Benchmarking Machine-Learning Forecasts of Low-Magnitude Seismicity in the Italy–Slovenia Border Region

Oleksandra Topal
oleksandra.topal@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Erik Calcina
erik.calcina@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Luciano Telesca
luciano.telesca@cnr.it
Institute of Integrated Methodologies for Earth Observation, CNR
Tito Scalo, Italy

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Reliable earthquake forecasting is an important scientific challenge with direct relevance to seismic monitoring and preparedness. Neural and tree-based models were benchmarked on 5784 earthquakes from the Italy–Slovenia cross-border seismic area (magnitudes 1.15–4.04, 2016–2023), using catalogue variables, satellite data and weather reanalysis. Three experiments examined next-earthquake regression for magnitude, epicentre location and inter-event time; earthquake-occurrence classification with an evaluation audit; and the same classification task in an expanded experiment with statistical assessment. Regression remained near simple-baseline performance: epicentre prediction improved by only 1.6% beyond recent-location averaging, with no further gain from the additional inputs. The evaluation audit showed that the strongest classification score was supported by repeated forecasts associated with one earthquake. The expanded experiment identified a possible five-day signal at magnitude 2.5 and above with support from 22 distinct earthquakes, but the result remained uncertain after all model configurations were considered, and the predicted probabilities did not match observed frequencies. Credible forecasting claims therefore require appropriate baselines, explicit counts of distinct earthquakes, dependence-aware statistical testing and calibration.

Keywords

earthquake forecasting, machine learning, low-magnitude seismicity, regression, classification, forecast evaluation

1 Introduction

The continuous movement of tectonic plates deforms the Earth's crust and accumulates stress that is intermittently released by earthquakes, most of which are low in magnitude. This study investigates whether low-magnitude seismicity can support reliable machine-learning forecasts in the Italy–Slovenia cross-border seismic area and its immediate surroundings (Figure 1). Earthquakes were selected from the Tarchini et al. catalogue [9] within a polygon covering the principal concentration of seismicity across the Italy–Slovenia border and at moment magnitude $M_w \geq 1.15$, the more conservative of two completeness estimates from a Gutenberg–Richter analysis [8]. The resulting study catalogue contains 5784 earthquakes, only 42 of which have $M_w \geq 3$, limiting the support available for evaluating higher-magnitude forecasts.

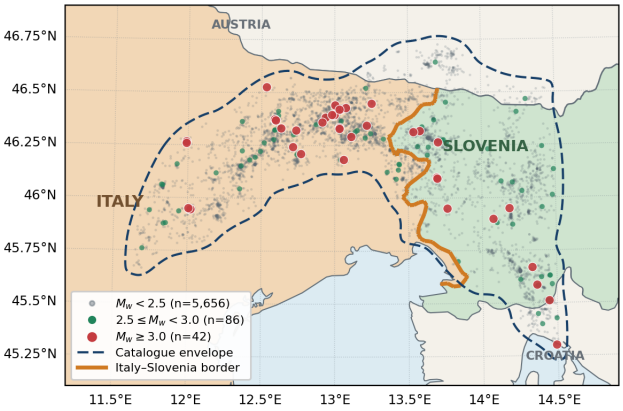


Figure 1: Spatial distribution of the 5784 earthquakes in the Italy–Slovenia cross-border seismic area and its immediate surroundings. The dashed outline marks the extent of the selected catalogue; national boundaries are shown for geographic orientation (Natural Earth 1:10m).

Two tasks were examined: next-earthquake regression for magnitude, epicentre location and inter-event time, and classification of whether an earthquake at or above a specified magnitude would occur within a fixed forecast horizon. Classification was evaluated through an audit of distinct supporting earthquakes and an expanded statistical assessment. Benchmarking against simple catalogue-based predictors extends event-level work on Italian seismicity [5] and responds to evidence that model complexity does not necessarily provide greater predictive skill [7].

