ReSearch: A Multi-Stage Machine Learning Framework for Earth Science Data Discovery

Youran Sun
Department of Mathematics
University of Maryland, College Park, MD, USA
sun1245@umd.edu

Yixin Wen*
Department of Geography
University of Florida, Gainesville, FL, USA
yixin.wen@ufl.edu

Haizhao Yang*
Department of Mathematics
Department of Computer Science
University of Maryland, College Park, MD, USA
hzyang@umd.edu

January 19, 2026

Abstract

The rapid expansion of Earth Science data from satellite observations, reanalysis products, and numerical simulations has created a critical bottleneck in scientific discovery, namely identifying relevant datasets for a given research objective. Existing discovery systems are primarily retrieval-centric and struggle to bridge the gap between high-level scientific intent and heterogeneous metadata at scale. We introduce **ReSearch**, a multi-stage, reasoning-enhanced search framework that formulates Earth Science data discovery as an iterative process of intent interpretation, high-recall retrieval, and context-aware ranking. ReSearch integrates lexical search, semantic embeddings, abbreviation expansion, and large language model–based reranking within a unified architecture that explicitly separates recall and precision objectives. To enable realistic evaluation, we construct a literature-grounded benchmark by aligning natural language intent with datasets cited in peer-reviewed Earth Science studies. Experiments demonstrate that ReSearch consistently improves recall and ranking performance over baseline methods, particularly for task-based queries expressing abstract scientific goals. These results underscore the importance of intent-aware, multi-stage search as a foundational capability for reproducible and scalable Earth Science research.

1 Introduction

Earth Science research seeks to understand complex, interconnected processes governing the atmosphere, oceans, land surface, cryosphere, and their interactions with human systems. This endeavor increasingly relies on large-scale observational datasets, reanalysis products, and numerical simulations produced by a diverse ecosystem of satellites, in situ instruments, and computational models. While the volume and diversity of Earth Science data have grown rapidly, the ability to efficiently identify, access, and integrate relevant datasets has not kept pace. As a result, data discovery has emerged as a critical bottleneck that constrains scientific productivity, reproducibility, and participation.

Efficient discovery of relevant Earth Science data is impeded by three interrelated challenges. First, datasets are distributed across repositories with inconsistent metadata schemas, causing identical variables to be labeled differently across sources. Second, a persistent semantic gap exists between high-level

*Corresponding author.

research objectives and the technical descriptors used in metadata, requiring substantial domain expertise to bridge. Third, as archives scale to petabytes, balancing recall and precision becomes increasingly difficult, demanding reasoning capabilities beyond simple keyword matching. These challenges are examined in detail in Section 2.

Recent advances in machine learning offer new opportunities to address these limitations. Semantic embedding models and large language models (LLMs) have demonstrated strong capabilities in representing meaning, handling linguistic variability, and interpreting natural language queries. These tools provide a promising foundation for bridging the gap between scientific intent and technical metadata. However, purely neural approaches often lack transparency, stability, and explicit control over recall and relevance. These limitations highlight the need for discovery frameworks that combine machine learning with principled search strategies, explicitly encoding the stages of reasoning that underlie scientific inquiry.

In this work, we introduce **ReSearch**, a multi-stage framework for Earth Science data discovery that decomposes the search process into query understanding, high-recall retrieval, and context-aware reranking. By combining lexical matching, semantic embeddings, and large language model reasoning within a unified architecture, ReSearch addresses metadata heterogeneity and the semantic gap while maintaining interpretability and reproducibility. The main contributions of this paper are summarized as follows.

- **Multi-stage, intent-aware search formulation.** We reformulate Earth Science data discovery as an iterative reasoning process that bridges high-level scientific intent and heterogeneous data repositories. ReSearch explicitly decomposes discovery into three stages, enabling scalable search while preserving scientific relevance.

- **Hybrid retrieval framework integrating machine learning and information retrieval.** We develop a unified search architecture that combines lexical matching, semantic embedding retrieval, abbreviation expansion, and large language model reasoning. This hybrid design improves robustness to inconsistent terminology and metadata heterogeneity common across Earth Science datasets.

