frontiers in Marine Science

REVIEW
published: 21 May 2019
doi: 10.3389/fmars.2019.00211

From the Oceans to the Cloud: Opportunities and Challenges for Data, Models, Computation and Workflows

Tiffany C. Vance[1], Micah Wengren[1]*, Eugene Burger[2], Debra Hernandez[3], Timothy Kearns[4], Encarni Medina-Lopez[5], Nazila Merati[6], Kevin O'Brien[7], Jon O'Neil[8], James T. Potemra[9], Richard P. Signell[10] and Kyle Wilcox[11]*

[1] US IOOS Program, NOAA/NOS, Silver Spring, MD, United States, [2] NOAA Pacific Marine Environmental Laboratory, Seattle, WA, United States, [3] Southeast Coastal Ocean Observing Regional Association, Charleston, SC, United States, [4] Great Lakes Observing System Regional Association, Ann Arbor, MI, United States, [5] School of Engineering, Institute for Infrastructure and Environment, The University of Edinburgh, Edinburgh, United Kingdom, [6] Earth Resources Technology, Inc., Silver Spring, MD, United States, [7] Joint Institute for the Study of the Atmosphere and Oceans, University of Washington, Seattle, WA, United States, [8] NOAA Big Data Project, Silver Spring, MD, United States, [9] University of Hawaii, Honolulu, HI, United States, [10] United States Geological Survey, Woods Hole, MA, United States, [11] Axiom Data Science, Anchorage, AK, United States

OPEN ACCESS

Edited by:
Justin Manley, Just Innovation Inc., United States

Reviewed by:
Thomas C. Gulbransen, Battelle, United States
Thomas Curtin, University of Washington, United States

***Correspondence:**
Tiffany C. Vance
tiffany.c.vance@noaa.gov
Micah Wengren
micah.wengren@noaa.gov

Specialty section:
This article was submitted to Ocean Observation, a section of the journal Frontiers in Marine Science

Received: *31 October 2018*
Accepted: *03 April 2019*
Published: *21 May 2019*

Citation:
Vance TC, Wengren M, Burger E, Hernandez D, Kearns T, Medina-Lopez E, Merati N, O'Brien K, O'Neil J, Potemra JT, Signell RP and Wilcox K (2019) From the Oceans to the Cloud: Opportunities and Challenges for Data, Models, Computation and Workflows. Front. Mar. Sci. 6:211. doi: 10.3389/fmars.2019.00211

Advances in ocean observations and models mean increasing flows of data. Integrating observations between disciplines over spatial scales from regional to global presents challenges. Running ocean models and managing the results is computationally demanding. The rise of cloud computing presents an opportunity to rethink traditional approaches. This includes developing shared data processing workflows utilizing common, adaptable software to handle data ingest and storage, and an associated framework to manage and execute downstream modeling. Working in the cloud presents challenges: migration of legacy technologies and processes, cloud-to-cloud interoperability, and the translation of legislative and bureaucratic requirements for "on-premises" systems to the cloud. To respond to the scientific and societal needs of a fit-for-purpose ocean observing system, and to maximize the benefits of more integrated observing, research on utilizing cloud infrastructures for sharing data and models is underway. Cloud platforms and the services/APIs they provide offer new ways for scientists to observe and predict the ocean's state. High-performance mass storage of observational data, coupled with on-demand computing to run model simulations in close proximity to the data, tools to manage workflows, and a framework to share and collaborate, enables a more flexible and adaptable observation and prediction computing architecture. Model outputs are stored in the cloud and researchers either download subsets for their interest/area or feed them into their own simulations without leaving the cloud. Expanded storage and computing capabilities make it easier to create, analyze, and distribute products derived from long-term datasets. In this paper, we provide an introduction to cloud computing, describe current uses of the cloud for management and analysis of observational data and model results, and describe workflows for running models and streaming observational data. We discuss topics

that must be considered when moving to the cloud: costs, security, and organizational limitations on cloud use. Future uses of the cloud via computational sandboxes and the practicalities and considerations of using the cloud to archive data are explored. We also consider the ways in which the human elements of ocean observations are changing – the rise of a generation of researchers whose observations are likely to be made remotely rather than hands on – and how their expectations and needs drive research towards the cloud. In conclusion, visions of a future where cloud computing is ubiquitous are discussed.

Keywords: ocean observation, ocean modeling and prediction, cloud, data management, archiving, technology

DEFINING CLOUD COMPUTING AND PATTERNS FOR ITS USE

The most widely used definition of cloud computing is in Mell and Grance (2011):

> Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Essential Characteristics:

- On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed.
- Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms.
- Resource pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model.
- Rapid elasticity: Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand.
- Measured service: Cloud systems automatically control and optimize resource use by leveraging a metering capability.

Service Models:

- Software as a Service (SaaS): The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure.
- Platform as a Service (PaaS): The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider.
- Infrastructure as a Service (IaaS): The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software (Mell and Grance, 2011).

For this paper, we define the cloud as shared, off-premises user configurable resources for data storage/discovery and computing.

Butler and Merati (2016) provide another view of the cloud. Following the framework of *A Pattern Language* (Alexander et al., 1977) and applications of patterns to object-oriented programming in Gamma et al. (1995), they define six patterns of cloud use:

Cloud-Based Scientific Data – Getting Data From the Cloud

Intent: Explores integrating the use of cloud-based data and how scientists can access large volumes of diverse, current and authoritative data, addresses the problem of locating and using large amounts of scientific data.

The Section "Architectures for Real-Time Data Management and Services for Observations" describes streaming data and an architecture for making it easy to gather data. Also see Johanson et al. (2016).

Cloud-Based Management of Scientific Data – Storing Data in the Cloud

Intent: Explores storing and managing data in the cloud. Addresses the problem of ever increasing data quantities with decreasing budgets for data management. Explores the ways scientific projects can meet data access and dissemination requirements such as the U.S. Public Access to Research Results (PARR) mandate (Holdren, 2013).

The section on the NOAA Big Data Project and open data and archiving are examples of this pattern. Also see Meisinger et al. (2009).

Computing Infrastructure for Scientific Research

Intent: Explores the ways in which cloud computing, in the form of PaaS or IaaS could be used as part of a research program and for teaching. It addresses the need for larger computational capabilities, especially under constrained budgets.

The modeling efforts described in later sections are examples of this pattern.

Analysis in the Cloud

Intent: Explores conducting analyses in the cloud. Addresses the problem of wanting to perform analyses on ever larger datasets

and on datasets from multiple sources. Explores the secondary question of ways scientific projects can standardize analysis tools among geographically distributed researchers.

The section entitled "Architectures for Real-Time Data Management and Services for Observations" is an example of this pattern, as are Henderson (2018) and Gorelick et al. (2017).

Visualization

Intent: Explores creating visualizations using cloud-based tools and making the visualizations available via the cloud. Addresses the need to visualize larger amounts of data and the opportunities provided by improved graphics processors and display devices such as VR headsets.

The section entitled "Workflow on the cloud" shows examples of this pattern as does Allam et al. (2018). Workflow tools can visualize more than just the data, they can reveal unexpected dependencies, bottlenecks, and participants' roles.

Results Dissemination in Real Time/Storytelling/Outreach

Intent: Explores ways in which cloud-based platforms and tools can be used to reach new audiences. Addresses the need to make research results rapidly available and relevant to a wide variety of audiences – scientific and non-scientific (Butler and Merati, 2016).

The US Integrated Ocean Observing System (IOOS) Regional Associations' work described below are examples of this pattern.

CURRENT USES – OBSERVATIONS AND MODELS IN THE CLOUD

To expand upon the patterns above, three specific use cases are presented – one focused on using the cloud to disseminate data, a second one describing how the IOOS Regional Associations use a number of patterns for their observational and model data, and a third one based on the European Copernicus Marine Environment Monitoring Service and the capabilities of Google Earth Engine and Google Cloud Datalab. These use cases are intended to provide a pragmatic introduction to using the cloud and specific implementations, to describe what data or outputs and analysis/modeling tools have been moved to the cloud, to show preliminary results and challenges, and to tell where we see these projects going.

Observational Data in the Cloud: The NOAA Big Data Project

The U.S. National Oceanic and Atmospheric Administration's (NOAA) Big Data Project (BDP), announced in 2015, is a collaborative research effort to improve the discoverability, accessibility, and usability of NOAA's data resources. NOAA signed five identical Cooperative Research and Development Agreements (CRADAs) with collaborators: Amazon Web Services (AWS), Google Cloud Platform (GCP), IBM, Microsoft Azure, and the Open Commons Consortium (OCC). The BDP is an experiment to determine to what extent the inherent value in NOAA's weather, ocean, climate, fisheries, ecosystem, and other environmental data can underwrite and offset the costs of commercial cloud storage for access to those data. The project also investigates the extent to which the availability of NOAA's data on collaborators' cloud platforms drives new business opportunities and innovation for U.S. industry.

The BDP facilitates cloud-based access to NOAA data to enhance usability by researchers, academia, private industry, and the public at no net cost to the American taxpayer. One example is the transfer of NOAA's Next Generation Weather Radar (NEXRAD) archive to cloud object stores. The entire NEXRAD 88D archive (∼300 TB, 20 M files) was copied from NOAA's National Centers for Environmental Information (NCEI) to AWS, Google and OCC in October 2015. Marine datasets include elements of the NOAA Operational Forecast System (OFS[1]), sea surface temperature datasets, NCEP/NCAR reanalysis data, and some National Marine Fisheries Service (NMFS) Trawl, Observer, and Essential Fish Habitat data. The full list of available datasets can be found at https://ncics.org/data/noaa-big-data-project/. Under the CRADA, collaborators are allowed to charge for the "marginal cost of distribution." To date, however, none of the collaborators has implemented this provision.

Following the NEXRAD release on AWS:

- In March 2016, users accessed 94 TB from NCEI and AWS combined, more than doubling the previous monthly maximum from NCEI.
- The amount of outgoing NEXRAD Level II data from NCEI has decreased by 50%.
- New analytical uses of the NEXRAD data became manifest – bird migration, mayfly studies.
- 80% of NOAA NEXRAD data orders are now served by AWS.

(Ansari et al., 2017)

Another approach has been the integration of NOAA data into cloud-based analytical tools, including GCP's hosting of NOAA's historical climate data from the Global Historical Climatology Network (GHCN). By offering access through Google BigQuery, from January 2017 to April 2017 1.2 PBs of climate data was accessed via an estimated 800,000 individual accesses. This occurred without Google or NOAA advertising the availability of the data.

