JGR Oceans

RESEARCH ARTICLE
10.1029/2024JC021565

Scenario Superposition Method for Real-Time Tsunami Prediction Using a Bayesian Approach

Saneiki Fujita[1], Reika Nomura[2], Shuji Moriguchi[2], Yu Otake[1], Randall J. LeVeque[2,3], and Kenjiro Terada[1,2]

[1]Department of Civil and Environmental Engineering, Tohoku University, Sendai, Japan, [2]International Research Institute of Disaster Science, Tohoku University, Sendai, Japan, [3]Department of Applied Mathematics, University of Washington, Seattle, WA, USA

Key Points:

- The proposed scenario superposition method achieves real-time tsunami waveform prediction via a linear combination of precomputed scenarios
- A set of parameters weighting each scenario is probabilistically estimated using a Bayesian approach
- Three-hour tsunami waveforms can be predicted using wave heights measured at existing gauge locations in less than 10 min

Correspondence to:
S. Fujita,
saneiki.fujita.s8@dc.tohoku.ac.jp

Citation:
Fujita, S., Nomura, R., Moriguchi, S., Otake, Y., LeVeque, R. J., & Terada, K. (2024). Scenario superposition method for real-time tsunami prediction using a Bayesian approach. *Journal of Geophysical Research: Oceans*, *129*, e2024JC021565. https://doi.org/10.1029/2024JC021565

Received 9 JUL 2024
Accepted 21 NOV 2024

Author Contributions:
Conceptualization: Saneiki Fujita
Data curation: Reika Nomura
Funding acquisition: Saneiki Fujita, Shuji Moriguchi, Kenjiro Terada
Methodology: Saneiki Fujita, Reika Nomura, Shuji Moriguchi, Yu Otake
Software: Saneiki Fujita
Supervision: Kenjiro Terada
Validation: Saneiki Fujita
Visualization: Saneiki Fujita
Writing – original draft: Saneiki Fujita
Writing – review & editing: Reika Nomura, Shuji Moriguchi, Yu Otake, Randall J. LeVeque, Kenjiro Terada

Abstract In this study, we propose a scenario superposition method for real-time tsunami wave prediction. In the offline phase, prior to actual tsunami occurrence, hypothetical tsunami scenarios are created, and their wave data are decomposed into spatial modes and scenario-specific coefficients by the singular value decomposition. Then, once an actual tsunami event is observed, the proposed method executes an online phase, which is a novel contribution of this study. Specifically, the predicted waveform is represented by a linear combination of training scenarios consisting of precomputed tsunami simulation results. To make such a prediction, a set of weight parameters that allow for appropriate scenario superposition is identified by the Bayesian update process. At the same time, the probability distribution of the weight parameters is obtained as reference information regarding the reliability of the prediction. Then, the waveforms are predicted by superposition with the estimated weight parameters multiplied by the waveforms of the corresponding scenarios. To validate the performance and benefits of the proposed method, a series of synthetic experiments are performed for the Shikoku coastal region of Japan with the subduction zone of the Nankai Trough. All tsunami data are derived from numerical simulations and divided into a training data set used as scenario superposition components and a test data set for an unknown real event. The predicted waveforms at the synthetic gauges closest to the Shikoku Islands are compared to those obtained using our previous prediction method incorporating sequential Bayesian updating.

Plain Language Summary In this study, we propose a new tsunami waveform prediction method that represents actual tsunami waveforms by superposing scenario waveforms resulting from a series of precomputed tsunami simulations. The prediction scheme uses a Bayesian approach to sequentially estimate the weight parameters assigned to the scenarios using actual real-time wave height data. At the same time, it identifies the probability distribution of the weight parameters and provides information to understand the forecast error. The actual wave heights of the target event can be predicted by multiplying of the wave heights in the scenario database by the estimated weight parameters. To verify the advantages of the proposed method, a demonstration test is conducted for the Shikoku region of Japan, which is threatened by a large tsunami risk due to the Nankai Trough subduction zone. We show that the proposed method can predict wave heights even when the target event is not present in the scenario database, and we show its superiority over the previously proposed method.

1. Introduction

Although early tsunami warning systems play a critical role in mitigating the catastrophic damage caused by tsunamis, accurate forecasts remain challenging especially for near-field tsunamis. Notable examples include the 2004 Indian Ocean earthquake and tsunami and the 2011 earthquake and tsunami off the Pacific coast of Tohoku (Koshimura & Shuto, 2015; Satake, 2014). Based on the lessons learned from these experiences, various forecasting methods have been developed in the last decade to ensure both accuracy and rapid performance, and notable studies are summarized in Table 1. In addition to physical computations of tsunami wave propagation (Musa et al., 2018), methods such as data assimilation (Gusman et al., 2016; Heidarzadeh et al., 2019; Maeda et al., 2015; Wang et al., 2017, 2018) and machine learning techniques (Alan et al., 2023; Liu et al., 2021; Makinoshima et al., 2021; Mulia et al., 2022; Rim et al., 2022; Xu & Wu, 2023) have been developed. These methods use observational data, such as offshore tsunami data or onshore geodetic displacements sampled by the global navigation satellite system (GNSS) (Tsushima & Ohta, 2014).

Table 1
Tsunami Prediction Methods and Their Characteristics

Methodology	Studies	Advantages	Challenges
Real-time tsunami simulation	Oishi et al. (2015) Musa et al. (2018)	Tsunami propagation computation with physical model using governing equations	Treatment of uncertainty contained in initial input data Balance of computational cost and model complexity
Data assimilation	Maeda et al. (2015) Gusman et al. (2016) Wang et al. (2017); Wang et al. (2018)	Tsunami propagation computation with physical model using governing equations More reliable real-time simulation by sequentially assimilating observation data	Balance of computational cost and model complexity
Database searching schemes	Gusman et al. (2014) Fauzi and Mizutani (2020) Nomura et al. (2022)	Model simplicity and low computational cost Transparency of database search algorithms	Need a comprehensive scenario database
Machine learning and Deep learning	Liu et al. (2021) Makinoshima et al. (2021) Mulia et al. (2022) Rim et al. (2022) Xu and Wu (2023) Alan et al. (2023) Heidarzadeh et al. (2019)	Immediacy of prediction calculation Model flexibility and applicability	Black-box nature Need a large amount of data sets and computational cost in the training phase
Probabilistic tsunami forecasting	Blaser et al. (2011); Blaser et al. (2012) Tatsumi et al. (2014)	Quantification of uncertainty latent in prediction models and observations	Calibration of model parameter settings Existence of unclear uncertainty, such as tsunami earthquakes and complex dynamics near coasts.

In addition to the abovementioned methods, probabilistic models have great potential in early forecasting due to their ability to quantify uncertainty as discussed by Selva et al. (2021). To date, probabilistic models, also known as probabilistic tsunami hazard assessments (PTHAs), have been used mainly for predisaster risk assessment. Several studies have quantified the relationship between tsunami heights and their exceedance probabilities in a specified period (Annaka et al., 2007; Fukutani et al., 2015; Geist & Parsons, 2006; Heidarzadeh & Kijko, 2011; Kotani et al., 2020; Mori et al., 2018; Omira et al., 2017). It is evident that probabilistic models contribute not only to prior assessment but also to real-time prediction. Although many forecasting methods provide a single outcome, probabilistic forecasting aims to account for the uncertainties latent in tsunami disasters and provide the probability distribution of forecast information. If credibility can be assessed through a reliable probabilistic scheme, better decisions can be made when issuing tsunami warnings.

Among previous studies, it is notable that some studies (Blaser et al., 2011, 2012; Tatsumi et al., 2014) have effectively quantified the probability distribution of predicted tsunami amplitudes via the Bayesian approach. Additionally, Nomura et al. (2022) proposed a sequential Bayesian update scheme for detecting the best-fit scenario in a precomputed database for an actual tsunami event. They demonstrated that the framework could select similar tsunami scenarios from a hypothetical scenario database of Nankai Trough earthquakes and tsunamis (Koshimura & Nomura, 2022). One advantage of such best-fit scenario detection techniques is their feasibility due to their low computational costs and transparency of algorithms. In fact, the Japan Meteorological Agency (JMA) employed scenario-searching schemes (Tatehata, 1997) earlier than such academic studies on best-fit scenario detection (Fauzi & Mizutani, 2020; Gusman et al., 2014). The National Oceanic and Atmospheric Administration (NOAA), which is responsible for most tsunami warnings in the Pacific Ocean, also employed a scenario-searching scheme (Gica, 2008) based on a precomputed tsunami database. In the Mediterranean areas, the Intergovernmental Coordination Group for the Tsunami Early Warning and Mitigation System in the North-Eastern Atlantic, the Mediterranean and Connected Seas (ICG/NEANTWS) manages tsunami warnings and mitigation strategies and uses decision matrices (Intergovernmental Oceanographic Commision, 2023), associating observed earthquake information with tsunami potential. These currently working tsunami warning systems use earthquake information, such as hypocenter location and magnitude, as input data for tsunami forecasting.

Nevertheless, there are limitations to early tsunami forecasting using such best-fit scenario detection methods. That is, these methods assume the existence of a "best-fit" scenario, but there is no guarantee that at least one

scenario in the database reasonably reproduces an actual event. Although the scenario database used by Nomura et al. (2022) assumes a rectangular-shaped uniform slip distribution, the actual slip distribution is heterogeneous. With various tools for generating random seismic ruptures (LeVeque et al., 2017; Nakano et al., 2020; Scala et al., 2020), we could make the scenarios more plausible, but this approach would simultaneously increase the diversity of scenarios in the database. As highlighted by Geist (2002), the complexity resulting from heterogeneous slip patterns leads to significant variations in nearshore wave information, even if the earthquake moment, location, and geometry remain identical. To ensure that the detected scenario is indeed the "best-fit" scenario, ideally, we would need numerous scenarios, numbering in the thousands. However, this may not always be feasible due to constraints on components such as storage or computational resources. To predict an actual event without a prepared similar scenario, it would be more reasonable to create a new scenario by superposing existing scenarios, but it is nontrivial to do this in the best way. Indeed, Nomura et al. (2024) reported the difficulty of forecasting to determine the optimal scenario by superposing scenarios using the relative fitness of each scenario calculated by the previous method (Nomura et al., 2022) as weights.

This study proposes a scenario superposition method for real-time tsunami wave prediction. Similar to the previous best-fit scenario detection method (Nomura et al., 2022), the proposed method has both offline and online phases, namely data preparation and real-time processing phases. The offline phase is similar to that of the previous scenario detection method, where hypothetical tsunami scenarios are prepared and then decomposed into spatial modes and scenario-specific coefficients. In contrast, the online phase is an extension of the previous method and thus a novel contribution of this study. Specifically, our prediction method assumes that waveforms of actual tsunami events can be generated by a linear combination of prepared scenarios' waveforms. Under the assumption, the set of weight parameters that allow for appropriate scenario superposition is identified by Bayesian update with real-time tsunami data obtained at synthetic gauges over time. The probability distribution of the weight parameters is simultaneously estimated and provides information that can be used to understand the prediction error.

To validate the performance and advantages of the proposed method, synthetic experiments targeting the Shikoku shore region of Japan are carried out. A hypothetical scenario database is generated from tsunami simulations of earthquakes triggered by fault rupture in the Nankai Trough subduction zone. Then, 25% of the database scenarios are randomly selected as the test set, and the remaining 75% of the scenarios are used as the training set. In all synthetic experiments, the waveform at the gauge closest to the Shikoku Islands is examined to determine the accuracy of the prediction. To demonstrate the capability of the proposed scenario superposition method, the predicted waveforms are compared with those obtained using the previous sequential Bayesian update method (Nomura et al., 2022), which determines the best-fit scenario from the database.

