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 13, 2026

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.

1.1 Challenges in Earth Science Data Discovery

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.

*Corresponding author.

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.

1.2 Opportunities with Machine Learning–Enhanced Search

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, enabling researchers to interact with data repositories using more expressive and flexible queries.

However, applying machine learning to Earth science data discovery introduces new challenges. Purely neural approaches often lack transparency, stability, and explicit control over recall and relevance. Without structured mechanisms to guide retrieval and ranking, such systems may produce results that are difficult to interpret or reproduce. 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.

1.3 Related Work

A growing body of work has explored machine learning and semantic technologies to improve Earth science data discovery and interoperability. Major data portals and federated archives provide centralized access to observational and model datasets, while recent efforts have investigated the use of semantic representations, embeddings, and knowledge graphs to bridge inconsistencies across repositories. These systems have significantly improved data accessibility but are largely designed for human-driven retrieval rather than intent-aware or reasoning-oriented discovery.

AutoClimDS [Jaber et al., 2025] represents a recent approach to climate data discovery that integrates a knowledge graph with a multi-agent architecture. Its search mechanism can be conceptualized as traversal over a **bipartite graph**, in which nodes are partitioned into two disjoint sets: *target nodes* (Dataset) and *auxiliary nodes* (e.g., DataCategory, Variable, ScienceKeyword). Edges exist only between these two sets, with no intra-set connections. Given a user query, the system first encodes the query using a sentence embedding model and retrieves the top-k most similar auxiliary nodes via approximate nearest neighbor search. It then follows predefined relationships (such as `hasDataCategory` and `hasVariable`) to reach connected Dataset nodes through one-hop traversal. Because Dataset nodes do not possess embeddings of their own, each dataset inherits the similarity score of its associated auxiliary node, with the maximum score retained when multiple paths exist. The final dataset ranking is determined by these inherited scores.

This design effectively leverages structured metadata relationships encoded in the knowledge graph and enables interpretable traversal-based discovery. However, the approach relies on the quality, completeness, and granularity of pre-defined auxiliary node categories. As a result, datasets that are poorly represented by existing taxonomies or that fall outside established semantic groupings may be difficult to discover. Moreover, the bipartite structure constrains discovery to a single-hop expansion from auxiliary concepts, limiting the system's ability to support iterative refinement or to adapt dynamically to diverse query formulations.

Beyond knowledge graph–based approaches, embedding-based semantic search and hybrid retrieval systems have been proposed to improve recall across large scientific archives. While these methods reduce sensitivity to lexical variation, they often conflate recall and ranking objectives and provide limited mechanisms for incorporating scientific context or structured constraints. Consequently, data

discovery in Earth science remains a largely manual, expertise-driven process, even in the presence of advanced machine learning techniques.

In contrast, the approach proposed in this work focuses on the *search intelligence layer* itself. Rather than relying on fixed taxonomies or single-step retrieval, ReSearch models data discovery as a multi-stage reasoning process that explicitly separates intent interpretation, high-recall candidate generation, and context-aware ranking. This design enables scalable and adaptable discovery across heterogeneous Earth science repositories while remaining compatible with downstream knowledge graph–based or agentic systems.

1.4 Contributions of This Paper

In this work, we introduce **ReSearch**, a multi-stage, machine learning–enhanced framework for Earth science data discovery. 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—query understanding, high-recall retrieval, and context-aware reranking—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–based 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 Methodology

2.1 Data Sources

To establish a robust foundation for climate research support, we have integrated four primary data repositories, including the NASA Common Metadata Repository (CMR), NOAA OneStop, the Coupled Model Intercomparison Project Phase 6 (CMIP6), and the ERA5 Reanalysis. These sources collectively represent a vast and diverse collection of climate information, ranging from satellite observations to model simulations.

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

Table 1: Overview of Integrated Climate Data Sources

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

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.

