International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct 2026, PP 01-6

# AI-Driven Unified Data Platform for Oceanographic, Fisheries, and Molecular Biodiversity Insights: A Review

**Rupnar Dhanshri Raju, Prof. Bangar Abhishek Popat,**
**Prof.Bhosale.Sachin Bajirao, Prof. Karad Vandana**
Dept. of Computer Engineering,
Jaihind collage of Engineering, Pune,India

**Abstract.** Marine ecosystems produce heterogeneous data from satellites, buoys, research vessels, fisheries activity, laboratory sequencing, and environmental DNA (eDNA) sampling. These sources can reveal ocean conditions, fish distribution, ecosystem change, and biodiversity, but are often stored in separate systems with incompatible formats and metadata. Artificial intelligence (AI) can support forecasting, anomaly detection, habitat modelling, fisheries analysis, and molecular classification. This review examines deep learning for marine water-quality assessment, wireless sensor networks, hybrid models for aquaculture, and molecular biodiversity workflows. It proposes a conceptual unified platform with secure ingestion, source-aware storage, metadata harmonisation, data fusion, AI analytics, and decision-support dashboards. Major challenges include uneven sampling, data interoperability, sensor drift, uncertain labels, model transferability, computational cost, and governance of sensitive fisheries and biodiversity data. The platform is a review-derived architecture rather than an implemented or experimentally validated system.

**Keywords:** Artificial Intelligence, Oceanographic Data, Fisheries, Molecular Biodiversity, Environmental DNA, Data Fusion, Marine Ecosystems

## I. Introduction

Oceans regulate climate, support food systems, provide livelihoods, and contain substantial biological diversity. Understanding marine ecosystems requires observations across different spatial and temporal scales: satellite products describe broad surface patterns, buoys measure local conditions, vessels collect in-situ observations, fisheries records describe human activity, and molecular methods detect organisms through biological traces. Each data source offers only a partial view. Integrating them can help researchers and managers interpret environmental changes and make more informed decisions.

The challenge is not simply the volume of data. Formats, units, coordinate systems, sampling intervals, taxonomic names, and access permissions differ across institutions. AI models may produce useful predictions, but unreliable input data or weak validation can create misleading outputs. A unified platform should therefore combine data engineering, scientific metadata, AI, visualisation, and governance rather than treating model accuracy as the only goal.

## II. Review Methodology

This focused narrative review is based on the supplied research documents covering deep learning for sustainable marine ecosystems, wireless sensor networks for water-quality monitoring, hybrid deep-learning models for aquaculture water quality,

and environmental DNA biodiversity assessment. The review organises the literature by data source, modelling task, integration requirement, and operational limitation. It is not a systematic review or meta-analysis, and no exhaustive database search or pooled performance estimate is claimed. Results from different studies are not directly comparable because their datasets, targets, and evaluation settings differ.

## III. Data Domains and Technologies

**1. Oceanographic observations**

Oceanographic datasets may include sea-surface temperature, salinity, dissolved oxygen, pH, turbidity, chlorophyll concentration, currents, waves, and meteorological variables. Satellite products provide wide coverage but can be affected by clouds, retrieval assumptions, and spatial resolution. Buoys and fixed stations provide repeated local measurements, while ship-based instruments may offer detailed vertical profiles. Combining these sources requires quality flags, units, timestamps, coordinate reference systems, and an explicit record of the method used to align observations.

**2. Fisheries and vessel data**

Fisheries information can include catch and effort records, species, gear type, landing locations, seasonal restrictions, and vessel-position data where collection and use are authorised. Such records can help examine fishing pressure, catch trends, and potential relationships between environmental conditions and species distribution. Reporting gaps and changes in fishing practices can bias the data. Commercially sensitive information and exact vessel locations may require aggregation, access controls, and clear sharing agreements.

**3. Molecular biodiversity and eDNA**

Environmental DNA refers to genetic material collected from environmental samples such as seawater, sediment, or other substrates. Sequencing and metabarcoding workflows can support the detection of organisms that may be difficult to observe directly. These workflows produce reads, quality scores, taxonomic assignments, and links to reference databases. Results depend on sampling design, contamination control, primer choice, sequencing depth, and reference-library coverage. A platform must preserve the relationship between each sample, its location, laboratory protocol, processing pipeline, and taxonomic interpretation.

