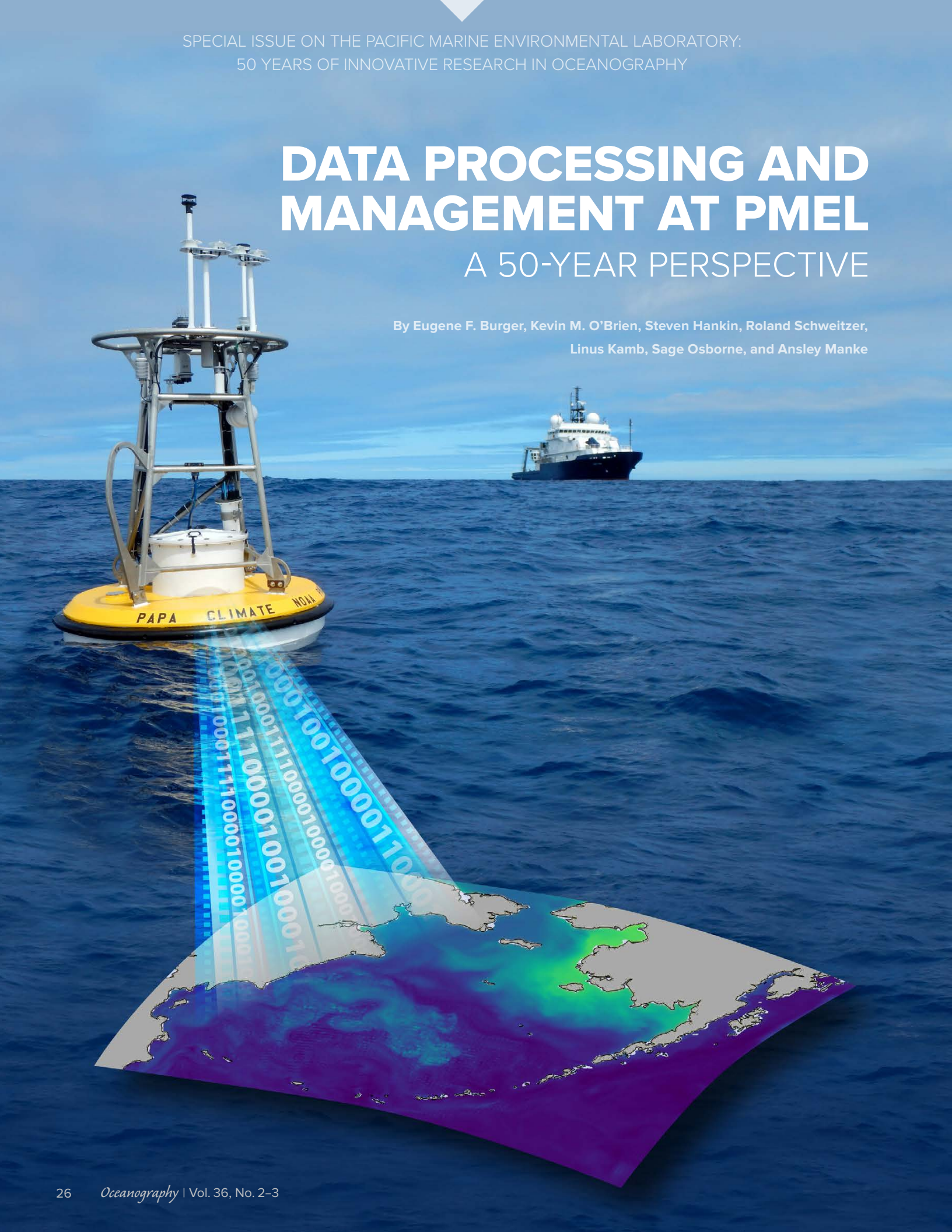
SPECIAL ISSUE ON THE PACIFIC MARINE ENVIRONMENTAL LABORATORY:
50 YEARS OF INNOVATIVE RESEARCH IN OCEANOGRAPHY