Thus far, the NOAA Big Data Project and the CRADA partners have published ∼40 NOAA datasets to the cloud. This has led to increased access levels for NOAA open data, higher levels of service to the data consumer, new analytical uses for open data, and the reduction of loads on NOAA systems. Some lessons learned to date include:

- There is demonstrable unmet demand for NOAA data – as additional services are made available, more total data usage is observed.
- Of equal value to NOAA's data is NOAA's scientific and analytical expertise associated with the data. By working with the CRADA partners to describe and reformat datasets, NOAA's expertise ensures that the "best" version

[1]https://docs.opendata.aws/noaa-ofs-pds/readme.html

of a data type or dataset is made available. If scientific questions arise, NOAA scientists can assist knowing exactly which version of the data is being used.

- Providing copies of NOAA's open data to collaborators' platforms to enable cloud-based access is a technically feasible and practical endeavor and it improves NOAA's security posture by reducing the number of users traversing NOAA networks to access data.
- Beyond the free hosting by cloud providers of several high-value NOAA datasets, another outcome of the NOAA BDP has been the development of an independent data broker entity or service that can facilitate publishing NOAA data on multiple commercial cloud platforms (**Figure 1**). The role of an intermediate "data broker" has emerged as a valuable function that enables the coordinated publishing of NOAA data from federal systems to collaborators' platforms, and could become a common Service supporting all of NOAA publishing data to the cloud.
- Integrating NOAA data into cloud-based tools, as opposed to simply making the original NOAA data files available, has great potential to increase usage. However, expertise and labor is required to properly load NOAA data into those tools.
- A defined commitment and level of service has emerged as a need for both NOAA and the collaborators for the partnership to be sustained.
- Noteworthy is the challenge in generating equal interest on the part of CRADA partners across all of the NOAA data domains. To date, weather related data has been the most requested as part of the NOAA BDP.

The NOAA Big Data Project is scheduled to end in May 2019. Looking toward the future, the BDP seeks, in discussions with current CRADA participants and NOAA managers, to define a sustainable partnership to continue providing cloud-based data access.

Cloud Use Within IOOS for Observational Data and Model Output

The Present

Within the Integrated Ocean Observing System (IOOS) enterprise, many Regional Associations (RA)[2] have migrated ocean observation data management and distribution services to the cloud. Cloud usage varies significantly between IOOS RAs, with some deploying most of their web service infrastructure on the cloud, some deploying infrastructure to shared data centers with or without cloud components, and others utilizing primarily on-premises infrastructure that sometimes includes a cloud backup capability.

IOOS Regional Associations' that have migrated some infrastructure to the cloud have focused on porting existing applications from their own infrastructure, and may not have re-architected to leverage the unique capabilities of cloud services. This represents an incremental approach to cloud adoption, as existing services and data on RA-owned hardware are migrated first, and then, as institutional familiarity with the cloud services grows, new features may be plugged in for better operation.

The most common use of cloud computing within IOOS' 11 RAs is for web applications and data access services. This includes data servers that provide both observation and forecast data to end users [e.g., THREDDS (Thematic Real-time Environmental Distributed Data Services), ERDDAP (Environmental Research Division's Data Access Program), and GeoServer], map-based applications, as well as standard web pages. IOOS RAs have

[2]https://ioos.us/regions

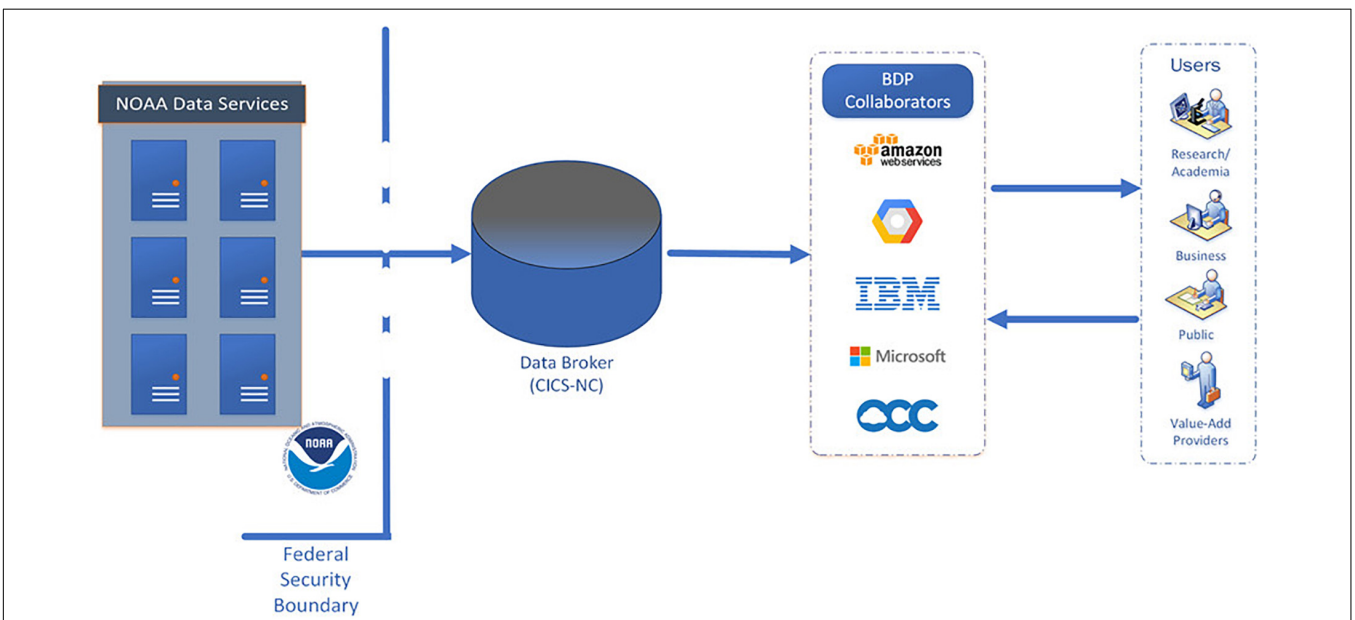

FIGURE 1 | Data broker architecture diagram for Cloud ingest (source: NOAA BDP). CICS-NC: Cooperative Institute for Climate and Satellites – North Carolina.

deployed THREDDS and ERDDAP servers on the cloud using both virtual machine and Docker runtime environments. The IOOS Environmental Data Server, or EDS[3], a web-mapping platform for oceanographic model visualization, is run on the cloud using the Docker platform. GLOS, the Great Lakes IOOS regional association, uses cloud-based virtual machines to run their buoy portal application and the Great Lakes acoustic telemetry system[4,5]. **Figure 2** depicts the number of RAs currently using, or planning to use within 2 years, the cloud for a particular use case.

Several RAs currently use or are actively investigating the cloud as a direct data ingest and storage service for near-real time observations. In this scenario, a server or service is deployed to a cloud-based resource as a direct ingest point for data telemetered from buoys or other sensors operated by the RAs or their affiliates. An example of this is GCOOS, the IOOS region for the Gulf of Mexico, and its affiliate Mote Marine Laboratory's use of a cloud-based instance of Teledyne Webb Research's Dockserver application. Dockserver receives data transmitted by a glider through the Iridium communications network and transfers it to the Internet. Mote's Dockserver has been cloud-based since 2010, receiving data packets in real-time from operational gliders via satellite downlink. Leveraging the cloud has provided a more stable operating environment for Mote's glider operations, and it is far less vulnerable to weather-related hazards than on-premises systems, especially if they are located on or near the coast.

[3]https://eds.ioos.us

[4]https://glbuoys.glos.us

[5]https://glatos.glos.us

GLOS is experimenting with transitioning their locally hosted near real-time data ingest system to a cloud-ready architecture. The primary change involves migrating from a custom sensor data ingest platform to one more suitable to leverage solutions such as AWS' Internet of Things (IoT) services. Currently, GLOS collects transmissions from deployed sensors in eXtensible Markup Language (XML) format via cellular modem to a locally managed secure file transfer protocol (SFTP) service, which then unpacks, stores, and distributes the data. In the new system, nearshore LoRaWAN (Long Range Wide Area Network) devices that connect to Internet-connected gateways may be used to transmit data using HTTP POST or MQTT (message queuing telemetry transport) to remote web services to read, store and re-publish the data. These web services could be more readily deployed on cloud platforms, or, if compatible, use the aforementioned IoT services provided by cloud vendors. GLOS will continue to investigate these pathways over the next 2 years along with its full-scale data center migration to the cloud.

The most significant value that cloud has provided to IOOS RAs to date is its reliability. CARICOOS, the IOOS region for the Caribbean, migrated much of their web presence and associated data services to AWS in 2015. The motivation for the move was mitigation of power grid reliability issues at their University of Puerto Rico's Mayaguez facility. Generator power proved insufficient, and the result was unreliable Internet, data flow, modeling, webpage, and THREDDS server uptimes.

