

WHEN THE GROUND STARTS TO MOVE: RAPID SEISMIC MAGNITUDE ESTIMATION IN CHILE USING MULTI-STATION PEAK DISPLACEMENTS AND ENSEMBLE MODELS

CUANDO LA TIERRA EMPIEZA A MOVERSE: ESTIMACIÓN RÁPIDA DE MAGNITUD SÍSMICA EN CHILE A PARTIR DE PICOS DE DESPLAZAMIENTO MULTIESTACIÓN Y MODELOS DE ENSAMBLE

Eliecer Peña-Ancavil^{1*}, Carla Angulo Larenas¹, Marta Isabel Gonzalez Baez¹, Felipe Ignacio Rodriguez Arriagada¹.

RESUMEN

La estimación rápida de magnitud es clave para la respuesta temprana ante terremotos y para módulos automáticos que requieren una cuantificación consistente del tamaño del evento poco después de su detección. Trabajos recientes muestran que amplitudes tempranas del movimiento del suelo, combinadas con la geometría fuente–estación, permiten aproximar la magnitud, y que los modelos de ensamble capturan relaciones no lineales con buena robustez. Sin embargo, suele faltar una descripción reproducible y auditable de la cadena completa que transforma miniSEED y StationXML en variables físicas comparables, con control explícito de calidad. Se presenta un flujo reproducible para Chile basado en estándares FDSN. La hipótesis es que un flujo físicamente consistente, con calibración por sensibilidad escalar, filtrado, integración numérica y corrección de deriva, produce características tempranas de desplazamiento con generalización temporal. El flujo automatiza selección de estaciones dentro de 500 km, construcción de tríos N/E/Z consistentes, verificación de integridad de descargas y metadatos, y extracción de amplitudes pico por componente, métricas horizontal y 3D, y distancias fuente–estación. El dataset final contiene 6641 pares evento–estación de 316 eventos (2020–2025) con Mww como objetivo. Se comparan ensambles de árboles frente a líneas base, con validación cruzada agrupada por evento y prueba temporal 2025. A nivel evento, agregando por mediana, se obtiene en 2025 desempeño típico MAE 0.15, RMSE 0.23 y R2 0.7. El aporte principal es un bloque trazable de estimación rápida condicionado a detección/localización, no una validación de extremo a extremo de latencias de alerta temprana.

Palabras clave: estimación rápida de magnitud, desplazamiento pico, pipeline reproducible, miniSEED, ensambles de árboles.

ABSTRACT

Rapid magnitude estimation is essential for early earthquake response and automated modules that require a consistent event size shortly after detection. Recent work shows that early ground-motion amplitudes combined with source–station geometry can approximate magnitude, and that ensemble models capture nonlinear relations with strong robustness. Yet many studies provide limited reproducible detail on the full chain that converts miniSEED and StationXML into comparable physical variables under explicit quality control. We present a reproducible FDSN-based workflow for Chile. Our hypothesis is that a physically consistent pipeline using scalar sensitivity calibration, filtering, numerical integration (single or double depending on units) and drift correction can produce early displacement features that generalize over time. The workflow automates station selection within 500 km, construction of consistent N/E/Z trios, integrity checks for downloads and metadata, and extraction of per-component peak amplitudes, horizontal/3D metrics, and source–station distances. The final dataset includes 6641 event–station pairs from 316 events (2020–2025) with Mww as target. We compare tree-based ensembles against baselines using event-grouped cross-validation and a 2025 temporal hold-out. At event level, median aggregation yields typical 2025 performance MAE 0.15, RMSE 0.23 and R2 0.7. The main contribution is a traceable module for rapid/operational magnitude estimation conditioned on external detection/location, not an end-to-end seconds-latency EEW validation.

Keywords: rapid magnitude estimation, peak displacement, reproducible pipeline, miniSEED, tree-based ensembles.

¹ Universidad Bernardo O'Higgins

* Corresponding Author: eliecer.r.pena.ancavil@gmail.com

1. INTRODUCTION

Chile lies on one of the most active convergent margins on the planet, where the Nazca plate subducts beneath the South American plate, generating recurrent moderate-to-large earthquakes (Aedo et al., 2024) (Geiß et al., 2023). This activity directly affects urban centers, critical infrastructure, and industrial systems; therefore, having rapid and reliable estimates of source parameters is essential for risk management and real-time decision-making (Pasten et al., 2022) (Devi D et al., 2024). In particular, an operational magnitude estimate shortly after the initial detection and location of an event conditions critical decisions (e.g., action thresholds, resource prioritization, and impact-scenario updates). In this context, systematic biases have clear operational consequences: underestimating magnitudes can lead to insufficient response, while persistent overestimation erodes trust and can trigger unnecessary actions (Cremen et al., 2022).

It is important to delimit the scope: this work does not present an earthquake early warning (EEW) system in the strict sense, nor does it evaluate second-level latencies (Wang et al., 2023) (Kolivand, 2024). The objective is to build and audit a reproducible workflow that, using real instrumental records and FDSN metadata, produces physically interpretable features and enables training rapid magnitude estimation models under an idealized temporal reference (catalog origin time). In an operational setting, this block would be integrated downstream of an external detection/location module; here, the focus is on predictive capability and temporal generalization of the approach, not on end-to-end online performance.

Early magnitude estimation often relies on ground-motion measures computed over a fixed analysis window after the event onset, such as peak displacement, velocity, or acceleration amplitudes, combined with some measure of source–station distance (Temneanu et al., 2025) (Noda et al., 2024) (Chandrakumar et al., 2022). However, these approaches face known limitations: saturation or reduced sensitivity at higher magnitudes, sensitivity to noise and instrumental variability, and difficulties generalizing when the network is heterogeneous in sensors, sampling rates, and site conditions (Wu et al., 2023) (Chandrakumar et al., 2022).

In real scenarios, these limitations depend not only on the chosen functional form, but also on the complete chain that transforms raw records into physically comparable variables across stations and events. Over the last decade, increased density of digital networks and the standardization of formats and services (e.g., miniSEED and StationXML within the FDSN ecosystem) have enabled automated workflows in which machine learning (ML) learns nonlinear relationships between early amplitudes, source–receiver geometry, and magnitude (Navarro et al., 2025) (Shen et al., 2025) (Wibisono & Naufal Rahmadika, 2025).

Within this framework, tree-based ensemble methods (Random Forest, Extra Trees, and gradient boosting variants such as LightGBM or CatBoost) are especially attractive because they capture nonlinear interactions, tolerate heterogeneous predictors, and are often robust to noise, while maintaining practical interpretability through feature importances and sensitivity analyses (Turino et al., 2025) (Joshi et al., 2022).

Even so, a relevant gap in the literature is that many works emphasize model architecture, but provide less detail on the end-to-end engineering required to operate with real data: automatic station selection, consistent N/E/Z triad construction, metadata verification, instrument-response calibration, stable displacement reconstruction, and quality-control criteria to prevent sensitivity or integration failures from dominating the training set (Navarro et al., 2025) (Arrais et al., 2022). This gap is particularly critical in contexts such as Chile, where diverse networks and instrumental configurations coexist, and where a pipeline must be reproducible and auditable if it is to be moved toward operational conditions.

This work presents and evaluates a reproducible pipeline, oriented toward operationally plausible conditions, for rapid/operational magnitude estimation in Chile from triaxial records and source–station geometry. The workflow starts from an event catalog and station inventories compatible with FDSN standards, automatically selects stations within a maximum radius, and downloads triaxial miniSEED records. Waveforms are then calibrated using instrumental sensitivity derived from StationXML, filtered, and integrated (single or double depending on the channel physical unit) to reconstruct displacement, applying drift corrections and physical plausibility criteria. Features are extracted over a fixed window relative to the catalog origin time, adopted as a

design decision to stabilize the maximum-displacement measurement and to make processing more robust to noise and edge effects; the aim is not to minimize latency, but to quantify predictive capacity under a reproducible processing definition. The result is a tabular dataset where each observation corresponds to a valid event–station pair, and the operational output is reported at the event level through robust aggregation (one magnitude per earthquake).

On this dataset, supervised regression experiments compare ensemble models (Random Forest, Extra Trees, LightGBM, CatBoost, and a classical boosting baseline) against explicit baselines (Dummy and Ridge). The evaluation is designed to be conservative and aligned with the operational goal of “one magnitude per earthquake.” First, a chronological split reserves 2025 as an independent test set to measure temporal generalization. Second, within 2020–2024, out-of-fold predictions are obtained using event-grouped cross-validation, preventing intra-event information leakage. Third, beyond record-level diagnostics, primary metrics are also reported at the event level by aggregating station predictions via the median; this prevents events with larger instrumental coverage from dominating metrics and approximates the required operational output: one magnitude per event.

It is important to emphasize the study scope. This work does not implement or validate an end-to-end system with real-time latency; in particular, it does not include station-level phase picking nor does it quantify triggering or communication times. The analysis assumes that basic source parameters (origin time and hypocentral location) come from an external detection/location module and, for experimental purposes, uses the catalog origin time as the reference. Therefore, the evaluation quantifies predictive capacity and temporal generalization (training on past events and testing on future events) but does not characterize EEW performance on second-level horizons. Explicit incorporation of detection/location uncertainty and the study of shorter, latency-oriented windows are identified as natural extensions. The specific objectives of this work are:

- Develop a reproducible pipeline that converts FDSN-compatible data (miniSEED and StationXML) into physically based features derived from reconstructed displacement and source–station geometry, including quality-control and physical-plausibility checks.
- Evaluate, under a temporal generalization protocol and intra-event dependence control, the performance of ensemble regression models against explicit baselines.
- Analyze error structure with respect to physically relevant variables (e.g., source–station distance and magnitude), using record-level diagnostics and event-level metrics via robust aggregation.

