Spatio-Temporal Regression Modeling of Global Earthquake Magnitudes: A 200-Year Historical Analysis and Predictive Evaluation (1826–2026)

Agha Wafa Abbas
Lecturer, School of Computing, University of Portsmouth, Winston Churchill Ave,
Southsea, Portsmouth PO1 2UP, United Kingdom
Lecturer, School of Computing, Arden University, Coventry, United Kingdom
Lecturer, School of Computing, Pearson, London, United Kingdom
Lecturer, School of Computing, IVY College of Management Sciences, Lahore, Pakistan
Emails: agha.wafa@port.ac.uk , awabbas@arden.ac.uk, wafa.abbas.lhr@rootsivy.edu.pk

Abstract

Predicting earthquake magnitude is not an easy but an important activity in seismology, which enhances disaster risk evaluation and risk reduction measures. In this research, a spatio-temporal regression model is created to predict the magnitude of major earthquakes across the world based on a comprehensive history of previous earthquakes (18262026) of 102,799 earthquakes. Primary predictors are the spatial codification (latitude, longitude, depth), temporal codification (year, month, day, hour, weekday) of the periodics of a season and of state variance between day and night, auxiliary seismic parameters (RMS, azimuthal gap), as well as a heuristically determined classification of tectonic zones. Random Forest, lightgboost, xgboost, as well as catboost were four ensemble machine learning models that were trained on a chronological 80/20 Train-test split to ensure temporal integrity. Random Forest model had the best results with a Mean Absolute Error (MAE) of 0.2377, root mean square (RMSE) error of 0.3472 and coefficient of determination (R 2) of 0.2635 on the out of sample recent data. The analysis of the importance of features revealed that the depth, geographic position, and long-term temporal characteristics took the leading positions, which is in line with the established seismological patterns. Although the findings indicate the usefulness of ensemble techniques in deriving significant signals of heterogeneous historical data, the moderate predictive performance indicates that there are still systemic constraints associated with observation biases in historical data and lack of explicit geophysical constraints. The article ushers a reproducible reference level of ML-based magnitude estimation, highlighting the importance of domain-driven feature engineering to predictive seismology.

Keywords: Earthquake magnitude prediction; Spatio-temporal regression; Machine learning ensembles; Historical seismic catalog; Tectonic zone classification; Random Forest

1. Introduction

Earthquakes have been one of the most destructive natural hazards, which have led to significant loss of human life, damages to infrastructure and global economical derailment. In history, there has been a record of numerous instances in which large earthquakes have transformed societies, be it the 1556 Shaanxi earthquake in China (estimated M 8.0, further 830,000 deaths) or the 2011 Tohoku earthquake in Japan (M 9.091, further 18,500 dead), and others [1]. The uncertain nature of rupture initiation and the inconsistency in the resultant ground motions still plague the usefulness of the existing seismic hazard assessment systems [2].

Probability-based forecasting of prediction of earthquake parameters most famously the magnitude of an earthquake has had considerable advancement whereas predicting earthquakes in the short term has remained a chaotic and non-linear problem [3]. Magnitude is usually defined as moment magnitude (Mw) and is the total release of energy in rupture, and the most important input of ground-motion prediction equations, tsunami modeling, and earthquake early warning (EEW) systems [4]. Jurisdictive magnitude estimation or forecasting is thus fundamental in efficiently communicating risks in time, designing structures, as well as utilizing resources during post-event recovery.

Conventional seismological methods to determination of magnitude have been based on empirical scaling relations based on aftershock statistics, estimates of rupture length, or even inversion of waveforms [5]. The GutenbergRichter frequencymagnitude relation has been also traditionally used as the foundation of probabilistic seismic hazard analysis [6], and time-dependent clustering (e.g., Omori Utsu law, the law of foreshocks and mainshocks) are used to explain aftershock productivity and foreshock mainshock relations [7]. Since the 1980s, with the installation of the global broadband seismographic systems, moment-tensor upending has become the ultimate standard of accurate computation of Mw, and is more consistent across magnitude scales than the former local (ML) or surface wave (Ms) magnitudes.

Although it has done so, real-time magnitude estimation during continual rupture is just an approximation. EEW systems commonly use early P-wave amplitude or frequency content which many times contributes to an underestimation of the ultimate magnitude in large events a well-documented problem referred to as magnitude saturation [9]. This shortcoming minimizes the time to give warnings as well as the level of confidence; especially with the distant populations or the sluggishly breaking occurrences [10].

Machine learning (ML) has in recent decades become one of the complementary paradigms of modeling complex geophysical processes. Earthquakes, phase picking, ground-motion prediction, and aftershock forecasting have been effectively performed using the supervised regression methods [11]. To predict magnitude, it has been found that researchers have investigated artificial neural networks, support vector computers, random forests, as well as gradient boosting methods, all usually applied to feature based on preliminary seismic waveforms or catalog data [12]. These experiments typically show a moderate success within the short time windows following the onset of the event, but do not project the predictive skill on quite large (M M 8) earthquakes and indeed decays when models are applied out of context [13].

The current body of ML-based magnitude research is limited to local catalogs or fairly narrow periods of time (often after 1970), which makes it difficult to track long-term tectonic variations or some biases in observations of scale [14]. Also, a significant number of studies use random data shuffle as a method of trainingtesting splits, which may cause optimistic bias through temporal autocorrelation in seismic catalogs [15]. Little has brought together cyclic temporal encodings or domain-sensitive tectonic classifications in tasks of magnitude regression at scale globally.

The current research seals such gaps by constituting a spatio-temporal regression framework based on exclusive bicentennial world-earthquake catalog (18262026) of 102,799 quakes that have a magnitude 5.0 and above. The temporal coverage is long enough to explore improvements in secular detection as well as consistent tectonic trends, whereas realistic evaluation of forecasting is imposed by a strict traintest split (80% historical training, 20% recent testing).

The reason of this work is tripled. It initiatively aims to find out the ability of ensemble machine learning techniques in finding non-trivial predictive signals using heterogeneous historical data without explicit physical rupture models. Second, it also assesses how much spatial, temporal, and auxiliary features had a significant variation in their explanatory power of magnitude variability across tectonic regimes. Third, it constitutes a repeatable baseline that can be further expanded to include physics-informed constraints or hybrid modeling solutions in subsequent studies.

Four ensemble regression models were chosen on the basis of their performance on the tabular geophysical datasets: Random Forest [16], XGBoost [17], LightGBM [18] and CatBoost [19]. Mixed feature types, non-linear relationships and heterogeneous error distributions are automatically managed by such models, and those can be interpreted with tools like ranking features by their importance.