2. Methodology

2.1. Scenario Superposition

Let us first assume that N_s hypothetical earthquake/tsunami scenarios are available and stored in a database. In this database, each scenario includes synthetic wave heights at N_g synthetic gauges, each comprising N_t snapshot data points as follows:

$$\boldsymbol{X}_j = \begin{bmatrix} \boldsymbol{x}_{t_1}^j & \boldsymbol{x}_{t_2}^j & \cdots & \boldsymbol{x}_{t_{N_t}}{}^j \end{bmatrix} \in \mathbb{R}^{N_g \times N_t}, \tag{1}$$

where the column vector $\boldsymbol{x}_t^j$ stores a snapshot of wave heights at time step t and has N_g components as follows:

$$\boldsymbol{x}_t^j = \begin{Bmatrix} x_{1,t}^j & x_{2,t}^j & \cdots & x_{N_g,t}^j \end{Bmatrix}^{\mathrm{T}} \in \mathbb{R}^{N_g}. \tag{2}$$

In addition, during a real event, the "actual" wave heights are assumed to be observed over time at the same locations as the synthetic gauge points immediately after the tsunamigenic fault rupture and are denoted as

$$\boldsymbol{y}_t = \begin{Bmatrix} y_{1,t} & y_{2,t} & \cdots & y_{N_g,t} \end{Bmatrix}^{\mathrm{T}} \in \mathbb{R}^{N_g}. \tag{3}$$

Following the same rule as in Equation 1, the snapshot vector at each observational time can be stored in one data matrix as follows:

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_{t_1} & \boldsymbol{y}_{t_2} & \cdots & \boldsymbol{y}_{t_{N_t}} \end{bmatrix} \in \mathbb{R}^{N_g \times N_t}. \tag{4}$$

Unlike the previous best-fit scenario detection method (Nomura et al., 2022), this study assumes that the real-time event is not similar to any scenario in the database. Instead, we suppose that the real-time event with the observation $\boldsymbol{Y}$ corresponds to a new scenario whose waveforms are created by superposition, that is, a linear combination of existing scenarios as follows:

$$\boldsymbol{Y} = \sum_{j=1}^{N_s} w^j \boldsymbol{X}_j, \tag{5}$$

where $\boldsymbol{w}$ is a vector whose components are weight parameters w^j. Notably, the weight parameters do not vary depending on time. From the superposition relation in Equation 5, it can be understood that the set of weight parameters is obtained as the solution to an inverse problem with actual and precomputed tsunami wave data.

The ultimate goal is therefore to identify a suitable set of weight parameters for scenario superposition within the shortest possible observation duration T, which is much shorter than the period corresponding to the number of snapshots, N_t. For this purpose, we extend the previous method (Nomura et al., 2022), which consists of two processes: the offline and online phases. As with the previous method, the offline phase should be completed before an actual tsunami occurs. The online phase, on the other hand, markedly improves on the previous method by performing Bayesian estimation as described in the following sections.

Remark 1 Nomura et al. (2024) tried to determine an optimal waveform by superposing scenarios using relative fitness of each scenario as weights. However, this prediction was found to be worse than the best-fit scenario detected by the previous method (Nomura et al., 2022). This is due to the estimation of weight parameters. Specifically, the weights in the previous method are determined based on the similarity that is obtained by comparing each scenario information with the actual event-specific information independently. In other words, the weight parameters were not identified to satisfy the superposition relation in Equation 5 even though the prediction was represented by this relation. In contrast, our prediction method incorporates this relation explicitly into the weight estimation process.

2.2. Offline Phase: Reduced-Order Model (ROM)

As explained in the previous section, we assume that N_s hypothetical tsunami scenarios are available. Combining the information of these scenarios into one matrix, the following training data matrix can be defined:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 & \cdots & \boldsymbol{X}_{N_s} \end{bmatrix} \in \mathbb{R}^{N_g \times (N_t \times N_s)}, \tag{6}$$

which stores the wave height snapshots of all the scenarios for all the time steps in the row direction. This data matrix $\boldsymbol{X}$ can be decomposed by singular value decomposition (SVD) (Klema & Laub, 1980), and a low-rank approximation of $\boldsymbol{X}$ is realized as follows:

$$\boldsymbol{X} = \boldsymbol{\Phi} \boldsymbol{D} \boldsymbol{V}^{\mathrm{T}} \approx \boldsymbol{\Phi}_r \boldsymbol{D}_r \boldsymbol{V}_r^{\mathrm{T}}, \tag{7}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{N_g \times N_g}$ and $\boldsymbol{V} \in \mathbb{R}^{(N_t \times N_s) \times N_g}$ are matrices composed of left and right singular column vectors arranged in the row direction, respectively. Additionally, the left singular matrix $\boldsymbol{\Phi}$ is referred to as the spatial mode matrix because it contains the spatial features latent in the data matrix. The singular values $\sigma_{j \in \{1, 2, \ldots, N_g\}}$ of $\boldsymbol{X}$ are stored in the diagonal entries of matrix $\boldsymbol{D}$ in descending order as follows:

$$\boldsymbol{D} = \begin{bmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_{N_g} \end{bmatrix}, \text{ where } \sigma_1 > \sigma_2 > \cdots > \sigma_{N_g}. \tag{8}$$

Here, $\boldsymbol{D}_r \in \mathbb{R}^{r\times r}$ is a dimensionally reduced matrix of $\boldsymbol{D}$ with only the first r singular values. Similarly, $\boldsymbol{\Phi}_r \in \mathbb{R}^{N_g\times r}$ and $\boldsymbol{V}_r \in \mathbb{R}^{(N_t\times N_s)\times r}$ have been reduced, and they consist of the corresponding r left and right singular column vectors, respectively.

The column vector $\boldsymbol{x}_t^j$ in each $\boldsymbol{X}_t$ in Equation 7 can be expressed as a linear combination with the dimension-reduced mode matrix $\boldsymbol{\Phi}_r$ as follows:

$$\boldsymbol{x}_t^j \approx \boldsymbol{\Phi}_r \boldsymbol{D}_r {\boldsymbol{v}_t^j}^{\mathrm{T}} = \boldsymbol{\Phi}_r \boldsymbol{a}_t^j, \tag{9}$$

where $\boldsymbol{v}_t^j$ is a right singular column vector in $\boldsymbol{V}_r$ that is associated with scenario j at time t and $\boldsymbol{a}_t^j \in \mathbb{R}^r$ is referred to as the set of "scenario coefficients" in this study. Notably, the spatial mode matrix $\boldsymbol{\Phi}_r$ is common to all the time steps of all the scenarios in the database, and the scenario coefficients $\boldsymbol{a}_t^j$ contain dynamic information at all time steps specific to the corresponding scenario. Such procedures of generating a data matrix and applying the SVD are based on previous studies (Fujita et al., 2024; Nomura et al., 2022). Additionally, the substitution of Equation 9 into Equation 5 yields the following expression of the observational wave snapshot:

$$\boldsymbol{y}_t \approx \sum_{j=1}^{N_s} \boldsymbol{\Phi}_r \boldsymbol{a}_t^j w^j. \tag{10}$$

The row-rank approximation through SVD is performed on the data matrix $\boldsymbol{X}$ as in Equation 9 and Equation 10, and the observations $\boldsymbol{y}_t$ are directly used in this scenario superposition relation.

2.3. Online-Phase: Tsunami Forecasting

2.3.1. Bayesian Estimation of Weight Parameters

First, considering an error in the scenario superposition relation, we rewrite Equation 10 as follows:

$$\boldsymbol{y}_t = \sum_{j=1}^{N_s} \boldsymbol{\Phi}_r \boldsymbol{a}_t^j w^j + \varepsilon_t, \text{ where } t \in \{t_1, t_2, \ldots, t_{N_t}\}. \tag{11}$$

Here, the error vector ε_t arises from some potential uncertainties, such as observation noise, ROM representation, the assumption of the superposition in Equation 5, and the assumption in generating hypothetical scenarios. For simplicity, we assume the error vector follows a Gaussian distribution with means $\boldsymbol{0}$ and covariance matrices $\boldsymbol{\Sigma}_\varepsilon$; that is, $\varepsilon_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon)$. For convenience, the element notation of Equation 11 is rewritten in matrix form as follows:

$$\boldsymbol{y}_t = \boldsymbol{\Phi}_r \boldsymbol{A}_t \boldsymbol{w} + \varepsilon_t, \tag{12}$$

where we define the components $\boldsymbol{A}_t$ and $\boldsymbol{w}$ as

$$\boldsymbol{A}_t = \left[\boldsymbol{a}_t^{j=1} \quad \boldsymbol{a}_t^{j=2} \quad \cdots \quad \boldsymbol{a}_t^{j=N_s}\right] \in \mathbb{R}^{r\times N_s}, \tag{13}$$

$$\boldsymbol{w} = \left\{w^{j=1} \quad w^{j=2} \quad \cdots \quad w^{j=N_s}\right\}^{\mathrm{T}} \in \mathbb{R}^{N_s}. \tag{14}$$

From the perspective of the inverse problem, the weight parameters can be regarded as coefficients of a set of the basis functions $\boldsymbol{\Phi}_r \boldsymbol{A}_t$. Also, the mode reduction via SVD decreases the condition number and improves the solution stability of the inverse problem by contracting sufficiently small singular values of the basis function matrix $\boldsymbol{\Phi}_r \boldsymbol{A}_t$; see Appendix A for more information. Additionally, it can be understood that the inverse problem of identifying the set of weight parameters in Equation 12 is regarded as a typical linear regression problem in which the model parameters are identified so that the regression model fits a known data set.

To solve this regression problem with a probabilistic model, we assume that the weight parameters are random variables following a probability density function. According to the Bayesian theorem, the conditional probabilities of the weights $\boldsymbol{w}$, given the real-time observations $\boldsymbol{y}_t$, are as follows:

$$p(\boldsymbol{w}\,|\,\boldsymbol{y}_t) = \frac{p(\boldsymbol{y}_t\,|\,\boldsymbol{w})\,p(\boldsymbol{w})}{p(\boldsymbol{y}_t)}, \tag{15}$$

where $p(\boldsymbol{w}\,|\,\boldsymbol{y})_t$ and $p(\boldsymbol{w})$ are the posterior and prior probability density distributions of the weight parameters, respectively. Additionally, $p(\boldsymbol{y}_t\,|\,\boldsymbol{w})$ is a likelihood function that assesses the fitness of a weight parameter set for the observations $\boldsymbol{y}_t$ and $p(\boldsymbol{y}_t)$ is a normalization factor. This procedure is recognized as a typical Bayesian linear regression problem (Bishop & Nasrabadi, 2006; Yoshida et al., 2021).

