A Logit-Based Binary Classifier of Tsunamigenic Earthquakes for the Northern Pacific Ocean

Alexey Konovalov (a.konovalov@geophystech.ru)
Far Eastern Geological Institute FEB RAS: Dal'nevostocnyj geologiceskij institut Dal'nevostocnogo otdelenia Rossijskoj akademii nauk https://orcid.org/0000-0003-2997-1524

Grigory Samsonov
Sakhalin State University

Andrey Stepnov
Far Eastern Geological Institute FEB RAS: Dal'nevostocnyj geologiceskij institut Dal'nevostocnogo otdelenia Rossijskoj akademii nauk

Research Article

Keywords: submarine earthquake, tsunami, logit, predictive model, binary classifier, early warning

Posted Date: March 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1421995/v1

A Logit-Based Binary Classifier of Tsunamigenic Earthquakes for the Northern Pacific Ocean

A.V. Konovalov[a*], G.A. Samsonov[b], A.A. Stepnov[a]

[a] Far East Geological Institute, Far Eastern Branch, Russian Academy of Sciences, Vladivostok, Russia

[b] Sakhalin State University, Yuzhno-Sakhalinsk, Russia

*e-mail: a.konovalov@geophystech.ru

In this study, logistic regression was used as a tool for deriving the binary classifier of tsunamigenic and nontsunamigenic earthquakes in the near-source zone. Earthquake source depth and moment magnitude were considered as predictors. The training dataset consisted of 767 M6.0+ submarine earthquakes, including 80 tsunamigenic and 687 nontsunamigenic events that occurred in the northern part of the Pacific Ocean from 1960 to 2020. The target area has already experienced significant and catastrophic tsunamis. The current analysis clearly showed that the data-driven logit model had a significantly lower false discovery rate relative to the threshold magnitude criteria that are widely used by tsunami warning agencies. At the same time, the balanced accuracy was about 71%, which suggests optimism for accurate tsunami forecasting based on the rapid interpretation of earthquake source parameters.

Keywords: submarine earthquake; tsunami; logit; predictive model; binary classifier; early warning

1 Introduction

Despite the fact that it is almost impossible to find the exact date of the first issued tsunami warning, the scientific principles for predicting this phenomenon were formulated in the 1920s (e.g., Finch 1924). Levin and Nosov (2016) report that probably the first successful tsunami forecast occurred on February 3, 1923, based on an interpretation of earthquake data.

S.L. Soloviev made a significant contribution to the development of the seismotectonic principle of tsunami forecasting. He proposed a magnitude–geographical criterion for the discrimination of tsunamigenic earthquakes (Soloviev and Shebalin 1959), which involved using ratios of the height (intensity) of the tsunami and the magnitude of the tsunamigenic earthquake. The analysis of the tsunamis produced by subduction earthquakes in the Far East region of Russia contributed to the assessment of the threshold intensity of the tsunami; intensities exceeding that threshold led to significant impacts in the target region. Based on the linear relationship between tsunami intensity and the magnitude of tsunamigenic earthquakes, a threshold magnitude corresponding to tsunami threshold intensity was selected.

Later, these empirical relations and threshold magnitudes were specified (e.g., Iida 1970; Soloviev 1972; Abe 1979; Gusyakov and Chubarov 1987; Boschetti and Ioualalen 2021). Current tsunami warning centers take into account not only the magnitude of a submarine earthquake but also the distance from the possible tsunami source to the coastal location (e.g., Amato et al. 2021). Such rules are represented by several hazard classes governing a certain action of emergency agencies.

In early studies (e.g., Iida 1970; Ivashchenko and Go 1973), the threshold magnitude as a function of the source depth was found by visual fitting of the line separating the two classes (tsunamigenic and nontsunamigenic) of submarine earthquakes on the magnitude–depth plot. These equations were never used in the tsunami early warning surveys. Of late, earthquake depth has been considered as a simplified indicator variable categorizing the seismic events into two groups: the first group includes events with a source depth of less than 100 km; the second, 100 km or more (e.g., Users Guide... 2017).

Nosov et al. (2018) developed a fully automatic service (Tsunami Observer) for tsunami forecasting. Earthquake source parameters such as the focal mechanism solution and source depth, reported by seismological agencies, were used to calculate the free surface deformation. The proposed method gave an estimation of the initial elevation of the water surface in the tsunami source region and the potential energy of the tsunami associated with it. Tsunami intensity was estimated based on the calibrated relation between the calculated tsunami energy and observed tsunami height. Then, the "threshold intensity" principle was applied to determine whether the submarine earthquake

would be tsunamigenic or not. This approach has a number of advantages over the current magnitude–geographical criteria. Its clarity lies in the use of additional parameters from numerical simulation. The disadvantage is the ambiguity of the nodal plane selection. The authors of the experimental service noted (Nosov et al. 2020) that tsunami energy estimates are highly sensitive to the choice of nodal plane for shallow earthquakes.

Cienfuegos et al. (2018) showed that the use of a simplified earthquake source model with uniform slip distribution led to conservative estimates of tsunami parameters. Epistemic uncertainties associated with spatial heterogeneity of earthquake slip play an important role in accurate tsunami prediction. Rational estimates of tsunami parameters in terms of probability can be obtained by stochastic simulation of the slip vector at the planar source and the position of the nodal plane in space, and subsequent numerical simulation of tsunami characteristics for each random realization of the finite fault model (Davies 2019). Probabilistic assessment of a tsunami hazard can also be realized using the Bayesian approach or the total probability theorem (Selva et al. 2021).

On the other hand, the focal mechanism is most commonly determined more slowly than the hypocenter parameters and the earthquake magnitude. For this reason, simulation-based approaches may lose their relevance in the near-source zone. Therefore, tsunami forecasting still remains an important practical issue of seismological observations based on rapid interpretation of earthquake source parameters.

In the current study, logistic regression (Cramer 2002) was used as a tool for the classification of tsunamigenic and nontsunamigenic earthquakes. We used a preprocessed dataset of submarine earthquakes that caused or did not cause tsunamis in the target region. Earthquake source depth and moment magnitude were considered as predictors. This approach was used here for the first time for developing the data-driven predictive model of tsunamigenic earthquakes.

