JOSE
The Journal of Open Source Education

Pythia Foundations: A community learning resource for Python-based computing in the geosciences

Brian E. J. Rose [1], **Robert R. Ford** [1], **Anderson Banihirwe** [2], **M. Drew Camron** [3], **John Clyne** [4], **Orhan Eroglu** [4], **Katelyn FitzGerald** [4], **Maxwell A. Grover** [5], **Julia Kent** [4], **Ryan May** [3], **Kevin Paul** [6], **Kevin R. Tyle** [1], **and Anissa Zacharias** [4]

1 Department of Atmospheric and Environmental Sciences, University at Albany (State University of New York) **2** CarbonPlan **3** NSF Unidata, University Corporation for Atmospheric Research **4** Computational Information Systems Lab., NSF National Center for Atmospheric Research **5** Environmental Science Division, Argonne National Laboratory **6** NVIDIA Corporation

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Submitted: 25 June 2025
Published: unpublished

Summary

Pythia Foundation (Rose et al., 2025) is the flagship product of the first phase of Project Pythia (Rose et al., 2023), a broad community effort to build, house, share, and maintain high-quality learning resources for Python-based computing in the geosciences. Project Pythia's central mission is to accelerate progress across the geosciences by reducing roadblocks to sharing technical knowledge, particularly related to scalable and reproducible data analysis in the cloud using the open-source Python software ecosystem.

Pythia Foundations is a geoscience-flavored introduction to essential tools in the scientific Python ecosystem and Pangeo (Abernathey et al., 2017) stack (e.g., JupyterLab, NumPy (Harris et al., 2020), Matplotlib (Hunter, 2007), Pandas (McKinney, 2010), (Pandas development team, 2024), Cartopy (Elson et al., 2024), Xarray (Hoyer & Hamman, 2017), Dask (Dask Development Team, 2016)), plus environment management tools (conda), basics of version control (git), and effective use of GitHub as an technical communication platform (Figure 1). It is a community-owned executable textbook backed by computational resources for automated health-checking and interactive use. It covers the foundational knowledge that is needed to get started with Python in the computational geosciences, as well as to become an effective citizen-practitioner in key open geoscience software ecosystems. It is intended for anyone from undergraduate students through established geoscientists who are relatively new to working in Python. The book assumes a basic knowledge of programming concepts, but a brief "Quickstart" lesson highlights distinctive features of Python for users migrating from other languages.

A distinguishing feature of Pythia Foundations is its rigorous quality control and maintenance. All Python code and external web links are tested nightly, and book contents are kept up to date as the software ecosystem and data sources evolve. Users can run the examples with a "one click" launch into a dedicated cloud-based Binder service (Project Jupyter et al., 2018).