Overall, this work aims to provide a feasibility demonstration centered on the complete procedure (from standardized raw data to honestly evaluated predictions) showing that a physically consistent, auditable pipeline based on open standards can produce early magnitude estimates with realistic temporal generalization, constituting an integrable block within post-detection workflows for rapid response and operational magnitude estimation in the Chilean subduction zone.

2. THEORETICAL FRAMEWORK

Rapid magnitude estimation is commonly grounded in the idea that earthquake size is reflected in early ground-motion amplitudes recorded at stations, once geometric spreading and attenuation effects related to source–receiver distance are accounted for. Operational formulations typically compute peak measures (displacement, velocity, or acceleration) within a fixed window after the onset and combine them with a distance proxy to infer magnitude (Temneanu et al., 2025) (Noda et al., 2024) (Chandrakumar et al., 2022). These peak-based variables are attractive because they are physically interpretable and can be computed with relatively low algorithmic complexity once a reference time is available.

In practice, several limitations make the full signal-processing chain inseparable from the estimator itself. Peak measures can lose sensitivity for larger magnitudes (e.g., saturation, waveform complexity, finite-window effects), and they are sensitive to noise levels and instrument/site variability, which complicates generalization across heterogeneous networks (Navarro et al., 2025) (Wu et al., 2023) (Chandrakumar et al., 2022). When

displacement is reconstructed by integrating velocity or acceleration, low-frequency drift and baseline offsets become especially relevant; without consistent detrending and filtering, integrated signals can exhibit non-physical trends that inflate peak displacement and contaminate learning signals.

Standardized data ecosystems (miniSEED for waveforms and StationXML for response metadata in the FDSN framework) enable end-to-end pipelines where instrument calibration and metadata validation are applied systematically, improving comparability across stations and events (Devi D et al., 2024) (Navarro et al., 2025) (Shen et al., 2025). This is critical in settings with mixed sensor types and sampling rates, because a model trained on uncalibrated or inconsistently processed records may inadvertently learn instrument artifacts rather than source physics (Navarro et al., 2025) (Arrais et al., 2022).

From a modeling perspective, modern ML approaches aim to capture nonlinear relationships among early amplitudes, distance/geometry, and magnitude (Devi D et al., 2024) (Navarro et al., 2025) (Shen et al., 2025). Tree-based ensembles are frequently used because they handle heterogeneous predictors well and can represent interactions without imposing a rigid parametric form; they are often robust under noisy predictors and still allow practical interpretability via feature importance and sensitivity analyses (Turino et al., 2025) (Joshi et al., 2022).

For operational relevance, evaluation protocols must respect the dependence structure induced by having multiple station records per event; otherwise, leakage can inflate performance. Grouped validation and event-level aggregation provide a principled way to align evaluation with the target output: one magnitude per earthquake, rather than many correlated station-level estimates.

3. MATERIALS AND METHODS

3.1 Study area and event catalog

This study focuses on earthquakes associated with the subduction of the Nazca plate beneath the South American plate along the Chilean margin and adjacent regions. An initial catalog was built from the USGS Comprehensive Catalog (USGS Earthquake Catalog, 2025), extracting for each event the origin time, hypocentral location (latitude, longitude, depth), and the reported magnitude along with its type (magType). The query and subsequent local filtering were restricted to the period 08-Jan-2020 to 16-Dec-2025 (UTC) and to a spatial window covering the Chilean margin and surroundings (approximately latitudes [-59.7, -11.9] and longitudes [-106.0, -60.0]). Under these criteria, the initial catalog considered by the pipeline contains 468 events.

This work uses exclusively the moment magnitude reported as M_{ww} , corresponding to moment tensor solutions based on W-phase and generally more homogeneous for moderate-to-large events (Han et al., 2025) (da Silva et al., 2022) (Dybing et al., 2024). Consequently, M_{ww} is adopted as the target variable (ground truth) in the regression experiments, avoiding mixing scales with different saturation behavior and different frequency bands.

3.2 Record availability and catalog curation

The USGS catalog filtered by $M_{ww} \geq 5.0$ defines the potential set of events, but the final usable set for regression depends on real instrumental availability and on whether triaxial records can be calibrated and reconstructed consistently. At this stage, an event is considered usable if at least one valid triaxial record (N/E/Z triad) can be reconstructed at a station within a maximum radius of 500 km, satisfying component consistency (Z, N, E) with the same channel prefix and the same sampling rate, as well as download integrity and metadata filters. In the final set, after applying the quality controls described below, event coverage ranged from 3 to 28 valid triaxial records (observed minimum and maximum; not imposed as inclusion criteria), with an average of 21 records per event.

After download, each trace is subjected to operational checks: existence of the miniSEED file and a minimum

size threshold to discard empty responses, completeness of StationXML metadata, valid instrumental sensitivity (e.g., discarding clearly invalid cases such as sensitivity equal to one), and waveform preprocessing/reconstruction without numerical failures. In addition, physical-plausibility thresholds are applied to reconstructed displacement (discarding clearly anomalous cases), and full triad consistency is required for an observation to be considered valid.

As a result of this curation, the final dataset comprises 316 events for which at least one valid record could be reconstructed, producing a total of 6641 event–station observations. The statistical unit used for ML is the event–station pair; therefore, a given event appears in multiple rows associated with different stations. This design is consistent with the main goal of evaluating model behavior at station level, while leaving open multi-station aggregation at event level as a natural extension.

3.3 Seismic networks and stations

Waveforms were obtained from broadband and short-period stations belonging to Chilean seismic networks, including the FDSN networks C (Universidad de Chile, Dept. de Geofísica (DGF), 1991), C1 (Universidad de Chile, 2012), and CX (GEOFON Program (GFZ German Research Centre for Geosciences), 2006), whose data are archived in FDSN data centers such as IRIS and GEOFON. Waveforms are downloaded in miniSEED format, which stores digital counts per channel. Physical interpretation and conversion to physical units are performed using StationXML metadata, which provide the instrumental response (sensitivity) and the physical units associated with ground motion (e.g., velocity or acceleration).

From StationXML files, a channel inventory table was built, including for each network–station–channel combination: sampling rate f_s , response physical units (as reported by response metadata), and the effective instrumental sensitivity (gain) used to calibrate counts into physical units. This information is consolidated into a table used during preprocessing to discard channels with incomplete metadata and to apply instrumental calibration reproducibly.

Operational definitions used in this work are as follows. A channel corresponds to the three-letter SEED code (e.g., BHN, BHE, BHZ). The channel family or prefix is defined as the first two letters of the code (e.g., BH or HN), grouping channels that share the same nominal band/instrument configuration. A triad corresponds to simultaneous availability of the three orthogonal components Z, N, and E for the same network and station, with the same prefix (e.g., BH) and the same sampling rate f_s ; when multiple variants exist, the triad is constrained to a consistent station configuration (e.g., same location code when available).

The inventory reveals that many stations provide multiple sensors and channels for a given site (e.g., short-period and broadband families, or redundant configurations) with different sampling rates. To guarantee adequate temporal resolution for early analysis, only channels with sampling rate strictly greater than 19 Hz (equivalently, $\Delta t < 0.05$ s) are considered. For each station, all triads satisfying the definition are generated; when multiple options exist, the triad with the highest sampling rate is selected as an operational prioritization rule, and the rest are discarded to keep a unique definition per station and reduce ambiguity.

Sensitivity and unit fields extracted from StationXML are also used as formal quality filters: channels with missing or inconsistent response, or clearly invalid sensitivities (e.g., sensitivity equal to one, non-finite, or non-positive) are excluded, since reliable calibration from counts to physical units is not possible in those cases. The resulting candidate station set consists of well-characterized N–E–Z triads, on which the source–station distance criteria are applied next.

3.4 Station selection and distance computation

For each catalog event, epicentral and hypocentral distances are computed between the source and all stations in the inventory. Horizontal (epicentral) distance is computed via the Haversine formula using event coordinates (φ_e, λ_e) and station coordinates (φ_s, λ_s) in radians. With $\Delta\varphi = \varphi_s - \varphi_e$ and $\Delta\lambda = \lambda_s - \lambda_e$, the great-circle distance dH is computed as:

$$a = \sin^2(\Delta\phi/2) + \cos(\phi_e) \cos(\phi_s) \sin^2(\Delta\lambda/2)$$

$$c = 2 \operatorname{atan2}(\sqrt{a}, \sqrt{1-a})$$

And $dH = R c$, with $R = 6371 \text{ km}$. To define vertical separation, hypocentral depth z_e (km, positive downward) from the catalog and station elevation h_s (m above sea level, positive upward) from the inventory are used. Converting h_s to km ($h_s/1000$) and adopting a consistent sign convention, the vertical distance is:

$$dV = z_e + h_s/1000$$

The 3D hypocentral distance is then:

$$d_{3D} = \sqrt{dH^2 + dV^2}$$

In addition to distances, geometric orientation variables are computed for each event–station pair. The azimuth from the station to the epicenter (bearing, clockwise from geographic north) is computed on the sphere using coordinates in radians. With $\Delta\lambda = \lambda_e - \lambda_s$:

$$y = \sin(\Delta\lambda) \cos(\phi_e)$$

$$x = \cos(\phi_s) \sin(\phi_e) - \sin(\phi_s) \cos(\phi_e) \cos(\Delta\lambda)$$

$$\text{Azimuth} = \operatorname{mod}(\deg(\operatorname{atan2}(y, x)), 360).$$

Optionally, the back-azimuth (event $\rightarrow$ station) is defined as $\text{BackAzimuth} = \operatorname{mod}(\text{Azimuth} + 180, 360)$. The geometric elevation angle is defined as:

$$\text{ElevAngle} = \deg(\operatorname{atan2}(dV, dH)),$$

so that 0° corresponds to purely horizontal separation and 90° to purely vertical separation. To reduce the dynamic range of distances, a logarithmic transform is used:

$$\text{LogDistHypo} = \log_{10}(d_{3D}[\text{km}] + 1),$$

where +1 avoids indeterminacy at $d_{3D} = 0$. Stations are filtered for each event to include only those within a maximum radius of 500 km. This threshold is adopted as a design decision oriented to early estimation scenarios: beyond this range, attenuation and lower signal-to-noise ratio reduce the usefulness of early amplitudes, and the operational value for rapid information diminishes. Within the selected subset, stations are ordered by increasing d_{3D} , and for each station the pipeline requires at least one complete N–E–Z triad with $f_s > 19 \text{ Hz}$ and valid metadata. In the final dataset, dH , dV , and d_{3D} are explicitly stored per event–station pair. In regression experiments, d_{3D} is used as the primary source–station distance measure, while dH and dV remain available to capture potential geometric effects.