2 Materials and methods

2.1 Methods

Logit analysis is widely used for binary data classification in geoscience (e.g., Zhu et al. 2017; Stepnova et al. 2021). In the binary logistic model, the logistic response function is determined by a binary dependent variable (Harrell, 2015), where the two possible outcomes are labeled as 0 and 1. The logistic response function predicts not the results themselves but the probability of their onset:

$$h_\theta(\boldsymbol{X}) = \frac{1}{1+e^{-\boldsymbol{\theta}^T\boldsymbol{X}}} = \mathrm{P}(Y = 1|\boldsymbol{X}, \boldsymbol{\theta}), \quad (1)$$

where $h_\theta(\boldsymbol{X})$ is the logistic response function, $\boldsymbol{X} = \{1, X_1, X_2, \dots, X_n\}^T$ is the column vector of independent variables (predictors), $\boldsymbol{\theta} = \{\theta_0, \theta_1, \theta_2, \dots, \theta_n\}^T$ is the column vector of the parameters to be estimated, n is the number of

independent variables, Y is a binary dependent variable that has two values $Y = 0$ and $Y = 1$, and $\mathrm{P}(Y = 1|\boldsymbol{X}, \boldsymbol{\theta})$ is the conditional probability of event 1 with the given $\boldsymbol{X}$ and $\boldsymbol{\theta}$, wherein

$$\mathrm{P}(Y = 0|\boldsymbol{X}, \boldsymbol{\theta}) = 1 - h_\theta(\boldsymbol{X}). \quad (2)$$

Each predictor can be set either by a continuous or by a categorical variable. Logistic regression measures the relationship between a categorical dependent variable and one or more predictors by estimating the logarithm of the odds:

$$Logit(\mathrm{P}) = \ln\frac{\mathrm{P}(Y = 1|\boldsymbol{X}, \boldsymbol{\theta})}{1-\mathrm{P}(Y = 1|\boldsymbol{X}, \boldsymbol{\theta})} = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_n X_n, \quad (3)$$

where $Logit(\mathrm{P})$ is the logit unit.

The given model follows the generalized Bernoulli scheme, in which a random variable takes only two values, 0 and 1, and their probabilities are determined by Eq. 2 and Eq. 1, respectively. Therefore, the probability distribution function of the variable $Y = y$ is given by

$$\mathrm{P}(y|\boldsymbol{X}, \boldsymbol{\theta}) = h_\theta(\boldsymbol{X})^y [1 - h_\theta(\boldsymbol{X})]^{1-y} \quad (4)$$

and the likelihood can be written as

$$L(y|\boldsymbol{X}, \boldsymbol{\theta}) = \prod_i \mathrm{P}(y_i|\boldsymbol{X}_i, \boldsymbol{\theta}) = \prod_i h_\theta(\boldsymbol{X}_i)^{y_i} [1 - h_\theta(\boldsymbol{X}_i)]^{1-y_i}, \quad (5)$$

where i runs from 1 to N (N is the sample size).

Maximization of the likelihood function is achieved by optimization methods. However, in practice, the log-likelihood is most commonly used instead of the original likelihood function. Since the logarithm is a monotonically increasing function, the maximum value of the logarithm of the probability occurs at the same point as the original probability function. Taking the logarithm of Eq. 5 gives:

$$L_{\log}(y|\boldsymbol{X}_i, \boldsymbol{\theta}) = N^{-1} \ln \prod_i \mathrm{P}(y_i|\boldsymbol{X}_i, \boldsymbol{\theta}) = N^{-1}(\sum_i y_i \ln h_\theta(\boldsymbol{X}_i) + (1 - y_i) \ln[1 - h_\theta(\boldsymbol{X}_i)]). \quad (6)$$

If Eq. 6 is multiplied by −1, the given expression will explain logistic losses. Minimizing the logistic loss function by the gradient descent method allows us to estimate the best parameters $\boldsymbol{\theta}$.

In our case, the moment magnitude M and the source depth D are predictors. The binary dependent variable is the tsunamigenic class label of submarine earthquakes, which has two values: $Y = \mathrm{true}$ and $Y = \mathrm{false}$. According to Eq. 3, the logarithm of the odds for tsunamigenic class is a linear combination of independent variables M and D:

$$Logit(\mathrm{P}) = \ln\frac{\mathrm{P}(Y = \mathrm{true}|X, \theta)}{1-\mathrm{P}(Y = \mathrm{true}|X, \theta)} = \theta_0 + \theta_1 * M + \theta_2 * D, \quad (7)$$

where θ_0, θ_1 and θ_2 are parameters of our model to be estimated.

The $Logit(\mathrm{P})$ is used as a classifier of tsunamigenic/nontsunamigenic events. This is achieved by using a threshold probability P_{th}: if the right-hand side of Eq. 7 with the given earthquake parameters is equal to or greater than $Logit(\mathrm{P}_{th})$, then the seismic event is categorized as tsunamigenic, with all other values belonging to the nontsunamigenic class.

For those classification problems that have a class imbalance, the threshold probability P_{th} may significantly differ from the default value 0.5. We used F-metrics as an indicator of the performance of a classifier that leads the tuning of the threshold level:

$$F_\beta = (1+\beta^2)\frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}. \quad (8)$$

$Precision$ and $Recall$ are defined as follows (see Eq. 9 and Eq. 10, respectively):

$$Precision = \frac{TP}{TP+FP}, \quad (9)$$

$$Recall = \frac{TP}{TP+FN}, \quad (10)$$

where *TP* (true positives) and *TN* (true negatives) are correctly predicted classes, and *FP* (false positives) and *FN* (false negatives) are type I and type II errors, respectively.

Parameter β is selected so that $Recall$ is considered β times more essential than $Precision$. When the number of true tsunamigenic events that may be incorrectly categorized as nontsunamigenic (type I error) is minimized, the β parameter can be different from 1. However, in the current study, it is taken as equal to 1. This means that $Precision$ and $Recall$ have the same significance or, in the other words, that the F1 score is the mean harmonic value between $Precision$ and $Recall$.

Generally, metrics in Eq. 8–10 are functions of P. Maximizing the F1 score with respect to the P allows us to select the optimal threshold probability P_{th} used in the binary classification.

2.2 Study area

The study area is limited by four angular points that form a polygon with the geographical coordinates 33–55°N and 138–166°E (see Fig. 1). This part of the Pacific Ocean encompasses possible tsunami sources that may cause significant impacts to the coastal locations of island arcs considered as near-source zones of tsunamis. The target area has already experienced significant and catastrophic tsunamis (e.g., Tanioka and Satake 1996; Mimura et al.

2011; Gusyakov 2011): the 1737 Kamchatka Tsunami, the 1886 Great Meiji Sanriku Tsunami, the 1952 North Kuril Tsunami and the 2011 Great East Japan Tsunami.

2.3 Data

The U.S. Geological Survey's earthquake database (U.S. Geological Survey 2021) was used in the current study. It is available at the website: https://earthquake.usgs.gov/earthquakes/search/. The time window from 1960 to 2020 was considered as a period of intensive increase in the number of instrumental sites. Initially, we used the M 4.5+ dataset. After statistical tests of the completeness level of the seismic database, the cutoff magnitude was increased to Mc 6 (see Sect. 2.4.1). Selection parameters for the earthquake depth were not limited.

As the seismological data centers report different types of magnitude such as Mw, Ms and mb, all magnitude estimates were unified to the Mw scale based on global intermagnitude relations (Gusev 1991). The magnitude conversion from M to Mw was achieved by polynomial approximation of the cubic type, where M is either Ms or mb:

$$\mathrm{Mw} = a * \mathrm{M}^3 + b * \mathrm{M}^2 + c * \mathrm{M} + d, \quad (11)$$

where a, b, c and d are the regression coefficients (see Table 1).

