



A Data-Quality-Aware Machine Learning Framework for Magnitude Homogenization in Indonesia's Earthquake Catalog

Desi Ramayanti^{1*}, Bias Yulisa Geni², Imam Mulya³

¹²³Department of Informatic Engineering, Universitas Dian Nusantara, Indonesia

*desi.ramayanti@undira.ac.id

Abstract

Indonesia's national seismological agency, BMKG, reports earthquake magnitudes under ten distinct, non-interchangeable scale labels in its public catalog (2020–2025, 65,788 events), with no published conversion to a common Mw-equivalent scale. Existing machine learning studies using this catalog have assumed a homogeneous magnitude scale, and instrument-quality metadata in the catalog has never been used as a covariate for magnitude-conversion bias or uncertainty. This study introduces a powerful hybrid magnitude-homogenization framework. For events with a native moment-magnitude anchor (the body-wave family), we perform self-supervised internal calibration and tie it to the global Scordilis relation. For events without such an anchor, the local/regional family, we apply a literature-based conversion using the Egyptian National Seismic Network formula. A quality-aware machine learning model is then trained to estimate homogenized magnitude with calibrated uncertainty, conditioned on instrument-quality features. The validation process uncovers and quantifies a selection bias of approximately 1.8 to 2.1 magnitude units in BMKG's internal moment-magnitude reporting, reflecting that natively Mw-labeled events are a non-representative, significance-selected subsample rather than ground truth. The quality-aware model reduces point-prediction error (RMSE) by 50 to 70% and achieves better-calibrated 80% prediction intervals than a baseline without quality features. This combination of a practical homogenization solution and an explicit diagnostic framework is intended to be of direct use not only to BMKG and Indonesian seismology, but to any national seismological agency maintaining a similarly heterogeneous magnitude catalog without paired, same-event, multi-scale readings.

Keywords: instrument-quality metadata; machine learning; magnitude homogenization; selection bias diagnostics; uncertainty calibration

How to Cite: D. . Ramayanti, B. Y. . Geni, and I. . Mulya, "A Data-Quality-Aware Machine Learning Framework for Magnitude Homogenization in Indonesia's Earthquake Catalog", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 5, pp. 1224–1232, Sep. 2026, doi: [10.29207/resti.v10i5.8137](https://doi.org/10.29207/resti.v10i5.8137).

Permalink/DOI: <https://doi.org/10.29207/resti.v10i5.8137>

Received: June 22, 2026

Accepted: September 5, 2026

Available Online: September 28, 2026

*This is an open-access article under the [CC BY 4.0 License](#)
Published by [Ikatan Ahli Informatika Indonesia](#)*

1. Introduction

Indonesia's national meteorological, climatological, and geophysical agency BMKG (Badan Meteorologi, Klimatologi, dan Geofisika) maintains the national earthquake catalog [1]. As in most operational seismic networks, BMKG reports whichever magnitude scale is easiest to compute from available stations [2]. The catalog lists ten magnitude types (M, MLv, Mw(mB), mb, Mw, ML, Mjma, Mw(Mwp), mB, and Mwp), with no published homogenization to a single scale such as moment magnitude (Mw).

This magnitude-type heterogeneity is a well-known practical problem in seismology [3], [4]. Values from different scales are not directly comparable, and any downstream use of the catalog inherits these inconsistencies unless corrected first [5]. Although

heterogeneity is inherent to the BMKG catalog, existing machine learning studies have treated the magnitude column as a single homogeneous scale, without explicitly homogenizing different magnitude types [6], [7].

The catalog already includes two metadata fields describing detection quality for each event: Phase Count, the number of seismic phases used to compute the event, and Azimuth Gap, the largest angular gap between stations around the epicenter. These fields are standard proxies for hypocenter-location uncertainty in seismology but, to our knowledge, have not been used as covariates in magnitude-homogenization procedures.

This leaves two questions open for the BMKG catalog: whether magnitude-type bias depends on instrument quality, and whether a homogenization model using this

information can improve both point estimates and uncertainty calibration.

Magnitude homogenization using statistical and machine learning methods has been widely applied in seismically active regions worldwide [8-10], with methods shifting from simple linear regression to more advanced techniques [11], [12].

A study in northern Algeria built an expanded moment-magnitude dataset using general orthogonal regression (GOR) with an error-variance ratio, yielding a regional empirical conversion relation [13]. A multi-stage homogenization of the Egyptian earthquake catalog converted various magnitudes to Mw using the best-fitting linear relation between magnitude and seismic moment for each pair.

The Andaman–Nicobar subduction zone found that Gaussian Process Regression (GPR) outperformed other regression models for magnitude-scale homogenization, indicating that ML methods are starting to surpass classical linear regression in Mw prediction [14].

Other studies have built ML-assisted catalogs for specific sequences, such as the 2022 offshore Adriatic sequence and the 2024 Kefalonia swarm [15], [16], and for regional compilations, including a 1520–2020 Central American catalog homogenized to Mw and deduplicated using quality-based priority criteria [17].

Across these studies, conversion relations are obtained by regressing same-event magnitudes reported on different scales or networks. Magnitude-homogenization has been widely applied in active seismic regions—Algeria, Egypt, Saudi Arabia, the Shillong Plateau, and the Andaman–Nicobar subduction zone—but, to our knowledge, not explicitly to the BMKG earthquake catalog [11], [14], [18].