3.5 miniSEED acquisition and organization

Waveforms were obtained via FDSN dataselect services in miniSEED format (Esposito et al., 2024) (Megna et al., 2022). Specifically, IRIS dataselect was used for networks C and C1, while GEOFON dataselect was used for network CX. For each selected station, the three orthogonal components (N, E, Z) associated with a candidate triad (same station, same channel prefix, same sampling rate) were requested according to the StationXML-derived inventory. To ensure the analysis window is fully contained within the downloaded record and to reduce failures due to windows starting too close to the origin, downloads were performed over a fixed 60-minute interval. Operationally, the download start time was defined from the event origin time t_0 as $t_{\text{start}} = t_0 - 15 \text{ min}$, then rounded down to the beginning of the immediately previous 10-minute block (UTC). The end time was computed as $t_{\text{end}} = t_{\text{start}} + 60 \text{ min}$. This interval is used only for acquisition and quality control; subsequent processing uses a much shorter window centered around the origin.

When a station offered multiple triad options (e.g., different channel families or sampling rates), triads were prioritized by decreasing sampling rate and download attempts were made in that order. A triad was considered valid only if all three components (N, E, Z) were successfully downloaded and passed a simple integrity check based on file size: each miniSEED file must exist on disk and exceed 10 kB, which helps discard empty responses or non-seismic content (e.g., error messages). If any component failed this check, the entire triad was discarded and the next triad option was tried.

miniSEED files were stored on disk organized by event under a dedicated subfolder, using a naming scheme that encodes network, station, channel, and sampling rate (e.g., NET_STA_CHA_FS.mseed). This file structure is complemented by one event file consolidating earthquake metadata, selected stations, source–station distances, and the miniSEED filenames associated with each N–E–Z triad.

3.6 Assumptions on source parameters and temporal reference

The proposed workflow focuses on feature engineering and regression for rapid/operational magnitude estimation under a defined temporal reference. In a real-time system, this stage would be coupled to an external rapid detection and location module responsible for estimating basic source parameters (origin time and hypocentral location) and, potentially, for providing phase picks per station. The design and validation of that module, as well as quantification of end-to-end latencies, are outside the scope of this work.

In the historical experiments presented here, time zero ($t = 0$) is defined using the origin time reported by the reference catalog (USGS). Therefore, reported performance quantifies regression performance under an idealized temporal reference and does not include degradation due to real uncertainties in detection, picking, or location. In operations, errors in the estimated origin time or picks would shift the analysis window and alter extracted features; explicit sensitivity analysis to such errors is left as future work.

In this version, traces are analyzed in a fixed window around the origin, from -30 s to $+240$ s. The pre-event segment ($t < 0$) is used only to estimate and remove baseline and to characterize noise; it does not imply access to future information, and in an online system it would be replaced by a buffer immediately preceding trigger time. The post-event segment stabilizes filtering and integration and reduces sensitivity to edge effects, which is particularly relevant when features are based on maximum reconstructed displacement amplitudes.

3.7 Waveform processing

Downloaded records in miniSEED correspond to digital counts. The goal of this stage is to transform each raw trace, reproducibly and traceably, into (i) a physically interpretable displacement series $D(t)$ and (ii) a compact set of features based on peak amplitudes within a fixed operational window relative to the catalog origin time. Instrument calibration (scalar sensitivity) and operational integration rule. Counts are converted to physical units using a homogeneous calibration based on the scalar sensitivity reported by StationXML. For each channel identified by (network, station, SEED code, and location when applicable), the effective instrumental sensitivity S is extracted, typically expressed as counts/(physical unit). Conversion is defined as:

$$x_{phys}(t) = \frac{x_{counts}(t)}{S}$$

As a minimal quality check, channels with non-finite, non-positive, or clearly invalid sensitivity (e.g., $S = 1$) are discarded because a traceable conversion cannot be guaranteed. The decision to apply single or double integration is defined operationally by combining (i) the channel units reported by StationXML and (ii) the channel instrument code (SEED). Units are normalized into an internal code (V, A), where V corresponds to velocity (m/s) and A to acceleration (m/s²). A channel is marked as accelerometric if (a) normalized units indicate acceleration and/or (b) the second letter of the SEED code is N (channels of the form *N*, typically accelerometers). Accelerometric channels undergo double integration (acceleration $\rightarrow$ velocity $\rightarrow$ displacement). Otherwise, the channel is treated as velocimetric and undergoes single integration (velocity $\rightarrow$

displacement).

Timestamps are converted to UTC and then to relative time with respect to event origin ($t = 0$). Processing and feature extraction are performed on a fixed window $[-30, +240]$ s. The pre-event segment ($t < 0$) is used to estimate and remove baseline/trend and to stabilize later stages (filtering and integration). If the pre-event portion lacks enough samples, a fallback segment is used (first 20% of the window).

On the calibrated signal, the pipeline applies: (i) baseline/trend removal estimated exclusively from the pre-event segment, (ii) edge tapering with a Tukey window, and (iii) zero-phase Butterworth bandpass filtering. The Tukey window uses $\alpha = 0.10$ over the full operational window. Bandpass filter rule. A 4th-order Butterworth bandpass filter is applied in zero phase (`filtfilt`). Sampling frequency is estimated as $fs = \text{round}(1/\text{median}(\Delta t))$. The high cutoff is:

$$f_{high} = \min(15, 0.45 fs),$$

leaving margin below Nyquist. The low cutoff baseline is:

$$f_{low,0} = \max(0.03, 0.7/\max(T, 1)),$$

where T is the effective duration (seconds) of the analyzed window. For accelerometric channels, a higher floor is imposed to reduce drift after double integration:

- $f_{low} = \max(f_{low,0}, 0.08)$ for acceleration
- $f_{low} = f_{low,0}$ for velocity

Displacement is reconstructed by cumulative numerical integration using the trapezoidal rule. Accelerometric channels are integrated twice; velocimetric channels are integrated once. Drift induced by offsets and very low-frequency components is controlled using a reproducible correction based on (i) detrending using only the pre-event segment and (ii) a high-pass filter after integration.

Before bandpass filtering, a first-degree polynomial trend is fitted and removed using only the pre-event segment (fallback: first 20% if needed). For acceleration, this pre-event detrending is applied again after the first integration (velocity; degree 1) and after the second integration (displacement; degree 2) to counter stronger drift tendencies associated with double integration. After integration, a 2nd-order high-pass Butterworth filter is applied in zero phase with cutoff:

$$f_{hp} = \max(0.02, 0.5 f_{low}),$$

and for acceleration an additional minimum $f_{hp} \geq 0.10$ Hz is enforced to stabilize displacement. Finally, displacement is re-centered by subtracting the mean of the pre-event segment (or fallback segment), forcing a near-zero baseline before origin time. Peak amplitude extraction and plausibility control. On reconstructed displacement series $D_N(t)$, $D_E(t)$, and $D_Z(t)$, peak amplitudes are computed within $[-30, +240]$ s:

$$D_{N,peak} = \max_t |D_N(t)|, D_{E,peak} = \max_t |D_E(t)|, D_{Z,peak} = \max_t |D_Z(t)|$$

Derived peak magnitudes are also computed:

$$D_{H,peak} = \max_t \sqrt{D_N(t)^2 + D_E(t)^2}, \quad D_{3D,peak} = \max_t \sqrt{D_N(t)^2 + D_E(t)^2 + D_Z(t)^2}$$

Each N–E–Z triad is marked valid only if (i) all three components pass the instrumental and preprocessing checks and (ii) $D_{3D,peak} < 1$ m, a threshold introduced to filter outliers associated with calibration failures, saturation, or numerical instability during integration. After feature extraction, time series are discarded and only

the peak amplitudes are retained as tabular features per event–station pair.

As shown in Figure 1, a representative N/E/Z triad illustrates the full processing effect by contrasting raw miniSEED counts (left) with reconstructed displacement after calibration, filtering, and integration (right). The vertical red line marks the catalog origin time $t = 0$, and the red markers indicate the peak displacement amplitudes extracted within the analysis window. A concise pipeline summary (conceptual algorithm). For each event, the workflow computes distances to candidate stations, filters stations within 500 km, builds consistent N/E/Z triads (same prefix and sampling rate), downloads miniSEED for N/E/Z, validates file integrity, extracts sensitivity/units from StationXML, calibrates counts to physical units, applies detrending, tapering, bandpass filtering, integrates to displacement (single/double), applies high-pass drift control and re-centering, computes peak displacement features, discards implausible cases ($D_{3D,peak} \geq 1$ m), and stores one row per valid event–station pair. The first valid triad per station (highest sampling-rate preference) is used.