The 2021 tsunami database provided by the US National Oceanic and Atmospheric Administration is free and available at the National Geophysical Data Center's website: https://www.ngdc.noaa.gov/hazel/view/hazards/tsunami/. The search criteria used for the tsunami dataset were the same as those used for the seismic database selection. The analysis of the completeness level of the tsunami database (see Sect. 2.4.2) indicated that events with a maximum water height of less than 10 cm could be missed due to the sensitivity of the ocean-bottom network. Therefore, tsunamis with reported maximum water heights of less than 10 cm or no reported height values were removed from the dataset.

2.4 Completeness check

2.4.1 Completeness level of seismic database

A statistical model developed by Ogata and Katsura (1993) was applied in the current study for estimating the magnitude of completeness of the earthquake catalogue. The μ (50% detection rate), b-value and σ can be simultaneously estimated by numerically maximizing the likelihood function:

$$L(M, \boldsymbol{p}) = \frac{W(M,\boldsymbol{p})q(M,\boldsymbol{p})}{\int_{-\infty}^{+\infty} W(M,\boldsymbol{p})q(M,\boldsymbol{p})dM}, \quad (12)$$

where $\boldsymbol{p}$ is a parameter vector (μ, b, σ) to be estimated, $W(M) = e^{-\beta M}$ with $\beta = b \ln(10)$ is the distribution based on the Gutenberg–Richter relation, and $q(M, \boldsymbol{p}) = \int_{-\infty}^{M} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx$ is the detection rate function.

Given the observed magnitudes $M_1, \ldots, M_N$, model parameters were estimated by minimizing the $\ln L(M, \boldsymbol{p}) = \sum_{i=1}^{N} \ln L(M_i, \boldsymbol{p})$ using the trust region method (Conn et al. 2000). The code is available on GitHub: https://gist.github.com/jamm1985/79b43a6dc1655a32e44ae64ee437d5db (accessed 10 February 2022).

The magnitude of completeness is determined by the equation

$$m_c(k) = \mu + k\,\sigma, \tag{13}$$

where k is the confidence level; k = 0, 1, 2, 3 means that 50%, 68%, 95% and 99% of events are detected above m_c.

According to Fig. 2, the model parameters with the given seismic dataset were well constrained. The magnitude of completeness (95% confidence level) determined from Eq. 13 was equal to m_c 6.

2.4.2 Completeness level of tsunami database

The cumulative counts depending on the log of the maximum water height (as a magnitude unit) were plotted for the analysis of the completeness level of the tsunami database (Fig. 3). The cumulative frequency–magnitude relation had a log-linear form. At the smallest magnitudes, the log-linear dependence was not fulfilled. It is usually assumed that the dataset is complete up to the threshold magnitude for which the log-liner dependence on the cumulative frequency–magnitude plot is fulfilled. According to the recurrence plot in Fig. 3, events with heights of less than 10 cm could occur more frequently, but were missed due to the sensitivity of the equipment. Various studies (e.g., Baba et al. 2004) indicate that the minimum tsunami height recorded by ocean-bottom equipment is about 5 cm. Therefore, the threshold value of the maximum water height reflects the detection bounds of the instrumentation network aimed at tsunami registration in the study area. We considered this value (10 cm) as the completeness level of the tsunami database.

2.5 Data preprocessing

We filtered the earthquake catalogue in order to remove nonsubmarine seismic events. In addition, at this stage, catalogs of submarine earthquakes and tsunamis were merged into one seismic dataset, in which an additional binary variable associated with the tsunamigenic class (true/false) was assigned to each seismic event. The following rules of association of events from the two datasets were adopted: date matching, matching of the epicenter locations with valid differences of up to 0.5 degrees, and matching of earthquake magnitudes with valid differences of up to 0.5. If events from the two subsets were associated, then the tsunamigenic variable was labeled as true; otherwise, it was

mapped as false. This means that the nonassociated events from the earthquake catalogue were automatically assigned a nontsunamigenic class.

During the merging of the seismic and tsunami databases into one training dataset, seven nonassociated events from the tsunami subset were identified. The earthquake source parameters of the three nonassociated tsunamis were outside the valid differences defined in the rules of association. The epicenters of the remaining four nonassociated tsunamigenic events according to the earthquake catalogue were located on land. These events were removed from the earthquake catalogue at the initial stage of data preprocessing. A detailed analysis of earthquake locations indicated that one of them, the catastrophic 1995 Neftegorsk earthquake (Mw 7.1), occurred in Sakhalin Island and could not have caused a tsunami in the ocean, while the other three events occurred in the immediate vicinity of the coastline and were accompanied by tsunami waves. This can be explained by effects related to errors in the epicenter location or the size of the earthquake source, when an area of the finite fault is located on land, while another area is underwater. We applied the corresponding adjustments to the training dataset. Maps of earthquake epicenters and tsunami sources, which were both removed and used in the current study after preprocessing, are shown in Fig. 1a and Fig. 1b, respectively.

Finally, after careful verification of the training dataset, we found that two events with Mw≥7.7 and source depths of less than 50 km were categorized as nontsunamigenic. Independent tsunami data centers (e.g., Web Encyclopedia on Natural Hazards 2021) reported the same information. The considered events occurred on 11.03.2011 in the aftershock area of the Great Japanese Earthquake (Mw 9.1), which occurred a few minutes earlier and caused catastrophic tsunami waves. Wave processes induced by the mainshock and by the largest aftershocks could be superimposed. Therefore, these aftershocks were excluded from the analysis since they could not be objectively attributed to one of the two classes.

The final database (see SI) consisted of 767 submarine earthquakes, including 80 tsunamigenic and 687 nontsunamigenic events (Fig. 4).

3 Results and discussion

Fig. 5 shows the distribution of submarine earthquakes on the 2D magnitude–depth plot. The class category is indicated by a color. It is clear that tsunamigenic earthquakes tend to be above a certain imaginary line, the parameters of which should be estimated.

Applying the logit tool to the joint earthquake–tsunami database, it was found that the conditional probability that a submarine earthquake with a given moment magnitude and source depth would cause a tsunami is given by the following expression:

$$h_\theta(\boldsymbol{X}) = \mathrm{Sig}(0.993 \cdot M - 0.0349 \cdot D - 7.3224), \quad (14)$$

where Sig is the sigmoid function determined in Eq. 1.

The probability in Eq. 14 increases with magnitude and decreases with depth.

As we are dealing with class–imbalanced data (there were many more records of nontsunamigenic earthquakes than tsunamigenic cases), it was necessary to select an optimal threshold probability according to Eq. 8. Fig. 6 shows that the F1 score reached the highest value at a probability of approximately 0.22, and in this way, the balance between $Precision$ and $Recall$ was achieved.

Substituting the obtained threshold probability (0.22) in Eq. 14, we obtain the threshold magnitude given by the functional dependence on the source depth of the submarine earthquake:

$$M_{th} = 5.8 + 0.035 \cdot D. \quad (15)$$

The predictive model of tsunamigenic earthquakes obtained in the current study (see Eq. 15), as well as proposed by K. Iida (1963), are shown in Fig. 5. Iida was the first to show that the threshold magnitude increases with the source depth according to a linear law. Eq. 15 reflects a well-known fact; however, the model coefficients were statistically justified for the first time and determined from a dataset over a long duration of instrumental observation.