DATA PROCESSING AND MANAGEMENT AT PMEL

A 50-YEAR PERSPECTIVE

By Eugene F. Burger, Kevin M. O'Brien, Steven Hankin, Roland Schweitzer, Linus Kamb, Sage Osborne, and Ansley Manke



ABSTRACT. Over the last 50 years, the landscape of marine data management has been transformed. Previously, each research project held its data privately and managed them as local files on disk; today, it is standard practice to share data collaboratively over the internet, often integrated with web tools that provide a global community of scientists with ready access to data analysis and visualization. NOAA Pacific Marine Environmental Laboratory (PMEL) developers and data managers have made and continue to make pivotal contributions toward this evolution. This article examines contributions that include a community-wide standard for metadata storage (e.g., climate and forecast [CF] metadata conventions), a widely used desktop computer tool (PyFerret), a pioneering web server providing visualization and analysis of distributed data (Live Access Server), tailor-made data management systems for uncrewed ocean platforms, and new developments in applications of machine learning to data quality control. We also describe the evolution of in-house PMEL data management, from PMEL developed tools to an open-science, interoperable data approach.

INTRODUCTION

Marine data are a part of the foundation for scientific efforts at the NOAA Pacific Marine Environmental Laboratory (PMEL). These data are collected by an in situ observing network that, at the time of PMEL's founding, was a far cry from the observing assets available today. As PMEL developed, deployed, and managed an ever-growing list of observing assets, the lab assembled a more diverse array of data in greater volumes. From these early beginnings, the data management landscape at PMEL has matured significantly, with PMEL at the forefront of many data management contributions that serve the wider ocean science community. Developers and data managers, along with PMEL science groups and the PMEL Science Data Integration Group (SDIG), continue to make contributions as we work toward a future with more data delivered from the array observing assets. Providing these data on shortened timelines is essential to ensure our understanding of and adaptation to climate change challenges and the needs of the broader scientific community.

THE EARLY YEARS

In the early days of PMEL, data silos were the norm, and paper records were common. Individual project scientists often implemented their own solutions to manage data. As data management practices have changed in the greater science community (Aronova et al., 2010), PMEL's data management has also evolved while at the same time making contributions that impact data management beyond this laboratory. PMEL data management activities remain diversified across the lab, but the PMEL Science Data Integration Group (SDIG), established in 2014, brings together PMEL staff dedicated to the development of data management solutions.

Three decades earlier, in 1984, PMEL hired early career oceanographer D.E. (Ed) Harrison to initiate a new PMEL research group, the Thermal Modeling and Analysis Project (TMAP). The goal of TMAP was to integrate in situ observations with model outputs from PMEL and other sources to better understand sea surface temperature anomaly patterns. Whereas in situ ocean observations are sparse and can capture only a fraction of the dynamics of the surrounding ocean, model outputs offer a spatially complete digital representation of ocean state from which the details of ocean dynamics can be explored. Author Steve Hankin was brought into TMAP to develop software capable of analyzing these massive modeled ocean data sets. What was needed was a tool for fast, interactive analysis and visualization. That tool, named Ferret, still has an enthusiastic community of users almost 40 years later.

THE DEVELOPMENT OF FERRET

The outputs of individual numerical ocean model runs in 1984 were hundreds of megabytes in size. Data sets were shipped to PMEL on large 9-track magnetic computer tapes by US mail. A single model run would require 10 to 30 reels of tape (Figure 1b), which would be stored in PMEL's large "tape vaults"—row upon row of seven-foot-tall (2.1 m tall) tape racks.

It wasn't feasible to analyze these data on PMEL's centralized DEC VAX mainframe computer (Figure 1a). New to the computer technology of the day, however, were high-speed intranets connecting graphic workstations—technologies that would later evolve into the Ethernet (the "wired" connections today) and desktop computers. TMAP introduced this generation of technology into PMEL.