Existing machine learning studies on the BMKG catalog treat magnitude as a single homogeneous scale, ignoring the distinct magnitude types in the data [19]. Meanwhile, instrument-detection-quality metadata is well established as a proxy for hypocenter-location uncertainty in both conventional seismology and modern ML pipelines, but its use has remained limited to that context [20]. This study tackles both gaps head-on, introducing a machine-learning-based magnitude-homogenization framework for the BMKG catalog that delivers bias correction and uncertainty estimates explicitly conditioned on instrument-quality metadata.

This study is organized around three hypotheses evaluated in Section 3. H1: bias from magnitude-type heterogeneity (pseudo_mw – Magnitude) depends on instrument-detection quality (Phase Count, Azimuth Gap) independent of Magnitude. H2: a quality-aware machine learning model conditioned on instrument-quality metadata achieves lower point-prediction error (RMSE) than a baseline using only Magnitude and mag_type_family. H3: the same quality-aware model achieves prediction intervals that meet or exceed the

nominal 80% coverage level more reliably than the baseline. These map directly onto the two open questions above: H1 addresses whether magnitude-type bias depends on instrument quality, and H2–H3 address whether a homogenization model using that information improves point estimates and uncertainty calibration, respectively.

2. Methods

The dataset for this study was obtained from the BMKG Earthquake Repository [4] and includes 65,788 seismic events in and around Indonesia recorded between 1 January 2020 and 30 December 2025. Each record originally contained 12 fields: event identifier, origin time, hypocenter coordinates (latitude/longitude), depth, magnitude value and type, two instrument-quality metadata fields (phase count and azimuthal gap), a free-text location string, and the reporting agency. Epicenters span latitudes -10.99° – 5.99° and longitudes 95.01° – 140.99° ; magnitudes 0.19–7.89 (mean 3.34, median 3.25); depths 0–750 km (median 14 km, mostly shallow crustal); phase counts 4–506 (median 22); and azimuthal gaps 8.6° – 351.9° (median 128.0°).

Before homogenization, the raw repository was checked for duplication and found to contain 1,266 fully duplicated rows which were removed, leaving 64,522 unique events. The duplicates were unevenly distributed 1,172 (92.6%) were associated with magnitude type “M,” indicating that the duplication likely stems from the repository’s export or indexing process for this magnitude type rather than genuine repeated seismic events.

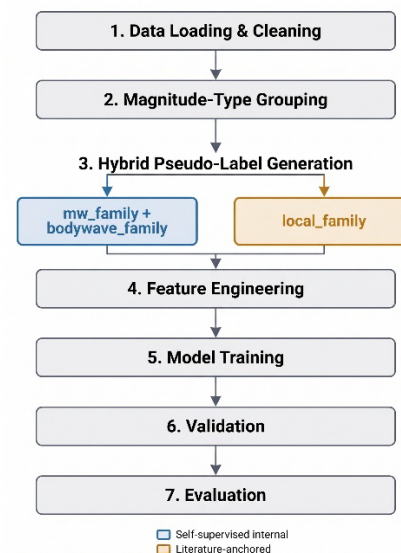


Figure 1. Research Stages and Hybrid Homogenization Pipeline

This is reported as a data-cleaning step, not a substantive finding. Twenty-three records missing a Location string were retained, since this field is not required for magnitude homogenization. Annual counts showed a strong anomaly: 2025 alone contributed 17,744 events after deduplication, about twice the

average of the other five years (2020–2024 ranged from about 7,200 to 10,900 events, averaging ~9,360). This discrepancy is consistent with an expansion or reconfiguration of the BMKG sensor network rather than a real doubling of seismic activity.

Figure 1 summarizes the end-to-end research pipeline, from data loading and cleaning to magnitude-type grouping, hybrid pseudo-label generation, feature engineering, model training, validation, and final evaluation of the research questions.

2.1 Data Preprocessing and Magnitude-Type Grouping

After deduplication, the ten raw magnitude-type labels were consolidated into three physically motivated families rather than treated as separate classes, preserving 99.994% of the cleaned dataset. Types are grouped by underlying scale family rather than exact label, and only the class that cannot be meaningfully merged is excluded.

The moment-magnitude family (*mw_family*: Mw, Mw(mB), Mw(Mwp), Mwp; $n = 1,385$ after deduplication) is the supervisory anchor set for homogenization; the body-wave family (*bodywave_family*: mb, mB; $n = 389$) groups broadband and short-period body-wave magnitudes on the same physical scale; and the local/regional family (*local_family*: M, MLv, ML; $n = 62,744$) contains the locally and regionally calibrated magnitudes that dominate the catalog.

The remaining type, Mjma ($n = 4$, 0.006% of the cleaned dataset), follows the Japan Meteorological Agency convention and is not directly comparable to the other scales without separate calibration, so it was excluded given its negligible size. The resulting modeling population, after this single minimal exclusion, comprises 64,518 records: 1,385 (*mw_family*), 389 (*bodywave_family*), and 62,744 (*local_family*).

2.2 Hybrid Magnitude Homogenization Framework

This study's main methodological contribution is a hybrid procedure for constructing the homogenization target, *pseudo_mw*, the Mw-equivalent magnitude used to train the downstream machine learning model. Because the BMKG repository lacks duplicate readings of the same event across different magnitude types or networks, classical homogenization cannot be applied directly.

