A Scalable Framework for Heterogeneous Environmental Data Management Using Smart Data Pipeline

Pratik Poudel
Florida International University
Miami, FL, USA
ppoud001@fiu.edu

Boyuan Guan
Florida International University
Miami, FL, USA
bguan@fiu.edu

Nicole Sanchez
Florida International University
Miami, FL, USA
nsanc170@fiu.edu

Kiavash Bahreini
Florida International University
Miami, FL, USA
kbahrein@fiu.edu

Wencong Cui
Florida International University
Miami, FL, USA
wecui@fiu.edu

Andres Lopez
Florida International University
Miami, FL, USA
alope489@fiu.edu

Hamed Najafi
Florida International University
Miami, FL, USA
hnaja002@fiu.edu

Zhaohui Fu
Florida International University
Miami, FL, USA
fujen@fiu.edu

Leonardo Bobadilla
Florida International University
Miami, FL, USA
bobadilla@cs.fiu.edu

Jason Liu
Florida International University
Miami, FL, USA
liux@fiu.edu

Abstract

Environmental data originates from diverse sources, posing challenges in management, processing, and visualization. This paper introduces a scalable, AI-driven data pipeline framework for environmental data management and discovery. The framework integrates workflow orchestration, automated data ingestion and processing, federated storage, and seamless geospatial visualization. It employs a Ceph-based storage system to handle large, heterogeneous datasets, leveraging its fault-tolerant, distributed architecture for high-performance storage across object, block, and file interfaces. To enhance data discoverability and interoperability, the framework incorporates Generative AI (GenAI) for automated metadata generation, reducing manual annotation overhead while improving real-time processing and cross-platform integration. Additionally, the system enables interdisciplinary collaboration through standardized metadata structures and scalable data federation. A case study using buoy data validates the framework's capabilities, including data processing, cleaning, and visualization. By addressing critical data integration and accessibility challenges, the system fosters a scalable, efficient, and intelligent research data-sharing ecosystem for environmental science studies.

PEARC '25, Columbus, OH, USA

ACM ISBN 979-8-4007-1398-9/25/07
https://doi.org/10.1145/3708035.3736017

CCS Concepts

• **Information systems** → **Data management systems**; *Information storage systems*; • **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → *Environmental sciences.*

Keywords

Environmental Data Management, Smart Data Pipeline, Ceph-based Storage System, Geospatial Visualization, Generative AI (GenAI)

ACM Reference Format:
Pratik Poudel, Boyuan Guan, Nicole Sanchez, Kiavash Bahreini, Wencong Cui, Andres Lopez, Hamed Najafi, Zhaohui Fu, Leonardo Bobadilla, and Jason Liu. 2025. A Scalable Framework for Heterogeneous Environmental Data Management Using Smart Data Pipeline. In *Practice and Experience in Advanced Research Computing (PEARC '25), July 20–24, 2025, Columbus, OH, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3708035.3736017

1 Introduction

Environmental challenges such as climate change, biodiversity loss, and natural disasters are escalating, necessitating robust, data-driven solutions to understand and mitigate their impacts. The field is increasingly driven by vast and complex datasets originating from remote sensing technologies, sensor networks, and computational models. As environmental challenges grow in scale and urgency, data-driven insights are essential for climate modeling, ecosystem monitoring, disaster response, and sustainability planning. Rapid proliferation of diverse datasets introduces significant challenges in data management, storage, interoperability, and accessibility [37]. Without a structured framework for handling such large-scale heterogeneous data, researchers and stakeholders face difficulties in

efficiently sharing and processing data, exploring correlations, and extracting meaningful insights, and making decisions.

Despite ongoing advancements in data infrastructure, existing data management systems exhibit several limitations. They struggle with automated metadata processing, lack interoperability across disciplines, and fail to provide scalable, efficient workflows for seamless data integration [4]. Studies have identified several challenges in managing large-scale geospatial datasets, including inconsistencies in metadata standards [15], scalability issues in real-time metadata extraction [5], and limited interoperability across Geographic Information System (GIS) platforms [16]. Manual metadata annotation remains a significant bottleneck, as automated solutions struggle with multi-resolution data integration and explainability [40]. Similarly, research on distributed storage solutions has emphasized the trade-offs between scalability and performance, but has not yet fully addressed the need for intelligent automation in data discovery and retrieval [27]. Furthermore, while interdisciplinary collaborations are essential for addressing global environmental challenges, the lack of standardized systems for metadata generation and cross-platform integration remains a major hurdle.

Although existing established data management frameworks, such as Kepler [22] and DataFed [38], are powerful tools for federating scientific data, they primarily function as access and federation middleware, leaving a critical gap between data accessibility and data usability. These systems excel at making distributed data available across institutional boundaries, but do not inherently address the intrinsic quality, formatting, or inconsistent metadata of the data itself. This gap means that researchers, after gaining access, are still burdened with the time-consuming and complex task of data curation before any analysis can begin.

A primary example of this challenge lies in ensuring data quality. Traditional rule-based cleaning methods offer precision but may not be robust, while statistical or probabilistic approaches, like HoloClean [34], handle uncertainties but contend with computational complexity. Machine learning models either require substantial labeled data (in case of supervised learning) or are difficult to interpret (for unsupervised learning), both of which present significant bottlenecks for diverse environmental datasets. A more promising LLM-based paradigm has emerged, leveraging pre-trained knowledge to handle complex semantic and contextual errors. Already frameworks, such as AutoDCWorkflow [21], Cocoon [45], and Retclean [1], have demonstrate the power of LLMs for purpose-driven workflow generation and combining semantic understanding with statistical methods. Our framework is designed to operationalize this approach within a unified ecosystem.

This paper introduces a scalable, AI-enhanced data framework that streamlines data ingestion, processing, and discovery by closing the gap between access and usability. The framework is built on *Envistor*, which facilitates inter- and intra-campus storage for collaborative research in environmental science, particularly for South Florida. Our proposed solution uses Generative AI (GenAI) [31] to automate data curation and metadata creation, ensuring improved data quality and discoverability. A Ceph-based distributed storage system [43] provides fault-tolerant, scalable storage for both structured and unstructured datasets, while integration with geospatial platforms such as ArcGIS Online [7] facilitates advanced visualization and analysis. By creating a unified data federation and curation framework, we overcome the data silos that impede interdisciplinary collaboration on complex environmental issues.