CARICOOS experienced a significant reduction in outages after the migration. During the 2017 hurricane season, they were able to provide near continuous uptime for their most essential data flows, data services, and web pages for use in planning and executing relief efforts. Despite widespread power

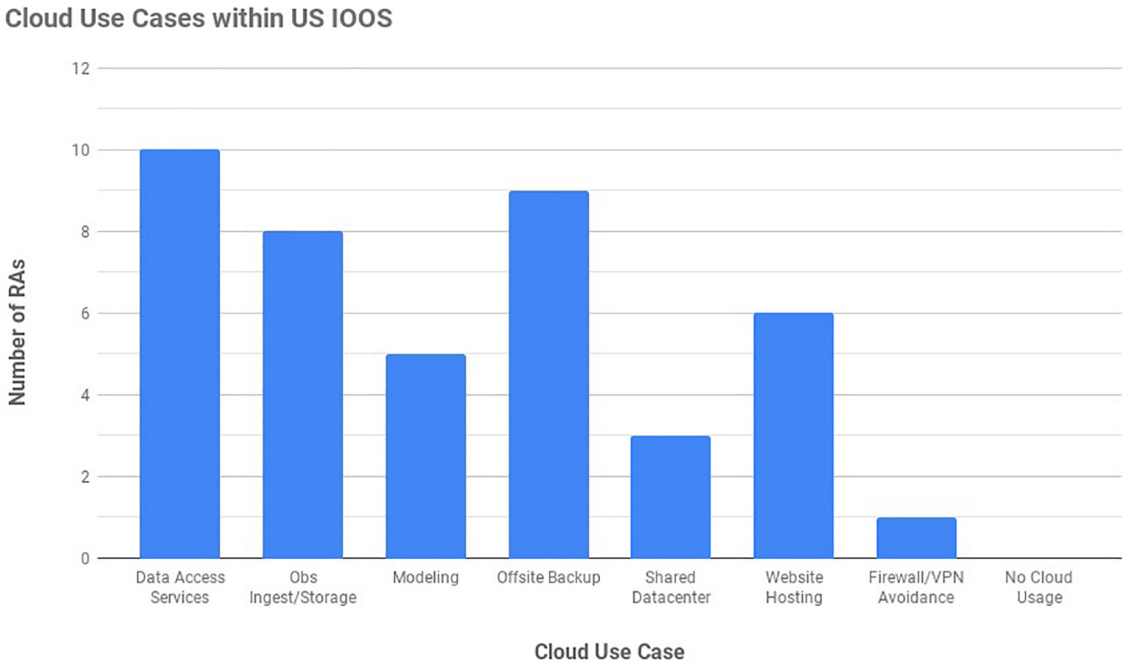

FIGURE 2 | Current and planned Cloud use within the IOOS Regional Associations.

outages and catastrophic damages sustained by Puerto Rico and other Caribbean islands during the hurricanes, CARICOOS' data buoys that had not been damaged in the storms were able to remain online.

Backup and redundancy are also common use cases for cloud computing. Since recovering from Hurricane Maria, CARICOOS has renewed their efforts to develop and test high-performance computing (HPC) ocean models in the cloud. CARICOOS' modelers have been experimenting with a regional high-resolution Finite Volume Community Ocean Model (FVCOM) on AWS, and next in line for migration are their Weather Research Forecast (WRF) implementations, Simulating Waves Nearshore (SWAN) and SWAN beach forecasts and an updated Regional Ocean Modeling System (ROMS). These models currently run on local servers and CARICOOS' goal is to maintain local-cloud redundancy in their operational modeling efforts.

Several of IOOS' RAs have organizational characteristics that affect decisions on whether or not to embrace the cloud. Several RAs share a common IT provider, which pools resources and runs its own self-managed data center similar to a cloud service. This data center is housed in a co-location facility and provides an expandable pool of compute nodes and other resources that allow the RAs to meet customer needs for data services. While no true cloud backup exists yet for this system, it is architected to allow a future cloud migration either in the case of emergency or if it makes economic sense to do so. Many IOOS RAs are affiliated with public universities or other research organizations that provide lower cost internal IT support and services, including data management and web publishing infrastructure. Due to these affiliations, the RAs can take advantage of considerable organizational investment in IT infrastructure and support that would have to be replicated in a cloud environment. In effect, this makes the decision to adopt the cloud an indirect one for these RAs: if their parent organizations or IT provider decide to make the move, they will be included.

The RA for the Pacific Islands, PacIOOS (Pacific Islands Ocean Observing System) is run primarily through the University of Hawaii (UH). The University provides IT infrastructure in the form of server rooms, cooling, network connectivity and firewalls at minimal costs (charged as an indirect cost to the grant). Thus, for an initial investment in hardware, PacIOOS established a variety of IOOS recommended data services, and obtained relatively secure data warehousing for individual observing system components – gliders, High Frequency Radars, model output, etc.

Two of PacIOOS' higher use datasets include real-time observations supplied by offshore wave buoys and forecasts from numerical models. PacIOOS THREDDS servers distribute hundreds of gigabytes of data per month of these data, risking large data egress costs on commercial cloud platforms and making the cloud not yet economically viable. Bandwidth and latency for data publishing is also a concern. PacIOOS forecast models generate about 15 GB/day in output. These models are run on UH hardware, and it is no problem getting the data between modeling clusters and the PacIOOS data servers, whereas bandwidth limitations might affect routine data publishing workflows to the cloud. High volume modeling input/output (I/O) can be handled efficiently on local hardware.

The Future

Challenges, cost barriers, and inertia aside, commercial cloud platforms increasingly offer novel services and capabilities that are difficult or impossible to replicate in an on-premises IT environment. Managed services, aka "software-as-a-service," provide flexibility and scalability in response to changes in user traffic or other metrics that are not easily replicated in self-owned systems and environments. "Serverless" computing, where predefined processes or algorithms are executed in response to specific events, offers a new way to manage data workflows, and are often priced extremely competitively when their unlimited elasticity and zero-cost for periods of non-operation are factored in. Event-based computing using serverless cloud systems is well suited to real-time observation processing workflows, which are inherently event-driven.

For IOOS, or other observing systems, the cloud may become compelling as these features are improved and expanded upon. Instead of data first being telemetered to a data provider's or RA's on-premises servers, it could be ingested by a cloud-based messaging platform, processed by a serverless computing process, and stored in a cloud-based data store for dissemination, all in a robust, fault- and environmental hazard-tolerant environment.

In summary, the motivations and benefits in adopting cloud-hosted services for IOOS RAs have so far been the following:

- Locally available computing infrastructure and/or power grids can be unreliable.
- The operational cost of cloud hosting can be lower. The cost of cloud hosting is highly dependent on a particular application, but IOOS could develop a set of best-practices to end up with lower costs for cloud hosting.
- Hardware lifecycle costs are reduced. The periodic replacement of critical server and network infrastructure is eliminated with cloud-hosted services.
- Cloud scalability can help meet user data request peaks.
- Greater opportunity for standardization exist by providing all RAs with a standard image for commonly used data services.

Undertaking a cloud migration is not without challenges, however. Data integrity on cloud systems must be ensured and characterized accordingly in data provenance metadata (see Section "Data integrity: How to Ensure Data Moved to Cloud Are Correct"). Users must have confidence in the authenticity and accuracy of data served by IOOS RAs on cloud providers' systems, and the metadata provided alongside the data must be sufficiently developed to allow this. The IOOS RA community will need to balance these and other concerns with the potential benefits both in choosing to move to the cloud and in devising approaches by which to do so.

Copernicus and Google: Earth Engine, Cloud and Datalab

Copernicus is the European Union's Earth Observation Program. It offers free and open information based on satellite and

in situ data, covering land, ocean and atmospheric observations, (European Space Agency, 2019a). Copernicus is made of three components: Space, *in situ*, and Services. The first component, "Space," includes the European Space Agency's (ESA) Sentinels, as well as other contributing missions operated by national and international organizations.

The second component of Copernicus, "*in situ*," collects information from different monitoring networks around Europe, such as weather stations, ocean buoys, or maps. This information can be accessed through the Copernicus Marine Environment Monitoring Service (CMEMS) (European Space Agency, 2019b). CMEMS was established in 2015 to provide a catalog of services that improved knowledge in four core areas for the marine sector: Maritime Safety, Coastal and Marine Environment, Marine Resources, and Weather, Seasonal Forecasting, and Climate. The *in situ* data is key to calibrate and validate satellite observations, and is particularly relevant for the extraction of advanced information from the oceans.

Sentinel data can be accessed through the dedicated Copernicus Open Access Hub (European Space Agency, 2019c), and can be processed using the Sentinel-2 and Sentinel-3 Toolboxes (Copernicus, 2019a,b), but Google Earth Engine (GEE) and Google Cloud (GC) provide a simplified environment to access and operate data online (Google Cloud, 2019; Google Earth Engine, 2019). The data is accessible through GC Storage and directly available using the GEE dedicated platform. The access and management of GEE is simplified using a Python API which interacts with the GEE servers through the GC Datalab (Google Cloud Datalab, 2019). The Datalab allows advanced data analysis and visualization using a virtual machine within the Google datacenters, allowing high processing speeds by means of open source coding. Moreover, the Datalab is also useful for machine learning modeling, which makes it very interesting when working with different marine *in situ* and satellite data combinations.

The main limitation of these set of tools is the lack of integration between some data sources and the virtual environment. At the moment, satellite data is stored in the cloud, but *in situ* data is just available through the dedicated Copernicus service, making the process of downloading and accessing this information not as straightforward as in the Earth observation case. However, the inclusion of machine learning techniques and a dedicated language for satellite data treatment makes the use of GEE very attractive, especially for academic and R&D applications. The Google computing capabilities make the GEE-GC-Copernicus combination a realistic option for future ocean observation applications.

OPERATIONAL CONSIDERATIONS

Costs

Comparing the cost of working in the cloud with traditional local computing is a challenge. Deciding which costs should be included to make an equitable comparison requires considering factors such as true infrastructure costs – not only the purchase of hardware but the costs of housing it, utility costs, the cost of personnel to run the system, and how long it will be before the system needs to be replaced. Systems hosted at universities may have unusually low costs due to State or other support – should these costs be used or should the true cost be calculated without external support? On the other hand, cloud providers may provide reduced cost resources or grants of compute time and storage to help new users move to the cloud. The terms and conditions of these grants may affect the long term costs of migrating and may also introduce concerns about data ownership.

Differences in use also affect estimating costs. A project that only involves storing data in the cloud is easy to price out, and the costs of the various types of storage can be balanced against the rapidity with which the data are needed. The benefits of compressing data and finding ways to reduce the size and frequency of data egress from cloud storage can be fairly easily determined. Maintenance of data in the cloud should also be included in cost estimations. The cost of running a model in the cloud is much harder to compute due to variables such as number of virtual machines, the number of cores, what compilers or libraries are needed, grid sizes and time steps, what output files need to be downloaded, and whether analyses of the output can be done in the cloud. If the model is to be used for real time forecasting, then the wall clock run time of the model is critical and may require more expensive compute options to ensure that runs are completed in time.

Molthan et al. (2015) described efforts to deploy the Weather Research and Forecast (WRF) model in private NASA and public cloud environments and concluded that using the cloud, especially in developing nations, was possible. Cost ran from $40–$75 for a 48-h simulation over the Gulf of Mexico.

Mendelssohn and Simons (2016) provide a cautionary tale on deploying to the cloud. As a part of the GeoCloud Sandbox, they deployed the ERDDAP web-based data service to the cloud and concluded that hosting the service in-house was still cheaper. They also found that, except in cases of one-time or infrequent needs for large scale computation, the limitations described in Section "Limitations/Barriers Imposed by U.S. Government Policies" could easily make use of the cloud untenable.

