Improving Access to Marine Glider Data to the Wider Community

Fisheries and Oceans Canada
Ocean Sciences Division
Sidney, BC, Canada

Julia Putko
V00889506
Work Term #2
Computer Science
juliaputko@uvic.ca
April 20, 2022

Table of Contents

List of Figures

Executive Summary

This report summarizes the investigation into different data management and accessibility methods of various organizations in comparison to the methods employed by the Canadian-Pacific Robotic Ocean Observing Facility (C-PROOF). C-PROOF's intended goals for the collected marine data include the sharing of data with a wide variety of user groups for the benefit of Canadians [1]. C-PROOF utilizes various methods to make glider data accessible. The primary methods are separated into data management and data access. Data management methods include accuracy of metadata conventions, naming conventions, and file types of data files. Data access methods refers to how data is shared with the wider community, including creating data acquisition methods for scientific partners, making data available on a public facing website and creating accurate and informative visualizations of data. The solutions employed by C-PROOF use concepts discussed in other data management plans applied by similar organizations that also manage large amounts of marine and environmental data, whose purpose is also to provide collected data to the wider community. C-PROOF prioritizes meeting metadata standards, organization and providing a public website that allows both members of outside scientific groups and members of the public to access the collected data.

Glossary

CCSM	Community Climate System Model.
C-PROOF	Canadian-Pacific Robotic Ocean Observing Facility
FERRET	An interactive netcdf visualization and analysis environment.
Glider	Autonomous marine vehicles used to collect data in the ocean.
GrADS	An interactive desktop visualization and analysis tool for NetCDF data.
Metadata	Data that gives information about data.
NCAR	National Center for Atmospheric Research.
NCO	NetCDF operators command line operators.
ncview	A simple NetCDF visual browsing tool.
NCL	The NCAR interpreted programming language analysis of scientific data.
NetCDF	(Network Common Data Form): A data format.
IOOS naming conventions	Conventions used by the Integrated Ocean Observing System (IOOS) program to assign identifiers.

1.0 Introduction

The data at the focus of this report comes from autonomous gliders operated by the Canadian-Pacific Robotic Ocean Observing Facility (C-PROOF), which move through the ocean off the coast of British Columbia, carrying instruments that collect data on the ocean state, including salinity and temperature, and transmit the data and location by satellite back to shore to be used by the wider community [1]. The purpose of the Canadian-Pacific Robotic Ocean Observing Facility (C-PROOF) is to understand the physical and biological ocean properties that drive variability in the ocean and in the climate, to better predict weather and climate changes across Canada [1]. Developing organized data management

practices and serving the data effectively is key to prevent delays in providing the collected data to the wider scientific community [1]. Accessibility requirements must consider the activities of both sets of users, their background information in the context of the data, and their intended uses, and solutions used depend on the amount of data, the type of data and who the data is meant to be used by. The intention of C-PROOF is to allow accessibility to data for users with a wide range of interests, ranging from users with unique data requirements, to users with little scientific or research background.
The practices and intentions employed by C-PROOF for data management and accessibility are compared to other practices that exist within the environmental and marine science research community and how these methods come to fruition in different areas, with the intentions of justifying the practices of C-PROOF.
The investigation into other data management methods in this report will attempt to encourage proper and most appropriate data management and accessibility practices for the current project, as well as serve the data in a way that best allows both the greater scientific community and the non-scientific public, to access marine glider data in the most appropriate forms that best serve their purposes. Solutions to prevent those delays in the case of C-PROOF involve creating efficient and standardized data management methods and creating data accessibility through a public facing website (cproof.uvic.ca), however various solutions exist in similar industries and projects.

2.0 Data Management and Accessibility Methods

2.1 Data Management

Science research projects or organizations whose focus or purpose involves data collection, require consistent data management methods to summarize intentions for data management over a course of a project. Methods used to organize, name, format, and make available data [2] are used in data management to make data more meaningful and usable outside the organization that accessed the data. The documentation aids in the preparation of data for future reuse, preservation, and sharing of data [2]. The management plans implemented depend on the project, the amount of data, and to which audiences the data will be served in the wider community.