2 Methodology

2.1 Regression of Next-Earthquake Properties

An earthquake event denotes one catalogue record with an origin time, epicentre location, depth and magnitude; *event* is used below as its shorter form. The regression experiments followed a staged design in which progressively richer representations were evaluated against simple catalogue baselines (Figure 2). In the initial stage, the C7 representation (Table 1) combined magnitude, latitude, longitude, depth and inter-event time with two cyclical variables encoding the earthquake origin time on a 24-hour cycle, and was evaluated using convolutional neural network–long short-term memory (CNN–LSTM) and random-forest models.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.98

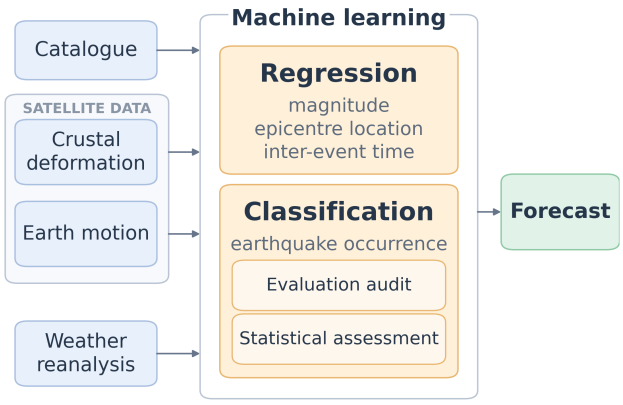


Figure 2: Benchmarking pipeline for next-earthquake regression and earthquake-occurrence classification using catalogue, satellite and weather-reanalysis data.

Because the catalogue-only models provided no improvement beyond the baselines, C7 was extended to M50 and evaluated with the same learners. M50 augmented C7 with a rolling Gutenberg–Richter b-value, which describes the relative frequency of small and larger earthquakes, its short-to-long-term change, and simplified estimates of tidal phase, amplitude and rate of change, together with 38 variables from additional data: 20 daily Global Navigation Satellite System (GNSS) displacement variables from ten stations [3], 11 European Ground Motion Service (EGMS) Interferometric Synthetic Aperture Radar (InSAR) motion variables for 2018–2022 [4], and 7 daily variables from the fifth-generation European Centre for Medium-Range Weather Forecasts reanalysis (ERA5) [6]. Same-day ERA5 matching, full-period EGMS summaries and full-catalogue median imputation made this a retrospective experiment.

Whereas C7 and M50 preserved each sliding event window as a sequence, the final stage transformed every M50 window into the tabular S300 representation by computing six summary statistics for each variable, thereby enabling evaluation with the gradient-boosted tree models CatBoost, LightGBM and XGBoost. Because the external sources were introduced together and S300 was evaluated with different learners, the experiment compared complete configurations rather than isolating the contribution of GNSS, EGMS or ERA5.

Table 1: Feature representations used for next-earthquake regression.

Set	Size	Representation
C7	7	Magnitude, latitude, longitude, depth, inter-event time and cyclical origin hour (2 variables)
M50	50	C7; b-value/change; tidal phase/amplitude/rate; 20 GNSS; 7 ERA5; 11 EGMS InSAR variables
S300	300	Six history summaries for each M50 variable

Regression samples used a sliding event window of the N preceding earthquakes to predict the next, advancing one earthquake at a time; models were trained through 2021 and evaluated in 2022–2023, which also guided model and input-history selection and S300 early stopping rather than serving as an untouched holdout. C7 and M50 each tested 11 values of N spanning 50–1500 and 50–2500 events, respectively, whereas S300 tested only $\{500, 1250, 2500\}$, leaving 500 and 1250 as the shared settings. Performance was measured by root-mean-square error (RMSE) against two baselines: the training mean, which predicts the mean of the target over the training period, and a moving average over the preceding N target values, representing recent-history persistence. The moving average was computed for C7 and M50, and its best RMSE is reported in Table 3 as a common reference. Epicentre RMSE used a Euclidean latitude–longitude norm in degrees; C7 and M50 values were multiplied by $\sqrt{2}$ to match the S300 convention.

