ECEE2026
18th European Conference on Earthquake Engineering

Benchmarking GEM's 2026 Global Seismic Risk Model through Historical Scenarios

Anirudh Rao[1], Jovana Borozan[1], Christopher Brooks[1], Al Mouayed Bellah Nafeh[1], Catalina Yepes-Estrada[1], Vitor Silva[1]
1 Global Earthquake Model Foundation, Via Ferrata 1, 27100 Pavia, Italy
anirudh.rao@globalquakemodel.org

Abstract

The Global Earthquake Model (GEM) Foundation is finalising the 2026 edition of its Global Seismic Risk Model. This paper presents a historical-scenario benchmarking framework that tests the complete rupture-to-loss chain while distinguishing discrepancies associated with rupture modelling, ground-motion modelling, exposure, vulnerability, historical-impact data, and differences in loss scope. The working inventory now contains 1,335 earthquakes from 1975 onward across 79 countries, spans magnitudes 5.0–9.1, and combines reviewed USGS finite-fault rupture representations with ruptures matched from GEM regional seismic source models using the GEESE (Global EarthquakE ScEnarios) algorithm. A high-constraint subset is used for detailed loss diagnostics, while the wider inventory supports geographic coverage and screening for regional patterns. For each scenario, the OpenQuake engine generates ground-motion fields using the regional ground-motion model logic trees employed in GEM's global seismic hazard model, with station conditioning where sufficient recordings are available. Modelled direct building and contents losses are compared with reported historical losses adjusted to 2025 conditions by considering changes in costs, built-up surface area, and seismic vulnerability since the event. Preliminary country-level comparisons show differing regional tendencies and substantial within-country dispersion, indicating that a single global scaling correction would be inappropriate. These preliminary patterns are being used diagnostically to prioritise targeted reviews of regional exposure, replacement-cost, and vulnerability assumptions, with model updates introduced only where recurrent discrepancies across multiple events are supported by independent evidence. The novelty of the framework lies in combining a globally consistent scenario workflow with an explicit discrepancy-attribution and model-change protocol in which historical events are used as structured diagnostics and model revisions are accepted only when independently supported and consistent across multiple relevant tests, rather than through event-specific scaling.

Keywords: seismic risk, model validation, earthquake scenarios, loss normalisation, OpenQuake

1. Introduction

Probabilistic seismic risk assessment (PSRA) models combine seismic source characterisation, ground-motion modelling, exposure information, and vulnerability relationships to estimate earthquake consequences. Global models must provide consistent coverage across many countries while retaining enough regional specificity to support risk mitigation, disaster-risk financing, engineering, and policy applications. In this paper, a catastrophe model means the end-to-end system that combines an earthquake rupture, ground-motion simulation, exposed assets, and vulnerability relationships to estimate building damage and direct economic loss. The present benchmark focuses on direct building and contents losses due to ground shaking; critical infrastructure failure, business interruption, wider indirect losses, and triggered or secondary perils are outside the modelled loss scope at the current stage.

GEM's 2026 Global Seismic Hazard and Risk Models are being prepared as the next release in a sequence of global products increasingly used beyond the research community. This broader use raises the standard for transparency and testing. Probabilistic loss models cannot be validated definitively using the limited and heterogeneous evidence available from past earthquakes [1]. The practical objective is instead to demonstrate that the model is internally consistent, scientifically defensible, reproducible, and robust when confronted with independent empirical information.

Historical scenarios provide one of the most useful empirical anchors for this process because model outputs can be confronted with observations from real events. The comparison is nevertheless non-stationary, as a contemporary risk model represents a different building inventory, replacement-cost level, and vulnerability state from those that existed when an older earthquake occurred. A benchmark must therefore

address both the modelled rupture-to-loss chain and the temporal adjustment and comparability of the historical impact.

This paper contributes three elements to GEM's wider validation programme [2]: a globally consistent framework for selecting and classifying historical scenario evidence; a reproducible OpenQuake workflow that compares contemporary risk model outputs with temporally adjusted reported impacts for historical events; and a diagnostic change-control procedure intended to improve model components without tuning them to individual historical losses.