According to Equation 12 and the nature of the error vector $\varepsilon_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon)$, the likelihood function $p(\boldsymbol{y}_t\,|\,\boldsymbol{w})$ or, equivalently, the conditional probability of $\boldsymbol{y}_t$ given $\boldsymbol{w}$, is represented as a multivariate normal distribution as follows:

$$p(\boldsymbol{y}_t|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}_t\,|\,\boldsymbol{\Phi}_r\boldsymbol{A}_t\boldsymbol{w},\, \boldsymbol{\Sigma}_\varepsilon). \tag{16}$$

For simplicity of calculating the posterior distribution, we introduce the following prior distribution, which is the conjugate of the Gaussian likelihood function $p(\boldsymbol{y}_t\,|\,\boldsymbol{w})$ with an unknown mean $\boldsymbol{\Phi}_r\boldsymbol{A}_t\boldsymbol{w}$ and known covariance $\boldsymbol{\Sigma}_\varepsilon$:

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_w,\, \boldsymbol{\Sigma}_w). \tag{17}$$

This is referred to as the Gaussian prior distribution with mean $\boldsymbol{\mu}_w$ and covariance $\boldsymbol{\Sigma}_w$. Then, recalling Equation 12, the marginal probability density function of the observations $\boldsymbol{y}_t$ is now defined as

$$p(\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{y}_t|\boldsymbol{\Phi}_r\boldsymbol{A}_t\boldsymbol{\mu}_w,\, \boldsymbol{\Sigma}_\varepsilon + \boldsymbol{\Phi}_r\boldsymbol{A}_t\boldsymbol{\Sigma}_w(\boldsymbol{\Phi}_r\boldsymbol{A}_t)^{\mathrm{T}}). \tag{18}$$

The substitution of Equations 16–18 into the Bayesian theorem in Equation 15 yields the following analytical expression of the posterior distribution:

$$p(\boldsymbol{w}|\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{w}|\hat{\boldsymbol{w}}_t,\, \boldsymbol{P}_t), \tag{19}$$

where.

$$\hat{\boldsymbol{w}}_t = \boldsymbol{\mu}_w + \boldsymbol{P}_t(\boldsymbol{\Phi}_r\boldsymbol{A}_t)^{\mathrm{T}}\boldsymbol{\Sigma}_\varepsilon^{-1}\left(\boldsymbol{y}_t - \boldsymbol{\Phi}_r\boldsymbol{A}_t\boldsymbol{\mu}_w\right), \tag{20}$$

$$\boldsymbol{P}_t = \left(\boldsymbol{\Sigma}_w{}^{-1} + (\boldsymbol{\Phi}_r\boldsymbol{A}_t)^{\mathrm{T}}\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi}_r\boldsymbol{A}_t\right)^{-1}. \tag{21}$$

For a detailed derivation of this equation, see Yoshida et al. (2018); Yoshida et al. (2021); Gelman et al. (2013); Bishop and Nasrabadi (2006). Thus, the proposed method sequentially estimates the expected value of the weight parameter set $\hat{\boldsymbol{w}}_t$ using the Bayesian estimation. At the same time, the posterior covariance matrix $\boldsymbol{P}_t$ associated with the reliability of the estimated weight parameters can be obtained. It should be noted here that the weight parameters are assumed to follow a Gaussian distribution to analytically and quickly estimate the posterior distribution.

2.3.2. Procedure of Sequential Updates

This section describes the specific procedure of sequential updating for the weight parameter set $\hat{\boldsymbol{w}}_t$ using the observations $\boldsymbol{y}_t$ that are measured at N_g offshore gauges. The entire procedure is summarized in Algorithm 1.

It is assumed that prior to the online phase, both the spatial mode matrix $\boldsymbol{\Phi}_\mathbf{r}$ and scenario coefficient matrix $\boldsymbol{A}_t$ have been prepared as the results of the offline phase. At the beginning of the online phase, the initial

observational wave snapshot $\boldsymbol{y}_{t_1}$ is obtained via measurements at N_g offshore gauges. Then, the covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ of the error vector ε_t is set as

$$\boldsymbol{\Sigma}_\varepsilon = \boldsymbol{I}, \tag{22}$$

where $\boldsymbol{I}$ is the identify matrix.

For weight parameter estimation, the Bayesian approach defined in Equations 16–18 (Lines 5–6 in Algorithm 1) requires two hyperparameters, $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\mu}_w$, to determine the form of the initial distribution of the prior probability $p(\boldsymbol{w})$. A possible method of identifying the optimal hyperparameters according to observed data was devised with the hierarchical Bayesian model (Gelman et al., 2013), which adaptively determines their components by referring to another probability distribution. However, the resulting process will be computationally expensive since analytical calculations are no longer possible, and there is no prior information on each scenario's likelihood at the initial step. Therefore, in this study, fixed values are determined for these parameters such that

$$\boldsymbol{\mu}_w = \boldsymbol{0}, \quad \boldsymbol{\Sigma}_w = \alpha \boldsymbol{I}. \tag{23}$$

Here, α is a parameter for defining an initial distribution range of weight parameters and set to 0.01 by trial calculations in this study. That is, both the mean and covariance are assumed to be uniform for all N_s scenarios.

By using the initial conditions above, the weight parameter set $\hat{\boldsymbol{w}}_t$ can be sequentially estimated as summarized as the loop in Lines 3–9 in Algorithm 1. Finally, at time step T, the scenario superposition as described in Equation 10 provides a snapshot of the wave height prediction for all time steps, that is, waveform prediction, as

$$\hat{\boldsymbol{y}}_t = \boldsymbol{\Phi}_r \boldsymbol{A}_t \hat{\boldsymbol{w}}_T. \tag{24}$$

Here, the covariance matrix of the prediction error can be estimated as

$$\boldsymbol{R}_t = \boldsymbol{\Phi}_r \boldsymbol{A}_t \boldsymbol{P}_T (\boldsymbol{\Phi}_r \boldsymbol{A}_t)^{\mathrm{T}}. \tag{25}$$

It should be emphasized that the predicted waveforms $\hat{\boldsymbol{y}}_t$ and their covariances $\boldsymbol{R}_t$ can be obtained for all time steps, $t \in \{t_1, t_2, \ldots, N_t\}$, within a limited observation duration T, which is expected to be much shorter than the observation time, as $T \ll t_{N_t}$.

Algorithm 1. Online Phase: Estimation of Weight Parameters

Require: Spatial mode matrix $\boldsymbol{\Phi}_r \in \mathbb{R}^{N_g \times r}$ and coefficient matrix $\boldsymbol{A}_{t \in \{1, 2, \ldots, T\}} \in \mathbb{R}^{r \times N_s}$.
Require: Observation vector of wave heights $\boldsymbol{y}_{t \in \{1, 2, \ldots, T\}}$.
1: $\boldsymbol{\Sigma}_\varepsilon = \boldsymbol{I}$.
2: $\boldsymbol{\mu}_w = \boldsymbol{0}$; $\boldsymbol{\Sigma}_w = 0.01\boldsymbol{I}$.
3: **for** $t = 1, \ldots, T \ll N_t$ **do**
4: // Lines 5–8 estimate the joint probability distribution of the weight parameters.
5: $\boldsymbol{P}_t = \left(\boldsymbol{\Sigma}_w^{-1} + (\boldsymbol{\Phi}_r \boldsymbol{A}_t)^{\mathrm{T}} \boldsymbol{\Sigma}_\varepsilon^{-1} \boldsymbol{\Phi}_r \boldsymbol{A}_t\right)^{-1}$
6: $\hat{\boldsymbol{w}}_t = \boldsymbol{\mu}_w + \boldsymbol{P}_t (\boldsymbol{\Phi}_r \boldsymbol{A}_t)^{\mathrm{T}} \boldsymbol{\Sigma}_\varepsilon^{-1} \left(\boldsymbol{y}_t - \boldsymbol{\Phi}_r \boldsymbol{A}_t \boldsymbol{\mu}_w\right)$
7: $\boldsymbol{\Sigma}_w \leftarrow \boldsymbol{P}_t$
8: $\boldsymbol{\mu}_w \leftarrow \hat{\boldsymbol{w}}_t$
9: **end for**
10: // Lines 11–12 generate waveform prediction.
11: $\hat{\boldsymbol{y}}_{t \in \{1, 2, \ldots, N_t\}} = \boldsymbol{\Phi}_r \boldsymbol{A}_t \hat{\boldsymbol{w}}_T$
12: $\boldsymbol{R}_{t \in \{1, 2, \ldots, N_t\}} = \boldsymbol{\Phi}_r \boldsymbol{A}_t \boldsymbol{P}_T (\boldsymbol{\Phi}_r \boldsymbol{A}_t)^{\mathrm{T}}$
13: **return** $\left(\hat{\boldsymbol{y}}_t, \boldsymbol{R}_t\right)$

3. Synthetic Experiments Targeting the Nankai Subduction Zone

The data matrix $\boldsymbol{X}$ and the observational snapshot $\boldsymbol{y}_t$ are obtained from numerical simulation and used as input for the offline and online phases instead of using the actual tsunami data according to Mulia and Satake (2021). In the following, we refer to $\boldsymbol{X}$ and $\boldsymbol{y}_t$ as "training" and "test" data, respectively, for convenience. To confirm the improved accuracy of the proposed method over the previous best-fit scenario detection-based method (Nomura et al., 2022), the same scenario database is used by both methods in the offline phase of the synthetic experiment, where reduced-order modeling by singular value decomposition is performed. In what follows, we refer to the results obtained using the previous method as N.2022. The online phases are different for each prediction method.

3.1. Target Area

Synthetic experiments can be effective for sites where there is a future tsunami risk but historical data are poor or unavailable. In this context, we target the Shikoku region of Japan, which is close to the Nankai Trough subduction zone, consisting of three main fault sources: Tokai, Tonankai, and Nankai (Ando, 1975, 1982). If the three source faults rupture simultaneously, the triggered tsunami is assumed to arrive on the Pacific coast of western Japan within a few minutes. Although the earthquakes triggered by these faults are known to occur periodically, there is no available tsunami evidence even for the most recent event in 1946 (Baba et al., 2002; Cummins et al., 2002). In addition, considering that the estimated damage can reach 200,000 casualties at most (Cabinet Office Japan, 2019), synthetic experiments targeting this region are also worth performing from a disaster mitigation perspective.

3.2. Synthetic Tsunami Scenarios With Hypothetical Earthquake Events

As shown in the left panel of Figure 1, the Nankai subduction zone is first discretized into 1,119 triangular subfaults assuming Slab 2 geometry (Hayes et al., 2018). Then, MudPy software (Melgar, 2020) implementing the Karhunen-Loève (KL) expansion (LeVeque et al., 2016) is used to calculate a variety of random lognormal slip distributions with a specified covariance matrix for the slip on each subfault relative to other subfaults. The parameters used are based on observed correlation lengths from past large subduction zone earthquakes (Melgar & Hayes, 2019). In addition, the other rupture parameters, such as strike, dip, rake angle, and depth, were also added to each subfault. See LeVeque et al. (2024) for the specific setting of each paramter and subfault location. In this work, we assume that this provides a reasonable probability distribution for sampling possible earthquakes to obtain the training data.

Using the created slip distribution patterns as input data, the time-dependent seafloor displacements (Comninou & Dundurs, 1975) are calculated with GeoClaw software (Clawpack Development Team, 2021). As an example, the slip distribution and seafloor deformation of the Mw 9.1 event (#2274) are shown in Figure 1. Using the seafloor displacements as input data, this software enables us to simulate the resultant tsunami wave propagation for 3 hr. The governing equations employed for GeoClaw are the nonlinear shallow water equations, and it implements the adaptive mesh refinement technique. The etopo1 data released by NOAA (Amante & Eakins, 2009) is used for the topography and bathymetry data in the simulation target area. In order to enrich the training and test data, we prepared 2,342 scenarios of synthetic Nankai Trough earthquakes/tsunamis with magnitudes ranging from Mw 8.1 to 9.1. Specifically, the numbers of scenarios are 400, 400, 399, 394, 392, and 357 for magnitudes 8.1, 8.3, 8.5, 8.7, 8.9, and 9.1.

