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

Bowen Shi
bowenshi@utexas.edu
Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin
Austin, Texas, USA

Abstract

Cascadia tsunamis may reach nearby coastlines within minutes. We seek real-time probabilistic forecasts of future wave heights from early seafloor-pressure observations. A physics-based Bayesian framework exploits linear time-invariant dynamics and a structured Gaussian prior for rapid inference, but this prior does not encode rupture-induced correlations. Incorporating earthquake information requires paired pressure and tsunami waveforms, yet separate earthquake and ocean simulations for every rupture are prohibitively expensive. We developed a distributed FFT-based source-to-waveform pipeline that reuses precomputed responses; on 64 GB200 GPUs, it maps one full-scale Cascadia rupture to waveforms in 24 ms after setup. We use one million generated ruptures to train a posterior-start flow that begins from the Gaussian posterior and learns an earthquake-informed correction. On 4,096 held-out events, this reduces forecast RMSE by 79% and CRPS by 86%.

Keywords

tsunami forecasting, Bayesian inverse problems, block Toeplitz matrices, multi-GPU computing, flow matching

1 Real-Time Probabilistic Forecasting

The Cascadia Subduction Zone is a megathrust fault off the Pacific Northwest, where earthquake-driven seafloor motion can generate destructive tsunamis. Bottom-pressure sensors provide early observations of the evolving acoustic and tsunami response. We use 175 sensors observed for 256 s to forecast 21 coastal wave-height trajectories through 1800 s with uncertainty.

Let $\mathbf{m}$ denote the seafloor velocity driving the acoustic–gravity model, $\mathbf{d}$ the pressure observations, and $\mathbf{q}$ the forecast quantities:

$$\mathbf{d} = F\mathbf{m} + \boldsymbol{\eta}, \qquad \mathbf{q} = F_q\mathbf{m}. \tag{1}$$

Because the dynamics are linear and time-invariant, F and F_q have block lower-triangular Toeplitz structure. A temporal FFT converts their convolutions into batched frequency-domain matrix–vector products, enabling fast operator application [1, 5].

With a structured Gaussian prior on $\mathbf{m}$, the conditional forecast is available in closed form [2]:

$$\mathbf{q} \mid \mathbf{d} \sim \gamma_{\mathbf{d}} = \mathcal{N}\left(Q\mathbf{d}, \Gamma_{\text{post}}^{(q)}\right). \tag{2}$$

This gives a fast, physics- and data-conditioned forecast. Its limitation is the prior: this structured Gaussian model does not encode the correlations created by realistic rupture extent, slip, rupture onset, and propagation.

2 From Stochastic Ruptures to Waveforms

To learn these missing correlations, we need many realistic earthquakes paired with their pressure and tsunami waveforms. FakeQuakes [4] samples rupture extent, hypocenter, spatially correlated slip, rupture-onset time, and rise time on 963 triangular subfaults, producing a slip-rate function $\mathbf{s}$. Separate elastic and acoustic–gravity simulations for every rupture make large training ensembles prohibitively expensive.

We instead precompute the impulse responses of both physical models. For each rupture, the source-to-waveform map becomes

$$\mathbf{z} = F_G\mathbf{s}, \quad \mathbf{m} = D_t\mathbf{z}, \quad \mathbf{d} = F\mathbf{m} + \boldsymbol{\eta}, \quad \mathbf{q} = F_q\mathbf{m}, \tag{3}$$

where F_G maps fault slip rate to seafloor displacement and D_t converts displacement to the bottom velocity driving the acoustic–gravity model. FFTMatvec evaluates both physical stages without repeated PDE solves.

The shared seafloor dimension is distributed across GPUs: each GPU stores part of F_G together with the corresponding part of F and F_q. The rupture source is broadcast, the local seafloor response is computed, and the second operator immediately maps it to local pressure and tsunami-waveform contributions. The seafloor field stays distributed in GPU memory; only waveform outputs are reduced.

3 Multi-GPU Source-to-Waveform Generation

At full Cascadia scale, the pipeline maps 963 subfaults through 2,416,530 seafloor points over 256 time steps. In the 64-output benchmark, the precomputed frequency-domain operators occupy 9.45 TiB.