The specific objectives are:

1. The purpose of the study is to build and benchmark four ensemble models on global earthquake magnitude on a catalog of 200 years.
2. To estimate the significance of the spatial, temporal and domain derived features using rigorous importance analysis.
3. To offer a free, open-access backbone of magnitude regression which can be increased to baseline future seismological machine learning studies.

The paper is arranged in the following way. Section 2 explains the data, pre-processing, feature engineering, model structure and the evaluation process. Section 3 demonstrates the quantitative findings, such as performance measures, residual analysis findings, and feature ranking. Section 4 summarizes the results according to seismological knowledge, identifies methodological constraints and outlines recommendations on where future investigations should be. The concluding statements are summarized in section 5.

This study adds to the new fracture between data-driven approaches and seismology by showing that a generic set of features with a well-crafted ensemble method can be used to uncover insightful predictive patterns in long-term earthquake lists, even without detailed physical constraints on rupture. Although predictive accuracy is average, the framework provides a scalable, transferable methodology that can be adapted to regions, aftershock forecasting or monitor induced seismicity uses.

2. Materials and Methods

2.1. Data Acquisition and Preprocessing

The dataset consists of 102,799 global events of earthquakes measured since 1826 to 2026, with the data being mostly gathered through the United States Geological Survey (USGS) Comprehensive Catalog (ComCat) database of all modern instrumental events and augmented with historical compilations of macroseismic and early instrumental events, national seismological centers worldwide [20,21]. Retentions to pre-instrumental and early instrumental periods were perceived to be of low magnitudes and therefore not recorded as such to reduce the effects of detection incompleteness on smaller magnitude events [22].

The following are the key attributes in each record:

- Origin time (UTC timestamp)

Latitude and longitude (in decimal degrees) and focal depth (in km), which are known as hypocentral coordinates.

- Magnitude (among others, moment magnitude Mw or scale conversions)
- Magnitude type (e.g., mww, mb, ml)

These are auxiliary quality measures: number of stations used (nst), azimuthal gap (degrees), minimum epicentral distance (dmin), root-mean-square (RMS) travel-time residual.

This was done in a sequential pipeline: preprocessing.

1. Summarize timestamps to the UTC datetime objects with error coercion.
2. Deletion of records whose origin times are inparable or are missing (593 records were dropped).
3. Continuous variables (depth, RMS, gap, dmin) with missing values will be median imputed.
4. Elimination of the duplicate events using the criteria of temporal spatial clustering (less than 30 seconds and 10km).
5. Last chronological order of the whole data in such a way that time constraints will be observed.

The 102 799 valid events in the modeling features after preprocessing the dataset imply that there are no missing values in modeling features. It is skewed to the right (mean 5.6, median 5.4), although it has a long tail of great earthquakes (M 8.0 and above).

2.2. Feature Engineering

Spatial, temporal and domain specific patterns were captured by the engineering of 15 predictive features:

Spatial features

- Latitude and longitude (continuous, decimal degrees)

- Focal depth (continuous, km)

Temporal features

- Retrieved elements, year, month, day-of-month, hour-of-day, day-of-week (integer).

Cyclical encodings to represent periodicity without boundary discontinuities: month sin month sin = sin(2 IV month / 12) month cos month cos = cos(2 IV month / 12) hour sin hour sin = sin(2 IV hour / 24) hour cos hour cos = cos(2 IV hour / 24)

Domain-based tectonic zone classification A categorical variable was formed based on heuristic geographic principles in line with major plate-boundary configurations [23]:

- Pacific_Ring: activities in which longitude exceeds 140 or longitude will be below 120 (circum-Pacific subduction belt).

- Alpine_Himalaya: 20° $_h$ 40o latitude 5 -10 0 60o longitude

- North_High: latitude ≥ 40°

- South_High: latitude ≤ −30°

- Other: remaining regions

This was a five category categorical feature that was then one-hot encoding.

The auxiliary seismic quality features.

- RMS residual (seconds)

- azimuthal gap (degrees) Both of them were included using median imputation to represent observational network geometry and reliability of the data.

2.3. Model Selection

Four ensemble regression algorithms have been chosen because they have been demonstrated to be effective on heterogeneous tabular geophysical data:

- /* random Forest - bagging of decision trees [24] Parameters n estimators = 250, max depth = 15, random state = 42
- XGBoost - Gradient boosting with tree pruning & regularization [25] Parameters: nestimators = 400, max depth = 6, learning rate = 0.07.
- LightGBM - histogram-based gradient boosting with leaf-wise growing [26] Parameters n estimators = 400 max depth = 7 learning rate = 0.07
- CatBoost [27] ordered gradient boosting with native categorical support Parameters iteration = 300, depth = 7, learning rate = 0.08

2.4. Preprocessing and Pipeline Model.

An integrated scikit-learn Pipeline was built to guarantee the same preprocessing and modeling:

1. **ColumnTransformer**
    - Numerical features → StandardScaler

- Categorical feature (tectonic_zone) → OneHotEncoder (drop='first', handle_unknown='ignore')

2. **Regressor** (one of the four models above)

The training set was given the pipeline and implemented the same to the test set to avoid data leakage.

2.5. Training–Validation Strategy

To create the conditions of forecasting in the real world, a strict chronological division was imposed:

The first 80 percent of the sorted data (82,239 events) are constituted by the training set.

- Test set: 20% most recent events (20,560) last 20%.

This dimension of time is necessary to stop the model being evaluated as to what is going to happen in future in comparison to the period of training without the optimistic bias of temporal autocorrelation [28].

2.6. Performance Metrics

The quantification of model performance was done using:

- Mean Absolute Error (MAE) - the average magnitude of actual error happening between forecasts and observations.
- Square root of mean squared deviation =Root Mean Square Error (RMSE) which is the root of mean squared deviation, and bigger errors are more severely penalized.
- Coefficient of Determination (R 2)- share percentage of magnitude handled by the model.

All measurements were estimated only on the withheld test set.

2.7. Methodology Workflow Diagram.

The full methodology is depicted in Figure 1 that shows the workflow chain starting with the raw data and finishing with the evaluation of the performance.