In each tsunami simulation, synthetic gauges are located at selected calculation points to record synthetic tsunami wave sequences. The number of synthetic gauges is $N_g = 62$, and their locations are shown in Figure 2; they were selected to be identical to the actual observational instruments near Shikoku Island, such as the Dense Oceanfloor Network System for Earthquakes and Tsunamis (DONET) (Kaneda, 2010) and the Nationwide Ocean Wave information network for Ports and Harbors (NOWPHAS) (Ports and Harbours Bureau, 2022). The observation network of DONET measures in situ water pressures, which can be converted to wave heights via the hydrostatic pressure, and its gauges are connected with seafloor cables around depths of 1,000–5,000 m. In contrast, gauges of the NOWPHAS network directly observe wave heights by ultrasonic wave gauges or Global Positioning System (GPS) buoys. They are installed relatively close to the coast, and the water depths are up to 500 m. The previous study (Nomura et al., 2022), which also focused on the tsunami prediction in the Shikoku region, considered 71 gauges. The gauge configuration consisted of 47 virtual gauges as well as 24 actual gauges of DONET and NOWPHAS. In contrast, we set the actual 62 gauges selected from DONET and NOWPHAS and eliminate virtual

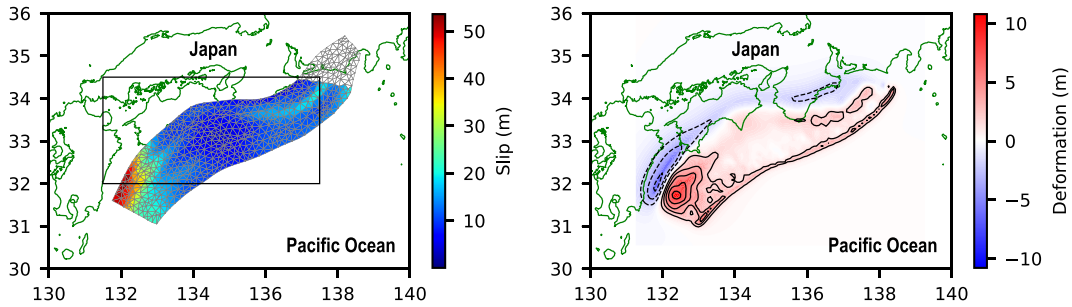

Figure 1. Left: Slip distribution of sample scenario #2274 (Mw 9.1) on the projected triangular subfaults of the Slab 2.0 model of Japan. The black rectangle shows a region in which synthetic observational gauges are installed as in Figure 2. Right: Final seafloor deformation for scenario #2274 simulated with the Okada model.

gauges to consider a more practical condition. Gauge A in the figure is one of the synthetic gauges used for observation and is selected as a target point for tsunami waveform prediction.

For validation purposes, 2,342 hypothetical earthquake/tsunami scenarios are divided into test and training data sets. Specifically, 586 scenarios, which correspond to 25% of the database, are randomly selected as test data, and the remaining 1,756 scenarios are used as training data. In summary, each training scenario data matrix $\boldsymbol{X}_{j\in\{1,2,\ldots,1756\}}$ has dimensions of $N_g \times N_t = 62 \times 2,160$, where N_t corresponds to $3\,[\mathrm{hr}] \times 3,600\,[\mathrm{sec/hr}]/5\,[\mathrm{sec}]$ in real time. Thus, the dimension of the data matrix $\boldsymbol{X} \in \mathbb{R}^{N_g \times (N_t \times N_s)}$ are $62 \times (2,160 \times 2,342)$. Additionally, there are 586 matrices of test scenarios, and each matrix $\boldsymbol{Y}_{i\in\{1,2,\ldots,586\}}$ has dimensions of $N_g \times N_t (= 62 \times 2,160)$.

3.3. Offline Phase

As explained in Section 2.2, the offline phase is carried out to extract the mode matrix $\boldsymbol{\Phi}_r$ and the coefficient matrix $\boldsymbol{A}_t$ from the training data matrix $\boldsymbol{X}$ via singular value decomposition. After decomposing the matrix $\boldsymbol{X}$, we need to determine the number of modes r needed to obtain the approximation in Equation 7. For this purpose, the contribution rate c_j of mode j is defined as

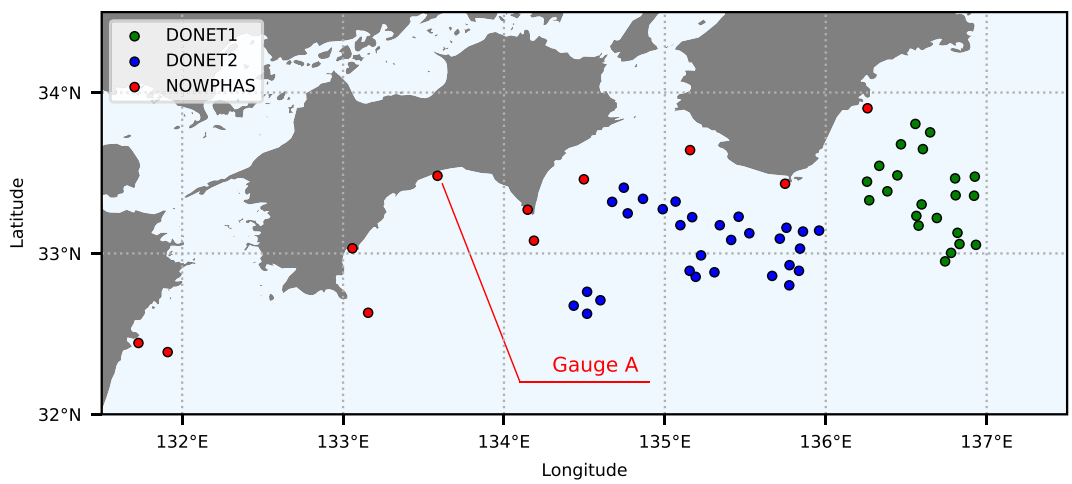

Figure 2. Synthetic gauge configuration in the targeted Shikoku shore region. The gauge locations are identical to those of the existing gauges DONET1, DONET2, and NOWPHAS as illustrated by the different colored circles. Gauge A is the target gauge for waveform prediction, and the training and test data of wave snapshots, denoted as $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, are sampled at all synthetic gauges, including Gauge A.

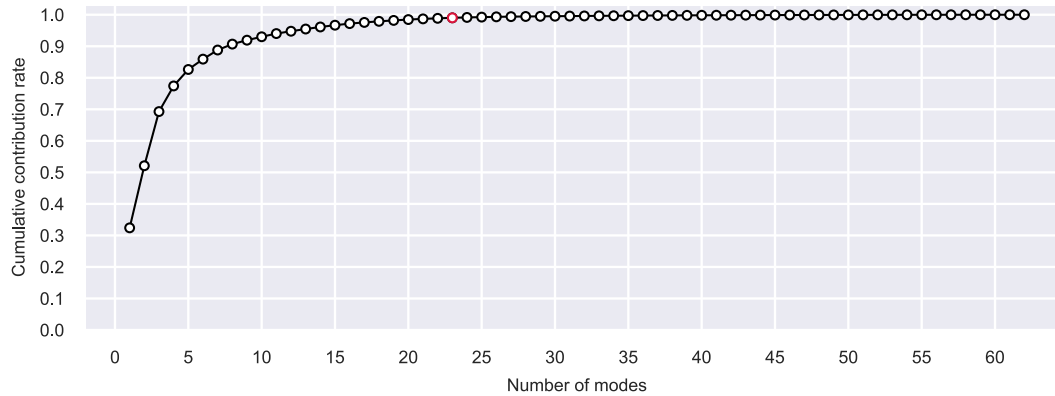

Figure 3. Cumulative contribution rate $\sum_{j=1}^{r} c_j$, where c_j is the contribution rate of mode j, defined as Equation 26. The red circle represents the 99% cumulative contributions at $r = 23$.

$$c_j = \frac{\lambda_j}{\sum_{i=1}^{n} \lambda_i} = \frac{\sigma_j^2}{\sum_{i=1}^{n} \sigma_i^2}, \tag{26}$$

where λ_i is the i-th eigenvalue of a covariance matrix $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ and is equal to the square of the corresponding singular value σ_i.

The obtained contribution rates are used to calculate the cumulative contribution rate $\sum_{j=1}^{r} c_j$, which generally expresses the ROM performance of $\boldsymbol{\Phi}_r \in \mathbb{R}^{N_g \times r}$ for the original data. Figure 3 shows the variation of the contribution rate with the number of modes. From this figure and specific values that are not given here, it is found that the cumulative contribution of the top 23 modes reaches 99% of the original data, which have a total of 62 $(=N_g)$ modes. In other words, the higher modes, from 24 to 62, cumulatively have less than 1% of the information and can be regarded as noise. Therefore, we employ $r = 23$ modes for the ROM representation in Equation 7 and for the online phases of both prediction methods.

3.4. Online Phase

During the online phase, the Bayesian estimation is carried out to evaluate the appropriate weight parameter set. As summarized on Lines 1–2 of Algorithm 1, the initial condition of the probability distribution is first set according to Equations 16–18. Figure 4 shows the initial probability density distribution of the weight parameters. Here, the contour plot shows the joint probability of the weight parameters for the training scenarios, and the two graphs of the sides show the probability of each weight parameter. Although only two training scenarios are sampled to show the initial probability density distribution in the two-dimensional space, the same initial distribution is given for all 1,756 training scenarios. As seen from these graphs and described in Section 2.3.2, the prior information on the weight parameters is characterized by a Gaussian distribution with zero means and covariance $0.01\boldsymbol{I}$.

With the observed wave heights $\boldsymbol{y}_t$, the weight parameter set is sequentially estimated at each time step as described in Equations 16–18. Figure 5 shows the variation in the distribution of the joint probability of $w_{j=1478}$ and $w_{j=1526}$ for test scenario #136 and that of $w_{j=2214}$ and $w_{j=2231}$ for test scenario #2274. Here, these two specific test events (#136 and #2274) have small and large amplitudes, respectively, and are selected from 586 test data to show the prediction performance. These weight parameters for each test scenario are the largest and second-largest components in $\hat{\boldsymbol{w}}_{T=600}$ and are commonly used as axes for all three panels, respectively. As seen from the figures, the probability density varies as the sequential Bayesian update proceeds. Although the density distribution of test scenario #136 does not significantly change, that of test scenario #2214 notably changes. More specifically, the posterior means, corresponding to the darkest red zone, of $w_{j=1526}$ and $w_{j=1478}$ tend to stay near $(0, 0)$, whereas the means of $w_{j=2214}$ and $w_{j=2231}$ are likely to reach approximately 0.15–0.20 at $T = 600$ s.