2.2 Search Strategy Design

We propose a multi-stage search engine architecture tailored for climate data discovery, designed to bridge the gap between natural language research queries and technical metadata schemas. The pipeline comprises three distinct stages, comprising Query Understanding, Retrieval, 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. Retrieval (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 BM25 for keyword-based matching to capture exact lexical matches in metadata fields, while simultaneously utilizing 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 both the indexed metadata and query terms 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.

2.3 Evaluation Dataset Construction

To rigorously evaluate the proposed search engine, we established a benchmark dataset derived from peer-reviewed climate 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.

3 Experiments

3.1 Metrics

We quantify system performance using three standard information retrieval metrics:

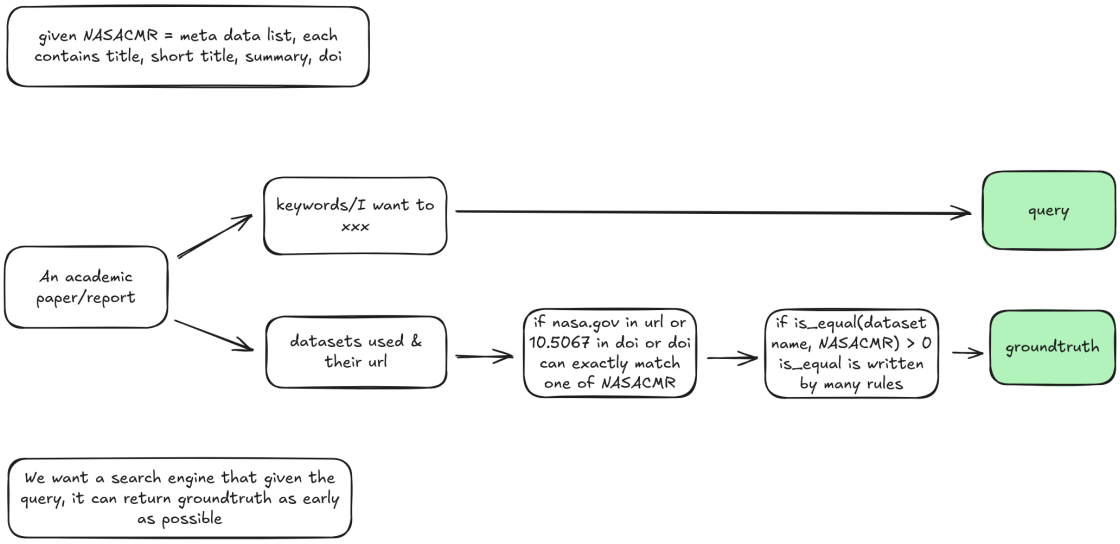

Figure 1: Evaluation dataset construction pipeline. From academic papers, we extract queries and dataset references, then match datasets to NASA CMR entries to establish ground truth for retrieval evaluation.

Table 2: Evaluation Dataset Statistics

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

- **Recall@K.** Measures the fraction of relevant datasets retrieved within the top K results. Given the high volume of climate data, ensuring relevant candidates are included in the initial retrieval set is critical.
- **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.

3.2 Results

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

Table 3: Retrieval Performance on Climate 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

3.3 Ablation Study: Effect of Query Rewriting Models

To investigate the impact of different LLMs on query rewriting performance, we compare three models: o1, o3, and GPT-4o. Table 4 presents the retrieval performance on task-based queries with different rewriting models.