- **Literature-grounded benchmark for Earth Science data discovery.** We construct an evaluation dataset derived from peer-reviewed Earth Science literature by aligning natural language research queries with datasets cited in published studies. This benchmark reflects authentic discovery scenarios and enables realistic assessment of recall, ranking quality, and robustness to semantic variation.

- **Empirical evaluation on large-scale Earth Science repositories.** Through extensive experiments, we demonstrate that ReSearch consistently outperforms baseline retrieval methods, particularly for task-based queries that represent high-level scientific objectives. The results highlight the importance of multi-stage search strategies for reliable and reproducible Earth Science research.

2 Problem Formulation

The efficient discovery of relevant Earth Science data is impeded by a combination of structural and semantic barriers arising from decentralized archiving practices, heterogeneous metadata standards, and the use of specialized scientific terminology across subdisciplines. These challenges hinder the translation of scientific questions into effective data queries and limit the scalability, reproducibility, and accessibility of Earth Science research workflows.

1. **Metadata Heterogeneity.** Earth Science datasets are distributed across numerous repositories that employ inconsistent metadata schemas and naming conventions. Identical physical variables are often labeled differently across sources; for example, precipitation may appear as “precip”, “rainfall”, or “pr”, while soil moisture, elevation, or land surface temperature exhibit similar inconsistencies. Such heterogeneity prevents direct matching through keyword-based search and necessitates additional semantic normalization before datasets can be meaningfully compared or integrated.

2. **The Semantic Gap Between Scientific Intent and Metadata.** A persistent disconnect exists between high-level Earth Science research objectives (e.g., “drought assessment”, “flood risk

analysis", or "cryospheric change detection") and the low-level technical descriptors used in dataset metadata. Translating abstract scientific intent into specific variables, products, and spatiotemporal constraints requires substantial domain expertise. Even modern semantic search systems struggle to perform this translation reliably, often failing to infer the precise data requirements implied by a user's research goal.

3. **Scale and Precision Trade-offs.** As Earth Science data archives grow to petabyte scales, achieving both high recall and high precision becomes increasingly difficult. Simple keyword matching frequently produces excessive noise, while overly restrictive filtering risks excluding relevant datasets. Effective discovery therefore requires nuanced reasoning over spatiotemporal coverage, variable semantics, and scientific context, capabilities that are largely absent from existing retrieval-centric systems.

3 Related Work

A growing body of work has explored the use of machine learning and semantic technologies to enhance Earth Science data discovery and interoperability [Addink and Leeflang, 2025, Parisi and Bratsas, 2025]. Major data portals and federated archives provide centralized access to observational and model datasets, while recent approaches have investigated semantic representations, embeddings, and knowledge graphs to bridge heterogeneity across repositories. Although these systems improve data accessibility, they are primarily designed for human-driven retrieval and offer limited support for intent-aware or reasoning-oriented discovery.

AutoClimDS [Jaber et al., 2025] integrates a knowledge graph with a multi-agent system to support climate data discovery. It structures metadata as a bipartite graph linking datasets to auxiliary concepts (e.g., variables, keywords), enabling interpretable, one-hop traversal-based retrieval. Dataset relevance is inferred indirectly via similarity to these auxiliary nodes. While effective, this approach depends heavily on fixed taxonomies and lacks flexibility for iterative search or adaptation to diverse query intents.

Beyond knowledge-graph-based systems, embedding-based semantic retrieval has been employed to improve coverage across large, lexically diverse archives. For example, Qiu et al. [Qiu et al., 2019] proposed an ontology-enhanced embedding model to support keyphrase extraction from geoscience literature. Ramachandran et al. [Ramachandran and Ramasubramanian, 2021] incorporated prediction-based embeddings to augment metadata in NASA Earth science repositories. More recently, Weckmüller et al. [Weckmüller et al., 2025] designed a multilingual embedding-based search framework for geo-textual datasets. These approaches enhance recall and language robustness, but often conflate retrieval and ranking, and offer limited mechanisms for integrating structured constraints or domain-specific reasoning.

