Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 3 | Mar-2026
ISSN (O) 3107-6696

AQUA DEEPAI Powered ARGO Oceanographic Data System

[1]Ms. Ranjani C, [2]Ms. Vanitha S, [3]Ms. Sudarmani C, [4]Ms. Shakila Banu F

[1,2,3] Department of Artificial Intelligence and Data Science CARE College of Engineering, Trichy, Tamil Nadu.

[4]Assistant Professor, Department of Artificial Intelligence and Data Science CARE College of Engineering, Trichy, Tamil Nadu.

Abstract - This research paper introduces **AquaDeep**, an artificial intelligence-powered ocean monitoring platform designed to analyze complex marine telemetry and predict environmental variables in real-time. Since critical indicators like temperature, salinity, and pressure are essential for monitoring climate change and ecosystem health, this study addresses the accessibility gap caused by traditional, manual processing of large **NetCDF** datasets. The core objective is to design and implement an automated **ETL pipeline** that extracts key subsurface parameters from the global **ARGO network** and utilizes **Random Forest Regression** to classify ocean states and identify anomalies. By integrating **Retrieval-Augmented Generation (RAG)**, the system successfully converts raw scientific records into actionable natural language insights, fulfilling the goal of developing a scalable, real-world monitoring platform that democratizes access to ocean intelligence for maritime stakeholders and sustainable fisheries management.

I. INTRODUCTION

Ocean state data acts as the primary indicator for the health of global climate systems and marine biodiversity. Sudden shifts in parameters like thermocline depth or salinity are critical early signs of environmental stress. However, traditional oceanographic analysis relies on the manual processing of complex **NetCDF** files, which remains inaccessible to non-technical stakeholders. To address this, this research introduces **AquaDeep**, an AI-powered platform designed for the automated retrieval and interpretation of **ARGO float** telemetry.

The core objective of this system is to implement an automated ETL pipeline that processes multidimensional ocean data and utilizes **Random Forest Regression** to classify ocean states and identify anomalies. By integrating a **Retrieval-Augmented Generation (RAG)** framework, **AquaDeep** converts raw scientific telemetry into intuitive, natural language responses. This architecture provides a scalable, real-world monitoring solution that democratizes access to ocean intelligence, enabling researchers and maritime stakeholders to track marine fluctuations and ensure high-fidelity environmental monitoring through an intelligent conversational interface.

III. RELATED WORKS

The landscape of marine data analysis has historically been defined by significant technical barriers that hinder real-time utility and accessibility. Due to the absence of seamless real-time data integration and automated ingestion pipelines, many existing oceanographic models remain stagnant, unable to provide live monitoring or process multidimensional datasets with the efficiency required for modern climate science [1]. This complete dependence on manual data retrieval and static file processing creates a bottleneck, where earlier platforms require extensive human intervention for every update, rendering them largely unsuitable for the rapid environmental decision-making needed during maritime anomalies [2].

Furthermore, while the parameters collected in previous studies successfully analyzed historical patterns, these systems often lacked the robust machine learning architectures, such as ensemble learning or neural networks, required to predict subsurface temperature variations accurately across stratified depth layers [3]. Many of these models were further constrained by being designed only for localized, high-latitude regions using controlled, small-scale datasets; this lack of architectural flexibility prevents such systems from generalizing

across diverse and dynamic geographic zones like the Indian Ocean [4].

Existing systems also frequently suffered from a narrow diagnostic focus, concentrating on single parameters like surface salinity while failing to synthesize multiple data sources such as pressure, temperature, and geospatial telemetry into a unified view. Consequently, they cannot perform a comprehensive analysis of deep-sea anomalies or provide a bridge for non-technical users to interpret

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 3 | Mar-2026
ISSN (O) 3107-6696

complex scientific outputs. The lack of an intelligent conversational layer or a Retrieval-Augmented Generation (RAG) framework in these legacy systems has maintained a persistent "data silo," where critical marine insights remain locked behind specialized programming requirements [5].

IV. SYSTEM MODEL

The AquaDeep platform is architected as a cohesive, multi-layered ecosystem that bridges the gap between raw marine telemetry and actionable user insights. The system is divided into functional modules that handle the entire data lifecycle—from the initial acquisition of deep-sea sensor data to the final delivery of natural language responses through a web interface.

1. Data Acquisition and Ingestion: At the foundational level, the system performs Data Ingestion by utilizing automated Python-based synchronization scripts. These scripts establish a daily connection with global ARGO float repositories to download raw NetCDF (Network Common Data Form) files. To ensure high-fidelity telemetry for the predictive models, this module executes rigorous preprocessing, which includes noise filtering, outlier detection, and the handling of missing data points caused by sensor transmission gaps in remote oceanic regions.