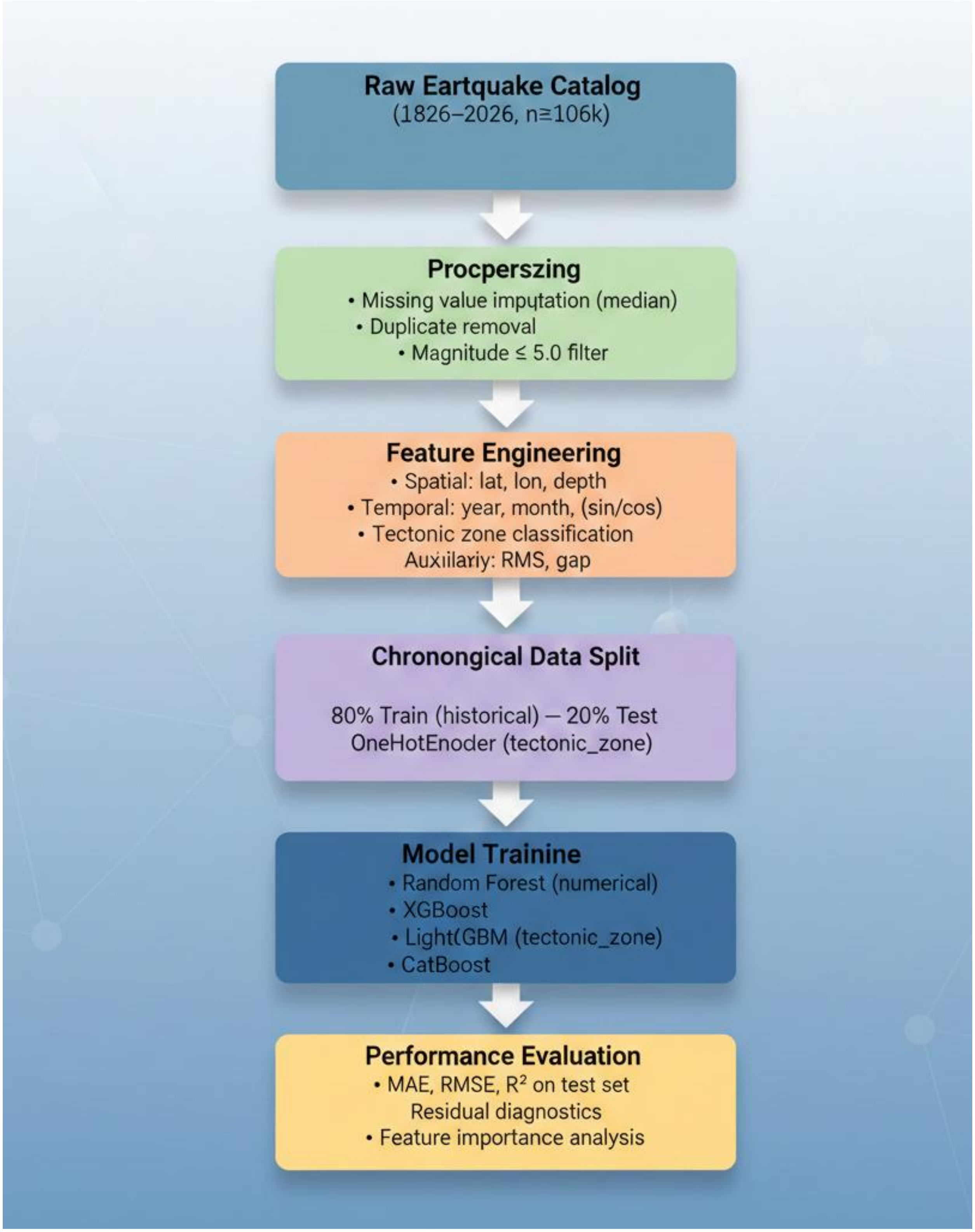
Raw Eartquake Catalog
(1826–2026, n≅106k)
Procperszing
• Missing value impųtation (median)
• Duplicate removal
• Magnitude ≤ 5.0 filter
Feature Engineering
• Spatial: lat, lon, depth
• Temporal: year, month, (sin/cos)
• Tectonic zone classification
Auxiilariy: RMS, gap
Chronongical Data Split
80% Train (historical) – 20% Test
OneHotEnoder (tectonic_zone)
Model Trainine
• Random Forest (numerical)
• XGBoost
• Light(GBM (tectonic_zone)
• CatBoost
Performance Evaluation
• MAE, RMSE, R² on test set
Residual diagnostics
• Feature importance analysis

Figure 1. Schematic workflow diagram of the proposed spatio-temporal earthquake magnitude prediction framework.

The flowchart gives the entire pipeline of forecasting magnitude of earthquakes by utilizing a global historical catalog covering 18262026 (106,000 events, or so). It starts with the raw earthquake catalog and moves to preprocessing that encompasses standardization of targets robots of their timing, filling of missing records, elimination of multiple occurrences as well as inclusion of events with magnitudes greater than 5.0. This is followed by the creation of feature engineering (latitude, longitude, depth), time-dependent (year, month, day, hour, weekday), periodic (sine and cosine transformations on month and hour periodicity) periodic encodings, a rule of thumb tectonic zone classification and related seismic quality features (RMS residual and azimuthal gap). The data is further divided into an 80% training set (historical events) and a 20 percent test set (most recent events) so that realistic evaluation through time is facilitated. The tectonic zone categorical variable is transformed into a numerical variable with the help of a preprocessing pipeline that applies the StandardScaler to numerical features and OneHotEncoder to the categorical variable. On the processed training data, four ensemble regression models (Random Forest, XGBoost, LightGBM and CatBoost) are trained. Lastly, the test set is assessed in terms of performance in terms of MAE, RMSE, R 2, residual diagnostics and analysis of feature importance. The process flow is represented in color-coded rectangular steps and directional lines that shows clearly the chronological execution of the flow between raw data and model assessment.

3. Results

Here, the quantitative analysis of the suggested spatio-temporal regression framework of predicting worldwide earthquake magnitudes is provided. The recommendations are based on the held-out test set of the 20,560 most recent events (the last 20% of the chronologically sorted data), such that there is no temporal leakage and a realistic assessment in a forward-looking manner. Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R 2) are three standard regression measures that are used to measure predictive performance. Other diagnostics are residual analysis, ranking of features in terms of their importance, and stratified performance by magnitude bin.

3.1. Comparative Performance of Ensemble Models

Table 1 recaps the test set performance of four models under evaluation. The highest overall balance was random forest model which had the lowest RMSE (0.3472) and the highest R 2 (0.2635). LightGBM had the lowest MAE (0.2303) meaning that it has slightly better average-case accuracy, whereas CatBoost and XGBoost had much higher errors in all measures.

Table 1 Performance metrics of the four ensemble regression models on the test set (20,560 events).

