PDSW23: 8th International Parallel Data Systems Workshop
Nov. 12, 2023

Compression of Scientific Simulation Data by Stochastic Basis Expansion - Example on Multiple Computer Systems

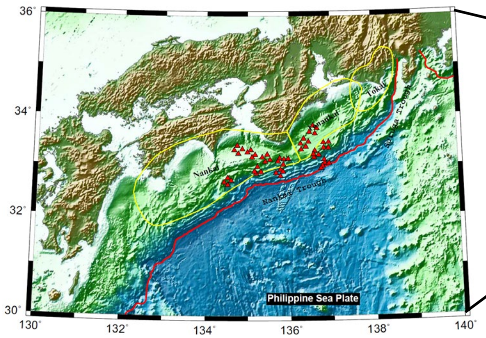
Kohei Fujita, University of Tokyo
Tsuyoshi Ichimura, University of Tokyo
Maddegedara Lalith, University of Tokyo
Muneo Hori, Japan Agency for Marine-Earth Science and Technology

Introduction

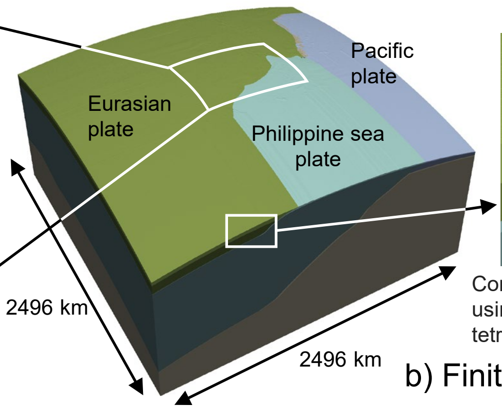
- High-performance parallel and distributed storage systems are being developed to enable scientific data storage and processing
  - Here, information about the structure and temporal evolution of phenomena is extracted by storing and processing data of the target phenomena
- For large-scale complex problems, the space in which the phenomenon takes place is especially large, and amount of data is enormous
  - Important to develop both high-performance data storage systems and algorithms that effectively utilize these systems for data compression, restoration, and analysis

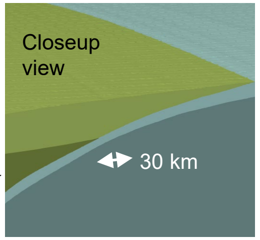
Example of large-scale complex problem

- Uncertainty quantification for earthquake simulation: important as crustal information is not known precisely
- Since the size of crustal deformation simulation is huge (1.3×10^{11} degrees of freedom), uncertainty quantification requiring 10000 cases of crustal deformation analysis cannot be conducted as data becomes $1.3 \times 10^{11} \times 10000 = 1.3 \times 10^{15}$ degrees of freedom
- We develop a method to generate compressed data of many cases of simulations by stochastic basis expansion to enable uncertainty quantification of such large-scale problems