It is well known that the source depth of shallow earthquakes is characterized by uncertainty, especially for distant events. In seismic bulletins, this parameter is most commonly fixed. Generally, the interplate type of seismicity predominates in the Japan–Kurile–Kamchatka subduction zone. The source depth of shallow interplate earthquakes is fixed in the range of 30–40 km. From Eq. 15, it follows that the threshold magnitude for the depth of 33 km corresponds to Mw 7.0. This estimate corresponds to the minimum threshold magnitude used in the tsunami warning survey of Russia and of many other centers.

The hypocenters of the crustal earthquakes are most commonly located at a depth of 10 km. In this case, the logit model predicted the threshold magnitude of Mw 6.2 (see Eq. 15). Some shallow earthquakes with M ~ 6, occurring in the coastal zone of the northwest part of the Pacific seismic belt, are accompanied by significant tsunami waves (e.g., Satake and Kanamori 1991; Konovalov et al. 2015). The results obtained in the current study correspond to the considerations expressed recently on clarifying the threshold magnitude of tsunamigenic earthquakes (e.g., Gusiakov 2011).

The performance of a binary classifier was tested using balanced accuracy (BA) and the false discovery rate (FDR). The false discovery rate indicates the counts of tsunamigenic earthquakes that may be incorrectly categorized

as nontsunamigenic. We did not split the training and validation subsets, due to the low number of positively categorized classes. The results from testing the model should be considered as the results obtained with the training dataset. The earthquake source depth of the validation subset was in the range of 0–250 km.

The given metrics are defined as follows:

$$BA = 0.5 \cdot \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right), \quad (16)$$

$$FDR = \frac{FP}{TP+FP}. \quad (17)$$

Table 2 summarizes the performance tests considering different models: the data-driven logit model (see Eq. 15), the Iida-1963 model (see Eq. 18) and the threshold magnitude criterion that is currently used by the Russian tsunami warning survey (see Eq. 18). The Iida-1963 model is applicable to tsunamis with an intensity greater than 1 (Iida 1963):

$$M_{th} = 6.46 + 0.019 \cdot D. \quad (18)$$

The threshold magnitude criterion is given by a simplified expression:

$$M_{th}=7. \quad (19)$$

As can be seen from Table 2, the Iida-1963 model had a higher balanced accuracy relative to the logit model. However, the Iida-1963 model had a relatively high false discovery rate; that is, the Iida-1963 model was mainly applicable to significant tsunamis. The threshold magnitude criterion is widely used by tsunami warning agencies around the world. However, in this case, the counts of missed warnings remained significant. The current analysis clearly showed that the data-driven logit model had a significantly lower false discovery rate at the 70% accuracy level.

For future studies, it would be advisable to consider additional categorical variables, such as the type of focal mechanism (strike-slip/reverse/normal/unknown), the tectonic type of earthquake based on the geographical location of the hypocenter (subduction/collision/rift/unknown) as well as continuous variables, for example, underwater depth in the earthquake epicenter.

Chicco and Jurman (2020) showed that the Matthews correlation coefficient produces more informative and truthful metrics for deriving the binary classifier, especially for imbalanced datasets. They suggest that it should be used instead of the F1 score. We suggest that further evaluation with the Matthews correlation coefficient using additional independent variables would help to significantly improve the performance of the binary classifier.

4 Conclusions

The logistic regression tool was used for deriving the data-driven binary classifier of tsunamigenic and nontsunamigenic submarine earthquakes. It was assumed that source depth and moment magnitude are the most important factors controlling tsunami generation in the near-source zone. The logit model coefficients are statistically justified and determined using preprocessed tsunami and earthquake catalogues that contain empirical data over a long duration of instrumental observation in the Northern Pacific Ocean. The advantages of the considered approach are not only the use of the data on the intensity of tsunamigenic earthquakes but also data of submarine earthquakes that did not cause tsunamis.

The threshold magnitude criteria are widely used by tsunami warning agencies. The current study clearly showed that the data-driven logit model had a significantly lower false discovery rate. At the same time, the balanced accuracy was about 71%, which suggests optimism for accurate tsunami forecasting based on the rapid interpretation of earthquake source parameters. The authors hope that the given model improves tsunami forecasting.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Data availability The primary data used in the current study are freely available at the indicated sources. Data that result through preprocessing are available as electronic supplementary material.

Code availability Visual Studio Code Source Editor and Python programming language were used in the current study. The following Python libraries were employed: NumPy, Matplotlib, Scikit-learn and Global-land-mask. Code is available upon request.

References

Gusev AA (1991) Intermagnitude relationships and asperity statistics. Pure Appl Geophys 136:515–527.

Gusiakov VK (2011) Magnitude-geographical criterion for operational tsunami prognosis: Analysis of application in 1958–2009. Seism Instr 47:203. https://doi.org/10.3103/S0747923911030078

Gusyakov VK, Chubarov LB (1987) Numerical simulation of tsunami excitation and propagation in the coastal zone. Izv AN SSSR. Physics of the Earth 21(11):53–64.

Ivashchenko AI, Go CHN (1973) Tsunamigennost' i glubina ochaga zemletryaseniya // V kn.: Volny tsunami. Yuzhno-Sakhalinsk: SakhKNII DVNTS AN SSSR 32:152–155 (in Russian).

Konovalov AV, Nagornykh TV, Safonov DA, Lomtev VL (2015) Nevelsk earthquakes of August 2, 2007 and seismic setting in the southeastern margin of Sakhalin Island. Russ J of Pac Geol 9:451–466. https://doi.org/10.1134/S1819714015060056

Nosov MA, Kolesov SV, Bolshakova AV, Nurislamova GN (2020) The Effect of the Choice of the Nodal Plane on Tsunami Energy Estimates. Moscow Univ Phys 75:501–506. https://doi.org/10.3103/S0027134920050197

Nosov MA, Kolesov SV, Bolshakova AV, Nurislamova GN, Sementsov KA, Karpov VA (2018) Automated system for estimation of tsunami hazard of an earthquake. Uch Zap Fiz Fak Mosk Univ 5:1850901 (in Russian). http://uzmu.phys.msu.ru/abstract/2018/5/1850901

Solov'ev SL (1972) Tsunami and Earthquake Recurrence in the Pacific Ocean. Tr SakhKNII 29:7–47 (in Russian).

Solov'ev SL, Shebalin NV (1959) Tsunami and Intensity of Kuril-Kamchatka Earthquakes. Izv AN SSSR. Ser Geofiz 8:1195–1198.

Abe K (1979) Size of great earthquakes of 1873-1974 inferred from tsunami data. J Geophys Res 84:1561–1568.

