SCEC Statewide California Earthquake Center

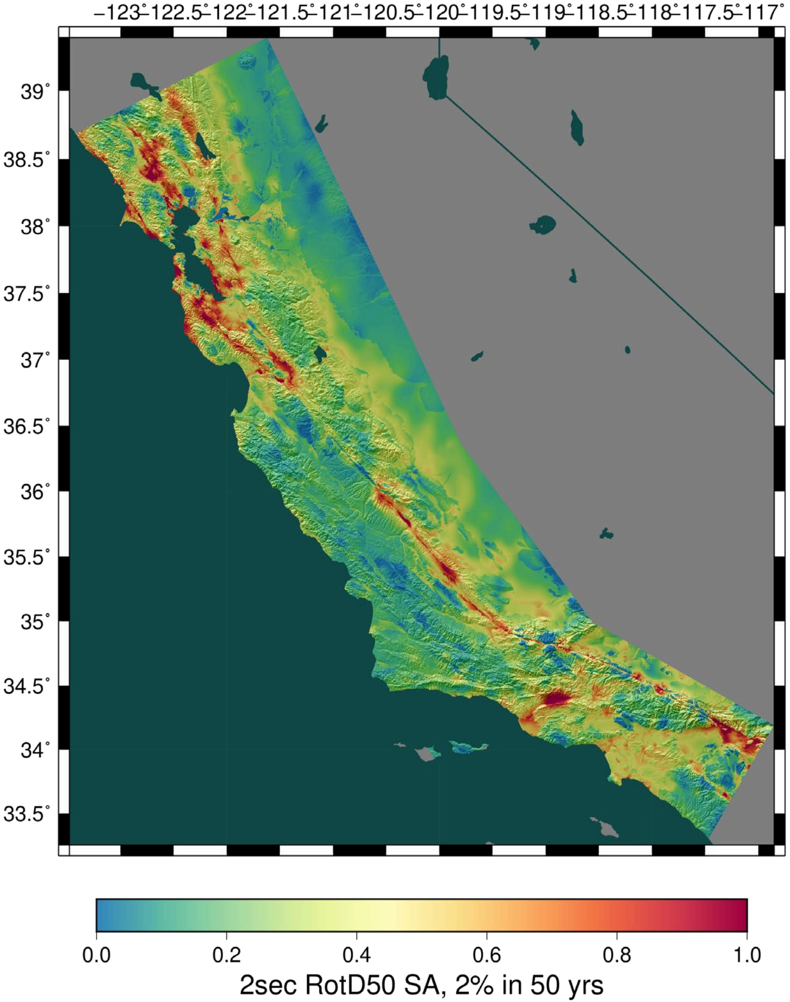
Managing Simulated Data Products from the CyberShake PSHA Platform

Scott Callaghan (SCEC) & the CyberShake Collaboration

September 2, 2024

Geo-INQUIRE Workshop on Data Lakes

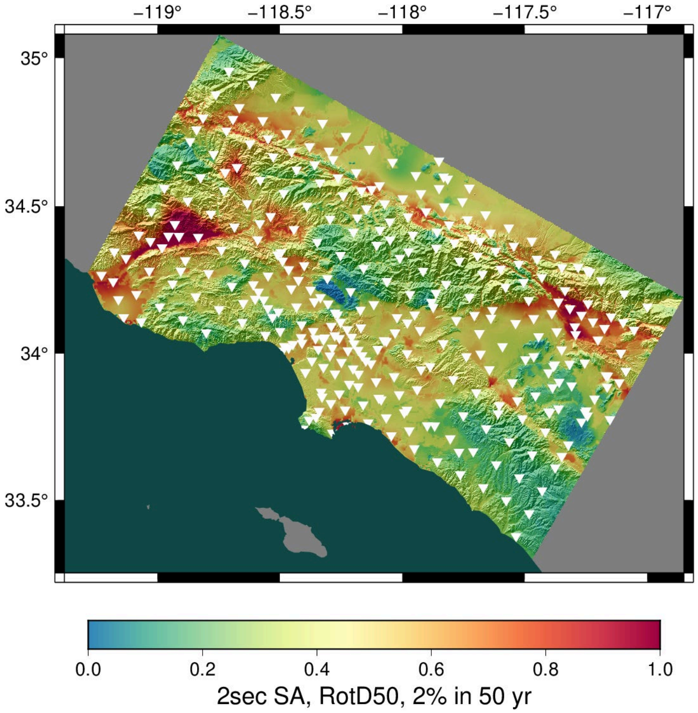
scottcal@usc.edu



- CyberShake Overview
- Data and Metadata
- Current CyberShake milestones
- Data challenges (and solutions)
- What’s next?
- Opportunities for collaboration