**. AI and data fusion**

AI tasks include forecasting time-series variables, classifying habitat conditions, detecting anomalies, estimating species distributions, analysing catch-effort relationships, and assisting sequence classification. Data fusion may combine satellite observations with in-situ measurements or link biodiversity detections with habitat characteristics. Fusion must not hide differences in resolution or confidence. Missing observations should remain marked as missing rather than being silently replaced with zero, and model outputs should carry provenance and uncertainty.

## IV. Literature Review

**1. Deep learning for marine water-quality assessment**

Fernández Gambín and colleagues study deep learning for water-quality assessment and forecasting in sustainable marine ecosystems. This line of work illustrates the potential of learned temporal and spatial patterns to estimate or forecast environmental variables. Deep models can capture nonlinear relationships, but they often require substantial, well-curated training data and careful evaluation. Ocean conditions vary by location, depth, season, and instrument; a model trained on one region may perform poorly elsewhere. A unified platform should therefore preserve model-training provenance and provide local validation before operational use.

**2. Wireless sensor networks for water quality**

The supplied review of wireless sensor networks discusses in-situ monitoring and networked measurement. Its central relevance is architectural: sensors, communication protocols, power management, and network coverage determine whether reliable observations can be collected over time. Battery life, biofouling, harsh weather, maintenance access, and intermittent connectivity can constrain field systems. A data platform should ingest device status and calibration history along with scientific readings so that model users can distinguish environmental changes from instrument faults.

**3. Hybrid deep learning for aquaculture**

Rasheed Abdul Haq and Harigovindan investigate hybrid CNN-LSTM and CNN-GRU models for aquaculture water-quality prediction. Their reported evaluation identifies CNN-LSTM as the strongest among the compared models for the studied datasets. Convolutional layers can learn local patterns and recurrent layers represent temporal dependencies, making hybrid architectures relevant to time-series analysis. Nevertheless, farm-based aquaculture conditions differ from open-ocean monitoring and wild-fisheries assessment. Reuse of a model requires a new target definition, representative data, and independent validation.

**4. Environmental DNA and biodiversity monitoring**

Research on environmental DNA and metabarcoding shows how molecular evidence can complement visual surveys and conventional sampling. eDNA can reveal the presence of taxa from collected material, but detection probability is affected by environmental transport, degradation, sampling volume, and laboratory workflow. A positive detection is not always a direct estimate of organism abundance, while a non-detection does not necessarily prove absence. AI-supported taxonomy should retain confidence values and links to reference sequences, and uncertain assignments should be flagged for expert review.

**5. Implications for an integrated platform**

The reviewed studies cover connected but distinct research streams. Oceanographic modelling depends on calibrated physical measurements; fisheries analysis depends on representative catch and effort data; molecular biodiversity depends on reproducible laboratory and bioinformatics pipelines. A unified platform can make these data discoverable and enable cross-domain questions, but it must preserve the scientific meaning of each source. A common portal should not force every data type into one simplified table.

## V. Comparative Analysis

| Research area | Typical data / methods | Contribution | Main limitation |
|---|---|---|---|
| Marine water quality | Satellite and in-situ time series; deep learning | Assessment and forecasting | Regional transfer and missing observations |
| Wireless sensor networks | Sensors, gateways, low-power links | Frequent local measurements | Power, fouling, connectivity, calibration |
| Aquaculture prediction | Water-quality sequences; CNN-LSTM / CNN-GRU | Short-term forecasting | Farm-specific conditions and target variables |
| Fisheries analytics | Catch, effort, vessel and environment records | Trend and habitat analysis | Reporting bias and sensitive data |
| Molecular biodiversity | eDNA samples, sequence reads, taxonomic references | Biodiversity detection support | Reference gaps and lab-process variability |

These research areas should be compared by task-specific metrics rather than one overall score. Forecasting may use MAE or RMSE, classification may use precision, recall, F1 and calibration, and ecological applications may require spatially independent validation and expert review. Operational measures such as data latency, missingness, coverage, reproducibility, and cost are also important.

## VI. Proposed Unified Platform Architecture

The proposed architecture is a conceptual design derived from the reviewed literature. It has not been implemented or experimentally validated.