Recent trends increasingly emphasize modeling data discovery as a multi-stage reasoning process, where components such as intent interpretation, candidate generation, and contextual ranking are decoupled. Such designs offer greater flexibility across heterogeneous repositories and can integrate with both semantic and agent-based systems. This work builds on these directions, seeking to extend the reasoning capacity of discovery systems beyond rigid taxonomies and single-step retrieval.

4 Methodology

4.1 Data Sources

To establish a robust foundation for Earth Science research, we have integrated four primary data repositories. These include the NASA Common Metadata Repository (CMR), NOAA OneStop, the Coupled Model Intercomparison Project Phase 6 (CMIP6), and the ERA5 (ECMWF Reanalysis v5) dataset. These sources collectively represent a vast and diverse collection of Earth observation and simulation data, ranging from satellite measurements to numerical model outputs.

Table 1 summarizes the scale and characteristics of the integrated data sources.

A significant challenge in integrating these sources is the heterogeneity of metadata and variable naming conventions. For example, precipitation is referenced by various identifiers such as "precipitation_rate" or "precipRate" across different datasets. To address this, our system employs metadata normalization and semantic mapping strategies to ensure consistent discovery capabilities.

Table 1: Overview of Integrated Earth Science Data Sources

Data Source	Type	Datasets	Data Volume
NASA CMR	Satellite Observations	≈ 54,000	PB scale
NOAA OneStop	Meteorological & Oceanographic	≈ 52,000	Hundreds of TB
CMIP6	Earth System Model Simulations	≈ 102,000	≈ 20 PB
ERA5	Reanalysis	47	≈ 5 PB

4.2 Search Strategy Design

We propose a multi-stage search engine architecture tailored for Earth Science data discovery. This architecture bridges the gap between natural language research queries and technical metadata schemas. The pipeline comprises three distinct stages, comprising Query Understanding, Recall, and Reranking.

Stage 0. Query Understanding The initial stage focuses on intent classification and query refinement. User queries are categorized into two types, Type A (specific data requests) and Type B (broad research goals). While Type A queries undergo standard spell correction, Type B queries are processed using LLMs to translate abstract research objectives (e.g., “flood analysis”) into specific, retrievable data requirements (e.g., “precipitation”, “storm surge”). This stage also extracts structured constraints, such as temporal and spatial ranges, to aid downstream filtering.

Stage 1. Recall To maximize recall, we implement a hybrid retrieval strategy. This involves a combination of structured filtering, utilizing the extracted spatiotemporal constraints, and a dual-path search mechanism. We employ Best Matching 25 (BM25), a widely used term-frequency-based ranking function, for keyword-based matching to capture exact lexical matches in metadata fields. Simultaneously, we utilize vector embedding search to identify semantically related datasets that may differ in terminology. To further bridge the semantic gap, we apply abbreviation expansion, which augments the indexed metadata by inserting full-form expansions after detected abbreviations (e.g., “MODIS” → “MODIS (Moderate Resolution Imaging Spectroradiometer)”). This ensures a comprehensive candidate set that captures both explicit and implicit relevance.

Stage 2. Reranking The final stage refines the candidate set using an LLM-based reranker. By evaluating the detailed metadata of retrieved datasets (including titles and summaries) against the nuanced context of the user’s original query, this stage assigns a relevance score to each dataset. This process effectively promotes the most pertinent resources to the top of the results, filtering out noise introduced during the high-recall retrieval phase.

4.3 Evaluation Dataset Construction

To rigorously evaluate the proposed search engine, we established a benchmark dataset derived from peer-reviewed Earth Science literature. This approach ensures that our evaluation reflects authentic research needs and terminologies.

Figure 1 illustrates the construction pipeline for our evaluation dataset. Starting from academic papers, we extract both research queries (keywords and “I want to” statements) and dataset references with their URLs. Datasets are matched to NASA CMR entries through URL pattern matching and fuzzy name matching rules, producing query-groundtruth pairs for benchmark evaluation.

