


# Guiding principles and implementation strategies for reproducible geoscientific data analysis

AUTHOR

Matthias Schlögl  





AFFILIATIONS

BOKU University

GeoSphere Austria

## Prioritization

Prioritization is essential to focus on the most impactful and achievable tasks in this checklist. We recommend applying the MoSCoW method<sup>1</sup>, which is widely used in product development and requirements engineering, to categorize actions by priority and feasibility:

-  **Must** have (M): Essential actions that are critical for achieving reproducibility and cannot be omitted.
-  **Should** have (S): Important actions that significantly enhance reproducibility but are not critical.
-  **Could** have (C): Desirable actions that are beneficial but less impactful.
-  **Won't** have [*this time*] (W): Actions that are not a priority, either because they are not applicable to the specific use case or are too complex to implement within the current scope.

In the context of geoscientific data analysis, the ease of implementation and priorities of actions may vary depending on use cases, research goals, team composition, and available resources. The MoSCoW method provides a systematic framework to assess and prioritize checklist actions and tailor them to specific project requirements. This allows to focus on essential tasks while remaining flexible to integrate additional improvements as resources allow.















### Tip

We offer recommendations based on our experience to serve as a starting point for less experienced researchers in prioritizing actions (as indicated in the 'Priority' column). However, users are encouraged to adapt these priorities to align with their specific project requirements, alternative priority assignments, or greater expertise. The 'Eval' column provides space for users to record their own evaluation of an action's importance and to mark the checkbox once the action has been addressed.

# Actions

## Overarching principles & planning

Consider this when initiating your research project, as these principles lay the foundation for a well-organized and transparent workflow. Establishing a clear folder structure, adopting open standards, and developing a data management plan (DMP) should be prioritized during the planning phase to ensure systematic organization and long-term accessibility. Documenting provenance and isolating irreproducible components should be ongoing throughout the project to maintain transparency. Automating repetitive tasks and adhering to tidy data principles will streamline processes and improve efficiency during data preparation and analysis.

Action	Outcome	Priority	Eval
 Use a clear folder structure and naming conventions.	Achieving systematic organization of project components for easier navigation and collaboration.	 S	<input type="checkbox"/>
 Adopt open standards and technologies.	Ensuring accessibility, transparency, and compatibility of data, tools, and publications.	 M	<input type="checkbox"/>
 Develop (and implement) an actionable data management plan (DMP).	Establishing a structured approach to managing data and code, including storage, security, sharing, and quality assurance.	 S	<input type="checkbox"/>
 Document provenance.	Enhancing the quality, reliability and trustworthiness of analysis artifacts by recording processing steps and computational details.	 M	<input type="checkbox"/>
 Isolate and document irreproducible workflow components.	Providing clarity on inherently irreproducible elements to enhance transparency.	 M	<input type="checkbox"/>
 Automate repetitive tasks.	Minimizing human error, saving time, and creating implicit documentation of workflow steps.	 S	<input type="checkbox"/>
 Adhere to tidy data principles.	Facilitating consistent and efficient data organization for analysis and sharing.	 S	<input type="checkbox"/>











## Data collection & governance

Consider this during data acquisition and preparation. Identify and describe data sources, limitations, and metadata early. Assessing data quality and using common open data formats with a defined spatial reference systems should be integral to data preprocessing, and will facilitate downstream analyses and interoperability. Establishing semantic clarity by adopting standardized terminologies from the outset is particularly beneficial in interdisciplinary teams. Consider the spatial or spatiotemporal characteristics of the data when designing models. Templates and checklists can be used throughout to document datasets and workflows systematically.

	Action	Outcome	Priority	Eval
🔍	Identify and describe your data.	Providing a comprehensive understanding of data sources, metadata, and limitations for informed analysis.	🔴 M	<input type="checkbox"/>
🌐	Consider the spatial / spatiotemporal nature of data.	Improving modeling accuracy and transferability by addressing spatial and spatiotemporal autocorrelation and using appropriate validation techniques.	🔴 M	<input type="checkbox"/>
✅	Assess data quality.	Ensuring reliable analysis by evaluating data accuracy, completeness, and consistency.	🔴 M	<input type="checkbox"/>
📄	Use common open data formats.	Enhance data interoperability and long-term usability with standardized formats like GeoTIFF, NetCDF or GeoPackage.	🔴 M	<input type="checkbox"/>
🗂️	Adopt taxonomies, controlled vocabularies, and ontologies.	Promoting semantic clarity and interoperability through standardized terminology.	🟡 S	<input type="checkbox"/>
📍	Use established location reference systems.	Ensuring unambiguous spatial localization by specifying CRS with standards like EPSG codes or WKT-CRS strings.	🔴 M	<input type="checkbox"/>
📋	Use templates and checklists.	Streamlining systematic documentation of models, datasets, and workflows for consistency and completeness.	🔵 C	<input type="checkbox"/>











# Analysis design & automation

Consider this when designing your analytical workflows and implementing automation. Tracking model development and documenting workflow rationale are essential for ensuring transparency and reproducibility during the modeling and analysis phases. Using script-based workflows and exploring literate programming tools will help create self-documenting and accessible analyses. For more complex workflows, dataflow programming and pipeline tools can be explored to enable scalability and modularity, particularly when handling large datasets or multi-step processes.