2. Scenario benchmarking within GEM's model validation framework

Styron et al. [2] describe GEM's 2026 model evaluation strategy as a hierarchy of component, integration, system-level, and acceptance tests. Component-level tests examine individual modules, including the seismic source model, ground-motion model, exposure model, and vulnerability model. Integration tests examine interactions between components such as the consistency of exposure and vulnerability under specified shaking, while system-level tests evaluate the complete hazard-to-loss chain. Acceptance tests add expert review, comparison with other models where available, and engagement with model users.

Historical scenarios are system-level diagnostics because they exercise the overall risk model under the rupture and ground-motion characteristics of real earthquakes, but they can also expose component-level deficiencies when the comparison is disaggregated. This hierarchy follows ideas from previous seismic-risk verification work, catastrophe-model validation practice, and robustness testing [3-6], and recognises that different tests are required to answer different questions about the model and its components.

Events are selected to span the major tectonic environments represented in GEM's regional seismic hazard models and to provide progressively different levels of observational constraint. The working inventory contains 1,335 earthquakes from 1975 onward, spans moment magnitudes 5.0–9.1, and covers 79 countries. Of these, 271 use USGS finite-fault or literature-based rupture information and 1,064 use a GEESE-matched rupture. Seismic station data are available for 546 events. The event hierarchy is summarised in Table 1.

Figure 1 shows the current 1,335-event set, which is expected to expand as modelling information is collected for additional events. Each star marks an event location, and the coloured footprints show the corresponding scenario peak ground acceleration (PGA). The spatial distribution demonstrates broad coverage of the principal seismic belts but also highlights an uneven empirical evidence base. Event coverage is densest in East Asia (412 events), Europe (184), Southeast Asia (144), North America (113), South America (110), and Oceania (105), while Africa (21) and Central Asia (14) remain comparatively sparse. This imbalance is treated as a limitation of the benchmark evidence rather than as a direct statement about confidence in the regional risk model.

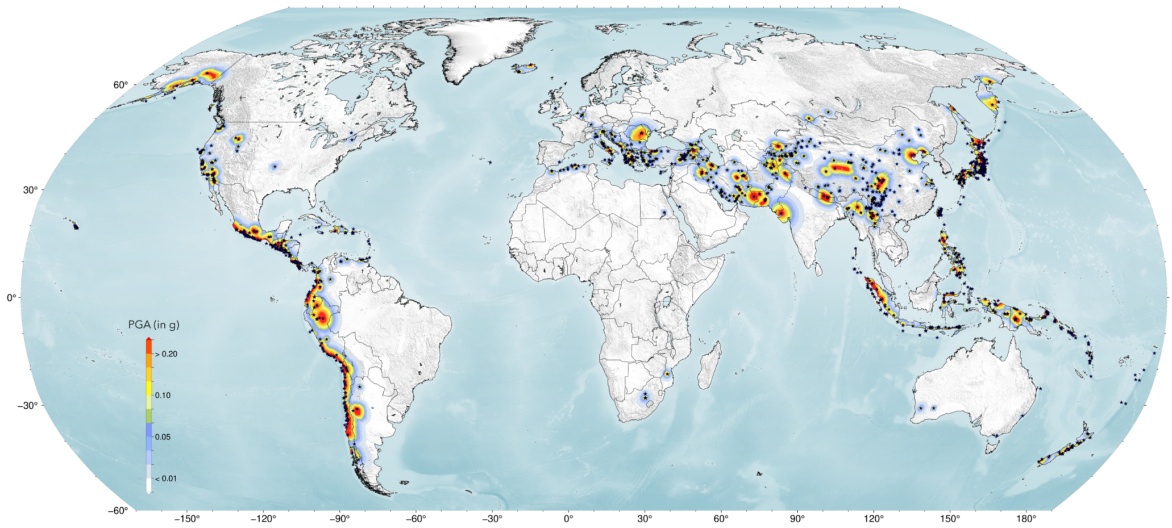


Figure 1 - Global distribution of 1,335 historical scenarios in the current benchmarking inventory. Stars indicate event locations; coloured footprints show scenario peak ground acceleration (PGA) in g, with warmer colours representing stronger shaking.