Siuta et al. (2016) looked at the cost of deploying WRF under updated cloud architectures and options and described how resource optimization could reduce costs to be equivalent to on-premises resources.

Generally, running in the cloud can be equal to, or possibly cheaper than running on-premises, but reaping the full benefits requires tuning and experimentation to get the best performance at the lowest cost.

Security

Security concerns are often cited as an impediment to cloud adoption, especially by government researchers. Adoption requires a shift in thinking on the part of institutional security managers from how to secure their resources, via mechanisms such as firewalls and trusted connections, and a shift on the part of users from hardware that they can see and manage to the more amorphous concept of unseen virtual machines. Unless on-premises infrastructures for data services are completely

isolated from public networks, the logical access requirements for on-premises and cloud-hosted infrastructure are very similar. Physical access to local infrastructure is visible, while physical access to cloud hosting solutions is less easily observed. Commercial cloud providers go to great effort to ensure the security of their data centers and users should ensure that these meet local IT requirements.

While the challenges are real, arguably the cloud can be the safest place to operate. Cloud operating systems are up to date on patches and upgrades, redundancy in disks ensures rapid recovery from hardware failures, tools such as Docker can containerize an entire environment and allow for rapid restarts in case of problems, and the fact that cloud systems need to meet commercial level security/data confidentiality requirements drives additional levels of system resilience. Coppolino et al. (2017) provide a good review of cloud security. While their paper is aimed more at business needs, their observations and conclusions are equally valid for scientific data. NIST also provides guidelines on security and privacy in the cloud (Jansen and Grance, 2011; Joint Task Force, 2017).

Limitations/Barriers Imposed by U.S. Government Policies

The dichotomy in U.S. Federal Government IT positions when policy is compared to strategy is evident with regards to cloud services. The U.S. Government proclaims an affinity for cloud services and has done so for the last 8 years (Kundra, 2011; American Technology Council, 2017). The biggest hurdle to cloud adoption has not been technical implementation, nor a lack of desire; it has been Federal IT policy. Offices using cloud services have had to deal with extensive re-engineering and documentation efforts to retroactively address IT requirements. While the merits of Federal IT policy are not under evaluation, it does not lend itself to rapid adoption for cloud services.

Here are notable policy barriers to Federal cloud adoption:

1. All cloud services used by the Federal Government must be FedRAMP approved. The Federal Risk and Authorization Management Program, FedRAMP, is a program established to ensure IT services are secure. While major cloud platform providers have undertaken the cost to ensure their FedRAMP certification, most smaller providers are not incentivized to spend the resources on FedRAMP approval.
2. All Federal IT traffic has to be routed via a Federally approved Trusted Internet Connection (TIC). This policy requirement is particularly onerous and restrictive to cloud adoption. It requires cloud users to configure or purchase dedicated secure routing between the cloud host provider and the end user. This places a large configuration burden and cost upon users, and might force the use of lower performing virtual private network (VPN) solutions. This requirement also negates the opportunity to leverage IT infrastructure co-location benefits with non-Federal collaborators due to the additional network latency added by the Federally derived network traffic routing. Plans are being developed to address the burdens of the TIC requirement (Federal CIO Council, 2018).
3. Federal cloud deployments are not exempt from any of the IT configuration/security requirements that apply to on-premises deployments. For example, the requirement for various monitoring and patch control clients to be installed on Federal IT systems is a hurdle as these clients are not available for many cloud platforms.
4. Procurement, especially the prescriptive nature of the Federal Acquisition Regulation (FAR) does not lend itself well to cloud adoption. Cloud providers innovate rapidly and, when developing contract requirements, it is impossible to know what future services may be available for a particular business need; thus handicapping some of the innovation potential of cloud solutions. Plans have been released for the US Government to develop cloud service catalogs to increase the efficiency with which the government can procure cloud services[6].
5. Budgeting, specifically in relation to cloud procurement, can be challenging. One of the primary advantages of cloud computing is the flexibility to scale resources based on demand. Budgeting in advance is therefore difficult or impossible: allocate too little and risk violating the Anti-Deficiency Act; allocate too much and risk needing to de-obligate unused funding at the end of the contract. NASA, as part of the Cumulus project on AWS, has developed monitoring functions for data egress charges (Pilone, 2018). In practice, they effectively operate without restriction until the budget limit is reached, and then shut down. This is not an optimal solution if users depend on continuous data availability.

These factors diminish the benefits of nimble deployment and increase the cost and complexity of Federal cloud applications.

Data Integrity

Data integrity addresses the component of data quality related to accuracy and consistency of a measurement. It is extremely important to ensure the quality of data that are used in the assessment of our environment. Broad confidence in the integrity of data is critical to research, and decisions driven by this research. The preservation of data integrity is an important consideration in the complete data lifecycle.

Software considerations for cloud hosted data processing and data management processes are no different from those hosted on-premises. Software should be tested and versioned, and the version of software used in the manipulation of the data should be cataloged in metadata. While most of the software on a cloud-hosted solution are bespoke solutions written for specific data needs, a component in a software architecture can depend on cloud host provided infrastructure. Often these solutions are unique to a particular cloud provider. Examples are the stores provided by popular commercial cloud platforms. Unlike commonly used open-source relational database servers or other storage frameworks, the inner working of these data stores are proprietary, and therefore opaque to the data manager. This raises the concern of the potential for data errors that could be

[6]https://cloud.cio.gov/strategy/

introduced and affect the integrity of hosted data. It is imperative that methods, such as periodic checksum verification, be applied to ensure data integrity are preserved over the lifetime of the cloud-hosted storage.

Due to the off-site nature of cloud hosting, serious consideration should be given to the preservation of data integrity during the data transmission from on-premises facilities to cloud hosting. The financial sector has placed a heavy emphasis on this subject and the environmental data sector can benefit greatly from tapping into methodologies and processes developed by other sectors. One such technology is Blockchain.

Blockchain, or digital general ledger technology, is a category of technologies that record transactions between two parties as digital encrypted records, or blocks. As each block contains a digital reference to the previous record as a cryptographic hash, these records create an immutable chain of transactions, or a blockchain. Blockchain implementations are often distributed, and by design can track transactions on many different computers. Many commercial Blockchain solutions are available, and this technology is widely used, especially in the financial sector. The transactions embedded in a blockchain, combined with the immutability of embedded metadata, makes this technology a favorable framework for data provenance tracking. In combination with digital checksums computed against the data embedded in the Blockchain entries, this technology can also support elements used to ensure data integrity. The decentralized nature of Blockchain makes it ideally implementable on cloud solutions and distributed data management systems. Blockchain is a complex topic, worthy of a discussion by itself. For an introduction to using blockchain in science, see Brock (2018) and Extance (2017).

These important considerations that could affect the short and long term integrity of the data are critical to maintain trust in data, but do not detract from the benefits of cloud hosted data processing and data storage.

EMERGING CLOUD TECHNOLOGIES FOR OBSERVATIONS AND MODELING

Architectures for Real-Time Data Management and Services for Observations

Rapidly growing volumes of application-, user-, or sensor-generated data, have led to new software tools built to process, store, and use these data. Whether the data are primary, as in the case of sensor-generated data streams, or ancillary, such as application-generated log files, software stacks have emerged to allow humans to understand and interpret these data interactively and downstream applications to monitor them continuously for abnormal behavior, change detection, or other signals of interest.

While observation data do not always constitute "big data," sensor data in general fits this classification, especially as the measurement frequency of the sensor increases. Low measurement frequency may be due to limitations in communication standards or speeds (i.e., satellite communications costs and the opacity of the ocean to radio frequencies) or in the data processing pipeline that prevent more frequent measurements, not limitations of the sensors themselves. Real-time data streaming applications have the potential to change this paradigm. Combined with server-based Edge computing and the scalability of cloud platforms as execution environments, there is the potential to measure ocean conditions on scales and at precisions not previously possible.

Cloud platforms also reduce the geographic risk associated with research-grade ocean observation systems. Typically, an institution deploys sensors into the ocean and communicates and/or downloads data from them via a "base station" – a physical computer at said institution. In extreme weather events – situations where ocean observing data are critical to decision-making – the stability of the physical computer can be compromised due to power outages, network connectivity and other weather-related nuisances. Putting the software required to keep observing systems running into a cloud system can mitigate most of the geographic risk and provide a more stable access point during events.

One processing model that adapts well to the cloud is stream processing, a technology concept centered on being able to react to incoming data quickly, as opposed to analyzing the data in batches. It can be simplified into three basic steps:

- Placing data onto a message broker
- Analyzing the data coming through the broker
- Saving the results

Stream processing is a natural fit for managing observational ocean data since the data are essentially a continuous time-series of sensor measurements. Data from ocean sensors, once telemetered to an access point, can be pushed to a data-streaming platform (such as Apache Kafka) for analysis and transformation to a persistent data store. Many streaming platforms are designed to handle large quantities of streaming data and can scale up by adding additional "nodes" to the broker as data volume increases. As data volume increases, the analysis may also need to increase. This can be done by increasing the resources available to the analysis code or by adding additional analysis nodes. Each streaming platform is different and has its advantages and disadvantages that should be taken into account before deciding on a solution. Vendor provided end-to-end systems include GCP Dataflow, AWS Kinesis, and Azure Stream Analytics.

An example cloud-architected system for ocean observation data handling system could use this workflow:

Stream system is spun up on cloud resources and, using the provided client tools, is hooked into receive a continuous stream of ocean observations from multiple stations.

Processing code is written using the provided client application programming interfaces (APIs) to:

1. Quality control the data – detect missing/erroneous data using Quality Assurance of Real Time Oceanographic Data (QARTOD) and other quality control software.

2. Alert managers and users based on pre-defined or dynamic conditions.
3. Calculate running daily, weekly and monthly means for each parameter.
4. Store processing results back onto the processing stream as well as in a vendor-supplied analytical-friendly data format, such as AWS Redshift or BigTable, for additional analysis.
5. Export data streams to Network Common Data Form (netCDF) files for archiving and hosting through access services.