2.1.1 Metadata

Large amounts of oceanographic data are not immediately usable. When dealing with large amounts of data it is necessary to create standards for data and metadata according to some convention so that automated analysis and efficient interaction with datasets is possible. A data management strategy applied to many datasets includes metadata for collected data and allows users to understand and use data files, and are essential for the automated analysis necessary to efficiently interact with large data collections [3]. The data management plan for the Community Climate System Model (CCSM), a global climate model, includes metadata as a part of their strategy to encourage automated analysis of their datasets [3]. C-PROOF data's intended use is to be shared, so having accurate and informative metadata is crucial. C-PROOF's data management plans exist in a proposal and in a more in depth in progress plan for the data. The plan within this document regarding data is compared to other existing publicly available data management plans for projects in similar areas of science and with similar purposes for their data and similar amount of data [3].

2.1.2 File Formats

A consideration to be made for metadata is which file formats make data most operable. File formats should allow for reuse, sharing and long term access to data. Converting the data into accessible file formats allows users to view and interact with the data through a large amount of software. There is a large number of publicly accessible software available for the interaction with NetCDF files including NCO, NCL, ncview, GrADS, FERRET, among others [4]. The CCSM project chose NetCDF as its data format due to the software tools that support this format [3]. Among other file formats, C-PROOF provides its data in NetCDF files that have metadata and attributes that should be compliant with US-IOOS standards, making the data accessible to existing visualization and analysis tools. Making data accessible

in file formats that do not require specialized software or hardware is an important consideration when trying to make data accessible for the scientific community.

2.1.3 Naming Conventions

Included in metadata, are the file naming conventions. Consistent naming conventions are important for accurately serving the data to the wider community. It is important that the names of files are consistent and are organized in a meaningful way making them easy to find. Clear organizational, standardized and accurate metadata allows for easy error detection, and long term management and storage of data. CCSM [3] uses the naming conventions outlined in a detailed document that accounts for different subsets of datum, including output files and processed datasets [6]. C-PROOF will use IOOS naming conventions for files, which includes proper labeling of different levels of processing. Similar conventions will be used as similar research projects [3]. C-PROOF's naming conventions identify the name of the project, unique identifiers for glider, project abbreviations, the start and end times of when the data was collected, and the type of data. Properly organized data with correct file names and complete and correct metadata are crucial for the organization of C-PROOF data, especially when accessing the data on web pages. Lack of organization in these areas can lead to data not appearing on the public facing website in the proper manner, making it difficult to search for and therefore use data. C-PROOF standards for metadata and data management that work best for the type of data provided and for the intended audience.

2.2 Data Access

There are various methods and platforms for sharing data in an accessible, discoverable or usable way with scientific groups and the non-scientific public. Providing accessible data informs the public and other researchers on the state of the oceans with the intention that, as a result, the global community will be able to better plan for the implications of the changing climate, and provide observational data for climate models and other tools critical for the understanding of the changing ocean state and making informed decisions [1]. There has been an increase in awareness and desire for engagement in climate change and environmental sciences. Public participation in science especially in climate and marine sciences is crucial to drive action intended to reduce the risks of climate change. The open flow of information and sharing data and results increases accuracy and verifiability of analyses and conclusions, and allows for the advancement of knowledge [7]. Sharing data effectively is crucial to accomplishing these goals. Not all groups that collect data make their data available in the same way, or at all. In some cases research groups may only be producing data to confirm their own hypotheses, and do not intend to share the specific data. It is still valuable to provide information for reuse, even in these cases, to aid in credibility and to encourage all scientific research data to be freely and publicly accessible and usable. Researchers may choose to only provide data by request, which usually requires contacting the researcher directly for access to the data. In this case, the researcher becomes responsible for the data, and may not have teh technical knowledge and resources to guarantee the data remains accessible and prioritizes the long term preservation of the data.