For *mw_family*, *pseudo_mw* is simply the reported magnitude, since these 1,385 events are already on the moment-magnitude scale. For *bodywave_family*, *pseudo_mw* was obtained by quantile matching to the *mw_family* reference. This self-supervised procedure, which used no external event data, implies a typical offset of +0.13 to +0.45 magnitude units over that range.

The same quantile-matching procedure was initially applied to *local_family*, mapping each event's

percentile in the *local_family* distribution to the corresponding value in the *mw_family* reference distribution. This yielded a median shift of +1.806 magnitude units indicating a methodological artifact rather than a true scale offset.

An initial diagnosis attributed the distortion to a support mismatch: *mw_family*, by design, only includes events large enough for BMKG analysts to compute a full moment magnitude (median 5.02, range ~3.30–7.90), whereas *local_family* spans the full range of recorded seismicity, dominated by small-to-moderate routine earthquakes (overall median 3.25). Because quantile matching uses percentile ranks rather than absolute values, the much wider spread of *local_family* magnitudes was compressed into the narrow range of *mw_family*.

Within this overlapping magnitude band, the median shift remained +1.529, showing the distortion is mainly due to density/selection, not range/support. At the same nominal magnitude, BMKG preferentially computes full moment magnitudes for events that are systematically more “significant” than typical catalog events of that size. Thus, *mw_family* is a selection-biased subsample within each magnitude band.

Because no internally self-supervised mapping could be validated for *local_family*, *pseudo_mw* for this family was derived from a published empirical conversion formula instead of internal BMKG data, using mathematical relations from the seismological literature but no externally sourced event records.

We adopted the piecewise linear relation from the Egyptian National Seismic Network (ENSN) [21], [22], rather than a global teleseismic relation, because ENSN is, like the BMKG local/regional network, an empirical regression fitted to local-network-recorded magnitudes rather than teleseismically recorded ones — a methodological match in how the input magnitudes are measured, not a claim of tectonic similarity between Egypt and Indonesia.

$$pseudo_mw = 0.69 \times M + 0.58, \text{ for } M \leq 3.0 \quad (1)$$

$$pseudo_mw = 0.95 \times M - 0.15, \text{ for } M > 3.0 \quad (2)$$

Equations 1-2 were applied to all *local_family* events ($n = 62,744$). Because it does not use the *mw_family* distribution as a reference, no events fall outside its scope, and all *pseudo_mw* values are defined. The resulting median shift was −0.317, within the range implied by the ENSN relation and other regional ML–Mw studies served as a final sanity check confirming the formula's applicability across the observed magnitude range.

Each record's *pseudo_mw* has a categorical flag, *homogenization_method*, set to *internal_validated* for *mw_family* and *bodywave_family*, or *literature_anchored* for *local_family*, marking the two mechanisms and their distinct epistemic bases. Together, these mechanisms yield full coverage: all

64,518 modeled records have a valid, non-missing pseudo_mw.

2.3 Feature Engineering and Model Architecture

Two models were trained and compared. The baseline is an ordinary least-squares regression of pseudo_mw on Magnitude and mag_type_family only. Because this baseline is expected to have very low point-prediction error, comparing models solely by RMSE is insufficient.

The main model is a quality-aware gradient-boosted quantile regression (LightGBM) trained on all nine features to jointly predict the 10th, 50th, and 90th percentiles (Q10/Q50/Q90) of pseudo_mw, using Q50 as the point estimate and the Q10–Q90 range as the nominal 80% prediction interval before calibration [23], [24].

Model performance was evaluated with two complementary validation schemes. The primary scheme is 5-fold stratified cross-validation by

mag_type_family, ensuring the minority bodywave_family group ($n = 389$) remains proportionally represented in every fold rather than risking near-empty folds with a purely random split.

A secondary robustness check uses a temporal holdout split, training on 2020–2024 and testing on 2025, showing that sensor network changes during the study period can cause models trained only on earlier years to generalize poorly.

3. Results and Discussions

Selection-bias diagnostics during pseudo-label construction directly shape the literature-anchored mechanism. Table 1 quantifies the diagnosed artifact. Two internal quantile-matching variants produce shifts of +1.806 and +1.529 magnitude units including the ENSN relation used for local_family. Relative to this literature-anchored baseline, the variants imply selection biases of +2.123 and +1.846 magnitude units.

Table 1. Selection-bias Diagnostic for local_family

Version	Scope	Shift (median)	Literature Benchmark	Est. selection bias
QM internal — global (initial)	All local_family ($n=62,744$)	+1.806	n/a	+2.123
QM internal — in-scope (Option A)	local_family within calibration range ($n=28,880$)	+1.529	n/a	+1.846
Literature-anchored, ENSN piecewise (final)	All local_family ($n=62,744$)	−0.317	−0.3 to +0.5 (ENSN/regional ML–Mw literature)	0.000 (baseline)

This gap is not an error or inconsistency in BMKG's reporting. It reflects an expected difference between two populations, events for which BMKG computes a full moment magnitude and the broader set of recorded seismicity. The methodological lesson is that treating such a restricted but significant subset as representative ground truth for self-supervised magnitude homogenization yields a distortion of about 1.8–2.1 magnitude units. This diagnostic is a key contribution of this study, revealing a failure mode that, to our knowledge, has not previously been documented for self-supervised magnitude-homogenization pipelines using operational earthquake catalogs.