To demonstrate the practicality of our smart data pipeline framework, we present a real-world use case based on data collected from research buoys deployed in South Florida. These buoys monitor key parameters like pH, temperature, salinity, and dissolved oxygen to assess water quality and identify contaminants in both shallow freshwater ecosystems and nearshore marine environments [42]. This example effectively illustrates the efficiency of our framework in data processing, cleaning, and visualization.

Overall, the paper introduces several key contributions which can be summarized as follows:

- A scalable, AI-powered data framework that automates data ingestion, processing, and discovery for heterogeneous environmental datasets;
- Integration of GenAI to enhance metadata creation, enabling efficient data organization and retrieval;
- Implementation of a Ceph-based distributed storage system, ensuring scalable, fault-tolerant, and high-performance storage for diverse data types;
- Interdisciplinary data convergence, facilitating seamless interoperability among research groups working on environmental challenges; and
- Demonstration of real-world applicability through a user case of environmental data monitoring, underscoring the framework's ability to support scientific research and data-driven decision making.

The remainder of the paper is organized as follows. In Section 2, we outline the main design considerations behind the data storage and data management framework. In Section 3, we provide an overview of the system architecture and describe the major components of our implementation. In Section 4, we focus on the smart data pipeline that provides seamless integration with both local and external computational resources, facilitates geospatial data processing and visualization, and incorporates AI and machine learning for data management and content discovery. To illustrate the utility of the proposed data storage and management framework and the smart data pipeline, in Section 5, we describe a study of water quality data collected continuously by several research buoys deployed in South Florida rivers and coastlines. In Section 6, we conclude the paper and provide future directions.

2 Design Consideration

2.1 Heterogeneous Data Sources

Environmental datasets originate from diverse sources, including IoT sensor networks, remote sensing platforms, climate models, and field surveys, each varying in format, spatial-temporal resolution, and metadata structure [36]. Climate models, like CMIP [9], generate high-dimensional datasets that require specialized storage for multi-resolution simulations [24]. In-situ sensor networks, such as ocean buoys and weather stations, demand real-time ingestion and processing for timely analysis [42]. These datasets exist across multiple storage paradigms, including object storage (such as Amazon S3 [20]) and file-based storage (such as NetCDF [25]). A robust data storage system must accommodate heterogeneity to ensure seamless data integration, retrieval, and analysis.

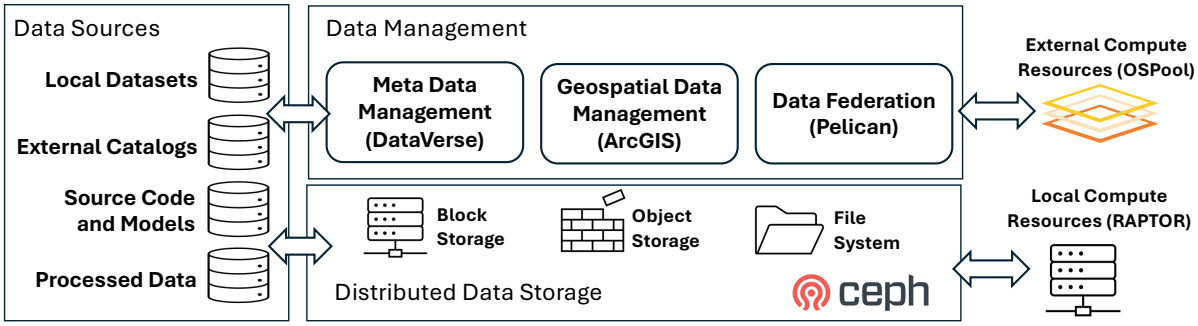

Figure 1: Envistor System Overview

2.2 Automated Metadata Management

Metadata is essential for organizing, retrieving, and integrating environmental datasets, yet manual annotation remains inefficient, especially for large-scale geospatial data. Existing platforms rely on labor-intensive tagging, leading to inconsistencies and inefficiencies, particularly for complex datasets, such as 3D geospatial and remote sensing imagery. The complexity of standardized metadata frameworks further hinders cross-platform interoperability. Naveen Kumar et al. [27] proposed AI-driven automation, including GenAI and feature-learning models, which offers a scalable solution by enhancing real-time processing and improving metadata tagging. However, the lack of standardized integration remains a challenge. A well-structured metadata framework is crucial for seamless data discovery, retrieval, and interdisciplinary collaboration.

2.3 Data Federation

The diverse datasets that environmental research relies on are often scattered across institutions, creating fragmented data ecosystems. The lack of a unified federation framework leads to data silos, inconsistent metadata, versioning conflicts, and redundant storage, thus limiting efficient integration and large-scale analysis. Without interoperable sharing mechanisms, critical datasets remain isolated, delaying real-time decision making in climate mitigation, biodiversity conservation, and disaster response. Federated data systems address these challenges by enabling seamless integration, enhancing accessibility, and fostering interdisciplinary collaboration for comprehensive environmental insights [14].

2.4 Integration with Computational Resources

Environmental data analysis often requires high-performance computational resources due to the large-scale and complex nature of the datasets. Traditional infrastructure is often insufficient to handle data-intensive operations. Examples include climate model downscaling for high-resolution climate projections, remote sensing analysis that involves terabyte- to petabyte-scale imagery data, and AI and machine learning applications for weather forecasting, land cover classification, and anomaly detection. While various computational infrastructures provide on-demand resource allocation for these tasks, they oftentimes operate in isolation, lacking integrated data and metadata management essential for collaboration and reproducibility.

3 System Architecture

The Envistor Data Repository and Discovery System is a data storage platform designed and developed to efficiently manage, process, and analyze heterogeneous datasets pertaining to coastal environmental science. The system ensures data discoverability, seamless application integration, and streamlined access for diverse end users. At a high level, as illustrated in Figure 1, the system incorporates three major components: data sources, data management, and distributed data storage.