2.2.1 Online Access

Public online tools for access to data allow for greater transparency to for what purpose data is being collected, while also taking into consideration the different levels of understanding and purposes of user groups. The NCAR climate data guide is provided through an public online website which provides access to hundreds of environmental data sets [8]. After navigating to a specific dataset, a user is able to access metadata, figures, and descriptions of strength and limitation and application of climate data. The CCSM data is provided on a public webpage that organizes its data under descriptive headings and links that allow direct download of data [9]. With the data, analysis and visualization software, as well as other interaction tools that might be used with the data are accessible on the same page as the data. Some

projects use repositories that host the data of multiple project's data. Data repositories uphold the integrity of data and the metadata, allowing for easier discovery and searchability.

The data sharing method employed by C-PROOF, like many other organizations, is in the form of a project website with general information on glider lines, descriptive figures and maps that communicate information about the marine glider date, and other relevant news and publication information. The existing C-PROOF website (cproof.uvic.ca) provides command line tools to download subsets of data, links to download data [Figure 1], an interactive map to view data locations [Figure 2], subsets of metadata shown on individual glider pages [Figure 3], and figures for the wider community to access data.

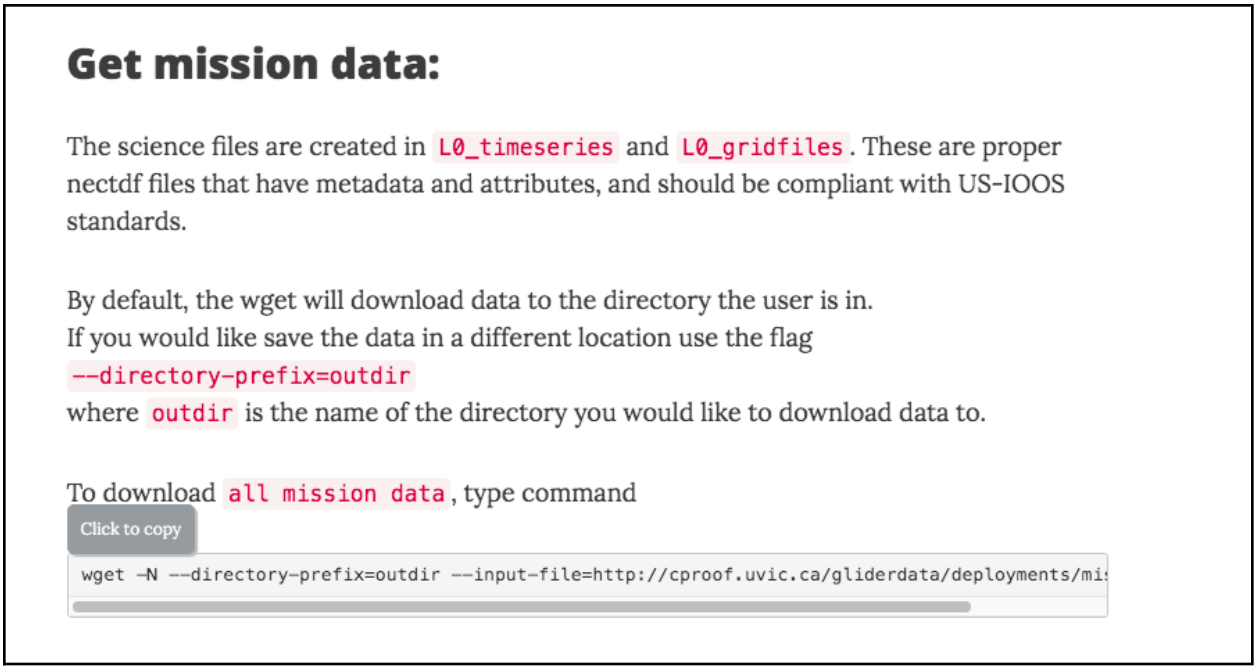

Figure 1. C-PROOF website Data Page command line tool instructions.

