# Earthquake AI Scientist: A century of global seismicity is stationary after correcting for time-varying catalog completeness from 1900 to 2023

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Reported global earthquake counts rise through the twentieth century, but whether this reflects a physical increase in seismicity or evolving detectability has remained unresolved. We compile a 1900–2023 global catalogue (37,331 significant events; minimum $M \approx 5.5$) and explicitly reconstruct a time-varying magnitude of completeness, $M_c(t)$, across instrumentation eras. Gutenberg–Richter parameters are estimated strictly above $M_c(t)$ with bootstrap uncertainty, and completeness-consistent rate tests are conducted globally and by tectonic regime. Once $M_c(t)$ is enforced, the apparent secular rise in counts disappears; decadal $b$-values are stable within uncertainty; and large-earthquake rates are statistically stationary at robust thresholds (notably $M \geq 7.5$–$8.0$). We further translate this baseline into decadal forecasts: for 2025–2034 we predict $N_{10}(M \geq 7.5) = 42.3$ events on average (90% interval $[32, 53]$) and $N_{10}(M \geq 8.0) = 8.3$ ($[4, 13]$), implying $\sim 56\%$ annual probability of at least one $M \geq 8.0$ event. These results resolve the long-standing debate over secular changes in global seismicity and establish a completeness-aware, uncertainty-quantified foundation for probabilistic hazard models, risk pricing and performance benchmarking of global monitoring networks. Beyond seismology, the framework generalises to observation-limited records (e.g., volcanic unrest, landslides, epidemiological surveillance, biodiversity monitoring), providing a template for bias-corrected trend detection and forecasting.

## Introduction

Global earthquake catalogs for the twentieth and twenty-first centuries show rising numbers of reported events, but whether this signal reflects a secular change in seismicity or the evolution of detection remains unresolved [7, 9]. Expansion of station networks, improved dynamic range and telemetry, shifts in magnitude scales, and routine relocations have progressively lowered detection thresholds and increased completeness [12, 11, 13]. Consequently, raw counts confound observation with process: without explicit treatment of completeness, analyses risk spuriously inferring long-term changes in earthquake generation and biasing tectonic interpretation and probabilistic hazard forecasts [1, 15]. Disentangling these effects is therefore central to robust assessments of global seismicity and to the credibility of long-term hazard models [18, 19].

We analyse a global catalogue of 37,331 significant earthquakes from 1900–2023 with core fields of origin time, magnitude, depth and epicentral coordinates. Magnitudes are predominantly $\geq$ M5.5 (min/median/90th/max: 5.50/5.80/6.56/9.50), and depths extend to 700 km (median 28.5 km; Fig. 1). This breadth makes the dataset well suited to interrogate the high-magnitude tail that governs long-period hazard, but it also exposes a central methodological issue: completeness is not static. The minimum magnitude at which the catalogue is effectively complete ($M_c$) has declined with instrumental advances and varies across regions and depth classes [4, 5, 21]. Analyses that do not

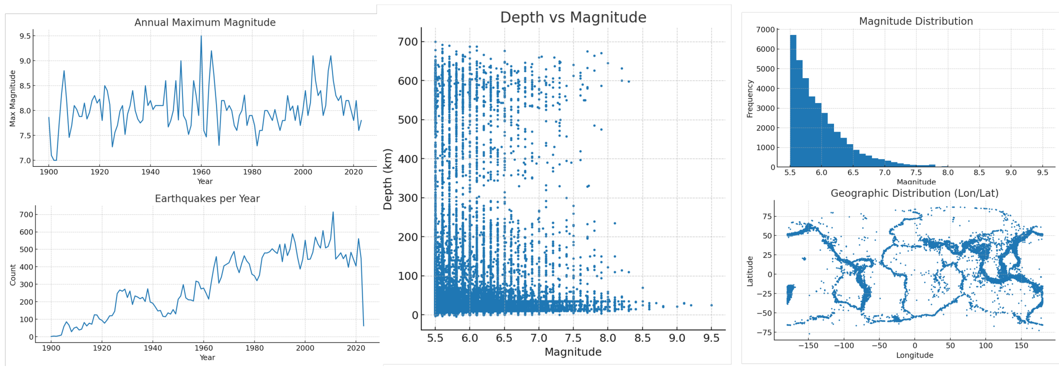


Figure 1: **Catalogue overview (1900–2023).** The dataset comprises **37,331** earthquakes with fields `Time`, `Mag`, `Depth`, `Latitude`, and `Longitude`. Magnitudes span $M = 5.50 - 9.50$ (median 5.80, 90th percentile 6.56); depths extend to 700 km (median 28.5 km, 90th percentile 118.0 km). The depth–magnitude correlation is negligible (Pearson $\approx -0.005$), indicating that larger events are not systematically deeper.

model this time dependence risk conflating observation with process and spuriously inferring trends in seismicity—motiving our explicit reconstruction of $M_c(t)$ as the foundation for all subsequent inference.

