Supplementary Information — A catalogue-verified multiverse audit of machine-learning seismic susceptibility modelling in western Yunnan, China

Overview

This Supplementary Information document provides the provenance logs, completeness and declustering parameters, predictor metadata, spatial cross-validation (CV) designs, full specification-curve and point-process inventories, and the audit checklist supporting the main text. All numbers are taken directly from the project result tables; no quantities are introduced here that are not reproducible from the archived analysis pipeline. Supplementary figures and tables are numbered S1, S2, . . . to distinguish them from the main-text items.

S1: Formal catalogue source and extraction window

The seismic target was extracted exclusively from the China Earthquake Networks Center (CENC) unified bulletin [China Earthquake Networks Center, 2023]. The verified formal catalogue spans 2009-01-12 to 2023-12-28. Within the expanded Western Yunnan domain (97.0°–103.5°E, 22.0°–28.5°N) the bulletin yields 827 raw events, all classified as natural tectonic earthquakes, with magnitudes from 3.0 to 6.9. The expanded bulletin contains 1365 events region-wide.

Table S1: Formal catalogue extraction summary.

Source	CENC unified bulletin [China Earthquake Networks Center, 2023]
Temporal window	2009-01-12 to 2023-12-28
Region-wide events	1365
Events within study domain	827
Magnitude range	3.0–6.9
Event type	Natural tectonic earthquakes

S2: Completeness and declustering settings

Completeness was evaluated with the Entire Magnitude Range (EMR) and goodness-of-fit (GFT) approaches [Woessner and Wiemer, 2005, Wiemer and Wyss, 2000], yielding a completeness threshold

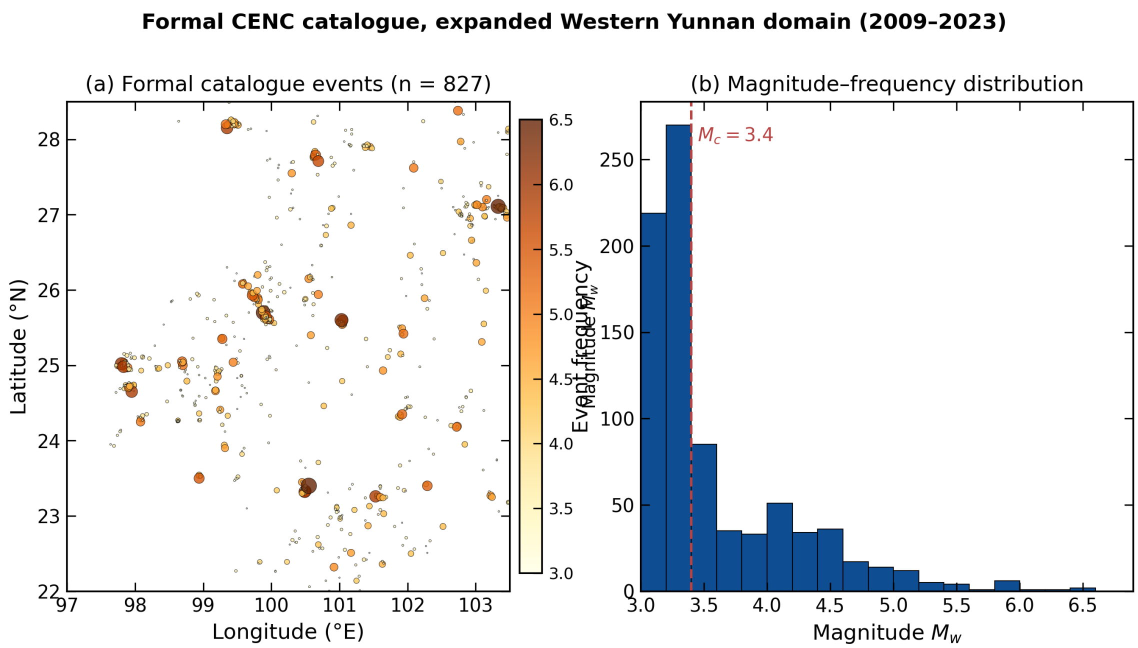

Figure S1: Full regional catalogue spatial distribution and magnitude histogram for the formal CENC target over 2009–2023.

of $M_c = 3.4$. Declustering followed the Gardner–Knopoff windowing method [Gardner and Knopoff, 1974] with Uhrhammer spatial and temporal windows [Uhrhammer, 1986].