For each sampled paper, we utilized an LLM (GPT-4o) to extract the datasets explicitly cited by the authors. These datasets constitute the ground truth and were aligned with the integrated metadata repositories (e.g., NASA CMR) via unique short name identifiers. To simulate diverse retrieval scenarios, we synthesized two distinct types of queries for each document:

- **Keyword-based Queries.** Aggregations of domain-specific keywords extracted from the text, simulating users searching with precise technical terminology.
- **Task-based Queries.** Natural language formulations (e.g., “I want to analyze...” statements) that describe high-level research objectives, representing users who may not be familiar with specific dataset identifiers.

Table 2 summarizes the evaluation dataset statistics. Note that one paper was excluded from evaluation as none of its datasets matched entries in NASA CMR.

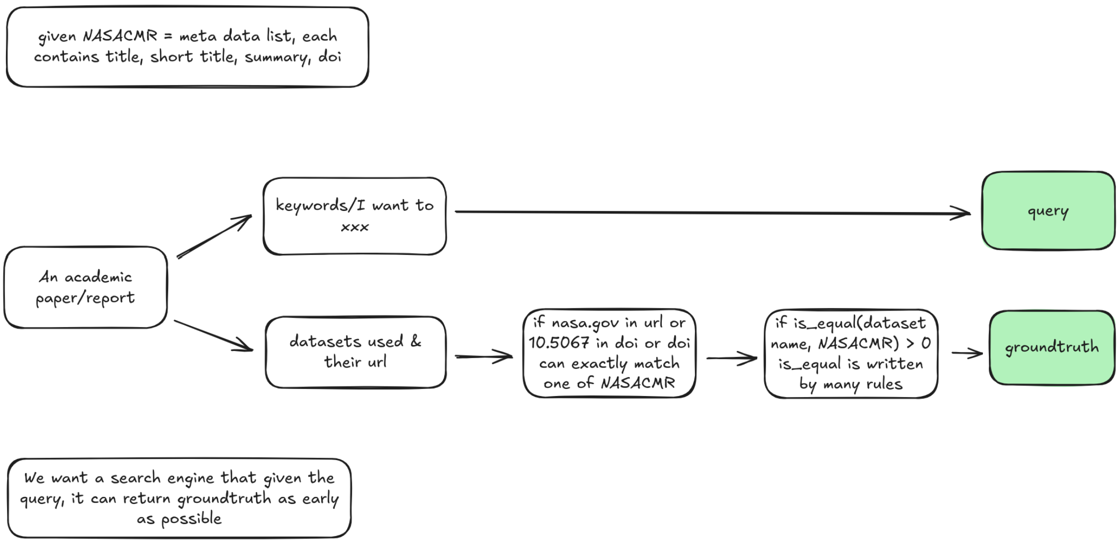
given NASACMR = meta data list, each contains title, short title, summary, doi
keywords/I want to xxx
query
An academic paper/report
datasets used & their url
if nasa.gov in url or 10.5067 in doi or doi can exactly match one of NASACMR
if is_equal(dataset name, NASACMR) > 0 is_equal is written by many rules
groundtruth
We want a search engine that given the query, it can return groundtruth as early as possible

Figure 1: Evaluation dataset construction pipeline. From academic papers, we extract queries and dataset references, then match datasets to NASA CMR entries to establish ground truth for retrieval evaluation.

Table 2: Evaluation Dataset Statistics

Item	**Count**
Sampled academic papers	15
Extracted keywords	80
Task-based queries (“I want to...”)	30
Datasets mentioned	49

5 Experiments

5.1 Metrics

We quantify system performance using three standard information retrieval metrics:

- **Recall@K.** Measures the fraction of relevant datasets retrieved within the top K results. For each query, Recall@K is computed as the number of ground truth datasets found within the top K results divided by the total number of ground truth datasets for that query. The final score is the macro-average across all queries.

- **Mean Reciprocal Rank (MRR).** Evaluates the ranking effectiveness by calculating the average of the reciprocal ranks of the first relevant result. This metric reflects the system’s ability to place a correct answer near the top of the list. It is defined as

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q}$$

where Q is the set of queries and rank_q is the rank of the first relevant document for query q.

