Big Blue Cloud: Architecture and a Marine Protected Area Pilot for Heterogeneous Ocean Data Integration

B. Jack Pan
Oceanography Group
Ocean Motion Technologies, Inc.
San Diego, USA
0000-0002-2620-7478

Alexander J. Orona
Data Science & AI Group
Ocean Motion Technologies, Inc.
San Diego, USA
contact@oceanmotion.tech

Paul Glick
Ocean Motion Technologies, Inc.
San Diego, USA
0000-0001-9841-6503

Leverett Bezanson
Ocean Motion Technologies, Inc.
San Diego, USA
contact@oceanmotion.tech

Abstract -- Ocean data is distributed across agency archives, project portals, local files, and live observing systems. Mature services make large holdings discoverable or computable, but project teams must still reconcile formats, specialized retrieval systems (such as APIs), spatial and temporal references and resolutions, access rules, provenance, and decision products across multiple systems. This paper presents the logical architecture and a pilot implementation of the Big Blue Cloud, a cloud-based integration workspace for heterogeneous large environmental data. The architecture separates intake and metadata registration, storage and curation, analytics and derived products, visualization, and dissemination, with cross-cutting access control and versioning while integrating AI to expedite some of these processes. The platform is positioned relative to NASA Earthdata Search, ERDDAP, Google Earth Engine, and the Copernicus Data Store rather than attempting to replace them. In a Ross Sea Region Marine Protected Area pilot, the workflow combines multiple data sources, such as CCAMLR zone boundaries, OSCAR surface currents, GFS winds, and NASA PACE ocean-color products, in a shared common operating picture. The paper also presents the necessity of openness and interoperability of a platform such as this: source datasets should retain their original licenses, public data and derived products can be exported in reusable formats, and project-restricted data must remain governed. The contribution is therefore a configurable integration and decision-support layer, together with a commercial stewardship model designed to sustain maintenance without requiring exclusive custody of users' data.

Keywords – big data, cloud data platform, data integration, ocean informatics, data provenance, interoperability

I. Introduction

Ocean observing systems now produce data through satellites, numerical models, research cruises, fixed and mobile in-situ platforms, autonomous vehicles, and regulatory or management programs. The principal difficulty is no longer simply the absence of data. It is the effort required to discover relevant holdings, reconcile heterogeneous formats and conventions, preserve provenance, combine public and restricted project data, and present the result in a form that scientists and decision-makers can interpret together. These integration costs recur in marine protected-area monitoring, field operations, coastal security, fisheries analysis, and other multi-institutional ocean workflows.

The FAIR principles emphasize that digital research objects should be findable, accessible under clearly stated conditions, interoperable, and reusable [1]. The CARE principles add that data governance must also consider collective benefit, authority to control, responsibility, and ethics [2]. Neither framework requires every dataset to be anonymous and unrestricted. Rather, both encourage explicit metadata, access conditions, provenance, and governance. These distinctions matter for ocean programs in which openly licensed satellite products may need to be analyzed beside embargoed research observations, community-governed information, or operationally sensitive data.

The Big Blue Cloud is being developed as a cloud-based workspace for this middle layer between authoritative data sources and project decisions. It is not intended to become a replacement archive for NASA, NOAA, CCAMLR, or other public institutions. Instead, it provides configurable connectors, analytical workflows, collaborative visualization, export and curation (such as documentation of queries and subsets of data used from public sources) around a particular research or operational question. The platform can therefore preserve external archives as systems of record while reducing the repeated manual work required to assemble cross-source products for a project team.

This paper makes several contributions: (1) It reports functional architecture and data flow at a level that can be evaluated independently of a particular cloud vendor. (2) It compares open platforms with established data services and identifies a complementary, rather than a competing role. (3) It describes a multi-source pilot for research and monitoring of the Ross Sea Region Marine Protected Area (RSRMPA). (4) It separates current evidence from untested claims and states the limitations of the present beta implementation. (5) It defines what is an open data platform when the project lead is a private company.