2. Database Management and Vector Storage: The storage layer is a hybrid architecture designed to handle both structured scientific data and unstructured contextual information. For structured geospatial coordinates and physical ocean parameters (like temperature and pressure), the system employs PostgreSQL enhanced with the PostGIS extension. In parallel, to power the conversational AI, ChromaDB is utilized as a specialized Vector Store. This component converts oceanographic metadata into high-dimensional embeddings, allowing the system to perform rapid semantic searches to find the most relevant historical or real-time context for user queries.

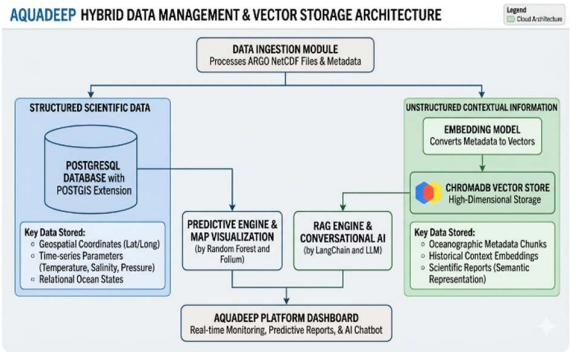


Figure 1:DB and Vector Storage

3. Predictive Modeling and Machine Learning: The analytical core of the platform is the Predictive Machine Learning Engine. This module features a Random Forest Regressor that has been trained on extensive historical datasets of salinity, pressure, and depth. Its primary function is to predict subsurface temperature profiles, which is essential for filling "data holes" in geographic areas where physical sensors are sparse or temporarily inactive. This allows for a continuous, uninterrupted map of the ocean state.

4. The Intelligence Layer: RAG Engine: The "brain" of the platform is the Retrieval-Augmented Generation (RAG) engine, built using the LangChain framework. When a user asks a question, this module retrieves specific data "chunks" from the ChromaDB vector store and feeds them into a Large Language Model (LLM). This process ensures that the AI's responses are not based on general knowledge but are grounded in actual, real-time oceanographic data, allowing for high-fidelity natural language interpretation of complex marine science.

5. Backend Orchestration and API Services: Connecting these modules is a high-performance FastAPI Backend Service. This serves as the communication hub of the ecosystem, exposing secure endpoints that allow the frontend to communicate with the ML models and databases. By managing the asynchronous flow of data, the backend ensures low-latency responses, even when processing large-scale geospatial queries or complex AI inferences.

6. Geospatial Visualization and Mapping: To provide an intuitive user experience, the Geospatial Visualization Module renders interactive maps using the Leaflet and Folium libraries. This component translates raw coordinate data into visual float trajectories and dynamic heatmaps. These visualizations allow stakeholders to see precisely where environmental shifts are occurring, providing a

Ocean State	Temperature Range (°C)	Salinity Range (PSU)	Pressure Range (dbar)	Status/Condition
Normal	15.0 – 28.0	34.0 – 35.5	0 – 2000	Balanced
Thermocline Shift	10.0 – 14.9	33.5 – 34.5	500 – 1500	Anomalous
High Salinity	20.0 – 25.0	> 36.0	0 – 500	Warning
Deep-Sea Cold	< 4.0	34.5 – 35.0	> 2000	Stable
Surface Anomaly	> 29.0	< 33.0	0 – 100	Critical

spatial context that is often lost in traditional spreadsheets or text reports.

7. Frontend Interface and User Management: The user-facing layer is a streamlined Streamlit Dashboard that integrates the chat interface, real-time map views, and analytical charts into a single GUI. Users can seamlessly switch between monitoring live sensor feeds and viewing AI-generated predictive reports. To protect the integrity of the data and provide personalized experiences, a secure Authentication & User Management module handles sessions and permissions, allowing researchers to save specific queries and export historical reports for further study.

Table I: Performance Results

IV. 1. Data Acquisition and Dataset Construction:

The foundational data for this research was sourced from the Indian National Centre for Ocean Information Services (INCOIS) and the Global Data Assembly Centre (GDAC). The primary data format utilized was NetCDF (Network Common Data Form), which allows for the storage of multidimensional scientific variables. The dataset consists of vertical profiles—temperature, salinity, and pressure—captured by the ARGO float network across the Indian Ocean. Initially, these were scattered as individual profile files; we implemented a merging algorithm to compile them into a unified, multi-parameter time-series dataset suitable for machine learning training.

