Science Storms the Cloud

C. L. Gentemann[1,2], C. Holdgraf[3,4], R. Abernathey[3,5], D. Crichton[6], J. Colliander[3,7,8], E. J. Kearns[9], Y. Panda[3], R. P. Signell[10]

[1]Farallon Institute, Petaluma, CA, [2]Earth and Space Research, Seattle, WA, [3]2i2c, Berkeley, CA, [4]International Computer Science Institute, Berkeley, CA, [5]Lamont Doherty Earth Observatory of Columbia University, Palisades, NY, [6]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, [7]Pacific Institute for the Mathematical Sciences, Vancouver, BC, Canada, [8]University of British Columbia, Vancouver, BC, Canada, [9]First Street Foundation, Brooklyn, NY, [10]US Geological Survey, Woods Hole, MA

Corresponding author: Chelle Gentemann (cgentemann@faralloninstitute.org)

Key Points:

- Science stands at the cusp of a new, open science, cloud-enabled era
- Advances in data, software, and computing are enabling transformational, interdisciplinary science, changing the realm of possible questions
- Deliberately designed open science communities can advance science and inclusivity simultaneously

Abstract

The core tools of science (data, software, and computers) are undergoing a rapid and historic evolution, changing what questions scientists ask and how they find answers. Earth science data are being transformed into new formats optimized for cloud storage that enable rapid analysis of multi-petabyte datasets. Datasets are moving from archive centers to vast cloud data storage, adjacent to massive server farms. Open source cloud-based data science platforms, accessed through a web-browser window, are enabling advanced, collaborative, interdisciplinary science to be performed wherever scientists can connect to the internet. Specialized software and hardware for machine learning and artificial intelligence (AI/ML) are being integrated into data science platforms, making them more accessible to average scientists. Increasing amounts of data and computational power in the cloud are unlocking new approaches for data-driven discovery. For the first time, it is truly feasible for scientists to bring their analysis to data in the cloud without specialized cloud computing knowledge. This shift in paradigm has the potential to lower the threshold for entry, expand the science community, and increase opportunities for collaboration while promoting scientific innovation, transparency, and reproducibility. Yet, we have all witnessed promising new tools which seem harmless and beneficial at the outset become damaging or limiting. What do we need to consider as this new way of doing science is evolving?

Plain Language Summary

For a long time, scientists have downloaded data and analyzed it on their computer. This made collaborating hard because other people didn't have access to the same data, software, and computer. It also gave scientists at big institutions with fast internet and lots of computers an advantage. Now, data are being put on the cloud, scientists are sharing their software, and anyone can access a computer on the cloud through their web browser. This makes it easier to collaborate because everyone can access the same data, software, and computer. Also, more people can access powerful computers and do science. This is a different way of doing science and there are potential drawbacks. We need to be careful that this new way of doing science actually advances science and includes more people so that we get better answers, faster.

1 Introduction

> *"New directions in science are launched by new tools much more often than by new concepts. The effect of a concept-driven revolution is to explain old things in new ways. The effect of a tool-driven revolution is to discover new things that have to be explained."*
> *--Freeman Dyson*

Since the advent of scientific computing, computers have driven major scientific breakthroughs. We have peered into deep space, developed models to predict our weather and climate, and sequenced the entire human genome. There is no question that computers have advanced science and improved lives. Yet, challenges around computing have frustrated researchers, driving efforts to improve efficiency through data standardization, development of common software tools, and connecting computers into a cluster. For many research topics, the Pareto Principle's 80/20 distribution 'rule' applies (Pareto, 1971). 80% of time on a project is spent 'data-wrangling' (downloading, storing, maintaining a private archive of data and developing software to access it), leaving only 20% for efforts to analyze results and share what was learned. This ratio changes depending on what level of institutional support is provided to

researchers, creating inequalities and barriers to research. Larger organizations may have invested in vast private data storage, powerful computer clusters, and technology support. At least in part, at top-tier institutions, cutting-edge transformational science is enabled by the infrastructure that these institutions have built up over decades, and this resource is not always available to others. In other words, while computers have undoubtedly advanced science, they have also perpetuated and strengthened some inequalities.

Challenges around data storage and management, a reliance on outdated programming languages, and limitations around access to powerful computers are barriers to accomplishing science. In this article we discuss how innovations in data access, software tools, and computer infrastructure are fundamentally changing how science is accomplished and who is able to participate. We believe this shift is going to change the realm of possible questions and our ability to answer them. The speed and impact of this shift will, in part, depend on whether this new way of doing science is able to empower more voices to yield better, stronger solutions.