Amato A, Avallone A, Basili R et al. (2021) From seismic monitoring to tsunami warning in the Mediterranean Sea. Seismol Res Lett 92:1796–1816. https://doi.org/10.1785/0220200437

Baba T, Hirata K, Kaneda Y (2004) Tsunami magnitude determined from ocean-bottom pressure gauge data around Japan. Geophysical Research Letters 31:L08303. doi: 10.1029/2003GL019397

Boschetti L, Ioualalen M (2021) Integrated tsunami intensity scale based on maxima of tsunami amplitude and induced current. Nat Hazards 105:815–839. https://doi.org/10.1007/s11069-020-04338-5

Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21:6. https://doi.org/10.1186/s12864-019-6413-7

Cienfuegos R, Catalán PA, Urrutia A, Benavente R, Aránguiz R, González G (2018) What can we do to forecast tsunami hazards in the near field given large epistemic uncertainty in rapid seismic source inversions. Geophysical Research Letters 45:1–12. https://doi.org/10.1029/2018GL076998

Conn AR, Gould NI, Toint, PL (2000) Trust region methods, Siam, pp 169–200

Cramer JS (2002) The Origins of Logistic Regression. Tinbergen Institute Working Paper. No 2002-119/4. P. 167–168.

Davies G (2019) Tsunami variability from uncalibrated stochastic earthquake models: tests against deep ocean observations 2006–2016. Geophysical Journal International 218(3):1939–1960. doi: 10.1093/gji/ggz260

Iida K (1963) Magnitude of tsunamigenic earthquake, aftershock area and area of tsunami origin. Geophysical papers dedicated to prof. Kenzo Sassa.

Iida K (1970) The generation of tsunamis and the focal mechanism of earthquakes. In: Adams, W.M. (ed.). Tsunamis in the Pacific Ocean. Honolulu (Hawaii), East-West Center Press, 3–18.

Finch RH (1924) On the Prediction of Tidal Waves. Monthly Weather Review 52(3):147–148.

Harrell FE (2015) Binary Logistic Regression. In: Regression Modeling Strategies. Springer Series in Statistics. Springer, Cham. https://doi.org/10.1007/978-3-319-19425-7_10

Levin BV, Nosov MA (2016) Physics of Tsunamis (Springer-Verlag, New York).

Mimura N, Yasuhara K, Kawagoe S, Yokoki H, So K (2011) Damage from the Great East Japan Earthquake and Tsunami – A quick report. Mitig Adapt Strateg Glob Change 16:803–818. https://doi.org/10.1007/s11027-011-9297-7

National Geophysical Data Center / World Data Service: NCEI/WDS Global Historical Tsunami Database. NOAA National Centers for Environmental Information. doi:10.7289/V5PN93H7 (accessed 30 June 2021).

Ogata Y, Katsura K (1993) Analysis of temporal and spatial heterogeneity of magnitude frequency distribution inferred from earthquake catalogues. Geophysical Journal International 113(3):727–738. https://doi.org/10.1111/j.1365-246x.1993.tb04663.x

Satake K, Kanamori H (1991) Abnormal tsunamis caused by the June 13, 1984, Torishima, Japan, earthquake. J Geophys Res 96(B12):19933–19939. doi: 10.1029/91JB01903

Selva J, Lorito S, Volpe M et al. (2021) Probabilistic tsunami forecasting for early warning. Nat Commun 12:5677. https://doi.org/10.1038/s41467-021-25815-w

Stepnova YA, Stepnov AA, Konovalov AV et al. (2021) Predictive Model of Rainfall-Induced Landslides in High-Density Urban Areas of the South Primorsky Region (Russia). Pure Appl Geophys. https://doi.org/10.1007/s00024-021-02822-y

Tanioka Y, Satake K (1996) Fault parameters of the 1896 Sanriku tsunami earthquake estimated from tsunami numerical modeling. Geophys Res Lett 23:1549–1552. doi: 10.1029/96GL01479

U.S. Geological Survey, 2021, Search Earthquake Catalog: https://earthquake.usgs.gov/earthquakes/search/ (accessed 30 June 2021).

Users Guide for the Pacific Tsunami Warning Center Enhanced Products for the Tsunami and other Coastal Hazards Warning System for the Caribbean and Adjacent Regions (CARIBE-EWS). IOC Technical Series, No 135. UNESCO/IOC, 2017.

Web Encyclopedia on Natural Hazards, Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Tsunami Laboratory, Novosibirsk, Russia http://tsun.sscc.ru/nh/list.html (accessed 30 June 2021).

Zhu J, Baise LG, Thompson EM (2017) An Updated Geospatial Liquefaction Model for Global Application. Bulletin of the Seismological Society of America 107(3):1365–1385. doi: 10.1785/0120160198

Figure captions

Fig. 1 Maps of study area in the Northern Pacific Ocean: **a** location of earthquake epicenters, M≥6, 1960–2020; **b** location of tsunami sources, 1960–2020.

Fig. 2 The observed ('+' symbol) and calculated (black line) probability density function of the magnitude for the earthquake catalogue, M≥4.6, 1960–2020. The best estimates of the model parameters are shown at the top of the panel.

Fig. 3 Cumulative frequency–magnitude distribution for the catalogue of tsunamis, 1960–2020. Tsunami magnitude is measured as $m_t = \log_2 h_{max}$, where h_{max} is the maximum water height.

Fig. 4 Location of submarine earthquakes used in the logit model. The earthquake catalogue was merged with tsunami database and filtered.

Fig. 5 Distribution of binary variable that categorizes the tsunamigenic class of submarine earthquake on the 2D magnitude–depth plot according to the training dataset (see Sect. 2.5). Predictive models of tsunamigenic earthquakes: data-driven logit model (black solid line) and Iida-1963 (black dotted line).

Fig. 6 F1 score against the threshold probability of the logit model.