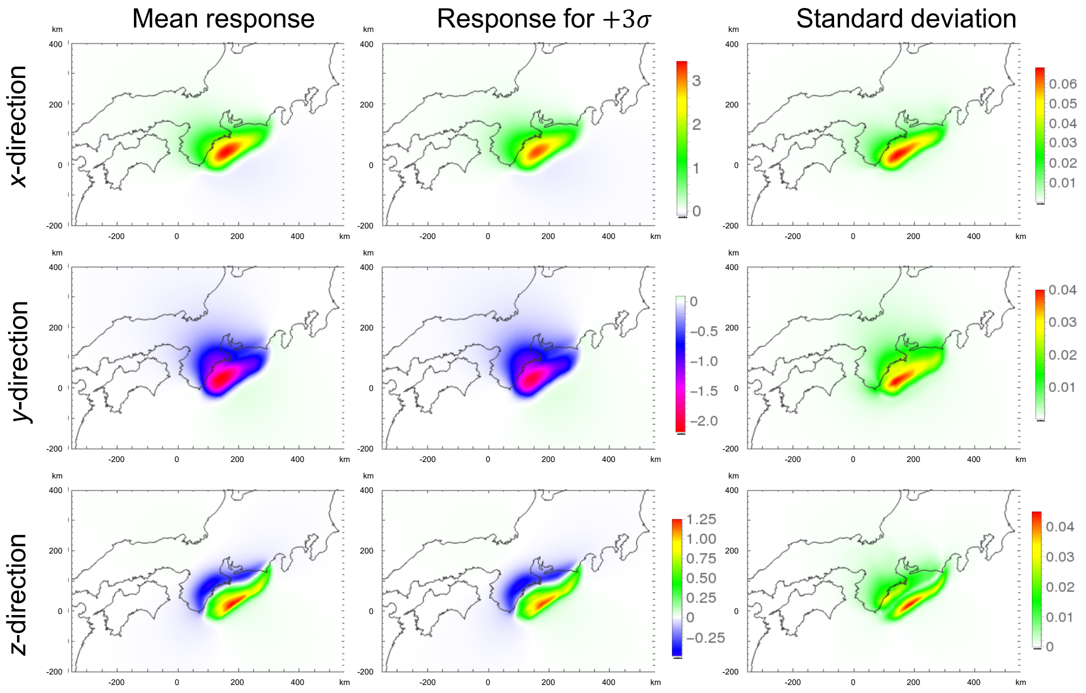
a) Target area



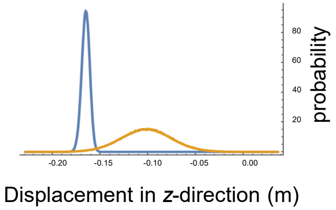


Complex geometry discretized using 375 m second-order tetrahedral elements

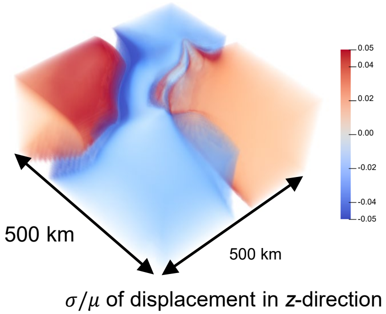
b) Finite-element model



c) Stochastic response at surface (displacement (m))



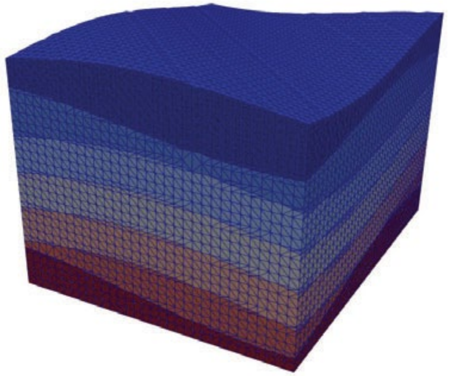
d) Stochastic response at two points on surface



e) Stochastic response shown in 3D

Reduction of Monte Carlo simulation data

- Expand results of many-case simulations using stochastic basis, and store only the coefficients for each basis
  - Response data expanded as $\boldsymbol{u} = \Sigma_{p=0}^{n_{PC}} \Psi_p \boldsymbol{u}^{(p)}$ in the stochastic space, where Ψ_p are the stochastic bases and $\boldsymbol{u}^{(p)}$ are components in physical space using finite-element discretization
- Example for crustal deformation problem:
  - Results obtained by running deterministic simulations (corresponding to a Monte Carlo simulation sample) and the response results reconstructed from the stochastic basis expansion match with high accuracy
  - Response results for many Monte Carlo simulations can be reconstructed accordingly