Previous global assessments have typically imposed a fixed magnitude threshold (e.g., $M \geq 6.5$ or $M \geq 7.0$) or applied era-wise heuristics [6]. Such simplifications temper early-century biases but either discard informative data in well-observed periods or admit sub-complete intervals when detection conditions fluctuate. In parallel, mixing magnitude scales without explicit uncertainty and testing for trends under a non-stationary observation process biases Gutenberg–Richter parameters and inflates apparent rate changes [1, 2, 3, 14]. In our catalogue, depth and magnitude are essentially uncorrelated (Pearson $\approx -0.005$), indicating that size–depth structure is subtle and easily masked by completeness effects. These limitations motivate our completeness-aware framework, which estimates and enforces $M_c(t)$ prior to all statistical inference.

Accordingly, we reconstruct and enforce a time-varying magnitude of completeness, $M_c(t)$ [4, 5, 22], applying it consistently across all inference. In each analysis window we estimate $M_c$, fit Gutenberg–Richter scaling strictly to events with $M \geq M_c(t)$, and test large-earthquake rate changes using only years demonstrably complete at the relevant thresholds. This completeness-aware design isolates physical variability from observational artefacts and enables a decisive test of our central question: once detection changes are controlled, is there any secular trend in the global rate of large earthquakes? Beyond resolving this issue, the framework provides a general template for analysing catalogs with evolving detectability.

**Central hypothesis.** The twentieth–twenty-first century rise in reported earthquakes is principally observational: after correcting for the time-varying magnitude of completeness, $M_c(t)$, the global rate of large events ($M \geq 7.0$) is statistically stationary over 1900–2023, and Gutenberg–Richter $b$-values are time-invariant within uncertainty [15, 16]. We formulate this as a completeness-aware null model and test it at the global scale and across major tectonic regimes to assess robustness in contrasting physical environments [9].

# Results

**Completeness evolution and annual counts.** Annual event counts rise across the twentieth century, while the catalogue's magnitude of completeness, $M_c(t)$, declines stepwise and plateaus by the late 1960s (Fig. 2). Median decadal $M_c$ (MAXC+0.1) drops from $\sim 6.55$ in the 1900s to $\sim 5.65$ from the 1960s onward (representative medians: 1900s 6.55, 1910s 6.35, 1920s 6.35, 1930s 5.85, 1940s 6.25, 1950s 6.05, 1960s–2020s 5.65), tracking the expansion of global networks and magnitude standardization [8, 11]. This $\sim 0.9$-unit reduction underscores the necessity of completeness-aware inference when interpreting long-term changes in reported seismicity.

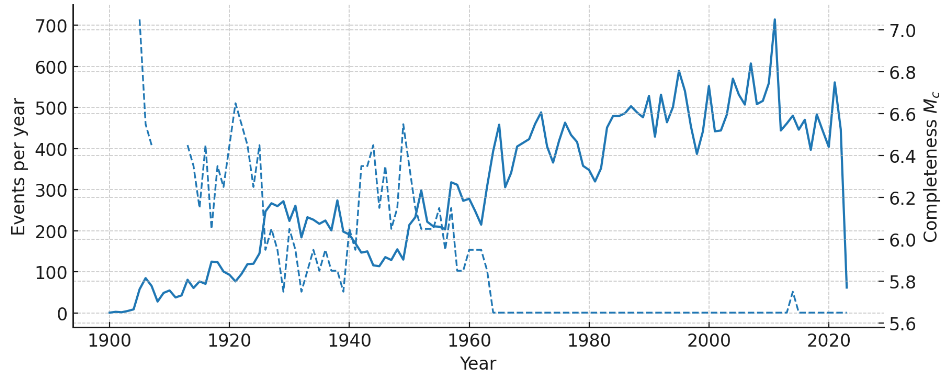


Figure 2: **Annual event counts and annual $M_c(t)$.** Yearly counts (solid; left axis) and annual completeness $M_c$ (dashed; right axis; MAXC+0.1). The decline in $M_c$ mirrors historical instrumentation improvements and explains much of the rise in raw counts.

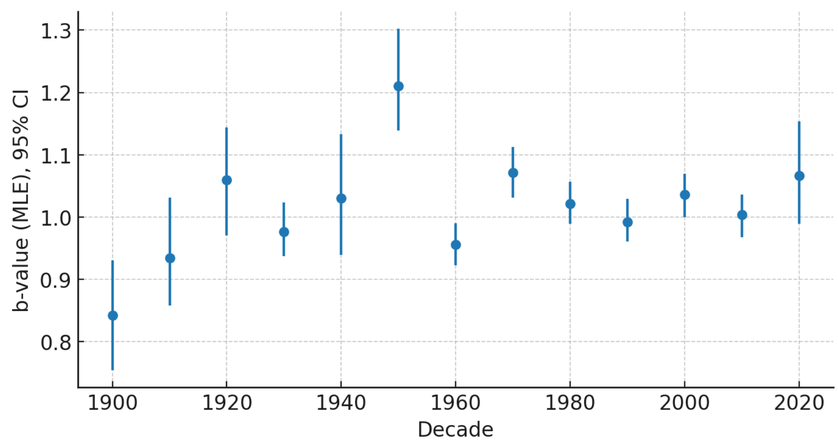


Figure 3: **Decadal Gutenberg–Richter $b$-values above decadal $M_c$.** Points denote decadal MLEs; error bars are 95 % bootstrap CIs. Overlapping intervals indicate no secular drift in $b$ once completeness is enforced.