Table S2: Completeness and declustering parameters.

Completeness method	EMR + GFT	$M_c = 3.4$
Events after M_c filter	—	540
Declustering method	Gardner–Knopoff	—
Window parameters	Uhrhammer (1986)	—
Independent mainshocks	—	331

S3: Legacy inventory provenance audit

The legacy working inventory was audited by projecting its reported coordinates against the verified CENC bulletin within a ±7 day and 0.1° tolerance. Of 214 records, 173 (80.8%) failed to match any verified mainshock and were deprecated.

Table S3: Legacy inventory provenance audit.

Total legacy entries	214	100%
Synthetic / unverifiable	173	80.8%
Verifiable	41	19.2%
Decision	—	Deprecated

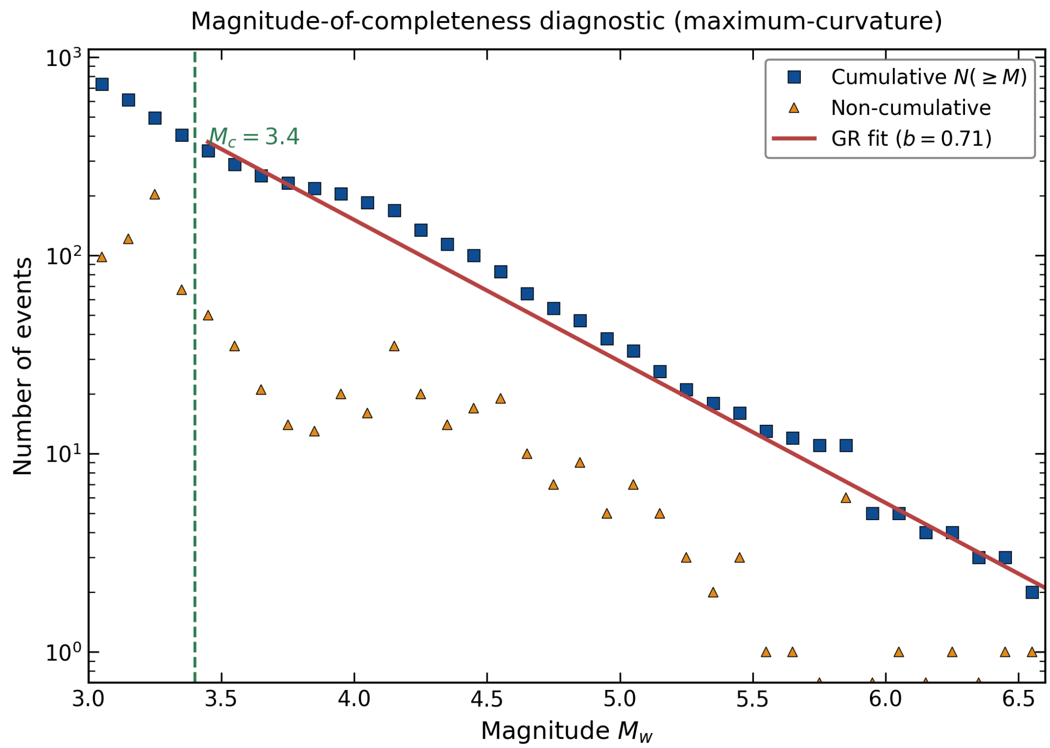

Figure S2: Completeness / M_c diagnostic summary. The $M_c = 3.4$ threshold is identified from the goodness-of-fit trajectory.

