~~ compartmentalization at the EGU GA programme may generate filter bubbles. We then demonstrate how NLP can be used to organize knowledge from EGU GA abstracts beyond traditional disciplines, benefiting conference participants, organizers, and the broader geoscience community. Using the abstracts from the EGU GAs (2020-2024), we create a *textual cartography* of the geoscience landscape, aiming to: (1) provide an overview of geoscience research presented at the EGU GAs, (2) guide participants to relevant research across

30    disciplinary boundaries, and (3) assist in organizing conference sessions.

## 2    Methods and Data

Our approach consists of 5 steps, including: (1) collecting abstracts from EGU, (2) computing similarities between abstracts, (3) visualizing the abstracts in a 2-dimensional space, (4) running a simulation to estimate the hypothesized filter effect, and (5) comparing the coherence of divisions created through clustering to the existing division structures.

35    To demonstrate our approach, we collected abstracts from the past five EGU GA. For each GA, we scraped the abstracts and corresponding ~~session data~~division. After removing withdrawn abstracts, ~~we obtained a total of~~ the dataset comprised 77,911 contributions ~~presented in~~ across 22 disciplinary ~~sessions~~ divisions (2020: 17728; 2021: 13368; 2022: 12129; 2023: 15772; 2024: 18914) ~~. We focused only on disciplinary sessions for several reasons:~~ and 3,663 abstracts from the Inter- and Transdisciplinary Sessions (ITS). We have not considered the ITS abstracts as a single division, as (1) ~~the~~ their content and

40    number of ~~inter- and transdisciplinary sessions vary significantly from year to year~~sessions may vary significantly across years, making it difficult to analyze trends consistently; (2) ~~while many scientists identify themselves as "interdisciplinary," they rarely align with specific EGU interdisciplinary session, which tend to be highly specialized and narrowly focused; (3) the share of these sessions is generally low~~ they represent a small fraction of total submissions (around 0.5 % of all ~~submissions). Thus, concentrating on disciplinary sessions provides a greater comparability across years~~abstracts). Instead, we reassigned

45    the ITS abstracts to their co-organizing divisions (e.g., ITS and NH) to account for their disciplinary interlinkages.

~~A~~ We then used a pre-trained language model ~~was used~~ to generate a text embedding for each abstract. An embedding is a high-dimensional vector of numerical values ~~which represents~~ representing the text's semantic meaning. These reveal relationships between abstracts as texts with similar content have similar embeddings. The pre-trained model requires no additional preprocessing (e.g. stop-word removal), as it is designed to ignore irrelevant parts and focus on meaningful patterns

50    in the text. Specifically, we employed the *distilroberta-base* language model (Liu, 2019), which was trained on large text corpora, enabling it to capture text structure, meanings, and context. This model was selected for its ability to capture the entire abstract and robust performance across diverse language tasks (Naseer et al., 2021; Briskilal and Subalalitha, 2022).

In order to visualize the text embeddings and create a textual cartography of geoscience research, we applied the U-MAP (Uniform Manifold Approximation and Projection) dimensionality reduction method (McInnes et al., 2018). This technique

55    projects the high-dimensional text embeddings onto a 2-dimensional space while preserving the local and global structure of

the data. The result is a map where similar abstracts ~~cluster together, with related content~~ are positioned closer and distinct content farther apart.

To investigate the hypothesized filter bubble effect, we conducted a simulation by randomly selecting 5,000 abstracts from each EGU GA. For each abstract, we identified the 10 most similar abstracts presented in the same year using cosine similarity (Kenter and De Rijke, 2015). We then analyzed the ~~sessions~~ disciplinary divisions associated with these similar abstracts to calculate the proportion that belonged in the same ~~session versus different sessions~~ versus different divisions. A higher proportion of similar abstracts from different ~~sessions~~ divisions suggests that EGU GA participants may miss out on potentially relevant contributions by focusing exclusively on their own ~~sessions~~ division.