Table 1 - Scenario sets used in the historical benchmarking workflow

Scenario set	Event source	Rupture and ground-motion treatment	Role in the workflow
High-constraint reviewed subset (86 events)	Reviewed finite-fault models, station data, and reported impacts	Reviewed rupture; regional GMM logic tree; station conditioning where data are sufficient	Primary detailed scenario-risk calculations and diagnostic comparison
Structured extension (272 events)	40 events with reviewed ruptures and limited station coverage; 232 events with station data but no reviewed finite fault	Reviewed rupture where available; otherwise GEESE-matched candidate rupture; regional GMM logic tree; station conditioning where data permit	Extends rupture and ground-motion diagnostics beyond the 86-event core
Broad working extension (977 additional events)	AtlasCat and supplementary event catalogues, including additions that improve geographic and tectonic coverage	USGS rupture where available or GEESE-matched rupture; variable observational constraints	Global coverage, prioritisation, and screening for regional patterns

Figure 2 summarises the end-to-end workflow. The three event tracks feed into rupture selection and regional ground-motion simulation, followed by application of the contemporary exposure and vulnerability models. Historical impacts are normalised to 2025 conditions and passed through a comparability gate before discrepancies are interpreted. Diagnosed issues are then routed to the relevant hazard, exposure, vulnerability, or impact-data workflow, and any accepted revisions are re-tested across the validation hierarchy.

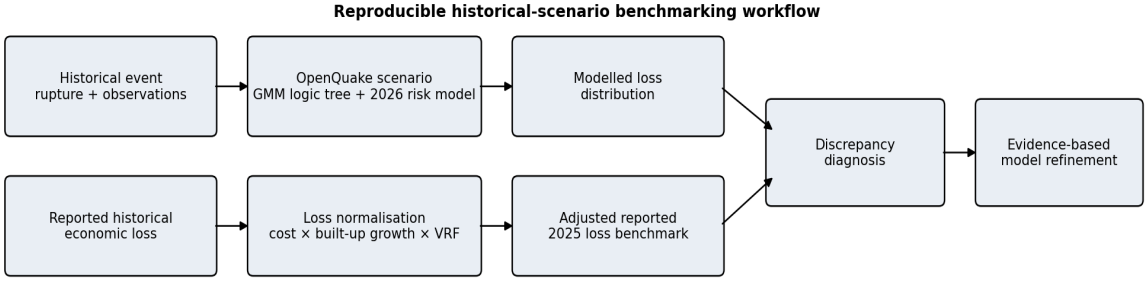


Figure 2 - Historical-scenario benchmarking and model-refinement workflow

3. Historical datasets, scenario calculations, and loss normalisation

The event inventory begins with the U.S. Geological Survey ShakeMap Atlas 4.0 and AtlasCat [7] and is supplemented by EM-DAT [8], the NOAA/NCEI Global Significant Earthquake Database [9], the Global Earthquake Impact Database (GEID) [10], and event-specific technical reports. The lower bound of 1975 reflects the availability of Global Human Settlement Layer (GHSL) built-up surface data and the greater consistency of inflation, exchange-rate, and exposure-growth information over the last five decades. The inventory is intentionally broader than the final quantitative loss benchmark: events with incomplete impact information remain useful for testing rupture selection, ground-motion simulation, geographic coverage, and the behaviour of model components under different tectonic settings.

For each event, the OpenQuake engine [11-13] combines a rupture representation with the GMM logic tree from the relevant regional GEM seismic hazard model, the 2026 global exposure model, and the corresponding vulnerability functions. Reviewed USGS rupture geometries are retained where available. Otherwise, GEESE [14] matches event magnitude, hypocentral location, and focal mechanism to candidate ruptures from the regional seismic source model; high-priority scenarios are then reviewed before use in detailed diagnostics. Ground-motion fields are generated in OpenQuake rather than imported from the ShakeMap Atlas so that the same GMMs used in the probabilistic model are exercised in the scenarios. The tectonic region type is inferred from the rupture and source-model context, epistemic uncertainty in the regional GMM logic tree is retained, and recordings at rupture distances within 100 km are used to condition the simulated fields when available. Events without adequate recordings are simulated without station conditioning.

The exposure model represents the contemporary spatial distribution, occupancy, replacement cost, and structural taxonomy of the building stock [15], while the vulnerability model translates shaking intensity into damage and loss for the relevant building classes [16-17]. Outputs are retained at several levels: total event loss, loss ratio, damage-state distribution, building-class contribution, and spatial summaries by first-level administrative division. These disaggregated outputs are essential for attributing aggregate discrepancies to particular model assumptions.

