Educational resources to accelerate analysis of remote sensing data using cloud resources with Xarray

Emma Marshall [1], **Deepak Cherian** [2], **Scott Henderson** [3], **Jessica Scheick** [4], **and Richard Forster** [1]

1 Department of Geography, University of Utah, Salt Lake City, UT, U.S.A. **2** Earthmover PBC **3** eScience Institute & Department of Earth and Space Sciences, University of Washington, Seattle, WA, U.S.A. **4** Earth Systems Research Center, University of New Hampshire, Durham, NH, U.S.A.

Summary

Advances in cloud computing, remote sensing, and engineering are transforming earth system science into an increasingly data-intensive field, requiring students and scientists to learn a broad range of new skills related to scientific programming, data management, and cloud infrastructure (Abernathey et al., 2021; Gentemann et al., 2021; Guo, 2017; Mathieu et al., 2017; Ramachandran et al., 2021; Wagemann et al., 2021; Wagemann et al., 2022). This work contains educational modules designed to reduce barriers to interacting with large, complex, cloud-hosted remote sensing datasets using open-source computational tools and software. The goal of these materials is to demonstrate and promote the rigorous investigation of n-dimensional multi-sensor satellite imagery datasets through scientific programming. These tutorials feature publicly available satellite imagery with global coverage and commonly used sensors such as optical and synthetic aperture radar data with different levels of processing. We include thorough discussions of specific data formats and demonstrate access patterns for two popular cloud infrastructure platforms (Amazon Web Services and Microsoft Planetary Computer) as well as public cloud computational resources for remote sensing data processing at Alaska Satellite Facility (ASF).

Statement of Need

Research on the transition to data-intensive, cloud-based science highlights the need for knowledge development to accompany technological advances in order to realize the benefit of these transformations (Abernathey et al., 2021; Gentemann et al., 2021; Guo, 2017; Martin Sudmanns & Blaschke, 2020; Mathieu et al., 2017; Palumbo et al., 2017; Radočaj et al., 2020; Ramachandran et al., 2021; Wagemann et al., 2021; Wagemann et al., 2022).

These educational modules address this need and are guided by principles identified in Diataxis (Procida, n.d.) in order to help analysts engage in data-driven scientific discovery using cloud-based data and open-source tools.

Content

We present two tutorials focusing on publicly available satellite imagery datasets. Both types of satellite imagery featured in these tutorials are 1) large in volume, 2) accessed as cloud-optimized data types and making use of cloud computing resources, and 3) have associated metadata that is crucial to data management and interpretation but can be complicated to work with.

The tutorials discuss remote sensing principles and note considerations when evaluating and interpreting data, such as resolution, possible distortions, and noise. We also compare two datasets derived from the same source data but created with slightly different processing pipelines in order to illustrate the impact of dataset selection and processing decisions on analytical outcomes. Throughout both tutorials, we emphasize the steps and skills involved with multi-dimensional data cube workflows and preparing data for analysis.

Instructional Design

We designed these tutorials to be accessible to users with various experiences and backgrounds. Emphasis is placed on discussing how to interact with and examine complex, cloud-optimized datasets rather than simply providing example code snippets. To facilitate skill-building, we include errors encountered during the development of the material and illustrate their solutions. The code examples contained within these tutorials highlight popular, robust, and well-maintained open-source tools and software, with a strong focus on the Python module Xarray, which is designed for working with n-dimensional array objects and well-suited to geospatial applications (Hoyer & Hamman, 2017). In addition, we use technology such as Jupyter Books, Jupyter Notebooks, and GitHub to make these tutorials accessible, participatory, and flexible (Executable Books Community, 2021; Kluyver et al., 2016).

Experience of use in teaching and learning situations

We aim for these resources to benefit learners in a range of settings. Anecdotally, we have been contacted by a number of professors and supervisors who use the tutorials when working with students. The ITS_LIVE tutorial has been successfully used as a lab exercise for students in an undergraduate course on quantitative methods in physical geography at the University of Utah. The Sentinel-1 RTC tutorial was used in an active remote sensing course at the University of Utah. At the same time, our goal is to reduce barriers to engaging in the scientific process: the emphasis on explanatory text and additional learning resources makes the tutorials accessible to learners outside of the traditional academic setting with the in-person guidance of an instructor. This includes imparting the idea that it is the responsibility of the data user to understand the nuances and limitations of different datasets. The modular nature of the tutorials allows users to identify specific areas that may be useful to them. Taken in full, the tutorials guide a novice user through each step required for developing a scientific workflow, from data access to exploratory analysis. We intentionally do not prescribe specific conclusions, instead focusing on accessibility and giving users the tools to guide their own exploration and analysis of complex and exciting datasets.