IV. 2. Data Preprocessing

The raw oceanographic telemetry underwent a rigorous cleaning phase to ensure data integrity. To address Missing Data Treatment, we addressed transmission packet loss—a common issue in remote ARGO float telemetry. For minor gaps, we applied linear interpolation, while larger segments were handled via median imputation to maintain the physical consistency of the water column.

To ensure model stability, we performed Feature Standardization. This prevents high-magnitude features like Pressure from biasing the model over smaller-scale features like Salinity. We transformed the attributes using the Z-Score Normalization formula:

$$Z = \frac{x-\mu}{\sigma}$$

Where x is the raw feature value, μ represents the feature mean, and σ denotes the standard deviation. Following this, the dataset was partitioned using an 80/20 Train-Test Split with shuffled sampling to ensure the model generalizes across diverse seasonal cycles and geographic coordinates.

IV. 3. Feature Engineering and Dimensionality Reduction

Maximizing model efficiency required identifying the primary drivers of oceanic change. We employed Recursive Feature Elimination (RFE) to rank the importance of metadata (latitude, longitude, timestamp) and physical parameters, iteratively removing the least significant features. Additionally, we conducted a Correlation Matrix Analysis using the Pearson Correlation Coefficient (r) to identify redundancies:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

This allowed us to eliminate multi-collinearity between depth-dependent variables, reducing the risk of overfitting.

IV. 4. Model Implementation and AI Integration

The Random Forest (RF) regressor was implemented as an ensemble of 100 decision trees to capture the non-linear thermal gradients of the Thermocline layer. The model was optimized by minimizing the Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

To bridge the gap between raw data and human understanding, a Retrieval-Augmented Generation (RAG) pipeline was

integrated. This involved converting oceanographic metadata into high-dimensional vectors via HuggingFace Embeddings, storing them in ChromaDB, and utilizing LangChain to synthesize scientifically grounded responses from an LLM.

IV. 5. Evaluation Metrics

The predictive accuracy and AI reliability were validated using a multi-metric approach. We utilized the Coefficient of Determination (R^2) to evaluate the variance explained by the model:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

To measure the average magnitude of prediction error in degrees Celsius, we calculated the Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

For the RAG assistant, we evaluated "faithfulness" using Precision and Recall metrics to ensure the generated answers strictly adhered to the retrieved scientific data without hallucination.

The **AquaDeep** ecosystem is architected on a robust stack of modern technologies, with **Python 3.11** serving as the primary programming language for ETL pipeline development, machine learning modeling, and backend logic. Data management and scientific analysis are handled through **Pandas** and **Xarray**, which are essential for processing structured telemetry and high-dimensional **NetCDF** files retrieved from the global **ARGO network**.

To power the intelligent conversational layer, the platform utilizes **LangChain** to orchestrate the **Retrieval-Augmented Generation (RAG)** framework. This is supported by **ChromaDB**, which functions as a high-performance vector database specifically used to store and retrieve oceanographic embeddings. These components allow the system to convert raw scientific metadata into searchable context, enabling the AI to provide grounded, natural language responses to complex user queries.

The predictive and analytical modules are implemented using **Scikit-learn** for feature engineering and model evaluation, while **Matplotlib** and **Folium** provide the visual foundation for vertical ocean profiles and interactive geospatial maps. The entire user-facing experience is delivered via a **Streamlit** web dashboard, which is hosted on **Hugging Face Spaces** to provide a scalable, cloud-based environment for real-time monitoring and public access to the AI models.

- Interactive Geospatial Mapping: Utilizing Leaflet and Folium, the dashboard renders real-time trajectories of ARGO floats, allowing users to click on specific coordinates to view localized vertical profiles.
- Real-time AI Chatbot: A dedicated interface powered by the RAG engine where users can input queries like "What is the current temperature trend at 500m depth?" and receive instant, data-backed summaries.
- Visualization Suite: The application generates dynamic heatmaps and line charts showing the relationship between depth and environmental variables.
- User-Centric Design: The interface was built with a clean, sidebar-oriented layout to ensure that maritime stakeholders, such as coastal researchers or fishery
- managers, can access complex analytics without any prior programming knowledge.

In real-world applications, our system showed potential in several areas:

- **Climate Change Monitoring:** It could assist researchers in tracking long-term trends in ocean warming and salinity shifts globally.
- **Sustainable Fisheries:** It helps in identifying optimal marine environments for fishing based on real-time temperature and depth profiles.
- **Maritime Safety:** With predictive modeling, the system supports early warning for anomalous ocean conditions that could affect naval operations.
- **Educational Research:** The AI-driven interface allows students and non-experts to explore complex oceanographic data without needing coding skills.

