# SCOPED: Seismic COmputational Platform for Empowering Discovery [Year 2]

NSF Cyberinfrastructure for Sustained Scientific Innovation (CSSI) Frameworks Project


**PI TEAM:**

Ian Wang (iwang@tacc.utexas.edu): high-performance seismology, filesystem, database
Ebru Bozdag (bozdag@mines.edu): full waveform modeling and tomography
Marine Denolle (mdenolle@uw.edu): ambient noise seismology and cloud computing
Felix Waldhauser (felixw@ldeo.columbia.edu): earthquake source precisions and catalogs
Carl Tape (ctape@alaska.edu) [PI]: full waveform modeling, earthquake sources and uncertainties

[TAAC/UT: OAC-CSSI 2103494]
[MINES: OAC-CSSI 2103621]
[UW: OAC-CSSI 2103701]
[Lamont/Columbia: OAC-CSSI 2103741]
[UAF: OAC-CSSI 2104052]


project website:
https://seisscoped.org/

Lamont-Doherty Earth Observatory
COLUMBIA UNIVERSITY | EARTH INSTITUTE

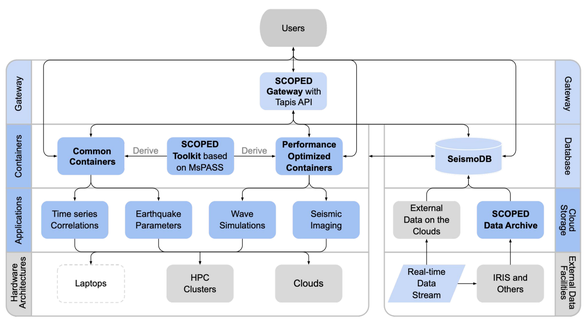
COLORADOSCHOOLOFMINES
EARTH ENERGY ENVIRONMENT

jupyterhub

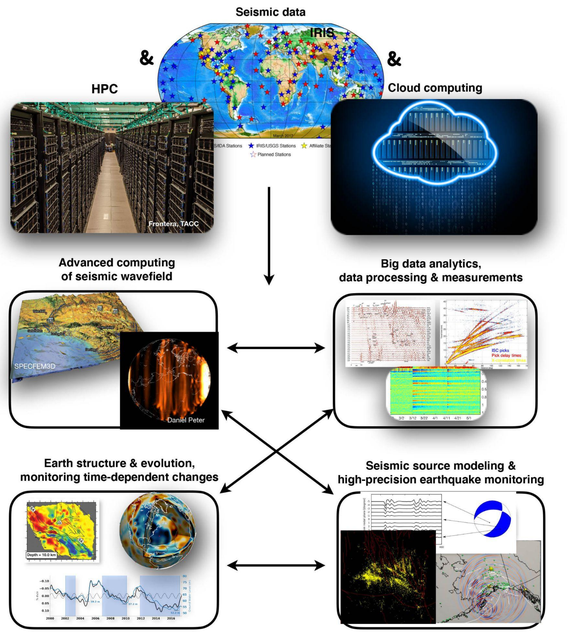
TACC

## Overview

Seismology is a data and model-intensive field. Data-driven computations benefit from Cloud Computing because observational seismology relies on horizontal scalability. Model-driven computation benefits from High-Performance Computing because simulations necessitate large memory and a high number of workers for efficient parallelization. Our project SCOPED (Seismic COmputational Platform for Empowering Discovery) aims to bridge both observational and theoretical fields by building a CyberInfrastructure (software, data) as a service to the seismic community. The challenges ahead are the benchmarking of open-source codes in cloud and HPC environments, the containerization of these codes adapted to various computing architectures, and the choice of data formats adapted to both Cloud and HPC computing, among others.




A schematic diagram of the proposed SCOPED Cyberinfrastructure that combines computation (left) and data (right). The SCOPED deliverables will be either fully functional (dark blue) or prototyped (light blue). The external components that SCOPED will interact with (gray) are either computing facilities, data archives, and users.