3.1 Bias and Instrument-Detection Quality

Bias from magnitude-type heterogeneity (bias = pseudo_mw – Magnitude) depends on instrument-detection quality, measured by Phase Count and Azimuth Gap. Figures 2 and 3 show bias versus Phase Count and Azimuth Gap for bodywave_family and literature_anchored, with clear, statistically significant correlations. For bodywave_family, bias increases with Phase Count ($r=+0.611$, $p<.001$) and decreases with Azimuth Gap ($r=-0.226$, $p=7\times 10^{-6}$). For literature_anchored, bias decreases with Phase Count ($r=-0.369$, $p<.001$) and increases slightly with Azimuth Gap ($r=+0.065$, $p<.001$).

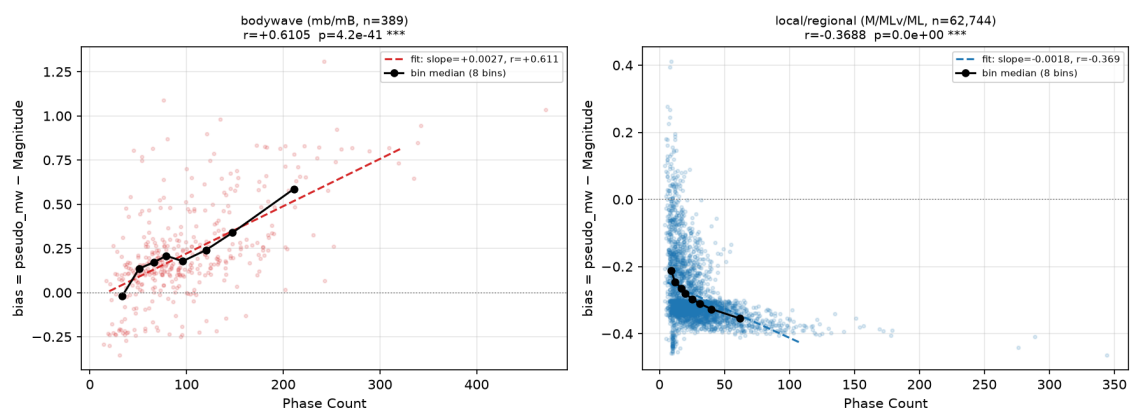


Figure 2. Bias (pseudo_mw – Magnitude) versus Phase Count

However, both bias and the two quality features are themselves correlated with Magnitude, since larger earthquakes are typically detected by more stations over

a wider azimuthal range, and bias is correlated with Magnitude by construction.

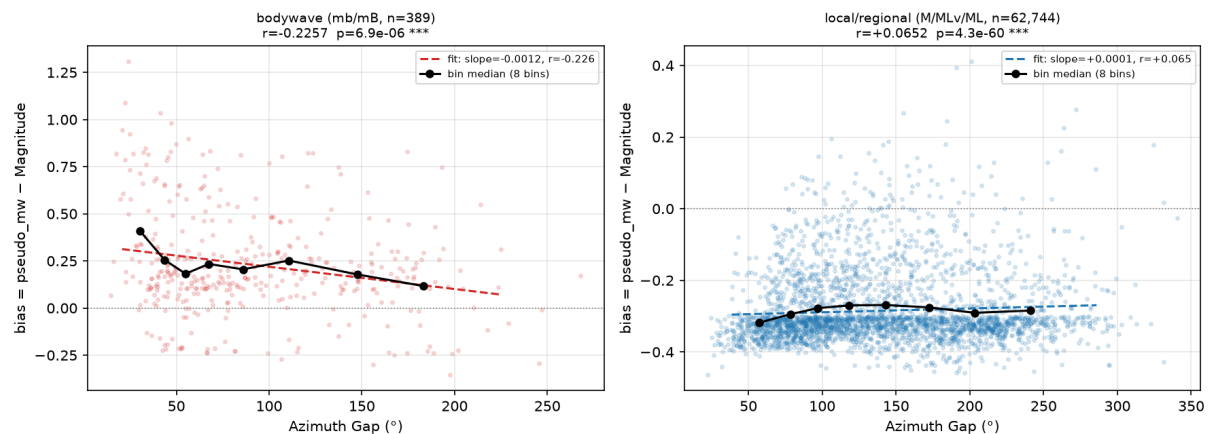


Figure 3. Bias (pseudo_mw – Magnitude) versus Azimuth Gap

Table 2 shows an OLS model of bias on Magnitude, Phase Count, Azimuth Gap, and their interactions with Magnitude, fitted separately by group, to test whether the apparent quality effect persists after this control. A reduced model without Magnitude (not shown) attains only $R^2=0.373$ for *bodywave_family* and $R^2=0.138$ for *literature_anchored*.

Table 2. OLS regression of bias on Magnitude, Phase Count, Azimuth Gap, and their interactions with Magnitude

Term	bodywave_family (n=389)	literature_anchored (n=62,744)
Azimuth Gap	0.00307 ***	-0.00027 ***
Magnitude × Phase Count	-0.00012 (ns, p=.321)	+0.00112 ***
Magnitude × Azimuth Gap	-0.00063 ***	+0.00008 ***
R ²	0.950	0.756

For *bodywave_family*, once Magnitude is controlled, Phase Count and its interaction with Magnitude become non-significant ($p=.218$, $p=.321$). However, the limited sample ($n=389$) reduces power, so this is only a suggestive null. For *literature_anchored*, Phase Count remains highly significant ($\beta=-0.00469$, $p<.001$) after controlling for Magnitude, but this effect disappears under a stricter test.