The element of Ferret's design that has made it unique among scientific data analysis applications is the use of "delayed

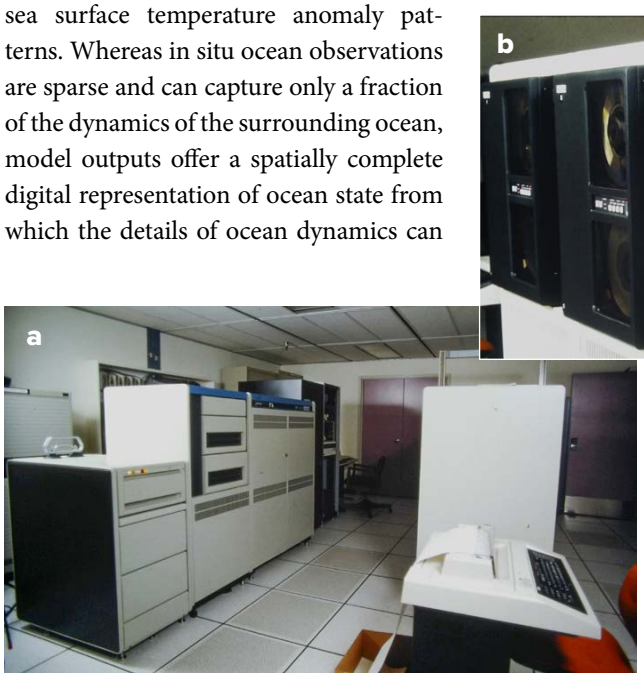


FIGURE 1. (a) The NOAA Pacific Marine Environmental Laboratory central DEC VAX computer room circa 1983. (b) Tape readers located in the central computer room. Note the size of the reels.

analysis," or "lazy evaluations." Much in the style of a mathematician laying out a proof, the Ferret user begins by defining the variables of interest with the word "Let." For example, if a model output contains a temperature variable, "TEMP," a Ferret user might define a climatological ocean temperature field by typing "Let Tclim = {an expression representing the time average of TEMP}" and then define the anomaly relative to this average by typing "Let Tanomaly = TEMP – Tclim". In doing so, the user has provided Ferret with a description of the desired analysis before actually processing any data. Only when that user requests a specific data product—say, a map or a time-series plot of this Tanomaly variable—over a designated region of space and time will number crunching occur. The requested product will typically be delivered quickly, because the volume of data to be processed is a mere fraction of the data set. The delayed evaluation approach lends itself to scientific spontaneity and creativity.

To modernize and extend the software, PMEL began developing PyFerret, whose initial version was released in 2016. PyFerret enables Python programmers to take full advantage of Ferret's analysis engine and capabilities for working with netCDF (https://www.unidata.ucar.edu/software/netcdf/) data sets. It also extends the capabilities of Ferret to make use of Python's up-to-date graphics libraries and extensive Python functions, while keeping all of Ferret's capabilities intact.

Since 2016, PyFerret has been extended to work with new climate and forecast (CF) standards for collections of so-called discrete sampling geometries (DSG), sets of time series, trajectories such as ship tracks, and collections of profiles. Hearkening back to Ferret's beginnings, this development brings to the present our ability to easily compare observational data and reference gridded data sets.

LIVE ACCESS SERVER

In October 1994, the TMAP group presented a new web server, the Live Access Server (LAS), at the Second International Conference on the World Wide Web in Chicago. LAS was so-named because it was among the first web applications ever to provide "live" graphics—visualizations produced on the fly to a user's custom specifications. LAS was effectively a web user interface to the Ferret application.

From the beginning, the goal of the LAS was to enable users to apply the analysis and visualization powers of Ferret to local and distributed data via an intuitive and highly interactive Web interface. LAS development closely followed the development of web application technologies, beginning with its earliest implementation using the Common Gateway Interface (CGI) and on to a bespoke Java Servlet technology, to existing Java web application frameworks like Apache Struts (versions 1 and 2), and finally to Grails, a "convention over configuration framework," all the while, leveraging the Java implementation of the Common Data Model for netCDF and the THREDDS Data Server from Unidata.