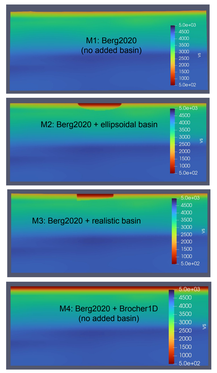

Representation of the unified and interconnected SCOPED Cyberinfrastructure described in the science components and cyberinsfrastructure. At the top, the global map of seismic stations represents one of several featured sources of data sets. Data are hosted and processed on HPC and on Cloud clusters, which overly pillars of wavefield simulations (left column) and data processing and analysis (right column). Cross-disciplinary research will facilitate the investigations of outstanding questions about Earth and fault dynamics

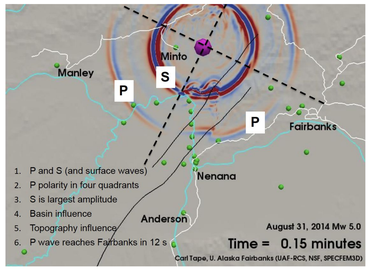
## Advancing science with SCOPED codes

SCOPED focuses on four seismological applications enabled by several open-source software packages. We seek to improve these packages, while also containerizing them for easier use by others.

### Seismic wavefield modeling with spectral-element method simulations

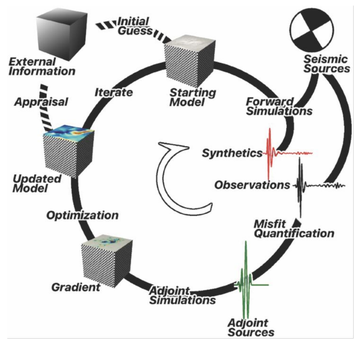
Wavefield simulations within 3D Earth models provide synthetic seismograms that can be compared with recorded seismograms, either to better understand earthquake sources or to improve the subsurface characterization of Earth's structure.




**Above**. Snapshot from a seismic wavefield simulation, showing an S wavefront entering into a deep sedimentary basin. **Left**. Cross sections of four 3D models used by Yuan Tian to investigate the influence of basin structures on the seismic wavefield.

### Seismic imaging using recorded seismograms and simulated seismograms



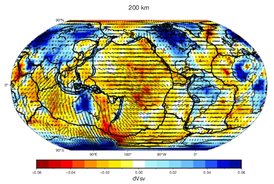

Iterative workflow for seismic imaging using 3D wavefield simulations and adjoint methods (Chow et al., 2020). This workflow is automated with the help of Pyatoa software.

The seismic imaging workflow represents an optimization problem whereby the misfit between recorded and simulated seismograms is minimized while iterating toward a more accurate representation of the subsurface Earth properties (Vp, Vs, density).

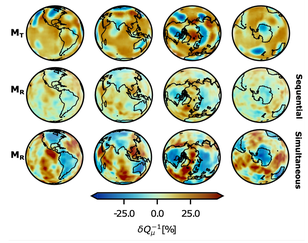
### Seismic imaging of the whole Earth

Seismic imaging of the entire Earth is computationally challenging, as it requires thousands of simulations within a large domain, as well as massive storage of temporary volumetric fields. Furthermore, the physics of Earth's interior must be addressed properly in the inverse problem as well as global data coverage problem.

We participated in ***Texascale days on TACC's Frontera system*** and successfully performed two iterations scaling up our simulations on ***~8000 nodes*** with our current resolution. The main takeaways are: 1) GPU computing is necessary to further increase number of earthquakes or the resolution of simulations. 2) ***ADIOS library*** and compression help reduce the ***IO*** challenges.




Azimuthally anisotropic global adjoint model GLAD-AZI-M50 (Bozdag, Örsvuran et al. in prep.) constructed on TACC's Frontera system.




Towards anelastic global FWI. We performed 3D global tests on TACC's Frontera system to explore the parameter trade-offs between wavespeeds and attenuation (Carmona et al. in revision for Geophys. J. Int.).

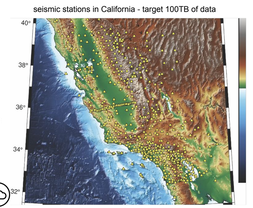
### Ambient noise seismology

The surface of the Earth is in a constant state of low-amplitude ground motion that can be used for monitoring and for seismic imaging.