II. Related Data Systems and Platform Positioning

Several mature systems already address essential portions of the ocean and Earth-data lifecycle. Some examples (Table I) include NASA Earthdata Search which provides discovery and access across billions of Earth observations, with keyword, temporal, and spatial search [3]. ERDDAP standardizes access to gridded and tabular datasets through URL-based subsetting, open protocols, multiple output formats, and server-generated maps or graphs [4]. Google Earth Engine combines a petabyte-scale public catalog with cloud computation, JavaScript, and Python interfaces [5]. These systems demonstrate that map layers, APIs, and heterogeneous data access are established capabilities.

The Big Blue Cloud is positioned differently in scale, operating purpose and usability. Its primary object is a project workspace that can assemble data from multiple upstream systems, preserve project-specific transformations, combine public and controlled-access information, and deliver a shared product to a defined user community. It is smaller than global archives and does not claim their breadth, authority, or public-service mandate. Its potential value lies in adaptation to a project, direct support from the platform team, and integration of data acquisition, provenance, analytics, visualization, and dissemination in one governed workflow.

TABLE I. REPRESENTATIVE PLATFORM COMPARISON

System	Primary role and relationship to the Big Blue Cloud
NASA Earthdata Search	Authoritative discovery and access across NASA and partner holdings; an upstream source and reference catalog.
ERDDAP	Standards-oriented publication, subsetting, and API delivery for gridded or tabular data; a possible source or downstream delivery endpoint.
Google Earth Engine	Planetary-scale geospatial catalog and computation with public and user assets; a complementary analysis environment and data source.
Copernicus Data Store (C3S/ECMWF)	An upstream source for environmental observations, reanalyses, forecasts, and projections. [6]
Big Blue Cloud	Project-configurable integration, provenance, controlled collaboration, visualization, and export across public archives, sensor streams, and partner data.

III. Architecture and Methods

A. Design Requirements

The design is driven by six requirements: (1) Source diversity: the workspace must accept gridded rasters, vector features, tabular observations, time series, model outputs, and streamed telemetry. (2) Reproducibility: users must be able to determine the source, version, acquisition time, processing parameters, and lineage of a derived product. (3) Governed collaboration: a project must be able to mix public data with partner-restricted or operational data without changing the rights attached to either and allow project participants to manage revocable access rights. (4) Decision-ready presentation: synchronized maps, animations, charts, and annotations must remain connected to the underlying data objects. (5) Portability and continuity: users must be able to export data and products in reusable formats so that the platform is not the sole custodian of the scientific record. (6) Stakeholder Access: the platform must allow derived products, such as dashboards and datasets, to be shared with a wide range of stakeholders, including non-technical stakeholders.

B. Five-Layer Logical Architecture

Fig. 1 shows the service-agnostic logical architecture. The intake and metadata layer registers public APIs, file uploads, scheduled acquisition jobs, and live streams. The storage and curation layer preserves source objects, creates normalized working representations, and maintains catalog metadata. The analytics layer performs temporal and spatial subsetting, harmonization, quality checks, transformations, and derived-product generation. The visualization layer renders maps, overlays, time animation, and linked charts. The dissemination layer supports governed workspaces, download, reports, and interfaces to external tools. Identity and role-based access, provenance and versioning, and operational monitoring apply across all five layers [7].

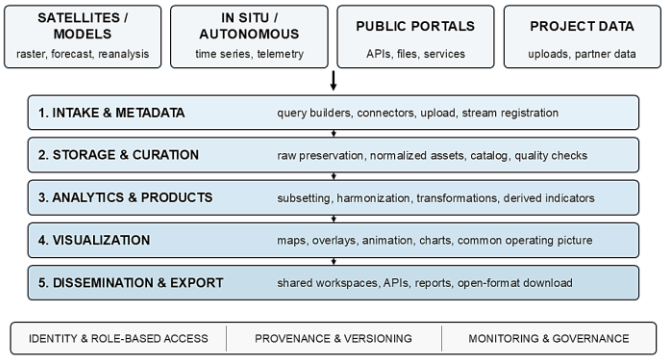


Fig. 1. Logical architecture of the Big Blue Cloud. Authoritative external archives remain systems of record; the platform supplies a project integration and decision-support layer.