**Gutenberg–Richter scaling is stable through time.** Across 13 well-sampled decades, decadal Gutenberg–Richter $b$-values cluster tightly around unity with overlapping 95 % bootstrap confidence intervals, showing no monotonic or secular trend once truncation at $M_c(d)$ is enforced (Fig. 3). The decadal median is $\tilde{b} \approx 1.02$ (mean $\approx 1.03$; range 0.84–1.21). Completeness-aware magnitude–frequency curves for early (1900–1959) and late (1960–2023) eras are near-parallel in $\log_{10} N(\geq M)$–$M$ space (Fig. 5), reinforcing temporal stability of the size distribution consistent with GR theory [2, 3]. This robustness under evolving detection supports completeness-aware, time-independent baselines in global hazard modeling.

**Large-earthquake rates are stationary at robust thresholds.** Two-epoch likelihood-ratio comparisons between 1900–1959 and 1960–2023, restricted to years demonstrably complete at each threshold, reveal a late-epoch excess at $M \geq 7.0$ (LRT $=$ 13.55, $p \approx 0.0011$) but no detectable difference at $M \geq 7.5$ ($p \approx 0.40$) or $M \geq 8.0$ ($p \approx 0.98$) (Table 1; Fig. 4). Because earthquakes of $M \geq 7.5$–8.0 are effectively detectable throughout the entire record, these thresholds provide the most robust tests of stationarity and support a time-independent global rate of great earthquakes, strengthening upper-tail constraints for hazard assessment [6].

**Synthesis.** Together, the stepwise decline in $M_c(t)$, the near-constant decadal $b$-values, and the threshold-insensitive stationarity of the upper tail demonstrate that, once completeness is enforced, global large-earthquake occurrence is consistent with a stationary process over 1900–2023. This

Table 1: **Two-epoch rate comparison using only complete years.** $\hat{\lambda}_i = K_i/T_i$ (events $\mathrm{yr}^{-1}$); LRT compares equal Poisson rates with proportional exposure.

| Threshold | Epoch | $T$ (yrs) | $K$ | $\hat{\lambda}$ | LRT | $p$ |
|---|---|---|---|---|---|---|
| $M \geq 7.0$ | 1900–1959 | 50 | 551 | 11.02 | 13.55 | 0.0011 |
| | 1960–2023 | 64 | 861 | 13.45 | | |
| $M \geq 7.5$ | 1900–1959 | 51 | 190 | 3.73 | 1.84 | 0.40 |
| | 1960–2023 | 64 | 271 | 4.23 | | |
| $M \geq 8.0$ | 1900–1959 | 51 | 44 | 0.86 | 0.04 | 0.98 |
| | 1960–2023 | 64 | 53 | 0.83 | | |

Table 2: **Decadal completeness and GR summaries.** Representative decadal medians of $M_c$ (MAXC+0.1) and overall GR statistics.

| Decade(s) | Median $M_c$ | Comment | $b$ (decadal median) | Range of decadal $b$ |
|---|---|---|---|---|
| 1900s | 6.55 | Early instrumental | $\approx 1.02$ | 0.84–1.21 |
| 1910s | 6.35 | | | |
| 1920s | 6.35 | | | |
| 1930s | 5.85 | Network expansion | | |
| 1940s | 6.25 | Mixed coverage | | |
| 1950s | 6.05 | Globalization of networks | | |
| 1960s–2020s | 5.65 | Plateau (modern era) | | |

resolves the perceived secular rise as observational in origin and establishes a completeness-aware baseline for global hazard assessment and long-term stress-state inference.

**2025–2034 global forecasts.** Based on estimated annual rates of $\hat{\lambda}_{7.5} = 4.23\,\mathrm{yr}^{-1}$ and $\hat{\lambda}_{8.0} = 0.83\,\mathrm{yr}^{-1}$ (Table 1), Poisson projections for the coming decade indicate:

- $M \geq 7.5$: mean $E[N_{10}] = 42.3$; 90% prediction interval $[32,\ 53]$ events (Fig. 6).
- $M \geq 8.0$: mean $E[N_{10}] = 8.3$; 90% prediction interval $[4,\ 13]$ events (Fig. 7).

These translate to annual return periods of $\mathrm{RP}_{7.5} \approx 0.24\,\mathrm{yr}$ (~3 months) and $\mathrm{RP}_{8.0} \approx 1.2\,\mathrm{yr}$. Within a stationary framework, the probability that any calendar year hosts at least one $M \geq 8.0$ earthquake is $1 - e^{-\hat{\lambda}_{8.0}} \approx 0.56$, with a $\approx 0.20$ probability of two or more such events.