JOSE
The Journal of Open Source Education

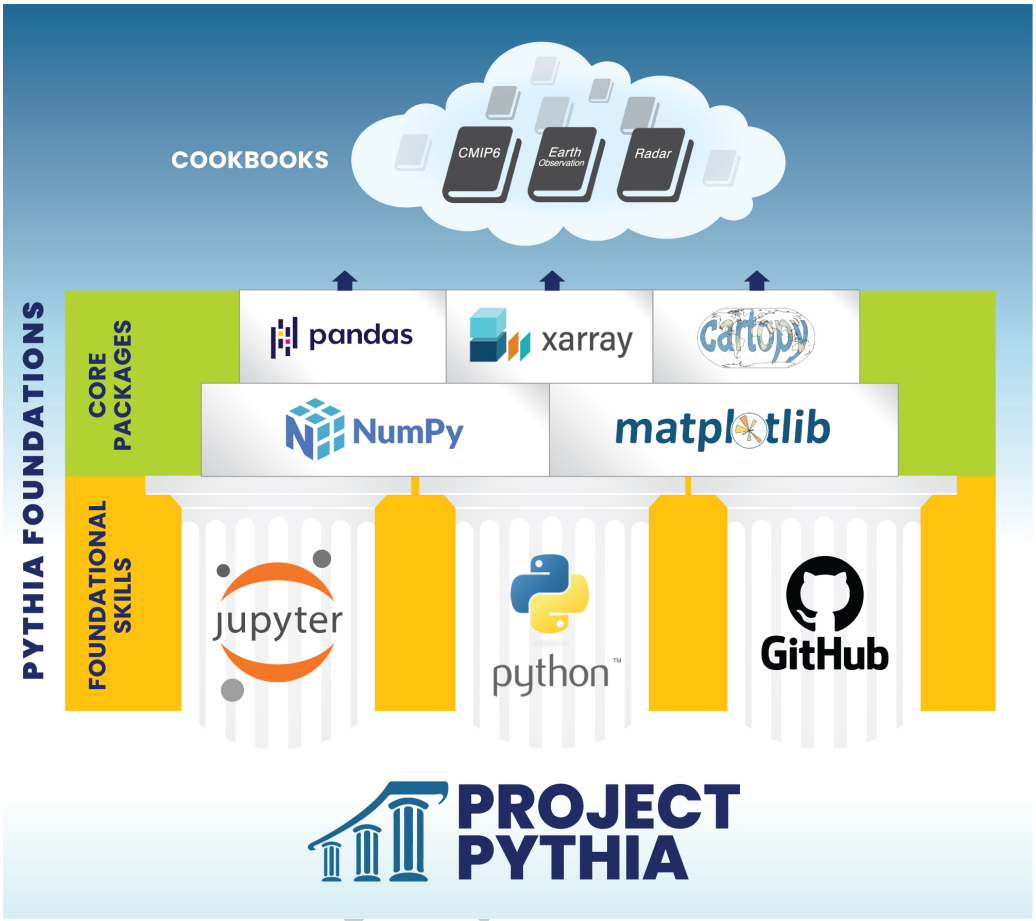

Figure 1: Figure 1: Pythia Foundations infographic

Figure 1: A schematic of the content and organization of Pythia Foundations. The book is divided into two major sections, "Foundational Skills" and "Core Scientific Python Packages", limited to those currently in broad use across multiple geoscience disciplines. Pythia Cookbooks house more advanced or domain-specific content that reference individual chapters from Foundations as prerequisites.

Statement of Need

Today's geoscientists require not only domain expertise but also proficiency with specialized software and high-level technical skills to effectively analyze, manipulate, and manage potentially vast volumes of digital data in a complex and ever-changing computing environment. The scientific Python ecosystem and the emergence of cloud computing have been game-changers for many, providing an abundance of open-source tools with wide ranging functionality. Ironically, however, this abundance is often untapped, and can be a source of great frustration. Scientists spend an inordinate amount of time pondering questions such as: Which tool or technology should I use? How do I use it? Can I trust it? Is it compatible with other tools in my workflow? Often, the answers are unclear, due to inadequate documentation or difficulty in finding relevant up-to-date working examples. The result is too much time spent navigating or avoiding technology—time that could have been spent productively doing science. Pythia Foundations fills this need by providing a trusted community-owned, web-accessible, geoscience-specific education and training resource for scientists and students at all career stages who want to know what tools to

JOSE
The Journal of Open Source Education

use and how to use them to explore their data.

The Foundations book embodies the FAIR principles (Wilkinson et al., 2016) that play a central role in open science. Findability is served by gathering geoscience-specific tutorials into a high-visibility community archive. Accessibility is served by our automated CI testing and integrated public binder. Tutorials and example code are largely Interoperable due to reliance on a common ecosystem of tools (e.g., NumPy and Xarray). Reusability is addressed through permissive licensing of book content and geoscience relevance of the examples, as well as our commitment to maintaining up-to-date working examples—an essential need in light of the widespread problem of rapid obsolescence of computational notebooks (Pimentel et al., 2019).

Content, instructional design, and usage

The scope of Pythia Foundations is limited to tools and packages that are currently in broad use across multiple geoscience disciplines; packages tailored to more narrow scientific domains are not covered in Foundations but may be suitable for a Cookbook. The book outline was designed collaboratively by the core author team, informed by community feedback, and drawing on our substantial collective experience in teaching Python-based scientific workflows in classrooms, workshops, and outreach events.

The book is organized into two main sections: Foundational skills and Core Scientific Python packages (Figure 1). The foundational skills section covers "getting started" skills such as how to install Python and manage environments and how to run Python code in JupyterLab. There is also a set of tutorials on the use of GitHub and git for version control and collaboration on open source projects. The scope of this section was chosen with the specific goal of enabling users to contribute back to Pythia Foundations.