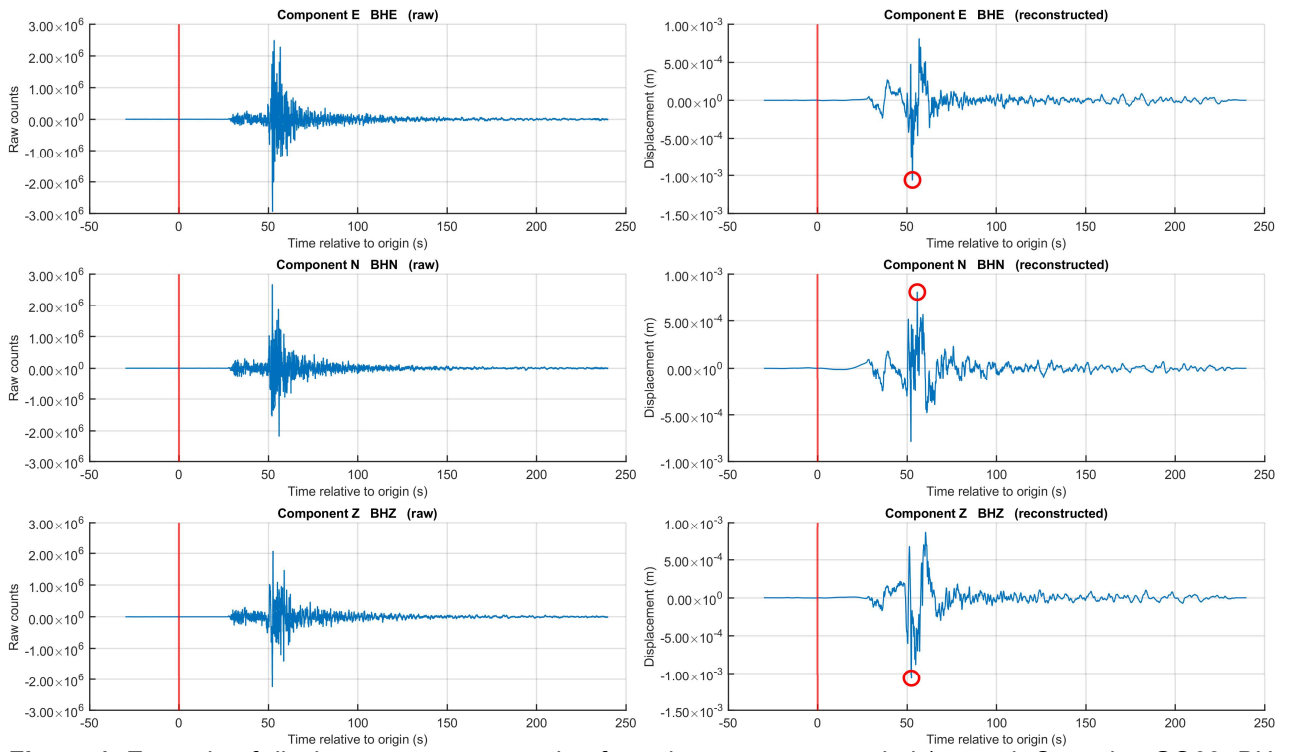


Figure 1. Example of displacement reconstruction for a three-component triad (network C, station GO03, BH channel family, 40 Hz) associated with an event of $M_{ww}=6.0$ with origin time 04-Jul-2021 01:29:15 UTC. Left: raw miniSEED digital counts. Right: reconstructed displacement (m) after scalar-sensitivity calibration, band-pass filtering, and numerical integration (single or double depending on channel type). The red vertical line marks $t=0$. Red circles indicate the maximum absolute displacement per component used as an input feature.

3.8 Dataset construction

The final dataset is organized as a table where each row corresponds to a valid event–station pair, i.e., a statistical unit defined by an earthquake and a station that recorded the three components N–E–Z correctly. Each row includes:

- Event identifiers (name or FileMat, origin time, hypocentral coordinates, and moment magnitude M_{ww}).
- Station identifiers (network, station code, and channel prefix).
- Component peak displacements: $D_{N,peak}$, $D_{E,peak}$, $D_{Z,peak}$.

- Derived peak measures: horizontal peak displacement $D_{H,peak}$ and 3D peak displacement $D_{3D,peak}$.
- Source–station distances: epicentral dH , vertical dV , and 3D hypocentral d_{3D} .

Peak amplitudes are expressed in meters and are obtained after calibration, filtering, and integration as described above. Distances are computed in kilometers. Log-transformed versions of distances $\log_{10}(d + 1)$ are also included to reduce skewness and improve regression stability when representing nonlinear relationships with magnitude. An N–E–Z triad is considered valid only if all three components pass calibration and preprocessing filters and if $D_{3D,peak} < 1$ m. This control eliminates unstable or miscalibrated cases.

After applying all filters, the consolidated dataset contains 6641 observations corresponding to 316 unique events. The structure supports two analysis levels: station level (each row as an observation) and event level (aggregating predictions or errors by median or mean across all stations for a given event). This duality allows evaluation both in terms of physical consistency per event and statistical generalization. The table is exported as a spreadsheet and constitutes the basis for the modeling experiments.

3.9 Regression models and experimental configuration

Rapid/operational magnitude estimation is formulated as a supervised regression problem. Let $\{(xi, yi)\}_{i=1}^N$ be the set of observations, where $xi \in \mathbb{R}^p$ is the feature vector for event–station pair i , and yi is the catalog moment magnitude. The goal is to approximate a function $f: \mathbb{R}^p \rightarrow \mathbb{R}$ such that $\hat{y}_i = f(xi)$ reproduces observed values. Train/test split and intra-event dependence control. Because the same earthquake generates multiple rows (one per station), evaluation is designed to prevent records from the same event appearing simultaneously in training and evaluation. A chronological split by event date is used: all event–station pairs from years 2020–2024 are assigned to training, and those from 2025 are reserved as the final test set. Within the training period, model selection and hyperparameter tuning are performed via event-grouped cross-validation using GroupKFold (5 folds) with group defined by FileMat. This keeps all rows from the same earthquake in the same fold, preventing intra-event leakage and approximating generalization to unseen events.

Models and baselines. Tree-ensemble regressors are evaluated, including Random Forest, ExtraTrees, CatBoost, and LightGBM (Kaushal et al., 2025) (Amerthen et al., 2025). Two baselines are used. The first is a trivial model (DummyRegressor) that predicts the mean magnitude observed in training. The second is a Ridge regression, serving as a robust linear baseline under collinearity and allowing assessment of whether nonlinear ensembles provide improvements beyond linear amplitude/distance–magnitude relationships.

Hyperparameter search and model selection. Initial model comparison and candidate selection are implemented in Python using PyCaret under the same event-grouped cross-validation scheme (fold_strategy="groupkfold" with fold groups defined by FileMat) (Sarangpure et al., 2023). For selected models, hyperparameters are optimized via Optuna to minimize the average RMSE under grouped CV. The search is run with a finite iteration budget and a fixed experimental seed.

Metrics and station-level vs event-level evaluation. Performance is quantified using MAE, RMSE, and R^2 . Beyond row-wise evaluation (station level), event-level evaluation is reported to reflect the final operational objective of estimating one magnitude per earthquake and to prevent events with more stations from dominating metrics. For an event e with station set Se , the aggregated prediction is defined as the median of station-level predictions:

$$\hat{y}_e = \text{median}\{\hat{y}_{e,s} : s \in Se\}.$$

MAE/RMSE/ R^2 are then computed using pairs $(y_e, \hat{y}_e)$. In training, this event-level evaluation is built from out-of-fold predictions under GroupKFold so that aggregation does not benefit from having trained on the same event. In the 2025 test set, event-level predictions are obtained by aggregating station predictions from the model trained on 2020–2024.

3.10 Reproducibility and data availability

Raw data used in this study are publicly available through FDSN services (IRIS and GEOFON) for networks C, C1, and CX. To facilitate reproducibility, the following are provided: (i) the exact list of USGS events used (IDs and metadata), (ii) a snapshot of the instrument metadata table derived from StationXML (sensitivity, units, and sampling rate), and (iii) the derived event–station dataset containing the final features (displacement peaks and geometry) and the target magnitude, together with code to replicate preprocessing, training, evaluation, and figure generation. All scripts and derived data products required to reproduce the preprocessing and modeling experiments are available in a public repository: <https://github.com/EliecerPA/PaperTdeUBO>.

4. RESULTS

Applying the pipeline described in Section materials and methods yields a tabular dataset in which each observation corresponds to a valid triaxial record (N/E/Z triad) associated with an event–station pair, while retaining instrumental metadata for traceability (e.g., channel family and sampling rate). For each row, the event moment magnitude is available (*mag* as ground truth), together with early amplitudes of reconstructed displacement by component (*PeakE*, *PeakN*, *PeakZ*) and source–receiver geometric variables (log-scaled hypocentral distance, depth, and azimuth/angle descriptors).

To evaluate generalization to future events and avoid temporal dependence, we adopt a chronological year-based split: 2020–2024 for training and 2025 as an independent test set. The full dataset contains 6,641 observations (event–station records), comprising 5,731 training records and 910 test records. In terms of unique earthquakes, this corresponds to 316 events, with 269 events in training and 47 events in testing (Table 1). This distinction matters because the dataset is expanded (a single event can contribute multiple rows); therefore, the number of distinct earthquakes conditions evaluation as much as the total number of records.

Table 1. Overall summary of the dataset used in the study and its chronological split. “Records per event” refers to the number of event–station pairs with a valid N/E/Z triad after quality-control checks (observed minimum and maximum; not imposed as an inclusion criterion).