2 Innovations in Data

> *"Paradigm shifts arise when the dominant paradigm under which normal science operates is rendered incompatible with new phenomena, facilitating the adoption of a new theory or paradigm." (Thomas, 1962)*

When scientists need data, they often turn to agency archive centers for access. Data is downloaded, stored locally, and networked to a computer for analysis. Large datasets can take weeks or months to download, and when a new version of the data is released, the process must be repeated. Many scientists at smaller institutions or in developing countries don't have the bandwidth or infrastructure to handle these data, limiting their ability to do science. Data are being moved from archive centers to vast cloud data storage facilities. NASA has partnered with Amazon Web Services (AWS) in a Space Act Agreement to provide free access to NASA datasets stored on AWS (NASA, 2020). NOAA has partnered with multiple public commercial cloud providers for their Big Data Project (BDP) to enable free, cloud-based data storage and access for users of the most popular portions of their data holdings (NOAA, 2020). Through these partnerships, data are public and can be freely accessed or downloaded (Text S1). These agencies are in the midst of a historic transformation in data access, moving data from archive centers to public commercial cloud and national cloud storage facilities adjacent to server farms.

Beyond where the data are stored, how the data are stored determines how fast and easy it can be accessed. For example, NOAA datasets made available through integrations with highly-scalable data warehouse tools such as Google's BigQuery have been observed to be used at rates 1000x greater than when they were only available from an agency archive (Kearns et al., 2018). Traditional databases are giving way to highly scalable formats that can accommodate heavy search loads with faster performance. Cloud-optimized data are organized into 'chunks' of data, making it possible to distribute the access to 100 Gigabytes (GB) across 100 machines. The open science community Pangeo Project (http://pangeo.io/) (Abernathey et al., 2017) is creating a 'Pangeo Forge', to crowdsource and automate the creation of cloud optimized data (https://github.com/pangeo-forge/roadmap). This change in how data are accessed is due to improvements in the software used to access data (Abernathey et al., 2020).

3 Innovations in Software

For decades, science has relied on fast compiled programming languages, such as Fortran and C, and commercial interpreted analysis languages such as Matlab, S-Plus, and Interactive Data Language (IDL). The reliance on expensive commercial software for scientific analysis directly reinforces the inequities between wealthy, privileged institutions and scientists and those from economically disadvantaged communities and the developing world. Also, these languages are rarely used outside of the science community and therefore 1) there simply aren't many people to ask for help when stuck on a problem and 2) there aren't many community-developed software tools (Fangohr, 2004; Momcheva & Tollerud, 2015). The open source languages Python and R have experienced a substantial growth in popularity over the last decade (Figure S1). Software based on an open source language encourages others to build open source tools that can be widely shared, incrementally improved, and adopted by large communities as they mature.

The shift in science towards using and participating in the development of open source software libraries has enabled rapid innovations and software improvements. Contributors to open source libraries help eliminate programming errors, improve documentation, and extend capabilities to broaden applicability. For example, the Python Xarray software library (Hoyer & Hamman, 2017). Xarray provides a powerful and easy-to-use toolkit for analysis of structured files common in Earth science (eg. Network Common Data Form (NetCDF), GRIB, and gridded raster). Xarray was built on top of other layers of the scientific python software ecosystem, specifically NumPy (Harris et al., 2020) and Pandas (McKinney, 2010). While only three software libraries required for a basic Xarray installation, there are 21 more optional ones, such as the plotting library Matplotlib (Hunter, 2007), analysis library Scipy (Virtanen et al., 2020), and parallel computing library, Dask (Rocklin, 2015).

The integration of these disparate software libraries in the service of Earth system science doesn't happen by magic. Coordinated efforts from funding agencies, such as NSF EarthCube's funding of the Pangeo Project, helped accelerate and coordinate the development of Xarray and Dask to meet the needs of science users. Other agencies are also recognizing the value of these software libraries to science. For example, in 2020, NASA released a request for proposals "for the improvement and sustainment of high-value, open source tools, frameworks, and libraries" (https://tinyurl.com/nasaE7OSS).

4 Innovations in Computation

To help scientists handle increasingly large and complex datasets, the default response by institutions is often to purchase a local computing cluster. While local computer clusters can be efficient and cost-effective when fully utilized, only a select few institutions can afford them. This excludes vast parts of the scientific community and creates a have-have/not situation. These are computing fortresses that only the lucky can enter. Much like a medieval fortress, the infrastructure ages rapidly, requires constant maintenance, and is not as agile as science often requires. The closed environment means that collaborating with outside investigators can be challenging, often there is an application process and a steep learning curve to understand the computational environment.