LAS capabilities continued to evolve along with the use of newer web application technology. Early versions allowed users to plot slices of data, first in latitude and longitude at a fixed depth and time, and later offered the ability to plot a slice of data in any one or two dimensions. LAS also allowed plots of vectors and plots of data sets defined on curvilinear grids with two-dimensional latitude and longitude coordinates. Users could also specify some analysis operations to be performed on the data before plotting the results. These operations included the ability to calculate many different summary statistics over time and over specific regions. LAS added features to allow plots from up to four different data sets in one display and to calculate the difference (with any necessary regridding) of three of the data slices from one reference slice.

The software development libraries used by the scientific computing community are continuously evolving. The interpretation of data format conventions, the mathematical expression syntax, and the lazy evaluation functionalities for scientific data sets found in Ferret and Java are now found in Python scripting language libraries (e.g., Xarray, cf_xarray, and cf_time). Some Python projects completely change conventional approaches to the development of web-hosted interactive user interfaces (Dash and Anvil are two such frameworks). This shift in off-the-shelf software functionality makes it easier to build applications in the style of the LAS today.

FROM TAPES TO ERDDAP—CHANGING PERSPECTIVES ON DATA MANAGEMENT

Applications and software developed by the TMAP group, such as Ferret and LAS described above, extended PMEL's impact on data management beyond PMEL itself as TMAP members' ideas and practices were adopted in other data management arenas.

A key breakthrough in data dissemination was the ability to access remote data as if they were locally hosted. Now called Open-source Project for a Network Data Access Protocol (OPeNDAP; https://www.opendap.org/), this effort was a precursor to data serving improvements developed in such tools as Thematic Real-time Environmental Distributed Data Services (THREDDS; https://www.unidata.ucar.edu/software/tds/) and Environmental Research Division's Data Access Program (ERDDAP; https://github.com/ERDDAP), both of which implement the Data Access Protocol (DAP; https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/data-access-protocol-2). This was a big leap from a time in which data discovery was accomplished by walking down the hall and talking to colleagues. The adoption of ERDDAP at PMEL has set the stage for many of the data management developments and breakthroughs that followed. TMAP also contributed widely to the development of community conventions (CF metadata) and data formats (network Common Data Form, or netCDF).

Access to distributed data sets greatly heightened the need for data

interoperability—standards designed to ensure that data sets created by one institution are readable by software at another. PMEL was one of the original authors of the 1995 Comprehensive Ocean-Atmosphere Research Data Service (COARDS) standard, which established conventions for the storage of simple gridded data sets in netCDF files. PMEL continued to be a core contributor over the following two decades as the COARDS standard grew into the widely used CF standard that addresses complex model coordinate systems and collections of in situ observations such as time series, cruise tracks, drifters, and profiles.

Similarly, developments by the TMAP group and later SDIG have served the data management needs of NOAA's Oceanic and Atmospheric Research (OAR) program offices. Examples of NOAA OAR program office support are explained in the sections below.

GLOBAL OCEAN MONITORING AND OBSERVATIONS PROGRAM SUPPORT

Over much of the life of SDIG and the former TMAP, NOAA's Global Ocean Monitoring and Observations Program (GOMO) has been a staunch supporter of the group's many data management activities. In particular, GOMO has provided funding for the data activities of the global surface carbon community through the Surface Ocean Carbon Dioxide Atlas (SOCAT), as well as the projects and activities of the Observing System Monitoring Center (OSMC; https://www.osmc.noaa.gov/). One of the innovations developed through the latter project that has had significant international impact is the Open Access to Global Telecommunication System (Open-GTS) project.

OSMC and Open-GTS

Early on, GOMO recognized the need to integrate data streams from the individual global observing networks, which generate tens of thousands of discrete observations every day. Initially, its main goal was to monitor the health and activity of the global observing system. As the Global Ocean Observing System (GOOS) matured, and the GOOS Observations Coordination Group was developed, a new focus arose on the need to significantly improve interoperability of these observations, but in near-real time as well as in delayed mode. This improved interoperability not only served individual networks but also ensured that external stakeholders had access to the data. Thus, the OSMC project evolved toward a goal of combining the discrete "networks" of in situ ocean observing platforms—for example, ships, surface floats, profiling floats, and tide gauges—into a single, integrated system.