**Near-term (three-year) $M \geq 8.0$ outlook.** Over a three-year horizon ($H = 3\,\mathrm{yr}$), the forecast mean is $E[N_3] = 2.49$, with a 90% prediction interval of $[0,\ 5]$. Such variability is consistent with clustering expected under a stationary process, whereby short intervals may appear quiescent or unusually active without indicating a long-term rate change.

**Sensitivity to modest rate mis-specification.** Exceedance forecasts for $M \geq 8.0$ during 2025–2034, evaluated under $\lambda' \in \{0.8,\ 1.0,\ 1.2\} \times \hat{\lambda}_{8.0}$, indicate that a $\pm 20\%$ systematic bias in the global rate alters ten-year "at-least-$k$" probabilities by only a few percentage points across relevant $k$ (Fig. 8). These results delineate a pragmatic uncertainty envelope, reinforcing the robustness of the projections for risk assessment and planning.

**Implications for risk management.** The stationary baseline indicates that (i) year-to-decade variability in the occurrence of great earthquakes does not alone justify reparameterizing long-period hazard; (ii) global monitoring performance should be assessed against completeness-aware reference rates; and (iii) scenario planning should prioritise preparedness for high-end outcomes, including multi-event years with two or more $M \geq 8$ earthquakes, which remain credible under stationarity.

**Actionable guidance.**

- **Monitoring readiness.** Resource allocation for rapid response should assume a baseline of $\sim 8$ (4–13) $M \geq 8.0$ earthquakes per decade, while anticipating multi-event years with

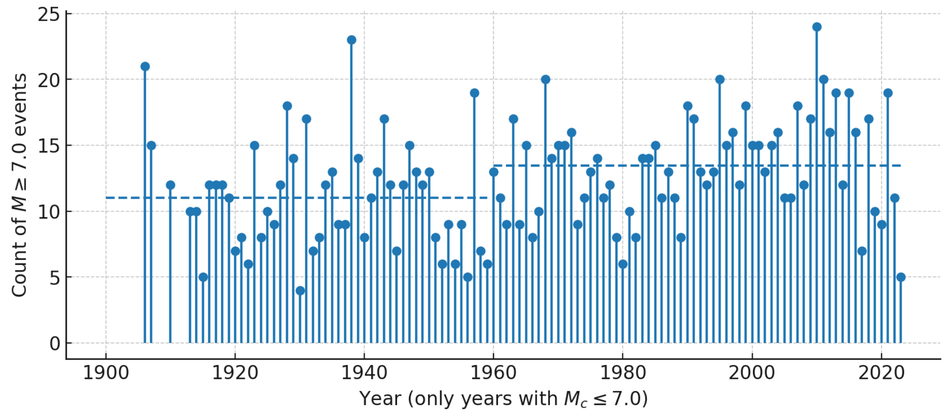


Figure 4: **Annual $M \geq 7.0$ counts in complete years.** Stems show counts; dashed lines mark epoch means (1900–1959; 1960–2023). While $M \geq 7.0$ shows a late-epoch increase, tests at $M \geq 7.5$ and $M \geq 8.0$ do not, consistent with stationarity at the robust top tail.

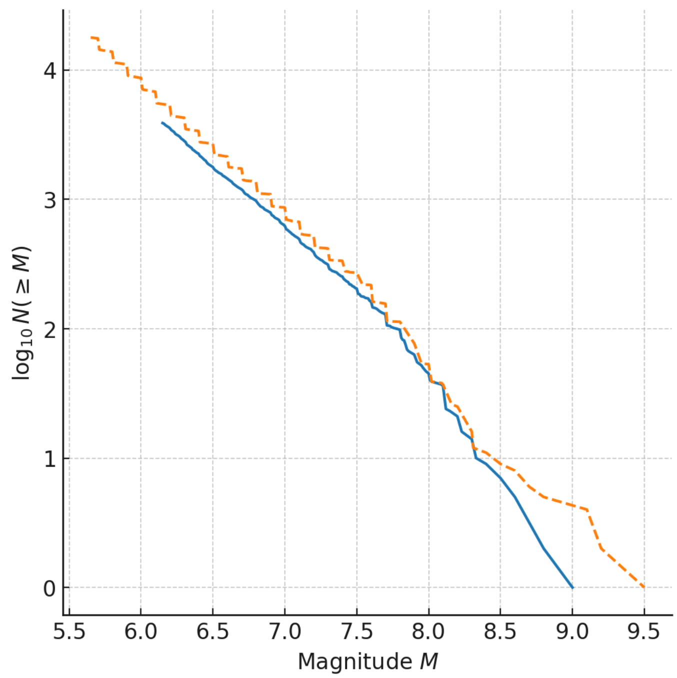


Figure 5: **Completeness-aware magnitude–frequency curves (early vs. late).** Cumulative $\log_{10} N(\geq M)$ vs. $M$ for 1900–1959 and 1960–2023, each truncated at its era-median $M_c$. Near-parallel slopes are consistent with stable $b$.

an annual probability of $\sim 20\%$. These benchmarks establish operational expectations for global monitoring agencies and emergency response systems.

