

AN OPEN ETL AND RAG PIPELINE FOR UNIFIED ACCESS TO HETEROGENEOUS CLIMATE DATA

Dávid Chomo¹, Jaromír Landa¹

¹Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1,
613 00 Brno, Czech Republic

DOI: https://doi.org/10.11118/ecos2026_005

ABSTRACT

Climate data are dispersed across numerous portals in incompatible formats, presenting a substantial barrier for researchers and analysts working on climate change and sustainability. This paper presents an open system that automates the processing of climate datasets end-to-end. The system covers the entire workflow, from the downloading of raw files through to format detection (NetCDF, GRIB, GeoTIFF, CSV, HDF5, and ZIP archives), and transformation into a text description with statistical characteristics. The system also creates vector embeddings and support natural natural language querying. The ETL pipeline is orchestrated by Dagster, the vector store is provided by Qdrant with an HNSW index, and response generation takes place in a multi-level retrieval-augmented generation (RAG) pipeline combining weighted Reciprocal Rank Fusion, Maximal Marginal Relevance diversification, and a Corrective RAG validation layer. The integration of a web interface, constructed on the foundations of Vue.js 3 and FastAPI, serves to facilitate the accessibility of this functionality to non-technical users. The system achieved an average composite score of 91% on a set of ten queries, with all tests passing the 80% threshold. Following a series of optimizations, the average query latency decreased from 120 to 180 seconds to approximately 26 seconds. The primary contribution of this work lies in the integration of components that currently function independently within existing climate RAG systems, thereby reducing the barriers to accessing standardized climate data for sustainability research.

Keywords: ETL pipeline, climate data, vector database, retrieval-augmented generation, semantic search, Qdrant, automation, Dagster, open data, embeddings

JEL Code: C81, C88, Q54



This article is licensed under CC BY-SA 4.0 Attribution-ShareAlike 4.0 International license.
<https://creativecommons.org/licenses/by-sa/4.0/deed.en>

1 INTRODUCTION

Climate change represents one of the most significant challenges of the 21st century, and any quantitative analysis of its manifestations or economic impacts depends on standardized climate data. However, this data is dispersed across numerous portals and stored in incompatible formats, which hinders its processing and utilization. The most significant sources comprise reanalyses, such as ERA5 (Hersbach *et al.*, 2020), CMIP6 model simulations (Eyring *et al.*, 2016), and observational datasets, including E-OBS (Cornes *et al.*, 2018) and CRU TS (Harris *et al.*, 2020). Each source is characterized by its own temporal granularity, spatial resolution, and metadata. Many portals require prior registration, and the data is often accessible only through specialized APIs. For researchers in the fields of climate economics, energy policy, or adaptation strategies, a significant portion of their project is spent on registration on portals, downloading data, converting formats, and extracting relevant variables, even before they can proceed to the actual analysis.

Rolnick *et al.* (2022) identified machine learning as a key tool for future climate research; however, easy access to input data remains a prerequisite. Existing open-source tools address only specific steps. Climate Data Operators enable operations on NetCDF and GRIB files, the GDAL library handles the conversion of georeferenced formats, and the Python library xarray (Hoyer and Hamman, 2017) facilitates work with N-dimensional arrays. These tools do not, however, share a common runtime; combining them requires substantial expertise, and their command-line orientation excludes users without programming skills.

The ETL (Extract, Transform, Load) pattern has been identified as a suitable solution for processing heterogeneous sources (Vassiliadis, 2009). This methodology involves the initial extraction of data, followed by its transformation (cleaning, normalization and enrichment), and subsequent loading into the target system. The alternative ELT model, which moves the transformation to the target repository (Kimball and Caserta, 2004), is not practical for climate data because the NetCDF and GRIB binary formats cannot be stored directly in a vector database. Instead, they must first be divided into spatial-temporal blocks, transformed into text descriptions, and only then embedded. Frameworks such as Apache Airflow, Prefect, or Dagster are considered suitable for orchestrating the ETL pipeline, with the latter using a declarative software-defined assets paradigm that reduces configuration requirements compared to traditional DAG-based workflow systems (Deelman *et al.*, 2015).