Characteristic	Value
Study period	2020–2025
Minimum magnitude (<i>mag</i>)	$M_w \geq 5.0$
Total events	316
Total rows (valid triaxial records)	6,641
Training (2020–2024):	events, rows 269, 5,731
Test (2025):	events, rows 47, 910
Average records per event	21.0
Records per event (observed minimum)	3
Records per event (observed maximum)	28

Figure 2 compares the magnitude distributions between training and test using consistent bins. In both subsets the magnitude range is identical ($5.0 \leq M_{ww} \leq 7.4$), and descriptive statistics are similar (e.g., comparable means and dispersions), which reduces the risk that the 2025 set is strongly out-of-distribution in magnitude. Consequently, performance differences between Train and Test are interpreted primarily as an effect of temporal generalization rather than an abrupt change in the target range.

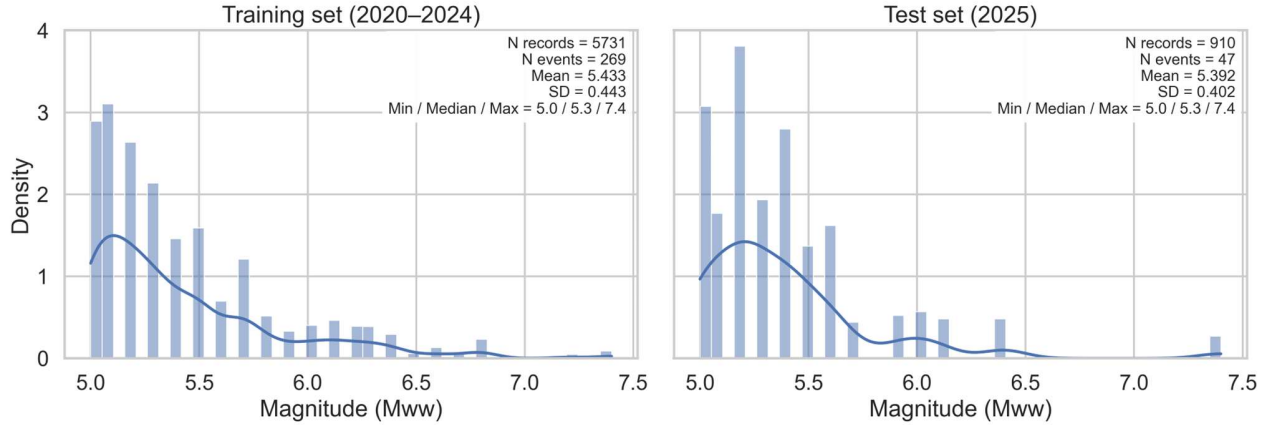


Figure 2. Magnitude distributions under the chronological split. Left: training set (2020–2024). Right: test set (2025). The same histogram bins are used in both panels to enable direct comparison.

In the main results we use a predictor variant referred to as the physical baseline, designed to maximize interpretability and reduce the risk of leakage from identifiers. Specifically, features are restricted to amplitudes (PeakE, PeakN, PeakZ) and source–receiver geometry (LogDistHypo, EQ_depth, ElevAngle_deg, Azimuth_deg), explicitly excluding identification or temporal columns such as FileMat, EQ_time, and free-text fields such as EQ_place. Instrumental and categorical metadata (e.g., network/station, band, sampling rate) are retained in the table for auditing and complementary analyses, but they are not used as predictors in the main configuration.

Figure 3 summarizes the linear dependence structure (Pearson correlation) among variables in the physical-baseline variant within the training set. As expected, the component peak amplitudes (PeakE, PeakN, PeakZ) exhibit high collinearity, consistent with being derived from the same instrumental record and capturing a common ground-motion field. These amplitudes also decrease with hypocentral distance (moderate negative correlations), consistent with attenuation and lower SNR at larger source–receiver separation. In contrast, magnitude shows relatively low marginal correlations with each individual predictor (on the order of 0.25 with amplitudes and near zero with several geometric variables), suggesting that the problem is not well described by a single isolated linear relationship, but rather by nonlinear combinations of observed amplitude and event–station geometry. This motivates the use of nonlinear ensemble models to capture predictor interactions in the following sections.

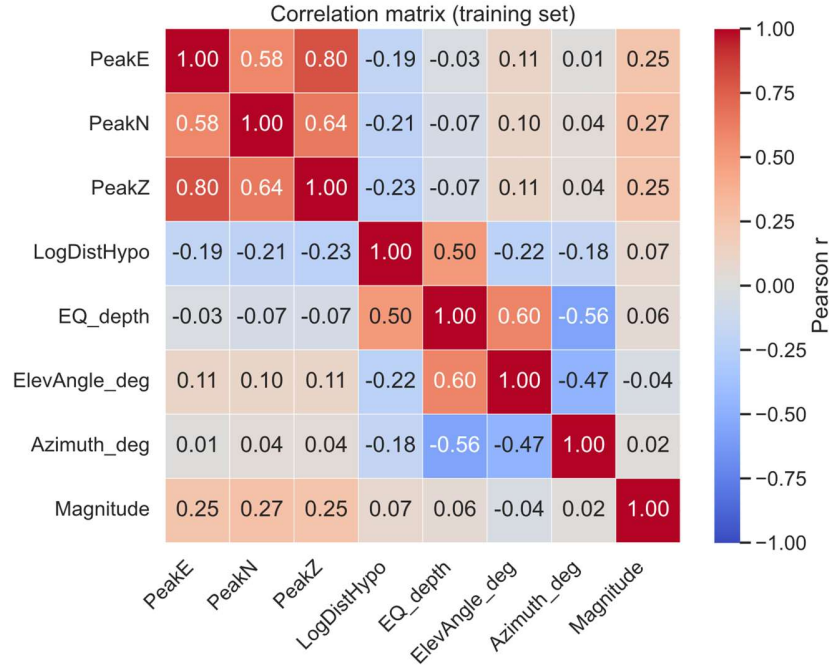


Figure 3. Pearson correlation matrix computed on the training set (2020–2024) for the physical-base feature variant (per-component peak amplitudes and geometric variables) together with the target magnitude. Component peak amplitudes show high collinearity ($r \approx 0.58$ – 0.80) and a moderate marginal relationship with magnitude ($r \approx 0.25$ – 0.27). Log hypocentral distance is negatively associated with amplitudes ($r \approx -0.19$ to -0.23) and exhibits geometric dependencies with depth and angles (e.g., $r \approx 0.50$ with depth and $r \approx 0.60$ between depth and elevation angle).

4.1. Models and evaluation protocol

This section reports regression results obtained with tree-ensemble models trained under the physical-baseline predictor set described above. Evaluated models include CatBoost, LightGBM, Random Forest, and Extra Trees; additionally, a linear Ridge regressor is included as a parametric control to contrast against nonlinear ensembles. The comparison is framed as a temporal generalization exercise, using 2020–2024 for training and 2025 as an independent test set. The reported metrics describe performance under a catalog-based temporal reference and do not measure EEW latencies.

To avoid optimistic estimates, training performance is computed from out-of-fold (OOF) predictions obtained via cross-validation within 2020–2024. That is, each OOF prediction comes from a model that did not use that observation for fitting. Model selection and hyperparameter tuning (when applicable) are performed exclusively within the training period via cross-validation, so the 2025 set does not influence model choice. Final performance is reported as direct evaluation on the 2025 hold-out using hyperparameters fixed from 2020–2024.

Because the dataset is expanded (one earthquake can contribute multiple rows due to multiple stations and/or multiple valid triads), cross-validation is defined at the event level to control intra-event dependence. Specifically, folds are built using the earthquake identifier (FileMat) as the grouping variable (GroupKFold), so that all records associated with a given event are assigned to a single fold. This protocol prevents training on some records from an earthquake and evaluating on other records from the same earthquake, yielding a more conservative and realistic estimate of generalization. Performance is quantified using MAE, RMSE, and R^2 :

$$MAE = (1/N) \sum |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{(1/N) \sum (y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - [\sum (y_i - \hat{y}_i)^2] / [\sum (y_i - \bar{y})^2]$$

Unless stated otherwise, this section reports event-level metrics (one magnitude per earthquake), computed over pairs $(y_e, \hat{y}_e)$. Record-level metrics (event–station) are used only for specific diagnostics (e.g., residual versus distance) and are reported as complements.

4.2. Record-level prediction and event-level aggregation

Each model is trained and produces predictions at record level (event–station pair): for each row r associated with event e , a prediction $\hat{y}_{e,r}$ is obtained. However, in a rapid/operational setting the final goal is a single magnitude per earthquake. Therefore, in addition to record-level metrics, we report an event-level evaluation by aggregating the multiple predictions available for the same earthquake. For an event e with the set of valid records Re , we use the median as a robust estimator:

$$\hat{y}_e = \text{median}\{\hat{y}_{e,r} : r \in Re\}.$$

Event-level MAE, RMSE, and R^2 are then computed over pairs $(y_e, \hat{y}_e)$, which (i) aligns evaluation with the operational output (one magnitude per event) and (ii) avoids events with more stations dominating global metrics. In training, event-level evaluation is built from OOF predictions so the model does not benefit from having seen the same earthquake during fitting. In the 2025 test set, event-level predictions are obtained by aggregating record-level predictions generated by the model trained on 2020–2024. This protocol enables consistent comparison of performance and temporal generalization under an operationally relevant output definition.