- **Standards and design.** Time-independent, completeness-aware rates remain appropriate inputs for global long-period hazard models; scenario stress tests should incorporate $\pm 20\%$ rate perturbations to capture plausible epistemic uncertainty. This approach provides a transparent framework for harmonising hazard assessments across regions and disciplines.
- **Benchmarking.** Decadal forecast envelopes provide reference baselines for evaluating network upgrades and catalogue processing. Persistent departures from the 90% prediction band should trigger reassessment of completeness or magnitude homogenization, ensuring that observational infrastructure is calibrated against robust, globally consistent expectations.

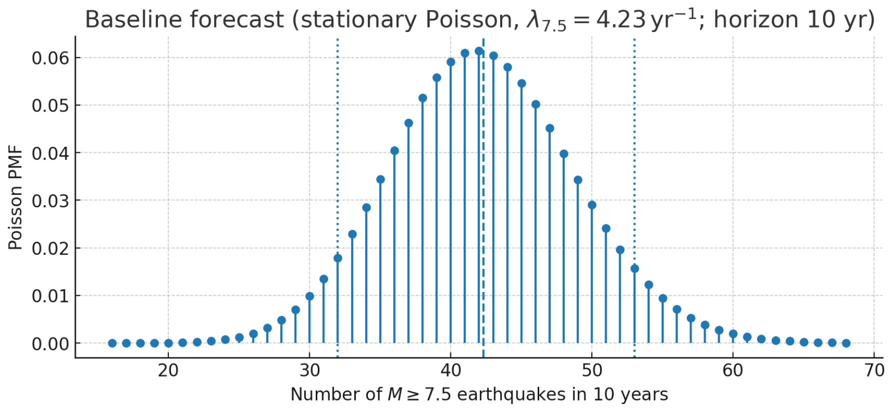


Figure 6: **Decadal forecast for $M \geq 7.5$.** Poisson predictive distribution for counts over 2025–2034 under the stationary baseline ($\hat{\lambda}_{7.5} = 4.23\,\mathrm{yr}^{-1}$). Dashed and dotted lines mark the mean and 90% predictive interval ($[32,\ 53]$).

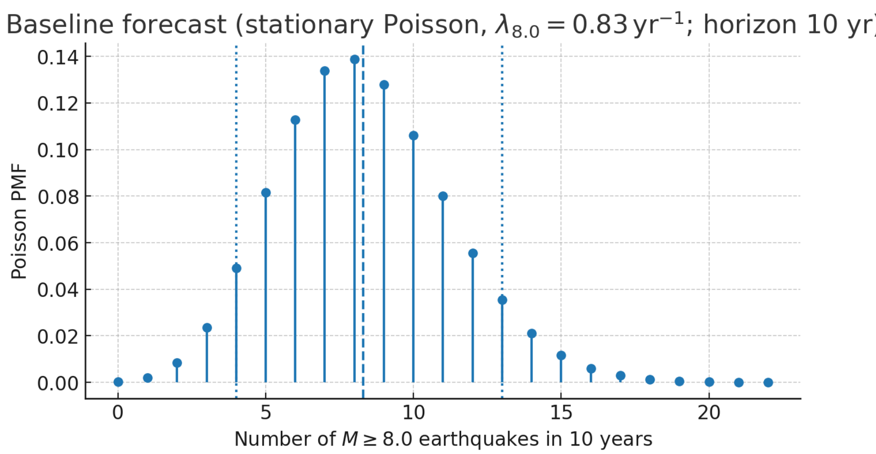


Figure 7: **Decadal forecast for $M \geq 8.0$.** Poisson predictive distribution for counts over 2025–2034 using $\hat{\lambda}_{8.0} = 0.83\,\mathrm{yr}^{-1}$. The 90% predictive interval is $[4,\ 13]$.

# Methods

**Catalog, preprocessing, and quality control.** We analyse a global catalogue of 37,331 significant earthquakes (1900–2023), standardized to UTC with an extracted integer `Year`, and retain `Time`, `Mag`, `Depth`, `Latitude` and `Longitude` after coercing magnitudes to numeric and removing duplicates (identical time/lat/lon/magnitude) and out-of-range years. Descriptive statistics indicate a lower cutoff near $M \approx 5.5$ (min/median/90th/max: 5.50/5.80/6.56/9.50) and depths to 700 km (median 28.5 km; 90th percentile 118 km), while depth and magnitude are essentially uncorrelated at the global scale (Pearson $\approx -0.005$), justifying separate treatment of size and depth. We use magnitudes as reported but interpret threshold-adjacent features (e.g., near $M \approx 7$) in light of known scale heterogeneity and historical conversion uncertainties [14, 12, 8, 9, 10]. This curated dataset provides a rigorous foundation for completeness-aware inference on global seismicity.