The Open-GTS project aimed to improve near-real-time data access by data distribution through the World Meteorological Organization's (WMO's) Global Telecommunication System (GTS). GTS is the exchange mechanism for oceanographic and marine meteorological data used by global weather forecasting centers. However, it is an intentionally closed system that is reliant on complex data formats. The Open-GTS project demonstrated an easier way for data producers to exchange their data through the GTS. Eventually, Open-GTS became a WMO pilot project as part of the WMO effort to evolve the GTS data exchange infrastructure, and it was also selected as a United Nations Ocean Decade endorsed activity.

SOCAT

The Surface Ocean CO_2 Atlas (SOCAT; Bakker et al., 2016) is a quality-controlled, global surface ocean carbon dioxide (CO_2) data set compiled from data gathered on research vessels, by the Ship of Opportunity Program, and from buoys and autonomous platforms. The initial idea to create a high-quality global surface ocean data set of the recalculation of the fugacity of CO_2 (fCO_2; Feely et al., 2023, in this issue) was agreed upon in 2007 at a workshop in Paris. The first version of the SOCAT synthesis product, which took four years to complete due to the manual techniques used, contained around six million observations. The second version followed two years later and added four million new observations. It quickly became clear that a more automated data ingestion and quality control structure would have to be implemented to improve the efficiency of creating the ever-growing SOCAT synthesis product. This increase in efficiency was also necessary in order to support annual global products, such as the Annual Carbon Budget developed by the Global Carbon Project, that depend on SOCAT for its contribution of ocean carbon data.

PMEL and SDIG, with the support of colleagues at the Bjerknes Climate Data Center, developed a data submission dashboard and quality control editor to facilitate more frequent release cycles. The data ingestion system was implemented for SOCAT version 3, and since that version, there have been annual releases of the SOCAT synthesis product—a total of 11 releases to date and over 35 million observations as of SOCAT version 2023. This data ingestion tool transformed the way carbon scientists and technicians built, and continue to build, the SOCAT data set.

OCEAN ACIDIFICATION PROGRAM SUPPORT

The benefit of a streamlined data submission workflow was clearly illustrated with the data submission dashboard developed for the SOCAT community. The NOAA Ocean Acidification Program (OAP) was established in 2009 through the Federal Ocean Acidification Research and Monitoring Act (FOARAM; 33 U.S.C. Chapter 50, Sec. 3701-3708). As this program began funding the collection of data and generation of synthesis products, the need was recognized for a streamlined mechanism for data submission to the Ocean Carbon and Acidification Data System (OCADS).

The SDIG philosophy of maximizing and adopting successful frameworks was applied here with the adaptation of

the SOCAT dashboard to provide an efficient data submission workflow for the OAP. The first phase of this effort was development of a platform for the convenient assembly of metadata. A web-based metadata tool (Figure 2) was then developed to allow users to enter the metadata attributes required by the OCADS template (Jiang et al., 2023).

With the OAP metadata tool in hand, the OAP funded the adaptation of the SOCAT dashboard for use with a greater diversity of data. This web-based tool, the Science Data Information System (SDIS), extended the SOCAT application to acceptance of all data types funded by the OAP, including biological, experimental, and model data. User authentication was added, and the metadata editor developed under phase 1 was integrated into the application. Users can now upload data, assemble metadata, and with a few button clicks, submit their data to the NOAA National Centers for Environmental Information (NCEI) OCADS data archive (https://www.ncei.noaa.gov/products/ocean-carbon-acidification-data-system).

Development of the SDIS application is ongoing. While data quality control is not explicitly required for data submission, the SDIS includes optional quality control features that allow users with biogeochemical data to assess the quality of the uploaded data.