The challenge of efficient search within large volumes of processed data is addressed by techniques based on vector embeddings—compact representations that preserve the essential features of the content. The Retrieval-Augmented Generation (RAG) technique, as formally defined in Lewis *et al.* (2020), combines vector search in a knowledge base with answer generation by a large language model. The development of specialized systems for the climate domain is an area of current research. ChatClimate (Vaghefi *et al.*, 2023) provides responses to inquiries regarding IPCC reports by employing GPT-4 with retrieval, while ClimateGPT (Thulke *et al.*, 2024) leverages ongoing pretraining from the Llama-2 model on climate texts. Furthermore, GeoGPT-RAG (Huang *et al.*, 2025) is designed for geoscience publications. A common limitation of these systems is that they are only compatible with text documents, rather than raw raster climate data. Despite the fact that the Spatial-RAG system (Yu *et al.*, 2025) extends RAG with spatial filtering, it is limited to the GeoJSON format and is component-closed.

The objective of this paper is to design and implement an end-to-end system that consolidates heterogeneous climate data sources into a common format and makes them accessible to users without programming skills via a web interface that supports natural language queries. The system has been designed to encompass the entire chain, from the downloading of raw files through to

automatic format detection, transformation, generation of text descriptions, and creation of vector embeddings, culminating in final RAG querying. The work poses three specific research questions: firstly, how to design an ETL pipeline for heterogeneous climate data formats; secondly, how to build a web interface with automatic update orchestration; and thirdly, whether vector embeddings derived from text descriptions of climate chunks are sufficiently accurate for semantic search over a real-world collection.

From the perspective of sustainability and climate policy, easier access to standardized data has a direct impact: a faster cycle between the formulation of a research question and its answer, easier replicability of analyses, and the opening up of climate resources to non-traditional users, such as students, journalists, local governments, or small analytical groups without their own data infrastructure. The work thus contributes to the ongoing discourse on open data infrastructures as a technical prerequisite for sustainability research.

The article is structured as follows. The section entitled “Methodology and Data” provides a detailed description of the research framework employed, the architectural design of the proposed system, the catalogue of the input dataset, and the evaluation methodology. The “Results” section presents a quantitative assessment of the RAG system’s quality and the performance characteristics of the ETL pipeline. The Discussion and Conclusions chapter provides an interpretation of the results within the context of existing climate RAG systems, identifies limitations, and outlines directions for future work.

2 METHODOLOGY AND DATA

This work follows the Design Science Research (DSR) methodology of Hevner *et al.* (2004), which is designed for the creation and evaluation of IT artefacts. The framework fits this work because the output is not only an analysis of the current state but primarily the design and implementation of a software artefact addressing the identified problem. The process was divided into three phases: firstly, an analysis of existing solutions was conducted through a literature review; secondly, a modular design and implementation of the system was undertaken; and thirdly, a final quality evaluation was made on a running endpoint.

The analysis is initiated with a catalogue comprising 233 records, which collectively represent climate datasets derived from diverse sources. The catalogue is divided into three categories of data. The ERA5 (Hersbach *et al.*, 2020), ERA5-Land, and CERRA combine observations with a physical model, thereby ensuring consistent coverage of the entire atmosphere. CMIP6 (Eyring *et al.*, 2016) and EURO-CORDEX model simulations are utilized for the projection of future climate scenarios. The E-OBS (Cornes *et al.*, 2018) and CRU TS (Harris *et al.*, 2020) observational datasets are based on station measurements. The datasets are available in six main formats (NetCDF, GRIB, GeoTIFF, CSV, HDF5, and ZIP archives) and vary in dimensionality, spatial and temporal granularity, and access methods. Of these 233 records, 167 are currently fully ingested end-to-end, while the remaining 66 are available at the metadata embedding level from the zero phase.