A common sentiment of users learning to work with n-dimensional array data, particularly in earth observation applications, is difficulty experienced in the 'data organizing' steps of preparing a dataset for analysis (usually, coercing observations segmented in time and/or space into a data cube with x,y, and time dimensions) (Marshall et al., 2023). The submitted tutorials place significant focus on these steps, explaining and demonstrating operations required to organize data in a way that facilitates subsequent analysis.

Story of the Project

The tutorials were initially developed while Emma Marshall interned with the Summer Internships in Parallel Computational Sciences (SIParCS) program at the National Center for Atmospheric Research (NCAR). Jessica Scheick, Scott Henderson, and Deepak Cherian were internship supervisors for this project. The internship was also supported by a NASA

Acknowledgments

References

Abernathey, R. P., Augspurger, T., Banihirwe, A., Blackmon-Luca, C. C., Crone, T. J., Gentemann, C. L., Hamman, J. J., Henderson, N., Lepore, C., McCaie, T. A., Robinson, N. H., & Signell, R. P. (2021). Cloud-native repositories for big scientific data. *Computing in Science & Engineering*, *23*(2), 26–35. https://doi.org/10.1109/MCSE.2021.3059437

Executable Books Community. (2021). *Jupyter book*. Zenodo. https://doi.org/10.5281/zenodo.4539666

Gentemann, C. L., Holdgraf, C., Abernathey, R., Crichton, D., Colliander, J., Kearns, E. J., Panda, Y., & Signell, R. P. (2021). Science Storms the Cloud. *AGU Advances*, *2*(2), e2020AV000354. https://doi.org/10.1029/2020AV000354

Guo, H. (2017). Big Earth data: A new frontier in Earth and information sciences. *Big Earth Data*, *1*(1-2), 4–20. https://doi.org/10.1080/20964471.2017.1403062

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

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). *Jupyter notebooks – a publishing format for reproducible computational workflows*. 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

Marshall, E., Cherian, D., & Henderson, S. (2023). *Tidy geospatial data cubes*. Zenodo. https://doi.org/10.25080/gerudo-f2bc6f59-034

Martin Sudmanns, S. L., Dirk Tiede, & Blaschke, T. (2020). Big Earth data: Disruptive changes in Earth observation data management and analysis? *International Journal of Digital Earth*, *13*(7), 832–850. https://doi.org/10.1080/17538947.2019.1585976

Mathieu, P.-P., Borgeaud, M., Desnos, Y.-L., Rast, M., Brockmann, C., See, L., Kapur, R., Mahecha, M., Benz, U., & Fritz, S. (2017). The ESA's Earth Observation Open Science Program [Space Agencies]. *IEEE Geoscience and Remote Sensing Magazine*, *5*(2), 86–96. https://doi.org/10.1109/MGRS.2017.2688704

Palumbo, I., Rose, R. A., Headley, R. M. K., Nackoney, J., Vodacek, A., & Wegmann,

M. (2017). Building capacity in remote sensing for conservation: Present and future challenges. *Remote Sensing in Ecology and Conservation*, *3*(1), 21–29. https://doi.org/10.1002/rse2.31

Procida, D. (n.d.). *Diátaxis documentation framework*. https://doi.org/https://diataxis.fr/

Radočaj, D., Obhođaš, J., Jurišić, M., & Gašparović, M. (2020). Global Open Data Remote Sensing Satellite Missions for Land Monitoring and Conservation: A Review. *Land*, *9*(11), 402. https://doi.org/10.3390/land9110402

Ramachandran, R., Bugbee, K., & Murphy, K. (2021). From Open Data to Open Science. *Earth and Space Science*, *8*(5). https://doi.org/10.1029/2020EA001562

Wagemann, J., Fierli, F., Mantovani, S., Siemen, S., Seeger, B., & Bendix, J. (2022). Five Guiding Principles to Make Jupyter Notebooks Fit for Earth Observation Data Education. *Remote Sensing*, *14*(14), 3359. https://doi.org/10.3390/rs14143359

Wagemann, J., Siemen, S., Seeger, B., & Bendix, J. (2021). A user perspective on future cloud-based services for Big Earth data. *International Journal of Digital Earth*, *14*(12), 1758–1774. https://doi.org/10.1080/17538947.2021.1982031