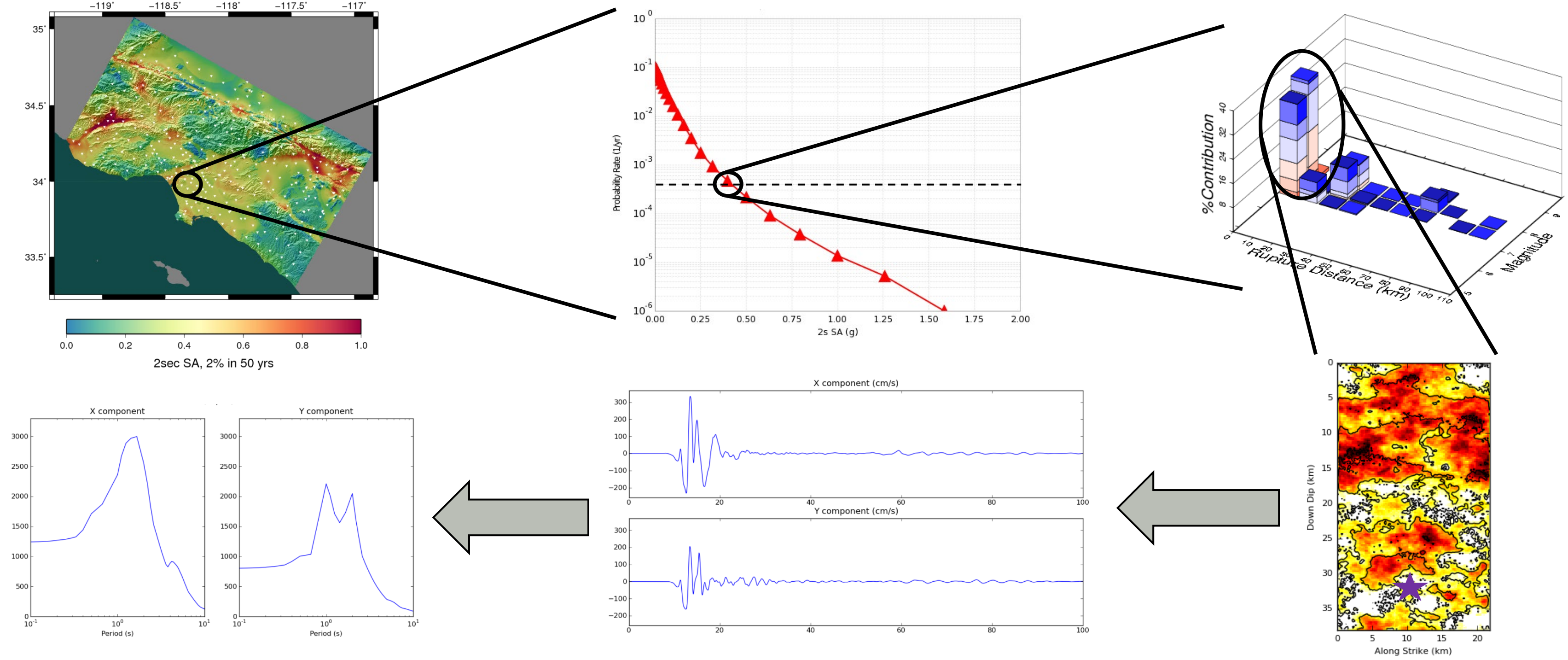
CyberShake overview

- SCEC-developed 3D physics-based probabilistic seismic hazard analysis (PSHA) platform
- Earthquake rupture forecast (ERF) provides list of relevant events + probabilities
- Reciprocity-based approach to simulate low-frequency seismograms for sites of interest
- Intensity measures derived from seismograms
- Hazard results from sites interpolated for map
- Optional stochastic high-frequency simulations to produce broadband models



Hazard map from most recent Southern California CyberShake Study, 22.12. Each triangle is a site location.