A five-phase classification system was designed for the systematic processing of heterogeneous sources. Phase 0 processes all 233 original records via text embeddings from the catalog metadata without downloading data. Subsequent phases cover 98 sources with direct HTTP downloads, 25 sources requiring registration, 71 sources accessible via API portals with seven specialized adapters (Copernicus CDS, ESGF, NOAA, NASA Earthdata, Marine Copernicus, EIDC, and NCAR RDA), and 39 sources without automated access, which remain at the metadata level. Point identifiers in the

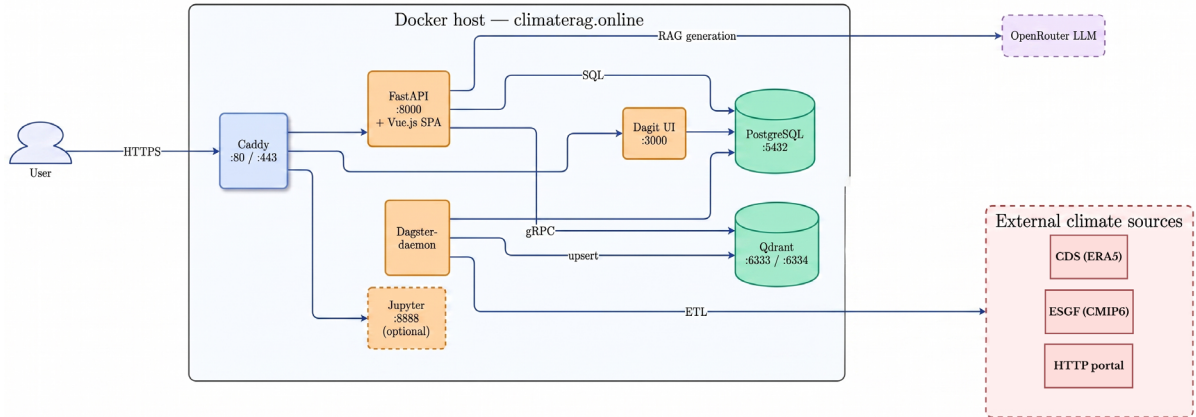


Fig. 1: The ClimateRAG system architecture. (Chomo, 2026)

vector database are generated deterministically from a UUID5 based on the source ID and chunk key, so that a repeated upsert from a higher phase overwrites the previous metadata representation without duplication.

The proposed ClimateRAG system is a distributed application with six main Docker services plus one optional service for interactive analysis. Figure 1 illustrates the overall architecture and interconnections of the components.

The entry point to the system is a Caddy reverse proxy with automatic HTTPS certificates. FastAPI handles the backend for the REST API and frontend serving, using asynchronous request processing. The ETL pipeline is orchestrated by Dagster using the software-defined assets paradigm, with scheduling via cron expressions and sensors for detecting updates. Relational metadata and processing statuses are stored in PostgreSQL, while vector embeddings are stored in Qdrant with an HNSW index (Malkov and Yashunin, 2020) with parameters $m = 16$ and $ef_construct = 100$. The frontend is a single-page application in Vue.js 3 with nine views covering resource management, update scheduling, a chat with RAG responses, and monitoring of ETL runs.

The xarray library (Hoyer and Hamman, 2017) is used to load and transform climate grids, with generator-based streaming, which ensures that files hundreds of megabytes in size are never loaded entirely into memory but are instead chunked progressively according to their granularity (90 time steps and 40×40 spatial cells). For each spatial-temporal chunk, descriptive statistics are calculated, and a text description is generated containing the variable name, units, geographic area, time interval, and statistical characteristics according to CF Conventions. This text description is then encoded by the BAAI/bge-large-en-v1.5 model (Xiao *et al.*, 2023) into a 1024-dimensional vector and stored in a Qdrant collection with metadata in the payload. The choice of text embedding over a direct numerical vector is crucial because a purely numerical representation cannot distinguish temperature from precipitation—the same number 15 can mean both 15 °C and 15 mm.