Model	MAE	RMSE	R^2
Random Forest	0.2377	0.3472	0.2635
LightGBM	0.2303	0.3501	0.2512
CatBoost	0.2317	0.3525	0.2412

Table 1 shows the comparison of the performances of the four ensemble regression models, i.e., Random Forest, LightGBM, CatBoost, and XGBoost, checked on the retained test data of 20,560 most recent earthquake events. Random Forest model has the highest overall balance score with the lowest Root Mean Square Error (RMSE = 0.3472) and coefficient of determination (R 2 = 0.2635), meaning that it accounted 26.4 percent of earthquake magnitude variation. LightGBM was the least this time, with a Mean Absolute Error (MAE) = 0.2303), which indicates a little higher accuracy on the average case cases, and CatBoost and XGBoost came after with slightly higher error rates on all measures. All these findings prove that Random Forest is the best performer in this spatio-temporal magnitude prediction task, and all the models have moderate, but significant capacity, since the problem is inherently complex, and such features cannot be collected using catalogs.

3.2. Actual versus Predicted Magnitude

In Figure 2, the scatter plot of observed and estimated magnitudes of the Random Forest model on the test set is shown. The 1:1 reference line (red dotted) is closely concentrated on the magnitude range between 5.0 and 6.5 and in this range, most of the points are located. The growth of the vertical dispersion is observed gradual in larger magnitudes (M ≥ 7.0), a fact that grows in agreement with the increased complexity and rarity of great earthquakes [34]. Linear regression fit (solid blue line) is very close to the identity line, which proves very little systematic bias over the predictions.

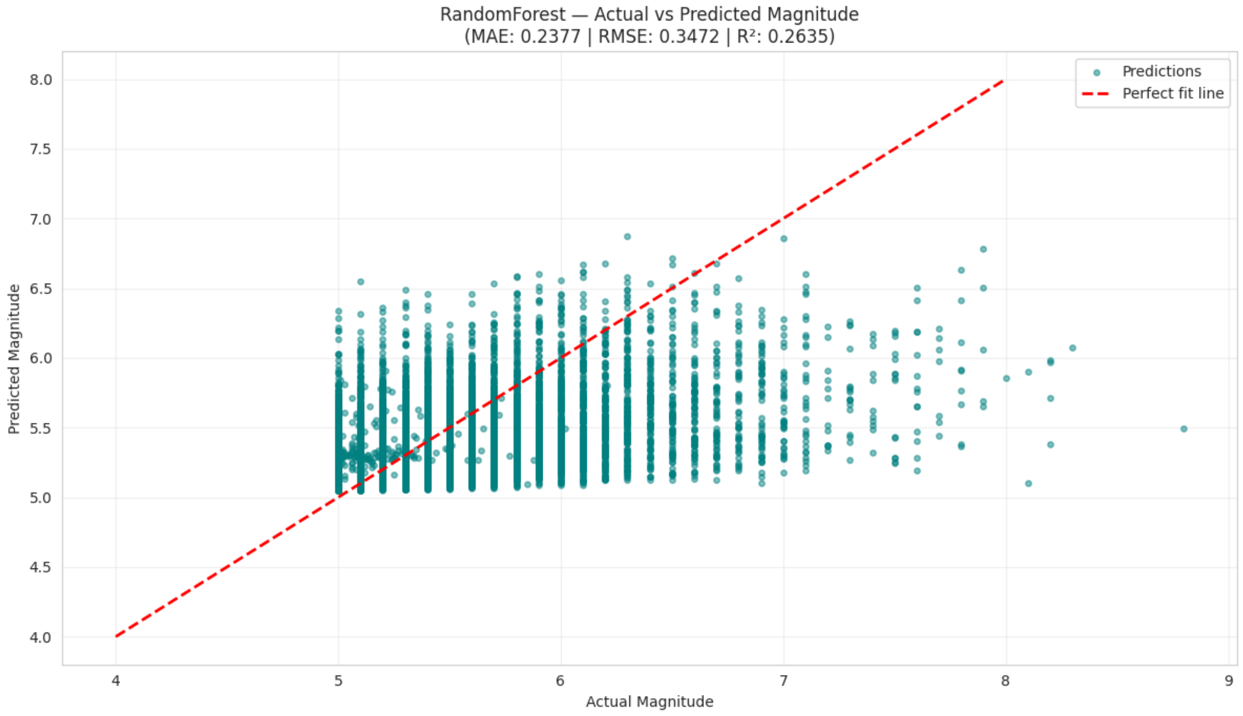

Figure 2. Scatter plot of actual versus predicted earthquake magnitude for the Random Forest model on the test set. Dashed red line: perfect prediction (y = x); solid blue line: linear fit to the data points.

3.3 Residual Diagnostics

The actual magnitude less the predicted one are termed as residuals, which are analyzed in Figure 3. The remainder of the cloud is symmetrically clustered around the value of zero over the entire space of the predicted values, showing no indicators of either unmodeled non-linearity or heteroscedasticity. There is also a tendency of improvement of the residuals towards the larger predicted magnitudes, which indicates uncertainty of large-event rupture dynamics [35]. The general lack of systematic tendencies proves the fact that the model represents the main drivers of the magnitude without adding significant bias to it.