**1. Layer 1—Data acquisition**
The platform accepts satellite products, buoy and sensor streams, research-vessel observations, fisheries logbooks, authorised vessel-position data, eDNA sample metadata, sequencing outputs, and external environmental datasets. Each source receives a persistent identifier and retains collection date, location, instrument or protocol, unit, and quality flag. Data owners and access restrictions should be recorded before ingestion.

**2. Layer 2—Secure ingestion**
Support real-time streaming and scheduled batch uploads through APIs, instrument gateways, and managed file transfers. At ingestion, validate schemas, timestamps, coordinate ranges, units, and mandatory metadata. Detect duplicates and record ingestion errors without silently dropping records. Apply authentication, encryption, role-based access, and audit logs. Edge devices can buffer data during network outages and synchronise later.

**3. Layer 3—Storage and metadata catalogue**
Keep immutable raw data separate from quality-controlled and model-ready datasets. A data lake can store original files, while time-series and geospatial databases support queries by time and location. A metadata catalogue should describe dataset owner, licence, resolution, collection method, update frequency, quality flags, and known limitations. Sequence reads and derived taxonomic results should link to sample records and processing versions rather than being flattened into generic sensor tables.

**4. Layer 4—Preprocessing and data fusion**
Preprocessing may standardise units, coordinate systems, time zones, taxonomic identifiers, and file formats. Spatial and temporal alignment must document the aggregation method, resolution, and uncertainty. Examples include comparing satellite estimates with buoy measurements, relating environmental conditions to catch records, and linking biodiversity detections with habitat variables. The platform should flag gaps and conflicts rather than implying that all sources agree. Quality-control rules should be domain-specific and reproducible.

**5. Layer 5—AI and analytics**
Candidate modules include time-series forecasting, habitat-suitability modelling, species-distribution analysis, catch-effort summaries, anomaly detection for unusual environmental events, and classification support for molecular biodiversity. Model selection must follow the decision question and available labels. Baseline methods should be evaluated alongside complex models. Use temporal or spatial holdouts, uncertainty estimates, calibration checks, and external validation where possible. Version datasets and models so that a prediction can be traced to its inputs and software configuration.

**6. Layer 6—Applications and decision support**
Dashboards can provide GIS maps, trend charts, search, APIs, reports, and alerts. Researchers may need raw metadata and uncertainty; fisheries managers may need catch-effort summaries; conservation teams may need biodiversity trends and habitat-risk indicators. Alerts should state what was observed, what was predicted, when the data were updated, the confidence level, and whether the result has been independently verified. Human review should remain part of high-impact decisions.

## Implementation Workflow

- Define a narrow initial use case, user group, and decision to be supported.
- Select data sources with documented permissions, quality, coverage, and update schedules.
- Create common metadata fields while preserving source-specific schemas.
- Build ingestion checks, raw-data storage, quality-controlled views, and access rules.
- Develop a baseline analytics model and evaluate it against a simple reference method.
- Expose maps and reports to domain experts and collect feedback before wider deployment.
- Monitor drift, missingness, security incidents, and operational costs after deployment.

## Challenges and Research Gaps

### Interoperability and data quality

Data sources may use different units, taxonomies, coordinate systems, file formats, and sampling schedules. The same parameter can be measured by instruments with different accuracy and calibration. Satellite products may contain retrieval uncertainty, sensor networks may drift, catch records may be incomplete, and sequence libraries may lack local species. Data harmonisation therefore requires explicit metadata and quality flags, not just format conversion. Common vocabularies and documented APIs can improve reuse, but domain experts must validate the meaning of converted fields.

### Spatial and temporal generalisation

Random train-test splits can overestimate performance when nearby observations or repeated measurements appear in both sets. Marine data often have strong spatial and seasonal dependence. Evaluation should hold out regions, seasons, or campaigns to test whether models transfer to new conditions. Where data are sparse, uncertainty should be made visible and predictions should not be presented with false precision. Model drift monitoring is necessary as climate, instruments, and fishing practices change.

### Explainability and scientific reproducibility

A model score alone does not explain why a forecast changed or which observations support it. Decision-makers need provenance, feature descriptions, confidence estimates, and the ability to inspect source records. Reproducibility requires versioned data transformations, model parameters, software dependencies, and evaluation procedures. Model cards or similar documentation can state intended use, known failure cases, and conditions under which a model should not be used.