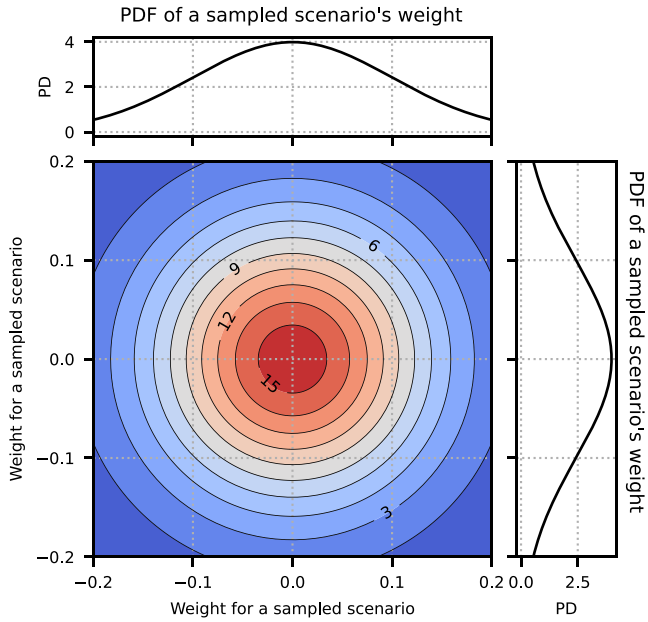

Figure 4. Initial prior distribution $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$ projected on a two-dimensional contour plot. The axes refer to the weights of two scenarios selected from the 1,756 training scenarios. The contour lines show the joint probability density distribution, and the upper and right side graphs show the probability density function for each scenario.

Moreover, both distributions tend to shrink compared with the initial distributions shown in Figure 4. Although only weights ranging from zero to 0.4 are shown in Figure 5, negative and larger positive values are also recorded for all 1,756 training scenarios.

Figure 6 shows the histograms of the mean values of the estimated weight parameters $\hat{\boldsymbol{w}}_{T=600}$ for test scenarios #136 and #2274. The initial mean values were set to zero for all the training scenarios, and the posterior weights at $T = 600$ were also distributed around zero. Additionally, the range of the weight values of Scenario #2274 is 10 times larger than that of Scenario #136, implying that the former spreads more widely than the latter. This tendency corresponds with the probability density variation shown in Figure 5, where the two largest posterior estimates, $\hat{w}_{j=2214}$ and $\hat{w}_{j=2231}$, of Scenario #2274 are much farther from zero than those of Scenario #136.

3.5. Tsunami Forecasting

Using the set of estimated weight parameters, $\hat{\boldsymbol{w}}_T$, obtained at an arbitrary observation time in the online phase, future waveforms at $t = (T, t_{N_t})$ that have not been observed can be predicted for an arbitrary test scenario by the superposition of training data for tsunami forecasting. For example, Figure 7 shows the superposed waveforms $\hat{\boldsymbol{w}}_T$ at Gauge A for Scenarios #136 (Mw 8.1) and #2274 (Mw 9.1) and their standard deviations according to Equations 16–18. As seen from this figure, the superposition or, equivalently, the linear combination with the estimated weight parameters effectively traces

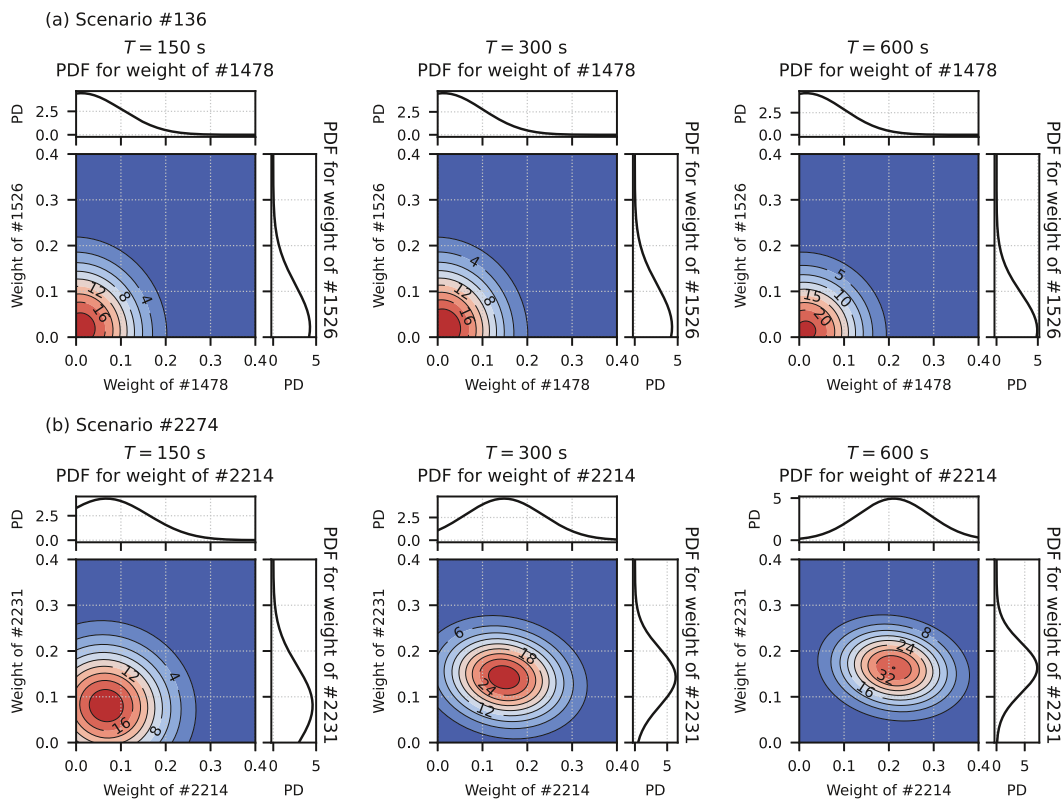

Figure 5. Bayesian estimates of weight parameters at representative time steps, $T = 150, 300, 600$ s. (a) Results for test scenario #136; (b) Results for test scenario #2274. Each contour plot shows the joint probability of the weight parameters that correspond to the largest and second-largest posterior estimates at $T = 600$ s.

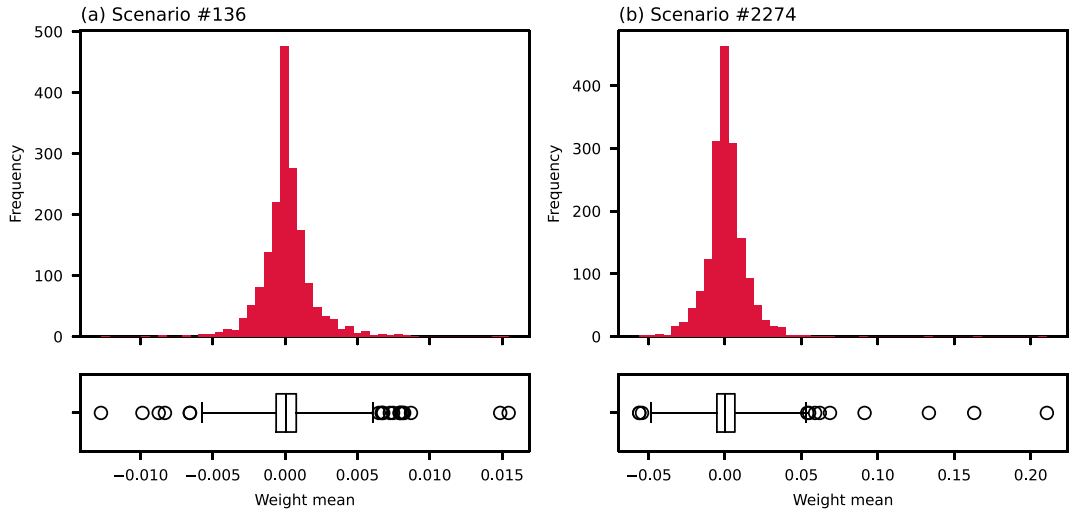

Figure 6. Distribution of the estimated weights for test scenarios #136 and #2274 with an observation duration of $T = 600$ s. The weight distribution is summarized as a histogram in the upper panel and a boxplot in the lower panel.

the trend of future waveforms in the range from the green vertical line to the right. The longer the observation duration is, the better the performance in both scenarios, as a matter of course. This figure also shows a comparison with the conventional method (N.2022) for different observation duration periods ($T = 150, 600$ s) to confirm the advantages of this method. The waveforms that are superpositions based on all training scenarios provide much better predictions for both observations than those of the most likely scenario detected using N.2022.

Another advantage of the proposed method is that treating the weight parameters as random variables allows us to refer to their probability distributions. The pink bands around the predicted waveforms in Figure 7 indicate the confidence intervals of $\pm 1\sigma$ based on the covariance matrix $\boldsymbol{R}_t$ of the superposing prediction. The true waveforms are within the pink bands even when they are not perfectly traced by the predicted waveforms. This suggests that the proposed method has great potential for robust decision-making in tsunami forecasting since it prevents the

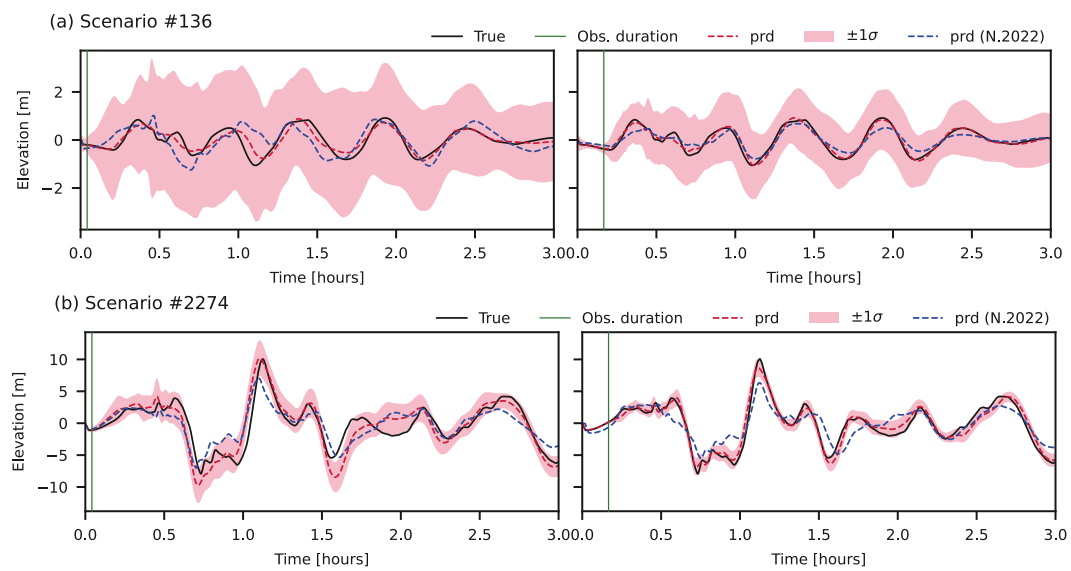

Figure 7. Predicted (superposed) waveforms at Gauge A and their standard deviations for two sampled test scenarios in comparison with the true waveforms: (a) test scenario of #136 (Mw 8.1); (b) test scenario of # 2274 (Mw 9.1). The waveforms are obtained by the proposed and previous (N.2022) methods using wave heights measured at all synthetic gauges during observation durations $T = 150$ s (left panels) and 600 s (right panels).