2.2 Earthquake-Occurrence Classification: Evaluation Audit

The limited improvement of next-earthquake regression over simple baselines motivated a coarser forecasting question: whether at least one earthquake above a selected magnitude would occur within a fixed forecast horizon. The first classification experiment also used sliding event windows, issuing one prediction after each catalogue event. Each window used the preceding N events as its input history and the following D days as its *forecast horizon*. For an event at time t_i, the binary label is

$$y_i(M, D) = \mathbf{1}\{\exists j > i : t_j \le t_i + D,\ M_{w,j} \ge M\}, \qquad (1)$$

where M is the magnitude threshold. A window is positive when at least one earthquake meeting that threshold occurs within its forecast horizon; such a future earthquake is called a *target earthquake*, and the window is negative when no such earthquake occurs. Performance was summarised by the area under the receiver-operating-characteristic curve (AUC). Because consecutive windows share $N - 1$ input events and their forecast horizons overlap, one target earthquake may produce several consecutive positive windows; the audit therefore linked each positive window to its target and counted the distinct target earthquakes.

The audit evaluated an artificial neural network (ANN), support vector machine (SVM) and random forest using input histories of $N = 100$, 200 or 300 events and forecast horizons of $D = 10$, 30, 60, 90 or 120 days. The whole-catalogue experiment used the full study catalogue with 99 variables summarising past seismicity: b-value and inter-event-time statistics, a magnitude histogram, and Allan-factor and detrended-fluctuation measures. It tested $M \in \{3.0, 3.5\}$, giving 90 model combinations.

To reduce the contribution of clustered seismic sequences, the audit was repeated with a declustered catalogue containing 1717 earthquakes at $M_w \ge 1.5$, obtained by nearest-neighbour declustering [10]. The declustering was not re-run here, so the catalogue was used as a fixed data product. Because only 11 of these earthquakes reach $M_w \ge 3.5$, the threshold $M = 2.5$ was added for measurable target support, giving 135 combinations with a related 95-variable representation. Each combination was repeated 100 times on a fixed chronological 70/30 split, with seeds affecting only model initialisation and class balancing.

2.3 Earthquake-Occurrence Classification: Statistical Assessment

The expanded assessment addressed the same occurrence question on the declustered catalogue, broadening the model comparison and adding First-In-Cluster (FIC) evaluation. FIC retains every negative window but only the first window in each consecutive run of positive windows; here *cluster* means adjacent positive labels, not a seismic cluster. Its AUC reduces the repeated weighting of positive forecasts without establishing independence, because the retained negative windows can still overlap.

Performance was reported as all-window AUC, FIC AUC and the number of distinct target earthquakes, against a causal persistence baseline that forecasts occurrence in the next D days whenever a qualifying earthquake occurred in the previous D days. Model scores were held fixed while labels were permuted

independently, in contiguous 20-row blocks, or circularly shifted, using up to 100 000 permutations (Table 2); Benjamini–Hochberg correction [2] gave the reported q-values. Calibration in the large was assessed by comparing the sum of predicted probabilities with the observed number of positive windows.

Table 2: Design of the expanded statistical assessment.

Factor	Values or procedure
Threshold M	2.5, 3.0, 3.5
Forecast horizon D	5, 7, 10, 15, 50 days
Input history N	50, 100, 200, 300, 500 events
Features	Seis26; Asim60 [1]; Full86
Learners	Nine linear, kernel, tree-based and neural models
Development	Three walk-forward folds, 2016–2021; decision-threshold choice and prioritisation
Evaluation	Full 2025-configuration grid scored on 2022–2023
Inference	Causal baseline; independent, 20-row block and circular-shift label permutations; bootstrap FIC AUC interval; within-family and global Benjamini–Hochberg correction

Seis26 denotes 26 timing, distance, Gutenberg–Richter, magnitude and moment features; Asim60 denotes 60 energy, timing, seismicity, magnitude-deficit, residual and recurrence features [1]; and Full86 concatenates both sets. The nine learners included logistic regression, elastic net, SVM, k-nearest neighbours (KNN), random forest, Extra Trees, XGBoost, histogram-based gradient boosting (HistGBM) and a multilayer perceptron. The $5 \times 3 \times 5 \times 3 \times 9 = 2025$ grid gave 135 model combinations per threshold–horizon family, and the comparison is exploratory because it was scored on the period used for selection.