On 64 NVIDIA GB200 GPUs within one NVL72 domain, one rupture is mapped to waveforms in 24 ms after setup, or 41.67 ruptures/s. This is 21.8× faster than distributed direct time-domain convolution and allows 10,000 full-scale ruptures to be processed in approximately four minutes.

Strong and weak scaling cover TACC Vista GH200, Vista GB200 NVL72, and TACC Horizon GB200 NVL4/InfiniBand, against ideal and Vista GB200 PyTorch references.

4 Posterior-Start Flow Matching

These ensembles provide samples from the earthquake-informed conditional law

$$\nu_{\mathbf{d}} = \mathrm{Law}(\mathbf{q} \mid \mathbf{d}).$$

Rather than starting flow matching [3] from an isotropic Gaussian, we start from the Gaussian posterior:

$$\mathbf{q}_0 \sim \gamma_{\mathbf{d}}, \qquad \frac{d\mathbf{q}_t}{dt} = v_\theta(\mathbf{q}_t, t, \mathbf{d}), \qquad \mathbf{q}_1 \sim \nu_{\mathbf{d}}. \tag{4}$$

A Diffusion Transformer conditioned on pressure observations learns the correction from the Gaussian forecast to the earthquake-informed distribution. The Gaussian posterior already captures the linear physics and data conditioning, while the flow learns the non-Gaussian rupture information missing from the prior.

5 Forecasting Results and Scalability

We trained on one million stochastic ruptures and evaluated on 4,096 held-out events with 64 samples per forecast. Noise standard deviation is 1% of the training pressure standard deviation. Over 290–1800 s, waveform RMSE decreases from 0.27422 to 0.05870 m (79%), and CRPS from 0.14136 to 0.02038 m (86%). Nominal 95% pointwise intervals attain 89% coverage.

In a separate 30,000-event, eight-draw evaluation, MSE falls from 0.0240 to 0.00378 m^2 as training grows from 10,000 to 965,904 ruptures. Training compute rises from 2.86×10^{17} to 2.74×10^{19} estimated FLOPs. With one seed and 60 epochs, data and compute scale together.

Together, FFT-based ensemble generation makes earthquake-informed training practical, while posterior-start flow matching uses those ensembles to improve a real-time physics-based probabilistic forecast.

Acknowledgments

This work was conducted with Sreeram Venkat and Stefan Henneking under the supervision of Omar Ghattas.

References

[1] Stefan Henneking, Sreeram Venkat, Veselin Dobrev, John Camier, Tzanio Kolev, Milinda Fernando, Alice-Agnes Gabriel, and Omar Ghattas. 2025. Real-Time Bayesian Inference at Extreme Scale: A Digital Twin for Tsunami Early Warning Applied to the Cascadia Subduction Zone. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '25)*. Association for Computing Machinery, New York, NY, USA, 60–71. doi:10.1145/3712285.3771787

[2] Stefan Henneking, Sreeram Venkat, and Omar Ghattas. 2026. Goal-Oriented Real-Time Bayesian Inference for Linear Autonomous Dynamical Systems with Application to Digital Twins for Tsunami Early Warning. *J. Comput. Phys.* 552 (2026), 114682. doi:10.1016/j.jcp.2026.114682

[3] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*. https://openreview.net/forum?id=PqvMRDCJT9t

[4] Diego Melgar, Randall J. LeVeque, Douglas S. Dreger, and Richard M. Allen. 2016. Kinematic Rupture Scenarios and Synthetic Displacement Data: An Example Application to the Cascadia Subduction Zone. *Journal of Geophysical Research: Solid Earth* 121, 9 (2016), 6658–6674. doi:10.1002/2016JB013314

[5] Sreeram Venkat, Milinda Fernando, Stefan Henneking, and Omar Ghattas. 2025. Fast and Scalable FFT-Based GPU-Accelerated Algorithms for Block-Triangular Toeplitz Matrices with Application to Linear Inverse Problems Governed by Autonomous Dynamical Systems. *SIAM Journal on Scientific Computing* 47, 5 (2025), B1201–B1226. doi:10.1137/24M1683172