Supplementary information

Hemispheric contrasts in recorded volcanic explosivity are sensitive to temporal adjustment and missing VEI

Bulletin of Volcanology

Rongzhen Dai
Independent Researcher, Yangzhou, Jiangsu, China
Correspondence: dairongzhen3@gmail.com

Supplementary methods and aggregate results for the manuscript. All model results below come from the executed reanalysis. The two derived analysis tables are supplied separately as Supplementary Data, with the reanalysis code, README and aggregate results in the accompanying code materials.

S1 Analysis population and missing covariates

The analysis contains 9,913 eruption episodes, of which 7,721 have a reported VEI and 753 satisfy VEI≥4. Retaining usable VEI, crustal thickness and nearest-oceanic age yields the fixed model sample of 7,595 episodes and 742 large outcomes. Catalogue inclusion defines the study population. Nine retained episodes have nominal start years before approximately 9700 BCE, the conventional Holocene boundary; three have a VEI and two enter the fixed model sample. We retain the archived dates and outcomes, including their mixed dating methods and source uncertainties. The comparison therefore concerns the archived episodes associated with catalogue-listed Holocene volcanoes.

Missing basement age is assigned 500 Ma and represented by a separate missingness indicator, affecting 700 NH and 402 SH episodes. Missing sediment thickness would be assigned zero before transformation; no retained episode requires this assignment. These fixed computational conventions are applied identically across the model comparisons.

The global target of the fixed-sample standardization is the empirical covariate distribution of these retained records. It is entry-weighted. The reanalysis does not estimate an equal-volcano-weighted contrast; clustering changes uncertainty, not the entry weights. The analysis remains conditional on inclusion in the catalogue. No exposure-time denominator has been constructed.

S2 Frozen model specification

The archived covariate table, eruptions_NG.csv, uses point assignment from the 1° CRUST1.0 grid for crustal and sediment thickness (km), the nearest nonmissing oceanic-age grid value within 5° for seafloor age (Ma), and point-in-polygon assignment to the Hasterok geological provinces with a mapped lastorogen value for basement age (Ma). The regressions use these frozen assignments. Spatial joins and the exact EarthByte grid release have not been independently revalidated; reproducibility therefore rests on the supplied assigned table, with its coarse spatial and geological resolution.

Let Y indicate VEI≥4 for a recorded eruption episode and H indicate a Northern Hemisphere volcano. The baseline model is a logistic regression with an intercept and the following predictors:

Predictor	Operational definition
Hemisphere	H=1 for NH, 0 for SH
Crust	crust_km−30
Oceanic-age proxy	log10(max(nearest_oceanic_age_Ma, 0.1))
Basement-age proxy	log10(max(basement_age_Ma after assigning missing values 500, 1))
Sediment proxy	log10(sed_km after assigning missing values 0, plus 0.1)
High-latitude proxy	1 if latitude>55° or latitude<−40°, otherwise 0

Predictor	Operational definition
Rift	tectonic-class indicator
Intraplate	tectonic-class indicator
Basement-age missing	indicator for an unavailable original basement-age value

The tectonic classifier lowercases the source field and tests for Subduction first, then Rift, then Intraplate, with Other otherwise. There is no separate Other dummy in the archived design: Other and Subduction therefore share the reference coding. The actual reference group therefore combines Subduction and Other. The fixed sample contains Subduction 6,496, Rift 565, Intraplate 533 and Other 1; the ambiguity-exclusion sample contains 5,657, 565, 514 and 1, respectively.

The main extended model adds an indicator for year≥1500 CE. The year≥1826 CE sensitivity marks a 200-year window relative to the 2026 extraction. The common row sample and other covariate definitions are fixed. Checks cover constant columns, matrix rank, separation, convergence and the observed-information matrix. The baseline design has 10 columns and each calendar-extended design has 11, including the intercept; no constant column was removed in these fits. All main fits passed the numerical checks.

The high-latitude term defines a geographical category for the fixed specification. Testing deglaciation mechanisms calls for reconstructed ice cover and regional timing. The nearest available oceanic age likewise serves as a coarse geographical proxy; direct slab-age interpretation requires volcano-specific measurements.

In the fixed-sample high-latitude sensitivity, replacing latitude>55° or latitude<−40° with |latitude|≥45° changes 758 indicators. With all other terms fixed, the post-1500 conditional OR increases from 1.162658882 to 1.244040649. A separate tectonic-priority comparison changes zero classifications in the actual sample. The primary models consistently use the first high-latitude definition and subduction-first tectonic classification.