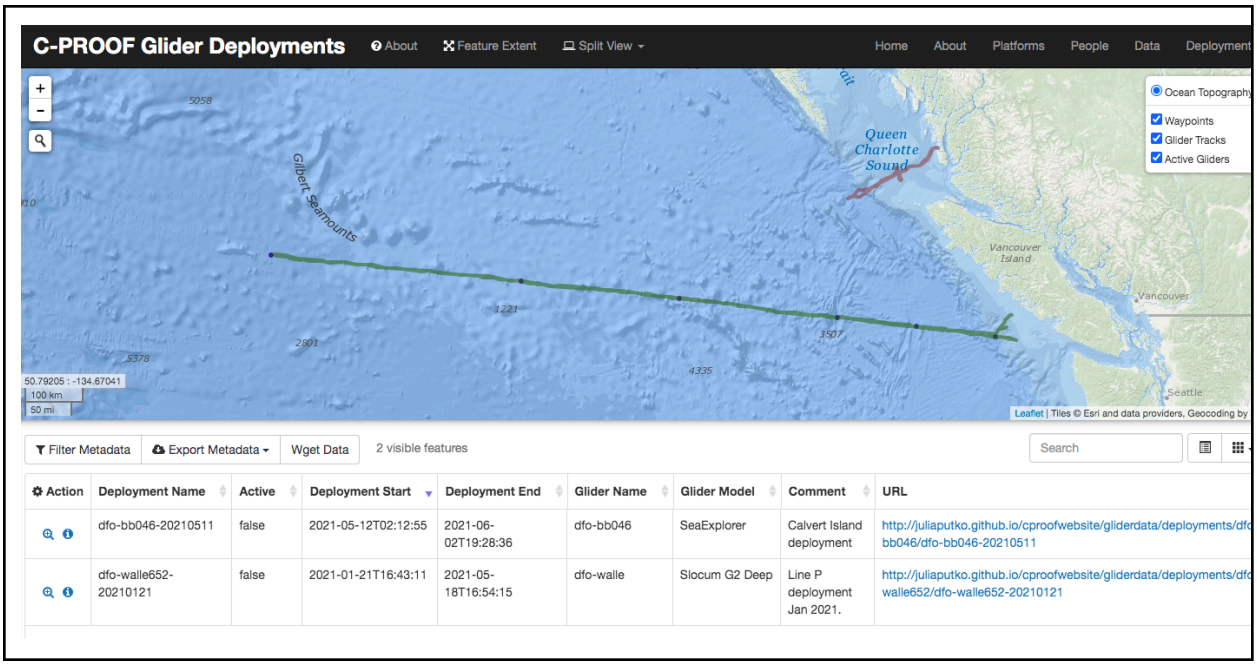

Figure 2. C-PROOF website Deployment Page

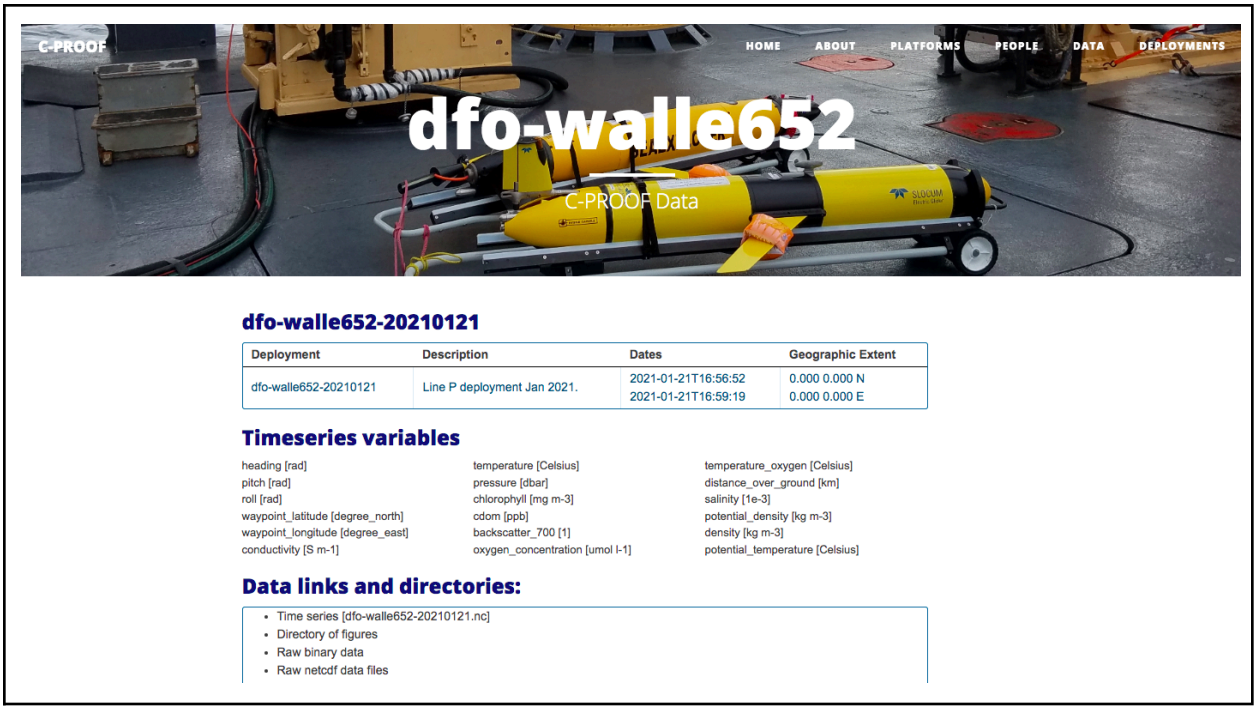

Figure 3. C-PROOF Mission Page for dfo-walle652-20210121 with related metadata.

C-PROOF's intention is also to supply processed data on a publicly available website to secondary repositories for archiving to ensure data is shared to scientists internationally effectively [1]. Updates to the written content describing different aspects of the project, additional data access methods, newly created figures, and improved discoverability and usability have been made to the C-PROOF website. Providing content in the form of media such as written content and other media contributes to education at all levels to contribute to understanding the impact of climate science, inspiring public

involvement with climate sciences and the protection of the ocean [10]. User testing was done within the C-PROOF group which allowed individuals informed on the purpose of the data for the wider marine science community to provide feedback on usability and discoverability and access to data. Providing multiple methods and avenues to access data allows for wider use and availability of data.

2.2.2 Figures

Figures provide standard information about the data collected that allows users to comprehend large amounts of data. The format and content of visualizations can be used to inform the public about the state of ocean observation and relay important data to the wider community.
Figures should be relevant and well described so that a wide variety of user groups can derive information from the visualizations. The Arm Data Quality Office, within the Cooperative Institute for Severe and High-Impact Weather Research and Operations, provides an online interactive visualization tool for their datasets [11],[12]. The tool allows the user to apply filters to the data, view the location of the available data, the different variable data, type of data, as well as the ability to condense the start and end date shown in the visualization. The visualization tool also provides additional data such as the metadata information of the datafile being viewed, and minimum, maximum, means and standard deviation of the variable.

C-PROOF, rather, uses static figures to describe its datasets through visualizations. Similar to the Arm visualization, the figures provided on the C-PROOF website are also gridded, have titles and axis labels, and accurately describe the data [Figure 4].