- **Mean Average Precision (MAP).** Assesses the overall quality of the ranking by considering the precision at each relevant document’s position. For a query q, the average precision is defined as

$$\text{AP}(q) = \frac{1}{|\mathcal{G}_q|} \sum_{k=1}^{|\mathcal{R}_q|} P_q(k) \cdot rel_q(k),$$

where $\mathcal{G}_q$ is the set of relevant documents for query q, $\mathcal{R}_q$ is the ranked list returned by the system, $P_q(k)$ denotes the precision at rank k, and $rel_q(k) \in \{0, 1\}$ indicates whether the item at rank k is relevant. MAP is computed as the mean of AP over all queries.

5.2 Results

Table 3 presents the preliminary evaluation results across different retrieval configurations.

Table 3: Retrieval Performance on Earth Science Literature Benchmark

Method	R@10	R@20	R@50	R@100	MRR	MAP
Keyword-based Queries (n=14)						
AutoClimDS	0.02	0.02	0.02	0.02	0.0714	0.0179
BM25	0.05	0.05	0.18	0.31	0.0248	0.0199
BM25+AbbrExp	0.12	0.15	0.27	0.32	0.0397	0.0327
Embedding Retrieval	0.02	0.02	0.22	0.25	0.0516	0.0223
Embd+AbbrExp	0.06	0.09	0.26	0.27	0.0646	0.0292
Task-based Queries (n=28)						
AutoClimDS	0.01	0.01	0.01	0.01	0.0714	0.0129
BM25	0.03	0.11	0.22	0.28	0.0274	0.0165
BM25+AbbrExp	0.04	0.11	0.22	0.28	0.0619	0.0257
Embedding Retrieval	0.05	0.08	0.17	0.25	0.0636	0.0315
Embd+AbbrExp	0.07	0.12	0.19	0.23	0.0489	0.0309

Note: AbbrExp = Abbreviation Expansion

6 Conclusion

In this paper, we presented **ReSearch**, a multi-stage, machine learning–enhanced framework for Earth Science data discovery that addresses a foundational bottleneck in data-driven research, namely translating high-level scientific intent into reliable and scalable access to heterogeneous data repositories. By reformulating data discovery as an iterative reasoning process, ReSearch explicitly separates intent interpretation, high-recall retrieval, and context-aware ranking, enabling effective search across large and inconsistently described Earth Science archives.

The proposed framework integrates complementary techniques from information retrieval and machine learning, including lexical matching, semantic embedding search, abbreviation expansion, and large language model–based reranking. This hybrid design improves robustness to metadata heterogeneity and linguistic variation, which are pervasive challenges in Earth Science data ecosystems. Through a literature-grounded evaluation derived from peer-reviewed studies, we demonstrated that ReSearch consistently outperforms baseline methods, particularly for task-based queries that reflect abstract scientific objectives rather than dataset-specific terminology.

Beyond empirical performance gains, this work highlights the importance of treating data discovery as a first-class component of scientific workflows. Errors or omissions at the discovery stage can propagate downstream, affecting analysis quality, reproducibility, and scientific conclusions. By improving the reliability and transparency of discovery, ReSearch supports more reproducible and inclusive Earth Science research, lowering barriers for interdisciplinary and data-driven investigations.

While this study focuses on Earth Science data repositories, the ReSearch framework is domain-agnostic and applicable to other scientific fields characterized by large-scale, heterogeneous datasets and evolving terminology. Future work will explore extensions of ReSearch to support the discovery of analysis methods, modeling workflows, and simulation strategies, as well as tighter integration with knowledge graph–based and agentic systems for end-to-end automated research pipelines. Together, these directions position ReSearch as a foundational search intelligence layer for next-generation, machine-learning-enabled scientific discovery.

A Evaluation Dataset Details

Table 4 presents the datasets extracted from the sampled Earth Science papers that were successfully matched to entries in the integrated data repositories.