Because *pseudo_mw* for *literature_anchored* is constructed from an exact, bias is entirely determined by that exact, leaving no room for any other variable, including instrument-quality, to explain its variance. This was confirmed by replacing it with the correctly specified piecewise model, which had a residual standard deviation of 7.71×10^{-16} (floating-point noise).

When we add Phase Count and Azimuth Gap to the correctly specified piecewise model, R^2 doesn't budge (0.00 increase), and the Phase Count coefficient collapses to 8.83×10^{-17} . The “significance” of Phase Count in Table 2 for *literature_anchored* is therefore purely a model-specification artifact. It misreads the real curvature at the true $M=3$ breakpoint and falsely credits it to Phase Count and its interactions as spurious “significance.”

It shows for *literature_anchored*, and with solid statistical evidence for *bodywave_family* that the striking visual and statistical pattern in the raw correlations in Figures 2–3 can be generated entirely by confounding with Magnitude, rather than by any true bias–quality interaction. In other words, a seemingly persuasive signal can be fully manufactured by an unexamined homogenization pipeline that fails to explicitly control for Magnitude.

3.2 Point Accuracy and Probabilistic Calibration

Contrary to expectations of similar point accuracy across models, the quality-aware main model attains much lower RMSE than the baseline in both groups, a 49.7% reduction for *bodywave_family* and 70.5% for *literature_anchored*.

This shows that a simple linear model with only these two inputs fits *pseudo_mw* less well than a gradient-boosted model using all features. The nonlinear ENSN relationship used for *literature_anchored* cannot be captured by a single linear term, but the quality-aware model still incidentally accounts for it

Table 3. Model Comparison

Group	n	RMSE Baseline	RMSE Main (Δ%)	NLL Baseline	NLL Main	ΔNLL
bodywave_family	389	0.310	0.156 (–49.7%)	+0.248	–0.782	+1.030
literature_anchored	62,744	0.053	0.016 (–70.5%)	–1.510	–3.153	+1.642
EVAL ALL (pooled)	63,133	0.059	0.020 (–66.0%)	–1.419	–3.138	+1.719

Table 3 shows Δ NLL of +1.030 for *bodywave_family* and +1.642 for *literature_anchored*. The larger gain for

literature_anchored matches its more deterministic target, which gives the quality-aware model more

exploitable structure. In contrast, *bodywave_family*'s smaller, noisier sample ($n=389$) and quantile-matched target leave more irreducible uncertainty.

3.3 Calibration of Confidence Intervals

Table 4 reports coverage at the 80% nominal level used throughout. Both groups slightly exceed the target: 84.1% for *bodywave_family* and 86.7% for *literature_anchored*. For *bodywave_family*, calibration improves over the uncalibrated baseline from 78.2% to 84.1%, showing the benefit of the per-group conformal sigma floor for a group.

The modest overshoot above 80% in both groups, seen across all confidence levels in Figure 4, follows directly from how the floor is defined. Because it is set at the 80th percentile of out-of-fold absolute residuals,

observations whose residuals fall below the floor from the full out-of-fold distribution are more likely to lie inside the interval, raising average coverage.

Table 4. Coverage of the Nominal 80% Prediction Interval

Group	n	Coverage P80 Baseline	Coverage P80 Main
<i>bodywave_family</i>	389	78.2%	84.1%
<i>literature_anchored</i>	62,744	84.2%	86.7%
EVAL_ALL (pooled)	63,133	86.5%	86.6%

Figure 4 shows two plots of observed versus nominal coverage across the full confidence range (5%–95%), one for each group (*bodywave_family*, *literature_anchored*). The pattern reflects a deliberately conservative calibration strategy and is presented here as the intended behavior of the design, not as a miscalibration that needs fixing.

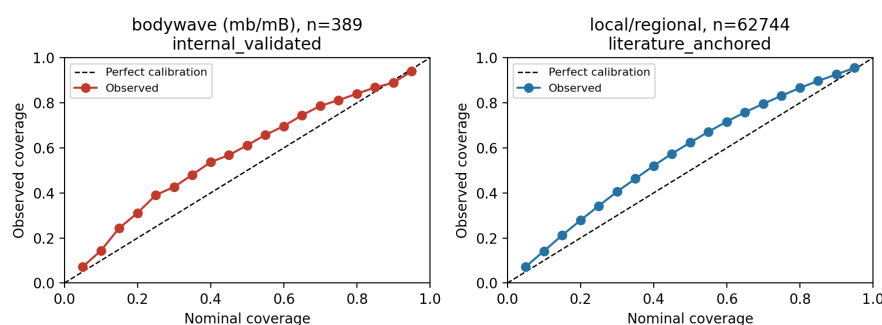


Figure 4. Calibration Curve: Observed vs. Nominal Coverage for the Quality-Aware Main Model

3.4 Synthesis and Practical Implications

The three main results—the selection-bias diagnostic, interval collapse during uncertainty calibration, and the falsification—may seem unrelated in pipeline order, but together they form one coherent methodological argument. The study's distinctive contribution is that all three issues were actively diagnosed and corrected, rather than left to silently distort the reported results.