3.1 Data Sources

Table 1 shows an example of different datasets we handle. In this section, we provide a description of the local datasets and the associated data types we are dealing with at the university, while the other categories are included and briefly described as well. We discuss the smart data pipeline by introducing the buoy dataset as a use case in a later section.

3.1.1 Local Datasets. Beginning with the data maintained by the university, local datasets encompass a diverse range of environmental, climate, and experimental research data, curated to support large-scale scientific investigations across multiple domains. The data sources include real-time and archival records of atmospheric conditions, marine and coastal monitoring, wind engineering experiments, disaster preparedness, and autonomous navigation. The dataset spans from high-resolution satellite imagery and fluid dynamics (PIV) data to climate model outputs, citizen science observations, and robotic navigation logs.

Storage requirements vary significantly, with some projects needing relatively small allocations (e.g., 5 TB for citizen science flooding data) while others demand extensive storage (e.g., 200 TB for wind research experiments). The data movement characteristics are tailored to the needs of each scientific driver—ranging from real-time ingestion of sensor data in marine and atmospheric monitoring to periodic large-scale uploads of satellite imagery and climate projections. To ensure longevity and accessibility, most data are curated following the FAIR (Findable, Accessible, Interoperable, and Reusable) principles with only a few (e.g., climate projections) adhering to only Findable and Accessible part of the principle [44]. The data are stored in different storage interface formats depending on the nature of the data. For example, the CREST buoy data is

Table 1: Example Datasets

Category	Dataset	Size	Usage	Storage Type	FAIR
Local Datasets	Biscayne Bay Monitoring	20 TB/yr	satellite data, weekly	File,Object	FAIR
	Marine Autonomy	30 TB/yr	marine missions, weekly	File	FAIR
	Topographic & Bathymetry Flood	120 TB/yr	TANDEM satellite collection, yearly	File	FAIR
	Saltwater Intrusion	100 TB/yr	Real-time archival	File,Object	FAIR
	Climate Projections	10 TB	Climate model downloads	File	FA
	Citizen Science (Flooding/Heat)	5 TB	Real-time SECOORA access	Object	FA
	CREST Buoy Data	50 TB	Frequent uploads by users	File,Object	FAIR
	Wall of Wind	200 TB	HD videos	Object	FAIR
Source Code / Models	Arcgis API for python	75MB	Version 2.4.0	File	FAIR
	Dataverse	30MB	Version 6.5	File	FAIR
	Project codes	10 MB	Code revisions, version control	File	FAIR
External Catalogs	CONUS404 Hydro-Climate Data	~10 TB	different time scales (hourly, daily, monthly), on-demand	File	FAIR
	EAR5 Atmospheric Reanalysis, GCM Models	~10 TB	Different resolutions, on-demand		

stored both as objects and files, and the Wall of Wind data is stored only as objects.

3.1.2 External Catalogs. External Catalogs are repositories that store and organize published datasets, often with Digital Object Identifier (DOI) entries [39], making them readily accessible for research and analysis. Similar to platforms like Dataverse [18], GeoPlan [12], and Water Atlas [30], these catalogs provide structured data, including GIS and environmental datasets. They serve as essential resources for researchers by offering curated, high-quality data for spatial analysis, urban planning, and ecological studies.

3.1.3 Source Code and AI/ML Models. Source code and AI/ML model data are typically smaller in size but essential for processing and analyzing existing datasets. This category includes source code, trained AI/ML models from experimental analysis, and GIS tools, such as source code using the ArcGIS API [7], Dataverse scripts [18], and various transformed climate AI models.

3.1.4 Processed Data. A processed dataset is a refined version of raw data (both local and external), enhanced through techniques like interpolation to improve resolution for localized research. For example, climate data from sensors or satellites often require processing to overcome the low resolution of general circulation models (GCMs), which are limited by computational complexity. By generating higher-resolution data, processed datasets can enable more precise analysis and decision making, making them essential for fields like climate modeling and urban planning.

3.2 Data Management

3.2.1 Metadata Management with Dataverse. The Data Management Platform integrates robust systems to facilitate comprehensive metadata management and research data dissemination. Key functionalities include metadata editing and creation, as well as the assignment of DOIs for research datasets, scripts, models, and supporting files. These features ensure dataset discoverability and accessibility across the Dataverse community [23], fostering collaboration and reuse of data in compliance with FAIR principles.

3.2.2 Data Federation Via Pelican. The Pelican Platform, a component of the Open Science Data Federation (OSDF), is an Open Science Grid (OSG) service that facilitates the hosting of data origins and globally distributed caches [26]. This platform enables seamless access, processing, and distribution of benchmark datasets stored within the Envistor's Ceph storage system across the Pelican federated computing network. Through its distributed framework, Pelican ensures data availability via server caches deployed at 28 sites across 19 institutions, complemented by origin servers spanning 13 sites across 11 institutions.

An origin server is responsible for sharing original datasets from an institution or entity with the broader research community, while cache servers host replicated copies of these datasets at geographically distributed locations. This caching mechanism significantly reduces data access latency, thereby enhancing the efficiency of research workflows. By federating data sources, Pelican enables global access to shared datasets, fostering collaboration among researchers worldwide. Pelican caches integrate seamlessly with computational infrastructure, extending their utility beyond data accessibility to facilitate large-scale computation. Researchers can leverage OSDF's computational resource allocators to conduct data-intensive analyses, benefiting from high-performance computing resources integrated into the OSDF ecosystem.

The Envistor System further enhances this capability by incorporating its own origin server, ensuring that datasets stored at Florida International University (FIU) are accessible to external researchers in compliance with FAIR principles. This integration allows external users to leverage OSDF infrastructure, including computational resources such as OSPool [32] and Chameleon Cloud [17], to perform advanced data processing and analysis.