Table 2 summarizes the main performance under an operationally relevant definition: one estimated magnitude per earthquake. On the 2025 test set (47 events), tree-based ensembles achieve bounded and mutually consistent errors, with MAE around 0.15 and RMSE in the 0.23–0.25 range, together with R^2 values around 0.68–0.73. LightGBM achieves the lowest test RMSE (0.233) and the highest R^2 (0.726) among evaluated models and is therefore adopted as the reference model for subsequent visualizations and analyses.

Table 2. Performance summary at the event level (one magnitude estimate per earthquake) for the physical-base feature variant. Train OOF (2020–2024): $N_{evt} = 269$. Test (2025): $N_{evt} = 47$. In Test, 95% confidence intervals (CI) for MAE and RMSE are estimated via event-level bootstrap ($B = 5000$).

Modelo	Train OOF (2020-2024)			Test (2025)		
	MAE	RMSE	R^2	MAE (CI 95%)	RMSE (CI 95%)	R^2
LightGBM	0.152	0.204	0.789	0.153 [0.108, 0.210]	0.233 [0.140, 0.324]	0.726
CatBoost	0.144	0.197	0.805	0.150 [0.104, 0.210]	0.238 [0.138, 0.342]	0.714
GBR	0.145	0.194	0.81	0.155 [0.108, 0.215]	0.239 [0.140, 0.338]	0.711
ExtraTrees	0.158	0.21	0.777	0.154 [0.106, 0.215]	0.246 [0.135, 0.363]	0.694
RandomForest	0.15	0.208	0.782	0.162 [0.113, 0.224]	0.251 [0.143, 0.356]	0.682
Ridge (tuned)	0.326	0.424	0.093	0.286 [0.208, 0.380]	0.415 [0.267, 0.572]	0.131
Dummy (mean)	0.342	0.447	-0.006	0.305 [0.222, 0.406]	0.445 [0.278, 0.627]	-0.002

Baselines play a methodological role rather than a competitive one. The Dummy model (constant prediction equal to the mean) yields approximately zero or negative R^2 on both Train and Test, indicating that the problem is not solved by a trivial estimator. Ridge (regularized linear control) marginally reduces error relative to Dummy

but remains substantially worse than ensembles (e.g., test RMSE 0.415 versus 0.233–0.251 for tree ensembles). This gap supports the conclusion that the relationship between early amplitudes and source–receiver geometry requires modeling nonlinear interactions.

Bootstrap confidence intervals on the 2025 test set further show that, with 47 events, statistical uncertainty is not negligible: the 95% CIs for MAE and RMSE overlap substantially across the different ensembles. Accordingly, the main result is not interpreted as one algorithm “beating” another by small margins, but as evidence that the complete pipeline (calibration, early displacement reconstruction, and extraction of compact physical predictors) enables magnitude estimation with realistic temporal generalization to 2025 under event-controlled evaluation.

4.3. Predicted vs. observed magnitude

To complement the numerical summary, Figure 4 shows the event-level predicted-versus-observed comparison for the reference model (LightGBM), including the identity line ($\hat{y} = y$). Panel A shows training performance using OOF predictions on 2020–2024, providing an honest estimate by preventing the model from evaluating earthquakes it has seen during fitting. Panel B shows temporal generalization on the 2025 hold-out.

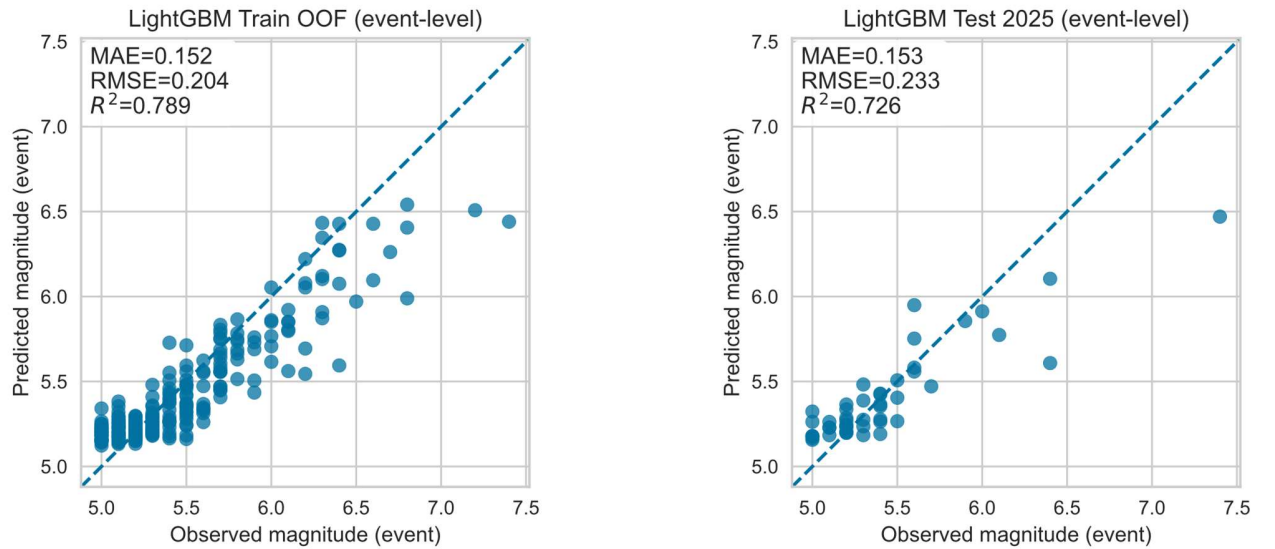


Figure 4. LightGBM as the reference model. Event-level predicted vs. observed comparison (one magnitude estimate per earthquake) with the identity line. Panel A: training period 2020–2024 using out-of-fold (OOF) predictions (honest evaluation with event-level GroupKFold). Panel B: direct evaluation on the 2025 dataset as a temporal hold-out.

In both panels, an approximately linear relationship is observed between true and estimated magnitudes, with bounded dispersion in the range where most of the dataset concentrates (mainly $M_{ww} \approx 5.0$ – 5.8). The gap between OOF and Test is moderate: MAE remains essentially stable, while RMSE increases on Test, consistent with this metric being more sensitive to a few higher-error events. Visually, the largest mismatches tend to occur toward the upper magnitude end; in some cases, underestimation is observed, which is expected when estimation is based on early information and limited event-level coverage.

The choice of LightGBM as the reference model is made for operational and parsimony reasons: among the ensembles evaluated, it achieves marginally the lowest RMSE and the highest R^2 on the test set, and is therefore used as the reference for the main visualizations and diagnostics. However, because differences among ensembles are small and their confidence intervals overlap widely, this work does not interpret the outcome as an algorithmic “victory,” but rather as evidence of robustness of the pipeline and temporal

generalization under event-level evaluation.

4.4. Error structure diagnostics

Beyond global performance, it is important to verify whether the error exhibits systematic structure with respect to physically meaningful variables. Figure 5 summarizes two complementary diagnostics on the 2025 set for LightGBM. Panel A shows the record-level residual $r = \hat{y} - y$ as a function of LogDistHypo. This plot is used to inspect distance-dependent bias and heteroscedasticity; ideally, residuals should be centered around zero without a marked trend. Here, residuals remain largely centered near zero across the distance range, with no clear monotonic trend, although dispersion and outliers increase in some ranges, consistent with variations in SNR, site effects, and instrumental uncertainty.

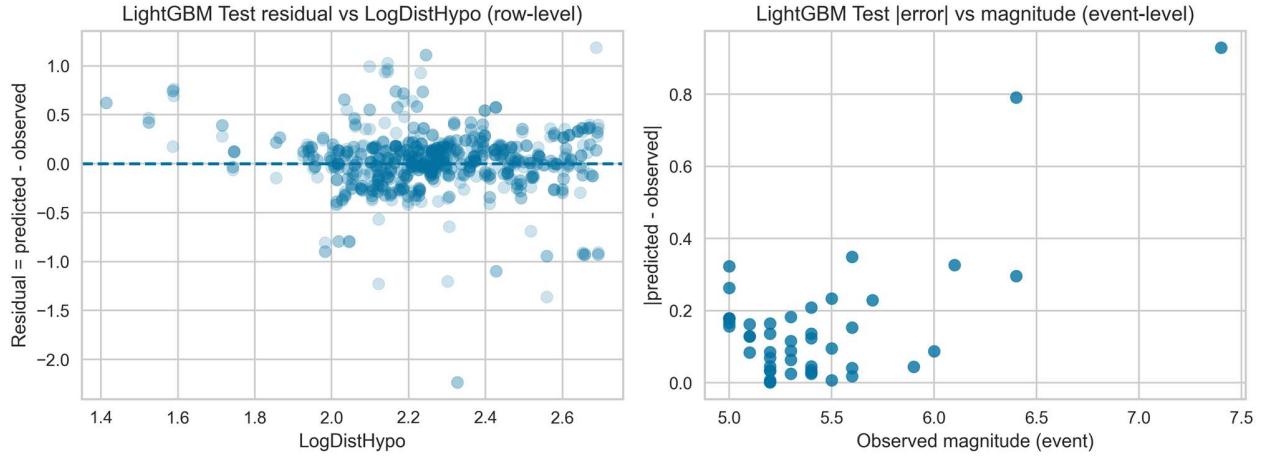


Figure 5. Compact error-structure diagnostics on the 2025 hold-out for LightGBM. Panel A: record-level residual $r = \hat{y} - y$ versus LogDistHypo $\log_{10}(d_{hypo}[km] + 1)$. Panel B: event-level absolute error versus observed magnitude (operational robustness: error control at the high-magnitude end).