LOOKING AHEAD

The data management activities highlighted in this article have helped to keep PMEL at the forefront of data management in NOAA. Data integrity has always been a key consideration, and development of new processes will ensure that PMEL continues this tradition.

NOAA's adoption of the commercial cloud in support of data management has been slow, largely due to policy and compliance requirements. The commercial cloud and ancillary services offered by commercial cloud vendors have the potential to change data workflows. By using the off-the shelf Internet of Things services offered by the major cloud vendors, data acceptance workflows can move from custom-designed, NOAA-developed software to platforms that offer high reliability and significantly scalable platforms to meet different demands. Work is underway at PMEL to experiment with these services and understand the benefits offered.

Another new frontier is the use of artificial intelligence and machine learning to address data management needs such as quality control/assurance. Recent developments have enabled the creation of explainable machine learning models that are suitable for applications in scientific research. These models' responses to input correspond to explicit mathematical formulae, allowing the model predictions to be understood in the same theoretical context as the data collected by in situ observing platforms. Based on this departure from "black box" models, PMEL's Science Data Integration Group has been developing tools to address the growing complexity and volume of in situ observation data. Once successful, these will liberate resources typically spent on searching for errors and reduce the overall time to disseminate the data.

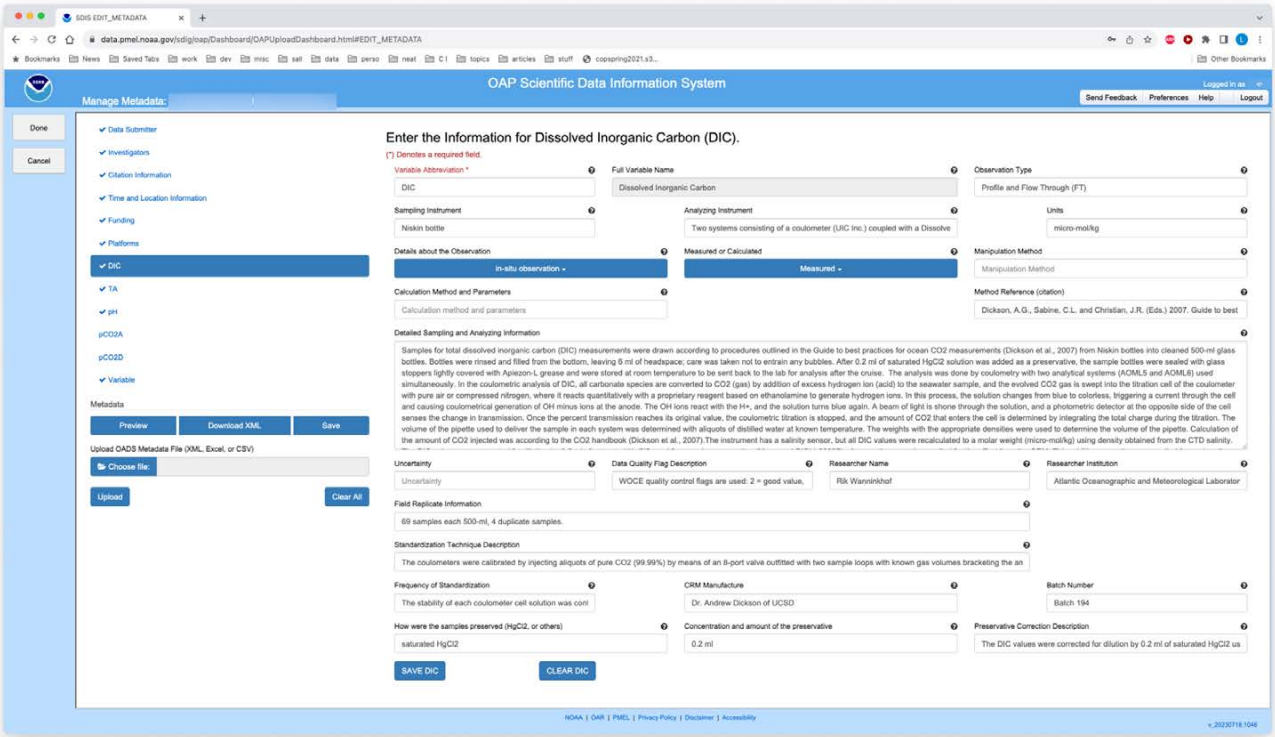


FIGURE 2. The SDIS Metadata Editor supports upload and editing of metadata records, including variable descriptions (aka Data Dictionary) shown here, as well as a presentation preview function and downloading of a metadata record that can be used as a template for future submissions.

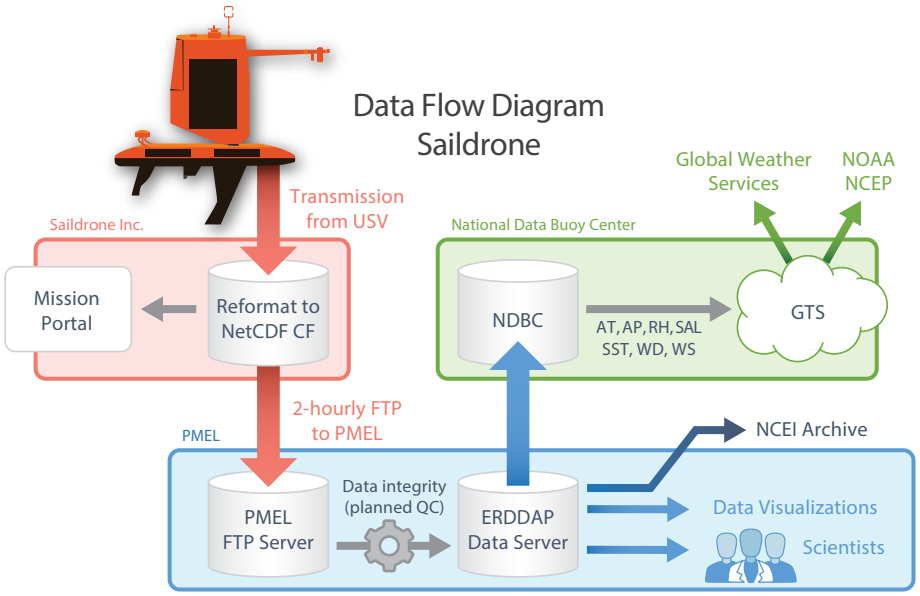