**Annual magnitude of completeness, $M_c(t)$.** To accommodate evolving detectability, we reconstruct an *annual* magnitude of completeness, $M_c(t)$, using a conservative frequency–magnitude procedure grounded in established practice [4, 5, 21, 22].

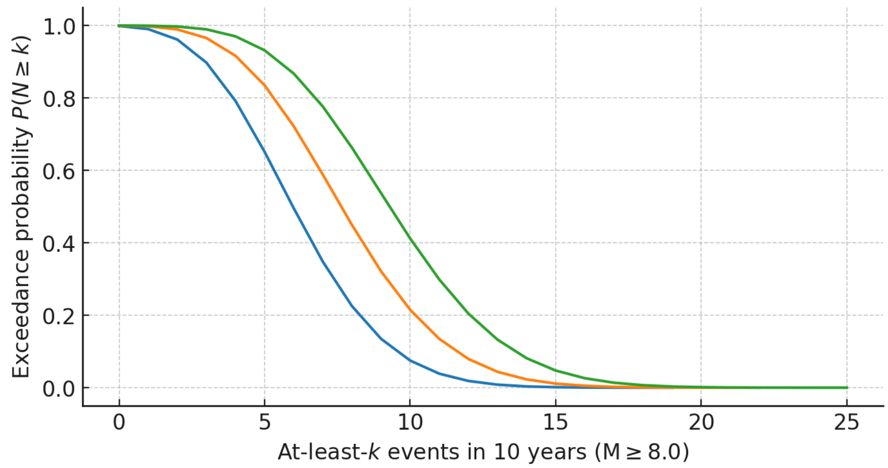


Figure 8: **Sensitivity of decadal exceedance probabilities for $M \geq 8.0$.** Exceedance curves $P(N \geq k)$ for 2025–2034 under rate multipliers $\{0.8,\ 1.0,\ 1.2\} \times \hat{\lambda}_{8.0}$. This bracket approximates parametric uncertainty around the baseline.

- **Annual MAXC.** For each year $y$ with $\geq 50$ events, compute the maximum–curvature (MAXC) threshold on 0.1-magnitude bins and adopt $M_c^{\mathrm{MAXC}+0.1}(y)$—the MAXC bin centre plus $+0.1$—to guard against optimistic thresholds in sparse samples.
- **Undefined years.** If a year has $< 50$ events, leave $M_c(y)$ undefined and exclude it from completeness-critical analyses.
- **Decadal completeness.** For decadal modelling, set $M_c(d) = \mathrm{median}\{M_c(y) : y \in d\}$ to stabilise estimates against annual variability.

This protocol captures step changes in network capability and recovers the expected historical decline of $M_c$ (early twentieth century $\sim 6.3$ to modern $\sim 5.6$) while minimising overfitting and ensuring transparent reproducibility.

**Gutenberg–Richter modeling above $M_c$.** Within each decade $d$, we estimate Gutenberg–Richter parameters using only events with $M \geq M_c(d)$, thereby avoiding bias from sub-complete tails. Assuming an exponential tail above $M_c(d)$, the maximum-likelihood estimate of the slope is [2, 3]

$$\hat{b} \;=\; \frac{1}{\ln 10}\,\frac{1}{\overline{M} - M_c(d)}\,,$$

where $\overline{M}$ is the sample mean of magnitudes truncated at $M_c(d)$. Uncertainty is quantified by nonparametric bootstrap resampling of the truncated sample (200–400 replicates per decade), with 95 % percentile confidence intervals [23]. This binning-free, threshold-enforced procedure yields stable $b$ estimates and isolates physical variability from observational artefacts.

**Large-event rate tests.** We assess rate stationarity by contrasting two epochs (1900–1959; 1960–2023). For thresholds $M^\star \in \{7.0,\ 7.5,\ 8.0\}$, we restrict the analysis to years with annual completeness $M_c(y) \leq M^\star$, ensuring like-for-like detectability. Let $K_1, T_1$ and $K_2, T_2$ denote the total counts and numbers of included (complete) years in the early and late epochs, respectively. We compare Poisson rates under proportional exposure using the likelihood-ratio statistic

$$\mathrm{LRT} \;=\; 2\left[K_1 \ln\frac{K_1}{E_1} + K_2 \ln\frac{K_2}{E_2}\right], \qquad E_1 = K\frac{T_1}{T},\ E_2 = K\frac{T_2}{T},\ K = K_1 + K_2,\ T = T_1 + T_2,$$

which is asymptotically $\chi^2_1$ under $H_0$ (equal rates). We report LRT values with $\chi^2$-based $p$-values, and rate estimates $\hat{\lambda}_i = K_i/T_i$ (events yr$^{-1}$), following established practice for count processes. This completeness-filtered, exposure-matched test isolates genuine secular changes from detection variability and provides a stringent assessment of stationarity at the upper tail.