Table 4: Comparison of Query Rewriting Models on Task-based Queries (n=5)

Method	R@10	R@20	R@50	R@100	MRR	MAP
$\text{Rewrite}_{\text{o1}}$ + BM25	0.00	0.00	0.00	0.00	0.0015	0.0009
$\text{Rewrite}_{\text{o3}}$ + BM25	0.00	0.00	0.00	0.00	0.0026	0.0020
$\text{Rewrite}_{\text{gpt-4o}}$ + BM25	0.00	0.00	0.00	0.05	0.0048	0.0014
$\text{Rewrite}_{\text{o1}}$ + Embedding	0.00	0.00	0.00	0.10	0.0057	0.0014
$\text{Rewrite}_{\text{o3}}$ + Embedding	0.00	0.05	0.10	0.10	0.0271	0.0070
$\text{Rewrite}_{\text{gpt-4o}}$ + Embedding	0.00	0.00	0.00	0.05	0.0033	0.0008

A Evaluation Dataset Details

Table 5 presents the datasets extracted from the sampled climate science papers used for evaluation. The "In CMR" column indicates whether the dataset was matched to an entry in the NASA Common Metadata Repository.

Table 5: Datasets Extracted from Evaluation Papers

Paper	Dataset	In CMR	Matches
1-s2.0-S0048969723041037-main.pdf	ERA5 reanalysis dataset	✗	–
	CRU TS v.4.06	✗	–
	GLEAM v3.6a	✗	–
	Resource and Environment Science and Data Center	✗	–
	Physical Sciences Laboratory	✗	–
essd-15-5449-2023.pdf	Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS, version 2)	✗	–
	Multi-Source Weighted-Ensemble Precipitation (MSWEP, version 2.8)	✗	–
	Potential evapotranspiration from the Global Land Evaporation Amsterdam Model (GLEAM, version 3.7a)	✗	–
	Hourly Potential Evapotranspiration (hPET)	✗	–
	Normalized Difference Vegetation Index (NDVI) from Moderate Resolution Imaging Spectroradiometer (MOD13C2)	✓	86
	Land surface temperature data from Moderate Resolution Imaging Spectroradiometer (MOD11C3)	✓	3
	Global SPEI database, SPEIbase v2.8	✗	–
High-resolution-observations-from-space-to-address.pdf	Copernicus 1-km surface soil moisture product	✗	–
	Theia 100-m soil moisture product	✗	–
	RT1-based TU Wien 1-km soil moisture product	✗	–
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
	NCEP/EMC 4KM Gridded Data (GRIB) Stage IV Data	✗	–
	Level-3 Aura/OMI Global Aerosol Data (OMAEROe)	✓	2
	ERA5 monthly averaged data on pressure levels from 1959 to present	✗	–
	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 multi-resolution terrain elevation data 2010 (GMTED2010)	✗	–