***We delivered a cloud-optimized solution for for Distributed Acoustic Sensing storage and processing that uses TileDB or Zarr. (Ni et al, in press in SRL)***

**HPC - Filesystems** for large-scale ambient noise cross correlations tested on Frontera (pers. comm. Kurama Okubo, Frontera). Use large hierarchical formats (H5) for data output, maximize local SSD or RAM for intermediate data products. Deploy over 500+ cores. High throughput on scratch filesystems. Parallelization in time (e.g., days of data) using distributed memory and station and station pairs using shared memory on nodes. We use SLURM to manage jobs on cluster.

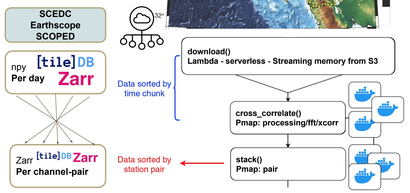
**Cloud - Object Storage:** When raw data is stored on object storage (e.g., S3). Optimal when using millions of small objects (e.g., mseed/npy day files < 1MB) due to the writing and parallelization needs. We use AWS Batch to manage jobs on instances. Use S3-optimized instances for improved throupughts.



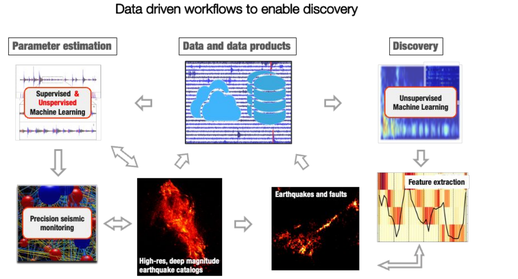

SCOPED is supporting the refactoring of Noisepy, an ambient noise seismology python package. The parallelization now uses distributed and shared memory. The back-end data formats are flexible and user-defined, we tested on H5, Numpy-tar NPZ, and cloud optimized data formats such as TileDB and Zarr.

Noisepy will be delivered to the community in fall 2023 with extensive testing and contributing guidelines. It has extensive continuous integration and improved code coverage, while preserving the original work of the package from Jiang and Denolle (2020).

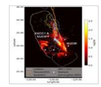
We will demonstrate this on the S3-hosted SCEDC data set. We have demonstrated such workflow on AWS for Distributed Acoustic Sensing.



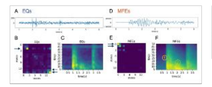

### Precision seismology: large-scale, high-resolution catalog production and analysis



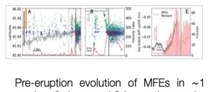

**Discovery of precursory mixed-frequency earthquakes** (MFE) prior and leading up to 2015 eruption of Axial volcano



***Supervised ML detection, cross-correlation timing***, and ***double-difference location*** are used to build a high-res catalog of 160,000 earthquakes. Blue dots show mixed-frequency earthquakes (MFEs) discovered with unsupervised ML. *Wang et al., 2023*



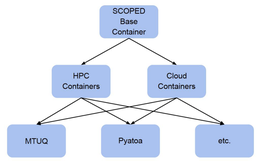

***Spectral feature extraction*** using ***unsupervised machine learning*** and hierarchical clustering: Waveform examples of one representative event in each cluster (A and D), stacked fingerprints (B and E), and stacked spectrograms (C and F) using 100 events in each cluster.



Pre-eruption evolution of MFEs in ~1 week, ~2 days, and 8 hours time scale. Dashed red curve shows binned seismic moment of the MFEs leading up to the eruption (dashed black line). Arrows in (B) point to northward and southward migration of MFEs prior and leading up to the eruption. *Wang et al. 2023*

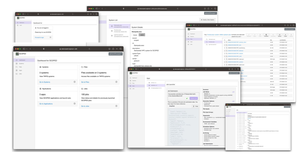
## SCOPED Container Registry

We now have adjtomo, MTUQ, pysep, specufex, and MsPASS added to the SCOPED Container Registry built with the SCOPED Base Container. We also added an external software, SeisSol, to our container collection using the base container.