S3 Model-standardized contrasts

For each fitted model, predict each retained row twice, changing H to 1 and to 0 while holding its other covariates fixed. Average each set of predictions over the same empirical target distribution to obtain standardized NH and SH recorded probabilities. Report the difference in percentage points and the probability ratio, together with the conditional odds ratio as a supplementary model coefficient.

These predictions are descriptive standardizations, not estimates of an intervention that moves a volcano between hemispheres. Poor overlap in geology or calendar composition can make them extrapolative. Regional coverage and the ambiguity-exclusion sensitivity describe some support limitations. No multivariable common-support trimming or propensity-overlap restriction was performed; the standardization may still extrapolate. A reduction of a conditional odds ratio is not translated into a causal percentage explained.

Intervals use the delta method applied to the model coefficients and the chosen clustered covariance. The target empirical covariate distribution is treated as fixed, so these intervals do not include uncertainty from resampling that distribution. Report volcano-cluster and region-cluster results separately. The primary sample has 661 volcano clusters and 30 regional clusters, with t reference degrees of freedom 660 and 29. The ambiguity-exclusion sample has 619 volcano clusters and 29 regional clusters, with degrees of freedom 618 and 28. Small and imbalanced cluster structures remain a limitation even with the t reference.

S4 Regional assignment and support

The archived regional fixed-effects code assigns a record to the first inclusive latitude–longitude box it matches. Several boxes overlap. In particular, Sunda–Banda fully contains the later Sulawesi–Halmahera box, and Africa-West fully contains the later Cameroon box. Those later labels are therefore

unreachable in that first-match implementation. The earlier Japan box also overlaps part of the Izu–Bonin–Marianas box. These are properties of the code, not findings about whether volcanism is absent from those regions.

Two region-assignment routes are retained for sensitivity: reproduce the original first-match algorithm exactly; and exclude records matching boxes with more than one distinct label using the original full box set. The second route changes the supported sample and is not described as a geological correction. It does not choose a new boundary or reorder boxes in response to significance. The ambiguity-exclusion route removes 858 of 7,595 episodes, leaving 6,737 episodes and 674 large outcomes. It does not remove every unsupported covariate combination.

The archived Mantel–Haenszel code instead filters each candidate box independently, permitting potential overlap. The actual local eligibility check found zero duplicated entries across the strata admitted to the reported pool. Sulawesi–Halmahera had 277 NH and three SH entries and failed the minimum-five-per-hemisphere criterion; Cameroon had 18 and zero. Thus the potential overlap did not produce actual duplicate use in this pool. Its uncertainty still requires consideration of repeated records and its target remains the restricted regional population.

The four-region MH benchmark is reproducible from the counts in Table S2. All 30 leave-region-out fits completed; their point differences span 0.4024–1.6286 percentage points, with the lower endpoint omitting Sunda–Banda and the upper endpoint omitting Kamchatka–Kurile. The range summarizes sensitivity across region exclusions.

S5 Missing recorded VEI

All records have a usable year in the local input. The missing-outcome prediction design contains an intercept, NH, post-1500 and their interaction. The undated indicator is constant and is omitted. No geological covariates enter this separate sensitivity model. For included records with missing VEI, use a pattern-mixture sensitivity in which the log odds of a large entry are shifted from a fitted observed-VEI model. Use the frozen offsets $-\log(4)$, $-\log(2)$, 0, $\log(2)$, $\log(4)$ separately for the two hemispheres and report all 25 combinations. Offset zero is an assumption, not a test establishing missing at random.

Also report extreme bounds obtained by assigning the missing binary outcomes to 0 or 1. These can be wide and should not be presented as likely scenarios. Sensitivity point estimates are not confidence intervals. No sampling confidence intervals are attached to this grid. A resampling extension would have to refit the prediction model, rather than varying only Bernoulli outcomes at a fixed probability.

The complete grid reports expected outcomes at each specified offset without random completion draws. It addresses missing VEI among catalogued episodes; omitted eruptions require a separate model of catalogue inclusion.

S6 LaMEVE scope audit

The archived LaMEVE extract has an age-before-present range of −64 to 2,560,000 years and therefore includes pre-Holocene material. Older records are not uniformly represented in its CE field. Hemisphere counts across this full extract and GVP Holocene volcano totals describe different time domains and denominator populations.

The reported analyses use GVP episodes with the defined outcomes and samples. An aligned LaMEVE comparison requires explicit age filtering, volcano mapping and a common denominator that accommodates its large-eruption selection. Shared regional research histories also enter comparisons of catalogue completeness.

S7 Model results and uncertainty

Table S1a. Standardized differences in percentage points and conditional odds ratios. All main models use 7,595 episodes and 742 large episodes. Both cluster intervals use t(G−1); the empirical prediction target is fixed.

