Informing Probabilistic Tsunami Forecasting with Massive Earthquake Ensembles: FFT-based Multi-GPU Waveform Generation and Posterior-Start Flow Matching

The University of Texas at Austin
Oden Institute for Computational Engineering and Sciences

Bowen Shi[a] and Omar Ghattas (advisor)[a,b]

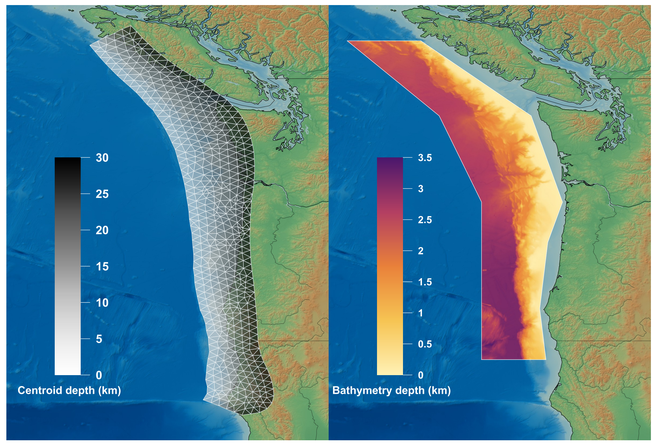
[a]*Oden Institute for Computational Engineering and Sciences, UT Austin,* [b]*Walker Department of Mechanical Engineering, UT Austin*

OPTIMUS
BEYOND FORWARD SIMULATION

Overview

Real-Time Probabilistic Tsunami Forecasting at Cascadia Subduction Zone

- The Cascadia Subduction Zone is a megathrust fault off the Pacific Northwest, where earthquake-driven seafloor uplift and subsidence can generate destructive tsunamis that reach nearby coastlines within minutes.
- Seafloor pressure sensors provide early observations of the evolving acoustic and tsunami response.
- **Goal: infer earthquake-induced seafloor motion and forecast future coastal wave heights with uncertainty.**



Computational domain of the Cascadia Subduction Zone. Left: fault-plane model for rupture modeling, with 963 triangular subfaults colored by centroid depth below sea level. Right: 3D acoustic–gravity tsunami simulation domain, discretized with 14,473,092 hexahedral elements and 2,416,530 seafloor grid points; color denotes bathymetric depth.

Real-time Bayesian Inference

Inverse and prediction problems

Infer seafloor velocity m from early pressure data d, and predict future tsunami wave heights q.

$$\mathbf{d} = \mathbf{F}\mathbf{m} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(0, \boldsymbol{\Gamma}_{\text{noise}}), \quad \mathbf{q} = \mathbf{F}_q \mathbf{m}$$

Gaussian prior and posterior [Henneking et al., 2025]

$$\mathbf{m} \sim \mathcal{N}(0, \boldsymbol{\Gamma}_{\text{pr}}), \qquad \mathbf{m} \mid \mathbf{d} \sim \mathcal{N}(\mathbf{m}_{\text{post}}, \boldsymbol{\Gamma}_{\text{post}})$$

$$\boldsymbol{\Gamma}_{\text{post}} = \left(\mathbf{F}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1} \mathbf{F} + \boldsymbol{\Gamma}_{\text{pr}}^{-1}\right)^{-1}$$

$$\mathbf{m}_{\text{post}} = \boldsymbol{\Gamma}_{\text{post}} \mathbf{F}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1} \mathbf{d}$$

Precompute a fast data-to-forecast map

$$\gamma_{\mathbf{d}} = \mathcal{N}(\mathbf{Q}\mathbf{d}, \boldsymbol{\Gamma}_{\text{post}}^{(q)})$$

$$\mathbf{Q} = \mathbf{F}_q \boldsymbol{\Gamma}_{\text{post}} \mathbf{F}^T \boldsymbol{\Gamma}_{\text{noise}}^{-1}, \qquad \boldsymbol{\Gamma}_{\text{post}}^{(q)} = \mathbf{F}_q \boldsymbol{\Gamma}_{\text{post}} \mathbf{F}_q^T$$

Why learn a correction?

Linear Gaussian inverse problem assumes earthquakes come from a Gaussian distribution, while real ruptures could be complicated.
We need to provide **non-Gaussian forecast distribution** conditioned on d, informed by potential earthquakes, and in real time

Can We Have All Three?

- **High-fidelity physics:** Accurate simulations are expensive. Data is hard to get.
- **Stochastic earthquakes:** Capturing source variability requires many rupture realizations.
- **Real-time uncertainty:** Deliver probabilistic forecasts within seconds.