FIGURE 3. The data acceptance and data management workflow for data from Saildrone "data buys" is diagrammed here.

To this end, machine learning advances are focused on generalizing quality assurance and quality control assessments across observing systems and locations. Current procedures for this rely on specialist knowledge in order to design tests and metrics, which can vary across sensors, platforms, locations, and times.

The emerging field of scientific machine learning holds a lot of potential for applications in environmental data management. Contemporary challenges range from data discovery to metadata standards, quality control, and readiness. By leveraging the rigorous vetting methods from more traditional scientific tools, future machine learning developments at PMEL will enable transparent and explainable practices that add greater value to environmental data.

PMEL pioneered the use of leased in situ observation platforms (Meinig et al., 2019), ushering in the period of commercial "data buys" for in situ observations. This approach allows researchers to focus on the data they need, and not on the management of the observing system.

SDIG has responded with a data acceptance and data dissemination workflow that includes integrity checking (Figure 3). Soon to be integrated is level-0 quality control that applies machine learning methodologies. The combination of all these steps into a fully automated workflow has created today's approach that accepts hundreds of data files delivered from commercial data-buy vendors annually. This is another PMEL and SDIG innovation that holds great promise for the future.

CONCLUSION

Data management solutions developed at PMEL to address specific requirements have been integrated to address more complex needs. These complex implementations have been deployed at PMEL, across NOAA, and across the United States at other agencies and educational institutions. Data acquisition and data management approaches are evolving, and SDIG has demonstrated the capability to evolve in order to support emerging data management needs over the next 50 years, and to continue serving both PMEL and the at-large data management community.