Figures

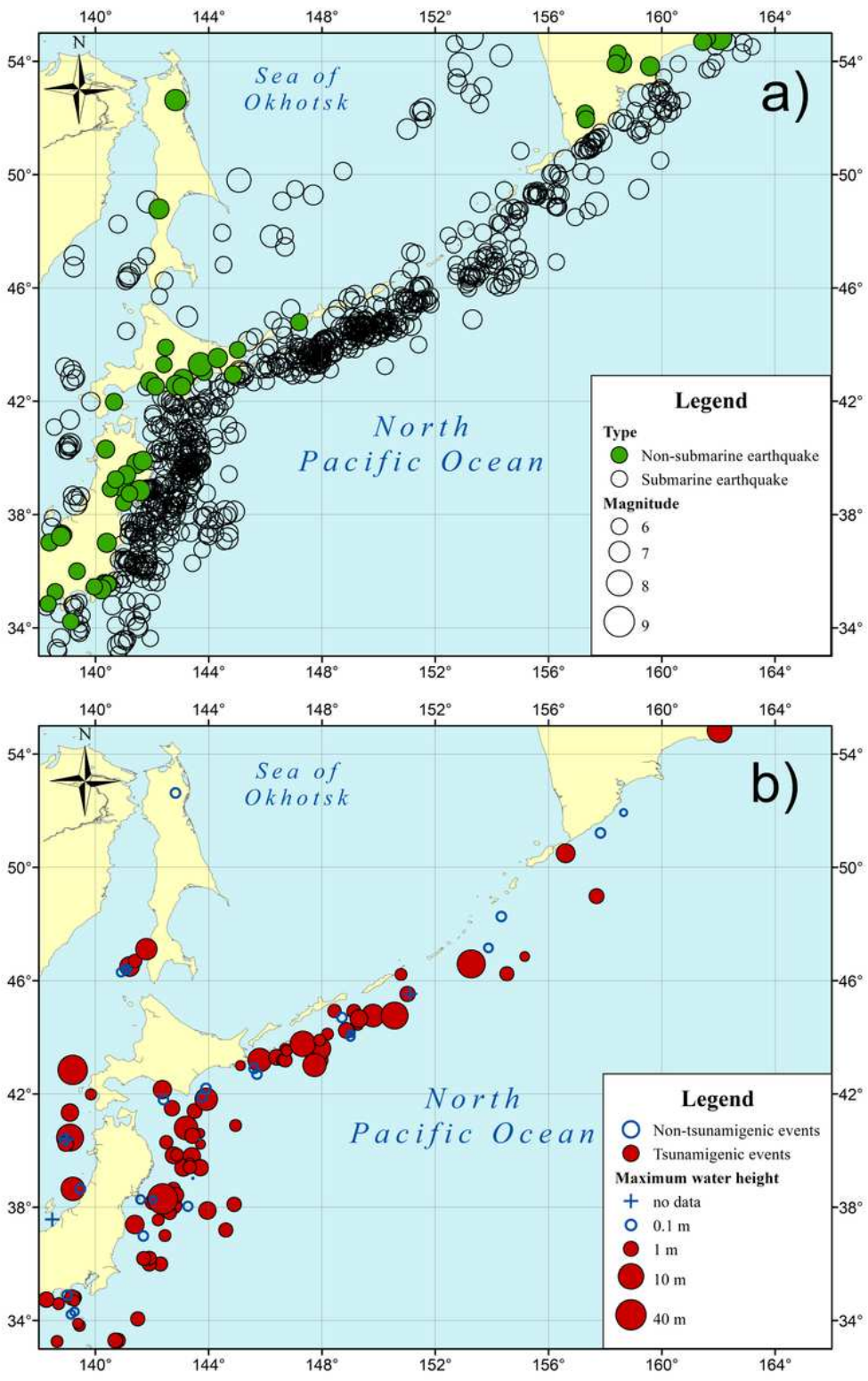

Figure 1

Maps of study area in the Northern Pacific Ocean: **a** location of earthquake epicenters, M≥6, 1960–2020; **b** location of tsunami sources, 1960–2020.

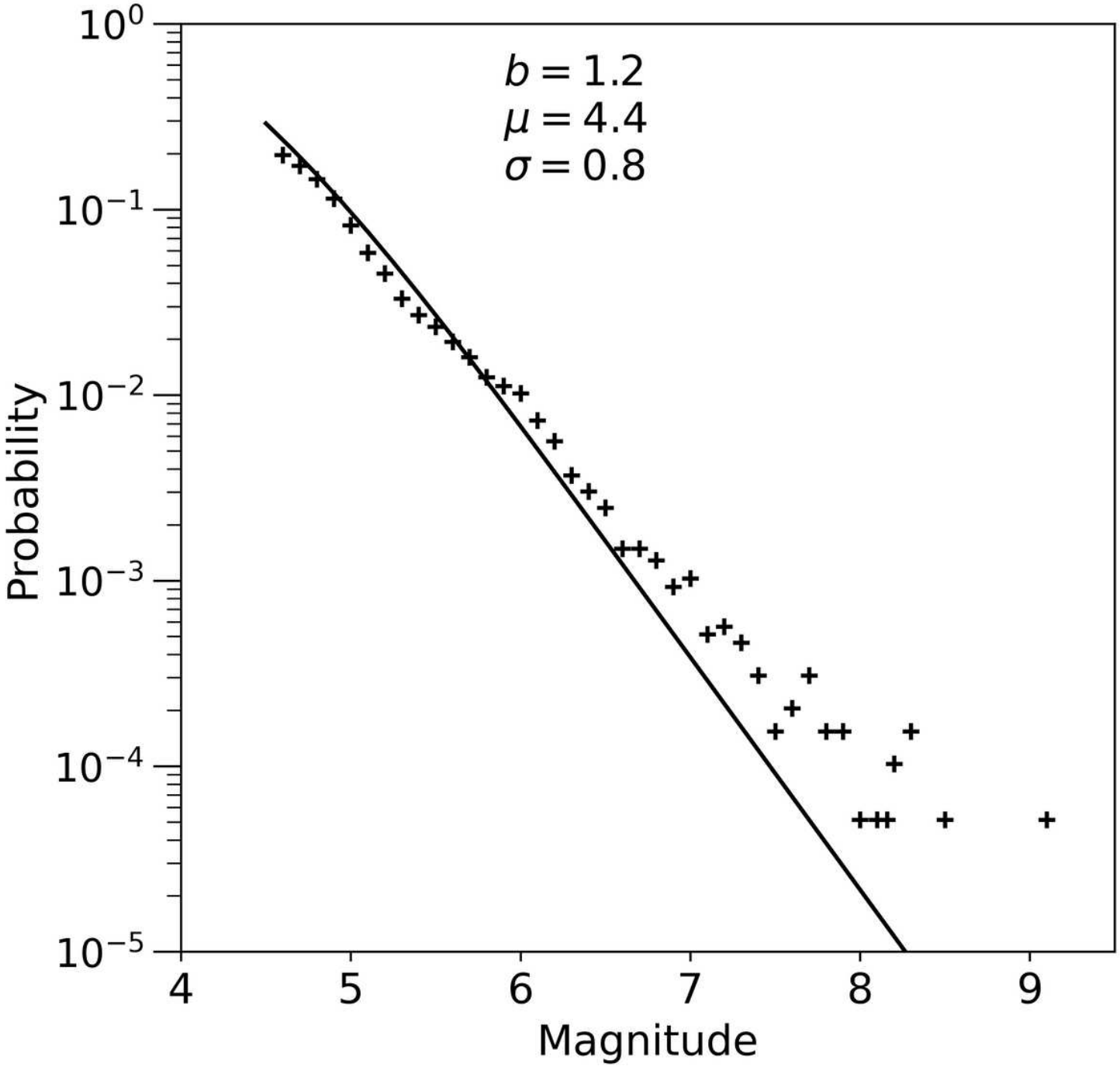

Figure 2

The observed ('+' symbol) and calculated (black line) probability density function of the magnitude for the earthquake catalogue, M≥4.6, 1960–2020. The best estimates of the model parameters are shown at the top of the panel.