3.2.3 Geospatial Data with ArcGIS. ArcGIS is a versatile Geographic Information System designed to support the creation of interactive maps, conduct spatial analysis, and handle geospatial data. Geospatial data is easily integrated into ArcGIS using a variety of tools, workflows, and platforms. Data from sources, such as satellite imagery and GPS devices, and external databases can be imported, processed, and analyzed within the ArcGIS environment. ArcGIS consists of several components. *ArcGIS Online* serves as the foundation for its cloud-based mapping and sharing capabilities [7]. It uses Role-Based Access Control, allowing administrators to define user roles and manage access to geographic data and services within the organization. For advanced geospatial tasks, *ArcGIS Pro* serves

as a desktop application for complex data processing and supports a wide range of data formats, enabling users to analyze spatial data, create maps, and manage geographic information effectively [8].

Supporting sub-components, including version control servers (e.g., AWS CodeCommit [2]) and web applications and digital twin platforms (e.g., NVIDIA OmniVerse [28]), provide enhanced functionality for data exploration, real-time analysis, and collaboration, enabling end users to utilize data more effectively for decision making.

3.3 Distributed Data Storage

3.3.1 Ceph Storage System. Ceph is a highly scalable, software-defined distributed storage system designed for fault tolerance, high availability, and efficient data distribution [43]. It operates on the Reliable Autonomic Distributed Object Store (RADOS), which replicates and distributes data dynamically across nodes using the Controlled Replication Under Scalable Hashing (CRUSH) algorithm. This decentralized architecture eliminates single points of failure and allows seamless scalability by adding storage nodes as needed. Unlike traditional storage solutions, Ceph supports object, block, and file storage, making it a versatile choice for cloud computing, high-performance computing (HPC), and large-scale data analytics.

A key advantage of Ceph is its self-healing and self-managing capabilities, which reduce administrative overhead while maintaining data integrity. It continuously monitors cluster health, automatically redistributing data in case of node failures to minimize downtime and ensure reliability. Ceph can also be deployed on commodity hardware, significantly reducing costs compared to proprietary storage solutions. However, despite its advantages, Ceph requires careful configuration and performance tuning, particularly in high-performance storage environments where resource management is critical. Nevertheless, its ability to provide scalable, cost-effective, and fault-tolerant storage has made it widely adopted in modern cloud and containerized infrastructures.

Initially deployed as a low-latency 20 TB storage prototype, our system was later expanded to 4 PB to accommodate increasing data demands. The prototype validated Ceph's high availability, fault tolerance, and scalability, ensuring continuous data availability through three-way replication across the cluster. Over time, Ceph evolved to store diverse datasets (with examples of local datasets shown in Table 1) across all three storage formats. In its early stages, for example, the system primarily housed buoy data using both file and object storage interfaces. Since buoy data is structured (tabular) data, it could be stored in either format, demonstrating Ceph's flexible storage capabilities.

Ceph's Amazon S3-compatible object storage enabled standardized data access and integration with technologies like Trino [41], which enhances object-read latency for tabular data. This integration facilitates faster data retrieval and improved performance, as a versatile solution for large-scale storage needs.

3.3.2 Computational Resources via OSPool and RAPTOR. The storage system hosts a diverse range of datasets that support research, data processing, computational analysis, and modification. These datasets are accessed by both internal and external users, each potentially with distinct requirements. Internal users, primarily within FIU, work with datasets that are not yet publicly available but must remain within the storage infrastructure for ongoing research. To process these datasets efficiently, project members require a computational environment capable of handling large and diverse data volumes.

Similarly, external users located outside FIU rely on data shared by the institution via Pelican for their research needs. To accommodate both internal and external users, the framework integrates scalable computing resources alongside its storage infrastructure, facilitating advanced data analysis, task execution, and high-performance computing (HPC) workloads. Two key computing platforms, RAPTOR (Reconfigurable Advanced Platform for Transdisciplinary Open Research) [11] and OSPool [32], serve these users by providing access to computational resources tailored to their specific requirements.

This integration enhances flexibility through dynamic, on-demand resource allocation. OSPool supports high-performance data processing through a job-submission model, where researchers request computational resources (with CPU and GPU specifications) as needed. As part of OSG, OSPool leverages contributed infrastructure from participating universities, offering a wide range of systems to support diverse computational requirements. RAPTOR, on the other hand, provides bare-metal access, enabling greater control over the computing environment. Researchers using RAPTOR share a leased infrastructure that consists of compute nodes, each equipped with up to three A100 GPUs, and use the hardware for compute-intensive tasks for a specified duration before relinquishing access to the next user.

This unified framework, which can be expanded in the future to include more computing and storage capabilities, ensures on-demand resource allocation, bare-metal configuration, and seamless scalability for distributed high-throughput computing capabilities. It supports research in AI, machine learning, and model optimization. By integrating storage with flexible computing resources, it establishes a robust foundation for future research and collaboration, benefiting both internal and external users.

4 Smart Data Pipeline

The Smart Data Pipeline leverages the system architecture and intermingles with the components to handle the data—from discovering data to data transformation workflow to performing computation, visualizing and federating the data/insights with the community. This Smart Data Pipeline facilitates end-to-end workflows for managing, transforming, and disseminating data efficiently.

4.1 AI Data Agent and Automated Workflow

The Smart Data Pipeline orchestrates automated workflows for seamless data processing, managed by an AI data agent. This agent, powered by tools like LangChain [19] and GenAI APIs (e.g., [31]), automates critical tasks including data cleaning, cataloging, formatting, and metadata creation, streamlining the preparation of datasets for publication. This workflow takes the data from the ingestion process and creates clean and publishable dataset that is ready to be shared using a generated DOI. The ingestion process includes handling data acquisition from various sources, ensuring data quality through automated validation checks, and standardizing datasets through Extract, Transform, Load (ETL) pipelines [44].

The ETL process enables structured and unstructured data to be transformed into standardized formats, facilitating integration with metadata management systems, such as Dataverse, which we chose to use for our system. Once processed, the datasets are safely stored and linked with metadata to ensure discoverability.