C. Ingestion and Normalization

A connector is defined by a source endpoint, authentication method when required, query schema, expected response format, and metadata-mapping rules. A user or scheduled workflow specifies the variable or product, geographic bounds, time interval and resolution or aggregation request. The platform stores the source response or a resolvable source reference before transformation. This separation makes it possible to rerun a transformation against the same input, or to identify when an upstream provider has revised a product.

Normalization does not force every dataset into one universal schema. Raster, vector, and observational data retain their native structural differences. The common contract instead standardizes discoverability and alignment fields: dataset identifier, source and access URL, acquisition timestamp, spatial and temporal footprint, coordinate reference system, variables and units, missing-value conventions, license or use constraint, checksum or version identifier, workspace access class, and lineage links. Format adapters then create analysis-ready representations appropriate to the data type, such as cloud-optimized raster assets, vector features, or time-indexed tables, while retaining the original source object or pointer.

D. Provenance, Versioning, and Derived Products

Each transformation produces a new derived object rather than silently overwriting its parent. The derived record references all input objects and records the operation, parameter set, execution time, software or workflow version, and responsible workspace. This object-level lineage supports three practical tasks: reproducing a figure after an upstream revision, comparing alternative parameterizations, and identifying which products are affected when a source is corrected. Versioning also applies to dashboard configurations so that a map used in a meeting can be reconstructed even if a later workspace contains newer layers.

The analytics layer is designed to support both low-code/no-code query workflows and programmatic methods. A graphical query builder reduces the need for users to write source-specific download scripts, while analysts can use reusable routines for subsetting, resampling, unit conversion, collocation, temporal aggregation, and statistical or machine-learning analysis. The

present paper evaluates the integration pathway and visualization products; it does not claim a new machine-learning algorithm.

E. Access Control, Collaboration, and Dissemination

The governing model separates three access classes. Public objects are derived from public-domain or openly licensed sources and can be disseminated subject to their source terms. Project objects are shared among named collaborators under project agreements. Restricted objects are visible only to authorized roles and may represent licensed commercial data or operational information. Access class is metadata attached to the object and workspace; it does not rewrite the underlying license. A public source remains public even when temporarily analyzed inside a restricted workspace, and restricted project data do not become public merely because they are displayed beside a public layer.

Collaboration is organized around a shared common operating picture rather than independent snapshots or local copies. Users view the same layer definitions, time window, color scale, annotations, and derived products, with versioned changes and role-appropriate visibility. Dissemination can take the form of a dashboard, an exported data subset, a static figure or report, or an API-mediated handoff to another analytical environment. The design objective is therefore interoperability and nonexclusive custody, not a closed catalog in which data can enter but cannot leave.

F. Reference Workflow

An end-to-end workflow follows six steps: (1) The user defines a scientific or operational question, region, time window, and candidate sources. (2) Source-specific connectors retrieve or register the requested objects. (3) The system validates metadata, spatial and temporal coverage, units, and basic quality flags. (4) The workflow creates normalized and derived objects while recording lineage. (5) The user assembles synchronized map and chart views and shares a versioned workspace with collaborators. (6) Selected data and products are exported or disseminated according to their licenses and access classes. This sequence is intentionally compatible with external archives and tools; the platform may perform every step or only the integration and presentation steps.