Action		Outcome	Priority	Eval
	Track model development.	Maintaining a detailed record of model lifecycles, hyperparameters, and metrics.	 S	<input type="checkbox"/>
	Document provenance and workflow rationale.	Providing transparency and justification for workflow design and data processing steps.	 S	<input type="checkbox"/>
	Use script-based workflows.	Ensuring consistency, repeatability, and self-documentation of analytical processes.	 S	<input type="checkbox"/>
	Explore literate programming tools.	Providing clear and accessible technical documentation of analytical steps and workflows with tools like Quarto, R Markdown or Jupyter.	 C	<input type="checkbox"/>
	Explore dataflow programming and pipeline tools.	Enabling scalable and modular workflows with tools like Snakemake or Apache Airflow.	 C	<input type="checkbox"/>






## Code & scientific computing

Consider this during the development and implementation of computational workflows. Using free and open-source software and defined environments ensures consistency and accessibility from the start. Leveraging collaborative development platforms and adhering to a consistent coding style will facilitate teamwork and maintainability. Following the DRY principle and considering the code life cycle will simplify maintenance and enhance long-term usability. Providing logging information and performing automatic testing should be integrated into the development process to ensure reliability and traceability. Version control systems (VCS) and data versioning should be adopted to manage changes effectively throughout the project.

Action	Outcome	Priority	Eval
 Use free and open-source software.	Promoting transparency, accessibility, and community-driven development.	● M	<input type="checkbox"/>
 Use defined environments.	Ensuring consistency by specifying details on operating systems, dependencies and configurations using tools like uv, renv or Docker.	● M	<input type="checkbox"/>
 Leverage collaborative development platforms.	Facilitating version control, issue tracking, and collaborative development with tools like GitHub or GitLab.	● S	<input type="checkbox"/>
 Adhere to a consistent coding style.	Improving code readability and maintainability with style guides and linters.	● S	<input type="checkbox"/>
 Don't repeat yourself (DRY).	Simplifying maintenance and reducing errors by modularizing and reusing code.	● S	<input type="checkbox"/>
 Consider code life cycle and maintenance.	Supporting long-term usability by documenting, testing, and updating code.	○ C	<input type="checkbox"/>
 Provide logging information.	Facilitating debugging and documentation by recording system details, parameters, and execution steps.	● M	<input type="checkbox"/>
 Perform automatic testing.	Verifying code correctness and functionality through automated tests.	○ C	<input type="checkbox"/>
 Adopt Version Control Systems (VCS).	Tracking and managing changes in code and data with tools like Git.	● S	<input type="checkbox"/>
 Version-control data.	Managing changes in large datasets effectively with tools like DVC or Git LFS.	○ C	<input type="checkbox"/>

## Publishing and reuse of results

Consider this during the final stages of your research, as you prepare to share your findings. Publishing all data, code, and results under appropriate licenses ensures that your work is accessible and reusable. Using machine-readable, well-documented, and stable formats will enhance the usability and longevity of your outputs. Depositing artifacts in long-preserving repositories guarantees persistent access, while adhering to data stewardship and management best practices ensures compliance with FAIR principles, promoting findability, accessibility, interoperability, and reusability.

	Action	Outcome	Priority	Eval
	Publish all data, code, and results.	Enabling precise replication and reuse of research by sharing all artifacts.	● M	<input type="checkbox"/>
	Publish under a permissive or copyleft license.	Facilitating reuse, modification, and redistribution of results with licenses like MIT or GNU GPL.	● M	<input type="checkbox"/>
	Publish in a machine-readable, well-documented, common, and stable format.	Ensuring long-term usability and interoperability of research outputs.	● M	<input type="checkbox"/>
	Publish in long-preserving repositories.	Guaranteeing persistent access to artifacts by depositing them in repositories like Zenodo or PANGAEA.	● S	<input type="checkbox"/>
	Adhere to data stewardship & management best practices.	Ensuring data is findable, accessible, interoperable, and reusable (FAIR principles).	● S	<input type="checkbox"/>

### Footnotes

1. Agile Business Consortium (2014): [The DSDM Agile Project Framework, Chapter 10: MoSCoW Prioritisation](#) [accessed: 15 August 2025] 