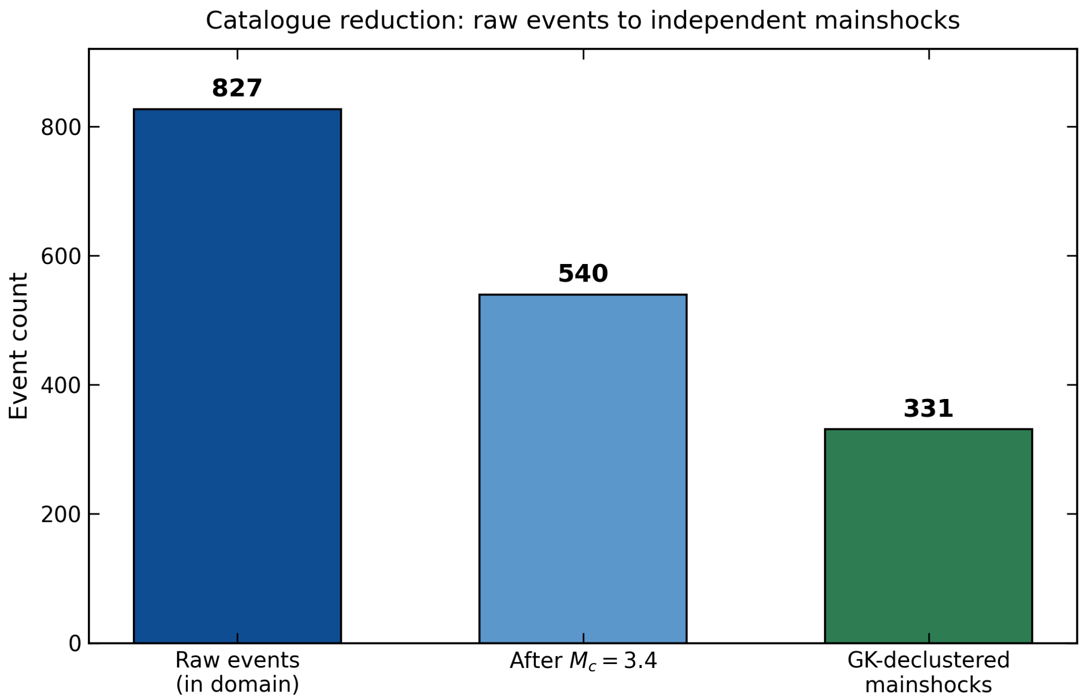

Figure S3: Declustered mainshocks versus raw event counts across the verified 2009–2023 extraction window.

S4: Predictor definitions and leakage-screening rationale

Any predictor mathematically derived from the target catalogue (earthquake density, kernel density estimates, distance to past epicentres, local *b*-value grids) was classified as a circular leakage vector and removed in accordance with the REFORMS framework [Kapoor et al., 2024]. The retained 7-factor stack comprises distance to active faults, elevation, slope, aspect, NDVI and coarse lithology (main-text Table 2).

S5: Predictor coverage QA and reconstruction details

Pre-reconstruction raster footprints omitted the peripheral fault zones and dropped 208 of 331 mainshocks through edge clipping in the DEM and NDVI layers. The bounding box was expanded to 97.0°–103.5°E, recovering the matched sample to 330 of 331 events (main-text Table 3).