3 Results

3.1 Regression of Next-Earthquake Properties

Table 3 reports the forecast error for each target against the simple baselines. Neither magnitude nor inter-event time separated from them; the largest inter-event-time gain was under a minute (13.832 versus 13.844 h). Only epicentre improved, and most of that was persistence: the 4.4% gain over the training mean fell to 1.6% against the moving average, and the near-identical C7 and M50 values showed that GNSS, EGMS and ERA5 added no location information. The best epicentre error was roughly 58–84 km at this latitude.

Each column reports the lowest RMSE attained by that representation over its own sweep of N. C7 and M50 searched 11 input-history lengths, compared with three for S300; learners also differed, with C7 and M50 using CNN–LSTM apart from random-forest epicentre estimates and S300 using LightGBM or CatBoost. The training-mean baseline also shifted with N; its smallest value was used as the strictest reference, and only for epicentre did this matter (0.793°–0.817°).

No input-history length changed this: the shortest and intermediate histories performed worse than the baselines for both magnitude and inter-event time, and only the longest matched them, reproducing the training mean rather than improving on it. Similar outcomes across sequence and tabular representations, external inputs and learners therefore indicate that the weak exact-property performance was not confined to one architecture, motivating the coarser occurrence analysis.

Table 3: Next-earthquake RMSE in the evaluation period; lower is better.

Target	Mean	Moving avg.	C7	M50	S300
Magnitude (M_w)	0.349	0.349	0.349	0.349	0.348
Inter-event time (h)	14.003	13.844	14.005	14.004	13.832
Epicentre (°)	0.793	0.770	0.770	0.769	0.758

3.2 Earthquake-Occurrence Classification: Evaluation Audit

Occurrence classification discriminated little better than chance on either catalogue: mean all-window AUC was 0.513 for the whole catalogue and 0.528 for the declustered one, with 44 of 90 and 57 of 135 model combinations below 0.5.

Table 4 relates these scores to the earthquakes behind them. Each window range spans the five forecast horizons, since a longer horizon makes more windows positive while the input-history length does not affect the label. The original results recorded positive-window counts but not the earthquakes producing them, so the labels were rebuilt from the catalogue; the rebuilt window counts matched the stored ones for all 45 combinations of input-history length, threshold and forecast horizon. Mean AUC was highest at $M = 3.5$, yet the positive windows on that row came from a single target earthquake, and the whole catalogue showed the same pattern, falling from 12 target earthquakes at $M = 3.0$ to two at $M = 3.5$. Apparent discrimination therefore coincided with the smallest event support.

Table 4: Declustered classification: mean AUC is highest at the threshold with the fewest supporting earthquakes.

Threshold M	Positive windows	Distinct earthquakes	Mean AUC
2.5	149–476	29	0.449
3.0	36–333	6	0.447
3.5	9–77	**1**	**0.688**

Individual combinations vary far more than the means provided in Table 4. At $M = 3.5$ the strongest combination was SVM with $N = 100$ and $D = 30$ d, averaging AUC 0.995 over 100 runs. Its 27 positive windows were all linked to the one target earthquake at that threshold, so the score reflects repeated forecasts before a single event rather than 27 independent successes.