Public commercial cloud computing solutions offer data storage and computing services, which can be provisioned and scaled by anyone, on demand. Three commercial providers (AWS,

Google Cloud Platform (GCP), and Microsoft Azure) dominate the market, but others also offer competitive solutions (Digital Ocean, Wasabi, OVH). Science funding agencies are also experimenting with operating their own clouds (e.g. NSF Jetstream). Unfortunately, accessing cloud resources requires specialized expertise. Configuring a 'computer on the cloud' involves selecting virtual machines, data storage, setting security access rules, monitoring costs, and other technical decisions. As scientific analyses are moved to the cloud, it is important that we do not re-create the same barriers that researchers currently experience with local computer clusters.

Fortunately, there is an ecosystem of tools, organizations, and communities that has grown around open and vendor-agnostic approaches to research. For example, JupyterHub provides an easily accessed common data science platform that removes interoperability as a barrier to collaboration. The computing environment, whether running on local or remote cloud infrastructure, can be accessed through any browser window. Other tools in the Jupyter ecosystem (such as JupyterLab and the Jupyter Notebook) provide domain- and vendor-agnostic interfaces for software development. These tools are already the default for most data scientists and are rapidly being adopted by others who require computational notebooks (Perkel, 2018). JupyterHubs can separate and consolidate the maintenance of running shared infrastructure from the act of doing science.

Managing secure, cost effective access to JupyterHubs for scientists will likely look different depending on how the research is funded. Some agencies have invested in an agency-managed cloud solution for their researchers (eg. NSF's JetStream). Institutions like the Norwegian Institute for Water Research maintain and manage a GCP JupyterHub available to all of their researchers. To promote open science in the social sciences, the Leibniz Institute for the Social Sciences provides free persistent JupyterHub environments (https://notebooks.gesis.org/). Several companies have formed to meet the needs of both industry and science, such as Coiled (https://coiled.io/), and Saturn (https://www.saturncloud.io/), and the nonprofit International Interactive Computing Collaboration (https://2i2c.org). These companies make managed cloud infrastructure accessible for smaller organizations and individuals, ensuring that large institutions or agency-affiliate researchers aren't at a 'cloud' advantage. The new scientific workflow (Figure 1) illustrates the shift in how science is accomplished on the cloud, whether the cloud is a public commercial cloud or national science cloud like JetStream.

5 Putting it all together

Open, cloud-based science is already starting to occur. In this section, we present an example of a key open source tool that advances science and several science results that do open, cloud-based science. Advancing reproducibility: The Binder project (https://mybinder.org) combines open software and cloud computing to advance reproducibility and simplify sharing among teams. Through a simple web browser window, Binder connects users, in one-click, with an interactive cloud-based JupyterHub that is running a user-specified collection of computational notebooks. With over 100,000 weekly users, this project is changing how scientists share reproducible analyses (Holdgraf, 2020; Text S2). New science: NOAA has collected over 200 TBs of whale calls using seafloor mounted acoustic recorders. Listening to the data would take over 19 years. Researchers developed a convolutional neural network to automate identification of beluga whale calls (https://github.com/microsoft/belugasounds) (Zhong et al., 2020). This entire dataset is being analyzed by scientists for the first time to understand where whales are, how they move, and how changing ocean conditions affect their

Figure 1. Science is changing as data, software, and computers are coming together on the cloud. Scientists can access massive cloud computing resources through a web browser window, effectively putting a super-computer into any internet-connected device.

population. Research to operations: Rapid estimation of hurricane strength and heading is critical to allocating emergency resources. Trained meteorologists estimate hurricane intensity using satellite imagery matched to known patterns. NASA artificial intelligence experts automated hurricane classification (http://hurricane.dsig.net/) (Pradhan et al., 2018), reducing the latency in communicating major threats to the public. Societal impact: Researchers combined NOAA and USGS open cloud-optimized data, open software, and cloud computing to produce flood risk scores for over 140 million properties in the U.S. (https://floodfactor.com/) (Kearns et al., 2020). These scores are easily communicated to and consumed by the American public, enabling complex science to be translated into simple, practical information products.

6 Challenges

How do we ensure that this new way of doing science does not just swap one system's challenges and inequalities for different ones? In some ways, our rush to expand into the cloud is already experiencing growing pains. In this section we discuss several challenges and provide additional discussion in Text S3.

The federal agencies that fund science move slowly, and while this provides stability that gives science a solid foundation, this inertia can also open up gaps in support when there is a major shift in community needs. For example, cloud-based datasets still require careful data curation, metadata standards, and data provision from trusted sources. By reducing the barriers to creating, publishing, accessing and using data, we may increase the potential for inadvertent mis-use by users not familiar with scientific data practices, version controls, and trusted repositories.