Reported historical losses are not compared directly with the contemporary model. The adjustment follows the preferred Method 3B of Karimzadeh et al. [18], which performed best in tests using paired historical and recent exposure models for Costa Rica, Italy, and Ethiopia. In compact form, the adjusted benchmark is

$$L_{adj,s} = L_{rep,t} \times F_{cost} \times F_{BU} \times \text{VRF}, \quad (1)$$

where $L_{rep,t}$ is the reported event-year loss, F_{cost} converts the loss from event-year USD to local currency, applies the country-specific Consumer Price Index to the 2025 reference year, and converts the result back to 2025 USD; F_{BU} is the ratio of 2025 to event-year built-up surface area derived from GHSL; and VRF is the vulnerability reduction factor, defined as the ratio of present-day to historical vulnerability. The vulnerability factor represents the shares of construction associated with No, Low, Moderate, and High Code regimes, the level of code enforcement, and the relative vulnerability reduction associated with each regime [18]. Enforcement is represented by five levels from fully enforced to not enforced, and partially enforced construction is split between the prevailing and preceding code classes.

This is a forward-normalisation approach: the 2026 risk model represents the contemporary building stock, while the historical loss is trended to a comparable present-day basis. Major earthquakes can accelerate the demolition of highly vulnerable buildings, the reconstruction of damaged or destroyed buildings, and changes in codes or enforcement. Such effects are represented only insofar as they are captured by subsequent built-up growth and the documented evolution of design regulations and enforcement; event-specific reconstruction programmes are not otherwise modelled explicitly. Where these assumptions are weak, the benchmark is interpreted with lower confidence rather than as an exact target.

Three confidence attributes are therefore tracked separately for each event: the quality of the shaking constraint, the reliability and scope of the reported impact, and the reliability of the temporal normalisation. Each is classified as high, moderate, or low rather than collapsed into a single score. This distinction is important because confidence in a historical benchmark is not equivalent to confidence in the regional risk model. Loss-scope compatibility is also recorded: the model currently estimates direct building and contents loss, whereas reported economic losses may include infrastructure, indirect losses, or triggered perils such as liquefaction, landslide, tsunami, or fire following earthquake. Where possible, contributions from these non-modelled components are identified and excluded. Primary quantitative interpretation nevertheless gives greatest weight to events with comparable loss scope.

These attributes are not averaged into a single regional validation score. Benchmark coverage, evidence quality, and residual consistency are reported separately because they answer different questions. A region with only a few high-quality historical events may have a weak empirical benchmark even if the underlying model is well supported by component-level evidence; conversely, a large number of poorly attributed loss estimates does not constitute strong validation evidence. Regions with sparse scenario coverage are therefore identified explicitly as evidence gaps.

Each adjusted reported loss is compared with the full simulated loss distribution rather than a single deterministic value. The median is used as the primary point comparison because event-loss distributions are typically right-skewed; the mean is also retained as the expected modelled loss. Percentile intervals and ratios of the adjusted reported loss to the modelled median and mean are used to assess whether discrepancies are large relative to model uncertainty.

4. Diagnostic benchmarking and model-refinement protocol

The key purpose of the workflow is to convert a mismatch into a diagnosis rather than an automatic calibration factor, thereby avoiding tuning the model to individual events. The baseline model and calculation settings are first frozen and versioned. The comparison is then examined in sequence: rupture representation and tectonic-region assignment; simulated shaking across the GMM branches; exposure location and value; building-type and code-level composition; exposure-to-vulnerability mapping; vulnerability functions; and, finally, the quality, normalisation, and scope of the reported impact. Spatial damage patterns and building-class contributions are used wherever available to distinguish alternative explanations for the same aggregate loss discrepancy, although such detailed damage and loss information is available only for a smaller subset of the complete event set.

The model parameters that may be refined are therefore explicit. Exposure revisions may affect the spatial allocation of assets, building-type fractions, construction-era or earthquake-resistant-design assignments, and replacement costs. Vulnerability revisions may affect the mapping from exposure classes to existing functions and, only when supported by independent empirical or analytical evidence, the functions themselves. Rupture, tectonic-region, or GMM issues are referred to the hazard-model validation workflow [2]. No event-specific scaling factor is applied to total losses, and vulnerability functions are not rescaled solely to force agreement with a single earthquake.