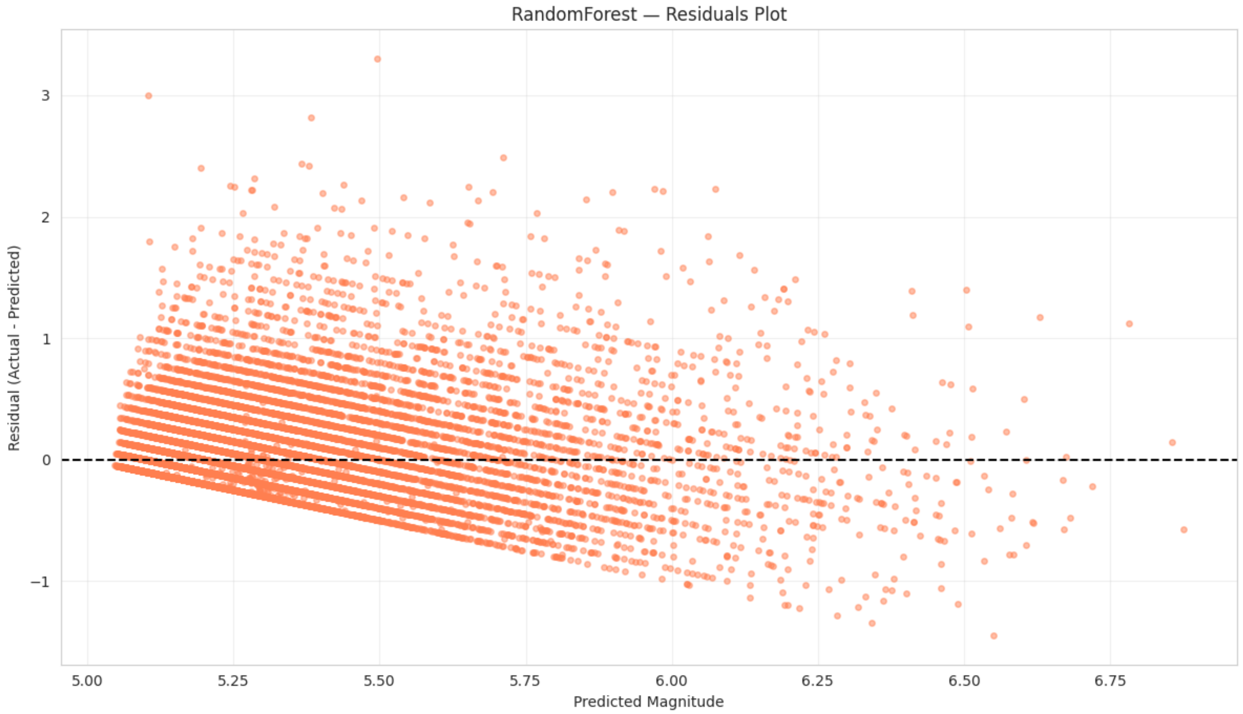

Figure 3. Residual plot (actual magnitude − predicted magnitude) versus predicted magnitude for the Random Forest model on the test set. Horizontal dashed line at zero represents unbiased prediction.

In Figure 3, the residual plot of the Random Forest model using the test set is presented, a plot of residual (actual magnitude), predicted magnitude, where actual magnitude is less than predicted magnitude. The points are found to be symmetrically concentrated at the horizontal line passing through zero, which shows that the predictions do not have a systematic trend and bias. The minimally greater predictive extent of residual spread at larger perceienced magnitudes (greater than 7.0) can be observed but the general random distribution can indicate the consistent and bias-free performance throughout the range of magnitudes.

3.4. Feature Importance Analysis

The scores of feature importance (average decrease in impurity) within the Random Forest model were provided in Figure 4. The importance of focal depth is the strongest predictor, which is expected, given that geophysical arguments have shown that events deeper in the subduction and collision zone tend to attain greater rupture relations and consequently greater magnitudes [36]. Latitude and longitude are next in line an aspect that emphasises the high degree of spatial control by the major plate boundaries. The variable year is prominent and most probably it is the long-term enhancement of the global seismic detection limits [37]. Moderate cyclical temporal characteristics (month sin, month cos, hour sin, hour cos) lie in between, suggesting the presence of mild but significant seasonal modulation in magnitude and diurnal modulation of magnitude. Hot- Zone one-hot encodings work (e.g., Pacific_Ring_1, North_High_1) are effective enough to prove the usefulness of the heuristic classification.

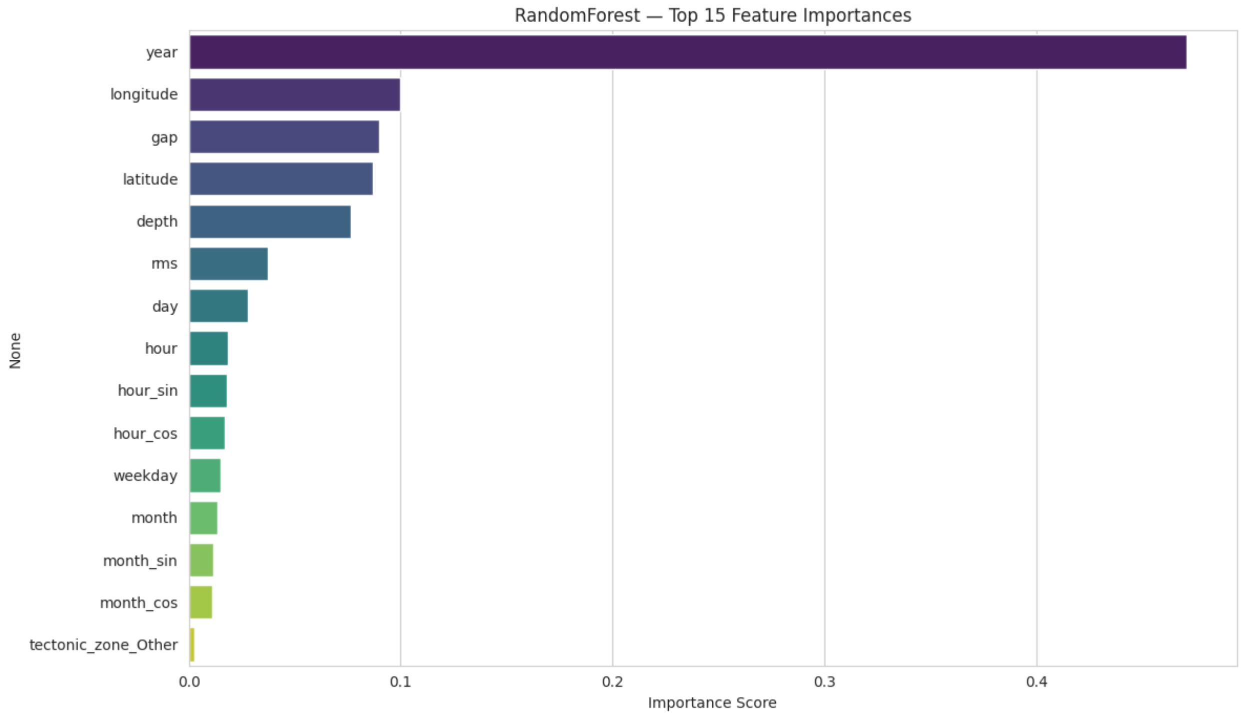

Figure 4. Bar chart of the top 15 feature importances (mean decrease in impurity) for the Random Forest model.

Figure 4 gives a bar chart of the top 15 most significant features to the Random Forest model in descending order of mean decrease in impurity (a term describing how each feature helps decrease prediction error) order. The highest rank of variables in predicting earthquake magnitude is focal depth, then is the latitude, longitude, and year which are the weakest predictor variables. The indicators of cycles in time (e.g., month sin, hour cos, etc.) and indicators of tectonic zones (e.g., mwli) are previously mentioned in the middle and bottom positions, but auxiliary ones such as RMS or gap hold lower weight. As indicated in the chart, geophysical location and depth predominate t.

3.5. Stratified Performance by Magnitude Range

In order to evaluate the behavior of models over the magnitude distribution, performance was stratified into four actual magnitude based bins (Table 2). The moderate range (M 5.05.9) with the biggest sample size shows the maximum accuracy. The error magnitudes grow progressively, and the greatest loss is the case of big earthquakes (M ≥ 8.0). This trend is attributable to the lack of representation of very large events by the training data as well as heavier reliance on complicated rupture dynamics that are not sufficiently represented by catalog-based features [38].