Panel B shows the event-level absolute error $|\hat{y}_e - y_e|$ as a function of the true earthquake magnitude. This diagnostic is operationally relevant because the final system output is one magnitude per event. The figure indicates low errors for most moderate events and an increase in error at the upper magnitude end, where the 2025 set contains few examples and the error is dominated by one or a small number of earthquakes. Together, both panels support that the model does not show strong distance-related bias, but does face the most difficult and expected scenario in rapid/operational estimation: large and rare events, where early information may saturate or be insufficient due to limited coverage.

Additional event-level residual diagnostics for the 2025 hold-out are shown in Figure 6 for the reference model (LightGBM). Panel A presents the residual distribution. Residuals remain centered close to zero overall (mean residual = -0.011; SD = 0.235), which is consistent with the absence of a strong global bias, although a negative tail is observed. Panel B shows residuals as a function of predicted magnitude. Most events remain within a relatively narrow residual range, but a moderate negative trend is visible, indicating that the largest predicted magnitudes are associated with a small number of stronger underestimation cases.

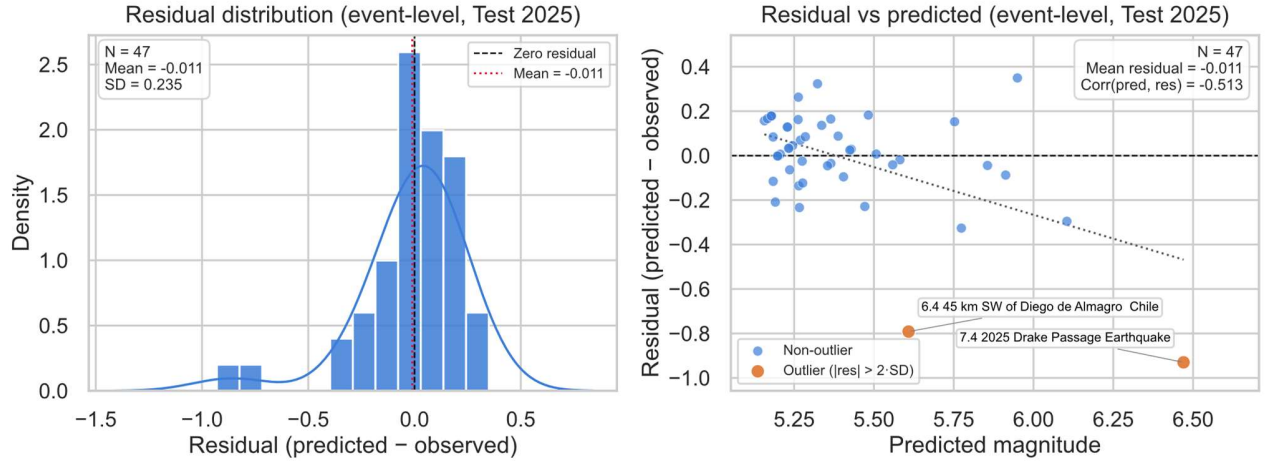


Figure 6. Additional event-level residual diagnostics on the 2025 hold-out for LightGBM. Panel A shows the residual distribution, including the zero-residual reference and the mean residual. Panel B shows residuals versus predicted magnitude, with event-level outliers highlighted using the criterion $|\text{residual}| > 2 \text{ SD}$.

To complement the graphical diagnostics, Table 3 summarizes the event-level outliers identified in the 2025 test set under the criterion $|\text{residual}| > 2 \text{ SD}$. Only two clear outliers were detected: the M 7.4 Drake Passage earthquake and the M 6.4 event located 45 km southwest of Diego de Almagro, Chile. Both correspond to marked underestimation cases. Notably, these events were aggregated from 13 and 18 valid event–station records, respectively, suggesting that the largest errors were not explained solely by extremely low coverage.

Table 3. Event-level outliers identified in the 2025 test set for LightGBM. Outliers were defined as events with $|\text{residual}| > 2 \text{ SD}$ in the event-level 2025 hold-out. The table reports the observed magnitude, predicted magnitude, residual, absolute error, and the number of event–station records used in the event-level aggregation.

Event ID	Observed magnitude	Predicted magnitude	Residual	Absolute error	Records used in aggregation
7.4 2025 Drake Passage Earthquake	7.4	6.470	-0.930	0.930	13
6.4 45 km SW of Diego de Almagro, Chile	6.4	5.609	-0.791	0.791	18

5. DISCUSSION

The central outcome of this study is not only that a specific model (e.g., LightGBM) achieves the best score in a table, but that a complete, auditable processing chain based on open standards can transform real seismic data (miniSEED + StationXML) into physically interpretable variables that enable early magnitude estimation with temporal generalization. In other words, the main contribution is the pipeline as a reproducible procedure: automatic station selection and consistent N/E/Z triad construction, download-integrity checks, calibration through instrumental sensitivity, filtering and integration to reconstruct displacement, drift correction, extraction of compact features, and construction of an event–station dataset evaluated under a protocol that controls intra-event dependence and avoids optimistic estimates.

Under this perspective, ensemble models act as an empirical verification that the proposed feature set retains information relevant to magnitude. In event-level evaluation (one magnitude per earthquake), tree-based ensembles achieve bounded errors on the 2025 temporal hold-out (e.g., $\text{MAE} \approx 0.15$, $\text{RMSE} \approx 0.23$, and $R^2 \approx 0.7$), while explicit baselines such as Dummy and Ridge perform substantially worse. This gap is not interpreted as an “algorithmic victory” of one method over another, but as evidence that the combination of early amplitude

and source–receiver geometry, extracted in a physically consistent manner, explains a substantial fraction of magnitude variability.

A key methodological element is the shift from row-dominated metrics to event-dominated metrics. In expanded datasets (multiple stations per earthquake), evaluating only at record level can overweight events with higher coverage and distort operational interpretation. Therefore, reporting primary metrics at event level using robust aggregation (median of record-level predictions) serves two purposes: it aligns reporting with the output required by a rapid/operational magnitude-estimation module (one magnitude per event), and it reduces sensitivity to instrumental outliers or individual stations with low signal-to-noise ratio. In addition, computing training metrics using event-grouped OOF predictions reinforces that evaluation is “honest,” because aggregation does not benefit from training on the same earthquake.

Error diagnostics support this reading. Residuals versus log hypocentral distance suggest no strong monotonic distance-dependent bias within the observed range, consistent with a model that jointly uses amplitudes and geometry to compensate for attenuation. However, dispersion and outliers appear in some distance ranges, which is expected for a heterogeneous network with site variability, historical instrumental differences, and fluctuations in SNR. Moreover, the increase in absolute error at the upper magnitude end is dominated by a small number of events, reflecting a structural difficulty: large earthquakes are rare and their early estimation is often more uncertain, whether due to partial saturation of early measures, limited instrumental coverage, or greater rupture complexity. In this sense, the observed behavior is consistent with what typically limits fixed-window early-magnitude systems. Additional event-level residual diagnostics reinforce this interpretation: residuals remain centered close to zero overall, but the negative tail and the identified outliers indicate that the largest errors are concentrated in a small number of higher-magnitude underestimation cases.

The simplifications adopted in the pipeline also condition interpretation. Instrument calibration is implemented through scalar sensitivity derived from StationXML, which is operationally simple and reproducible, but it is not equivalent to full instrument-response deconvolution in the frequency domain. This may introduce residual differences across sensors and bands, especially when metadata are incomplete or when response variations are not captured by an effective gain. Likewise, reconstruction via integration is sensitive to low-frequency drift; although corrections are applied (high-pass filtering and baseline removal), some records remain at risk of instability, which motivates the inclusion of physical-plausibility criteria and outlier filters. These are not “implementation details” but an integral part of the procedure’s value: they formalize how to control typical failure modes when working with real data.

Another relevant limitation is the temporal reference. Window centering uses the catalog origin time; in operation, detection, phase picking, and rapid location introduce uncertainties that shift windows and alter extracted features. Therefore, the reported results quantify the approach’s potential under an idealized reference and represent a necessary step before evaluating sensitivity to pick and location errors. Similarly, although spatial coverage is broad, the magnitude range is dominated by moderate events ($M_{\text{ww}} \geq 5.0$), and evidence at higher magnitudes depends on few examples, increasing statistical uncertainty. This is reflected in bootstrap confidence intervals on the 2025 set, which overlap among ensembles and discourage fine-grained conclusions about small differences between models.

Overall, the results support that a moderate-complexity approach (based on FDSN standards, explicit quality control, and physically interpretable features) can serve as a plausible block for early/operational magnitude estimation in Chile. The emphasis on an end-to-end pipeline and event-level evaluation with temporal generalization provides a solid basis for operational extensions: multi-station schemes with distance/quality weighting, multi-model combinations, and evaluation under detection and location uncertainty.

6. CONCLUSIONS

This work presented a reproducible pipeline for early seismic magnitude estimation in Chile, designed to operate on real data and open standards: StationXML inventories and triaxial miniSEED records. The workflow automates selection of nearby stations, consistent construction of N/E/Z triads with adequate sampling rate,

verification of download integrity, calibration through instrumental sensitivity, filtering and integration (single or double depending on physical units) to reconstruct displacement, drift correction, and extraction of compact, physically interpretable features (component peaks, horizontal and 3D displacement, and source–station distances including logarithmic transforms). The result is an auditable and traceable event–station dataset suitable for regression experiments under conservative protocols. On this dataset, tree-based ensemble models (Random Forest, Extra Trees, LightGBM, CatBoost, and a classical boosting method) were evaluated together with explicit baselines (Dummy and Ridge).