A primary function of the AI data agent is to ensure data quality with data cleaning. For data cleaning, the agent incorporates RetClean, a validated, LLM-powered tool [1]. We selected RetClean because its support for Retrieval-Augmented Generation (RAG) aligns with our direction in building a custom vector knowledge base from our data corpus, and its effectiveness is documented. By integrating with specialized tools like RetClean, the AI data agent can handle data cleaning while managing the broader workflow, from data ingestion and transformation into standardized formats to linkage with metadata management systems like Dataverse.

Furthermore, real-time monitoring and alerting mechanisms help promptly address failures or anomalies, maintaining a reliable end-to-end data ingestion process. By automating these traditionally manual workflows, the smart data pipeline can significantly reduce the effort required to prepare and clean datasets, which aligns with best practices in research data management [13].

4.2 Discovering Content through Dataverse

To facilitate efficient data discovery and sharing, the system integrates with Dataverse, a widely adopted open-source repository designed for managing and disseminating research data [23]. FIU has its own in-house dataverse instance where users can explore datasets and metadata through an intuitive interface, allowing easy access to curated resources for researchers, thus enabling efficient data exploration and reuse [10]. Dataverse supports features such as dataset versioning, metadata enrichment, and access control, making it a reliable solution for both individual researchers and large-scale collaborative projects. By indexing datasets with rich metadata, users can search, retrieve, and cite data, fostering reproducibility and knowledge dissemination across disciplines.

4.3 Integration with Computational Resources

In addition to generating clean, publishable datasets, the smart data pipeline integrates computational resources to serve both external and internal users for efficient data processing of tasks, such as modeling and simulation, machine learning, and data visualization. In particular, external users can use OSPool, provided by the Open Science Grid Foundation [32], which offers a SLURM-like job submission system [35] for resource reservation, scheduling, and automated notifications. Meanwhile, FIU researchers have access to RAPTOR, a dedicated HPC cluster integrated with JupyterHub [33], where users are assigned to projects and share computational environments without duplicating datasets or reconfiguring software. Each user maintains a private scratch space that remains isolated unless explicitly shared. By consolidating data and computation within a single framework, the smart data pipeline ensures reproducibility, consistent run-time environments, and streamlined collaboration, allowing users to build on one another's work with reduced overhead.

4.4 Visualization and Exploration through GIS

To enhance data exploration, the system is integrated with GIS applications, such as digital twins, ArcGIS Online, and ArcGIS Portal. By combining metadata management from Dataverse with the advanced visualization capabilities of GIS tools, the system bridges the gap between data cataloging and real-time geospatial analysis.

These GIS applications support data-driven decision making in various domains, including urban planning, environmental monitoring, and disaster response. Digital twins facilitate dynamic simulations for trend analysis and resource optimization [46], while ArcGIS Online and ArcGIS Portal centralize geospatial data management, enabling the creation, analysis, and dissemination of advanced mapping products [6].

Furthermore, ArcGIS Dashboards and digital twins enhance interactivity, allowing users to dynamically explore spatial trends, simulate environmental changes, and gain deeper insights into complex datasets. By integrating metadata engines, such as Dataverse, with GIS systems, the framework promotes real-time data sharing, automated geospatial layer publication, and enhanced visualization capabilities. These features empower the researchers and other stakeholders to engage in more effective, data-driven decision making across a wide range of disciplines.

5 A Case Study

5.1 Buoy Data Workflow

To demonstrate the smart data pipeline's capabilities, we executed a workflow on a real-world dataset from South Florida, comprising 11 raw data files from 10 distinct buoy locations. The pipeline processed a total of 496,292 rows of sensor data collected from October 2018 to May 2024. The buoy data exemplifies the smart data pipeline's capability to automate data processing on multiple platforms for analysis and visualization. Figure 2 represents the workflow. The smart data pipeline in this case uses OpenAI API, LangChain, and Python scripts to clean the data, format the data, and generate metadata.

5.1.1 Notification System. The smart data pipeline can help create different data summaries that give quick insights on the data that user is dealing with. Although this is rather insignificant in the implementation, these digestible summaries can help the user understand the data before performing further processing and analysis. The system incorporates an email notification system to deliver the summary reports to target users upon request.

5.1.2 GIS Integration and Visualization. The pipeline generates heatmaps, shapefile, etc., by publishing the processed data to the ArcGIS Online portal [6]. In particular, we use ArcGIS API for Python for spatial data pre-processing and heatmap generation. ArcGIS API is a powerful library that enables seamless interaction with ArcGIS services for spatial analysis, data management, and visualization. Such enhancements improve the accuracy and usability of geospatial data for analysis. Once processed, the datasets are published, where they become accessible through a variety of interactive tools, including dashboards, story maps, and web maps [7]. These visualization tools support data-driven decision making in various domains, including climate research, ocean monitoring, and

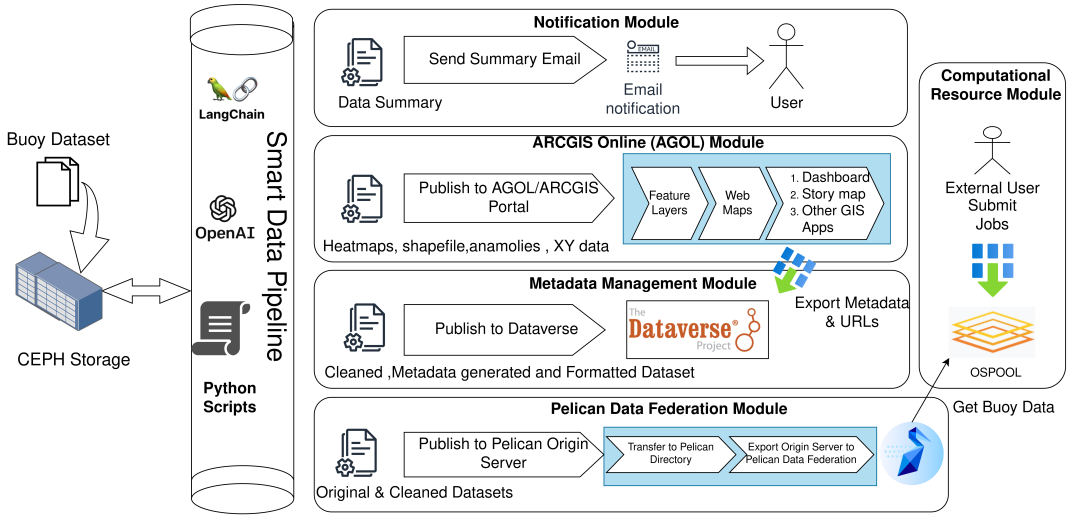

Figure 2: Buoy Data Workflow.

disaster response. The pipeline also generates metadata and URLs ready to be exported to Dataverse for metadata management.