Model	NH percent	SH percent	Difference pp	Volcano CI pp	Region CI pp
No period term	12.2015	6.2052	5.9963	2.3777 to 9.6150	3.2943 to 8.6984
Post 1500	10.0798	9.0397	1.0400	−2.0841 to 4.1641	−1.3460 to 3.4260
Post 1826	10.5531	8.0575	2.4956	−0.9023 to 5.8936	0.4790 to 4.5123
Post 1500 ambiguity excluded	10.5069	9.0024	1.5045	−1.6332 to 4.6422	−0.8506 to 3.8595

Table S1b. Conditional odds ratios for the same fits.

Model	Conditional OR	Volcano OR CI	Region OR CI
No period term	2.150632	1.368835 to 3.378945	1.474441 to 3.136929
Post 1500	1.162659	0.738596 to 1.830196	0.825447 to 1.637630
Post 1826	1.418774	0.872530 to 2.306993	1.075253 to 1.872042
Post 1500 ambiguity excluded	1.238655	0.792420 to 1.936178	0.892294 to 1.719462

The ambiguity-exclusion row has a different target sample: n=6,737 and 674 large outcomes. Probability ratios for the first three rows are 1.966346, 1.115049 and 1.309729. Their volcano-cluster intervals are 1.321468–2.925925, 0.801590–1.551085 and 0.897459–1.911384, respectively; their regional-cluster intervals are 1.402945–2.755998, 0.867371–1.433452 and 1.056569–1.623546. The ambiguity-exclusion probability ratio is 1.167120, with volcano interval 0.842600–1.616624 and regional interval 0.916144–1.486850.

S8 Four region benchmark

Table S2. Verified scored-entry counts for eligible MH strata. Latitude and longitude intervals are inclusive, in degrees.

Region	Latitude	Longitude	NH large	NH other	SH large	SH other
Sunda–Banda	−11 to 6	95 to 130	11	288	31	942
East Africa Rift	−12 to 20	28 to 45	3	47	5	120
Northern Andes	−5 to 12	−82 to −65	26	89	35	137
Galápagos	−2 to 1	−92 to −89	1	12	1	66

The pooled sample contains 1,814 scored entries and 113 large entries: NH 41/477 and SH 72/1,337. The MH odds ratio is 1.210382380, with Robins–Breslow–Greenland log-variance 0.045816285 and a normal-reference 95% interval of 0.795656–1.841281. The calculation was independently reproduced from the four contingency tables and cross-checked against the archived pooling function. No record belongs to multiple eligible strata. This RBG interval addresses stratified contingency-table variance, not repeated episodes within a volcano; it is a descriptive benchmark rather than the primary dependence-aware inference.

S9 Complete missingness sensitivity grid

Table S3. Fractions and NH–SH differences among all 9,913 recorded episodes after expected-value completion of missing VEI. Multipliers apply to the odds in missing outcomes relative to the fitted observed-outcome model. Percentages are 100 times the probabilities; differences are percentage points. All 25 combinations are shown. They have no assigned probability weights.

NH odds multiplier	SH odds multiplier	NH percent	SH percent	Difference pp
0.25	0.25	10.9817	7.4718	3.5099
0.25	0.5	10.9817	8.6391	2.3426
0.25	1	10.9817	10.2130	0.7687
0.25	2	10.9817	11.9504	−0.9687
0.25	4	10.9817	13.5301	−2.5484
0.5	0.25	12.9982	7.4718	5.5265
0.5	0.5	12.9982	8.6391	4.3591
0.5	1	12.9982	10.2130	2.7852
0.5	2	12.9982	11.9504	1.0478
0.5	4	12.9982	13.5301	−0.5319
1	0.25	16.0928	7.4718	8.6210
1	0.5	16.0928	8.6391	7.4537
1	1	16.0928	10.2130	5.8798
1	2	16.0928	11.9504	4.1424
1	4	16.0928	13.5301	2.5627
2	0.25	20.0986	7.4718	12.6268
2	0.5	20.0986	8.6391	11.4595
2	1	20.0986	10.2130	9.8856
2	2	20.0986	11.9504	8.1482
2	4	20.0986	13.5301	6.5685
4	0.25	24.3131	7.4718	16.8414
4	0.5	24.3131	8.6391	15.6740
4	1	24.3131	10.2130	14.1001
4	2	24.3131	11.9504	12.3627
4	4	24.3131	13.5301	10.7830

At multiplier 1 in both hemispheres, the expected completed fractions are 16.0928% and 10.2130%. These preserve each hemisphere's own period mix and are not the covariate-standardized probabilities in Table S1. Three cells have a negative difference; 3/25 is not a probability that the physical contrast reverses.