Continued on next page

Table 5 – continued from previous page

Paper	Dataset	In CMR	Matches
	Present and future Köppen-Geiger climate classification maps at 1-km resolution	✗	–
	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
	Climate Prediction Center Global Unified Gauge-Based Analysis of Daily Precipitation (CPC-Global)	✗	–
	Global Precipitation Climatology Project (GPCP)	✗	–
	ASTER GDEM	✗	–
	CMPA dataset	✗	–
	Global Satellite Mapping of Precipitation (GSMaP)	✗	–
Improvements in Precipitation Product with Newer NASAGPM IMERG Algorithm on TRMMTMPA Data During Summer Monsoon Period over India.pdf	IMERGF-V7	✗	–
	IMERGL-V7	✗	–
	IMERGF-V6	✗	–
	IMERGL-V6	✗	–
	TMPA-3B42 V7	✗	–
	TMPA-3B42RT V7	✗	–
	IMD-SRG	✗	–
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
	SMAP Level 3	✗	–
	Downscaled SMAP 1 km	✗	–
	GRACE RL06	✓	38
Tropical deforestation causes large reductions in observed precipitation.pdf	CHIRPS	✗	–
	CMORPH	✗	–
	CPC	✗	–
	CRU	✗	–
	ERA5	✗	–
	GPCC	✗	–
	GPCP	✓	14
	GPM	✓	476
	JRA	✗	–
	MERRA-2	✓	109
	NOAA (PREC/LAND)	✗	–
	PERSIANN (CCS, CDR, CCS-CDR, PDIR-NOW)	✗	–
	TRMM	✓	120
	UDEL	✗	–
	Global Forest Change (GFC) version 1.9	✗	–
Water Resources Research - 2024 - Yao - Leveraging ICESat ICESat-2 and Landsat for Global-Scale Multi-Decadal.pdf	ICESat L2 Global Land Surface Altimetry product (GLAH14)	✗	–
	ICESat-2 L3A Land and Vegetation Height product (ATL08)	✗	–
	Global Surface Water Data Set (GSW) version 1.3	✗	–
	Global Lake and Reservoir Area Time Series Data Set (GLATS)	✗	–
	Hydroweb	✗	–
	Database for Hydrological Time Series of Inland Waters (DAHITI)	✗	–
	Global Reservoir and Lake Monitor (G-REALM)	✗	–
jwh0221162.pdf	TMPA	✓	12
	CHIRPS v2.0	✗	–
	CMORPH	✗	–
	PERSIANN	✗	–
	GSMaP	✗	–
	GPM/IMERG	✓	9
	MSWEP	✗	–
	GPCC	✗	–
	CRU	✗	–
	GHCN-M	✗	–
	PREC/L	✗	–
	UDEL	✗	–
	CPC-Global	✗	–
	SM2RAIN	✗	–
	RS-PM	✗	–
	MOD16 ET	✗	–
	PT-JPL	✗	–
	GLEAM	✗	–
	ALEXI-DisALEXI	✗	–
	ESI	✗	–
	Landsat	✗	–
	AVHRR/GIMMS	✗	–
	MODIS	✓	1113
	VIIRS	✓	854
	Sentinel-2	✗	–
	AMSR-E	✗	–
	AMSR2	✗	–
	SMOS	✗	–
	SMAP	✓	203
	Sentinel-1	✗	–
	ASCAT	✗	–
	FY3-B	✗	–
	CHIRTS	✗	–
	ASTER	✓	551
	VIIRS	✓	854
	Sentinel-3	✗	–
	ECOSTRESS	✓	47
	GRACE	✓	80
	SWOT	✗	–
	Jason-2/3	✗	–
	ENVISAT	✗	–
	Topex/Poseidon	✓	6
	GRDC	✗	–
016002_1.pdf	MODIS MOD16A2	✗	–
	SSEBop	✗	–
	MODIS MYD11A2	✓	2
	MODIS MYD07	✓	3
	MODIS MYD06	✓	3
	MODIS MCD43B3	✓	0
	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
	GridMET	✗	–
	TopoWx	✗	–
	PRISM	✗	–
	MPI ET	✗	–
science.adh9607.pdf	MERIT Hydro	✗	–
	HydroSHEDS	✗	–
	TanDEM-X	✗	–