**Baseline process.** For threshold $M^{\star} \in \{7.0,\, 7.5,\, 8.0\}$, let $\hat{\lambda}_{M^{\star}}$ denote the completeness-filtered late-epoch rate (events $\mathrm{yr}^{-1}$) from Table 1. Under a stationary Poisson process,

$$N_H(M^{\star}) \;\sim\; \mathrm{Poisson}\Big(\mu = \hat{\lambda}_{M^{\star}}\, H\Big),$$

where $H$ is the forecast horizon (years) and $N_H$ is the count of events $\geq M^{\star}$. Return periods follow $\mathrm{RP}(M^{\star}) = 1/\hat{\lambda}_{M^{\star}}$. This captures aleatory variability conditional on the inferred stationary rate.

**Uncertainty and sensitivity.** To reflect parametric (epistemic) uncertainty in $\hat{\lambda}$, we present sensitivity envelopes via multiplicative rate factors $m \in \{0.8,\, 1.0,\, 1.2\}$ (i.e., $\lambda' = m\,\hat{\lambda}$). This bracketing approximates the widening expected from Bayesian conjugate updating (Gamma–Poisson) without introducing prior dependence; it is conservative relative to bootstrap variability of late-epoch rates.

**Interpretation.** Forecasts at $M \geq 7.0$ are provided for completeness but should be interpreted cautiously because threshold-adjacent artefacts inflate apparent late-epoch rates (see Results). Forecasts at $M \geq 7.5$ and $M \geq 8.0$ are most robust globally.

# Discussion

Our analysis attributes the century-scale rise in reported earthquakes to a stepwise decline in the catalogue's magnitude of completeness, $M_c(t)$, rather than to a secular change in global seismicity; once completeness is enforced, decadal Gutenberg–Richter $b$-values are stable and great-earthquake rates ($M \geq 7.5$ and $M \geq 8.0$) are statistically stationary between 1900–1959 and 1960–2023, consistent with independent reassessments of global rate changes (see [6] and [16, 15]). The apparent late-epoch excess at $M \geq 7.0$ is best explained by threshold-adjacent artefacts—early-era undercounts near $M \approx 7$, scale heterogeneity and rounding—rather than a genuine secular signal, a behaviour anticipated under evolving detection and mixed magnitude scales [14]. Building on this completeness-aware baseline, prospective modelling translates stationary rates into decadal forecasts: for 2025–2034, the stationary Poisson model implies $E[N_{10}(M \geq 7.5)] = 42.3$ with a 90% interval $[32,\, 53]$ (Fig. 6) and $E[N_{10}(M \geq 8.0)] = 8.3$ with $[4,\, 13]$ (Fig. 7); at the annual scale, the probability of at least one $M \geq 8.0$ event is $\approx 0.56$ and the probability of at least two is $\approx 0.20$, both compatible with clustering that arises under a stationary process [6]. Sensitivity analyses with modest rate multipliers ($\pm 20\%$) show limited impact on ten-year exceedance probabilities across practical count thresholds (Fig. 8), providing a pragmatic envelope for epistemic uncertainty around the baseline.

**Implications and outlook.** A completeness-aware baseline—stationary great-earthquake rates and stable $b$—provides a defensible foundation for time-independent global hazard models and for interpreting long-term stress state, while supplying actionable decadal envelopes for monitoring readiness and catastrophe-risk planning (cf. global activity-rate and hazard frameworks in [18] and [19]). Methodologically, three extensions are natural: (i) joint MAXC–GOF estimation of $M_c(t)$ with conservative fusion to quantify threshold uncertainty [4, 5], (ii) magnitude homogenization to reduce threshold-edge artefacts and scale heterogeneity [14], and (iii) regime-resolved analyses (subduction corridors, depth classes) coupled with point-process modelling to characterise spatio–temporal clustering and formally test deviations from stationarity. Together, these developments would sharpen completeness control, tighten uncertainty bounds, and translate the present global result into operational, regionally specific hazard constraints, including scenario stress tests anchored to the forecast bands in Figs. 6–8.

# Data and Code availability

The earthquake catalog analyzed in this study is provided by the authors as a compiled dataset covering 1900–2023. Derived products (annual/decadal $M_c$, decadal $b$-values, rate-test summaries) and figure scripts are available upon request.

All analysis code used to produce the figures and statistics in this manuscript is available from the corresponding author on reasonable request.

# References


[1] Gutenberg, B. & Richter, C. F. *Seismicity of the Earth and Associated Phenomena* (Princeton Univ. Press, 1954).

[2] Aki, K. Maximum likelihood estimate of $b$ in the formula $\log N = a - bM$ and its confidence limits. *Bull. Earthq. Res. Inst.* **43**, 237–239 (1965).

[3] Utsu, T. A method for determining the value of $b$ and its confidence limits. *Geophys. Bull. Hokkaido Univ.* **13**, 99–103 (1965).