Extreme outcome bounds are NH 8.5087–35.3863% and SH 5.8495–18.8419%, permitting differences from −10.3332 to 29.5368 percentage points. They reflect unconstrained missing outcomes among included records only.

S10 Source snapshots and sample construction

The position CSV contains 1,215 rows and 1,215 valid volcano identifiers, all with parseable coordinates. Its header identifies Volcanoes of the World 5.3.5 and a download time of 29 April 2026 at

08:41 AM. The eruption worksheet inside the archived workbook identifies version 5.3.5 and a download date of 24 April 2026. The filename date is not used to replace that embedded date.

The same workbook's Info sheet supplies a citation for version 5.2.8, dated 6 May 2025. Its listed DOI has the registered title ‘Volcanoes of the World, v.5.2.0’ (released 6 June 2024): https://doi.org/10.5479/si.GVP.VOTW5-2024.5.2. This inconsistency is present in the archived source. We report the data-table labels and embedded dates without claiming that this DOI identifies the 5.3.5 snapshot. The raw-source metadata and assigned tables are necessary to distinguish the analyzed extraction from later catalogue releases.

The eruption worksheet has 9,918 rows, all Confirmed Eruption. Five rows, representing five different volcano identifiers, are absent from the position table and are excluded by the join, giving 9,913. No additional Discredited or non-Confirmed records are removed. Among the 9,913, 7,721 have VEI and 753 reach VEI≥4. The fixed covariate filter removes 126 scored records and 11 large records, leaving 7,595 and 742. Of these model records, 4,722 are NH and 2,873 SH. Supplementary Data includes byte-preserved copies of the source worksheet and position export alongside the assigned analysis tables. Nine retained episodes have nominal start years before approximately 9700 BCE, with the earliest at 11345 BCE. Their source date methods and uncertainty fields remain in the worksheet; three have VEI and two enter the fixed model sample. The analysis retains the catalogue population and archived date values.

S11 Regional and covariate support

Table S4. NH and SH counts in the fixed model sample under the archived first-match region rule. These are records, not volcano counts. The 25 one-sided labels and uneven bilateral counts warn against interpreting global standardization as matched common support.

Region	NH records	SH records
Aegean	18	0
Africa-East-Rift	50	58
Africa-West	18	0
Aleutian-Alaska	320	0
Antarctica	0	9
Antilles	65	0
Azores-MAR	41	0
C-Andes	0	88
Cascades	96	0
Central-America	567	0
Galapagos	13	67
Hawaii	206	0
Iceland-MAR	274	0
Izu-Bonin-Marianas	112	0
Japan	1000	0
Kamchatka-Kurile	728	0
Med-Italy	236	0
Mexican-volc-belt	133	0
N-Andes	115	172
NZ-Taupo	0	190
Other	201	171
PNG-Solomon-Bismarck	0	222

Region	NH records	SH records
Philippines	191	0
Reunion-Mauritius	0	188
Ryukyu-Taiwan	39	0
S-Andes-Patagonia	0	421
S-Sandwich	0	20
Sunda-Banda	299	973
Tonga-Kermadec	0	158
Vanuatu	0	136

Table S5. Marginal covariate support in the transformed model variables. C=crust_km−30; O=log10(max(oceanic-age proxy,0.1)); B=log10(max(basement-age proxy after filling,1)); S=log10(sed_km+0.1). These summaries do not demonstrate sufficient joint covariate overlap.

Hemisphere	Variable	Min	P5	Median	P95	Max
NH	C	−23.570000	−19.250000	0.000000	12.330000	20.500000
SH	C	−23.540000	−19.730000	1.000000	27.670000	40.170000
NH	O	−1.000000	0.870989	1.744840	2.352588	2.422655
SH	O	−1.000000	1.013719	1.823930	2.130977	2.204066
NH	B	1.698970	1.698970	1.845098	2.698970	2.698970
SH	B	1.698970	1.698970	1.845098	2.698970	2.698970
NH	S	−1.000000	−1.000000	−0.221849	0.574031	0.934498
SH	S	−1.000000	−0.823909	0.000000	0.662758	0.724276

Table S6. Joint support for broad tectonic class and the high-latitude proxy. Only one Other-class observation is present, in the NH.

Tectonic class	High latitude proxy	NH records	SH records
Subduction	0	3562	2379
Subduction	1	447	108
Rift	0	163	117
Rift	1	280	5
Intraplate	0	260	242
Intraplate	1	9	22
Other	0	1	0

No additional test or model alteration was selected from these diagnostics. The standardization remains model-dependent, including predictions for covariate combinations that may be poorly supported within a hemisphere.

Supplementary reference

Robins JM, Breslow NE, Greenland S (1986) Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. Biometrics 42:311–323. https://doi.org/10.2307/2531052