3.3 Earthquake-Occurrence Classification: Statistical Assessment

The expanded assessment identified one low-threshold candidate for detailed evaluation, whose significance depended on the assumed dependence structure and whose predicted probabilities did not match observed frequencies. This candidate, HistGBM with $N = 500$, $M = 2.5$, $D = 5$ d and Full86, was selected from the 2025 model combinations (Table 5), with an all-window AUC of 0.650 and a FIC AUC of 0.707 (bootstrap 95% interval 0.580–0.823) against 0.456 for the causal persistence baseline. FIC reduced 74 positive windows among 494 to 22, which correspond to 22 distinct target earthquakes, so the all-window set carried about 3.4 positive windows per target earthquake. The baseline sat below chance not through an inverted sign but because FIC keeps the positive windows furthest from a trailing event.

Independent label permutation yielded $p = 4.6 \times 10^{-4}$ and within-family $q = 0.062$, the block null yielded $q = 0.016$, and the circular-shift null yielded $q = 0.305$; all three are reported because only the block null cleared 0.05 and it was selected from among them. Under global correction no configuration was significant, the smallest q being 0.243.

The candidate was also miscalibrated in the large: summed over the 494 windows, its predicted probabilities implied 131.1 positives against 74 observed, consistent with a base-rate shift from 0.191 in training to 0.150 in the evaluation period.

Its support is likewise bounded by the search design, since q-values correct within a threshold–horizon family rather than across the complete grid and the configuration was selected on

the period it was scored on. The higher thresholds add no support of their own, resting on six and one retained positive windows with no qualifying cell in 14 of the 15 families. Nearby input histories of $N = 300–500$ produced similar FIC AUCs without reaching $q < 0.05$ under any null, and the candidate therefore requires evaluation in a later chronological period not used for model selection.

Table 5: Selected results from the statistical assessment. The first row is the candidate; the others illustrate high AUC with sparse support. n_{FIC} is the number of retained positive windows. q is within-family under the independent label permutation; the block and circular-shift nulls give 0.016 and 0.305 for the candidate.

Configuration	Model	All-window AUC	FIC AUC	q	n_{FIC}
$M = 2.5$, $D = 5$ d, $N = 500$, Full86	HistGBM	0.650	0.707	0.062	22
$M = 3.0$, $D = 15$ d, $N = 300$, Asim60	KNN	0.633	0.722	0.895	6
$M = 3.5$, $D = 5$ d, $N = 100$, Full86	Extra Trees	1.000	1.000	0.108	1

4 Discussion and Conclusion

Taken together, the experiments indicate that forecasting conclusions depended less on model complexity than on the specificity of the target and the unit of evaluation. Next-earthquake regression recovered little information beyond simple catalogue baselines, whereas the less specific occurrence task produced a possible five-day signal at magnitude 2.5 and above across 22 distinct target earthquakes; at the same time, the evaluation audit showed that the strongest high-threshold AUC was supported by 27 positive windows linked to a single earthquake. Occurrence classification could therefore appear more successful than exact-property regression, but its evidential strength depended on how many distinct earthquakes supported that discrimination.

The limited regression gains should be interpreted within the tested data integration design. Daily GNSS and ERA5 values and approximately six-day InSAR observations were aligned with event-indexed histories, which may have obscured short-lived or delayed relationships; consequently, the absence of improvement from these inputs concerns their present representation rather than their broader physical relevance. Nevertheless, recent-location averaging explained most of the apparent epicentre gain, while the inter-event-time improvement remained below one minute, indicating that the tested models learned little beyond catalogue persistence.

For occurrence classification, AUC over all windows could count the same target earthquake repeatedly because consecutive windows covered overlapping forecast horizons. FIC AUC reduced this repetition by retaining only the first positive window in each consecutive run, but it did not establish independence because negative windows could still overlap. The five-day candidate at $M = 2.5$ reached FIC AUC 0.707, yet its significance varied with the permutation scheme, no configuration survived correction across the full grid, and its predicted probabilities did not match the observed frequencies; it should therefore be regarded as exploratory rather than as evidence of operational forecast skill.