IV. Ross Sea Region Marine Protected Area Pilot

A. Scientific and Governance Context

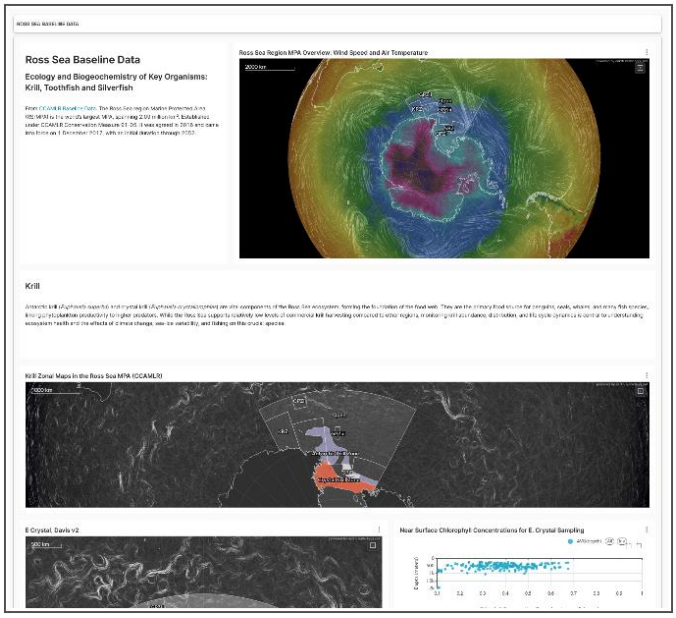
CCAMLR Member States agreed to the RSRMPA in 2016, and the measure entered into force in 2017. The protected area is divided into zones with different management provisions and is intended to protect ecosystem processes, biodiversity, predator-prey relationships, and areas important to Antarctic toothfish while promoting research and monitoring [8]. Effective evaluation requires physical, biological, fisheries, and governance information collected by different national and institutional programs. A coordinated framework for the region has therefore identified data science and cyberinfrastructure as one of the central elements needed for synthesis and science-policy engagement [9].

The Ross Sea Research Coordination Network is an international, interdisciplinary collaboration that connects policy engagement, community partners, data science, modeling, and observations [10]. The pilot uses the Big Blue Cloud as a shared integration workspace for this community. The immediate objective is not to replace contributing archives or to declare one definitive ecosystem assessment. It is to make heterogeneous evidence easier to inspect against a common set of MPA boundaries, time windows, and provenance records in preparation for research coordination and review activities [11], [12].

B. Operational Workflow and Observed Utility

A pilot session begins by selecting the RSRMPA and a date or aggregation interval. The system registers the relevant source versions and retrieves or links the requested data. Spatial data are reprojected for a common map view, but the original coordinate reference system and resolution remain part of the metadata. Vector fields can be animated, raster layers are rendered with source-appropriate color scales, zone polygons and project observations are overlaid, and charts are linked to the active spatial or temporal selection. A saved workspace records the layer stack and display state so that collaborators can revisit the same evidence rather than reconstructing it from individual files (Fig. 2).

Fig. 2. Compact view of the Ross Sea pilot dashboard. The workspace combines management boundaries, physical fields, biological layers, and linked charts; access to the full beta workspace is governed by project role.



The current evidence is implementation-level and qualitative. The platform has demonstrated that multiple source families can be assembled into one versioned workspace and has supported data-synthesis discussions associated with RSRMPA collaborators. The result is useful because it shortens the path from a question - for example, what physical conditions coincide with a biological observation inside a particular management zone - to a common visual product that can be inspected by participants from different disciplines.

C. Relation to Live Ocean Observations

The longer-term architecture also includes direct observations from autonomous or wave-powered buoy systems.

In a separate 2023 Puget Sound deployment, the development team validated direct upload from a marine platform to a partner-hosted cloud via the Big Blue Cloud, providing an early device-to-cloud pathway for the broader Ocean of Apps concept where third-party service providers can participate in the Big Blue Cloud [13]. The Ross Sea pilot reported here currently uses remote and archived data rather than a live Antarctic sensor network.

V. Evaluation, Evidence, and Limitations

The pilot can be evaluated against the design requirements without asserting that development is complete. For source diversity, the public example integrates vector boundaries, gridded raster products, and time-varying vector fields from four independent providers. For reproducibility, the architecture records source identifiers, time ranges, versions, and transformation lineage, and dashboard states are saved as versioned workspace configurations. For governed collaboration, public and project-controlled objects can coexist under separate access classes. For decision-ready presentation, the system links maps, animation, and charts around one spatial and temporal context. For portability, the design supports export and external interfaces rather than requiring exclusive platform custody.