CyberShake Data Layers



Hazard Map

Hazard Curve

Disaggregation

Intensity Measures

Seismograms

Rupture Realization

Data Products

- Seismograms (historically 2-component) for each event for each site
    - Base raw data product
- Peak shaking measures
    - Used to be geometric mean; now RotD50 and RotD100
    - Subset (~25%) stored in relational database for quick access
- Durations
    - 5-75%, 5-95%, others
    - ~25% stored in relational database
- Disaggregations, hazard curves, hazard maps
    - Aggregate data products

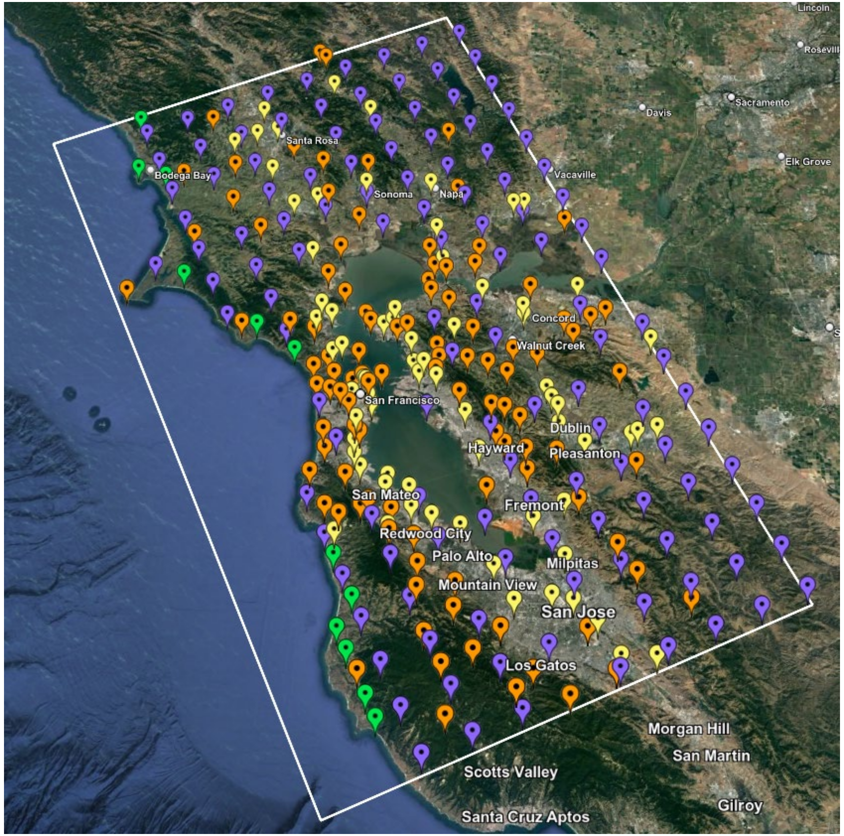
Metadata

- Seismic
  - Maximum frequency
  - Site info
  - Event information (magnitude, hypocenter, fault name)
  - Velocity model
  - Rupture generator
  - Tracked in database

- Simulation-based
  - Mesh dimensions
  - Timestep size, number of timesteps
  - Tracked in database, on wiki
- Runtime-based (provenance)
  - Execution system
  - Code version
  - Command-line arguments
  - Runtime
  - Tracked by workflow system (Pegasus-WMS, HTCondor)

Study 24.8

- Began latest CyberShake study last Tuesday
- Updated broadband simulations for the San Francisco Bay Area
- Improved velocity model
- Similar configuration to Study 22.12
- New data products:
  - 3-component seismograms
  - Vertical response spectra
  - Period-dependent durations



Challenge: Large Data Lake Size

From Study 22.12

Data Product	Records per study	Number of files per study	Data size per study
Low-frequency seismograms	200 million	2 million	15 TB
Low-frequency IMs	10 billion	6 million	<1 TB
Broadband seismograms	200 million	2 million	60 TB
Broadband IMs	30 billion	6 million	<1 TB
Aggregate products	3,000	3,000	<1 TB
Total	**40 billion**	**16 million**	**75 TB**