The architectures described above provide a number of tools to better support data stewardship and management when setting up a new system and workflow in the cloud. Some of these needs and opportunities will be described in later sections on data provenance, data quality and archiving. Migrations of existing applications have taught helpful lessons in coherently answering the question "hey wait, who's responsible for these data?" as they move along the pipeline from signals to messages to readings in units to unique records to collated data products to transformed information. Migration will require reexamining data ownership – is it correctly documented, will moving to the cloud intentionally or unintentionally transfer ownership to another entity, and who will maintain the data in the cloud – and how useful the data are for further computations or analyses. The following section addresses some of these questions and challenges.

Modeling Workflows in the Cloud

The traditional workflow for ocean modeling is to run a simulation on an HPC cluster, download the output to a local computer, then analyze and visualize the output locally. As ocean models become higher resolution, however, they are producing increasingly massive amounts of data. For example, a recent one-year simulation of the world ocean at 1 km resolution produced 1PB of output. These data are becoming too large to be downloaded and analyzed locally.

The cloud represents a new way of operating, where large datasets can be stored, then analyzed and visualized all in the cloud in a scalable, data-proximate way. Data doesn't need to leave the cloud, and can be efficiently accessed by anyone, allowing reproducibility of results as well as supporting innovative new applications that efficiently access model data. Moving analysis and visualization to the cloud means that modelers and other researchers need only lightweight hardware and software. The traditional high-end workstation can be replaced by a simple laptop with a web browser and cell-phone-hotspot-level Internet connection.

With these benefits come new challenges, however, some cultural, some technical and some institutional. We will examine the benefits of the Cloud for each component of the simulation workflow and then discuss the challenges.

Simulation and Connectivity Between Nodes

Numerical models solve the equations of motion on large 3D grids over time, producing 4D (time, depth, latitude, longitude) output. To reduce the time required to produce the simulation, the horizontal domain is decomposed into a number of small tiles, with each tile handled by a different CPU in a parallel processing system. Because the information from each tile needs to be passed to neighboring tiles, interprocess communications require high throughput and low latency.

For large grids that require many compute nodes, this traditionally has meant using technologies such as Infiniband. Of the major cloud providers, Microsoft Azure offers Infiniband (200 Gb/s), Amazon offers an Enhanced Network Adaptor (20 Gb/s) and Google offers no enhanced networking capability. Because cloud providers provide nodes with sizes up to 64 cores, however, smaller simulations can be run efficiently without traversing nodes. In many cases, simulations with hundreds of cores perform reasonably well on non-specialized cloud clusters, depending on how the simulation is configured.

Storage

Model results are traditionally stored in binary formats designed for multidimensional data, such as NetCDF and hierarchical data format (HDF). These formats allow users to easily extract just the data they need from the dataset. They also allow providers the ability to chunk and compress the data to optimize usage and storage space required.

While these formats work well on traditional file systems, they have challenges with object storage used by the Cloud (e.g., S3). While NetCDF and HDF files can simply be placed in object storage and then accessed as a filesystem by systems like FUSE, the access speed is very poor, as multiple slow requests for metadata are required for each data chunk access. This has given rise to new ways to represent data that use the NetCDF and HDF data models on the Cloud. The Zarr format, for example, makes access to multidimensional data efficient by splitting each chunk into a separate object in cloud storage, and then representing the metadata by a simple JSON (JavaScript Object Notation) file.

With cloud storage, there are no limitations on dataset size, and the data is automatically replicated in different locations, protecting against data loss. A large benefit of storage data on the Cloud is that the buckets are accessible via HTTP (HyperText Transfer Protocol), so efficient access to the data is possible without the need for web services like THREDDS or OPeNDAP (Open-source Project for a Network Data Access Protocol).

Analysis

Analysis of model data on the Cloud is greatly enhanced by frameworks that allow parallel processing of the data (e.g., Spark, Dask)[7]. This takes advantage of the Cloud's ability to allow arbitrary scale up processing. An analysis that takes 100 min on one processor costs the same as an analysis that takes 1 min on 100 processors. The analysis runs on the Cloud, near the data, and with server/client environments like Jupyter, the only data transferred are images and javascript objects to the user's

[7]http://docs.dask.org/en/latest/spark.html

browser. The Pangeo (2018) project is developing a flexible, open-source, cloud-agnostic framework for working with big data on the Cloud, using containers and container orchestration to scale the system for number of users and number of processors requested by each user.

Visualization

Display of data on large grids or meshes is challenging in the browser, but new technologies like Datashader allow data to be represented directly if the number of is polygons is small, but represented as dynamically created images if the number of polygons is large (**Figure 3**). Signell and Pothina (2019) used the Pangeo framework with these techniques to analyze and visualize coastal ocean model data on the Cloud.

Challenges

There are several challenges with moving to cloud simulation, storage, analysis and visualization of model data. Likely, the largest is the apparent cost. Computation can appear expensive because local computing is often subsidized by institutional overhead in the form of computer rooms, power, cooling, Internet charges and system administration. Storage is often expensive but offers increased reliability and the benefit of sharing your data with the community, essentially getting a data portal for free (Abernathey, 2018). The main challenge therefore might be getting institutions and providers to calculate the true cost/benefit of local vs. cloud computing and storage.

THE FUTURE IN THE CLOUD: OPPORTUNITIES AND CHALLENGES

Open Data Hosting

With continued growth in volume of both ocean observations and numerical ocean model output, the problem of how and where to efficiently store these data becomes paramount. As described earlier, the commercial cloud can accept massive volumes of data and store them efficiently in object storage systems, while also enabling new data analysis approaches (Pangeo, 2018). Setting aside costs, migrating open data to commercial cloud provider platforms offers clear technical advantages, but we must consider the potential pitfalls alongside the benefits.

If we take the assumption that all ocean data generated by IOOS, NOAA, or other publicly funded scientific organizations should be freely available and accessible for public use, as stewards of these data we must consider any downstream implications of where we store these data, including storing them on the cloud. Fair and equitable access to ocean data for users,

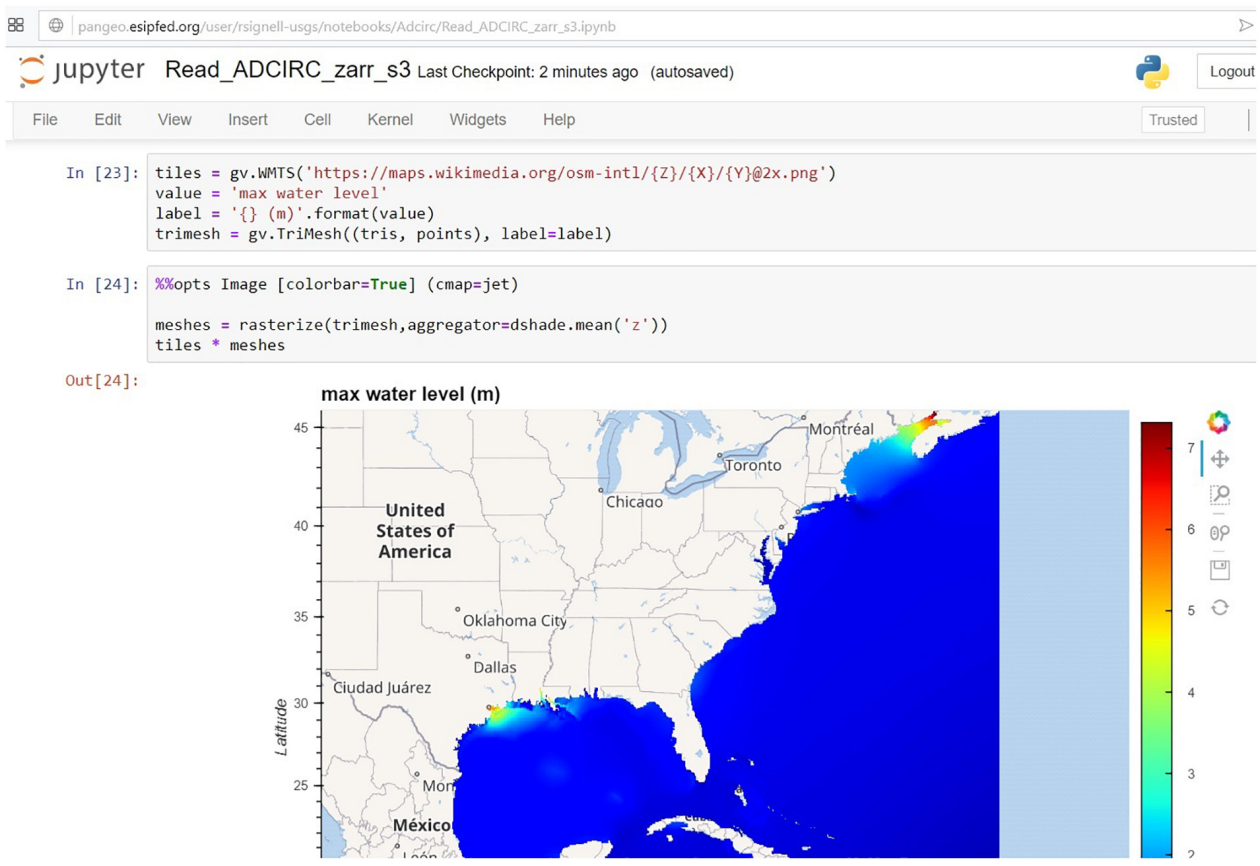

FIGURE 3 | Hurricane Ike simulation on a nine million-node mesh, displayed using Datashader in a Jupyter notebook.

assurance of long-term preservation, archival and continuous access, and flexibility for users to choose the environment in which they use the data, are all factors we must consider.

Already, some earth observing organizations (EOSDIS, 2018) are anticipating that the sheer growth in size of the data they collect will make it prohibitively expensive and complex to host within their own data centers. In this situation, commercial cloud storage services offer an effectively infinite ability to scale to meet their projected data storage needs. Similarly, we can expect ocean data holdings to someday eclipse our abilities to efficiently manage the systems to store them.

As a result, organizations may choose to migrate both the primary public copy of their data, as well as the standards-based services – such as OPeNDAP or Open Geospatial Consortium (OGC) Web Coverage Service (WCS) – users depend on for access, to the cloud. Discontinuing on-premises data hosting entirely can eliminate the need to maintain increasingly complex systems, relying instead on cloud vendors' almost infinite scalability. While this may seem a transparent change for end users, it may not be; there may be indirect implications of this choice on users of our data.