### Privacy, governance, and responsible access

Fisheries data may expose commercially sensitive activity, and exact biodiversity locations may be sensitive for endangered species. Access policies should reflect data ownership, legal restrictions, licences, and conservation needs. Role-based access, aggregation, audit logs, retention policies, and data-sharing agreements can reduce risk. AI outputs should support rather than replace the judgement of scientists and public authorities. Decisions affecting fishing restrictions or conservation priorities need transparent evidence and a process for contesting errors.

## Evaluation Strategy and Future Scope

A prototype should begin with one measurable use case, such as a validated water-quality forecast for selected coastal stations or a searchable catalogue linking eDNA samples to collection metadata. Evaluation should report data coverage, ingestion latency, missingness, query performance, model error, calibration, cost, and expert usefulness. The system should be tested on independent periods or locations and compared with existing workflows. Feedback from researchers, fisheries officers, and conservation practitioners can reveal whether the interface supports actual decisions.

Future extensions may include edge AI for remote stations, autonomous underwater vehicles, improved satellite-sensor fusion, explainable AI, federated analysis across institutions, near-real-time biodiversity reporting, and decision tools for sustainable

fisheries. These features should be added only when there is evidence of measurable value and a plan for maintenance. A complex platform is not inherently better if data are sparse or users cannot sustain it.

## VII. Limitations and Conclusion

This narrative review uses a limited set of supplied materials and does not claim a systematic search or new experimental results. The proposed platform remains conceptual. The literature supports the potential of deep learning for environmental forecasting, sensor networks for repeated observations, and molecular methods for biodiversity monitoring. However, data heterogeneity, uncertainty, uneven sampling, and limited external validation remain significant barriers. An effective unified platform must combine secure ingestion, source-aware metadata, reproducible preprocessing, suitable AI models, transparent uncertainty, and responsible governance. Its value should be demonstrated through independent evaluation and improved decision workflows, not assumed from the presence of AI.

## References

1. **Á. Fernández Gambín, "Sustainable Marine Ecosystems: Deep Learning for Water Quality Assessment and Forecasting," IEEE Access, vol. 9, pp. 121343–121366, 2021, doi: 10.1109/ACCESS.2021.3109216.** This research focuses on applying deep learning techniques for marine water quality assessment and forecasting. It explains how Artificial Intelligence (AI) can analyze environmental data and identify patterns in changing marine conditions. The study highlights the importance of predicting water quality parameters for protecting aquatic ecosystems.
**"Wireless Sensor Networks for Water Quality Monitoring: A Comprehensive Review," IEEE Access, vol. 11, pp. 95120–95143, 2023, doi: 10.1109/ACCESS.2023.3308905.** This paper presents a review of Wireless Sensor Networks (WSNs) used for water quality monitoring. It explains how sensors collect environmental parameters and transmit the collected information to a central monitoring system. The study discusses communication technologies, sensor accuracy, energy consumption, and network performance.
2. **V. P. Harigovindan, "Water Quality Prediction for Smart Aquaculture Using Hybrid Deep Learning Models," IEEE Access, vol. 10, pp. 60078–60095, 2022, doi: 10.1109/ACCESS.2022.3180482.** This research focuses on predicting water quality in smart aquaculture systems using hybrid deep learning models. It explores how machine learning and deep learning techniques can analyze environmental measurements and identify complex patterns in water quality data. Such models can help estimate future water conditions and provide useful information for aquaculture management.
3. **M. A. Thomsen and E. Willerslev, "Environmental DNA—An Emerging Tool in Conservation for Monitoring Past and Present Biodiversity," Biological Conservation, vol. 183, pp. 4–18, 2015, doi: 10.1016/j.biocon.2014.11.019.** This paper discusses Environmental DNA (eDNA) as an emerging technique for biodiversity monitoring and conservation. Organisms release DNA into their surroundings through skin cells, waste, mucus, and other biological materials. Researchers can collect environmental samples, such as water, and analyze the DNA to identify species that may be present in an ecosystem.
4. **D. R. Rupnar, "AI-Driven Unified Data Platform for Oceanographic, Fisheries, and Molecular Biodiversity Insights,"** seminar presentation, Department of Computer Engineering, Jaihind College of Engineering, Kuran, supplied course material, 2026.