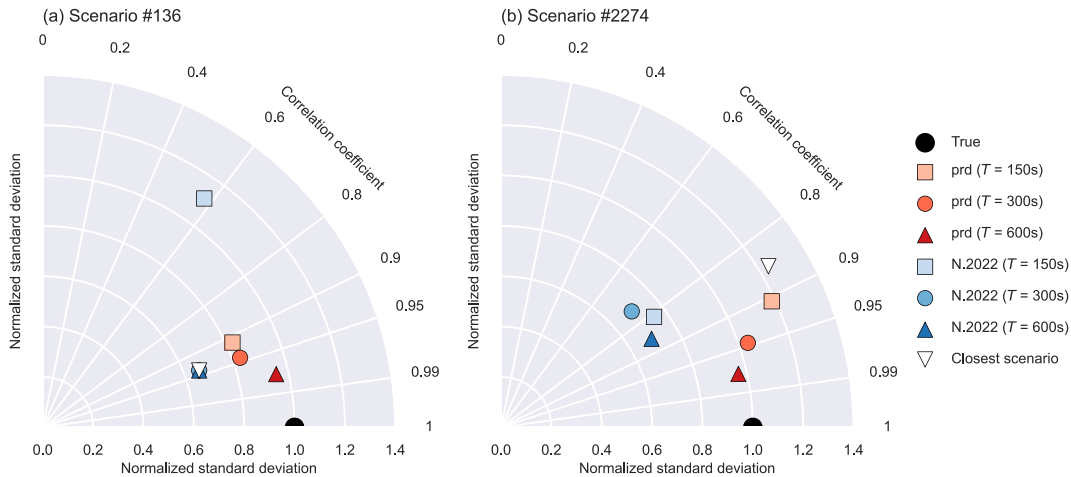

Figure 8. Taylor diagram summarizing the accuracy of the predicted waveform at Gauge (a) (a) Scenario #136; (b) Scenario #2274. The black marker shows the true value; the orange, red, and dark red markers show the waveforms superposed by the proposed method; and the light blue, blue, and dark blue markers show the waveforms detected by the previous method (N.2022). The rectangular, circular, and triangular shapes indicate the results for three different observation durations, $T = 150$ s, 300 and 600 s, respectively. The inverted white triangle in each panel corresponds to the waveform at Gauge A for the closest scenario that satisfies Equation 27.

issuance of tsunami warnings based on unreliable forecasts. In other words, such probabilistic information provided by the proposed method is expected to minimize the possibility of underestimating tsunamis by referring only to predicted values.

Despite these promising results, further efforts must be made to address the relatively low reliability of predictions for earthquakes of smaller magnitudes. The range of the pink bands in Figure 7a is large compared to the wave amplitudes in Scenario #136, which originated from an Mw 8.1 earthquake. Indeed, the confidence interval of $\pm 1\sigma$ exceeds five m, almost three times the maximum wave amplitudes for this scenario. On the other hand, the confidence interval for Scenario #2274, with a Mw 9.1 earthquake, is only half of the maximum wave height. These results are attributed to the fact that the considered uncertainties, that is, the error vector ε_t and the initial distribution of the weight parameters, $p(\boldsymbol{w})$, are set uniformly for all the test scenarios. Thus, there is room to solve this problem by adaptively setting the prior distribution of the weight parameters instead of using the Gaussian distribution employed in this study and explicitly considering other sources of uncertainties.

To more quantitatively compare the prediction results of the proposed and previous methods, Taylor diagrams (Taylor, 2001) are created for each test case as shown in Figure 8. The advantage of the Taylor diagram is that it provides a comprehensive visualization of the similarities between two time series. The fan-shaped single diagram summarizes three evaluation indices: the normalized standard deviation, correlation coefficient, and Euclidian distance (root-mean-squared error). In this figure, the black marker on the horizontal axis represents the test case itself, that is, the true value. The polar axis shows the normalized standard deviation and represents the scale difference in the wave amplitude from the test data. Additionally, the polar angle indicates the correlation coefficient between the two referenced time series data. If a predicted waveform at Gauge A has a similar upward/downward trend to the test data, the resulting point is plotted at a smaller counterclockwise angle from the horizontal axis. In addition, the radial distance from the black marker represents the root-mean-squared error of the predicted waveform at Gauge A. Thus, in simple terms, the closer the plot is to the black marker, the more accurate the predicted waveform is. Since the proposed and N.2022 methods predict waveforms of the entire time series, prediction errors can occur within the observation window. To fairly evaluate such errors, the three indices are computed using the entire time series, including observation periods.

In both Taylor diagrams for different durations, the dark red, red, and orange markers, which correspond to the waveforms predicted as superpositions of the training data by the proposed method, are closer to the black marker than the light blue, blue, and dark blue makers are, respectively, which are predicted by N.2022. In addition, the proposed method increases the accuracy of the results for longer observation periods as indicated by the circle and

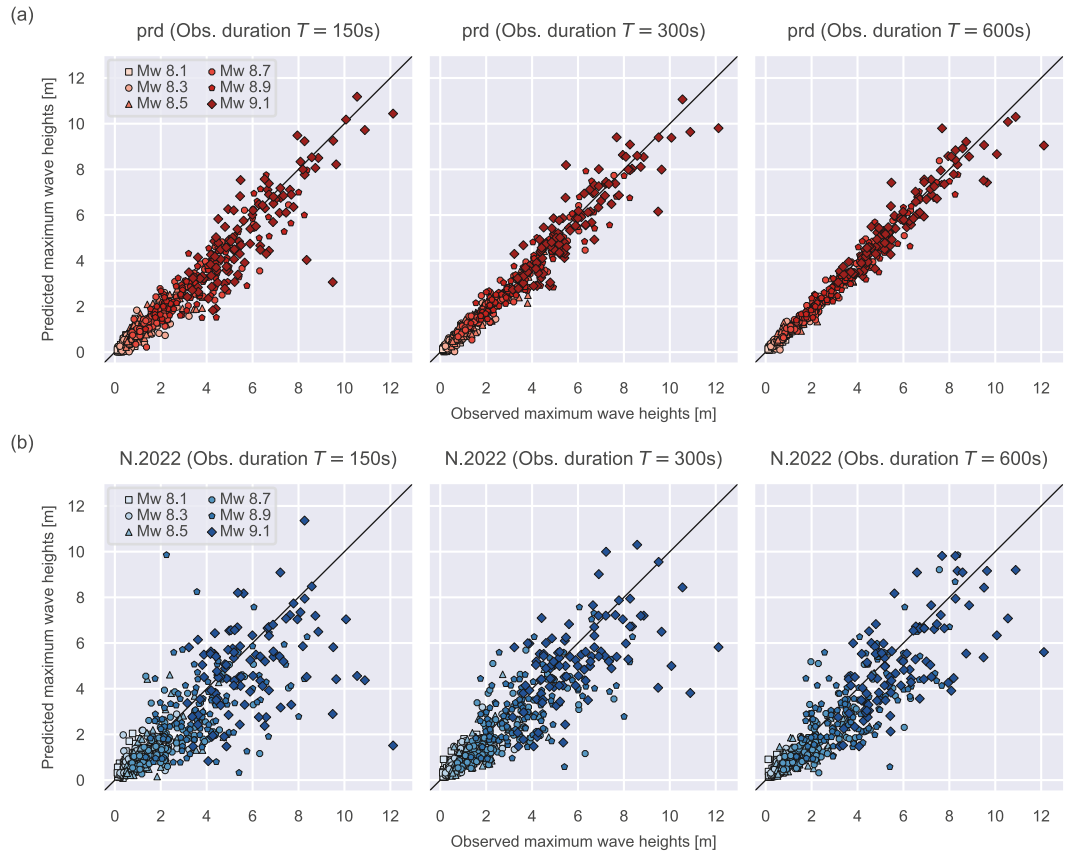

Figure 9. Comparison between the observed and predicted maximum wave heights for all the test scenarios using the proposed method and the method of N.2022. (a) Proposed method; (b) method of N.2022. The observation durations are $T = 150$, 300 and 600 s from left to right.

triangle markers. In contrast, improvements are not necessarily visible for N.2022 since the blue plots showing the results of the previous method are close to each other.

To quantitatively evaluate whether the achieved accuracy is acceptable, we extract and compare the scenario with the most similar waveform in the training database. In Figure 8, the inverted white triangle in each panel corresponds to the waveform at Gauge A for the closest scenario j, which satisfies the following relationship:

$$j = \underset{j \in \{1, \ldots, N_s\}}{\operatorname{argmin}} \|\boldsymbol{X}_j - \boldsymbol{Y}\|_{\mathrm{F}}, \tag{27}$$

where $\| \bullet \|_{\mathrm{F}}$ is the Frobenius norm for all 3 hr wave heights. The extracted scenario j is located much farther from the true plot (black marker) than the blue and red triangles for both test cases. This result means that the proposed method is not only superior to the previous method but also makes more accurate predictions than are possible with the straightforward scenario search method using complete observation data. It is also concluded that the scenario superposition approach proposed in this study has a strong advantage for tsunami waveform prediction even when a scenario similar to the actual event does not exist.

To further confirm the advantage of the present approach, we compare the predicted results with the true values in terms of the maximum wave height at Gauge A for all 586 test cases. Figure 9 shows the scatter plots for this comparison. Similar to the results of the Taylor diagram, the plots approach the diagonal as the observation duration increases for the proposed method, whereas they remain scattered even for $T = 600$ s for the previous method as shown in Figure 9b. In particular, the large-magnitude scenarios represented by the diamond-shaped markers (Mw 9.1), which usually result in larger wave amplitudes, are over/underestimated in Figure 9b. Such nonnegligible prediction errors are eliminated by the proposed method given sufficient observational duration;

see Figure 9a. In addition to the superiority over the previous work (N.2022), comparison studies with other prediction schemes would provide a comprehensive understanding of the prediction performance of the proposed method, and therefore, such a study will be conducted in the future.

Remarks 2 From the above demonstration of tsunami waveform prediction, we argue that the proposed scenario superposing scheme is superior over the previous one (Nomura et al., 2022). However, since the weight parameter estimation is based on the regression problem, the prediction may lose its reliability if the solution of the inverse problem described in Subsection 2.3 is unstable. For example, when two or more training scenarios have similar waveforms at all synthetic gauges, the solution may be indeterminate because two or more basis functions become almost identical. This leads to the existence of multiple possible sets of weight parameters and makes their probability distribution have a very large variance. As another example, if the number of training scenarios is insufficient, appropriate solutions may not be available, and an accurate prediction will not be achieved. It should be noted that this study assumes that cases like these examples do not occur. Guaranteeing prediction accuracy in such cases where the solution is unstable will be a future challenge.

4. Conclusion

In this study, we proposed a tsunami scenario superposition method that predicted an actual event by a linear combination of all the scenarios in the database with appropriate weight parameters. The Bayesian approach was used to search for the best weight parameter set as the solution of a linear regression problem.

Specifically, the scenario coefficients and spatial modes were extracted by singular value decomposition (SVD) in the offline phase and used in the online phase to model the observational wave snapshots by a linear combination. In the online phase, the set of suitable weight parameters for scenario superposition was found as the solution to the Bayesian regression problem with real-time tsunami data obtained at the synthetic gauge over time.

Synthetic experiments targeting the Nankai subduction zone were performed to validate the advantages of the proposed method. In the experiments, hypothetical earthquake and tsunami scenarios were first generated, and their wave sequences were recorded through a series of numerical simulations. Then, the proposed and our previous methods were tested on the synthetic data for comparison. The results showed that the proposed method performed better than the previous method in predicting the waveform near the Shikoku coast and provided more accurate predictions with longer observation durations. In addition, the predicted waveforms of the proposed method were more accurate than those of the most similar scenario detected using all 3 hr wave data. This fact indicates the advantage of scenario superposition in situations where the closest scenario in the database is not similar to the actual event. Accurate predictions of the maximum wave heights were also made for all test cases within a few hundred seconds, such as 150, 300, and 600 s. In addition, the treatment of the weight parameters as random variables with the Bayesian approach allowed us to refer to the confidence intervals of the predicted waveforms. These confidence intervals are assumed to be beneficial for the local community in determining tsunami mitigation with minimal uncertainty.