As highlighted earlier, open data that is available on a commercial cloud platform enables users to deploy massively parallel analyses against them. This is becoming known as the "data-proximate" computing paradigm (Ramamurthy, 2018; Pangeo, 2018). This is a breakthrough capability and one very likely to facilitate new discoveries from data that were either not previously possible, or not readily available at costs manageable for most users.

Data-proximate analyses such as these are most efficiently done only when the user provisions computing resources on cloud platform where the data resides. If we move our public, open data to a cloud provider, we may as a result be determining cloud platform suitability for end users looking to run these types of analyses, limiting them to only use the cloud host of our choice. If we move our open data exclusively to a single cloud provider, we lose a degree of impartiality as data brokers compared to when we self-host. We are in effect incentivizing users to come with the data to a particular provider.

Furthermore, the possibility is very real that if one ocean data provider organization selects Amazon, while another selects Google, and yet another selects Microsoft, it will be impossible for a single end user to run data-proximate analyses efficiently without first performing a data migration step to bring each source dataset to a common cloud platform for their compute workflows. If the data are massive, migration likely would not even be an option if the user does not have substantial resources to pay to self-host a copy of it. Fragmentation of data between competing clouds has the potential to negate, or at least lessen, the potential gains of "bringing the compute to the data" as the phrase goes.

These are both issues that deserve recognition and an effort to resolve as the earth observation community undertakes a migration to the cloud. As publishers and stewards of publicly funded open data, it is our responsibility to ensure fair and equitable access to the data for users (Project Open Data, 2018). Budgets, however, are limited, and because of the cost implications for storage and data egress from the cloud, we may not be willing or able to pay to replicate our data on multiple cloud platforms. If we were to try, what criteria would we use to determine which providers to use? Costs and performance would be typical selection factors for contract solicitations, but if, for our users' sake, we must factor in our decision which provider or providers other data publishers have used to host their data on, the picture gets complicated. As soon as data is moved from institution-owned systems, the calculus to determine "fair and equitable" changes.

So what can be done? Standard practice is for individual data provider organizations to sign contracts with commercial cloud providers to host their open data at prearranged costs to the organization. This accomplishes the individual organization's goal to migrate to the cloud, however, it does nothing to address either of the above issues. As an earth observation open data community as a whole, perhaps we can leverage the inherent value of our data to advocate for solutions that meet both our and our user communities' needs.

A technical solution to these issues might be to encourage cooperation among the cloud providers to replicate cloud-hosted public open data on their own. In exchange for signing a pay-for-hosting contract as described above, they could require the provider offer a free data replication service to their competitors, separate from standard data egress channels. Costs for such a service would be borne by the cloud providers, hence a strong push as a community might be necessary to spur its development. Downstream cloud providers would have to self-host open datasets they were not being paid to host in the first place, and each would have to make a cost/benefit decision whether to replicate the data or not, similar presumably to the decision made by participants in NOAA's BDP to host a particular dataset – it would need to have value to their business.

From a technical perspective, a service like this might resemble the following:

- The cloud provider hosting the data would provide a free, authenticated API endpoint that a well-known competitor cloud provider would be able to access to pull new data as it arrives. There would need to be a means to restrict access to this service to legitimate competitor cloud providers in order to prevent standard users downloading the data from circumventing egress charges due to be paid to the primary provider.
- A notification service would allow the downstream cloud provider to subscribe to receive update notifications of new or modified datasets, helping ensure data is kept in sync from one provider to the next.
- Finally, checksums for each data granule or "object" in a cloud data store would be generated and provided via the API to ensure integrity of the duplicated data, or a blockchain-based technology, as described in Section " Data integrity: How to Ensure Data Moved to Cloud Are Correct," might be deployed to accomplish this.
- Costs for the data replication service would either be paid by the original cloud provider hosting the organization's data, or passed back to the original data

publishing organization. In the case of the data publishing organization, this might need to be a part of the organization with archiving/stewardship responsibilities and accompanying funding/budget or a funding agency with similar requirements and responsibilities.

Stepping back from the technical, large organizations such as NOAA that publish many open datasets of high-value to users – including potential future customers for cloud providers – can leverage this value to negotiate free or reduced-cost hosting arrangements with the providers. NOAA's Big Data Project is an attempt to accomplish this. The "data broker" concept conceived by the NOAA BDP is a possible independent, cloud-agnostic solution to the one dataset-one cloud problem, making it easier for competing cloud providers to replicate to their own platforms agency open data published there.

It remains to be seen how much NOAA open data cloud providers are willing to host at no cost to the agency because of the BDP. If NOAA can negotiate free hosting of less commercially valuable data in exchange for technical expertise, including a data broker service, to assist in transfer of its more valuable data, then this lessens the cost of hosting data on multiple clouds and may be the best solution to avoid fragmenting NOAA's data among the clouds. For smaller data publishers, however, this negotiation is less likely to be feasible, and a single-cloud migration is the most likely option for them. In this case, a cloud-to-cloud data replication service might be the best path to ensure their data remains truly "open," and platform-neutral.

A final option for open data providers concerned about equitability is whether to eschew the cloud entirely, or to continue self-hosting data that they also move to the cloud. Either approach carries risks, either of technical obsolescence in the former, or excessive cost and technical complexity to manage two parallel data hosting systems in the latter – likely a deal-breaker for most budget-restricted data publishers.

Time may tell, as more public open data is moved from institution-owned systems to commercial cloud providers, whether cloud-by-cloud data fragmentation or the risk of inadvertently forcing users in co-locating computation on clouds alongside data are significant or not. The open data community as a whole, however, should plan a cloud migration carefully and considerately, and avoid the potential for adverse effects on our users.

Sandboxes in the Cloud for Modeling and Development

Adapting an on-premises high-performance computing environment used to execute ocean model simulations to a commercial cloud environment can be a challenging undertaking. Commercial cloud platforms, however, offer services not available in standard HPC environments that offer significant returns on investment for ocean modeling once the time and effort is taken to leverage them properly.

A computing sandbox is an isolated computing environment where researchers and others may test and develop new applications and workflows. In the computer security realm, they are a place where code and tools can be downloaded and examined without risking malicious damage to operational systems. In research, a sandbox can be a way for a funding source or other entity to provide computing resources to new users of the cloud so they can try-out migrating to the cloud. The sandbox is a communal resource and users are allowed to spin up new virtual machines with limited oversight. Cloud sandboxes support agile development and allow users to try out new ideas. The sandbox is usually a dynamic resource in that virtual machines are expected to be in use for short periods, may be taken down unexpectedly, and should not be used for operational applications. They are a place to "fail small."

One of the first cloud sandboxes for ocean research was the Federal Geospatial Data Consortium (FGDC) cloud sandbox, started in 2011[8]. This sandbox provided IaaS and PaaS and explored the logistics of managing a shared resource. Applications included particle tracking of larval fish, spatial data warehousing, and deployment of an ERDDAP installation.

From 2011 to 2014, the European Commission funded GEOSS interoperability for Weather, Ocean and Water (GEOWOW) project explored expanding the Global Earth Observation System of Systems (GEOSS) in general and the GEOSS Common Infrastructure (GCI). Deployment was on the Terradue Developer Cloud Sandbox and the results provided better access to datasets by centralizing their location and getting them out from behind firewalls and a sandbox for access to and processing of these datasets (Combal and Caumont, 2016)[9].

Molthan et al. (2015) used a NASA private cloud sandbox to explore deploying the WRF weather model and to test various system configurations before re-deploying on a commercial cloud for full scale testing. The UK Met Office Visualization Lab does all their work in the cloud and can be thought of as an all-encompassing sandbox. Because of the operational and security requirements of the main Met Office IT infrastructure, the only way they are able to explore cutting edge technologies and applications is by doing all their work in the cloud (Robinson et al., 2016).

The Earth Science Information Partners (ESIP) has created a *de facto* sandbox by creating an organizational account with Amazon Web Services to enable members to start using the cloud painlessly. Hack for the Sea[10] has set up a sandbox for use during their hackathons but also makes it available to a wide variety of marine professionals and non-professionals for cloud compute and storage at greatly reduced cost. Ocean Networks Canada (ONC) has The Oceans 2.0 Sandbox, which is for internal use. The goal of this sandbox is to bring computing closer to the data by making it easy for ONC scientists to upload and use scripts on the same cloud resources where their data reside.

IOOS is creating a Coastal Ocean Modeling sandbox to enable researchers to explore transitioning their models to the cloud. The aim of the sandbox is to make computing resources available and to foster a community

[8]https://www.fgdc.gov/initiatives/geoplatform/geocloud

[9]https://cordis.europa.eu/result/rcn/171980_en.html

[10]https://github.com/hackforthesea/welcome/

of researchers with expertise in migrating models to the cloud and running them there. The sandbox is intended to serve as a transitional location for models that will eventually be run on NOAA/National Weather Service computing resources or NOAA-wide cloud resources and it will replicate the operational computing environment wherever possible.

As more users explore migrating to the cloud, sandboxes will remain an important tool for easing the transition from research to operations and will provide an important place for experimentation and development. They will also be used as a commons to create and nurture communities and as a place to exchange experiences and techniques. Funding agencies and others can provide these sandboxes in the same way they currently provide communication tools and meeting spaces to support projects.

Cloud Providers as an Archive, or an Archive

Archiving – Not Necessarily the End Point

For many projects, archiving is where data are preserved, typically after a project is completed. Many publications or granting agencies require the researchers to archive their data to make them accessible so that others might reproduce research results, or to provide a safe haven for data that are danger due to lack of funds to continue data stewardship or preservation. For others, archives or data repositories are places where collaborators can contribute similar data sets for sharing and curation – the pattern of "management of scientific data" mentioned above. Formal Archives, less restrictive archives, and repositories improve the chances of data being re-used and shared with a wider audience for reanalysis or use in new ways. Cloud computing provides new ways to make archiving work for research and collaboration and the cloud can host new kinds of archives and repositories.

While many researchers provide data to Archives of record National Archives and Records Administration (NARA) or specialty federated archives, the cloud allows for new groups and collaborators to build archives or repositories using commodity cloud with less oversight and levels of governance. While this seems contrary to the idea of a formal Archive, new tools and certifications developed by data management communities can give data providers considering submitting data a sense that the archive has been vetted, evaluated and made trustworthy for data submission and download. One example of this is the World Data System's Core Trust Seal certification process. This organization has certified repositories for their ability to steward, curate, and provide data submitters and providers with open data access in a reliable manner for a long-term stewardship.