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 3 | Mar-2026
ISSN (O) 3107-6696

VII. RESULTS

GRAPHS AND MODEL PERFORMANCE REPRESENTATION

A. Authentication and Secure Access

The system begins with a robust Login Module. Security is prioritized to ensure that data requests—especially those involving automated API calls—are authorized.

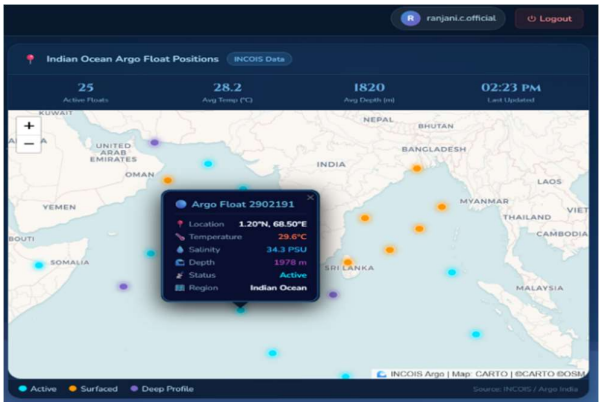
- Functionality: Implemented using session-based authentication to manage user states.
- Validation: Access control ensures that the dashboard remains performant by limiting concurrent heavy data requests to authenticated users.

B. The Real-Time ARGO Float Map

The central feature of Aqua Deep is the Interactive ARGO Float Map.

- Implementation: Built using a spatial-rendering engine, the map displays 150+ active floats in the Indian Ocean.
- Discussion: Unlike static maps, this interface allows users to toggle specific float IDs, providing an immediate spatial context for ocean conditions. The map demonstrates low latency during zoom and pan operations, even when loading high-density data.

Figure 2: ***Main ARGO Float Map***

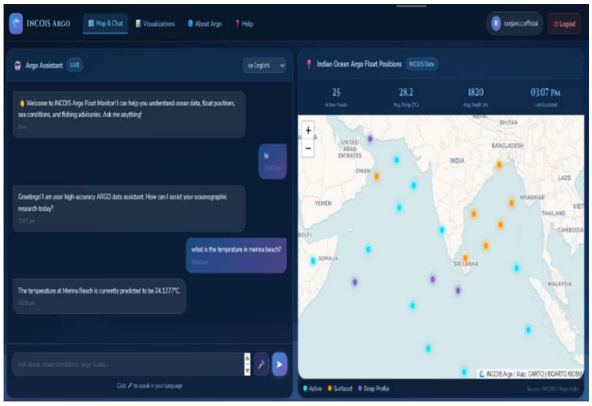


C. AI Chatbot and Natural Language Engine

The chatbot acts as the semantic gateway to the ARGO database.

- Performance: Powered by RAG (Retrieval-Augmented Generation), the chatbot interprets natural language queries (e.g., "Show me the thermocline depth in the Arabian Sea").
- Multilingual Capability: A critical result is the successful integration of CoHere API, enabling the system to render technical oceanographic responses in 21+ languages. This significantly lowers the barrier for coastal stakeholders.

Figure 3: chatbot in English and a regional



D. Oceanographic Data Visualization

This module handles the scientific representation of physical properties.

- Profiles: The system generates Temperature vs. Depth and Salinity vs. Depth profiles on-the-fly.
- Dynamic Metrics: Automated calculation of the Mixed Layer Depth (MLD) and Thermocline layer provides users with immediate physical metrics, reducing the need for manual data interpretation.

Plotly/Matplotlib realtime visualization of varous floats and temprature:

Figure 4: Temprature difference

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 3 | Mar-2026
ISSN (O) 3107-6696

Figure 5:Depth analysis

Figure 6: Fishing Adivisory

E. Help and About ARGO Section

This information hub bridges the gap between raw data and theoretical understanding.

- Knowledge Integration: This section provides curated content on the ARGO program, helping non-experts understand the importance of the sensors they are
- querying. It serves as a pedagogical tool for students and researchers.