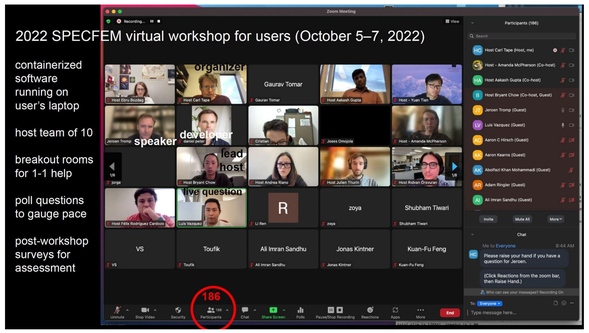

## SCOPED Gateway

The SCOPED Gateway prototype has been created through the utilization of Tapis. Currently, we have integrated MTUQ and MsPASS as HPC applications added into the gateway. Ultimately, this will enable all SCOPED users to access containerized applications and execute them on either Cloud or HPC systems.



## Training users of SCOPED codes

SCOPED contributed toward two training workshops: **Specfem** on October 5–7, 2022 (Bryant Chow, Carl Tape), and **High-Performance Seismology Cybertraining** on May 9–12 (Marine Denolle). These free workshops were attended by hundreds and included downloadable containerized software for users to gain research-level experience.




Zoom screenshot from a free virtual workshop on Specfem3D, held October 5-7, 2022, and hosted by UAF. The screenshot shows a range of participants, including the organizer (Carl Tape), the guest speaker (Jeroen Tromp), the principal developer (Daniel Peter), the lead host (Bryant Chow), a participant asking a question (Luis Vasquez), and the host team: Ebru Bozdag, Aakash Gupta, Yuan Tian, Amanda McPherson, Andrea Riano, Julien Thurin, Ridvan Orsvun, and Felix Cardozo. At this starting time of the workshop, there were 186 participants.

**Summary of attendance**

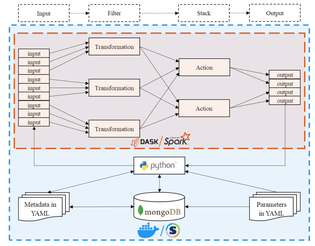
There were over 200 interested participants, 85 online participants the first day, then a steady 30-50 participants throughout the week. We had 175 Slack members and over 2.6K slack messages during the live debugging sessions.
The project resulted in a Youtube Channel and are writing a JupyerBook with the materials

JupyterBook

YouTube recordings


## Establishing the SCOPED platform

We have established a base software container for SCOPED that contains basic Python packages, as well as ObsPy, pysep, and MsPASS (see below). The container can run on the gateway, offering both batch and interactive modes through Jupyter Lab.




MsPASS: Massive Parallel Analysis System for Seismologists:

- Core implementation in Python and C++
- Data management using MongoDB
- Scalable parallel processing framework using Spark or Dask
- All components are containerized that can be pulled by Docker or Singularity for distribution
- Data provenance support at global and object levels

### References and Links


SeisNoise: https://github.com/tclements/SeisNoise.jl
NoisePy: https://github.com/mdenolle/NoisePy
Ambient Noise on the cloud: https://github.com/Denolle-Lab/C4-Project.jl
MSPASS: https://www.mspass.org/
Mtuq: https://uafgeotools.github.io/Mtuq/index.html
Specfem2d: https://github.com/geodynamics/Specfem2d
Specfem3d: https://github.com/geodynamics/Specfem3d
Specfem3d_Globe: https://github.com/geodynamics/Specfem3d_Globe
Pysep: https://pysep.readthedocs.io/en/latest/
Pyatoa: https://pyatoa.readthedocs.io/en/latest/;
Seisflows: https://seisflows.readthedocs.io/en/latest/
SPECUFEX: https://github.com/Specufex/specufex
BPMF: https://github.com/ebeauce/Seismic_BPMF
HYPODD: https://www.ldeo.columbia.edu/~felixw/HYPODD/
QUAKEFLOW: https://github.com/wayneweiqiang/QuakeFlow
PHASENET: https://github.com/wayneweiqiang/PhaseNet


SCOPED website: https://seisscoped.org/
SCOPED container registry: https://github.com/orgs/SeisSCOPED/packages
SCOPED base container: https://github.com/SeisSCOPED/container-base


The SCOPED project is supported by the National Science Foundation via the Cyberinfrastructure for Sustained Scientific Innovation program of the Office of Advanced Cyberinfrastructure; award numbers to each PI institution are listed at the top.