Finally, to evaluate how well the EGU GA abstracts were grouped, we compared the clustering quality when using the disciplinary ~~sessions~~ divisions against using the k-means clustering algorithm (Rodriguez et al., 2019), considering the same number of clusters as the number of disciplinary sessions (n=22). The average Silhouette coefficient (Dinh et al., 2019) was used to assess cluster coherence, where values close to 1 indicate well-separated clusters and negative values ~~suggest miss-assigned abstracts~~ indicate no coherent cluster structures. This comparison ~~enabled~~ allowed us to assess which method ~~more effectively clusters abstracts addressing similar issues~~ clusters abstracts more effectively that address similar topics.

## 3 Results

Figure 1a shows the research landscape of the EGU GA ~~research from 2020 to 2024~~ from 2020-2024, featuring a total of 77,911 abstracts. Each abstract is colored by their respective disciplinary ~~session~~ division, highlighting the relationships among them. For example, abstracts from the natural hazard (NH) and climate (CL) ~~sessions~~ divisions are spread across different clusters. A qualitative inspection of this map reveals not only such macro-level structures but also the similarities and differences on a more fine-grained level~~, beyond the broad disciplinary sessions~~. For instance, the contributions most similar to a research abstract on modelling farmers' irrigation preferences (Heilemann et al., 2024) include an abstract evaluating irrigation demand in the same case-study area (Fallah-Mehdipour and Dietrich, 2024) and an abstract on the modelling of global irrigation water demand (Beier et al., 2024). These were presented ~~at different sessions~~ in different divisions and thus would potentially be missed if an attendee only consulted one division and the sessions are not co-organized.

Our findings confirmed the presence of a filter bubble effect, with varying degrees across the EGU GA ~~disciplinary sessions~~ divisions. On average, participants who restrict their attention to abstracts within their own ~~session~~ division potentially miss 44% of the 10 ~~contributions most relevant to them~~ abstracts most similar to their own abstract. This percentage varies depending on the ~~session the participant is considering. Sessions~~ division considered. Divisions that are rather separated, such as Atmospheric Sciences (AS) and Solar-Terrestrial Sciences (ST), ~~reduce the likelihood of missing related research~~ have a lower likelihood for attendees to miss relevant contributions (Fig. 1b). Conversely, ~~sessions~~ divisions with contributions widespread across different topics, such as Geosciences Instrumentation & Data Systems (GI) and Nonlinear Processes in Geosciences (NP), reduce the share of relevant contributions covered by the own ~~session.~~ division. We found only a weak correlation (Spearman's Rho = 0.38) between the number of abstracts of each division and the previously discussed share of relevant presentations. This

suggests the impact of the size of a division on a potential filter bubble effect is limited and unlikely to strongly influence our findings.

Adding the abstracts from the ITS sessions to their co-organizing divisions significantly increased the number of missed relevant contributions in the divisions of: AS, CR, BG, HS, and SSS. This may be due to the lower thematic coherence of the ITS abstracts, which makes their text less likely to align with that of the disciplinary divisions that co-organize the ITS. Yet, these findings underscore the value of the ITS sessions themselves: the described lower thematic coherence is important for bringing together contributions from diverse fields, facilitating cross-disciplinary exchange.

When comparing different ways of organizing the abstracts, we found that grouping ~~the abstracts using disciplinary sessions~~ them by disciplinary divisions yields an average silhouette coefficient of -0.17. In contrast, applying the k-means clustering algorithm results in an average silhouette coefficient of 0.39. This suggests that using a statistical clustering approach to group ~~sessions produces significantly~~ abstracts produces more coherent clusters ~~, where abstracts addressing similar topics or methods are more likely to be grouped together~~ than the divisions.

## 4  Discussion

Abstracts presented at EGU GA every year provide a snapshot of current geoscientific research. Yet, the increasing number of conference ~~contributions~~abstracts, over 18,000 in 2024, makes it challenging for participants and organizers to keep track of relevant contributions. By creating a textual cartography of more than 77,000 abstracts presented in the last ~~5~~ five GAs, we show here how NLP can help capture the broad range of subjects discussed at EGU GA, as well as identify relevant contributions outside of the field or discipline the participant identifies with.

Our results indicate the presence of a bubble effect, where EGU GA ~~attendants~~ attendees may miss contributions addressing similar problems from a different perspective presented outside their own ~~sessions~~division. This effect is evident, particularly for abstracts submitted to Nonlinear Processes in Geosciences (NP) and Geosciences Instrumentation & Data Systems sessions (GI). As a result, participants ~~restricting~~ who restrict their participation to ~~these sessions~~ their own divisions may be exposed only to content they are familiar with, due to the restrictive nature of session compartmentalization.