REFERENCES

Aronova, E., K.S. Baker, and N. Oreskes. 2010. Big science and big data in biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) network, 1957–present. *Historical Studies in the Natural Sciences* 40(2):183–224, https://doi.org/10.1525/hsns.2010.40.2.183.

Bakker, D.C.E., B. Pfeil, C.S. Landa, N. Metzl, K.M. O'Brien, A. Olsen, K. Smith, C. Cosca, S. Harasawa, S.D. Jones, and others. 2016. A multi-decade record of high quality fCO_2 data in version 3 of the Surface Ocean CO_2 Atlas (SOCAT). *Earth System Science Data* 8:383–413, https://doi.org/10.5194/essd-8-383-2016.

Feely, R.A., L.-Q. Jiang, R. Wanninkhof, B.R. Carter, S.R. Alin, N. Bednaršek, and C.E. Cosca. 2023. Acidification of the global surface ocean: What we have learned from observations. *Oceanography* 36(2–3):120–129, https://doi.org/10.5670/oceanog.2023.222.

Jiang, L.Q., A. Kozyr, J.M. Relph, E.I. Ronje, L. Kamb, E. Burger, J. Myer, L. Nguyen, K.M. Arzayus, T. Boyer, and others. 2023. The Ocean Carbon and Acidification Data System. *Scientific Data* 10:136, https://doi.org/10.1038/s41597-023-02042-0.

Meinig, C., E.F. Burger, N. Cohen, E.D. Cokelet, M.F. Cronin, J.N. Cross, S. de Halleux, R. Jenkins, A.T. Jessup, C.W. Mordy, and others. 2019. Public private partnerships to advance regional ocean observing capabilities: A saildrone and NOAA PMEL case study and future considerations to expand to global scale observing. *Frontiers in Marine Science* 6:448, https://doi.org/10.3389/fmars.2019.00448.

ACKNOWLEDGMENTS

This article is dedicated to the data managers and software developers at PMEL who have contributed to the foundations of data management at PMEL and the PMEL Science Data Integration Group. PyFerret developers Steve Hankin, Ansley Manke, Karl Smith, Jerry Davison, and Kevin O'Brien developed a tool that is used globally. PyFerret development was supported by NOAA's Geophysical Fluid Dynamics Laboratory (GFDL) on the memorandum of understanding established between PMEL and GFDL. SOCAT, OSMC, and OceanSITES development was supported by the NOAA Global Ocean Monitoring and Observation program, grant number GC03-648. The NOAA Ocean Acidification funded development of metadata and data submission applications. Funding for machine learning and artificial intelligence was provided by the NOAA National Environmental Satellite, Data, and Information Service. This is PMEL contribution number 5498.

AUTHORS

Eugene F. Burger (eugene.burger@noaa.gov) is Research Services Division Director, NOAA Pacific Marine Environmental Laboratory (PMEL), Seattle, WA, USA. **Kevin M. O'Brien** is Senior Research Scientist, Cooperative Institute for Climate, Ocean, and Ecosystem Studies (CICOES), University of Washington, and NOAA PMEL, Seattle, WA, USA. **Steven Hankin** (retired) was at NOAA PMEL, Seattle, WA, USA. **Roland Schweitzer** is the owner of WeatherTop, College Station, TX, USA. **Linus Kamb** and **Sage Osborne** are both at CICOES, University of Washington, and NOAA PMEL, Seattle, WA, USA. **Ansley Manke** (retired) was at NOAA PMEL, Seattle, WA, USA.

ARTICLE CITATION

Burger, E.F., K.M. O'Brien, S. Hankin, R. Schweitzer, L. Kamb, S. Osborne, and A. Manke. 2023. Data processing and management at PMEL: A 50-year perspective. *Oceanography* 36(2–3):26–31, https://doi.org/10.5670/oceanog.2023.230.

COPYRIGHT & USAGE