The first mechanism is selection bias in how events receive ground-truth labels. The initial quantile-matching design assumed that events BMKG reports as Mw (*mw_family*) were representative of seismicity at each magnitude. Comparing these results with literature-reported offset magnitudes showed that the +1.806/+1.529 shifts were three to ten times too large to reflect a true scale effect, revealing *mw_family* as a selection-biased rather than representative subsample.

The second mechanism is confounding from the modeling pipeline's design rather than the data. Defining *pseudo_mw* as an almost deterministic function of Magnitude lets a sufficiently expressive model learn that function nearly perfectly, collapsing its uncertainty estimate. Taken at face value, this yielded a very poor apparent NLL (mean 177.6 across cross-validation folds for the pooled ALL population), which would naively suggest the quality-aware model is far worse at uncertainty estimation than a baseline that ignores quality features.

The third mechanism is the most direct illustration of the throughline. The raw correlations in Figures 2–3 are not statistically spurious, but they also do not show an instrument-quality effect on bias, because Phase Count and Azimuth Gap are correlated with Magnitude, and bias is constructed as a function of Magnitude.

These three cases are not united by subtle signals, but by the fact that a less skeptical pipeline might have reported each as a positive result. Beyond its specific homogenization framework and hypothesis tests, this study's broader methodological contribution is to show that self-supervised and pseudo-labeled magnitude homogenization must actively and structurally question its own intermediate results, and to demonstrate how to do so in practice.

It is important to distinguish the hypothesis test on direct bias–quality interaction (H1) from the justification for retaining a nonlinear, quality-aware model. The rejection of H1 shows that, once Magnitude is properly controlled, Phase Count and Azimuth Gap do not independently explain variation in the diagnosed bias term — a statement about the causal structure of one intermediate diagnostic quantity. It does not imply that the quality-aware model is unnecessary for the study's predictive task.

First, neither homogenization target is an affine function of Magnitude: *local_family* follows the fixed piecewise ENSN relation (Eq. 1–2), and *bodywave_family* follows a quantile-matched warping

with an offset ranging from +0.13 to +0.45 magnitude units across the observed range. A single global linear term, as used in the OLS baseline, cannot represent either relationship, so the RMSE reductions reported in Table 3 (−49.7% and −70.5%, respectively) are attributable primarily to the model's capacity to approximate these known nonlinear target functions, not to the two instrument-quality covariates in isolation.

Second, the OLS baseline yields only a single conditional-mean estimate and cannot natively produce the calibrated Q10/Q50/Q90 prediction intervals evaluated in Section 3.3; the quantile-regression objective, independent of the quality features' individual explanatory power, is what makes the interval-calibration comparison possible at all.

Third, retaining Phase Count and Azimuth Gap alongside Magnitude and `mag_type_family` adds negligible computational overhead to a nine-feature boosted-tree model already required for the two reasons above, and both features remain part of the per-group conformal sigma-floor procedure used for interval calibration in Section 3.3. We therefore treat the H1 test as a diagnostic finding about instrument quality's causal role in bias, separate from the architectural justification for the quality-aware model itself

For BMKG and similar national seismological agencies, this provides a clear rule for when internal, self-supervised homogenization is sufficient and when it must be anchored in the literature. The `homogenization_method` flag retained throughout the released catalog serves as an explicit, machine-readable transparency marker for downstream users, so they need not rediscover or assume which magnitude values rest on which epistemic basis.

3.5 Limitations

Two limitations follow directly from how `pseudo_mw` is constructed and should be stated explicitly rather than left implicit in the selection-bias diagnostic of Section 3. First, `pseudo_mw` is a constructed reconstruction target, not an independently validated Mw-equivalent ground truth. For `bodywave_family` it is obtained by self-supervised quantile matching against `mw_family`, and for `local_family` it is obtained from a literature-anchored piecewise relation (Eq. 1–2).

Consequently, the point-accuracy and calibration results in Sections 3.2–3.3 should be read as evidence that the quality-aware model reconstructs these two internally-defined targets more accurately and with better-calibrated uncertainty than a linear baseline — not as evidence of accuracy against an independently measured moment magnitude.

Second, the BMKG repository used in this study contains no paired, same-event, multi-scale readings and no shared event identifier with external catalogs such as USGS or ISC, so external validation against manually verified Mw values from those sources would

require an independent event-association procedure (matching by origin time, epicenter, and magnitude proximity) with its own error characteristics; this falls outside the scope of the present study and is left, as stated in for future work.

4. Conclusions

This study addresses a previously overlooked heterogeneity problem in machine learning applications using Indonesia's BMKG earthquake catalog. The catalog reports magnitudes under ten distinct, non-interchangeable scales, with no published conversion to a common Mw-equivalent scale. To close this gap, this study proposed a hybrid magnitude-homogenization framework with a machine learning model that conditions its point estimates and uncertainties on instrument-detection-quality metadata (Phase Count, Azimuth Gap).

This work yields three main findings. First, the validation process used to build the homogenization target revealed and quantified a selection bias of about 1.8–2.1 magnitude units in BMKG's internal moment-magnitude reporting. Events labeled Mw are not a representative magnitude sample but those deemed important enough for full moment-magnitude computation.

Second, the resulting hybrid framework—internally validated against the global Scordilis (2006) relation for body-wave events and based on a literature-anchored regional conversion (Egyptian National Seismic Network) for local/regional events—produced a fully homogenized catalog of 64,518 events with complete magnitude conversion.