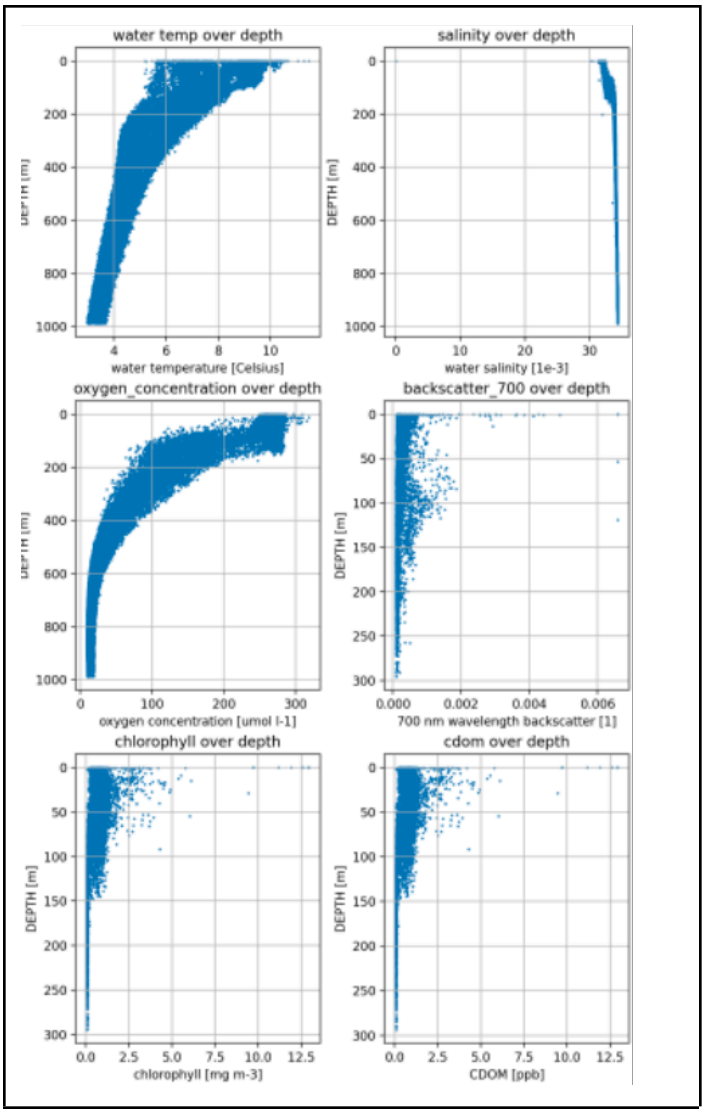

Figure 4. Dfo-walle652-20210121 variable plot.

3.0 Conclusion

Providing meaningfully accessible data that is available to a wide range of user groups aids in the advancement of science and research, which as a result helps to implement better scientific understanding into ocean models and technologies, supports scientific collaborators across the globe, and encourages public understanding about the state of the ocean. Different solutions for glider data management are employed by C-PROOF in the form of organized data management and an accessible public website. Providing consistent metadata minimizes confusion, allows for organization in all areas where the data might be used, and aids in the discovery and use of large collections of data. Online web pages provide ease of use and efficiency when trying to access data. Creating different methods for accessing data, such as various data acquisition methods and visualization tools allows for more meaningfully accessible data, especially in the case of large amounts of complex datasets. It is ideal to have a well documented data management plan that encompasses both the preprocessing of data and the methods in making it available to the wider community.

4.0 Recommendations

There are continuous changes that can be made, and should be made to continuously make data management and accessibility more available to the research and scientific community, but also the

non-scientific public. Incorporating more informative metadata, such as version history in the metadata of NetCDF files, which includes reprocessing and metadata changes for old missions, is a possible implementation into C-PROOF's data management plan [5]. It is crucial to acknowledge that the public is willing to engage with information on the state of the ocean, and to continue to provide accessibility in ways that involve users with less scientific or technological background, but still allows them to benefit from information about scientific data. Further work on the public facing website can also be done, including further additions to the glider mission pages and other public facing pages on the C-PROOF website that include visualizations and information that captures public interest. In order to move towards open access to data for all projects that include data collections, recognizing and rewarding researchers that contribute to data collection is crucial. To better serve the users of accessed datasets, information about the intended use of the data from users downloading the data could also be collected in order to demonstrate how the project's data is serving the scientific and global community. Access to data requires consistent upkeep of accessibility requirements, as well as continued work remedying missing data and developing methods for error checking in the preprocessing of data.

References

Cited References

[1] Klymak, J., "Canadian Pacific Robotic Ocean Observing Facility Innovation Fund 2017 Proposal," October 2016. [Online document]. [Accessed: 20-April-2022].