Addressing Storage and Accessibility

Enterprise data management has long used physical offsite data storage as method of data protection. Deploying data into the cloud uses the same concept but can be spun up quickly and can scale up storage and access to fit the needs of the users. Data retrieval can be slower for off-premises backups because data providers/enterprises may have written the data to tapes or other media which are slower to access or a user may have chosen deep storage in the cloud.

Cloud storage companies realize that not all data are created equal. Some data sets may not be accessed after archiving as often as others, based on the content, size and other means of data access. Variations in data storage choices are given names which allude to how the data are accessed – e.g., data storage that is hot or near, would be data accessed more frequently than data storage with names like cold or glacier storage which are for infrequently accessed data. In some cases, historical data sets that need to be archived to fulfill a mandate do not require direct read access and can be put into a deeper level of archive. For other data sets used by multiple collaborators which are still being analyzed or curated (metadata improved, error checked), the nearby or hot storage options makes sense. Costs of archiving data (and subsequent access) vary based on the storage decisions made by the data management mandates and colder storage typically costs less than warmer. Analytics on data access, user behavior modeling and download versus data browsing may help cloud engineers and data providers determine when to shift data sets from deeper colder storage to a warmer storage. This may be event-based (extreme weather-based reanalysis of data or new/older operational instrument comparisons). This also allows data managers and cloud engineers to estimate and budget for shifts in the archival access loads based on data usage. Data virtualization, used by commercial entities, is another method to keep down cloud costs – the data sets that are larger, and not used often, can be virtualized and produced on demand when needed, saving storage costs.

For many, the startup cost for archival storage and access may seem high, but prices continually come down and are dependent on the access requirements. Not having to purchase on-premises hardware should bring the costs down and as data volume increases and the capacity of data center does not increase as quickly, cloud archival storage makes more sense.

NASA EOSDIS – Cloud Archiving

NASA's Earth Observing System Data and Information System (EOSDIS) project has recently moved their data stores to the cloud (EOSDIS, 2018). The project includes earth observational data sets from several distributed active archives for users in the scientific community. The arguments for moving into the cloud are not so much that cloud resources are less expensive than the traditional on-premises data storage and archiving solutions, but that by putting the archival stores in the cloud, these data sets are closer to the cloud-based compute power that many using earth science resources require. There is no longer the need to download data to one's desktop to do complex analysis; the data, computation, analytics, and visualization are all in one place.

Challenges to Using the Cloud for Archiving

Success of projects like EOSDIS for cloud archiving may lead to other distributed archives prototyping projects to test the

efficacy of cloud storage as a way to provide data archiving as a service (DAaaS), but issues around data stewardship, certification, data retention policies and data governance may need to be addressed prior to transitioning from traditional archiving models.

Best practices around data archiving recommend that data formats for archiving be open and well-documented – flat files, simple geospatial files and JSON objects that are easily described and machine and human readable. While this makes sense in a traditional archive, these formats are not natively cloud friendly due to inefficiencies in reading large files in these formats. This may require a shift by data managers and archivists in what they are willing to archive or require an extra step to transform data to make it more cloud amenable.

The long-term storage of data as NOAA acquires commercial cloud infrastructure requires consideration of the long-term viability of the cloud provider. NCEI considers long-time storage of data to cover a period of 75 years. It is very likely that the business model of the cloud provider will change over this period. This may affect the costs and benefits of cloud-based storage. Similarly, the underlying data storage technology will most likely change over this period. Scientific data archive centers that act on behalf of future generations of scientists, should consider deep local storage of a verified "master" copy of data.

PREPARING FOR THE NEXT WAVES OF USERS, TECHNOLOGIES, AND POLICIES

The technical changes we have discussed are paralleled by human changes. In the past, marine research was mainly hands on – researchers went out and collected samples or measurements while on a cruise. As Kintisch (2013) has observed, graduate students are now less likely to ever go to sea – they are working with satellite data, model outputs and aggregated datasets. Research groups have become virtual (Robinson et al., 2016; Wigton, 2016) and an entire research program can be conducted in the cloud – from communication to gathering data to analysis and the distribution of results. It behooves us to consider the human dimensions of cloud adoption as well as the technical strengths and weaknesses of the cloud.

TABLE 1 | Waves of ocean data users and developers.

Wave	Connectivity	Work patterns and needs	Technology best practices that reflect their work patterns and information needs
The Future	Always online, always connected	Understand continuous information, less interested in data that produced the information Less likely to understand or be interested in the *entire workflow* rather they will focus on aspects of it: analysis, visualization relevance and context of mashing up a variety of information sources. Challenge will be ensuring that the information they receive still supports deep science	Expectation of immediacy and continuous connectivity to information Not tolerant of long processing times or difficulty in getting an answer to their question Demand speed of cloud or edge computing
Digital Nomads	Connect from anywhere, applications centric	Understand continuous data collection, require information Comfortable with technology and view it as extension of themselves. Digital Nomads have a different relationship with information than technical or research users?	May be frustrated by organizational inertia when it comes to adopting, leveraging and embracing the Cloud
Technical Experts	Hard core large pipeline connectivity	Understand entire workflow – sensor to results Want to know the details of what they are doing, are inclined to make updates and fixes themselves, and want to know the path of data from observation to use. They are very familiar with the conversion of data to information and intimately understand every nuance of that conversion	Well documented workflows in the cloud can enable them to share techniques and standardize paths. Need high capacity (cloud) computing resources and fast connectivity
Researchers	Lighter connectivity to the Internet	Understand final results of data processing and expect quality data. Want to use data and need to be able to trust their quality but may not want to know all the details. They may be comfortable with a reliance on the automatic conversion of data to information	Need trusted data in the cloud and efficient, expandable and easily shared analysis and visualization tools Need high capacity (cloud) computing resources and fast connectivity
Digitally divided	Limited connectivity – either permanently or situational	Need information but may not have access to HPC etc. May be members of other waves working as first responders during disasters or at sea Need information and require technologies that will enable them to mitigate issues such as low bandwidth or missing or destroyed communication infrastructure	Need cloud-based technologies that support intermittent connectivity and asynchronous communication. Need trusted data in the cloud and efficient, expandable and easily shared analysis and visualization tools

As we look at the factors and variables affecting the move of organizations to the Cloud for storage, resources, processing, and security, it is important to recognize the bi-directional impact of the next waves of users. Individual and institutional attitudes are shifting, prompting a new wave of Cloud users (**Table 1**). Many have never known a professional or personal environment without the concept of being "online." They are mobile-savvy, always connected, hyper-aware of shifting technology trends and readily willing to adopt emerging technology. Organizations should adopt best practices that reflect their work patterns and information needs, guaranteeing a much higher likelihood of their adoption and continuation of the data science integrity that is a core trait of the ocean science community.

A Technological (R)evolution

A technology evolution is equally underway and it will dramatically affect ocean science and the Cloud. Autonomous vehicles, swarm robotics, Edge computing (aka sensor-based computing), *in situ* communications such as cabled observatories, and global availability of low cost, high bandwidth communications will disrupt ocean data collection and distribution. Data will be processed at the source of collection, sensors and vehicles will autonomously make decisions based on that processed data, science-based machine learning[11,12] and the results will be broadcast in near-real time once the sensor is able to contact the Internet. Consumers will be a mix of human and machine end-points.

Future users will enjoy a near continuous Internet experience for non-submerged devices and cabled observatories and the human and non-human consumption of the information that they generate. The emerging Ambient Internet will see the Internet effectively disappear as it becomes connected to nearly everything, and in particular, devices, vehicles and sensors used for marine data collection. This will naturally lead to more data being collected. We are already nearing a tipping point of data volumes exceeding the human capacity to process it all. Automated and cloud-based processing have been slow to materialize in this industry, and what is likely to happen is that it will be eclipsed by sensor-based processing.

Edge computing sits squarely in the realm of the Ambient Internet, leveraging machine learning, artificial intelligence, advanced processing and computing power on the devices themselves and communicating securely and selectively with the Cloud for data transfer. Human reliance on the raw data itself will depreciate over time as the volume of collected data becomes untenable. Confidence levels will increase as machine-learning algorithms consistently produce better results. There will be a nexus when confidence in machine based acquisition, processing and delivery of information exceeds that of the human equivalent.

Tension Between the Current and the Future

As use of the Cloud becomes more widespread, a natural tension between users of local resources and cloud users will be created. Early adopters should feel compelled to prepare best practices to ease adoption for future waves of users. Some of these include embracing the notion of true Digital Nomads. They will not be beholden to any particular platform, operating system, application or physical space. They will have never known an Internet that was not in their pocket, available to them at all times, without constraint. Their expectations of immediacy will be unparalleled and is not centric to the data itself, but rather to the information that it possesses. The emerging Future wave is going to be comfortable with artificial intelligence and augmented reality with a blurred line of information derived from humans or machines. They will live and work in a world of *augmented intelligence*, where artificial intelligence, machine learning and deep learning assist the human experience.

The behavioral characteristics of the next wave(s), coupled with mainstream information technology and the inevitable reduction in cost and increase in proliferation of smart enabled marine based sensors, will cement the fundamental shift in data-information relationships. Sensor based processing with *results* transmitted to the Cloud will force emerging ocean knowledge workers to have an *information* centric mindset rather than a *data* centric one. This mindset will make it easier for others to receive the information they need without needing access to massive datasets. Cloud hosted weather models are a prime example where advances led by Technical Experts and Digital Nomads can benefit Researchers and the Digitally Divided.

Relevance to Cloud and Policy Today

Creation of the structures needed to support the work of all of the waves to do their jobs and advance ocean science is multi-faceted and need not be considered a monumental effort. Evolving and adapting the current mindsets around the Cloud, Edge computing, artificial intelligence, machine learning and augmented reality will dramatically alter the landscape for future workers. Engaging with industry, both traditional and non-traditional, will help spur the innovation. It will also significantly enhance the quantity and quality of data that is collected as the private sector works to produce more inexpensive and more capable devices that work in a connected world.

Data policies also need to shift. Following the old mantra of "collect it, process it, publish it and store it" will not work in an environment of constantly updated information, huge data volumes and increased access through widespread and continuous Internet coverage. Although data security will remain highly important, there will be demands to make data more available so non-human means can interrogate it, learn from it and apply those results to data banks of valuable information. These approaches need to percolate through all levels of organizations in order to create a culture of innovation and preparedness.