5.1.3 Metadata Management. The pipeline ingests raw buoy datasets stored in Ceph, with Dataverse playing a central role in automating metadata creation. Each dataset is assigned a Digital Object Identifier (DOI), ensuring persistent and unique identification for enhanced discoverability and citation. This structured metadata framework facilitates centralized dataset management, making it easier for researchers to search, access, and reference the data. By automating metadata generation, the system ensures compliance with best practices in FAIR principles. In addition to the raw datasets, Dataverse is also capable of creating catalogs from the exported metadata and URLs from ArcGIS.

5.1.4 Data Processing and Federation. To maintain consistency between stored datasets, their corresponding metadata and federated access to it, the pipeline integrates Pelican Platform, which facilitates seamless linkage between Ceph and Dataverse with outside world. Pelican integration with the framework ensures efficient data federation, synchronizing original, processed datasets with their respective metadata records. This integration streamlines the transfer of refined datasets for visualization, publication, and further computational analysis. By leveraging this federation mechanism, users can seamlessly access the latest processed data while preserving historical versions for provenance and reproducibility.

5.1.5 AI and Computational Resources. The raw buoy dataset, stored in Ceph Storage, requires cleaning, formatting, and metadata generation before publication on OSPool. An AI-driven pipeline, leveraging Python scripts, OpenAI API, and LangChain, automates the process, which also include DOI assignment. Once processed, the dataset is made available to OSPool through the Pelican Data Federation, accessible via a URL endpoint. This integration enables external researchers to execute computing and data analysis tasks on the dataset as if it were locally available, eliminating the need for manual data transfers. Researchers can leverage OSPool's SLURM-like job submission system to schedule programs and perform data analysis on the buoy dataset. This capability lays the foundation for advanced modeling of environmental patterns, oceanographic simulations, and data-driven visualizations, expanding the potential applications of the dataset in scientific research.

5.1.6 End-User Accessibility. End users can explore and access buoy datasets and visualizations through various interfaces, such as Dataverse, Pelican Protocol, and GIS applications. This multi-platform approach ensures broad accessibility, allowing users to interact with data in ways that cater to their specific needs. Researchers can retrieve datasets for in-depth analysis, policy-makers can utilize GIS visualization for informed decision-making, and educators can employ interactive maps for instructional purposes. By supporting diverse use cases, the system facilitates collaboration across disciplines, making scientific data more readily accessible.

5.2 Buoy Data Cleaning

To showcase the practical extensibility of the Smart Data Pipeline, we adapted the methodology from RetClean [1] for data cleaning and applied it to the buoy data use case. We used the native user interface to demonstrate the LLM capacity for context-aware data fixing capacity. We used RetClean's baseline LLM-powered cleaning function to identify and repair anomalies such as out-of-range sensor values and inconsistent data formatting. It used world knowledge and considered features like location and time to dynamically provide an estimate for the inconsistent data.

Figure 3 demonstrates the effectiveness of RetClean in identifying and repairing anomalous salinity (*SAL-ppt*) readings from real-time water quality data collected in a coastal estuarine environment. The tool successfully detected all invalid or suspicious entries and generated contextual estimates based on environmental factors and geolocation. Here, we highlight two examples to illustrate RetClean's reasoning capabilities:

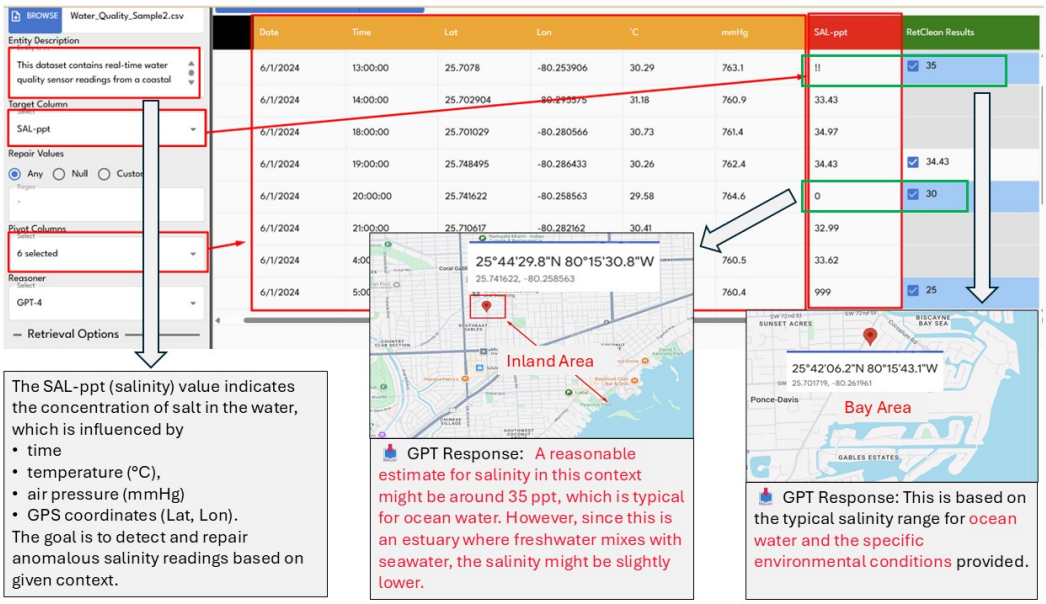

Figure 3: Retclean Data cleaning

- **Inland Area – Invalid Zero Value:** A salinity reading of 0 ppt in an estuarine region is highly unlikely. RetClean correctly recognized this as an invalid input and estimated the salinity to be 30 ppt, considering the inland location and typical freshwater-seawater mixing conditions.
- **Bay Area – Missing Value ("!!"):** For a coastal reading with missing salinity, RetClean inferred a likely value of 35 ppt, based on standard ocean water conditions during summer in Miami, with regular temperature (30.91°C) and air pressure.