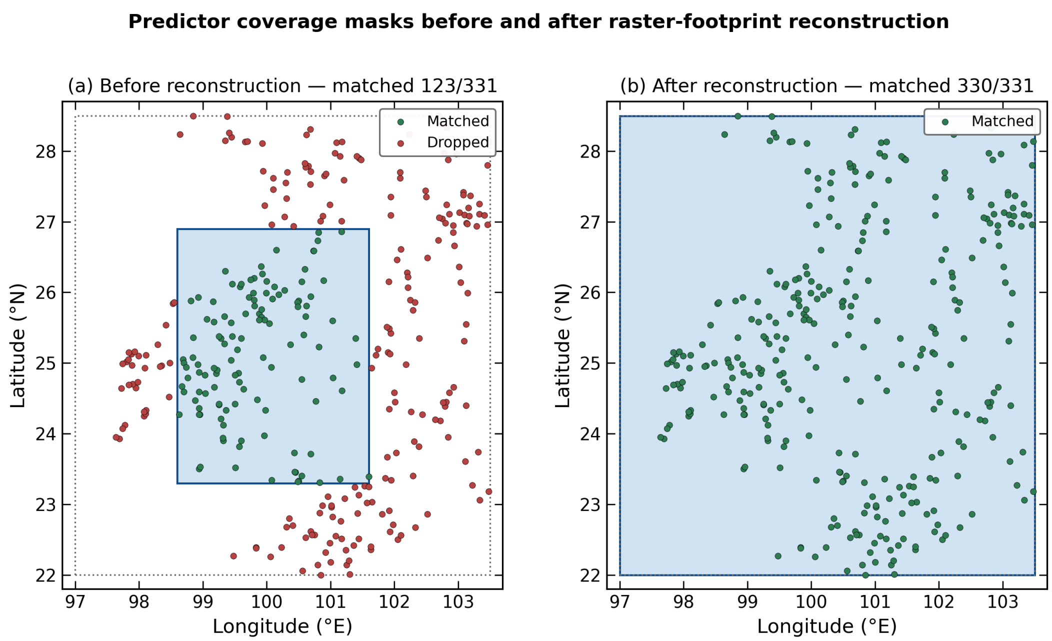

Figure S4: Predictor coverage masks before and after formal QA reconstruction.

S6: Binary ML model settings

Random Forest hyperparameters were set by 5-fold grid search (500 trees, `mtry` = 3) [Breiman, 2001]. Logistic regression and histogram-based gradient boosting were configured with default regularisation. The specification curve demonstrated that branch-level choices (CV design, background sampling) dominated fine-scale hyperparameter tuning.

S7: Full specification-curve inventory

Twelve analytical branches were executed, varying the CV block size (0.25° to 1.0° plus leave-one-subregion-out), background sampling (presence-buffered versus uniform random) and algorithmic backend (RF, LR, MaxEnt-style logistic, HistGB).