To tackle this, we propose an online application designed to recommend EGU GA ~~participant's other contributions~~participants' other contributions, independent of the ~~session~~ divisions to which they were submitted~~to~~. Using interactive visualizations such as the map in Fig 1a, EGU GA participants can identify relevant research contributions beyond their immediate field (e.g. addressing the same problem but using different methods, or using the same method for different problems). This tool aims to facilitate interdisciplinary exploration by surfacing relevant contributions that attendees might not discover if they restrict their participation to a particular session or division. This can foster interdisciplinary connections leading to innovative work and higher productivity (Portenoy et al., 2022; Specht and Crowston, 2022). During EGU GA 2024, we tested a prototype of such an application (Sup. Fig. 1) and received positive feedback on the relevance of the suggested talks.

## 5 Limitations and future research

In our analysis of filter bubbles, we recognize that the assumption that GA attendees attend only sessions on their own division is an oversimplification. In reality, participants engage with ITS and neighboring divisions. However, due to the large volume of abstracts and sessions per division (for example, in 2024, the median number of abstracts per disciplinary division was 638, with a maximum of 2,681), manually identifying all relevant contributions across divisions is infeasible. Another key limitation is that our scraped dataset includes only co-organizing groups, not the primary organizers, which may affect the results, particularly for ITS sessions, where the roles of each division vary.

Looking ahead, we plan to conduct a more in-depth comparison of models and consider fine-tuning existing approaches to improve accuracy. Also, to address these limitations, we are currently conducting a survey to more explicitly measure attendee behavior and evaluate the assumptions underlying our analysis by incorporating user feedback. This will allow us to assess whether the suggested contributions are perceived as meaningful by EGU GA attendees.

## 6 Conclusion

We showed that clustering the abstracts ~~at EGU GA compared to the disciplinary sessions achieved a significantly better clusteringi.e.~~ from the EGU GA using k-means, compared to relying on disciplinary divisions, achieved a better clustering, i.e., more semantically coherent sessions. While these results are ~~indeed provocativeand radical~~provocative, our intent ~~here~~ is not to suggest ~~the replacement of sessions~~ replacing divisions or other formal groupings. Large conferences, such as EGU GA~~require such organizational structures in order~~, require these organizational structures to coordinate and design a coherent program. ~~Instead,~~ We also recognize the role of authors in selecting the most appropriate session, often influenced by community ties. Hence, the computational approach suggested here is ~~rather intended as a co-pilot~~seen more as a tool in a co-design process, helping organizers and participants ~~to~~ make more informed decisions when ~~facing~~ faced with a vast amount of information. In fact, the authors of this study are collaborating with the EGU Programme Committee and the Copernicus Conference Manager to experiment with additional tools that can be integrated into the conference system and enhance the attendee experience.

Beyond the application provided here, the developed textual cartography opens multiple new research avenues for studying geoscientific research. Our results ~~imply new possibilities in~~ suggest new possibilities for assessing trends across different ~~geosciences sub-fields~~ geoscience subfields over time, identifying emerging research areas, and tracking shifts in scientific focus that ~~may~~ might otherwise go unnoticed. By "bursting bubbles" that isolate information, our method and online application reveal the interconnectedness of geosciences research, supporting collaboration beyond traditional boundaries.

*Code and data availability.* The code and data for computing the embedding, comparing similarity, and creating the figures for this publication are documented in https://git.ufz.de/sodoge/egu-abstract-embeddings

An interactive visualization of the data is available at https://taiscarvalho.github.io/egu-umap-viz/

## Appendix A

### A1

*Author contributions.* JS: conceptualisation, methodology, investigation, data curation, formal analysis, writing (original draft preparation, review, and editing), project administration, visualisation. TNC: conceptualisation, methodology, visualization, data curation. MMdB: conceptualisation, writing (original draft preparation, review and editing).

*Competing interests.* The contact author has declared that none of the authors has any competing interests.