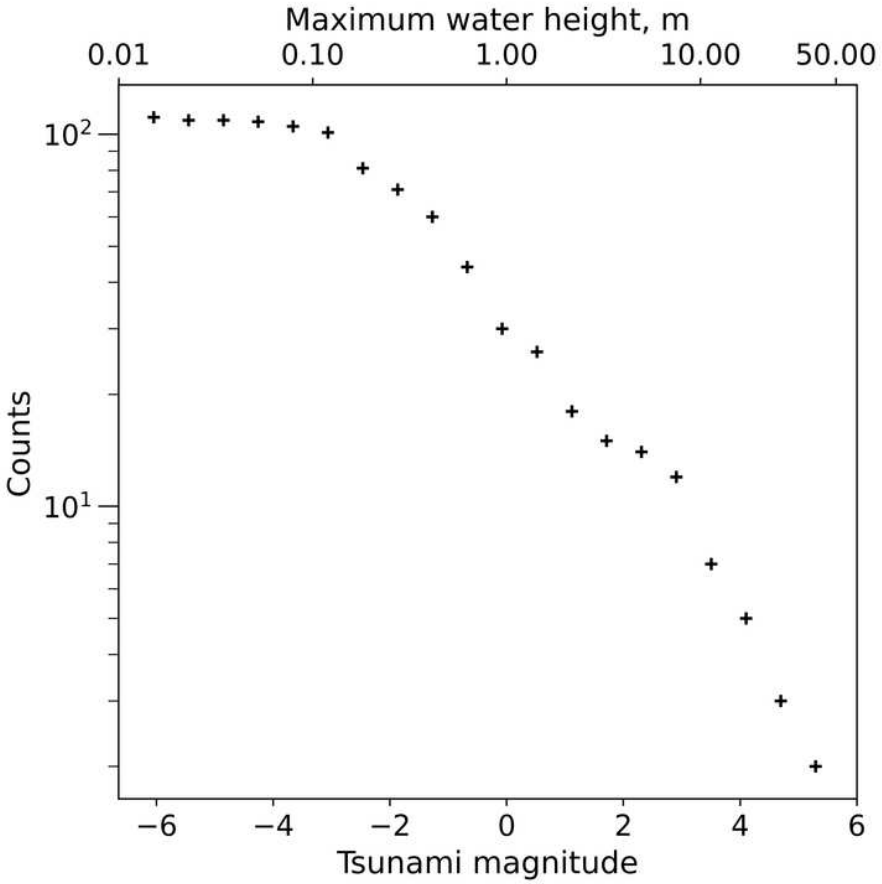

Fig. 3 Cumulative frequency–magnitude distribution for the catalogue of tsunamis, 1960–2020. Tsunami magnitude is measured as $m_t = \log_2 h_{max}$, where h_{max} is the maximum water height.

Figure 3

See figure for legend.

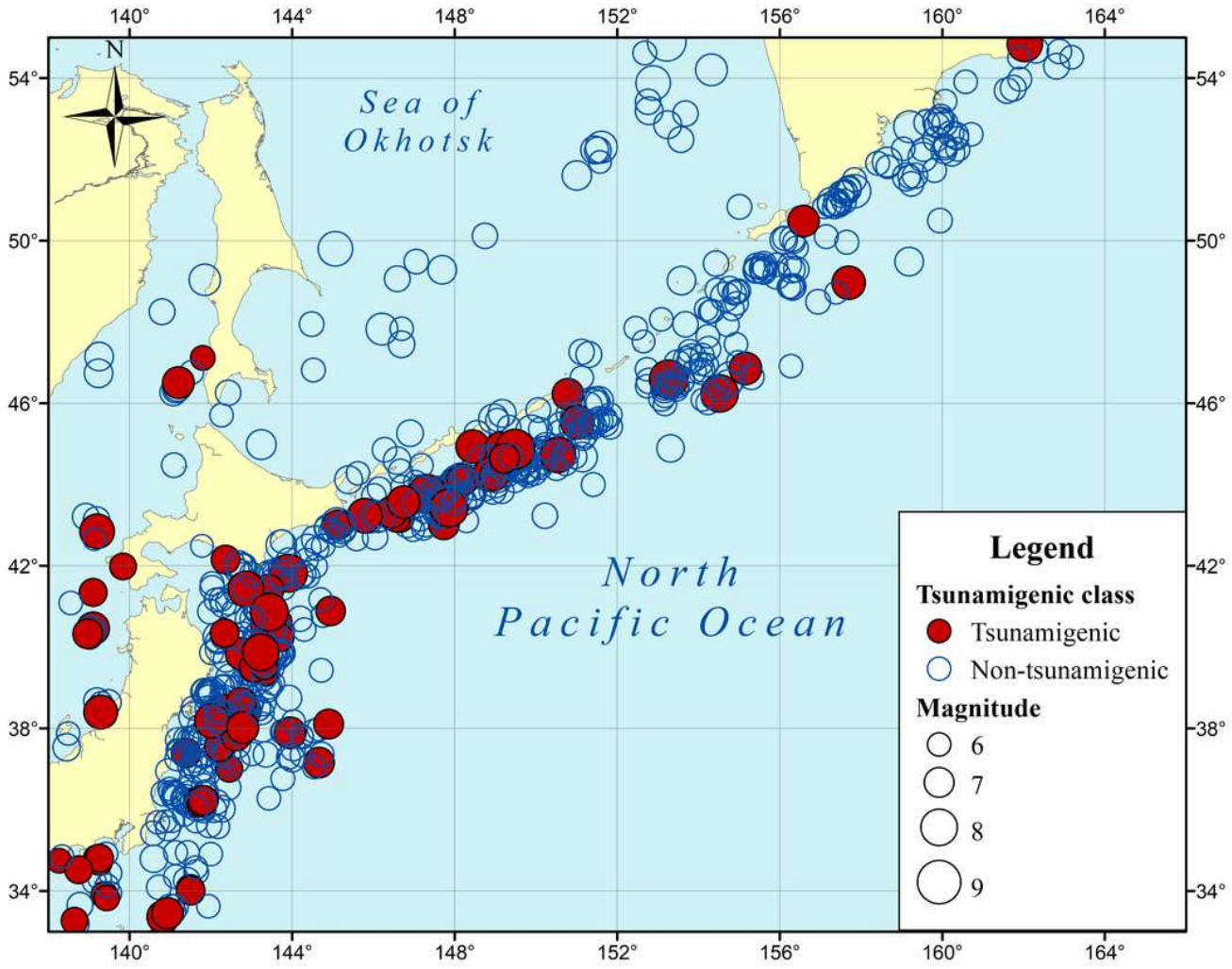

Figure 4

Location of submarine earthquakes used in the logit model. The earthquake catalogue was merged with tsunami database and filtered.