The query pipeline combines several stages. LLM-driven query analysis extracts structured filtering attributes (dataset, variable, hazard, time window) while simultaneously generating variants of the expanded query. For each variant, vector search is performed in Qdrant via batch gRPC calls. The results are merged using the weighted Reciprocal Rank Fusion method (Cormack *et al.*, 2009), followed by diversification using the Maximal Marginal Relevance algorithm (Carbonell and Goldstein, 1998) with parameter $\lambda = 0.55$. The Corrective RAG layer uses a fast

LLM call to assess whether the found chunks are sufficient for a response, and if the results are insufficient, it rewrites the query and repeats the search. In the current version, the response is generated via the OpenRouter service using the Claude Sonnet 4.6 model.

A custom deterministic script was designed for evaluation, which sends test queries to a running endpoint and evaluates the response against expected facts using regular expressions and set-overlap, without any LLM involved in the scoring. The script measures five metrics: Context Relevance, Faithfulness, Answer Correctness, Retrieval Precision@5, and Source Diversity. The first four form a weighted composite score. Source Diversity is measured but not included in the composite because, for queries with a single authoritative source, forced diversity would degrade it without contributing to the quality of the answer. The script runs on a set of ten queries. These tests do not reflect what would “suit” the system, but rather questions that a climatologist might actually ask. The originally considered RAGAS framework (Es *et al.*, 2024) with an LLM-as-judge approach was not used because, in the specific climate domain, it is not guaranteed that the judge understands terms such as SPEI or lwe_thickness, and the variance between runs would complicate iterative optimization.

3 RESULTS

The catalogue currently features 167 sources (out of the original 233 records) that are ingested automatically from start to finish. The Qdrant vector collection contains around 1.53 million points with 1,024-dimensional embeddings. The HNSW index enables sub-second search queries, even with this volume of data.

The quality of the RAG system was measured iteratively. The baseline version, which combined HNSW search, cross-encoder reranking and generation via a large language model, achieved an average composite score of 78%. However, there were several major issues: a cross-domain query on the relationship between drought and wildfires failed due to the retrieval system’s inability to return data from multiple thematic areas simultaneously; a CO₂ query was confused by the variables ,co2_concentration‘ and ,tcco‘ (total column carbon monoxide) in the embedding space; and latency hovered around 100 seconds per query.

The optimization process was executed across a total of eleven iterations, with each alteration being appraised utilizing the same test suite. Architecturally, several substantial modifications were introduced: multi-agent query decomposition generating three alternative variants and topic-focused variants for multi-topic queries; batch gRPC search instead of sequential calls; weighted RRF with primary weight 1.0, variant weight 0.6 and topic-focused query weight 0.9; MMR diversification with $\lambda = 0.55$; conditional data boost for multi-topic queries; a system prompt modification permitting linking of data to the query even with partial results; and a Corrective RAG validation layer. The cross-encoder reranker was eliminated from the primary processing pipeline due to its extended execution time of 40 to 50 seconds per query on the CPU, and its negligible impact on the composite score, contributing less than 1 percentage point. However, it was retained within the code as an optional fallback mechanism. The final deterministic run with a temperature of 0.0 achieved an average composite score of 91% and a success rate of 10/10, indicating that all tests passed above the 80% threshold.

Table 1 presents a comparison of the mean values of individual metrics between the baseline and the final deterministic run.

Tab. 1 Comparison of metrics is made between the baseline and the final deterministic run (Chomo, 2026)

Metric	Baseline	Final	Change (pp)
Context Relevance	89%	99%	+10
Faithfulness	86%	97%	+11
Answer Correctness	79%	81%	+2
Retrieval Precision@5	52%	86%	+34
Source Diversity	73%	74%	+1
Composite	78%	91%	+13

The most significant enhancement was observed in Retrieval Precision@5, with an increase of 34 percentage points. This finding corroborates the hypothesis that the majority of the observed quality gain was attributable to enhanced retrieval of pertinent segments, rather than to enhanced answer generation. A Context Relevance score approaching 100% indicates that, for the majority of tests, all returned chunks are deemed relevant. The accuracy of the answers remained at 81%, which can be attributed to the fact that the deterministic grader requires the exact occurrence of keywords in the answer, and the model occasionally uses a synonym (for example, “warm” instead of “warmest”), which counts as a mismatch even though the answer is semantically correct. The actual answer correctness is therefore higher than the nominally measured 81%. Concurrently, the discrepancy between the shift in Precision@5 and Answer Correctness indicates that the primary ongoing constraint of the system is no longer retrieval, but rather the generation component of the pipeline.