Method: Posterior-Start Flow Matching

Offline: prepare the reference and learn the correction

Gaussian reference
$\gamma_{\mathbf{d}} = \mathcal{N}(\mathbf{Q}\mathbf{d}, \boldsymbol{\Gamma}_{\text{post}}^{(q)})$
Precompute Q and the covariance.

Training pairs
Stochastic ruptures + GPU generator
$\mathcal{D} = \{(\mathbf{d}_i, \mathbf{q}_{1,i})\}_{i=1}^{N}$

Conditional flow matching

$$\mathbf{q}_\tau = (1-\tau)\mathbf{q}_0 + \tau \mathbf{q}_1, \quad \mathbf{q}_0 \sim \gamma_{\mathbf{d}}, \quad \tau \sim \mathcal{U}(0,1)$$

$$\min_\theta \mathbb{E}\left[\|\mathbf{v}_\theta(\mathbf{q}_\tau, \tau, \mathbf{d}) - (\mathbf{q}_1 - \mathbf{q}_0)\|_2^2\right]$$

learned $\mathbf{v}_\theta$

Online: forecast

Gaussian start
$\mathbf{q}_0 \sim \gamma_{\mathbf{d}}$

Learned flow
$\frac{d\mathbf{q}_\tau}{d\tau} = \mathbf{v}_\theta(\mathbf{q}_\tau, \tau, \mathbf{d})$
$\tau: 0 \to 1$

Non-Gaussian forecast

Source-to-Waveform Generation on GPUs

Stochastic ruptures
Sample slip + rupture timing
$\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(N)}$

$\mathbf{F}_G$
elastic Green's functions

Seafloor motion
displacement → velocity
$\mathbf{m} = \mathbf{D}_t \mathbf{u}$

$\mathbf{F}$, $\mathbf{F}_q$
ocean adjoint responses

Pressure + waveforms
$\mathbf{d} = \mathbf{F}\mathbf{m} + \boldsymbol{\eta}$
$\mathbf{q} = \mathbf{F}_q \mathbf{m}$

Block Toeplitz maps

$$\begin{bmatrix} \mathbf{A}_0 & 0 & 0 \\ \mathbf{A}_1 & \mathbf{A}_0 & 0 \\ \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 \end{bmatrix}$$

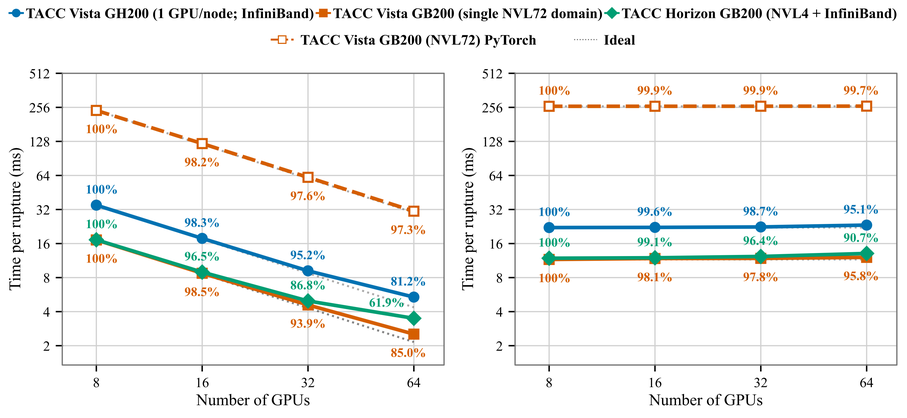
embed + zero-pad → **Circulant** → temporal FFT → **Block diagonal** → **GPU batched GEMV**

$$\hat{\mathbf{y}}_\ell = \hat{\mathbf{A}}_\ell \hat{\mathbf{x}}_\ell$$

IFFT + truncate

Both maps reuse precomputed Fourier blocks; seafloor fields stay on GPUs [Venkat et al., 2025].

Multi-GPU Scalability for Source-to-Waveform Generation

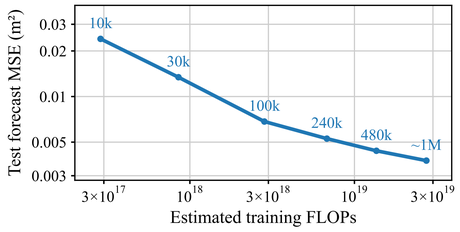


Strong scaling (left): 60 subfaults, 4 outputs. **Weak scaling (right): 480 subfaults, 32 outputs.**

At full Cascadia scale, 64 GB200 GPUs in one NVL72 domain map 963 subfaults through 2,416,530 seafloor points to 64 waveforms over 256 timesteps. Resident operators use 9.45 TiB HBM.