Continued on next page

Table 5 – continued from previous page

Paper	Dataset	In CMR	Matches
	World Ocean Circulation Experiment (WOCE)	✗	–
	Argo	✗	–
	Regional Cabled Array	✗	–
	TOGA/TAO/TRITON	✗	–
	CalCOFI	✗	–
	Ocean Observatories Initiative (OOI)	✗	–
	ERA5	✗	–
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
1-s2.0-S0034425720301620-main.pdf	Landsat 5	✗	–
	Landsat 7	✗	–
	Landsat 8	✗	–
	Shuttle Radar Topography Mission (SRTM)	✗	–
	Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010)	✗	–
	RapidEye	✗	–
Xu__2022__Environ.__Res.__Lett._17_074013.pdf	ICESat GLAH14	✓	1
	ICESat-2 ATL13	✗	–
s41467-022-32830-y.pdf	ERA5 reanalysis data	✗	–
	ARC-Lake	✗	–
	CCI Lakes	✗	–
	AMSR-E/2	✗	–
science.abo2812.pdf	Landsat 5 Thematic Mapper	✗	–
	Landsat 7 Enhanced Thematic Mapper-plus	✗	–
	Landsat 8 Operational Land Imager	✗	–
	ICESat	✗	–
	ICESat-2	✗	–
	CryoSat-2	✗	–
	ENVISAT	✗	–
	Jason series 1-3	✗	–
	SARAL	✗	–
	Sentinel 3	✗	–
	Global Reservoir Bathymetry Dataset	✗	–
	Global Surface Water (GSW) dataset	✗	–
	HydroLAKES database	✗	–
	Georeferenced global Dams And Reservoir dataset (GeoDAR)	✗	–
	Global Lake area, Climate, and Population (GLCP) dataset	✗	–
	HydroSHEDS dataset	✗	–
	Climatic Research Unit (CRU) data	✗	–
	ECMWF Reanalysis v5 (ERA5)	✗	–
	Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2)	✓	109
	Multi-Source Weighted-Ensemble Precipitation (MSWEP)	✗	–
	Global Historical Climatology Network data	✗	–
	Global Land Evaporation Amsterdam Model (GLEAM)	✗	–
	Global Reach-scale A priori Discharge Estimates for SWOT (GRADES)	✗	–
	Reconstructed human water use data	✗	–
1-s2.0-S0022169420306892-main.pdf	MODIS Terra Surface Reflectance Imagery (MOD 09)	✓	0
	LEGOS Hydroweb Altimetry Data	✗	–
	GRACE RL05 Time-Variable Gravity Solutions from CSR	✗	–
	GLDAS Noah Land Surface Model	✗	–
	WGHM (WaterGAP Global Hydrology Model)	✗	–
essd-16-201-2024.pdf	GloLakes historical and near-real-time lake storage dynamics data from 1984 to current	✗	–
	Joint Research Centre's Global Surface Water Dataset (GSWD)	✗	–
	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
	BLUEDOT water observatory	✗	–
	HydroLAKES database	✗	–
1-s2.0-S1569843224000487-main.pdf	Hydroweb	✗	–
	Global Reservoir and Lake Monitor database (G-REALM)	✗	–
	Database for Hydrological Time Series of Inland Waters (DAHITI)	✗	–
	Mountain data	✗	–
	HydroBASINS dataset	✗	–
	Global Reservoir and Dam (GRanD) database	✗	–
	Global Georeferenced Database of Dams (GOODD)	✗	–
	Georeferenced Global Dams and Reservoirs (GeoDAR)	✗	–
	SWOT River Database (SWORD)	✗	–
	Permafrost distribution data	✗	–
	Global dataset of small and medium-sized lakes	✗	–
	ERA5-Land	✗	–
	ISIMIP2b	✗	–
essd-14-2463-2022.pdf	HydroSat	✗	–
	GFZ Data Services	✗	–
	Envisat GDR-v3	✗	–
	Saral GDR T	✗	–
	ICESat-2 ATL13 v3	✗	–
	CryoSat-2 SIR GDR	✗	–
	Jason-1 GDR	✗	–
	Jason-2 (PISTACH) GDR	✗	–
	Jason-2 GDR	✗	–
	Jason-3 GDR	✗	–
	Sentinel-3A NTC	✗	–
	Sentinel-3B NTC	✗	–
	ITSG-Grace2018	✗	–
	MODIS MOD09Q1	✓	2
	Global Surface Water Dataset	✗	–
Quart J Royal Meteoro Soc - 2022 - Lavers - An evaluation of ERA5 precipitation for climate monitoring.pdf	ERA5	✗	–
	Tropical Rainfall Measuring Mission TRMM/3B43	✓	1
	Global Precipitation Climatology Centre	✗	–
	Climate Prediction Center Global Daily Unified Gauge-Based Analysis of Precipitation	✗	–
	Climate Hazards Group Infrared Precipitation with Stations	✗	–
	Next-Generation Radar (NEXRAD)	✗	–

References

[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.