New approaches to communicating 'data-best-practices' and how to identify trusted sources are already needed because data are already on the cloud. Scientists require more training in software best practices and in how to share software for reproducible results. Existing data archives that scientists are already familiar with, along with groups focused on education (eg. Openscapes, https://www.openscapes.org/ and The Carpentries, https://carpentries.org/), could play a central role in advancing this data and software literacy, but this will require prioritized support from agencies.

A central question that must be resolved to realize the vision of a large-scale migration to cloud-based science is who pays for the cloud computing and does this create incentives that affect science? Traditionally, the cost of computing infrastructure has been borne primarily by funding agencies (e.g. NSF, via grants to individual PIs as well as large-scale facilities) and by research institutions (via institutional support for computing hardware and support staff). As organizations shift budgets to pay for cloud infrastructure, it raises the question of which services or infrastructure should be removed from within the university. How can we ensure that cloud infrastructure is utilized in partnership with local infrastructure, so that their relative strengths are utilized, rather than an "all or nothing" proposition? Accessing secure, scalable, cloud computing requires technical expertise and ongoing cost oversight. How do we ensure that access to cloud computing doesn't simply replicate a situation where science is restricted to the privileged, well-funded, connected, few? There is also a risk of becoming too dependent on providers of cloud computing services. Who should be the service providers in this new cloud-native world? Programs like NSF's cloud bank (https://www.cloudbank.org/), companies like Coiled, and non-profit organizations like 2i2c can all play the role of an intermediary between the cloud providers and individual scientists, giving the scientific community greater leverage and control over their infrastructure choices.

Finally, as we advocate for open science it is important to recognize that openness that advances science is not a pure product of technology, it is a product of practices, norms, and community behavior around that technology. Just as new technology requires designing new workflows, it is important to deliberately design a new community infrastructure that is welcoming to a more diverse community, strategically directs support and community dynamics to include marginalized groups, and recognizes how previous power dynamics in science act to exclude groups from participation. As an example, the Pangeo Project defines itself as a "community platform", emphasizing both a focus on cutting edge open science and building community dynamics that are open, inclusive, deliberate, and that balance power across the many stakeholders in the ecosystem. Participants are asked to abide by a Code of Conduct (eg. https://tinyurl.com/pangeoCC; Text S4). How is this work to create inclusive open communities that advance science prioritized when this work isn't recognized as a contribution to science in most academic, commercial, and agency performance or tenure evaluations?

7 Conclusions

> *"A new generation of information technology tools and services holds the potential of further revolutionizing scientific practice... These tools and services will have maximum impact when used within an open science ecosystem" (National Academies of Sciences, 2018)*

Data, software, computers. These tools are already being combined to advance science, but to really enable transformational science, open science has to be the core design principle

integrated into all efforts moving forward. Open science is “research conducted openly and transparently” (National Academies of Sciences, 2018). Open data makes results reproducible. Open software creates community tools that advance science faster and can reduce the effort to reproduce and build on results. Open compute means building data science platforms and software services that have an open infrastructure that is entirely vendor agnostic and is accessible to anyone.

There is now a rich ecosystem of easily-accessible data, server-side computation, open source software tools, and one-click-to-compute cloud computing data science platforms that enable research at a scale and ease unimaginable only a few years ago (Text S5). Practically, for scientists, the effect of these changes is to vastly shrink the amount of time spent acquiring and processing data, freeing up more time for science. This shift in paradigm is lowering the threshold for entry, expanding the science community, and increasing opportunities for collaboration, while promoting scientific innovation, transparency, and reproducibility. Communities can work together to reduce barriers and create a powerful force for innovation. The more diverse the minds working together, the better chance we have to identify and remove barriers to innovation. Building open science on the cloud creates that same innovative community but without many of the previous barriers to collaborations.The community is open and scientists can collaborate with anyone, regardless of their affiliation, nationality, or location. Potentially, this transformation may free up researchers' time for science and create a space where more leaps in our understanding will be common and breakthrough interdisciplinary collaborations can flourish. Technology can be a two-edged sword, adding new barriers as it removes older ones. As we move towards this new way to do science, designing our new playground around open science will enable honest conversations around barriers to participation in science and help us move forward, both faster and together.

Acknowledgments, Samples, and Data

This paper originated from a series of discussions between the co-authors on the transformative nature of cloud computing and open, inclusive science. The original draft was written by Gentemann with edits and comments from all co-authors. We wish to thank Marisol Garcia Reyes and Jeffrey Dorman, AGU editor Bjorn Stevens, reviewer Geert Jan van Oldenborgh, reviewer Daniel Nowacki, and 2 anonymous reviewers for their thoughtful analysis and suggestions. Gentemann received support from the Inter-agency Implementation and Advanced Concepts Team (IMPACT) at NASA's Marshall Space Flight Center. No data was used in this paper.