[2] "Data Management Plans," *Portagenetwork.ca.* [Online]. Available at: https://libguides.uvic.ca/ld.php?content_id=36072978. [Accessed: 19-April-2022].

[3] "CCSM Data Management Plan," August 2005, [In Revision November 2010]. [Online] Available at: https://www.cesm.ucar.edu/experiments/data.mgmt.plan.050803.html. [Accessed: 19-April-2022].

[4] "Software for Manipulating or Displaying NetCDF Data," *Unidata*, UCAR. [Online document]. Available at: https://www.unidata.ucar.edu/software/netcdf/software.html. [Accessed: 19-April-2022].

[5] Dosser, H. V., "CPROOF Data Management Plan Version 0.1," March 30 2022.

[6] "CCSM3 Output Filename Requirements," *CESM*, Ucar. [Online document]. Available at: http://www.cesm.ucar.edu/experiments/filename_conventions.html. [Accessed: 19-April-2022].

[7] National Academy of Sciences (US), National Academy of Engineering (US) and Institute of Medicine (US) Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. "Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age," *Washington (DC): National Academies Press* (US); 2009. 3, Ensuring Access to Research Data. [Online]. Available at: https://www.ncbi.nlm.nih.gov/books/NBK215271/. [Accessed: 19-April-2022].

[8] "NCAR Climate Data Guide Content with Tag: Ocean: NCAR - Climate Data Guide," *NCAR Climate Data Guide Content with Tag: Ocean | NCAR - Climate Data Guide*. [Online]. Available At: https://climatedataguide.ucar.edu/variables/ocean. [Accessed: 19-April-2022].

[9] "Climate Data at the National Center for Atmospheric Research," *NCAR Climate Data Gateway Atom*. [Online]. Available at: http://www.earthsystemgrid.org/. [Accessed: 19-April-2022].

[10] L. Mabon, "Making Climate Information Services accessible to communities: What can we learn from environmental risk communication research?," *Urban Climate*, vol. 31, p. 100537, 2020.

[11] "Data" ARM. [Online]. Available at: https://www.arm.gov/data/https://adc.arm.gov/discovery/#/results/meas_category_code::svcprop. [Accessed: 19-April-2022].

[12] "DQ-Zoom PLotter." ARM. [Online]. Available at: https://dq.arm.gov/dq-zoom/?ds=asi30qcecorM1.s1&variable=latent_heat_flux&sdate=20171105&edate=20171105&coordinate=. [Accessed: 19-April-2022].

General References

D. Kapiszewski and S. Karcher, "Making Research Data Accessible," in The Production of Knowledge: Enhancing Progress in Social Science, C. Elman, J. Gerring, and J. Mahoney, Eds. Cambridge: Cambridge University Press, 2020, pp. 197–220.

Data Policy, Ocean Networks Canada. [Online]. Available at: https://www.oceannetworks.ca/data-tools/data-help/data-policy [Accessed: 5-April-2022].

Gertler,L.,A., Bullock, J.G., "Reference Rot: An Emerging Threat to Transparency in Political Science," https://johnbullock.org/papers/referenceRot.pdf
gertler and bullock, 2017, 167)

Government Accountability Office. 2007." Climate Change Research: Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research." September. Available at http://www.gao.gov/new.items/d071172.pdf [Accessed: 5-April-2022].

"ONC Data Policy." Ocean Networks Canada [Online]. Available at: https://www.oceannetworks.ca/sites/default/files/pdf/Data%20Policy%202021-05-03.pdf [Accessed: 5-April-2022].

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. "The FAIR Guiding Principles for scientific data management and stewardship." Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

Willis, Josh & National Center for Atmospheric Research Staff (Eds), Feb 27 2017, "The Climate Data Guide: Argo Ocean Temperature and Salinity Profiles." Available at: https://climatedataguide.ucar.edu/climate-data/argo-ocean-temperature-and-salinity-profiles [Accessed: 5-April-2022].