When interpreting the results, it is important to keep in mind that the test set contains only ten questions, so minor differences between adjacent iterations fall within the expected noise for such a small sample size. At $n = 10$, a single improved test shifts the overall score by 10 percentage points; differences of less than approximately 20 percentage points are not statistically significant in the sense of the Wilcoxon test. The highest scores are achieved by the per-test analysis for queries on drought (94%) and sea level (95%), where the system accurately identifies a single authoritative source. The lowest mean test scores were observed in queries concerning extreme heat (88%) and the atmosphere (89%), where the catalogue contains numerous datasets with analogous semantics and retrieval, thereby dispersing attention among them. Cross-domain queries (e.g. Canadian wild-fires, Pakistani floods) achieved 89% and 90% accuracy, respectively, although they require linking data from multiple datasets.

The second tracked quantity is response latency. The baseline pipeline averaged 120 to 180 seconds per query, which is unacceptable for a web interface. After optimization, the median response time stabilized at 25.9 seconds across the ten test queries in the final deterministic run — an improvement of roughly 80%. Two changes drove most of the gain. First, batch gRPC search collapsed sequential variant queries into a single Qdrant call, cutting the search time from 60–100 seconds to about 5. Second, removing the cross-encoder reranker from the hot path saved 40–50 seconds per query.

To evaluate the performance of the ETL pipeline, four real-world datasets from the catalogue were tested. The processing of the CMIP6 dataset with three variables (tas, pr, psl) required approximately 95 seconds and yielded 3,124 chunks with a total size of approximately 50 MB. The CRU TS dataset was processed in approximately 2.3 minutes, yielding 5,718 chunks and 120 MB of output data. The E-OBS dataset, comprising four variables (tg, tn, tx, rr), yielded approximately 33,000

chunks, processed in batches that completed in approximately 32 seconds each. The processing of metadata for all 233 records in the zero phase was completed in under 30 seconds. Key optimizations in the ETL phase included batch GPU embedding, which delivered a 10–20-fold speed enhancement compared to sequential execution, and generator-based streaming, which maintains a constant memory footprint for processing regardless of the input file size.

4 DISCUSSION AND CONCLUSIONS

The implemented system encompasses the entire workflow, from the downloading of raw climate datasets through to automatic format detection, transformation into text-numeric descriptions, embedding generation and storage, and finally RAG querying via a web interface. The validation results on a test set of ten queries (average composite score of 91%, success rate of 10/10) demonstrate that the text embeddings of climate chunks are sufficiently accurate for this type of query.

A comparison of the proposed solution with existing climate RAG systems reveals significant distinctions, primarily in terms of the breadth of formats it supports and its ability to directly process raw climate data, in addition to text documents. ChatClimate (Vaghefi *et al.*, 2023) is a system that utilizes GPT-4 with retrieval to address queries regarding IPCC reports; however, it is exclusively compatible with PDF documents and is not open-source. ClimateGPT (Thulke *et al.*, 2024), which was trained via follow-up pretraining on climate texts, functions with scientific literature rather than raw datasets. GeoGPT-RAG (Huang *et al.*, 2025) is conceptually the closest, utilizing its own GeoEmbedding model built on Mistral-7B and a georanker trained on spatial data. However, its RAG system targets geoscientific publications, not climate grids, and the system as a whole is not publicly available. Spatial-RAG (Yu *et al.*, 2025) is a system that explicitly addresses spatial filtering via geometric queries. However, it is limited to the GeoJSON format and is a closed-source system. The proposed system’s key advantage lies in its combination of dimensions, as evidenced by the fact that none of the compared solutions cover the entire chain from downloading raw climate rasters through transformation and embedding to final web-based querying. Furthermore, the system is open-source.