Evaluation was designed to reflect a direct operational goal: obtaining one robust magnitude per event, avoiding intra-event leakage and testing temporal generalization. Record-level predictions were aggregated by the median to produce an event-level magnitude estimate and to prevent uneven event coverage from distorting overall performance. Under this framework, ensembles achieve consistent performance on the 2025 hold-out (on the order of MAE $\approx$ 0.15, RMSE $\approx$ 0.23, and $R^2 \approx$ 0.7 at event level), clearly outperforming the baselines. Given the overlap of bootstrap confidence intervals, the result is interpreted as evidence of robustness of the complete procedure rather than a definitive advantage of a specific algorithm. Additional event-level residual diagnostics showed no strong global bias, while also identifying a small number of clear underestimation outliers in the 2025 hold-out, concentrated in larger-magnitude events.

Looking forward, three directions emerge as natural extensions to bring the approach closer to fully operational conditions. First, incorporate multi-station schemes that weight stations by distance, record quality, or prediction consistency, to reduce the impact of outliers and improve stability for large events. Second, explicitly evaluate pipeline sensitivity to uncertainty in detection, picking, and location by shifting windows and perturbing source parameters to approximate operational settings. Third, enrich the feature set with additional spectral or temporal descriptors while maintaining the physical traceability and quality-control criteria that characterize the proposed workflow.

In summary, the main contribution is an end-to-end, physically consistent demonstration grounded in FDSN standards, showing that it is feasible to move from raw waveforms and instrumental metadata to early magnitude estimates with realistic temporal generalization, establishing a solid baseline for developing integrable modules for rapid response and, potentially, as a component within rapid/operational estimation architectures once detection/picking is incorporated and end-to-end latency is evaluated.

REFERENCES

- Aedo, D., Melnick, D., Cisternas, M., & Brill, D. (2024). Tectonic control on great earthquake periodicity in south-central Chile. *Communications Earth & Environment*, 5(1), 703. <https://doi.org/10.1038/s43247-024-01834-5>
- Amerthen, S., Aron Nambikai Fernando, J., & Praba, M. (2025). Seismopredict: Advanced Machine Learning for Earthquake Damage Assessment. *2025 International Conference on Intelligent Systems and Computational Networks (ICISCN)*, (pp. 1-7). Bidar, India. <https://doi.org/10.1109/ICISCN64258.2025.10934698>
- Arrais, S., Urquiza-Aguiar, L., & Tripp-Barba, C. (2022). Analysis of Information Availability for Seismic and Volcanic Monitoring Systems: A Review. *Sensors*, 22(14), 5186. <https://doi.org/10.3390/s22145186>
- Chandrakumar, C., Prasanna, R., Stephens, M., & Tan, M. L. (2022). Earthquake early warning systems based on low-cost ground motion sensors: A systematic literature review. *Frontiers in Sensors*, 3, 1. <https://doi.org/10.3389/fsens.2022.1020202>
- Cremen, G., Galasso, C., & Zuccolo, E. (2022). Investigating the potential effectiveness of earthquake early warning across Europe. *Nature Communications*, 13(1), 639. <https://doi.org/10.1038/s41467-021-27807-2>
- da Silva, L. d., Montecino Castro, H., Aguayo Arias, M. I., González-Rodríguez, L., Rodríguez-López, L., & Cotias Simões, L. M. (2022). Signals of Surface Deformation Areas in Central Chile, Related to Seismic Activity—Using the Persistent Scatterer Method and GIS. *Applied Sciences*, 12(5), 2575.

- <https://doi.org/10.3390/app12052575>
- Devi D, R., Govindarajan, P., & N, V. (2024). Towards real-time earthquake forecasting in Chile: Integrating intelligent technologies and machine learning. *Computers and Electrical Engineering*, 117, 109285. <https://doi.org/10.1016/j.compeleceng.2024.109285>
- Dybing, S., Yeck, W., Cole, H., & Melgar, D. (2024). Rapid estimation of single-station earthquake magnitudes with machine learning on a global scale. *Bulletin of the Seismological Society of America*, 114(3), 1523-1538. <https://doi.org/10.1785/0120230171>
- Esposito, M., Marzorati, S., Belli, A., Ladina, C., Palma, L., Calamita, C., Pantaleo, D., & Pierleoni, P. (2024). Low-cost MEMS accelerometers for earthquake early warning systems: A dataset collected during seismic events in central Italy. *Data in Brief*, 53, 110174. <https://doi.org/10.1016/j.dib.2024.110174>
- Geiß, C., Priesmeier, P., Aravena Pelizari, P., Soto Calderon, A. R., Schoepfer, E., Riedlinger, T., Villar Vega, M., Santa María, H., Gómez Zapata, J. C., Pittore, M., So, E., & Fekete, A. (2023). Benefits of global earth observation missions for disaggregation of exposure data and earthquake loss modeling: evidence from Santiago de Chile. *Natural Hazards*, 119(2), 779-804. <https://doi.org/10.1007/s11069-022-05672-6>
- GEOFON Program (GFZ German Research Centre for Geosciences). (2006). *Plate Boundary Observatory Network Northern Chile (IPOC Seismic Network) (Network code CX)*. <https://doi.org/10.14470/PK615318>
- Han, G., Wang, D., & Sun, L. (2025). Estimating Moment Magnitudes of Moderate-to-Large Earthquakes in Western China Based on the Direct P Waves. *Bulletin of the Seismological Society of America*, 1. <https://doi.org/10.1785/0120250079>
- Joshi, A., Vishnu, C., & Krishna Mohan, C. (2022). Early detection of earthquake magnitude based on stacked ensemble model. *Journal of Asian Earth Sciences*: X, 8, 100122. <https://doi.org/10.1016/j.jaesx.2022.100122>
- Kaushal, A., Gupta, A. K., & Sehgal, V. K. (2025). Hybrid CatBoost and SVR Model for Earthquake Prediction Using the LANL Earthquake Dataset. *Informatica*, 49(14). <https://doi.org/10.31449/inf.v49i14.6524>
- Kolivand, P. S. (2024). A systematic review of Earthquake Early Warning (EEW) systems based on Artificial Intelligence. *Earth Sci Inform*, 17, 957–984. <https://doi.org/10.1007/s12145-024-01253-2>
- Megna, A., Cimini, G. B., Marchetti, A., Pagliuca, N. M., & Santini, S. (2022). A Waveform Dataset in Continuous Mode of the Montefeltro Seismic Network (MF) in Central-Northern Italy from 2018 to 2020. *Data*, 7(12), 169. <https://doi.org/10.3390/data7120169>
- Navarro, A., Castro-Artola, O., García-Guerrero, E., Aguirre-Castro, O., Tamayo Pérez, U., López-Mercado, C., & Inzunza Gonzalez, E. (2025). Recent Advances in Early Earthquake Magnitude Estimation by Using Machine Learning Algorithms: A Systematic Review. *Applied Sciences*, 15, 1. <https://doi.org/10.3390/app15073492>
- Noda, S., Iwata, N., & Korenaga, M. (2024). Improving the Rapidity of Magnitude Estimation for Earthquake Early Warning Systems for Railways. *Sensors*, 24(22), 7361. <https://doi.org/10.3390/s24227361>
- Pasten, D., Saravia, G., Vogel, E., & Posadas, A. (2022, 12). Information theory and earthquakes: Depth propagation seismicity in northern Chile. *Chaos, Solitons and Fractals*, 165(2), 112874. <https://doi.org/10.1016/j.chaos.2022.112874>
- Sarangpure, N., Dhamde, V., Roge, A., Doye, J., Patle, S., & Tamboli, S. (2023). Automating the Machine Learning Process using PyCaret and Streamlit. *2023 2nd International Conference for Innovation in Technology (INOCON)*, (pp. 1-5). Bangalore, India. <https://doi.org/10.1109/INOCON57975.2023.10101357>
- Shen, X., Hou, B., Lu, J., & Li, S. (2025). Comparative Analysis of Deep Learning Methods for Real-Time Estimation of Earthquake Magnitude. *Applied Sciences*, 15(5), 2587. <https://doi.org/10.3390/app15052587>
- Temneanu, M. C., Donciu, C., & Serea, E. (2025). Self-Contained Earthquake Early Warning System Based on Characteristic Period Computed in the Frequency Domain. *Applied Sciences*, 15(16), 9026. <https://doi.org/10.3390/app15169026>
- Turino, T., Saputro, R., & Karyono, G. (2025). Comparative Analysis of Decision Tree, Random Forest, Svm, and Neural Network Models for Predicting Earthquake Magnitude. *Jurnal Teknik Informatika (Jutif)*, 6, 755-774. <https://doi.org/10.52436/1.jutif.2025.6.2.2378>
- Universidad de Chile. (2012). *Red Sismológica Nacional (RSN) (Network code C1)*.

- <https://doi.org/https://www.fdsn.org/networks/detail/C1/>
Universidad de Chile, Dept. de Geofísica (DGF). (1991). C: *Chilean National Seismic Network*.
<https://www.fdsn.org/networks/detail/C/>
USGS Earthquake Catalog. (2025). *United States Geological Survey*.
<https://earthquake.usgs.gov/earthquakes/search/>
Wang, Y., Li, X., Wang, Z., & Liu, J. (2023). Deep learning for magnitude prediction in earthquake early warning,. *Gondwana Research*, 123, 164-173. <https://doi.org/10.1016/j.gr.2022.06.009>
Wibisono, A., & Naufal Rahmadika, R. (2025). Comparative Evaluation of Database Systems for High-Volume Seismic Prediction Data Management in Real-Time Applications. *Jurnal Ilmu Komputer Dan Informasi*, 18(2), 251–259. <https://doi.org/10.21609/jiki.v18i2.1539>
Wu, Y.-M., Mittal, H., Lin, Y.-H., & Chang, Y.-H. (2023). Magnitude determination using cumulative absolute absement for earthquake early warning. *Geoscience Letters*, 10(1), 1. <https://doi.org/10.1186/s40562-023-00314-6>