# References

Beier, F., Heinke, J., Bodirsky, B. L., Müller, C., Ostberg, S., Karstens, K., Abrahao, G., Popp, A., and Lotze-Campen, H.: Multiple cropping in global-scale Land-Use Models and the role of Irrigation, Tech. rep., Copernicus Meetings, 2024.

Briskilal, J. and Subalalitha, C.: An ensemble model for classifying idioms and literal texts using BERT and RoBERTa, Information Processing & Management, 59, 102 756, 2022.

Burt, R. S.: Structural holes and good ideas, American journal of sociology, 110, 349–399, https://doi.org/https://doi.org/10.1086/421787, publisher: The University of Chicago Press, 2004.

Callaghan, M. W., Minx, J. C., and Forster, P. M.: A topography of climate change research, Nature Climate Change, 10, 118–123, https://doi.org/https://doi.org/10.1038/s41558-019-0684-5, publisher: Nature Publishing Group UK London, 2020.

Chen, H., Wang, X., Pan, S., and Xiong, F.: Identify topic relations in scientific literature using topic modeling, IEEE Transactions on Engineering Management, 68, 1232–1244, https://doi.org/10.1109/TEM.2019.2903115, publisher: IEEE, 2019.

De Battisti, F., Ferrara, A., and Salini, S.: A decade of research in statistics: A topic model approach, Scientometrics, 103, 413–433, https://doi.org/https://doi.org/10.1007, publisher: Springer, 2015.

Dinh, D.-T., Fujinami, T., and Huynh, V.-N.: Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient, in: Knowledge and Systems Sciences: 20th International Symposium, KSS 2019, Da Nang, Vietnam, November 29–December 1, 2019, Proceedings 20, pp. 1–17, Springer, https://doi.org/https://doi.org/10.1007/978-981-15-1209-4_1, 2019.

EGU: List of General Assemblies, https://www.egu.eu/meetings/general-assembly/meetings/, 2024.

Fallah-Mehdipour, E. and Dietrich, J.: Evaluating irrigation demand forecasts from S2S/agro-hydrological modelling with field experiments in Northern Germany in the context of farmer decision support, Tech. rep., Copernicus Meetings, 2024.

Heilemann, J., Nagpal, M., Werner, S., Klassert, C., Klauer, B., and Gawel, E.: More Droughts, More Irrigation? Modeling the Adaptive Behavior of German Farmers to Hydrometeorological and Socioeconomic Change, Tech. rep., Copernicus Meetings, 2024.

Kenter, T. and De Rijke, M.: Short text similarity with word embeddings, in: Proceedings of the 24th ACM international on conference on information and knowledge management, pp. 1411–1420, https://doi.org/https://doi.org/10.1145/2806416.2806475, 2015.

Kittur, A., Yu, L., Hope, T., Chan, J., Lifshitz-Assaf, H., Gilon, K., Ng, F., Kraut, R. E., and Shahaf, D.: Scaling up analogical innovation with crowds and AI, Proceedings of the National Academy of Sciences, 116, 1870–1877, https://doi.org/https://doi.org/10.1073/pnas.1807185116, publisher: National Acad Sciences, 2019.

Liu, Y.: Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692, https://doi.org/https://doi.org/10.48550/arXiv.1907.11692, 2019.

McInnes, L., Healy, J., and Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426, https://doi.org/https://doi.org/10.48550/arXiv.1802.03426, 2018.

Naseer, M., Asvial, M., and Sari, R. F.: An empirical comparison of bert, roberta, and electra for fact verification, in: 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 241–246, IEEE, https://doi.org/10.1109/ICAIIC51459.2021.9415192, 2021.

Portenoy, J., Radensky, M., West, J. D., Horvitz, E., Weld, D. S., and Hope, T.: Bursting scientific filter bubbles: Boosting innovation via novel author discovery, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1–13, https://doi.org/https://doi.org/10.1145/3491102.3501905, 2022.

Robertson, R. E., Green, J., Ruck, D. J., Ognyanova, K., Wilson, C., and Lazer, D.: Users choose to engage with more partisan news than they are exposed to on Google Search, Nature, 618, 342–348, https://doi.org/https://doi.org/10.1038/s41586-023-06078-5, publisher: Nature Publishing Group UK London, 2023.