From a technical standpoint, the most significant finding of this study is that text embeddings of climatic chunks outperform purely numerical representations. A statistical 8-dimensional vector (mean, standard deviation, minimum, maximum, percentiles) is incapable of distinguishing variables with a similar numerical scale, resulting in temperature and precipitation chunks occupying proximity in the latent space. The resolution of this issue is achieved through the provision of a textual description that encompasses units, attributes in accordance with CF Conventions, and statistical characteristics. This approach is substantiated by the observation that sentences such as “mean temperature 15.3 °C” and “total precipitation 15.3 mm” possess divergent semantics.

The work has several limitations. The BAAI/bge-large-en-v1.5 embedding model has been trained on English texts; consequently, queries in Slovak or Czech have lower accuracy. The multilingual alternative, BGE-M3, was not utilized due to incompatibility with ONNX Runtime. The test set of ten queries is inadequate for deriving statistically robust conclusions, and the deterministic keyword-based grader imposes penalties on semantically correct answers formulated using synonyms. Dependence on a single LLM backend (OpenRouter) is a single point of failure; on service outage the endpoint degrades to raw retrieval results, without a generated answer. Furthermore, Juhasz *et al.* (2024) also measured a policy violation rate of 27 to 30% on four tested models using climate RAG, indicating that the errors are not merely a local phenomenon of a specific implementation.

From the perspective of climate policy and sustainability research, easier access to standardized climate data can broaden who is able to use it: municipal analytics teams, journalists, educational institutions, and small academic groups without dedicated data infrastructure. Climatologists and researchers benefit from a shorter onboarding phase, students from simpler exploration in teaching, and downstream LLM applications in the climate domain from a ready-to-use data source. The open-source nature of the system, together with reproducible Docker Compose deployment, supports analysis replicability - important for the academic credibility of climate work.

The article demonstrates that an integrated open-source system for the automated processing of heterogeneous climate data with a RAG layer over a real-world dataset of over one million points is technically feasible and evaluably accurate within the framework of external validation. The paper explicitly identifies and quantifies compromises in language availability (English embedding model), the scope of evaluation ($n = 10$ queries), and response latency (26 seconds). For the majority of these compromises, a proposed next step exists, indicating that the system is not merely a demonstration of a working prototype, but also a starting point for real-world deployment and domain extensions.

Acknowledgements

This paper was supported by the project CZ.02.1.01/0.0/0.0/16_017/0002334 Research Infrastructure for Young Scientists, this is co-financed from Operational Programme Research, Development and Education.

Several AI tools were used in the preparation of this article. They assisted with summarizing source material, translating passages between languages, and identifying typographical and stylistic issues in the text. All output produced by these tools was reviewed and verified. Summaries were checked against the original sources, translations were assessed for accuracy, and suggested corrections were evaluated before being accepted. The tools served as an aid to the writing process; the final content reflects the authors' judgment and remains the authors' responsibility.

REFERENCES

- CARBONELL, J., GOLDSTEIN, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ,98)*. New York: ACM, pp. 335–336. <https://doi.org/10.1145/290941.291025>
- CORMACK, G. V., CLARKE, C. L. A., BÜTTCHER, S. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ,09)*. New York: ACM, pp. 758–759. <https://doi.org/10.1145/1571941.1572114>
- CORNES, R. C., VAN DER SCHRIER, G., VAN DEN BESSELAAR, E. J. M., JONES, P. D. 2018. An ensemble version of the E-OBS temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*. 123(17), 9391–9409. <https://doi.org/10.1029/2017JD028200>
- CHOMO, D. 2026. *Automatizácia ETL procesov pre spracovanie klimatických údajov z webových zdrojov*. Master's thesis. Brno: Mendel University in Brno, Faculty of Business and Economics.