Overall, we argue that our extended method for real-time tsunami prediction has advantages over the previous scenario detection method in terms of tsunami mitigation. More appropriate prior probability distribution settings have been mentioned as a topic to be investigated in the future to address the large uncertainties relative to small tsunami wave amplitudes. Another probabilistic regression model, such as Bayesian Lasso, may also be a potential tool for improving the probability distribution settings. Also, we need to explicitly define a variety of uncertainties caused by assumptions in prediction models and hypothetical scenario generation to quantify the uncertainty inherent in forecasts. Once these remaining tasks are completed, early warning of a nearshore tsunami by scenario superposition will be a more feasible and promising tool for tsunami-prone communities such as those in the Shikoku region of Japan.

Appendix A: Effect of Singular Value Decomposition in Scenario Superposition

The effect of the low-rank approximation via SVD in Equation 10 and Equation 12 is tested. The low-rank approximation reduces the condition number of the matrix $\boldsymbol{X}$ and induces a robust prediction model to observation noises. To test the prediction with noise conditions, we first generated 1,000 cases of noise-added test data, considering test scenario #2274, which is the same scenario tested in Section 3.4. Specifically, synthetic Gaussian

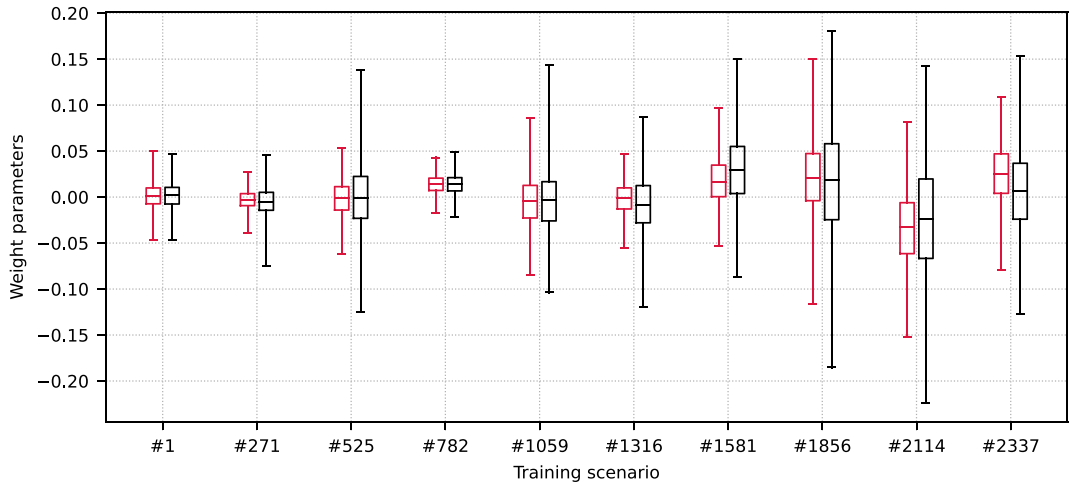

Figure A1. Comparison of the estimated weight parameters between the tsunami predictions with SVD (red) and without SVD (black). The weight parameter estimation is conducted with 1,000 cases of synthetic noise conditions.

noise with mean zero and covariance $\boldsymbol{I}$ were added to the original test data matrix $\boldsymbol{Y}$. Using the noise-added observations as input, we carry out two cases of tsunami predictions: w/SVD and w/o SVD.

Figure A1 shows the variety of estimated weight values of sampled training scenarios through 1,000 case predictions with and without SVD conditions. Generally, the weights estimated without SVD are distributed more widely, which means the weight estimation is more sensitive to input noise. Figure A2 shows the waveform prediction results at Gauge A (Figure 2) through 1,000 cases. The mean values of the results, shown with black triangles, show almost similar accuracy. It should be noted that only slight differences between the results with and without SVD can be seen since the Bayesian approach is used to estimate a weight parameter set. The Bayesian linear regression with a Gaussian prior, used in this study, is known to be equivalent to linear regression with l2 regularization. Therefore, the prediction through the Bayesian linear regression is robust to input noise, and the application of SVD leads to a slight improvement.

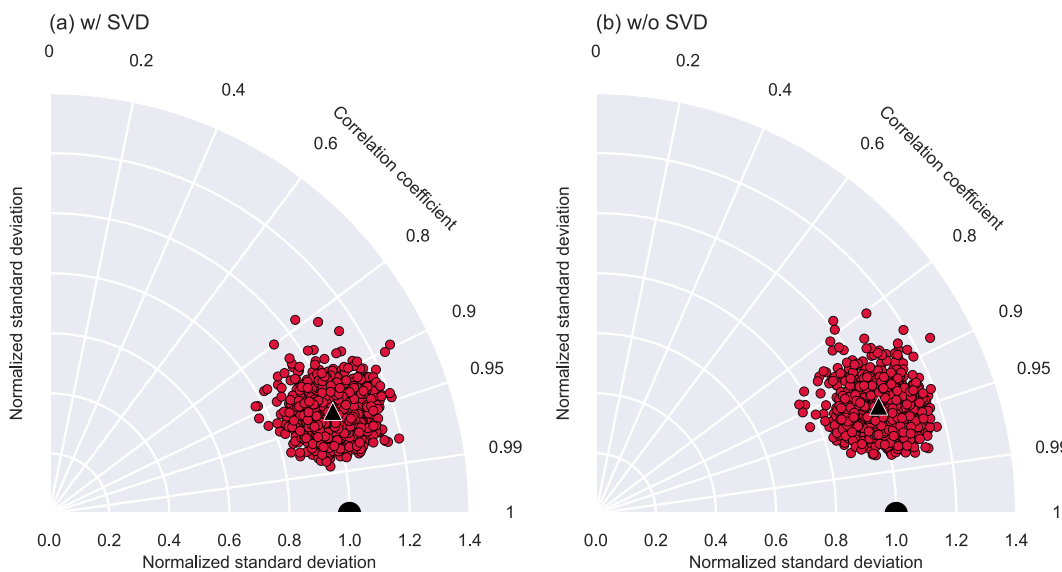

Figure A2. Comparison of the prediction results at Gauge A with and without SVD. The red circles show the prediction results of 1,000 cases with synthetic noise conditions, and the black triangles show the mean values of 1,000 predictions.

Data Availability Statement

The python codes used for all numerical experiments in this work are available at https://doi.org/10.5281/zenodo.14029637 (Fujita, 2024). All synthetic wave sequences of 2342 hypothetical scenarios, and the locations of the synthetic gauges are archived in https://doi.org/10.5281/zenodo.12696847 (LeVeque et al., 2024).

- Fujita, S. (2024). Python codes for tsunami prediction with the scenario superposition (Version v1.0.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.14029637.
- LeVeque, R., Nomura, R., & Fujita, S. (2024). Synthetic data set of 2342 earthquake/tsunami scenarios targeting the Nankai trough subduction zone (Version v0.0.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.12696848.

Acknowledgments
The authors gratefully acknowledge Diego Melgar for assistance in developing the KL expansion used in this study. We also thank Donsub Rim for useful discussions and for helping to generate and organize some of the data used in this study. This work was partially supported by JST, the establishment of university fellowships toward the creation of science technology innovation, Grant JPMJFS2102, by JSPS KAKENHI Grant JP23KJ0112, and by the Core Research Cluster of Disaster Science at Tohoku University (Designated National University).

References

Alan, A. R., Bayındır, C., Ozaydin, F., & Altintas, A. A. (2023). The predictability of the 30 october 2020 İzmir-samos tsunami hydrodynamics and enhancement of its early warning time by lstm deep learning network. *Water*, *15*(23), 4195. https://doi.org/10.3390/w15234195

Amante, C., & Eakins, B. (2009). Etopo1 1 arc-minute global relief model: Procedures, data sources and analysis. *NOAA Technical Memorandum NESDIS NGDC-24. National Geophysical Data Center, NOAA*. https://doi.org/10.7289/V5C8276M

Ando, M. (1975). Source mechanisms and tectonic significance of historical earthquakes along the nankai trough, Japan. *Tectonophysics*, *27*(2), 119–140. https://doi.org/10.1016/0040-1951(75)90102-x

Ando, M. (1982). A fault model of the 1946 nankaido earthquake derived from tsunami data. *Physics of the Earth and Planetary Interiors*, *28*(4), 320–336. https://doi.org/10.1016/0031-9201(82)90089-9

Annaka, T., Satake, K., Sakakiyama, T., Yanagisawa, K., & Shuto, N. (2007). Logic-tree approach for probabilistic tsunami hazard analysis and its applications to the Japanese coasts. In *Tsunami and its hazards in the indian and pacific oceans* (pp. 577–592).

Baba, T., Tanioka, Y., Cummins, P. R., & Uhira, K. (2002). The slip distribution of the 1946 nankai earthquake estimated from tsunami inversion using a new plate model. *Physics of the Earth and Planetary Interiors*, *132*(1–3), 59–73. https://doi.org/10.1016/s0031-9201(02)00044-4

Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning. *Springer*, *4*(4).

Blaser, L., Ohrnberger, M., Krüger, F., & Scherbaum, F. (2012). Probabilistic tsunami threat assessment of 10 recent earthquakes offshore sumatra. *Geophysical Journal International*, *188*(3), 1273–1284. https://doi.org/10.1111/j.1365-246x.2011.05324.x

Blaser, L., Ohrnberger, M., Riggelsen, C., Babeyko, A., & Scherbaum, F. (2011). Bayesian networks for tsunami early warning. *Geophysical Journal International*, *185*(3), 1431–1443. https://doi.org/10.1111/j.1365-246x.2011.05020.x

Cabinet Office Japan. (2019). Disaster management in Japan. Retrieved from https://www.bousai.go.jp/jishin/nankai/taisaku_wg/pdf/1_sanko2.pdf. in Japanese, accessed 17 10 2023.

Clawpack Development Team. (2021). *Clawpack Software*. Retrieved from https://doi.org/10.5281/zenodo.4503024

Comninou, M., & Dundurs, J. (1975). The angular dislocation in a half space. *Journal of Elasticity*, *5*(3–4), 203–216. https://doi.org/10.1007/bf00126985

Cummins, P. R., Baba, T., Kodaira, S., & Kaneda, Y. (2002). The 1946 nankai earthquake and segmentation of the nankai trough. *Physics of the Earth and Planetary Interiors*, *132*(1–3), 75–87. https://doi.org/10.1016/s0031-9201(02)00045-6

Fauzi, A., & Mizutani, N. (2020). Machine learning algorithms for real-time tsunami inundation forecasting: A case study in nankai region. *Pure and Applied Geophysics*, *177*(3), 1437–1450. https://doi.org/10.1007/s00024-019-02364-4

Fujita, S. (2024). Python codes for tsunami prediction with the scenario superposition (Version v1.0.0) [Computer software]. *Zenodo*. https://doi.org/10.5281/zenodo.14029637

Fujita, S., Nomura, R., Moriguchi, S., Otake, Y., Koshimura, S., LeVeque, R. J., & Terada, K. (2024). Optimization of a tsunami gauge configuration for pseudo-super-resolution of wave height distribution. *Earth and Space Science*, *11*(2), e2023EA003144. https://doi.org/10.1029/2023ea003144

Fukutani, Y., Suppasri, A., & Imamura, F. (2015). Stochastic analysis and uncertainty assessment of tsunami wave height using a random source parameter model that targets a tohoku-type earthquake fault. *Stochastic Environmental Research and Risk Assessment*, *29*(7), 1763–1779. https://doi.org/10.1007/s00477-014-0966-4

Geist, E. L. (2002). Complex earthquake rupture and local tsunamis. *Journal of Geophysical Research*, *107*(B5), ESE–2. https://doi.org/10.1029/2000jb000139

Geist, E. L., & Parsons, T. (2006). Probabilistic analysis of tsunami hazards. *Natural Hazards*, *37*(3), 277–314. https://doi.org/10.1007/s11069-005-4646-z

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gica, E. (2008). Development of the forecast propagation database for noaa's short-term inundation forecast for tsunamis (sift).