A next evaluation should use predefined tasks and independent participants. Candidate measures include time to locate and collocate specified sources, percentage of derived products reproducible from recorded lineage, round-trip fidelity after export, number and severity of unresolved connector failures, dashboard task-completion rate, and user-rated clarity of access and licensing. The comparison should include both a baseline workflow using source portals and local scripts and an integrated workflow using the platform. Publishing those methods and anonymized results would make future claims of improved time-to-insight testable.

VI. Openness, Portability, and Commercial Stewardship

A. Separating Open Data from Open-Source Software

Earlier descriptions used the phrase "open data platform" broadly. To avoid ambiguity, this paper reserves "open data" for data or content that anyone is free to access, use, modify, and share under an open license, subject at most to conditions that preserve provenance and openness [14]. "Open-source software" is a separate, defined term that requires source-code availability and license rights to redistribute and modify the software [15]. The platform's openness is instead evaluated along four operational dimensions: (1) License fidelity: public-domain or openly licensed data retain source attribution, rights, and use conditions. (2) Portability: users can export eligible data and derived products in reusable, machine-readable formats and move them to other tools. (3) Interoperability: the platform can query or link external archives rather than requiring all data to be permanently copied into a proprietary catalog. (4) Governed participation: project participants determine access to data for which they hold authority, consistent with FAIR accessibility and CARE governance. Under this model, the platform can support open science without requiring that every dataset, workspace, or line of application code is public.

The most durable scientific assets are the datasets, metadata, provenance records, and derived products -- the contents of the bookshelf rather than the shelf itself. The platform is therefore designed so that an organization can import its own data, combine them with public sources, export the resulting objects, and continue analysis elsewhere. External archives remain authoritative, and source identifiers remain attached to derived products. These properties are practical safeguards against a walled garden and are more directly testable than a general claim that a platform is "open."

B. Why Use a Commercial Steward?

Commercial stewardship is not presented as inherently superior to a university, government, nonprofit, or community model. Each can succeed or fail. The commercial model offers a particular alignment of responsibility and incentives: a named operator is accountable for uptime, security, support, bug resolution, documentation, and product improvement; users can make support and service expectations part of a contract; and recurring revenue, contract funding, retained earnings, and indirect-cost recovery can be deliberately allocated to engineering and platform sustainment. In a small company, management can make that allocation directly and users can reach the engineers responsible for a workflow.

Open-source and academic software can be exceptionally reliable and long lived, especially when they have strong institutions and funded maintainers. Source-code availability alone, however, does not guarantee maintenance. A survey of 1,149 research-software users and developers identified recurring needs for better funding models, training, and credit for software work [16]. Interviews with maintainers of established open-source projects similarly identified maintenance labor as a depletable resource and funding as an important enabler [17]. A commercial steward converts maintenance from an optional contribution into paid work, although users should still judge the steward by support performance, transparency, security, and exit rights.

Continuity risk is not unique to private companies. Public programs can be reorganized, university projects can lose grant support or key personnel, and open-source projects can become unmaintained. The appropriate response is not to assume one organizational form is permanent. It is to reduce dependency through source-of-record links, documented lineage, exportable assets, nonexclusive data custody, versioned interfaces, and clear deprecation and continuity policies. Before broad public release, the platform should publish an access and licensing statement, supported export formats, interface documentation, a service-support process, and a continuity plan. Those commitments would make the commercial stewardship claim auditable.

VII. Conclusion

This paper presents the Big Blue Cloud as an interoperable, project-configurable integration workspace rather than a new global archive. Its five-layer architecture connects heterogeneous intake, curation, analytics, visualization, and dissemination while applying provenance, versioning, and access control across the workflow. The RSRMPA pilot provides concrete evidence that management boundaries,

satellite ocean color, modeled winds, surface currents, and governed project layers can be combined into a shared common operating picture.