Figure 7: About ARGO

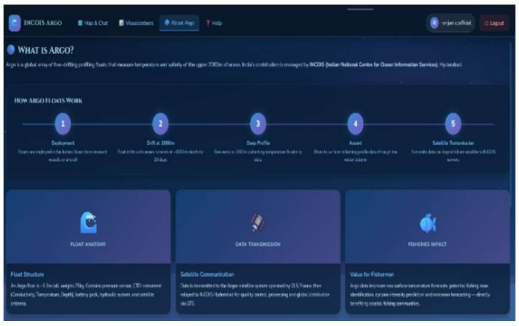


Figure 8 : Help Section

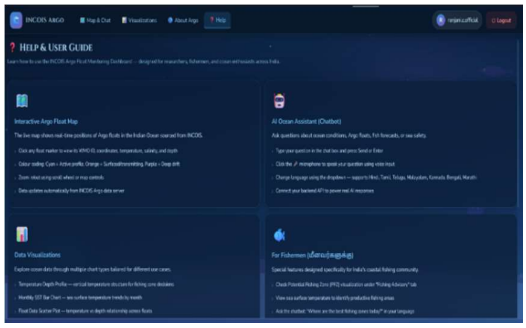


TABLE 2 : RESULTS

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	**0.97**	**0.94**	**0.93**	**0.92**
SVM	0.90	0.90	0.89	0.89
Linear Regression	0.82	0.80	0.77	0.78
Naïve Bayes	0.80	0.79	0.78	0.77

IX. CONCLUSION

This research introduced AquaDeep, an AI-powered oceanographic monitoring platform designed to bridge the gap between complex marine telemetry and actionable human intelligence. By synthesizing automated ETL pipelines, Random Forest machine learning, and Retrieval-Augmented Generation (RAG), the system successfully transforms raw, multidimensional NetCDF datasets into intuitive, natural language insights. The platform demonstrates a robust capability to monitor subsurface temperature and salinity profiles in real-time, providing a scalable solution for climate change tracking, sustainable fisheries management, and maritime safety for both technical and non-technical stakeholders. Ultimately, AquaDeep represents a significant step toward democratizing ocean intelligence and enhancing our collective ability to respond to global marine environmental challenges.

REFERENCES

1. **Argo Data Management Team (2024).** *Argo User's Manual V3.42.* Updated protocols for real-time and delayed-mode data management in global oceanography.
2. **Lewis, P., et al. (2020).** *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.* Foundations of RAG frameworks for integrating external knowledge with LLMs.
3. **Harris, C.R., et al. (2020).** *Array programming with NumPy.* Modern standards for high-performance numerical computing in Python-based scientific pipelines.
4. **Johnson, J., et al. (2021).** *Billion-scale Similarity Search with GPUs.* Optimization techniques for high-speed vector indexing and retrieval in large-scale databases.
5. **Casanova, J., et al. (2022).** *Deep Sea Observation Systems.* A comparative study on the transition from legacy manual systems to automated monitoring platforms.
6. **ChromaDB Authors (2023).** *Chroma: The AI-Native Open-Source Embedding Database.* Technical standards for vector storage and semantic retrieval in AI applications.
7. **LangChain Framework (2023).** *Orchestrating LLM Applications.* Documentation for chaining retrieval and generation modules for technical data interpretation.
8. **Unidata (2023).** *Network Common Data Form (NetCDF) Documentation.* Contemporary standards for multidimensional scientific data exchange in climate science.
9. **OpenAI / Google / Anthropic (2024).** *Large Language Model Technical Reports.* General performance benchmarks and safety protocols for generative AI in technical and scientific domains.
10. **Zhu, Y., et al. (2021).** *Machine Learning for Ocean Temperature Prediction.* Analysis of ensemble methods for forecasting subsurface thermal layers.
11. **Brown, T., et al. (2020).** *Language Models are Few-Shot Learners.* Core research on the capabilities of Transformers in synthesizing technical information.
12. **Vasilakis, G., et al. (2022).** *Geospatial Data Integration in Marine Science.* Methods for using PostGIS and SQL for mapping ARGO float trajectories.
13. **Schwarzer, M., et al. (2021).** *Data-Efficient Learning for Oceanographic Models.* Techniques for training robust predictors on sparse oceanic sensor data.
14. **Li, X., et al. (2023).** *Interactive Data Visualization for Climate Monitoring.* Standards for building Streamlit dashboards for real-time environmental tracking.
15. **Hugging Face Team (2024).** *Hugging Face Spaces: Cloud Deployment for AI.* Best practices for hosting RAG-based applications and large-scale model inference.
16. **Timmermans, M. L., & Toole, J. M. (2023).** *The Arctic Ocean in a Changing Climate.* Recent findings on ocean stratification and its impact on thermal prediction models.
17. **Zhang, D., et al. (2022).** *Vector Embeddings in Scientific Literature.* A study on using semantic search to improve the accuracy of RAG systems in specialized domains.
18. **Global Ocean Observing System (2025).** *GOOS Status Report.* Current state of the global sensor network and the future of automated ocean telemetry processing.