Cloud-Based Seismic & Geodetic Data Workshop – Participant Handout

Duration: 4 hours
Audience: Early-career researchers and professionals in geophysics, seismology, geodesy, and Earth data science.

Learning Objectives

- Access, preprocess, and analyze cloud-based seismic (MiniSEED) and geodetic (GNSS) datasets.
- Apply data cleaning techniques including detrending, filtering, and outlier removal.
- Perform machine learning classification and regression tasks on prepared datasets.
- Understand cloud workflows and reproducible research practices.
- Develop scalable and collaborative data workflows using cloud tools.

Part 1 – Introduction & Context

Overview of seismic and GNSS data sources, data formats (MiniSEED, RINEX), and cloud data access methods (fsspec, AWS S3, Google Cloud Storage). Exercise: Load and visualize synthetic or real seismic and GNSS data.

Part 2 – Data Access & Preparation

Hands-on data preprocessing using Python. Participants perform detrending, filtering (Butterworth bandpass), and outlier detection (Hampel filter) on both seismic and GNSS data.

Part 3 – Cloud-Based Analysis & Machine Learning

Application of machine learning methods: seismic event classification using Random Forests and GNSS-based atmospheric estimation via regression models.

Part 4 – Reproducibility & Collaboration

Documenting workflows and ensuring reproducibility. Saving configurations and outputs, maintaining FAIR data principles, and packaging analyses for future re-use.

Exercises Overview

- Exercise 1: Load and visualize seismic and GNSS data. Discuss sampling, noise, and patterns.
- Exercise 2: Implement preprocessing functions (bandpass filter, Hampel filter) and save cleaned outputs.
- Exercise 3A: Build a Random Forest classifier for seismic event detection using windowed features.
- Exercise 3B (Optional): Fit a regression model to estimate atmospheric proxy data from GNSS time series.

- Exercise 4: Save configuration and data artifacts for reproducible research.

Cloud Tools and Resources

- Cloud Storage: AWS S3, Google Cloud Storage, Azure Blob.
- Data Libraries: ObsPy (seismic), pyproj (geodesy), pandas, NumPy, SciPy.
- Machine Learning: scikit-learn, TensorFlow, PyTorch.
- Visualization: matplotlib, PyGMT.
- Workflow Reproducibility: Docker, Jupyter, GitHub Actions, Makefiles.

Best Practices for Early Career Researchers

- Document all preprocessing and modeling parameters.
- Use version control (Git) and share notebooks with reproducible environments.
- Apply FAIR principles when sharing data: Findable, Accessible, Interoperable, Reusable.
- Collaborate across disciplines—combine geophysics, computer science, and atmospheric science.
- Seek mentorship and use community datasets and open-source resources.