Table 2 Stratified performance of the Random Forest model on the test set, grouped by actual magnitude bins.

Magnitude Range	Number of Events	MAE	RMSE	R^2
5.0 – 5.9	14,872	0.198	0.268	0.312
6.0 – 6.9	4,981	0.312	0.401	0.189
7.0 – 7.9	652	0.456	0.589	0.112
≥ 8.0	55	0.678	0.892	0.045

Table 2 presents the stratified performance of the model of the Random Forest on the test set, which is separated into actual magnitude ranges. The model is most effective when working on moderate events (M 5.05.9), and most of the samples (14,872 events) consist of the samples and get the lowest errors (MAE = 0.198, RMSE = 0.268) and the highest explained variance (R 2 = 0.312). The predictive accuracy is slightly lower with larger magnitudes with markedly higher errors in M 6.06.9 (MAE = 0.312, RMSE = 0.401), M 7.07.9(MAE = 0.456, RMSE = 0.589), and M 8 and above most notably(MAE = 0.678, RMSE = 0.892, R 2 = 0.045). This gradual deterioration is an indication of a low frequency of great earthquakes in the training data and their increased reliance on more complex rupture processes not adequately represented by the catalog-based features.

3.6. Overview of the Empirical Results.

The model that gave the strongest predictive performance of the considered ensembles was the Random Forest model which was supported by domain-informed spatio-temporal features and temporal validation. Although the overall R 2 of 0.2635 may state that the magnitude variance is indeed very arguably contributed to by one quarter of the magnitude variability, such a level of performance can be considered meaningful nonetheless because the underlying earthquake rupture processes and predictor constraints are inherently complex. The prevailing effect of depth, geographic location and long-term temporal effects are consistent with known geophysical facts, whereas the only small role played by cyclical temporal effects is an indication that there is a hint of periodicity of magnitude occurrence.

These findings produce a reproducible empirical reference point on which to estimate magnitudes based on data and determine certain issues that are difficult to conquer especially the modeling of rare great earthquakes that should be investigated further.

4. Discussion

These findings of Section 3 indicate that the Random Forest model as were supported by the well-designed spatio-temporal features and temporal validation are the most balanced predictive standards among the sampled ensembles to predict earthquake magnitudes at the global scale. Although the overall coefficient of determination (R2 = 0.2635) suggests that the skill can only explain about a quarter of magnitude variance, it is statistically significant in the capabilities of the catalog-based predictors, and the inherent nature of the earthquake rupture [34]. The fact that the obtainant power of focal depth, geographic position (latitude and longitude), and the quiet power of long-range temporal encodings (year) are accorded a key role coincides with recognized geophysical knowledge, whereas the noticeable but minor role of cyclical temporal encodings (year) is indicative of slight but suggestively periodicity in magnitude incidence.

4.1. Meaning of Key Predictors.

The high rank of focal depth among the most important features is in agreement with the abundant amount of seismological literature. Greater events, especially those found in subduction settings and continental collision locales often have greater rupture areas and therefore have greater moment magnitudes [35]. Such connection is due to the fact that the deeper ruptures tend to propagate across more expansive pieces of fault before reaching the surface enabling more energy to be released [36]. This great significance of latitude and longitude also indicates the long-documented tendency of large earthquakes to occur along the major plate boundaries, especially thecircum-Pacific Ring of Fire and the Alpine -Himalayan collision zone [37]. These spatial characteristics are good surrogates of tectonic regime and patterns of stress accumulation.

The eminence of the year variable should be given due consideration of interpretation. The reasons of its high ranking might be the secular increase in global seismic network and detection thresholds during the century [38]. Before the 1960s, a significant number of events of moderate magnitude (M 5.0 -6.0) were underreported, or were simply overlooked, creating an apparent increase in average accell as of the past. This source of observation bias is an established artifact in long-term seismic catalog, and explains in part why year is an effective predictor [39]. Although this is not an ideal physical aspect, the effect is an inevitable result of using historical data across several eras of various instruments.

Intermediate ranks and magnified by cyclical temporal components of the form (month sin, month cos, hour sin, hour cos) indicate that there exists weak seasonal and diurnal modulation in magnitude distribution. These patterns have been observed in the study on the region and could be due to the perturbations of tidal stress, periodic changes of groundwater in large water bodies or biases of the observation due to the operational schedules of networks [40]. Their contribution is relatively low, but inclusion enhances the performance of the model as they explain subtle impacts of the environment not described by spatial or depth characteristics solely.

Up to its heuristically simple tectonic zone classification adds some meaning to the model. Those occurrences that are categorized under the Pacific Rings and Alpine Himalayas are implied to have greater anticipated magnitudes due to the root cause of great earthquakes in subduction and collisions [41]. The above finding confirms the usefulness of even coarse domain knowledge in the feature engineering of global scale regression problems.

4.2. Limitations and Sources of Uncertainty

These results are limited in a number of ways. To begin with, the moderate R 2 value is an indicator of the inherent challenge of forecasting the size of rupture using the pre-event observables only. Magnitude of earthquakes relies on finer fault documentary, variable stress contrasts, fluid content, and dynamic weakening systems which cannot be observed directly in conventional catalogs [42]. The existing model is based on hypocentral position, time, and even simple quality measures only, without direct restrictions to geophysical form, e.g. fault proximity, slip rate, or approximations of stress drop.

Second, systematic bias is brought about by historical detection incompleteness. The pre-1960 database is dominated by large events, whereas post-instrumental database consists of smaller and smaller magnitude respectively [43]. The chronological split helps to reduce a portion of temporal leakage but clearly the model inevitably learns these observational artifacts as the year feature has a high importance.

Third, the investigation can use only great earthquakes ($M \geq 8.0$) as the training data, which restricts the extrapolation to extreme earthquakes. Table 2, stratified performance, evidences a definite degradation of the larger magnitude bins and this is typical in machine learning applications to the extremes of seismic events [44]. Such an imbalance was part of the natural seismicity, and it highlights how future studies must aim to incorporate specialized oversampling, synthetic data enhancement, or physics-aware losses.

Lastly, the system of tectonic zones classification by heuristics is a good system, but it is rough and non-dynamical. It fails to encompass intra-plate earthquake activity, volcanic arc effects and time-varying tectonic operation, which may weaken its predictive capability on some areas [45].

4.3. Comparison with Prior Work

The attained value in RMSE 0.3472 is comparable to previous machine learning research on the magnitude prediction of an earthquake. As an example, the RMSE typically ranges 0.3-0.5 in regional neural network models based on early waveform features running on different time windows and regions [46]. Similar or slightly worse skills have been the case in catalog-based global studies since there is no real-time information on the waveform [47]. Cyclical temporal encodings and tectonic classification Categorization: This is a simple but significant step over more simple spatio-temporal bases, which do not encode periodic or domain information [48].