While the inferred values may not be perfect, RetClean consistently provided reasonable and justifiable estimations, demonstrating its ability to leverage geospatial awareness and environmental cues for robust data repair.

6 Conclusion and Future Plan

This paper proposes a smart data pipeline framework designed to enhance environmental research. The system integrates Ceph-based storage, AI-driven metadata generation, and geospatial visualization. By automating metadata creation and ensuring interoperability, the framework minimizes manual effort and enhances data discoverability.

A case study involving buoy data demonstrates the scalable and extensible architecture of this data management framework. We successfully integrated the RetClean tool for buoy data cleaning, showcasing the framework's modular design and its readiness to incorporate other advanced AI tools. As the framework evolves, it will remain a versatile tool for climate modeling, disaster response, and sustainability planning. Its ability to expand to new use cases is expected to further enhance its adaptability, enabling broader dataset integration and improved efficiency.

Our immediate future work involves extending the smart data pipeline to our new production Ceph cluster from OSNexus. This cluster ensures high redundancy and reliability to safeguard against failures. We will focus on four key tasks. First, we plan to expand dataset integration to include more environmental research data. This expansion will broaden the system's applicability in climate research and environmental monitoring. Second, we will implement optimized caching mechanisms and expanded node configurations to enhance data access times and enable better scalability. Third, we will complete the development of our internal vector database using existing project data and documentation. This will allow us to fully leverage the Retrieval-Augmented Generation (RAG) capabilities of integrated tools like RetClean. To assess its accuracy, we will conduct a comprehensive performance evaluation. Finally, we will implement a structured user management system, such as SAML [29] and CILogon [3], for secure, project-based data access. This system will provide fine-grained permission control, ensuring efficient collaboration while maintaining data security.

Acknowledgments

This work was partially funded by the National Science Foundation through grants OAC-2322308 and IIS-2331908. We would like to express our gratitude to our collaborators who participated in data collection and management. Additionally, we would like to thank the OSPool team for their assistance in integrating the Pelican implementation with our system.

References

[1] Mohammad Shahmeer Ahmad, Zan Ahmad Naeem, Mohamed Eltabakh, Mourad Ouzzani, and Nan Tang. 2023. RetClean: Retrieval-based data cleaning using foundation models and data lakes. *arXiv preprint arXiv:2303.16909* (2023).
[2] Amazon Web Services. 2025. AWS CodeCommit. https://aws.amazon.com/codecommit/
[3] Jim Basney, Terry Fleury, and Jim Gaynor. 2013. CILogon: A federated X.509 certification authority for cyberinfrastructure logon. *XSEDE '13: Proceedings of the Conference on Extreme Science and Engineering Discovery Environment* (2013).