- DEELMAN, E., VAHI, K., JUVE, G., RYNGE, M., CALLAGHAN, S., MAECHLING, P. J., MAYANI, R., CHEN, W., FERREIRA DA SILVA, R., LIVNY, M., WENGER, K. 2015. Pegasus, a workflow management system for science automation. *Future Generation Computer Systems*. 46, 17–35. <https://doi.org/10.1016/j.future.2014.10.008>
- ES, S., JAMES, J., ESPINOSA ANKE, L., SCHOCKAERT, S. 2024. RAGAS: automated evaluation of retrieval augmented generation. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations*. pp. 150–158.
- EYRING, V., BONY, S., MEEHL, G. A. *et al.* 2016. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6): experimental design and organization. *Geoscientific Model Development*. 9(5), 1937–1958. <https://doi.org/10.5194/gmd-9-1937-2016>
- HARRIS, I., OSBORN, T. J., JONES, P., LISTER, D. 2020. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*. 7(1), 1–18. <https://doi.org/10.1038/s41597-020-0453-3>
- HERSBACH, H., BELL, B., BERRISFORD, P. *et al.* 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*. 146(730), 1999–2049. <https://doi.org/10.1002/qj.3803>.
- HEVNER, A. R., MARCH, S. T., PARK, J., RAM, S. 2004. Design science in information systems research. *MIS Quarterly*. 28(1), 75–105. <https://doi.org/10.2307/25148625>.
- HOYER, S., HAMMAN, J. 2017. xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software*. 5(1). <https://doi.org/10.5334/jors.148>.
- HUANG, F., WU, F., ZHANG, Z. *et al.* 2025. GeoGPT-RAG technical report. arXiv preprint. [arXiv:2509.09686](https://arxiv.org/abs/2509.09686).
- JUHASZ, M., DUTIA, K., FRANKS, H. *et al.* 2024. Responsible retrieval augmented generation for climate decision making from documents. arXiv preprint. [arXiv:2410.23902](https://arxiv.org/abs/2410.23902).
- KIMBALL, R., CASERTA, J. 2004. *The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data*. Indianapolis: Wiley.
- LEWIS, P., PEREZ, E., PIKTUS, A. *et al.* 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Advances in Neural Information Processing Systems (NeurIPS)*.
- MALKOV, Y. A., YASHUNIN, D. A. 2020. Efficient and robust approximate nearest neighbor using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 42(4), 824–836. <https://doi.org/10.1109/TPAMI.2018.2889473>.
- ROLNICK, D., DONTI, P. L., KAACK, L. H. *et al.* 2022. Tackling climate change with machine learning. *ACM Computing Surveys*. 55(2), 1–96. <https://doi.org/10.1145/3485128>.
- THULKE, D., GAO, Y., PELSER, P. *et al.* 2024. ClimateGPT: towards AI synthesizing interdisciplinary research on climate change. arXiv preprint. [arXiv:2401.09646](https://arxiv.org/abs/2401.09646).
- VAGHEFI, S. A., WANG, Q., MUCCIONE, V. *et al.* 2023. ChatClimate: grounding conversational AI in climate science. *Communications Earth & Environment*. [arXiv:2304.05510](https://arxiv.org/abs/2304.05510).
- VASSILIADIS, P. 2009. A survey of extract-transform-load technology. *International Journal of Data Warehousing and Mining*. 5(3), 1–27. <https://doi.org/10.4018/jdwm.2009070101>.
- XIAO, S., LIU, Z., ZHANG, P. *et al.* 2023. C-Pack: packed resources for general Chinese embeddings. arXiv preprint. [arXiv:2309.07597](https://arxiv.org/abs/2309.07597).
- YU, D., BAO, R., NING, R. *et al.* 2025. Spatial-RAG: spatial retrieval augmented generation for real-world geospatial reasoning questions. arXiv preprint. [arXiv:2502.18470](https://arxiv.org/abs/2502.18470).

Contact information

Dávid Chomo: e-mail: xchomo@node.mendelu.cz

Jaromír Landa: e-mail: jaromir.landa@mendelu.cz