The order of diagnosis is important: a total-loss mismatch should not be assigned to vulnerability if the simulated shaking is inconsistent with recordings, nor should replacement costs be modified to compensate for a rupture or GMM bias. Similarly, a historical estimate that includes infrastructure, business interruption, or secondary-peril losses is not an appropriate direct target for a building-and-contents loss model. The workflow therefore applies a comparability gate before model changes are considered: shaking consistency, impact scope, and temporal-normalisation quality must first be judged adequate for the intended diagnostic.

A proposed revision is accepted only when the diagnosed deficiency is supported independently and the change is consistent across multiple relevant scenarios, regions, building classes, or component tests. The revised component is then re-evaluated through the applicable component, integration, and system-level regression tests. This change-control procedure allows scenario benchmarking to enhance the scientific integrity of the global model while limiting overfitting to a small and heterogeneous historical sample.

5. Preliminary findings and diagnostic examples

Figure 3 presents the updated preliminary comparison for the countries with currently available loss simulations. These results are a snapshot of the most recent iteration and remain subject to change while the benchmarking exercise is in progress. The results are stratified by country because a single global scatter plot can hide compensating biases: underestimation in one region and overestimation in another may cancel in aggregate. The solid black line represents 1:1 agreement between modelled and adjusted reported loss, while the dashed blue line summarises the country-level trend. The panels are used as screening diagnostics rather than regional performance grades because sample size, event magnitude, loss scope, and evidence quality differ among countries.

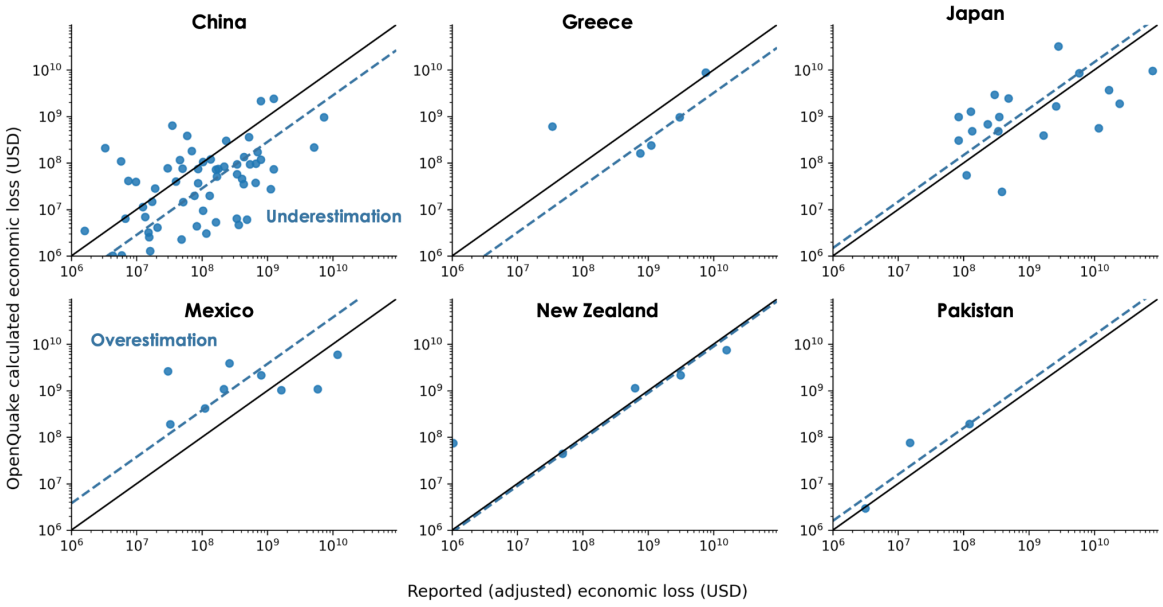


Figure 3 - Preliminary country-level comparison of OpenQuake modelled direct structural and contents losses with adjusted reported historical economic losses. All values are expressed in 2025 USD. Solid black lines indicate 1:1 agreement; dashed blue lines summarise the country-level trend. The panels contain the currently available simulated events for China, Greece, Japan, Mexico, New Zealand, and Pakistan.