References

Abernathey, R., Paul, K., Hamman, J., Rocklin, M., Lepore, C., Tippett, M., et al. (2017). Pangeo NSF Earthcube Proposal. https://doi.org/10.6084/m9.figshare.5361094.v1

Abernathey, R., Augspurger, T., Banihirwe, A., Blackmon-Luca, C. C., Crone, T. J., Gentemann, C. L., et al. (2020). Cloud-Native Repositories for Big Scientific Data. *Authorea*. https://doi.org/10.22541/au.160443768.88917719/v1

Fangohr, H. (2004). A Comparison of C, MATLAB, and Python as Teaching Languages in Engineering. In M. Bubak, G. D. van Albada, P. M. A. Sloot, & J. Dongarra (Eds.), *Computational Science - ICCS 2004* (pp. 1210–1217). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-25944-2_157

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Holdgraf, C. (2020). A 2019 retrospective from the Binder Project. *Jupyter Blog*, https://blog.jupyter.org/a-2019-retrospective-from-the-binder-project-57a449517362

Hoyer, S., & Hamman, J. (2017). xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software*, 5(1), 10. https://doi.org/10.5334/jors.148

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kearns, E., Amodeo, M., & Porter, J. (2020). Do You Know Your Home's Flood Risk? *Eos*, *101*. https://doi.org/10.1029/2020EO146389

Kearns, Edward, Glass, Shane, Brown, O., Brannock, J., Simonson, A., & Simonson, A. (2018). Making Data Available on the Cloud for Decision Support Applications through NOAA's Big Data Project. *In 98th American Meteorological Society Annual Meeting*. Austin, Texas: AMS.

McKinney, W. (2010). Data Structures for Statistical Computing in Python (pp. 56–61). *Presented at the Python in Science Conference*, Austin, Texas. https://doi.org/10.25080/Majora-92bf1922-00a

Momcheva, I., & Tollerud, E. (2015). Software Use in Astronomy: an Informal Survey. ArXiv:1507.03989 [Astro-Ph]. Retrieved from http://arxiv.org/abs/1507.03989

NASA. (2020, 30). Current Space Act Agreements. NASA. Retrieved from https://www.nasa.gov/partnerships/about.html

National Academies of Sciences, E. (2018). Open Science by Design: Realizing a Vision for 21st Century Research. https://doi.org/10.17226/25116

NOAA. (2020, July). NOAA's Data Strategy Maximizing the Value of NOAA Data. NOAA. Retrieved from https://nrc.noaa.gov/Portals/0/Final%20Data%20Strategy.pdf?ver=2020-07-02-122524-377

Pareto, V. (1971). Manual of political economy. A.S. Schwier, Trans.

Perkel, J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. *Nature*, 563(7732), 145–147.

Pradhan, R., Aygun, R. S., Maskey, M., Ramachandran, R., & Cecil, D. J. (2018). Tropical Cyclone Intensity Estimation Using a Deep Convolutional Neural Network. *IEEE Transactions on Image Processing*, *27*(2), 692–702. https://doi.org/10.1109/TIP.2017.2766358

Raspaud, M., Hoese, D., Dybbroe, A., Lahtinen, P., Devasthale, A., Itkin, M., et al. (2018). PyTroll: An Open-Source, Community-Driven Python Framework to Process Earth Observation Satellite Data. *Bulletin of the American Meteorological Society*, *99*(7), 1329–1336. https://doi.org/10.1175/BAMS-D-17-0277.1

Raspaud, M., Hoese, D., Lahtinen, P., Dybbroe, A., Finkensieper, S., Holl, G., et al. (2020). pytroll/satpy: Version 0.22.0. *Zenodo*. https://doi.org/10.5281/zenodo.3888695

Rocklin, M. (2015). Dask: Parallel computation with blocked algorithms and task scheduling. *In Proceedings of the 14th python in science conference*. Citeseer.

Thomas, K. (1962). The Structure of Scientific Revolutions. Retrieved September 14, 2020, from https://philpapers.org/rec/KUHTSO-10

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Zhong, M., Castellote, M., Dodhia, R., Ferres, J. L., Keogh, M., & Brewer, A. (2020). Beluga whale acoustic signal classification using deep learning neural network models. *The Journal of the Acoustical Society of America*, 147(3), 1834. https://doi.org/10.1121/10.0000921