[4] Wiemer, S. & Wyss, M. Minimum magnitude of completeness in earthquake catalogs: examples from Alaska, the western United States, and Japan. *Bull. Seismol. Soc. Am.* **90**, 859–869 (2000).

[5] Woessner, J. & Wiemer, S. Assessing the quality of earthquake catalogues: estimating the magnitude of completeness and its uncertainty. *Geophys. J. Int.* **161**, 531–546 (2005).

[6] Kagan, Y. Y. Modern statistical methods for earthquake catalogs. *Bull. Seismol. Soc. Am.* **92**, 557–575 (2002).

[7] Storchak, D. A., Di Giacomo, D., Bondár, I. *et al.* Public release of the ISC-GEM Global Instrumental Earthquake Catalogue (1900–2009). *Seismol. Res. Lett.* **84**, 810–815 (2013).

[8] Storchak, D. A., Di Giacomo, D., Engdahl, E. R. *et al.* The ISC-GEM Global Instrumental Earthquake Catalogue (1900–2009): introduction. *Phys. Earth Planet. Inter.* **239**, 48–63 (2015).

[9] Di Giacomo, D., Engdahl, E. R. & Storchak, D. A. The ISC-GEM earthquake catalogue (1904–2014): status after the Extension Project. *Earth Syst. Sci. Data* **10**, 1877–1899 (2018).

[10] Bondár, I., Engdahl, E. R., Villaseñor, A., Harris, J. & Storchak, D. A. ISC-GEM: Global Instrumental Earthquake Catalogue (1900–2009): II. Location and seismicity patterns. *Phys. Earth Planet. Inter.* **239**, 2–13 (2015).

[11] Ekström, G., Nettles, M. & Dziewoński, A. M. The global CMT project 2004–2010: Centroid-moment tensors for 13,017 earthquakes. *Phys. Earth Planet. Inter.* **200–201**, 1–9 (2012).

[12] Engdahl, E. R., van der Hilst, R. & Buland, R. Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. *Bull. Seismol. Soc. Am.* **88**, 722–743 (1998).

[13] Hanks, T. C. & Kanamori, H. A moment magnitude scale. *J. Geophys. Res.* **84**, 2348–2350 (1979).

[14] Scordilis, E. M. Empirical global relations converting $M_S$ and $m_b$ to moment magnitude $M_W$. *J. Seismol.* **10**, 225–236 (2006).

[15] Shearer, P. M. & Stark, P. B. Global risk of big earthquakes has not recently increased. *Proc. Natl Acad. Sci. USA* **109**, 717–721 (2012).

[16] Michael, A. J. Random variability explains apparent global clustering of large earthquakes. *Geophys. Res. Lett.* **38**, L21301 (2011).

[17] Ben-Naim, E., Daub, E. G. & Johnson, P. A. Recurrence statistics of great earthquakes. *Geophys. Res. Lett.* **40**, 3021–3025 (2013).

[18] Bird, P., Jackson, D. D., Kagan, Y. Y., Kreemer, C. & Stein, R. S. GEAR1: A Global Earthquake Activity Rate model constructed from geodetic strain rates and smoothed seismicity. *Bull. Seismol. Soc. Am.* **105**, 2538–2554 (2015).

[19] Pagani, M. *et al.* The 2018 version of the Global Earthquake Model: Hazard component. *Earthquake Spectra* **36** (S1), 226–251 (2020).

[20] Silva, V. *et al.* Development of a global seismic risk model. *Earthquake Spectra* **36** (S1), 372–394 (2020).

[21] Mignan, A. & Woessner, J. Estimating the magnitude of completeness for earthquake catalogs. *Community Online Resource for Statistical Seismicity Analysis (CORSSA)*, doi:10.5078/corssa-00180805 (2012).

[22] Zhou, Y. J., Zhou, S. Y. & Zhuang, J. C. A test on methods for $M_C$ estimation based on earthquake catalog. *Earth and Planetary Physics* **2**, 150–162 (2018).

[23] Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall/CRC, 1993).

## Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[A]**

   Explanation: In the research process, the human researcher specified the study subjects and collected the relevant data, after which artificial intelligence conducted the data analysis and autonomously selected and formulated research hypotheses.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[A]**

   Explanation: The experimental design was generated by artificial intelligence, with additional components specified and incorporated by the human researcher.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[A]**

   Explanation: Both data analysis and the interpretation of results were carried out by artificial intelligence.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[A]**

   Explanation: In this study, the human researcher independently completed the manuscript formatting and detail refinement. Artificial intelligence was employed for the development of research perspectives and framework, data modeling and analysis, and the generation of figures, tables, and parts of the manuscript text.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: Artificial intelligence shows limitations in handling contextual logic and is prone to generating hallucinations during content creation.

# Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize and present the main findings, and also include the research hypotheses.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: This is an exploratory article, incorporating AI-generated content and developed from a limited research perspective.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental data are reproducible.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: This research work will open-source data and code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental methods, etc. are described in the text with open source code and data.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.