Third, the quality-aware machine learning model outperformed a baseline without instrument-quality features in both point accuracy (RMSE reduced by 50–70%) and probabilistic calibration (80% interval coverage of 84–87% vs. 78.2% for body-wave events).

At the same time, this study frankly reports that the hypothesized direct interaction between magnitude bias and instrument quality (H1) was not supported after controlling for confounding by Magnitude—specifically, for literature-anchored events ($R^2=1.0$, β Phase Count ≈ 0). It treats this null result as strengthening, not weakening, the credibility of its diagnostic approach.

Beyond the specific homogenization framework and hypothesis results, this study's main contribution is methodological: it demonstrates that self-supervised and pseudo-labeled catalog homogenization must include active, structural scrutiny of intermediate results rather than relying on early sanity checks alone.

This combined practical homogenization method, and explicit diagnostic framework is meant for direct use by BMKG, Indonesian seismology, and any national agency with a similarly heterogeneous magnitude catalog lacking paired multi-scale readings.

Future work should emphasize regionally recalibrating the Indonesian conversion relation, cross-validating against overlapping USGS, ISC, or GCMT catalogs, and conducting a dedicated benchmark of competitive individual learners (including Random Forest, Extra Trees, Support Vector Regression, XGBoost, and CatBoost) and ensemble methods (weighted averaging, stacking, or dynamic ensemble selection) against the pseudo_mw catalog produced here, with statistical comparison of performance differences across cross-validation folds.

Declarations

AI Usage Statement

The authors used Claude solely to improve the readability and language of the manuscript; all scientific content, analysis, and conclusions are the authors' own.

Author Contributions

D.R: conceptualization, writing, original draft; B.Y.G: data curation, formal analysis; C.D: methodology, review and editing.

Funding Statement

This research was supported by an internal grant from Universitas Dian Nusantara (Grant Number: 11/34/H-SPK/I/2026).

Conflict of Interest Statement

The authors declare no conflict of interest.

Data and Code Availability

The processed dataset and the analysis code supporting the findings of this study are not publicly available due to institutional policy. Data and code are available from the corresponding author upon reasonable request