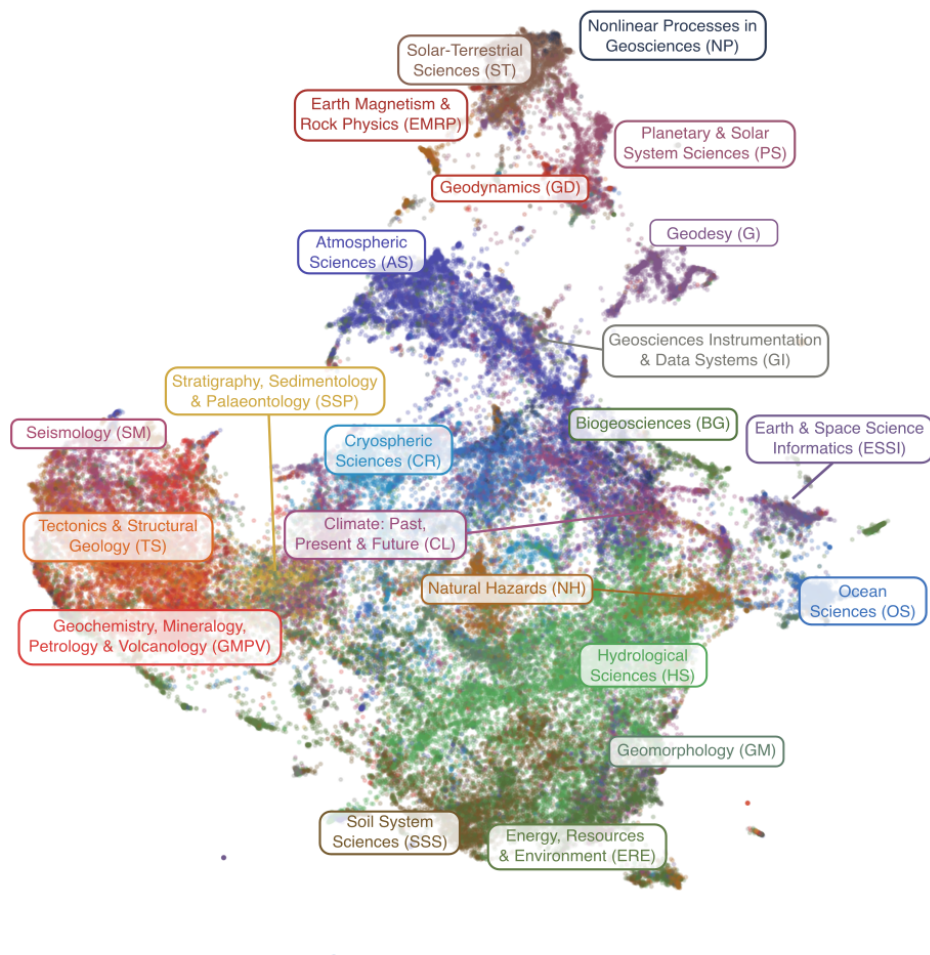
195 Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., and Rodrigues, F. A.: Clustering algorithms: A comparative approach, PLOS ONE, 14, e0210 236, https://doi.org/10.1371/journal.pone.0210236, 2019.

Sodoge, J. and de Brito, M. M.: Computational Social Sciences in Human–Water Research, in: Elgar Encyclopedia of Water Policy, Economics and Management, pp. 50–52, Edward Elgar Publishing, 2024.

Specht, A. and Crowston, K.: Interdisciplinary collaboration from diverse science teams can produce significant outcomes, PLoS One, 17,
200 e0278 043, https://doi.org/https://doi.org/10.1371/journal.pone.0278043, publisher: Public Library of Science San Francisco, CA USA, 2022.

Spohr, D.: Fake news and ideological polarization: Filter bubbles and selective exposure on social media, Business information review, 34, 150–160, https://doi.org/https://doi.org/10.1177/0266382117722446, publisher: SAGE Publications Sage UK: London, England, 2017.

Yau, C.-K., Porter, A., Newman, N., and Suominen, A.: Clustering scientific documents with topic modeling, Scientometrics, 100, 767–786,
205 https://doi.org/https://doi.org/10.1007/s11192-014-1321-8, publisher: Springer, 2014.