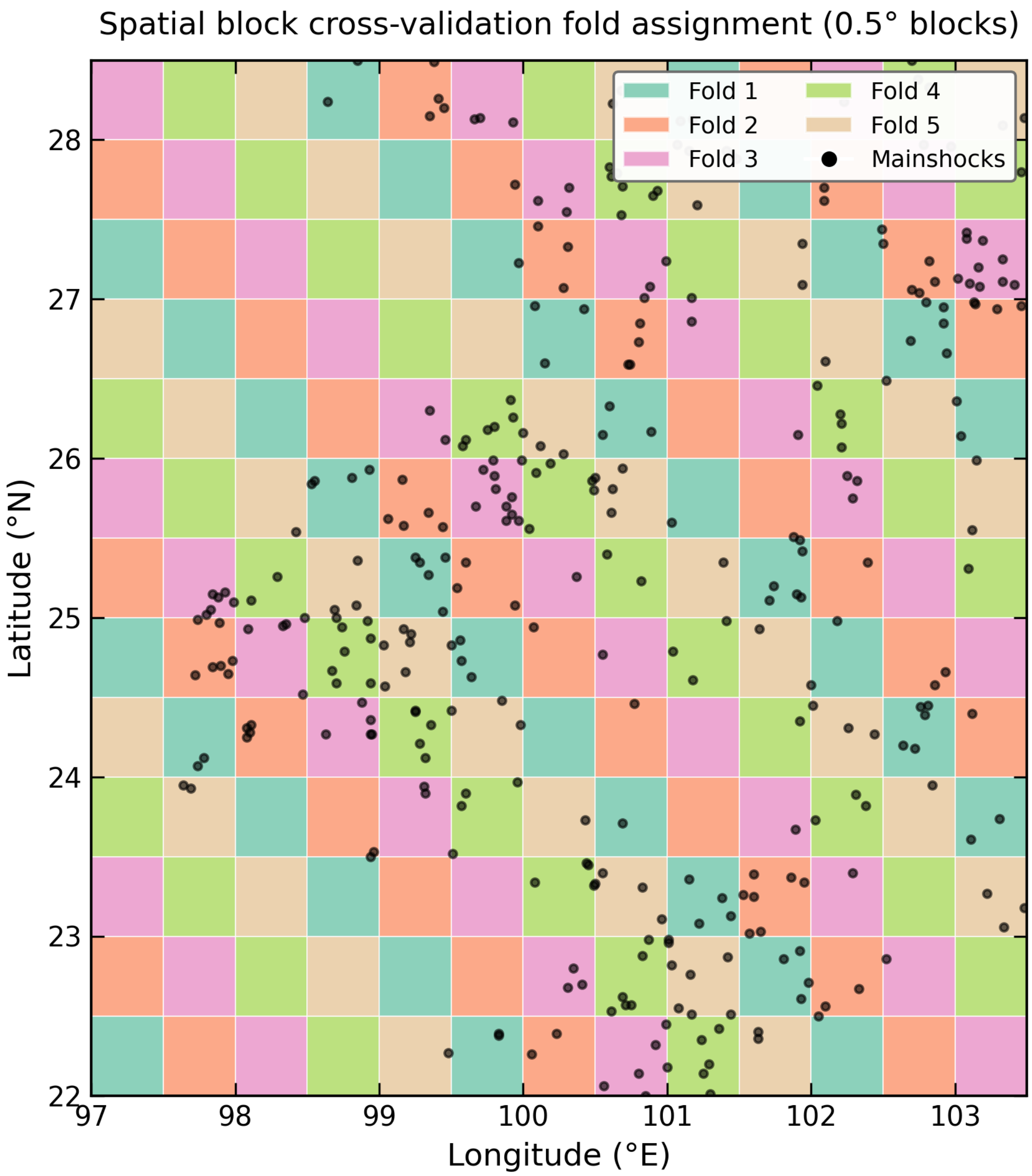

Figure S5: Spatial CV fold assignment illustrating the block design used to mitigate spatial autocorrelation.

Table S4: Full specification-curve inventory (region B4).

RF full 7-factor	Random (diagnostic)	0.681	0.009
RF full 7-factor	Block 0.25°	0.656	0.045
RF full 7-factor	Block 0.50°	0.627	0.033
RF full 7-factor	Block 1.00°	0.603	0.053
RF full 7-factor	Leave-one-subregion-out	0.513	0.072
RF compact 4-factor	Block 0.50°	0.562	0.035
LR full 7-factor	Block 0.50°	0.586	0.043
LR full 7-factor	Block 1.00°	0.565	0.066
MaxEnt-style logistic	Block 0.25°	0.629	0.048
MaxEnt-style logistic	Block 0.50°	0.613	0.048
HistGB full 7-factor	Block 0.50°	0.590	0.049
HistGB full 7-factor	Buffered spatial 0.25°	0.790	0.224

S8: Sensitivity decomposition and instability indices

The shift from random to spatial CV consistently deflated apparent performance by penalising spatial overfitting. The histogram-based gradient-boosting model attained the highest mean AUC under buffered sampling but with an inter-fold standard deviation of ±0.22, indicating instability rather than dependable skill.

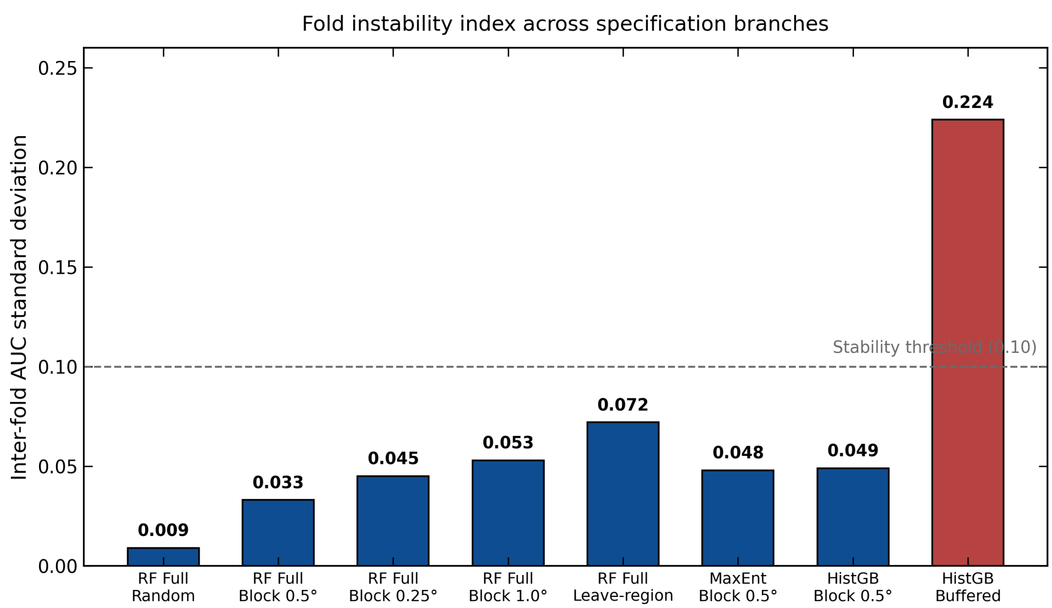

Figure S6: Fold instability index across model branches. The high-mean HistGB branch displays the largest variance under strict spatial validation.