The six panels in Figure 3 show materially different patterns. The China panel contains the largest sample and exhibits broad dispersion, with the country-level trend below the 1:1 line and hence an overall tendency toward underestimation. Greece shows a similar tendency, particularly at higher reported losses. Mexico trends above the 1:1 line, indicating overestimation for the available events. The limited New Zealand sample lies close to the 1:1 line, while Japan shows both substantial overprediction and underprediction and therefore cannot be summarised by a single bias without further stratification. The Pakistan sample is too small to support a regional conclusion. These patterns should be interpreted as screening diagnostics because the number, magnitude, loss scope, and evidence quality of events differ substantially among countries.

The value of the country panels is that repeated behaviour across several independent events is more informative than an isolated outlier. Persistent underestimation can motivate checks of replacement costs, missing exposure, building-class composition, vulnerability mapping, or reported-loss components outside the model scope. Persistent overestimation can motivate checks for overvalued exposure, overly vulnerable class

assignments, or incomplete historical reporting. Before any change is made, the shaking field and benchmark comparability are reviewed. A model revision is considered only if the same deficiency is supported by independent evidence and remains visible across multiple events or component tests. The dashed country trends therefore guide diagnosis but are not themselves calibration factors.

The next stage is to add event-level uncertainty intervals and the three evidence attributes described in Section 3 to the plotted points. This will allow the final benchmark to distinguish between a large residual for a high-confidence, loss-scope-compatible event and an equally large residual associated with uncertain historical reporting or secondary perils. The frozen 2026 release candidate will then be summarised using log residuals, the fractions of events within factors of 2, 5, and 10, and coverage of the simulated loss intervals.

6. Conclusions

Historical scenarios provide a practical empirical benchmark for GEM's 2026 Global Seismic Risk Model when they are treated as structured evidence rather than pass/fail tests. The working inventory now exceeds 1,300 events and combines a high-constraint core with progressively broader rupture and impact coverage. This architecture allows the same framework to support detailed loss diagnosis, ground-motion and rupture testing, and the identification of geographic evidence gaps. The workflow links reviewed and source-model-derived ruptures, regional GMM logic trees, the contemporary global exposure and vulnerability models, and a tested procedure for adjusting reported historical losses to 2025 conditions while making benchmark confidence and loss-scope compatibility explicit.

The updated country-level comparison indicates that a single global scaling correction would be inappropriate: the six country panels show different tendencies and substantial within-country dispersion. The principal contribution of the workflow is therefore the traceable path from discrepancy to evidence-based model refinement. Potential changes are restricted to identifiable exposure, vulnerability, or hazard assumptions; event-specific tuning is excluded; and accepted revisions must remain consistent across multiple tests. The final benchmark will use the frozen 2026 release candidate and report quantitative performance statistics together with event-level evidence and loss-scope classifications.

References

[1] Crowley, H. M.; Stafford, P. J.; Bommer, J. J.: Can earthquake loss models be validated using field observations?, Journal of Earthquake Engineering, 12(7), 2008, pp. 1078-1104. https://doi.org/10.1080/13632460802212923

[2] Styron, R.; Rao, A.; Silva, V.; Yepes-Estrada, C.; Brooks, C.; Nafeh, A. M. B.; GEM Seismic Hazard and Risk Teams: Validation and Performance Evaluation of GEM's Global Seismic Hazard and Risk Models, 13th U.S. National Conference on Earthquake Engineering, 2026.

[3] Crowley, H. M.; Dabbeek, J.; De Maio, F. V.; Despotaki, V.; Faravelli, M.; Borzi, B.; Silva, V.; Martins, L.; Kalakonas, P.; Weatherill, G. A.; Riga, E.; Karatzetzou, A.; Pitilakis, K. D.; Anastasiadis, A.; Pitilakis, D.; Fotopoulou, S.; Michelini, A.; Faenza, L.: Testing and verification of the European Seismic Risk Model (ESRM20), 2020.

[4] Risk Management Solutions, Inc.: Principles of Model Validation: United States Hurricane Model, Newark, California, United States, 2012.

[5] Guin, J.: What makes a catastrophe model robust, in: Risk Modeling for Hazards and Disasters, Elsevier, 2018, pp. 47-62.