doi:10.1145/2484762.2484791
[4] Rebekka Benfer and Jochen Müller. 2024. Semantic digital twin creation of building systems through time series based metadata inference–A review. *Energy and Buildings* (2024), 114637.
[5] Laura Díaz, Carlos Granell, Michael Gould, and Joaquín Huerta. 2011. Managing user-generated information in geospatial cyberinfrastructures. *Future Generation Computer Systems* 27, 3 (2011), 304–314.
[6] Environmental Systems Research Institute (ESRI). 2025. *ArcGIS API for Python*. https://developers.arcgis.com/python/latest/
[7] Environmental Systems Research Institute (ESRI). 2025. ArcGIS Online. https://www.esri.com/en-us/arcgis/products/arcgis-online
[8] Environmental Systems Research Institute (ESRI). 2025. ArcGIS Pro. https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview
[9] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. 2016. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development* 9, 5 (2016), 1937–1958. doi:10.5194/gmd-9-1937-2016
[10] Florida International University. 2025. FIU Dataverse. https://dataverse.fiu.edu/
[11] Florida International University. 2025. Raptor-Reconfigurable Advanced Platform for Transdisciplinary Open Research. https://raptor.fiu.edu.
[12] University of Florida GeoPlan Center. 2022. Florida Geographic Data Library (FGDL). https://fgdl.org.
[13] Harvard University. 2025. Harvard Dataverse. https://dataverse.harvard.edu/
[14] Jeffery S Horsburgh, David G Tarboton, Michael Piasecki, David R Maidment, Ilya Zaslavsky, David Valentine, and Thomas Whitenack. 2009. An integrated system for publishing environmental observations data. *Environmental Modelling & Software* 24, 8 (2009), 879–888.
[15] Fei Hu, Mengchao Xu, Jingchao Yang, Yanshou Liang, Kejin Cui, Michael M Little, Christopher S Lynnes, Daniel Q Duffy, and Chaowei Yang. 2018. Evaluating the open source data containers for handling big geospatial raster data. *ISPRS International Journal of Geo-Information* 7, 4 (2018), 144.
[16] Mohsen Kalantari, Hamed Olfat, and Abbas Rajabifard. 2010. Automatic spatial metadata enrichment: reducing metadata creation burden through spatial folksonomies. In *Global Spatial Data Infrastructures 12 World Conference: Realising Spatially Enabled Societies*. Citeseer.
[17] Kate Keahey, Jason Anderson, Paul Ruth, Jacob Colleran, Cody Hammock, Joe Stubbs, and Zhuo Zhen. 2019. Operational Lessons from Chameleon. In *Proceedings of the Humans in the Loop: Enabling and Facilitating Research on Cloud Computing (HARC '19)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3355738.3355750
[18] Gary King. 2007. An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods & Research* 36, 2 (2007), 173–199.
[19] LangChain Contributors. 2023. LangChain. https://github.com/hwchase17/langchain.
[20] Thomas J. Leeper. 2020. *aws.s3: AWS S3 Client Package*. R package version 0.3.21.
[21] Lan Li, Liri Fang, and Vetle I Torvik. 2024. AutoDCWorkflow: LLM-based data cleaning workflow auto-generation and benchmark. *arXiv preprint arXiv:2412.06724* (2024).
[22] Bertram Ludascher, Ilkay Altintas, Chad Berkley, Daniel Higgins, Efrat Jaeger, Matthew B. Jones, Edward A. Lee, Jing Tao, and Yang Zhao. 2006. Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience* 18, 10 (2006), 1039–1065. doi:10.1002/cpe.994
[23] D-Lib Magazine. 2011. The dataverse network: an open-source application for sharing, discovering and preserving data. *D-lib Magazine* 17, 1/2 (2011).
[24] Gerald A Meehl, Richard Moss, Karl E Taylor, Veronika Eyring, Ronald J Stouffer, Sandrine Bony, and Bjorn Stevens. 2014. Climate model intercomparisons: Preparing for the next phase. *Eos, Transactions American Geophysical Union* 95, 9 (2014), 77–78.
[25] Mike Mesnier, Gregory R Ganger, and Erik Riedel. 2003. Object-based storage. *IEEE Communications Magazine* 41, 8 (2003), 84–90.
[26] Morgridge Institute for Research. 2025. Pelican Platform. https://pelicanplatform.org/
[27] R Naveenkumar, Shantanu Bhadra, L Haldurai, S Visalakshi, and Nitin Kumar. 2024. Environmental Artificial Intelligence paving the way for a greener and more resilient planet using machine learning. *Journal of Computational Analysis and Applications* 33, 7 (2024).
[28] NVIDIA Omniverse Development Team. 2025. NVIDIA Omniverse. https://www.nvidia.com/en-us/omniverse/
[29] OASIS. 2005. SAML 2.0. https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security.
[30] University of South Florida Water Institute. 2025. Water Atlas. https://wateratlas.usf.edu.
[31] OpenAI. 2023. OpenAI API. https://openai.com/api/.
[32] OSG. 2006. OSPool. doi:10.21231/906P-4D78
[33] Project Jupyter. 2023. JupyterHub. https://jupyterhub.readthedocs.io/en/stable/.
[34] Theodoros Rekatsinas, Xu Chu, Ihab F Ilyas, and Christopher Ré. 2017. Holoclean: Holistic data repairs with probabilistic inference. *arXiv preprint arXiv:1702.00820* (2017).
[35] SchedMD. 2002. SLURM: Simple Linux Utility for Resource Management. https://slurm.schedmd.com. Accessed: 2025-02-02.
[36] John L Schnase, Daniel Q Duffy, Glenn S Tamkin, Denis Nadeau, John H Thompson, Cristina M Grieg, Mark A McInerney, and William P Webster. 2017. MERRA analytic services: Meeting the big data challenges of climate science through cloud-enabled climate analytics-as-a-service. *Computers, Environment and Urban Systems* 61 (2017), 198–211.
[37] John L Schnase, Glenn Tamkin, David Fladung, Scott Sinno, and Roger Gill. 2011. Federated observational and simulation data in the NASA Center for Climate Simulation data management system project. In *Proceedings of the iRODS user group meeting 2011: Sustainable policy-based data management, sharing, and preservation*. 17–18.
[38] Suhas Somnath, Scott Klasky, Glenn S. Lockwood, Thomas R. Maier, Marcus D. Hanwell, Steven P. Miller, Alexander J. McCaskey, John L. Turner, John M. Campbell, Mark R. Berrill, Thomas R. Evans, Michael R. Matheson, Thomas R. Womack, Nicholas D. Darnell, Thomas R. Allison, Thomas R. Williams, Thomas R. Baird, Thomas R. Miller, Thomas R. Maier, Thomas R. Evans, and Thomas R. Williams. 2019. DataFed: Towards reproducible research via federated data management. In *Proceedings of the 6th Annual Conference on Computational Science and Computational Intelligence (CSCI 2019)*. IEEE, 1312–1317. doi:10.1109/CSCI49370.2019.00245
[39] Joan Starr, Eleni Castro, Merc Crosas, Michel Dumontier, Robert Downs, Ruth Duerr, Laurel Haak, Melissa Haendel, Ivan Herman, Simon Hodson, Joe Hourclé, John Kratz, Jennifer Lin, Lars Nielsen, Amy Nurnberger, Stefan Pröll, Andreas Rauber, Simone Sacchi, Arthur Smith, and Tim Clark. 2015. Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science* 1 (05 2015). doi:10.7717/peerj-cs.1
[40] Adam J Stewart, Caleb Robinson, Isaac A Corley, Anthony Ortiz, Juan M Lavista Ferres, and Arindam Banerjee. 2022. Torchgeo: deep learning with geospatial data. In *Proceedings of the 30th international conference on advances in geographic information systems*. 1–12.
[41] Trino Software Foundation. 2025. Trino, a query engine that runs at ludicrous speed. https://trino.io/
[42] Kassidy Troxell, Bradley Schonhoff, Mark Kershaw, Milena Ceccopieri, Todd Crowl, and Piero Gardinali. 2024. Near real-time monitoring of the Biscayne Bay Watershed (South Florida, USA) and major tributaries by water quality research buoys. *Science of The Total Environment* 955 (2024), 177203.
[43] Sage Weil, Scott A Brandt, Ethan L Miller, Darrell DE Long, and Carlos Maltzahn. 2006. Ceph: A scalable, high-performance distributed file system. In *Proceedings of the 7th Conference on Operating Systems Design and Implementation (OSDI'06)*. 307–320.
[44] Mark D. Wilkinson, Michel Dumontier, IJsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bosco da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (2016), 160018. doi:10.1038/sdata.2016.18
[45] Shuo Zhang, Zezhou Huang, and Eugene Wu. 2024. Data cleaning using large language models. *arXiv preprint arXiv:2410.15547* (2024).
[46] Yan Zhang. 2024. *Digital Twin: Architectures, Networks, and Applications*. Springer Nature Switzerland, Cham. doi:10.1007/978-3-031-51819-5