**A** Textual cartography of the geoscience landscape in the 2020-2024 EGU GAs

**B** Share of relevant contributions in the same disciplinary session

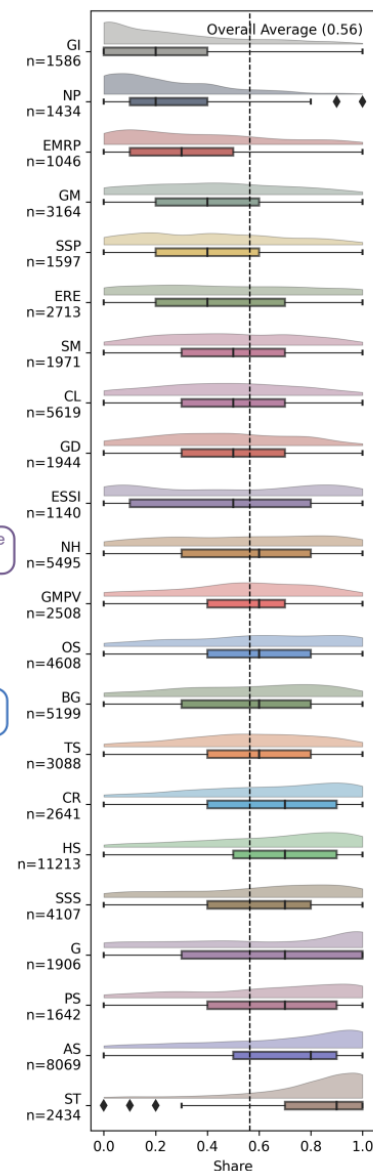



**Figure 1.** A: Textual cartography of 77,911 abstracts ~~presented at~~ from the EGU ~~GA~~ GAs 2020-2024. Each dot represents an abstract, where a shorter distance between dots indicates a greater similarity in content. B: Assessment of the potential filter bubble effect due to the ~~session~~divisions' ~~s~~compartmentalization. Each observation represents a randomly selected abstract. Its value indicates the proportion of the 10 most similar abstracts ~~that belong to~~ in the same ~~disciplinary session~~ division as the selected abstract. Higher values suggest a stronger tendency for relevant abstracts to be concentrated within the same ~~session~~division. *n* denotes the total number of abstracts in each session between 2020-2024.