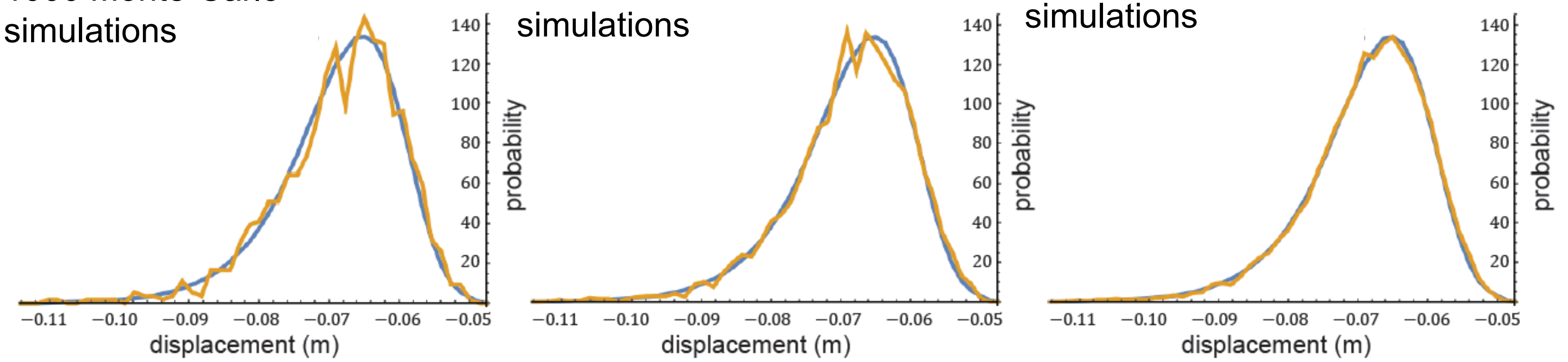
Problem with normal distribution in material properties

Compute displacement response to input fault slip

1000 Monte-Carlo simulations

5000 Monte-Carlo simulations

10000 Monte-Carlo simulations



Blue line: probability distribution obtained by the developed method

Results for 10000 Monte Carlo simulations can be reconstructed with small data size

Data size reduction rate and computation cost

- Data storage size can be reduced by 10000 / 286 = 35-fold by use of the proposed method:
  - 10000 cases of Monte Carlo simulations are typically required for considering the convergence of the response probability distribution
  - For a problem with 10 random variables, third-order stochastic basis expansion, which is considered sufficient for convergence of response, can be stored with the data size equivalent to 286 cases of deterministic simulations
- However, high-order stochastic expansion leads to large computation costs

Reduction of computation cost

- Based on the expansion calculations using stochastic bases conducted in the numerical analysis field, we developed a scalable algorithm that can efficiently utilize high-performance computer systems [1] for suppressing analysis costs involved in high-order expansions
  - Implemented on CPU-based Fugaku: up to 1.3×10^{11} degrees of freedom (DOF) x 10000 cases = 1.3×10^{15} DOF is reduced by 35-fold to 3.7×10^{13} DOF using nearly full Fugaku (147456 nodes)
  - Implemented on GPU-based Frontier: up to 3.7×10^{14} DOF data set is reduced by 35-fold using 5120 nodes of Frontier

[1] T. Ichimura, et al., “Extreme Scale Earthquake Simulation with Uncertainty Quantification,” SC22 Gordon Bell Prize Finalist, 2022.

Summary

- Developed a method that generates compressed large-scale simulation data by use of stochastic basis expansion
- Development of a scalable algorithm led to a 35-fold reduction in data size for problems up to 10^{15} degrees of freedom by use of CPU-based Fugaku and GPU-based Frontier
- With the developed method, we enabled uncertainty quantification of large-scale crustal deformation problem of 1.3×10^{11} degrees of freedom

Acknowledgements

- This work was supported by MEXT as "Program for Promoting Researches on the Supercomputer Fugaku" (Largescale numerical simulation of earthquake generation, wave propagation and soil amplification, JPMXP1020200203). This work was supported by JST FOREST Program (Grant Number JPMJFR215Q, Japan). This work was supported by JSPS KAKENHI Grant Numbers JP18H05239, JP18H03795, JP22K12057, JP23H00213.
- This work used computational resources of Fugaku provided by the RIKEN Center for Computational Science (Project ID: hp220171, hp230203). We thank members of RIKEN Center for Computational Science and Research Organization for Information Science and Technology for their support and coordination of full node scale simulations on Fugaku.
- This work used computational resources of Frontier provided by the OLCF Director's Discretion Project (Project ID: geo146). We thank members of the support team of Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory for their prompt help in the use of Frontier.