A template notebook and contribution guide is provided for new content, encouraging consistency of style and organization. Each chapter includes explicit prerequisites, references, and estimated learning time. The book is intended primarily for self-study and reference, backed by the interactive Binder or deployed on user machines following the detailed guidance in the book. From web-based metrics, Pythia Foundations served roughly 29,000 users in 111 countries during calendar year 2024.

Subsets of the book contents have been modified and repackaged for various workshops and short courses. A few examples include the 2022 EarthCube-AMGeO Hackathon, the ERAD 2024 Open Radar Science Shortcourse (Ladino et al., 2024), the Climatematch Academy international virtual summer school (annually since 2023), and in Spanish-language translation for a Colombian hydrometeorological workshop in 2023 (Ladino et al., 2023). Co-authors Rose and Tyle have integrated material from Foundations into the formal curriculum for several semester-length undergraduate and graduate level courses at the University at Albany.

Computational infrastructure

The book is deployed as an easy-to-navigate website using JupyterBook (Executable Books Community, 2020) and MyST-MD (Cockett et al., 2025), including "one-click" Binder links to interactive versions of every chapter. It features complete reproducibility: source materials are stored in a GitHub repository as unexecuted Jupyter notebooks, and all content is recreated in a bespoke computational environment during nightly builds and whenever the book pages are re-rendered. A full preview of the executed and rendered book is created whenever a change is proposed via a Pull Request. Development of the novel notebook publishing infrastructure enabling this full reproducibility was driven by the Pythia team's need to collaborate on a large computational document. The build-

and-preview automation that our team developed while authoring Foundations is now in wide use by the community of Cookbook creators. The automation notably includes the ability to route notebook execution through the same Binder environment offered to users, guaranteeing that the output of the automated builds are identical to those that users see when running code examples interactively.

Future plans

Pythia Foundations is a living document and is receiving continuous updates (Cockett et al., 2024) and improvements, both from the core author team and the broader community of user-contributors. On the content side, Project Pythia is simultaneously fostering a growing collection of more advanced and domain-specific tutorials in our crowd-sourced community Cookbook gallery, with explicit links to prerequisites from Foundations. We anticipate periodic reviews of the Cookbook collection to identify cross-cutting content that should be abstracted back to Foundations, e.g., common data access patterns or analysis workflows.

The computational and publishing infrastructure for Foundations is also continuously evolving. As of this writing, Foundations and all other Pythia content has just undergone a significant refresh and upgrade with the migration to JupyterBook 2 which is based on the MyST-MD publishing engine (Cockett et al., 2025). Among the compelling new functionality unlocked by this transition is a rich content cross-referencing and embedding model that will enable more modular reuse and repacking of Foundations content tailored to specific courses or audiences.

Acknowledgements

The authors gratefully acknowledge support from the broad open geoscience communities of Project Pythia and Pangeo for their feedback, suggestions, pull requests, and enthusiasm. Development and maintenance of Pythia Foundations was supported by the U.S. National Science Foundation (NSF) awards 2026899, 2026863, 2324302, 2324303 and 2324304. The Pythia BinderHub is deployed on Jetstream2 (Hancock et al., 2021) at Indiana University through allocations EES230007 and SEE240014 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program (Boerner et al., 2023), which is supported by NSF grants 2138259, 2138286, 2138307, 2137603, and 2138296.

References

Abernathey, R., Paul, K., Hamman, J., Rocklin, M., Lepore, C., Tippett, M., Henderson, N., Seager, R., May, R., & Del Vento, D. (2017). *Pangeo NSF Earthcube Proposal*. https://doi.org/10.6084/m9.figshare.5361094.v1

Boerner, T. J., Deems, S., Furlani, T. R., Knuth, S. L., & Towns, J. (2023). ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support. *Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good*, 173–176. https://doi.org/10.1145/3569951.3597559