24 ms per rupture; 10,000 waveform realizations in 4 minutes after setup.
> 10,000 times faster than per rupture solver with states resident 🚀

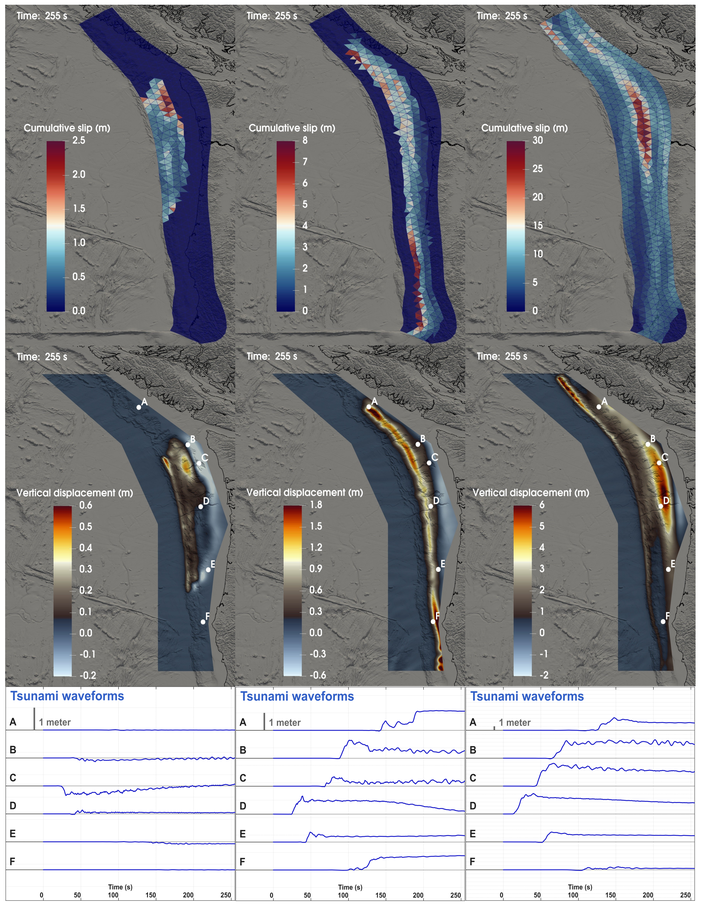
Scaling with Massive Earthquake Ensembles



More training ruptures expose a wider range of earthquake variability.

Waveform MSE significantly drops as training grows from 10k to 1M earthquakes.

From Stochastic Ruptures to Tsunami Waveforms

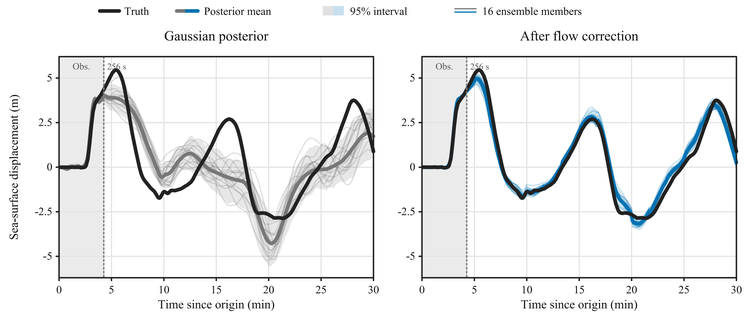


Three stochastic earthquake realizations: cumulative fault slip (top), vertical seafloor displacement at final time 255s (middle), and tsunami waveforms (bottom)—the input, intermediate field, and output of the source-to-waveform pipeline. Scan the QR code for visualizations.

Slip-rate function

Seafloor displacements

Online Forecasting Results



Forecasts for a held-out event, conditioned on the first 256 s of pressure observations. The posterior-start flow was trained on 1M stochastic ruptures. It corrects the mean waveform and concentrates the ensemble. Shading: 95% intervals from 64 draws; 16 samples shown.
Across 4,096 test events: 79% lower RMSE and 86% lower CRPS than Gaussian. Pointwise 95% coverage: 89%.

Acknowledgments: we thank Stefan Henneking and Sreeram Venkat at UT Austin for their collaboration and technical contributions.