References

- [1] bmkg, "InaTEWS Earthquake Repository," bmkg.go.id.
- [2] A. Prasetyo, R. F. Aji, and W. S. Wibowo, "Assessing Information Security Awareness Among Indonesian Government Employees: A Case Study of the Meteorology, Climatology, and Geophysics Agency," *Journal of Information Systems Engineering and Business Intelligence*, vol. 11, no. 2, 2025, doi: [10.20473/jisebi.11.2.126-142](https://doi.org/10.20473/jisebi.11.2.126-142).
- [3] Hans Raj Wason, "A Statistical Analysis of Completeness of Earthquake Data around Dehradun city and its Implications for Seismicity Evaluation," *15th World Conference on Earthquake Engineering (15WCEE)*, 2012.
- [4] J. Woessner, J. L. Hardebeck, and E. Hauksson, "Theme IV – Understanding Seismicity Catalogs and their Problems What: What is an instrumental seismicity catalog," *Community Online Resource for Statistical Seismicity Analysis*, no. September, 2010.
- [5] I. Grigoratos, V. Poggi, L. Danciu, and R. Monteiro, "Homogenizing instrumental earthquake catalogs – a case study around the Dead Sea Transform Fault Zone," *Seismica*, vol. 2, no. 2, 2023, doi: [10.26443/seismica.v2i2.402](https://doi.org/10.26443/seismica.v2i2.402).
- [6] F. Hisyam, A. Susilo, M. Anshori, and M. F. R. Hasan, "Spatio-Temporal Variation Seismicity Pattern in East Java Between 2002 and 2022 Based on the b-value and Seismic Quiescence z-value," *Trends in Sciences*, vol. 21, no. 4, 2024, doi: [10.48048/tis.2024.7608](https://doi.org/10.48048/tis.2024.7608).
- [7] N. Mubarak MR, S. Susandri, and A. Zamsuri, "Geodetically-Enhanced Hybrid GRU with Adaptive Dropout and Dynamic L2 Regularization for Earthquake Parameter Prediction in Indonesia," *Jurnal Teknik Informatika (Jutif)*, vol. 7, no. 2, pp. 1483–1499, Apr. 2026, doi: [10.52436/1.jutif.2026.7.2.5478](https://doi.org/10.52436/1.jutif.2026.7.2.5478).
- [8] P. Adetokunbo et al., "Statistical and Machine Learning Methods for Magnitude Scale Conversion and Gutenberg-Richter Analysis: Insights from Regional Seismicity of Morocco," 2025, doi: [10.2139/ssrn.5139022](https://doi.org/10.2139/ssrn.5139022).
- [9] D. Di Giacomo, D. Bindi, S. Parolai, and A. Oth, "Residual analysis of teleseismic P-wave energy magnitude estimates: inter- and intrastation variability," *Geophys. J. Int.*, vol. 185, no. 3, pp. 1444–1454, Jun. 2011, doi: [10.1111/j.1365-246X.2011.05019.x](https://doi.org/10.1111/j.1365-246X.2011.05019.x).
- [10] E. Yee and W. Park, "Moment Magnitude Homogenization Relations in the South Korean Region from 1900 to 2020," *Applied Sciences (Switzerland)*, vol. 12, no. 14, 2022, doi: [10.3390/app12147215](https://doi.org/10.3390/app12147215).
- [11] H. M. Hussein et al., "Integration and magnitude homogenization of the Egyptian earthquake catalogue," *Natural Hazards*, vol. 47, no. 3, pp. 525–546, Dec. 2008, doi: [10.1007/s11069-008-9237-3](https://doi.org/10.1007/s11069-008-9237-3).
- [12] S. N. Dybing, W. L. Yeck, H. M. Cole, and D. Melgar, "Rapid Estimation of Single-Station Earthquake Magnitudes with Machine Learning on a Global Scale," *Bulletin of the Seismological Society of America*, vol. 114, no. 3, pp. 1523–1538, Jun. 2024, doi: [10.1785/0120230171](https://doi.org/10.1785/0120230171).
- [13] F. Bellale, A. Talbi, and S. Maouche, "A unified MW parametric earthquake catalog for Algeria and adjacent regions (PECAAR)," *Mediterranean Geoscience Reviews*, vol. 4, no. 4, pp. 427–443, Dec. 2022, doi: [10.1007/s42990-022-00085-w](https://doi.org/10.1007/s42990-022-00085-w).
- [14] H. M. Hussein et al., "Integration and magnitude homogenization of the Egyptian earthquake catalogue," *Natural Hazards*, vol. 47, no. 3, pp. 525–546, Dec. 2008, doi: [10.1007/s11069-008-9237-3](https://doi.org/10.1007/s11069-008-9237-3).
- [15] A. K. Bala and U. Saha, "Exploration of Regression Models for Homogenization of the Earthquake Magnitude Scales to Study Earthquakes in the Andaman-Nicobar Subduction Zone," *Journal of Volcanology and Seismology*, vol. 19, no. 5, pp. 490–507, Oct. 2025, doi: [10.1134/S0742046325700228](https://doi.org/10.1134/S0742046325700228).
- [16] A. Costanzo, "A New Catalogue and Insights into the 2022 Adriatic Offshore Seismic Sequence Using a Machine Learning-Based Procedure," *Sensors*, vol. 25, no. 1, 2025, doi: [10.3390/s25010082](https://doi.org/10.3390/s25010082).
- [17] V. Anagnostou, E. Papadimitriou, V. Karakostas, and T. Bäck, "Investigating the 2024 Swarm-Like Activity Offshore Kefalonia Island, Aided by Machine Learning Algorithms," *Pure Appl. Geophys.*, vol. 182, no. 9, 2025, doi: [10.1007/s00024-025-03766-3](https://doi.org/10.1007/s00024-025-03766-3).
- [18] C. Gamboa-Canté et al., "Seismicity in Central America (1520–2020) and Earthquake catalog compilation for seismic hazard assessments," *Bulletin of Earthquake Engineering*, vol. 22, no. 15, 2024, doi: [10.1007/s10518-024-02059-9](https://doi.org/10.1007/s10518-024-02059-9).
- [19] A. Boudebouda and A. Athmani, "Expanding moment magnitude dataset for earthquake magnitudes homogenization," *Acta Geophysica*, vol. 72, no. 6, 2024, doi: [10.1007/s11600-024-01317-3](https://doi.org/10.1007/s11600-024-01317-3).
- [20] S. Susandri, F. B. Nasution, F. Fajrizal, and Saparudin, "Spatial-Temporal Analysis of Earthquakes in Indonesia with Deep Learning Models: Performance Evaluation of CNN, LSTM, and Hybrid CNN-GRU," *Jurnal RESTI*, vol. 9, no. 5, 2025, doi: [10.29207/resti.v9i5.6538](https://doi.org/10.29207/resti.v9i5.6538).
- [21] U. Wijaya, Kusri, and A. H. Muhammad, "Indonesian Seismic Mitigation using Earthquake Predicted Artificial Intelligence Model," in *ICOLACT 2022 - 5th International Conference on Information and Communications Technology: A New Way to Make AI Useful for Everyone in the New Normal Era*, 2022, doi: [10.1109/ICOLACT55506.2022.9972091](https://doi.org/10.1109/ICOLACT55506.2022.9972091).
- [22] S. S. R. Moustafa, A. M. Faried, and M. H. Yassien, "Characterizing Egyptian National Seismic Network station sites using genetic optimization for microtremor data inversion," *J. Seismol.*, vol. 28, no. 6, 2024, doi: [10.1007/s10950-024-10245-0](https://doi.org/10.1007/s10950-024-10245-0).

- [23] M. Veisi *et al.*, "Seismic velocity and anisotropy of the uppermost mantle beneath Madagascar from Pn tomography," *Geophys. J. Int.*, vol. 126, no. 1, 2021.
- [24] G. Papacharalampous, H. Tyralis, N. Doulamis, and A. Doulamis, "Uncertainty estimation of machine learning spatial precipitation predictions from satellite data," *Mach. Learn. Sci. Technol.*, vol. 5, no. 3, p. 035044, Sep. 2024, doi: [10.1088/2632-2153/ad63f3](https://doi.org/10.1088/2632-2153/ad63f3).
- [25] A. D. Tsampas, E. M. Scordilis, C. B. Papazachos, and G. F. Karakaisis, "A Homogeneous Earthquake Catalog Of Intermediate-Deep Focus Global Seismicity: Completeness And Spatio-Temporal Analysis," *Bulletin of the Geological Society of Greece*, vol. 50, no. 3, 2017, doi: [10.12681/bgsg.11833](https://doi.org/10.12681/bgsg.11833).