These conclusions are also bounded by evaluation-period reuse, evaluation of the complete classification grid, and independent label permutations that did not preserve temporal dependence. Moreover, the two catalogues do not begin from the same earthquakes, the declustered one starting at $M_w \geq 1.5$ and the whole one at $M_w \geq 1.15$, and the experiments on them used different thresholds and feature representations; their results are therefore reported separately rather than compared. In the study catalogue only 42 earthquakes reached $M_w \geq 3$ and none exceeded 4.04. A credible confirmation therefore requires a later chronological evaluation period not used for model selection, inputs restricted to information available when each forecast is issued, probabilities that match observed frequencies and explicit counts of distinct target earthquakes, so that predictive discrimination can be distinguished from catalogue persistence and repeated evidence around a small number of earthquakes.

Acknowledgements

This research was financially supported by the Slovenian Research and Innovation Agency (ARIS) via the Research Program Knowledge Technologies (grant P2-0103) and the Gravity project AI for Science (grant GC-0001), and by the bilateral CNR–JSI project 2026–2027 “Deep Learning-based Probabilistic Real-time Earthquake Prediction for the Italy–Slovenia cross-border seismic area”.

Data licences. EGMS and ERA5 contain modified Copernicus information: EGMS was cropped and downsampled to the study area, while ERA5 was spatially averaged and aggregated daily. Neither the European Union, the European Commission nor the European Centre for Medium-Range Weather Forecasts endorses these modifications or is responsible for their use.

The authors acknowledge the use of large language models for language optimisation. While these tools contributed to refining the presentation, all conceptual frameworks, analyses and interpretations remain the sole responsibility of the authors.

References

[1] Khawaja M. Asim, Sayed S. R. Moustafa, Iftikhar Azim Niaz, Eslam A. Elawadi, Talat Iqbal, and Francisco Martínez-Álvarez. 2020. Seismicity analysis and machine learning models for short-term low magnitude seismic activity predictions in Cyprus. *Soil Dynamics and Earthquake Engineering*, 130, 105932. doi:10.1016/j.soildyn.2019.105932.

[2] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 1, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

[3] Geoffrey Blewitt, William C. Hammond, and Corné Kreemer. 2018. Harnessing the GPS data explosion for interdisciplinary science. *Eos*, 99. doi:10.1029/2018EO104623.

[4] European Environment Agency. 2023. European ground motion service (EGMS), level-3 ortho. Copernicus Land Monitoring Service. (2023). https://egms.land.copernicus.eu/.

[5] Jesús Gonzalez, Wen Yu, and Luciano Telesca. 2022. Gated recurrent units based recurrent neural network for forecasting the characteristics of the next earthquake. *Cybernetics and Systems*, 53, 2, 209–222. doi:10.1080/01969722.2021.1981637.

[6] H. Hersbach et al. 2023. ERA5 hourly data on single levels from 1940 to present. (2023). doi:10.24381/cds.adbb2d47.

[7] Arnaud Mignan and Marco Broccardo. 2020. Neural network applications in earthquake prediction (1994–2019): meta-analytic and statistical insights on their limitations. *Seismological Research Letters*, 91, 4, 2330–2342. doi:10.1785/0220200021.

[8] Arnaud Mignan and Jochen Woessner. 2012. Estimating the magnitude of completeness in earthquake catalogs. *Community Online Resource for Statistical Seismicity Analysis*. doi:10.5078/corssa-00180805.

[9] Gabriele Tarchini, Luca Moratto, and Angela Saraò. 2025. A comprehensive moment magnitude catalog for the northeastern Italy region. *Seismological Research Letters*, 96, 4, 2714–2723. doi:10.1785/0220240303.

[10] Ilya Zaliapin, Andrei Gabrielov, Vladimir Keilis-Borok, and Henry Wong. 2008. Clustering analysis of seismicity and aftershock identification. *Physical Review Letters*, 101, 1, 018501. doi:10.1103/PhysRevLett.101.018501.