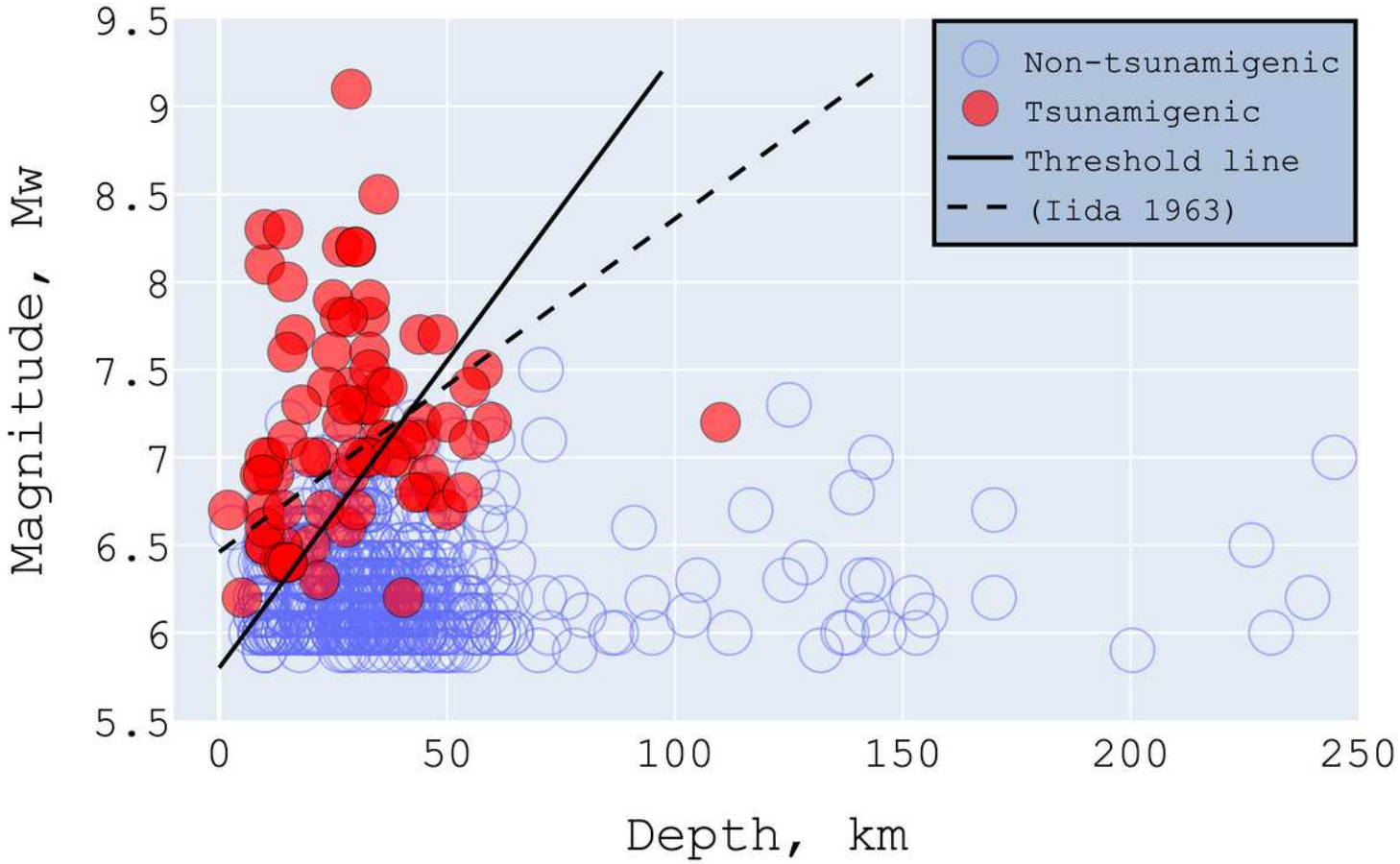

Figure 5

Distribution of binary variable that categorizes the tsunamigenic class of submarine earthquake on the 2D magnitude–depth plot according to the training dataset (see Sect. 2.5). Predictive models of tsunamigenic earthquakes: data-driven logit model (black solid line) and Iida-1963 (black dotted line).

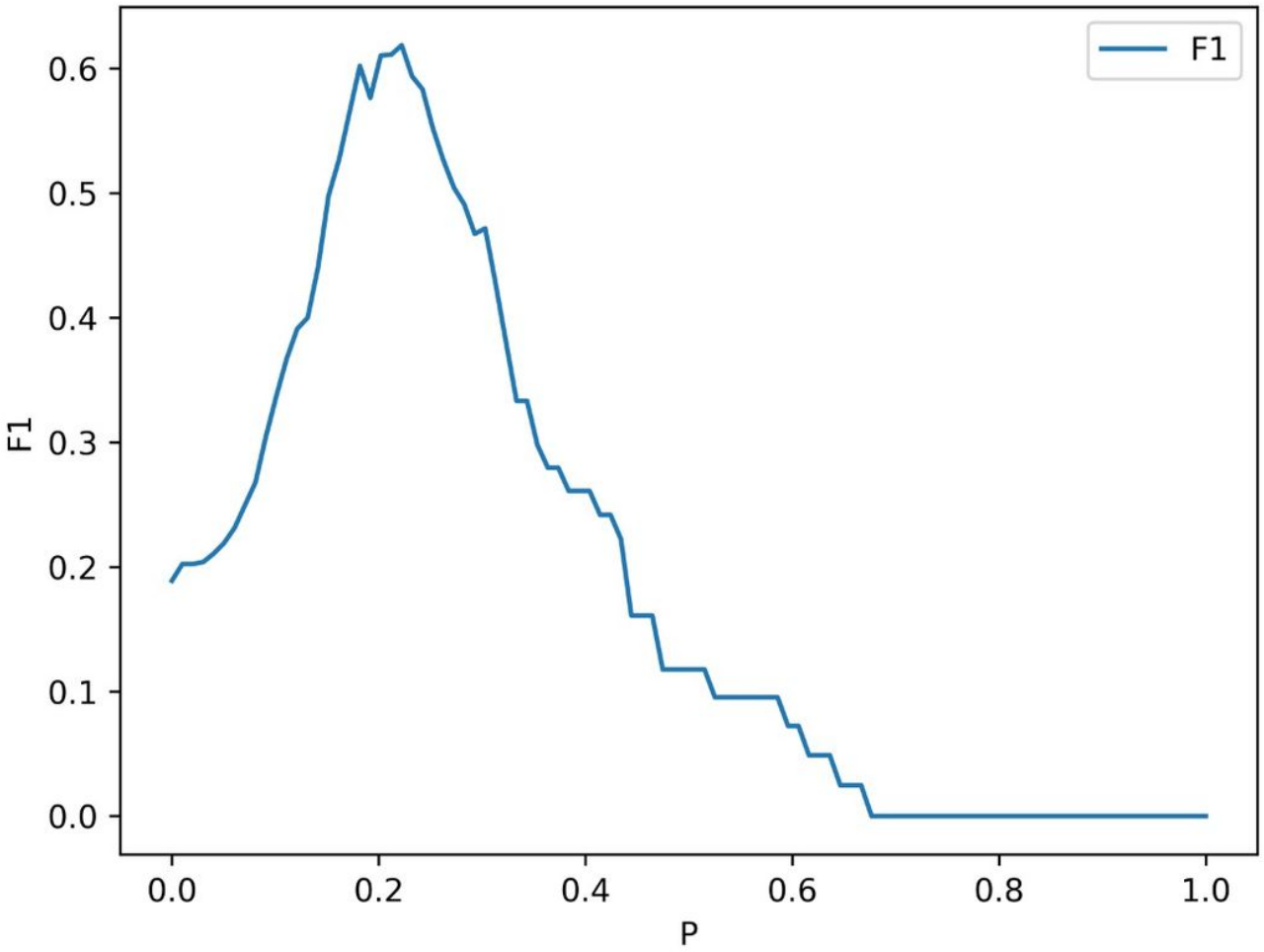

Figure 6

F1 score against the threshold probability of the logit model.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Finalearthquakelistwithtsunamiindicator.csv