Few would disagree that there is an enormous brain-trust resident in organizations all over the world. Intricate knowledge of data formats, sensor types, performance nuances, and metadata standards (or lack thereof) are just some of the elements. There needs to be a concerted effort to increase

[11] https://ieeexplore.ieee.org/document/7959606

[12] https://www.hydrol-earth-syst-sci.net/22/5639/2018/

documentation, standardization and openness in multiple areas in order to propagate and persist this knowledge.

New commercial opportunities may develop that focus on supporting new waves of workers. Encouraging proposals, new grants, funding for innovation and joint partnerships that stimulate research, commercialization and productization of emerging technology are a beginning. Existing companies also have an opportunity to embrace open data and standards, develop automated processing and better support Edge devices as they come online.

The transformation that is occurring does not just involve the Cloud. It is part of a larger technology movement for smarter, smaller and more computing power all around us. The Cloud is only one piece of the transformation and remaining focused on the Cloud at the expense of Edge computing, smart devices, artificial intelligence, automated processing and information centric workflows will not adequately prepare for the next waves of marine scientists. The combination of people, process, and technology – including the Cloud – must be interfaced effectively to develop ocean data and information systems necessary to observe and predict our oceans, lakes and coasts of the future.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This is PMEL contribution 4873.

ACKNOWLEDGMENTS

The authors wish to thank both of our reviewers and Alison Appling and Ellyn Montgomery for thorough reviews and helpful suggestions which strongly benefited the manuscript.

REFERENCES

Abernathey, R. (2018). *Step-by-Step Guide to Building a Big Data Portal*. Available at: https://medium.com/pangeo/step-by-step-guide-to-building-a-big-data-portal-e262af1c2977 (accessed December 1, 2018).

Alexander, C., Ishikawa, S., Silverstein, M., Jacobson, M., Fiksdahl-King, I., and Shlomo, A. (1977). *A Pattern Language: Towns, Buildings, Construction*. Oxford: Oxford University Press.

Allam, S., Galletta, A., Carnevale, L., Bekri, M. A., El Ouahbi, R., and Villari, M. (2018). "A Cloud Computing Workflow for Managing Oceanographic Data," in *Advances in Service-Oriented and Cloud Computing. ESOCC 2017. Communications in Computer and Information Science*, Vol. 824, eds Z. Mann and V. Stolz (Cham: Springer).

American Technology Council (2017). *Report to the President on Federal IT Modernization*. Available at: https://itmodernization.cio.gov/assets/report/Report%20to%20the%20President%20on%20IT%20Modernization%20-%20Final.pdf (accessed December 1, 2018).

Ansari, S., Del Greco, S., Kearns, E., Brown, O., Wilkins, S., Ramamurthy, M., et al. (2017). Unlocking the potential of NEXRAD Data through NOAA's big data partnership. *Bull. Am. Meteor. Soc.* 99, 189–204. doi: 10.1175/BAMS-D-16-0021.1

Brock, J. (2018). *Could Blockchain Unblock Science? Nature Index*. Available at: https://www.natureindex.com/news-blog/could-blockchain-unblock-science (accessed February 7, 2019).

Butler, K., and Merati, N. (2016). "Analysis patterns for cloud centric atmospheric and ocean research," in *Cloud Computing in Ocean and Atmospheric Sciences*, 1st Edn, eds T. C. Vance, et al. (Orlando, FL: Academic Press, Inc.).

Combal, B., and Caumont, H. (2016). "Supporting marine sciences with Cloud services: technical feasibility and challenges," in *Cloud Computing in Ocean and Atmospheric Sciences*, 1st Edn, eds T. C. Vance, et al. (Orlando, FL: Academic Press, Inc.).

Copernicus (2019a). *Sentinel-2 Toolbox*. Available: https://sentinel.esa.int/web/sentinel/toolboxes/sentinel-2 (accessed February 7, 2019).

Copernicus (2019b). *Sentinel-3 toolbox*. Available: https://sentinel.esa.int/web/sentinel/toolboxes/sentinel-3 (accessed February 7, 2019).

Coppolino, L., D'Antonio, S., Mazzeo, G., and Romano, L. (2017). Cloud security: emerging threats and current solutions. *Comput. Electr. Eng.* 59, 126–140. doi: 10.1016/j.compeleceng.2016.03.004

EOSDIS (2018). *EOSIS Cloud Evolution*. Available at: https://earthdata.nasa.gov/about/eosdis-cloud-evolution (accessed December 18, 2018).

European Space Agency [ESA] (2019a). *Copernicus*. Available: www.copernicus.eu (accessed February 7, 2019).

European Space Agency [ESA] (2019b). *Copernicus Marine Environment Monitoring Service*. Available: http://marine.copernicus.eu/ (accessed February 7, 2019).

European Space Agency [ESA] (2019c). *Copernicus Open Access Hub*. Available: https://scihub.copernicus.eu/ (accessed February 7, 2019).

Extance, A. (2017). Could bitcoin technology help science? *Nature* 552, 301–302. doi: 10.1038/d41586-017-08589-4

Federal CIO Council (2018). *From Cloud First to Cloud Smart*. Available at: https://cloud.cio.gov/strategy/ (accessed December 15, 2018).

Gamma, E., Helm, R., Johnson, R., and Vlissides, J. (1995). *Design Patterns: Elements of Reusable Object-oriented Software*. Reading: Addison-Wesley.

Google Cloud (2019). Available: https://cloud.google.com/ (accessed February 7, 2019).

Google Cloud Datalab (2019). Available: https://cloud.google.com/datalab (accessed February 7, 2019).

Google Earth Engine (2019). Available: https://earthengine.google.com/ (accessed February 7, 2019).

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Rebecca, M. (2017). Google Earth Engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* 202, 18–27. doi: 10.1016/j.rse.2017.06.031

Henderson, S. (2018). *Cloud Native Geoprocessing of Earth Observation Satellite Data with Pangeo*. Available at: https://medium.com/pangeo/cloud-native-geoprocessing-of-earth-observation-satellite-data-with-pangeo-997692d91ca2 (accessed December 15, 2018).

Holdren, J. (2013). *Increasing Access to the Results of Federally Funded Scientific Research*. Available at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf (accessed December 21, 2018).

Jansen, W., and Grance, T. (2011). *Guidelines on Security and Privacy in Public Cloud Computing*. Available at: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-144.pdf (accessed December 22, 2018).

Johanson, A. N., Flögel, S., Dullo, W., and Hasselbring, W. (2016). "OceanTEA: Exploring Ocean-derived Climate Data Oceantea: Exploring Ocean-derived Climate Data Using Microservices," in *Proceedings of the 6th International Workshop on Climate Informatics*, Boulder, CO.

Joint Task Force (2017). *Security and Privacy Controls for Information Systems and Organizations*. Available at: https://csrc.nist.gov/CSRC/media//Publications/sp/800-53/rev-5/draft/documents/sp800-53r5-draft.pdf (accessed December 19, 2018).

Kintisch, E. (2013). A sea change for U.S. Oceanography. *Science* 339, 1138–1143.

Kundra, V. (2011). *Federal Cloud Computing Strategy*. Available at: https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/egov_docs/federal-cloud-computing-strategy.pdf (accessed February 7, 2019).

Meisinger, M., Farcas, C., Farcas, E., Alexander, C. H., Arrott, M., de la Beaujardiere, J., et al. (2009). Serving ocean model data on the cloud. *Oceans* 2009, 1–10.

Mell, P., and Grance, T. (2011). *The NIST Definition of Cloud Computing*. Gaithersburg, MD: National Institute of Standards and Technology. doi: 10.6028/NIST.SP.800-145

Mendelssohn, R., and Simons, B. (2016). "A Distributed, RESTful Data Service in the Cloud in a Federal Environment: A Cautionary Tale," in *Cloud Computing in Ocean and Atmospheric Sciences*, 1st Edn, eds T. C. Vance, et al. (Orlando, FL: Academic Press, Inc.).

Molthan, A. L., Case, J. L., Venner, J., Schroeder, R., Checchi, M. R., Zavodsky, B. T., et al. (2015). Clouds in the cloud: weather forecasts and applications within cloud computing environments. *Bull. Am. Meteor. Soc.* 96, 1369–1379. doi: 10.1175/BAMS-D-14-00013.1

Pangeo (2018). *Pangeo: An Open Source Big Data Climate Science Platform*. Available at: https://figshare.com/articles/Pangeo_NSF_Earthcube_Proposal/5361094 (accessed December 21, 2018).

Pilone (2018). *FOSS4G NA 2018: How NASA is Building a Petabyte Scale Geospatial Archive in the Cloud*. Available at: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20180003014.pdf. (accessed December 24, 2018).

Project Open Data (2018). *Principles*. Available at: https://opendatacharter.net/principles/ (accessed December 14, 2018).

Ramamurthy, M. K. (2018). *Data-Proximate Computing, Analytics, and Visualization Using Cloud-Hosted Workflows and Data Services*. Available at: https://ams.confex.com/ams/98Annual/webprogram/Paper337167.html (accessed December 20, 2018).

Robinson, N., Hogben, R., Prudden, R., Powell, T., Tomlinson, J., Middleham, R., et al. (2016). "How we used cloud services to develop a 4D browser visualisation of environmental data at the Met Office Informatics Lab," in *Cloud Computing in Ocean and Atmospheric Sciences*, et al Edn, ed. T. C. Vance (Orlando, FL: Academic Press, Inc.).

Signell, R. P., and Pothina, D. (2019). Analysis and visualization of coastal ocean model data in the cloud. *J. Mar. Sci. Eng.* 7:110. doi: 10.3390/jmse7040110

Siuta, D., West, G., Modzelewski, H., Schigas, R., and Stull, R. (2016). Viability of cloud computing for real-time numerical weather prediction. *Weather Forecast.* 31, 1985–1996. doi: 10.1175/WAF-D-16-0075.1

Wigton, R. S. (2016). "Forces and Patterns in the Scientific Cloud: Recent History and Beyond," in *Cloud Computing in Ocean and Atmospheric Sciences*, et al Edn, ed. T. C. Vance (Orlando, FL: Academic Press, Inc.), 35–41. doi: 10.1016/B978-0-12-803192-6.00006-2

Conflict of Interest Statement: KW is employed by Axiom Data Science. NM is employed by ERT Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.