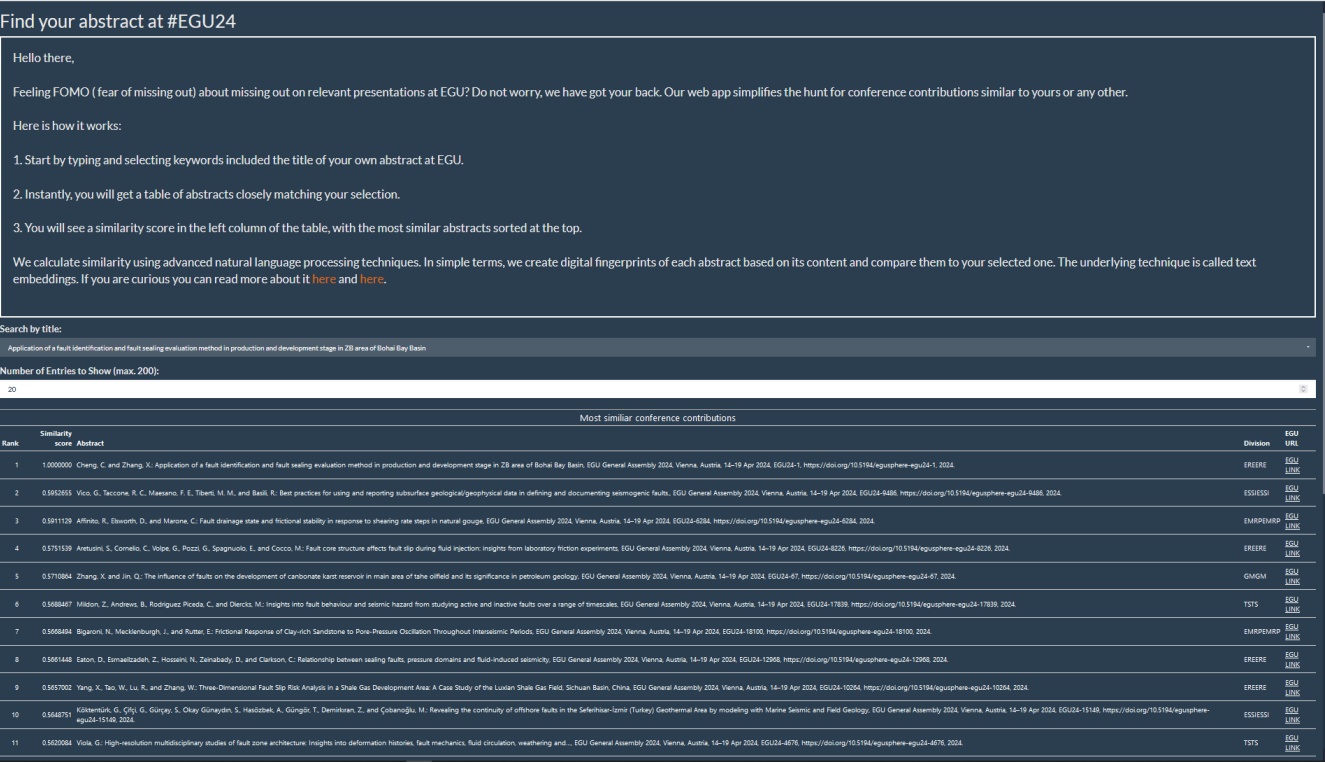



**Figure A1.** Screenshot of a web application for EGU GA participants to explore relevant abstracts. Participants can search for a particular submission and receive a list of the most similar abstracts. Additionally, an interactive map can be displayed similar to Fig. 1A.

### Create text embeddings for EGU abstracts

Pre-trained language model
distilroberta-base

### Build textual cartography of geoscience research

U-MAP dimensionality reduction

### Investigate the bubble effect

Compute the cosine similarity to identify the 10 most similar abstracts

### Cluster the abstracts using k-means

Cluster 1
Cluster 2
Cluster 3
Cluster n

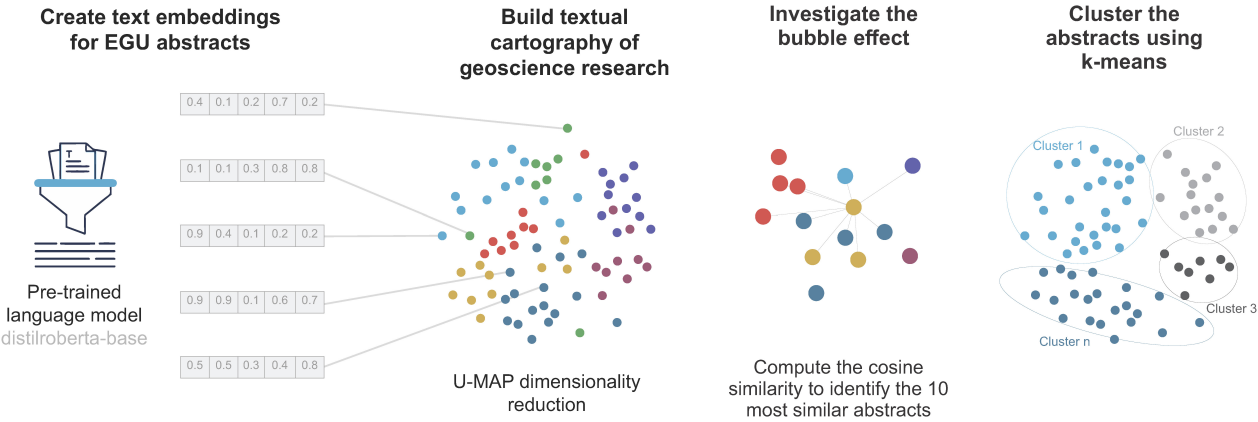



**Figure A2.** Research approach which consists of 5 steps, including: (1) collecting abstracts from EGU, (2) computing similarities between abstracts, (3) visualizing the abstracts in a 2-dimensional space, (4) running a simulation to estimate the hypothesized filter effect, and (5) comparing the coherence of divisions created through clustering to the existing division structures.