Table 4: Datasets Extracted from Evaluation Papers (Matched Only)

Paper	Dataset	Matches
essd-15-5449-2023.pdf	Normalized Difference Vegetation Index (NDVI) from Moderate Resolution Imaging Spectroradiometer (MOD13C2)	86

Continued on next page

Table 4 – continued from previous page

Paper	Dataset	Matches
	Land surface temperature data from Moderate Resolution Imaging Spectroradiometer (MOD11C3)	3
sui-et-al-2024-global-scale-assessment-of-urban-precipitation-anomalies.pdf	GPM IMERG Final Precipitation L3 1 day 0.1 degree x 0.1 degree V06	1
	Level-3 Aura/OMI Global Aerosol Data (OMAEROe)	2
	MOD11C3 MODIS/Terra Land Surface Temperature/Emissivity Monthly L3 Global 0.05Deg CMG V006	3
	MCD12C1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 0.05Deg CMG	2
	Global Human Settlement Layer: Population and Built-Up Estimates, and Degree of Urbanization Settlement Model Grid	1
Regional_scale_intelligent_precipitation_gap.pdf	Integrated Multi-satellite Retrievals for Global Precipitation Measurement-Early (IMERG-Early)	2
Regional analysis of the 2015–16 Lower Mekong River basin drought using NASA satellite observations.pdf	GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V06	1
	GRACE RL06	38
Tropical deforestation causes large reductions in observed precipitation.pdf	GPCP	14
	GPM	476
	MERRA-2	109
	TRMM	120
jwh0221162.pdf	TMPA	12
	GPM/IMERG	9
	MODIS	1113
	VIIRS	854
	SMAP	203
	ASTER	551
	ECOSTRESS	47
	GRACE	80
	Topex/Poseidon	6
016002_1.pdf	MODIS MYD11A2	2
	MODIS MYD07	3
	MODIS MYD06	3
	MODIS MOD13Q1	2
	MODIS MYD13Q1	2
	MODIS MYD03	3
	MODIS MYD05_L2	3
	MODIS MYD06_L2	3
	MODIS MYD07_L2	3
	MODIS MYD11_L2	4
	MODIS MYD11A1	2
1-s2.0-S0034425717301967-main.pdf	Landsat 5	30
	Landsat 7	18
	Landsat 8	4
essd-12-1141-2020.pdf	Global Lake/Reservoir Surface Inland Water Height GREALM V.2	1
	Global Lake/Reservoir Surface Inland Water Area Extent V2	1
	Lake and Reservoir Storage Time Series V2	1
Xu_2022_Environ._Res._Lett._17_074013.pdf	ICESat_GLAH14	1
science.abo2812.pdf	Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2)	109
essd-16-201-2024.pdf	ATLAS/ICESat-2 L3A Along-Track Inland Surface Water Data, Version 6 (ATL13)	3
	ATLAS/ICESat-2 L3A Along Track Inland Surface Water Data Quick Look, Version 6 (ATL13QL)	1
	Global Reservoirs and Lakes Monitor (GREALM)	1
essd-14-2463-2022.pdf	MODIS MOD09Q1	2
Quart J Royal Meteoro Soc - 2022 - Lavers - An evaluation of ERA5 precipitation for climate monitoring.pdf	Tropical Rainfall Measuring Mission TRMM/3B43	1

References

[Addink and Leeflang, 2025] Addink, W. and Leeflang, S. (2025). Towards a common language for digital specimens: Exploring the opends data model. In *Biodiversity Information Science and Standards*. Pensoft.

[Jaber et al., 2025] Jaber, A., Zhu, W., Jayavelu, K., Downes, J., Mohamed, S., Agonafir, C., Hawkins, L., and Zheng, T. (2025). Autoclimds: Climate data science agentic ai – a knowledge graph is all you need.

[Parisi and Bratsas, 2025] Parisi, E. and Bratsas, C. (2025). From data to decision. Technical report, Aristotle University of Thessaloniki.

[Qiu et al., 2019] Qiu, Q., Xie, Z., Wu, L., and Li, W. (2019). Geoscience keyphrase extraction algorithm using enhanced word embedding. *Expert Systems with Applications*, 126:157–169.

[Ramachandran and Ramasubramanian, 2021] Ramachandran, R. and Ramasubramanian, M. (2021). Augmenting data systems with prediction-based embeddings. *IEEE Transactions on Geoscience and Remote Sensing*.

[Weckmüller et al., 2025] Weckmüller, D., Dunkel, A., and Burghardt, D. (2025). Embedding-based multilingual semantic search for geo-textual data in urban studies. *Journal of Geovisualization and Spatial Analysis*, 9(1).