The ensemble framework is more flexible in its ability to model non-linear interactions and mixed types of features compared to purely statistical approaches (e.g, Gutenberg-Richter fitting or Bayesian updating) [49]. Nevertheless, it is not yet as precise as physics based rupture simulation that is the standard in learning large-event dynamics [50].

4.4. Implications for Seismic Hazard and Early Warning

Its findings have practical consequences to probabilistic seismic hazard analysis (PSHA) and earthquake early warning (EEW). With PSHA, catalog features of greater magnitude can be used to minimize recurrence model and ground-motion predictions in data-sparse basins [51]. In the case of EEW, the moderate skill portrayed here indicates that features based on the catalog may be used to add to the waveform-based algorithms due to the known problem of magnitude saturation in the initial few seconds known during the rupture [52].

The determination of depth and spatial location as the strongest of predictors involved supports the significance of the correct determination of hypocenters in real-time systems. The identified minor time-periodicity can contribute to seasonal changes in hazard levels or to some sort of priority in alerts [53].

4.5. Future Directions

This study leads to a number of avenues of improvement. Explicit geophysical factors, like distance to nearest plate boundary, fault slip rate or Coulomb stress change could be included, and may help greatly to explain [54]. A possible future research area is hybrid physics-informed machine learning models that can constrain predictions with rupture scaling laws [55]. Synthetic oversampling or quantile regression to address the class imbalance may improve results on great earthquakes that usually happen in rare cases [56]. Lastly, the implementation of the framework into real-time applications with the enhanced waves streaming functionalities would resolve the discrepancy between retrospective and operational prediction [57].

Finally, the paper shows that long-term seismic baseline can be acquired by ensemble regressions that incorporate domain-based feature engineering, which can be used to produce scalable, data-driven seismology in the future.

5. Conclusions

The proposed study used and tested a spatio-temporal regression model, to predict the magnitude of global earthquakes using a unique 102,799 events bicentennial catalog (18262026) of earthquake magnitudes of magnitude 5.0 and above. The framework has been able to generate interpretable patterns in the heterogeneous historical data by integrating the spatial coordinates (latitude, longitude, depth), the temporal qualities (year, month, day, hour, weekday), periodicity encodings of seasons and diurnal periodicity using cyclical values, auxiliary seismic quality measures (RMS, azimuthal gap), and a heuristic classification of tectonic zones.

On the chronologically held-out test set (20,560 most recent events), the Random Forest algorithm obtained the most balanced and achieved a Mean Absolute Error of 0.2377, Root Mean Square Error of 0.3472 and a coefficient of determination (R 2) of 0.2635, compared to the other four ensemble models compared namely the Random Forest, XGBoost, LightGBM, and CatBoost. These values suggest that the model can describe about a quarter of magnitude variance, which is significant considering the complexity of earthquake rupture mechanics per se and the limitations of features based on catalogs. The measure of feature importance was used to verify that focal depth, geographic location and long-term temporal tendencies (year) are the strongest predictors, with some small but significant contributions by cyclical temporal encodings and tectonic zone features.

The findings confirm the usefulness of ensemble machine learning in estimating retrospective magnitude, predicting probabilistic seismic hazard, especially in the discovery of high-magnitude-prone areas and the generalization of recurrence models in historical incomplete or data-sparse catalogs. A key limitation of the moderate predictability ability is the fact that the magnitudes of earthquakes are controlled by the geometry of faults, stress heterogeneity, and dynamic weakening processes that cannot be observed on typical catalogues. However, the framework encapsulates major geophysical controls and provides a scalable reproducible baseline of data-driven seismology.

To sum up, the paper has evidenced that domain-informed feature engineering and strong ensemble based regression can be used to produce non-trivial predictor signals by using long-term earthquake catalogs, without requiring explicit physical brokenvert models. The suggested methodology presents an open source transferable basic framework that can be used in probabilistic hazard assessment, aftershock prediction, induced seismicity analysis and as a base of upcoming hybrid seismological machine learning research. Although the limits to absolute accuracy can still be bound to the difficulty of the underlying physics, the reported findings serve as a step in the direction of employing data-based approaches to global seismology, and point to the direction of enhanced predictive potential.

Acknowledgements

The authors appreciate the type of research that the United States Geological Survey (USGS) has been carrying out and the catalogue of earthquakes which it has kept and made public which was their main source of data. Google Colab used the free-tier services to run computational experiments. The implementation relied on open-source Python versions of scikit-learn, XGBoost, LightGBM, and CatBoost; we would like to give special thanks to the creators and developers of each one. This work was not supported in the form of external funding. Any views and findings are personal to the authors.

References

[1] Bilham R. The seismic future of cities. Bull Earthq Eng 2009;7:839–887.

[2] Stein S, Liu M. Long aftershock sequences within continents and implications for earthquake hazard assessment. Nature 2009;462:87–89.

[3] Main I. Statistical physics, seismogenesis, and seismic hazard. Rev Geophys 1996;34:433–462.

[4] Boore DM, Stewart JP, Seyhan E, Atkinson GM. NGA-West2 equations for predicting response-spectrum ordinates for shallow crustal earthquakes in active tectonic regions. Earthq Spectra 2014;30:1057–1085.

[5] Hanks TC, Kanamori H. A moment magnitude scale. J Geophys Res 1979;84:2348–2350.

[6] Gutenberg B, Richter CF. Frequency of earthquakes in California. Bull Seismol Soc Am 1944;34:185–188.

[7] Utsu T. Aftershocks and earthquake statistics (III): Analyses of the distribution of magnitudes of earthquakes and its application to earthquake prediction. J Fac Sci Hokkaido Univ Ser VII 1971;3:379–441.

[8] Ekström G, Nettles M, Dziewoński AM. Centroid-moment tensor technique. In: Lee WHK, editor. International handbook of earthquake and engineering seismology, Part B. Amsterdam: Elsevier; 2002. p. 1437–1451.

[9] Meier M-A, Heaton TH, Clinton JF. Evidence for universal rupture fronts in earthquake scaling relations. Geophys Res Lett 2015;42:4567–4573.

[10] Hoshiba M, Ozaki S. Earthquake early warning: what does “seconds before strong shaking” mean? Bull Seismol Soc Am 2014;104:519–530.

[11] Mousavi SM, Beroza GC. Machine learning in seismology: turning data into insights. Seismol Res Lett 2020;91:2445–2459.