Gusman, A. R., Sheehan, A. F., Satake, K., Heidarzadeh, M., Mulia, I. E., & Maeda, T. (2016). Tsunami data assimilation of cascadia seafloor pressure gauge records from the 2012 Haida gwaii earthquake. *Geophysical Research Letters*, *43*(9), 4189–4196. https://doi.org/10.1002/2016gl068368

Gusman, A. R., Tanioka, Y., MacInnes, B. T., & Tsushima, H. (2014). A methodology for near-field tsunami inundation forecasting: Application to the 2011 tohoku tsunami. *Journal of Geophysical Research: Solid Earth*, *119*(11), 8186–8206. https://doi.org/10.1002/2014jb010958

Hayes, G. P., Moore, G. L., Portner, D. E., Hearne, M., Flamme, H., Furtney, M., & Smoczyk, G. M. (2018). Slab2, a comprehensive subduction zone geometry model. *Science*, *362*(6410), 58–61. https://doi.org/10.1126/science.aat4723

Heidarzadeh, M., & Kijko, A. (2011). A probabilistic tsunami hazard assessment for the makran subduction zone at the northwestern indian ocean. *Natural Hazards*, *56*(3), 577–593. https://doi.org/10.1007/s11069-010-9574-x

Heidarzadeh, M., Wang, Y., Satake, K., & Mulia, I. E. (2019). Potential deployment of offshore bottom pressure gauges and adoption of data assimilation for tsunami warning system in the western mediterranean sea. *Geoscience Letters*, *6*(1), 19. https://doi.org/10.1186/s40562-019-0149-8

Intergovernmental Oceanographic Commision. (2023). North-Eastern atlantic, the mediterranean and connected seas tsunami early warning and mitigation system (NEAMTWS): 2030 strategy. *IOC. Technical series*, 171.[

Kaneda, Y. (2010). The advanced ocean floor real time monitoring system for mega thrust earthquakes and tsunamis-application of donet and donet2 data to seismological research and disaster mitigation. In *Oceans 2010 mts/ieee seattle* (pp. 1–6).
Klema, V., & Laub, A. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, *25*(2), 164–176. https://doi.org/10.1109/tac.1980.1102314
Koshimura, S., & Nomura, R. (2022). Data from 666 earthquake/tsunami scenario simulations targeting Nankai subduction. *Zenodo*. Retrieved from https://doi.org/10.5281/zenodo.6785643
Koshimura, S., & Shuto, N. (2015). Response to the 2011 great east Japan earthquake and tsunami disaster. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *373*(2053), 20140373. https://doi.org/10.1098/rsta.2014.0373
Kotani, T., Tozato, K., Takase, S., Moriguchi, S., Terada, K., Fukutani, Y., et al. (2020). Probabilistic tsunami hazard assessment with simulation-based response surfaces. *Coastal Engineering*, *160*, 103719. https://doi.org/10.1016/j.coastaleng.2020.103719
LeVeque, R. J., Nomura, R., & Fujita, S. (2024). Synthetic dataset of 2342 earthquake/tsunami scenarios targeting the nankai trough subduction zone (Version v0.0.1) [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.12696848
LeVeque, R. J., Waagan, K., González, F. I., Rim, D., & Lin, G. (2016). Generating random earthquake events for probabilistic tsunami hazard assessment. *Pure and Applied Geophysics*, *173*(12), 3671–3692. https://doi.org/10.1007/978-3-319-55480-8_2
LeVeque, R. J., Waagan, K., González, F. I., Rim, D., & Lin, G. (2017). Generating random earthquake events for probabilistic tsunami hazard assessment. *Global Tsunami Science: Future Times*, *I*, 3671–3692. https://doi.org/10.1007/978-3-319-55480-8_2
Liu, C. M., Rim, D., Baraldi, R., & LeVeque, R. J. (2021). Comparison of machine learning approaches for tsunami forecasting from sparse observations. *Pure and Applied Geophysics*, *178*(12), 5129–5153. https://doi.org/10.1007/s00024-021-02841-9
Maeda, T., Obara, K., Shinohara, M., Kanazawa, T., & Uehira, K. (2015). Successive estimation of a tsunami wavefield without earthquake source data: A data assimilation approach toward real-time tsunami forecasting. *Geophysical Research Letters*, *42*(19), 7923–7932. https://doi.org/10.1002/2015gl065588
Makinoshima, F., Oishi, Y., Yamazaki, T., Furumura, T., & Imamura, F. (2021). Early forecasting of tsunami inundation from tsunami and geodetic observation data with convolutional neural networks. *Nature Communications*, *12*(1), 2253. https://doi.org/10.1038/s41467-021-22348-0
Melgar, D. (2020). Mudpy. Retrieved from https://github.com/dmelgarm/MudPy
Melgar, D., & Hayes, G. P. (2019). The correlation lengths and hypocentral positions of great earthquakes. *Bulletin of the Seismological Society of America*, *109*(6), 2582–2593. https://doi.org/10.1785/0120190164
Mori, N., Goda, K., & Cox, D. (2018). Recent process in probabilistic tsunami hazard analysis (ptha) for mega thrust subduction earthquakes. In *The 2011 Japan earthquake and tsunami: Reconstruction and restoration: Insights and assessment after 5 years* (pp. 469–485).
Mulia, I. E., & Satake, K. (2021). Synthetic analysis of the efficacy of the s-net system in tsunami forecasting. *Earth Planets and Space*, *73*, 1–11. https://doi.org/10.1186/s40623-021-01368-6
Mulia, I. E., Ueda, N., Miyoshi, T., Gusman, A. R., & Satake, K. (2022). Machine learning-based tsunami inundation prediction derived from offshore observations. *Nature Communications*, *13*(1), 5489. https://doi.org/10.1038/s41467-022-33253-5
Musa, A., Watanabe, O., Matsuoka, H., Hokari, H., Inoue, T., Murashima, Y., et al. (2018). Real-time tsunami inundation forecast system for tsunami disaster prevention and mitigation. *The Journal of Supercomputing*, *74*(7), 3093–3113. https://doi.org/10.1007/s11227-018-2363-0
Nakano, M., Murphy, S., Agata, R., Igarashi, Y., Okada, M., & Hori, T. (2020). Self-similar stochastic slip distributions on a non-planar fault for tsunami scenarios for megathrust earthquakes. *Progress in Earth and Planetary Science*, *7*(1), 1–13. https://doi.org/10.1186/s40645-020-00360-0
Nomura, R., Fujita, S., Galbreath, J. M., Otake, Y., Moriguchi, S., Koshimura, S., et al. (2022). Sequential bayesian update to detect the most likely tsunami scenario using observational wave sequences. *Journal of Geophysical Research: Oceans*, *127*(10), e2021JC018324. https://doi.org/10.1029/2021jc018324
Nomura, R., Vermare, L. A. H., Fujita, S., Rim, D., Moriguchi, S., LeVeque, R. J., & Terada, K. (2024). On the performance of sequential bayesian update for database of diverse tsunami scenarios. Retrieved from https://arxiv.org/abs/2407.03631
Oishi, Y., Imamura, F., & Sugawara, D. (2015). Near-field tsunami inundation forecast using the parallel tunami-n2 model: Application to the 2011 tohoku-oki earthquake combined with source inversions. *Geophysical Research Letters*, *42*(4), 1083–1091. https://doi.org/10.1002/2014gl062577
Omira, R., Matias, L., & Baptista, M. (2017). Developing an event-tree probabilistic tsunami inundation model for ne atlantic coasts: Application to a case study. *Global Tsunami Science: Future Times*, *I*(12), 3775–3794. https://doi.org/10.1007/s00024-016-1367-z
Ports and Harbours Bureau. (2022). Real-time NOWPHAS (the Nationwide Ocean Wave information network for Ports and HArbourS). Retrieved from https://www.mlit.go.jp/kowan/nowphas/index_eng.html. Accessed on 21 July 2022.
Rim, D., Baraldi, R., Liu, C. M., LeVeque, R. J., & Terada, K. (2022). Tsunami early warning from global navigation satellite system data using convolutional neural networks. *Geophysical Research Letters*, *49*(20), e2022GL099511. https://doi.org/10.1029/2022gl099511
Satake, K. (2014). Advances in earthquake and tsunami sciences and disaster risk reduction since the 2004 indian ocean tsunami. *Geoscience Letters*, *1*(1), 1–13. https://doi.org/10.1186/s40562-014-0015-7
Scala, A., Lorito, S., Romano, F., Murphy, S., Selva, J., Basili, R., et al. (2020). Effect of shallow slip amplification uncertainty on probabilistic tsunami hazard analysis in subduction zones: Use of long-term balanced stochastic slip models. *Pure and Applied Geophysics*, *177*(3), 1497–1520. https://doi.org/10.1007/s00024-019-02260-x
Selva, J., Lorito, S., Volpe, M., Romano, F., Tonini, R., Perfetti, P., et al. (2021). Probabilistic tsunami forecasting for early warning. *Nature Communications*, *12*(1), 5677. https://doi.org/10.1038/s41467-021-25815-w
Tatehata, H. (1997). The new tsunami warning system of the Japan meteorological agency. *Perspectives on tsunami hazard reduction: Observations, Theory and Planning*, 175–188. https://doi.org/10.1007/978-94-015-8859-1_12
Tatsumi, D., Calder, C. A., & Tomita, T. (2014). Bayesian near-field tsunami forecasting with uncertainty estimates. *Journal of Geophysical Research: Oceans*, *119*(4), 2201–2211. https://doi.org/10.1002/2013jc009334
Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000jd900719
Tsushima, H., & Ohta, Y. (2014). Review on near-field tsunami forecasting from offshore tsunami data and onshore gnss data for tsunami early warning. *Journal of Disaster Research*, *9*(3), 339–357. https://doi.org/10.20965/jdr.2014.p0339
Wang, Y., Satake, K., Maeda, T., & Gusman, A. R. (2017). Green's function-based tsunami data assimilation: A fast data assimilation approach toward tsunami early warning. *Geophysical Research Letters*, *44*(20), 10–282. https://doi.org/10.1002/2017gl075307
Wang, Y., Satake, K., Maeda, T., & Gusman, A. R. (2018). Data assimilation with dispersive tsunami model: A test for the nankai trough. *Earth Planets and Space*, *70*, 1–9. https://doi.org/10.1186/s40623-018-0905-6

Xu, H., & Wu, H. (2023). Accurate tsunami wave prediction using long short-term memory based neural networks. *Ocean Modelling*, *186*, 102259. https://doi.org/10.1016/j.ocemod.2023.102259

Yoshida, I., Tasaki, Y., Otake, Y., & Wu, S. (2018). Optimal sampling placement in a Gaussian random field based on value of information. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, *4*(3), 04018018. https://doi.org/10.1061/ajrua6.0000970

Yoshida, I., Tomizawa, Y., & Otake, Y. (2021). Estimation of trend and random components of conditional random field using Gaussian process regression. *Computers and Geotechnics*, *136*, 104179. https://doi.org/10.1016/j.compgeo.2021.104179