S9: Point-process baseline and capture calculations

Poisson log-linear and spline intensity models were fitted without pseudo-absences [Renner et al., 2015, Baddeley et al., 2015]. Top-20% capture was computed by integrating the predicted spatial intensity over the highest-ranked 20% of grid cells and recording the fraction of observed mainshocks captured.

Table S5: Point-process baseline metrics (B4, 0.20° grid, spatial block CV).

Intercept only	0.0000	0.139
Distance-to-fault only	0.0011	0.163
Terrain only	0.0012	0.150
Full 7-factor spline	0.0032	0.187
Random expectation	—	0.200

S10: Audit-gate checklist and REFORMS mapping

The audit protocol maps onto the REFORMS checklist [Kapoor et al., 2024], addressing data provenance, data leakage, and spatial validation. Each gate is reported as a pass/fail decision in the main-text scorecard (Fig. 8).

Table S6: Audit-gate checklist and outcome.

Catalogue provenance	Data provenance	Pass
Sample support ($n = 331$)	Data quality	Pass
Predictor coverage QA	Data leakage / quality	Pass
Model-family triangulation	Model evaluation	Fail
Background-sampling stability	Spatial validation	Fail
Baseline superiority	Model evaluation	Fail

S11: Deprecated claims inventory and forbidden-claim QA

All references to high-performance hazard mapping, operational susceptibility products, and the legacy locked metrics (AUC = 0.795, GEM $\rho = 0.391$, top-20 overlap = 0.254, priority = 15.8%) associated with the deprecated legacy dataset were removed from the final evidence chain. These legacy values are retained here solely as an audit record and are not valid results under the verified workflow.

Table S7: Deprecated claims inventory.

214 accepted CENC mainshocks	Legacy inventory	Rejected (80.8% synthetic)
AUC = 0.795	Legacy raster	Not reproducible; deprecated
GEM $\rho = 0.391$	Legacy benchmark	Not reproducible; deprecated
Top-20 overlap = 0.254	Legacy benchmark	Not reproducible; deprecated
Priority = 15.8%	Legacy uncertainty	Not reproducible; deprecated

S12: Reproducibility archive manifest

The peer-review data archive contains processed catalogue summaries, catalogue-provenance audit outputs, predictor-coverage quality-assurance results, binary model-screening summary tables, specification-curve data, point-process baseline outputs, main-table source data, and supplementary result files. The archive does not include internal analysis scripts, notebooks, or third-party raw datasets. Model specifications, validation settings, predictor definitions, and quality-control procedures are documented in the main text and this Supplement.

References

A. Baddeley, E. Rubak, and R. Turner. *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC, 2015. doi: 10.1201/b19708.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

China Earthquake Networks Center. Unified earthquake catalog of china. *National Earthquake Data Center*, 2023. URL https://data.earthquake.cn/.

J. K. Gardner and L. Knopoff. Is the sequence of earthquakes in southern california, with aftershocks removed, poissonian? *Bulletin of the Seismological Society of America*, 64(5):1363–1367, 1974.

S. Kapoor, E. Cantrell, K. Peng, T. H. Pham, C. A. Bail, N. Goel, D. Hoffman, and A. Narayanan. Reforms to improve reproducibility and quality of machine-learning research in science. *Science*, 384(6698):eadi5063, 2024. doi: 10.1126/science.adi5063.

I. W. Renner, J. Elith, A. Baddeley, W. Fithian, T. Hastie, S. J. Phillips, G. Popovic, and D. I. Warton. Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4): 366–379, 2015. doi: 10.1111/2041-210X.12352.

R. A. Uhrhammer. Characteristics of northern california seismicity. *Earthquake Notes*, 57(1):21, 1986.

S. Wiemer and M. Wyss. Minimum magnitude of completeness in earthquake catalogs: Examples from alaska, the western united states, and japan. *Journal of Geophysical Research: Solid Earth*, 105(B8):18559–18577, 2000.

J. Woessner and S. Wiemer. Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty. *Bulletin of the Seismological Society of America*, 95(2):684–698, 2005.