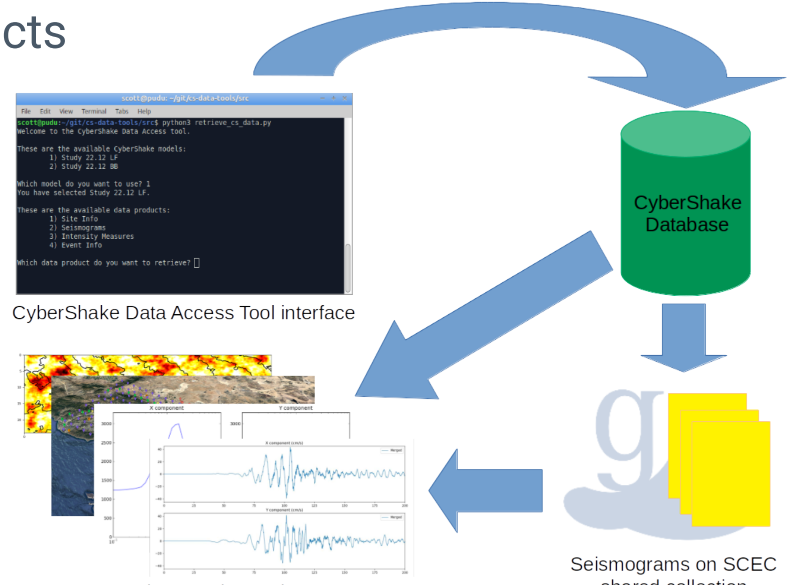
- Data currently stored at Center for Advanced Research Computing at USC
- Plan to migrate to DesignSafe at Texas Advanced Computing Center

Challenge: Support Community Access

- Key contribution of CyberShake is the creation of the dataset for later use
- Dozens of researchers interested in working with CyberShake data
  - Internal: members of the CyberShake collaboration
  - External: members of the broader SCEC, engineering, and preparedness communities
- Describe what data products are available
- Different users desire different levels of access:
  - Nicely packaged data
  - Interactive interface
  - API for scripting

Challenge: Identify and Deliver Data Subsets

- Size of the dataset makes full download difficult
  - Most users don’t need it all anyway
  - Query interface needed to help users select subsets
- Metadata must be delivered with data products
  - Documentation necessary
- Developed CyberShake Data Access Tool
  - Python-based, open source
  - Prompts users with questions to create filters
  - Users can bypass interactive components for use with scripting
  - Delivers database products, seismograms, and seismic metadata
  - https://github.com/SCECcode/cs-data-tools/



scott@pudu: ~/git/cs-data-tools/src
File Edit View Terminal Tabs Help

```
scott@pudu:~/git/cs-data-tools/src$ python3 retrieve_cs_data.py
Welcome to the CyberShake Data Access tool.

These are the available CyberShake models:
        1) Study 22.12 LF
        2) Study 22.12 BB

Which model do you want to use? 1
You have selected Study 22.12 LF.

These are the available data products:
        1) Site Info
        2) Seismograms
        3) Intensity Measures
        4) Event Info

Which data product do you want to retrieve? 
```

CyberShake Data Access Tool interface

CyberShake Database

Seismograms on SCEC shared collection

Data products and metadata

Challenge: On-Demand Data Products

- Not all possible data products are created at study time
- Rupture slip time histories
- Synthetic ShakeMaps
- Disaggregations at additional return periods
- Intensity measures on disk, but not in database
- How to support user generation of data products? Gateway? Quakeworx?
  No implemented solution to this challenge yet

Challenge: Human Resources

- Difficult to obtain funding for scientific software development in the US
- Limited resources for facilitating delivery of data products to users
  - Minimize CyberShake developer involvement
  - Easy-to-use interfaces
  - Documentation, tutorials
  - Extensible
- Balance between targeting new scientific milestones and improving usefulness of existing data

Looking Ahead

- Study 24.8 to finish in about 2 months
- CyberShake data lakes will continue to grow
  - 2 Hz deterministic runs targeted for 2025
  - Integrate non-linear forward simulations
  - Quantify uncertainty of velocity model and high-frequency codes through additional simulations
- Looking for ways to remove barriers to usage
  - Improved documentation
  - Migration to DesignSafe (DOI, access to DesignSafe tools)
  - Increase awareness in potential users

Collaboration and Standardization Opportunities

- File formats + converters
  - CyberShake uses custom binary data formats
  - Move to more common format? (HDF5, ASDF, …)?
  - Regardless of format, standard converters will be needed
- Capture and distribution of simulation parameters
  - Identify standard simulation parameters that are:
    - Of interest to users
    - Needed for reproducibility
  - Distribute along with other metadata when data is delivered
- What level of reproducibility do we seek?
- If formats and metadata are similar, opportunities for common tools
  - Single point-of-entry for users to access multiple data lakes

Thanks!

U.S. National Science Foundation

NSF

USGS
science for a changing world

PG&E

U.S. DEPARTMENT OF ENERGY

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

ISI
Information Sciences Institute

USC | Advanced Research Computing
Enabling scientific breakthroughs at scale

TACC
TEXAS ADVANCED COMPUTING CENTER

Pegasus