[6] Riga, E.; Karatzetzou, A.; Apostolaki, S.; Crowley, H. M.; Pitilakis, K. D.: Verification of seismic risk models using observed damages from past earthquake events, Bulletin of Earthquake Engineering, 19(2), 2021, pp. 713-744.

[7] Marano, K. D.; Hearne, M.; Jaiswal, K. S.; Thompson, E. M.; Worden, C. B.; Wald, D. J.: ShakeMap Atlas 4.0 and AtlasCat: An archive of recent and historical earthquake ShakeMaps and impacts for global hazard analyses and loss model calibration, Seismological Research Letters, 2023. https://doi.org/10.1785/0220220324

[8] Delforge, D.; Wathelet, V.; Below, R.; Sofia, C. L.; Tonnelier, M.; van Loenhout, J. A. F.; Speybroeck, N.: EM-DAT: the Emergency Events Database, International Journal of Disaster Risk Reduction, 124, 105509, 2025. https://doi.org/10.1016/j.ijdrr.2025.105509

[9] NOAA National Centers for Environmental Information/World Data Service: NCEI/WDS Global Significant Earthquake Database, NOAA National Centers for Environmental Information, 2026. https://doi.org/10.7289/V5TD9V7K

[10] Karim Zadeh, Z.; Silva, V.; Yepes-Estrada, C.; Caruso, M.; So, E.; Wald, D.; Jaiswal, K.; Marano, K.: Development of a Global Earthquake Impact Database (GEID) for the Verification and Calibration of Loss Models, Earthquake Spectra, 42(1), e70006, 2026.

[11] Pagani, M.; Monelli, D.; Weatherill, G.; Danciu, L.; Crowley, H.; Silva, V.; Henshaw, P.; Butler, L.; Nastasi, M.; Panzeri, L.; et al.: OpenQuake engine: An open hazard (and risk) software for the Global Earthquake Model, Seismological Research Letters, 85(3), 2014, pp. 692-702.
[12] Silva, V.; Crowley, H.; Pagani, M.; Monelli, D.; Pinho, R.: Development of the OpenQuake engine, the Global Earthquake Model's open-source software for seismic risk assessment, Natural Hazards, 72(3), 2014, pp. 1409-1427. https://doi.org/10.1007/s11069-013-0618-x
[13] Rao, A. S.; Yepes-Estrada, C.; Johnson, K. L.; Todorović, L.; Pagani, M. M.; Simionato, M.; Styron, R.; Silva, V.; Crowley, H. M.: Evolution of the OpenQuake Engine: Enhanced capabilities, collaborative development, and global adoption, Earthquake Spectra, 2025. https://doi.org/10.1177/87552930251378157
[14] Brooks, C.; Pagani, M.; Villani, M.; Johnson, K.; Styron, R.; Bayliss, K.: Global EarthquakE ScEnarios (GEESE): An OpenQuake Engine-Based Rupture Matching Algorithm and Scenarios Database for Seismic Source Model Testing and Rapid Post-Event Response Analysis, Seismica, 4(2), 2025. https://doi.org/10.26443/seismica.v4i2.1654
[15] Yepes-Estrada, C.; Calderon, A.; Costa, C.; Crowley, H.; Dabbeek, J.; Hoyos, M. C.; Martins, L.; Paul, N.; Silva, V.: Global building exposure model for earthquake risk assessment, Earthquake Spectra, 39(4), 2023, pp. 2212-2235.
[16] Aljawhari, K.; Nafeh, A. M. B.; Silva, V.: An Improved Global Vulnerability Model for Seismic Risk Assessment: Part 1 - Structural Vulnerability, Bulletin of Earthquake Engineering, in press, 2026. https://doi.org/10.21203/rs.3.rs-7935593/v1
[17] Nafeh, A. M. B.; Aljawhari, K.; Silva, V.: An Improved Global Vulnerability Model for Seismic Risk Assessment: Part 2 - Non-Structural and Contents Vulnerability, Bulletin of Earthquake Engineering, in press, 2026.
[18] Karimzadeh, Z.; Silva, V.; Borozan, J.; Rao, A.: Developing a Global Database of Adjusted Economic Losses for the Verification of Loss Models, in preparation.