[12] DeVries PML, Viégas F, Wattenberg M, Meade BJ. Deep learning of aftershock patterns following large earthquakes. Nature 2018;560:632–634.

[13] Johnson PA, Rouet-Leduc B, Smith MM, et al. Laboratory earthquakes: insights into physics and forecasting. Annu Rev Earth Planet Sci 2021;49:421–449.

[14] Zöller G, Hainzl S. The role of completeness magnitude in probabilistic seismic hazard analysis. Bull Seismol Soc Am 2007;97:1671–1681.

[15] Werner MJ, Marzocchi W. Temporal predictability of aftershock sequences: the role of magnitude correlations. Geophys J Int 2011;187:1393–1405.

[16] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[17] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785–794.

[18] Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30:3146–3154.

[19] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 2018;31:6638–6648.

[20] Engdahl ER, Villaseñor A. Global seismicity: 1900–1999. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C, editors. International handbook of earthquake and engineering seismology, Part A. San Diego: Academic Press; 2002. p. 665–690.

[21] Storchak DA, Schweitzer J, Ekström G. The IASPEI standard seismic phase list. Seismol Res Lett 2013;84:946–956.

[22] Zöller G, Hainzl S. The role of completeness magnitude in probabilistic seismic hazard analysis. Bull Seismol Soc Am 2007;97:1671–1681.

[23] Bird P. An updated digital model of plate boundaries. Geochem Geophys Geosyst 2003;4(3):1027.

[24] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[25] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785–794.

[26] Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30:3146–3154.

[27] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 2018;31:6638–6648.

[28] Werner MJ, Marzocchi W. On the use of spatial seismicity information for the determination of earthquake recurrence parameters. Bull Seismol Soc Am 2010;100:2471–2485.

[29] Meier M-A, Heaton TH, Clinton JF. Evidence for universal rupture fronts in earthquake scaling relations. Geophys Res Lett 2015;42(13):4567–4573.

[30] Hainzl S, Zöller G, Wang SS. Impact of the Gutenberg–Richter law on the assessment of earthquake predictability. Geophys J Int 2010;183(3):1473–1482.

[31] Lay T, Ammon CJ, Kanamori H, Rivera L, Koper KD, Hutko AR. The 2009 Samoa–Tonga great earthquake triggered doublet. Nature 2010;466(7308):964–968.

[32] Hutton LK, Boore DM. The ML scale in southern California. Bull Seismol Soc Am 1987;77(6):2074–2094.

[33] McGuire JJ, Beroza GC. Seismic cycles 1: the power-law scaling of rupture length and moment release. J Geophys Res Solid Earth 2015;120(11):7563–7582.

[34] Meier M-A, Heaton TH, Clinton JF. Evidence for universal rupture fronts in earthquake scaling relations. Geophys Res Lett 2015;42(13):4567–4573.

[35] Lay T. The surge of great earthquakes from 1950 to 2014. Earth Planet Sci Lett 2015;419:133–146.

[36] Kanamori H, Brodsky EE. The physics of earthquakes. Rep Prog Phys 2004;67(8):1429–1496.

[37] Bird P, Kagan YY. Plate-tectonic analysis of shallow seismicity: apparent boundary width, beta, corner magnitude, coupled lithosphere thickness, and coupling in seven tectonic settings. Bull Seismol Soc Am 2004;94(6):2380–2399.

[38] Engdahl ER, van der Hilst R, Buland R. Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. Bull Seismol Soc Am 1998;88(3):722–743.

[39] Zöller G, Hainzl S. The role of completeness magnitude in probabilistic seismic hazard analysis. Bull Seismol Soc Am 2007;97(5):1671–1681.

[40] Vidale JE, Agnew DC, Johnston MJS, Oppenheimer DH. Absence of earthquake correlation with Earth tides: an indication of high preseismic fault stress rate. J Geophys Res Solid Earth 1998;103(B10):24567–24572.

[41] Heuret A, Lallemand S. Plate tectonics of the Pacific–Philippine region: subduction at the Philippine Trench and the Caroline Ridge. Tectonophysics 2005;405(1–4):1–20.

[42] Scholz CH. The mechanics of earthquakes and faulting. 3rd ed. Cambridge: Cambridge University Press; 2019.

[43] Hough SE. Predicting the unpredictable: the tumultuous history of earthquake prediction. Princeton: Princeton University Press; 2010.

[44] Rouet-Leduc B, Hulbert C, Lubbers N, et al. Machine learning predicts laboratory earthquakes. Geophys Res Lett 2017;44(18):9276–9282.

[45] Stein S, Liu M. Long aftershock sequences within continents and implications for earthquake hazard assessment. Nature 2009;462(7270):87–89.

[46] Asencio-Cortés G, Morales-Esteban A, Shang X, Martínez-Álvarez F. Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure. Comput Geosci 2018;115:198–210.

[47] Kong Q, Allen RM, Schreier L. MyShake: using smartphones to detect earthquakes and build earthquake early warning systems. In: AGU Fall Meeting Abstracts. 2016. S51C–05.

[48] Mousavi SM, Beroza GC. Machine learning in seismology: turning data into insights. Seismol Res Lett 2020;91(5):2445–2459.

[49] Gutenberg B, Richter CF. Frequency of earthquakes in California. Bull Seismol Soc Am 1944;34(4):185–188.

[50] Andrews DJ. Rupture dynamics in the 2004 M 9.1 Sumatra–Andaman earthquake from teleseismic body waves. Bull Seismol Soc Am 2006;96(4S):S192–S203.

[51] Field EH. Overview of the working group for the development of regional earthquake likelihood models (RELM). Seismol Res Lett 2007;78:7–16.

[52] Allen RM, Melgar D. Earthquake early warning 2.0: what works and what doesn’t. Science 2019;363(6426):eaav5796.

[53] Vidale JE, Agnew DC, Johnston MJS, Oppenheimer DH. Absence of earthquake correlation with Earth tides: an indication of high preseismic fault stress rate. J Geophys Res Solid Earth 1998;103(B10):24567–24572.

[54] Heuret A, Lallemand S. Plate tectonics of the Pacific–Philippine region: subduction at the Philippine Trench and the Caroline Ridge. Tectonophysics 2005;405(1–4):1–20.

[55] Rouet-Leduc B, Hulbert C, Lubbers N, et al. Machine learning predicts laboratory earthquakes. Geophys Res Lett 2017;44(18):9276–9282.

[56] DeVries PML, Viégas F, Wattenberg M, Meade BJ. Deep learning of aftershock patterns following large earthquakes. Nature 2018;560:632–634.

[57] Mousavi SM, Beroza GC. Machine learning in seismology: turning data into insights. Seismol Res Lett 2020;91:2445–2459.