Cockett, R., Koch, F., Purves, S., Hollands, A., Wang, Y., Grandmont, D., Holdgraf, C., Andrea, Müller, J.-H., Lyon, S., Le, C., Madge, J., wwx, Reden, S., Geng, Y., Lovett, R., Roald-Arbøl, M., McKay, M., Brett, M., … Driscoll, T. (2025). *Mystmd v1.4.0*. Zenodo. https://doi.org/10.5281/ZENODO.14805610

Cockett, R., Purves, S., Koch, F., & Morrison, M. (2024). Continuous Tools for Scientific Publishing. *Scipy*, 121–136. https://doi.org/10.25080/NKVC9349

Dask Development Team. (2016). *Dask: Library for dynamic task scheduling*. https://dask.org

Elson, P., Andrade, E. S. de, Lucas, G., May, R., Hattersley, R., Campbell, E., Comer, R., Dawson, A., Little, B., Raynaud, S., scmc72, Snow, A. D., lgolston, Blay, B., Killick, P., lbdreyer, Peglar, P., Wilson, N., Andrew, … Kirkham, D. (2024). *Cartopy v0.24.1*. Zenodo. https://doi.org/10.5281/ZENODO.1182735

Executable Books Community. (2020). *Jupyter Book*. Zenodo. https://doi.org/10.5281/ZENODO.2561065

Hancock, D. Y., Fischer, J., Lowe, J. M., Snapp-Childs, W., Pierce, M., Marru, S., Coulter, J. E., Vaughn, M., Beck, B., Merchant, N., Skidmore, E., & Jacobs, G. (2021). Jetstream2: Accelerating cloud computing via Jetstream. *Practice and Experience in Advanced Research Computing 2021: Evolution Across All Dimensions*, 1–8. https://doi.org/10.1145/3437359.3465565

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hoyer, S., & Hamman, J. (2017). Xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software*, *5*(1). 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

Ladino, A., Moral Méndez, A. del, Javornik, B., Michelson, D., Wolfensberger, D., Ghiggi, G., DeHart, J., Figueras i Ventura, J., Giles, J., Mühlbauer, K., Grover, M., Dixon, M., Jackson, R., Collis, S., Cha, T.-Y., & contributors, E. O. R. S. S. (2024). *ERAD 2024 Open Radar Science Shortcourse*. Zenodo. https://doi.org/10.5281/zenodo.13694511

Ladino, A., Rivera, N., & Grover, M. (2023). *Aladinor/Atmoscol2023: Steable version 0.1.0-Beta*. Zenodo. https://doi.org/10.5281/ZENODO.8316796

McKinney, Wes. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Pandas development team. (2024). *Pandas-dev/pandas: Pandas*. Zenodo. https://doi.org/10.5281/ZENODO.3509134

Pimentel, J. F., Murta, L., Braganholo, V., & Freire, J. (2019). A Large-Scale Study About Quality and Reproducibility of Jupyter Notebooks. *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, 507–517. https://doi.org/10.1109/MSR.2019.00077

Project Jupyter, Bussonnier, Matthias, Forde, Jessica, Freeman, Jeremy, Granger, Brian, Head, Tim, Holdgraf, Chris, Kelley, Kyle, Nalvarte, Gladys, Osheroff, Andrew, Pacer, M., Panda, Yuvi, Perez, Fernando, Ragan-Kelley, Benjamin, & Willing, Carol. (2018). Binder 2.0 - Reproducible, interactive, sharable environments for science at scale. In Fatih Akici, David Lippa, Dillon Niederhut, & M. Pacer (Eds.), *Proceedings of the 17th Python in Science Conference* (pp. 113–120). https://doi.org/10.25080/Majora-4af1f417-011

Rose, B. E. J., Clyne, J., May, R., Munroe, J., Snyder, A., Eroglu, O., & Tyle, K. (2023). *Collaborative Research: GEO OSE TRACK 2: Project Pythia and Pangeo: Building an inclusive geoscience community through accessible, reusable, and reproducible workflows*. https://doi.org/10.5281/zenodo.8184298

Rose, B. E. J., Kent, J., Tyle, K., Clyne, J., Banihirwe, A., Camron, D., May, R., Grover, M., Ford, R. R., Paul, K., Morley, J., Eroglu, O., Kailyn, L., & Zacharias, A. (2025). *Pythia Foundations*. Zenodo. https://doi.org/10.5281/zenodo.14759276

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18