The present evidence supports the architecture and multi-source implementation, but not a claim of superiority over mature public platforms or a quantitative improvement in scientific outcomes. The public-facing catalog, documentation, benchmarking, and user evaluation remain incomplete. By explicitly distinguishing open data from open-source software and by prioritizing export, source-license fidelity, and nonexclusive custody, the platform can be commercially sustained without becoming a walled garden. The next phase should turn these design intentions into published policies and measurable performance results.

ACKNOWLEDGMENT

The authors thank the Ross Sea Region MPA Research Coordination Network and participating scientific and policy partners for requirements input and pilot feedback. The Ross Sea platform effort is supported in part by NASA Award 24-BIODIV24-0074, and the model management module is supported in part by NSF Award 2423329.

REFERENCES

[1] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, vol. 3, Art. no. 160018, 2016, doi: 10.1038/sdata.2016.18.

[2] S. R. Carroll et al., "The CARE Principles for Indigenous Data Governance," Data Science Journal, vol. 19, Art. no. 43, 2020, doi: 10.5334/dsj-2020-043.

[3] NASA Earth Science Data and Information System, "Earthdata Search." [Online]. Available: https://search.earthdata.nasa.gov/.

[4] NOAA CoastWatch, "ERDDAP griddap documentation." [Online]. Available: https://coastwatch.pfeg.noaa.gov/erddap/griddap/documentation.html.

[5] Google, "Google Earth Engine." [Online]. Available: https://developers.google.com/earth-engine.

[6] Copernicus Climate Change Service (C3S), ECMWF, "Climate Data Store." [Online]. Available: https://cds.climate.copernicus.eu/.

[7] Ocean Motion Technologies, Inc., "Big Blue Cloud." [Online]. Available: https://www.oceanmotion.tech/bigbluecloud.html.

[8] Commission for the Conservation of Antarctic Marine Living Resources, "Marine Protected Areas (MPAs)." [Online]. Available: https://www.ccamlr.org/en/science/marine-protected-areas-mpas.

[9] C. Brooks et al., "Building a coordinated framework for research and monitoring in large-scale international marine protected areas: The Ross Sea region as a model system," Conservation Letters, vol. 17, no. 6, Art. no. e13053, 2024, doi: 10.1111/conl.13053.

[10] Ross Sea Research Coordination Network, "Research Coordination Network." [Online]. Available: https://www.rosssearesearch.org/research-coordination-network.html.

[11] C. M. Brooks et al., "The Ross Sea, Antarctica: A highly protected MPA in international waters," Marine Policy, vol. 134, Art. no. 104795, 2021, doi: 10.1016/j.marpol.2021.104961.

[12] C. M. Brooks and D. G. Ainley, "A Summary of United States Research and Monitoring in Support of the Ross Sea Region Marine Protected Area," Diversity, vol. 14, no. 6, Art. no. 447, 2022, doi: 10.3390/d14060447.

[13] U.S. Department of Energy, "Ocean Motion Technologies: Leveraging DOE Resources to Harness the Power of the Ocean," Oct. 25, 2023. [Online]. Available: https://www.energy.gov/cmei/water/articles/ocean-motion-technologies-leveraging-doe-resources-harness-power-ocean.

[14] Open Knowledge Foundation, "Open Definition 2.1." [Online]. Available: https://opendefinition.org/od/2.1/en/.

[15] Open Source Initiative, "The Open Source Definition." [Online]. Available: https://opensource.org/osd.

[16] J. C. Carver, N. Weber, K. Ram, S. Gesing, and D. S. Katz, "A survey of the state of the practice for research software in the United States," PeerJ Computer Science, vol. 8, Art. no. e963, 2022, doi: 10.7717/peerj-cs.963.

[17] J. Linaker, G. J. P. Link, and K. Lumbard, "Sustaining maintenance labor for healthy open source software projects through human infrastructure: A maintainer perspective."