Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Rongzhen Dai

dairongzhen3@gmail.com

Research Article

Keywords: eruption catalogue, volcanic explosivity, Global Volcanism Program, recording bias, missing data

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11603899/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

This study quantifies the sensitivity of the hemispheric contrast in recorded volcanic explosivity to temporal coverage. We compare Volcanic Explosivity Index (VEI) ≥ 4 fractions in an archived Smithsonian Global Volcanism Program catalogue using fixed-sample standardization and missing-VEI sensitivity analysis. Among 7,721 VEI-scored eruption episodes, 11.64% of Northern Hemisphere (NH) entries and 6.72% of Southern Hemisphere (SH) entries reach VEI ≥ 4, with a crude odds ratio of 1.83 and proportion ratio of 1.73. Logistic models use the same 7,595 episodes and empirical covariate distribution throughout. Adding a post-1500 CE indicator reduces the standardized NH–SH difference from 6.00 to 1.04 percentage points, with a volcano-cluster 95% confidence interval of − 2.08 to 4.16. A post-1826 indicator gives a difference of 2.50 percentage points; its interval includes zero under volcano clustering and excludes zero under regional clustering. Excluding geographically ambiguous assignments gives a difference of 1.50 percentage points. In a separate analysis of all 9,913 recorded episodes, three of 25 specified missing-VEI scenarios reverse the hemispheric contrast. The comparison quantifies how calendar specification, geographical support and missing-outcome assumptions affect this catalogue result. Reporting the episode denominator, temporal composition and dependence structure makes the hemispheric comparison reproducible and its interpretation explicit.

Introduction

Global eruption inventories support comparisons across regions and periods, and their entries reflect heterogeneous observation systems. Recent activity may be monitored directly, whereas prehistoric eruptions are reconstructed from deposits whose preservation and recognition depend on environment, accessibility and geological investigation. Small eruptions are particularly liable to be absent from older inventories. The LaMEVE compilation and reassessments of postglacial volcanism have emphasized the consequences of these differences for global comparisons (Crosweller et al. 2012; Brown et al. 2014; Watt et al. 2013).

The hemispheres differ in volcanic geography and in their histories of observation and geological study. Differences in volcanic populations, period composition and the recording of small eruptions can accompany a higher fraction of large eruptions in one hemisphere’s catalogue. Studies of LaMEVE have quantified regional under-recording probabilities (Rougier et al. 2018) and examined the unequal contributions of individual volcanoes to the record (Deligne et al. 2017). Mead and Magill (2014) estimated regional change points in VOTW recording completeness, while Sheldrake et al. (2020) related eruption-size distributions to multidimensional tectonic settings in volcanic arcs. These studies establish why a global comparison needs an explicit denominator and treatment of recording history.

We quantify the NH–SH contrast in the fraction of recorded episodes reaching Volcanic Explosivity Index (VEI) ≥ 4 under specified calendar, geographical and missing-VEI choices. The main comparison holds both the episode sample and the empirical covariate target fixed while adding a calendar term. This isolates the change associated with temporal specification in this particular catalogue comparison.

Regional and missing-attribute analyses then show how its interpretation changes with support and unobserved outcomes.

We report three targets with explicit denominators: the crude contrast among episodes with known VEI, predictions standardized to one fixed empirical covariate distribution, and expected fractions among all included episodes under specified missing-VEI scenarios. The analyses are exploratory; the model comparison and sensitivity grid were fixed before the new calculations.

Data and methods

Catalogue unit and sample

We used archived extracts of the Smithsonian Global Volcanism Program (GVP) Volcanoes of the World catalogue assembled for this study. Data headers identify version 5.3.5, with an embedded download date of 24 April 2026 for the eruption worksheet and 29 April 2026 for the position CSV. The workbook's suggested citation instead names version 5.2.8; we document this source inconsistency rather than assign an unverified version-specific DOI (Online Resource 1, Supplementary Methods S10). The source position table contains 1,215 Holocene volcanoes. The eruption worksheet contains 9,918 entries, all labelled Confirmed Eruption; five entries, each belonging to a different volcano absent from the position table, fail the identifier join, leaving 9,913 episodes. The position-table size is not the number of distinct volcanoes represented in every analysis. Of these episodes, 7,721 have a VEI and 753 reach VEI ≥ 4. Hemisphere is assigned from the parent volcano's latitude, with positive values classified as Northern Hemisphere (NH) and negative values as Southern Hemisphere (SH). 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.

The unit is a GVP eruption episode, which can contain several explosive phases. GVP describes renewed activity after more than three months of inactivity as a new eruption; shorter pauses can remain within one episode (Global Volcanism Program, undated). We apply this catalogue unit consistently. VEI combines evidence on eruption magnitude whose availability and precision vary across eras, environments and eruptive styles (Newhall and Self 1982). We retain absent VEI values as missing.

The outcome is whether an episode reaches VEI ≥ 4, and the scored-episode denominator includes both large and smaller recorded episodes. The model sample excludes 126 scored episodes, including 11 large outcomes, because required covariates are unavailable. The resulting fractions describe the composition of recorded episodes. Physical incidence rates require an observation-time denominator.

Fixed sample and model comparison

The primary models use the same 7,595 episodes, including 742 VEI ≥ 4 episodes, with usable VEI, crustal thickness and nearest-oceanic age. Missing basement age is assigned 500 Ma with a missingness indicator, affecting 700 NH and 402 SH episodes. The sediment-thickness rule assigns zero before transformation if needed; all retained episodes have observed sediment thickness. The same assignments are used in each model.

The baseline logistic model includes hemisphere, four geological covariates, a high-latitude indicator, tectonic-class indicators and the basement-age missingness indicator. The archived table links crustal and sediment thickness to CRUST1.0 (Laske et al. 2013), oceanic age to EarthByte seafloor-age grids (EarthByte, undated), and basement age to the geological provinces of Hasterok et al. (2022). With thicknesses in kilometres and ages in Ma, the transformations are crustal thickness − 30, log10(max(oceanic age, 0.1)), log10(max(filled basement age, 1)) and log10(sediment thickness + 0.1). Online Resource 1, Supplementary Methods S2, documents the assigned covariates. The high-latitude indicator is latitude > 55° or latitude < − 40°. Tectonic classification tests subduction first, then rift, then intraplate. The reference coding combines subduction and Other; only one retained episode belongs to Other.

We add an indicator for year ≥ 1500 CE to the baseline model and use year ≥ 1826 CE as a second temporal specification. The latter marks a 200-year window relative to the 2026 catalogue extraction (2026 − 200=1826). Both thresholds define calendar contrasts rather than dates of globally complete recording; calendar composition includes documentation, preservation and temporal variation in volcanism.

For each model, we predict each retained episode twice, setting hemisphere to NH and SH while keeping its remaining covariates unchanged, and average over the same empirical distribution, using marginal standardization (Muller and MacLehose 2014). We report standardized probabilities, their difference in percentage points and their ratio. Conditional odds ratios are secondary. These are descriptive model-standardized associations; where covariate support differs between hemispheres, prediction requires extrapolation. Comparing conditional odds ratios alone would additionally be affected by noncollapsibility (Schuster et al. 2021).

Dependence and geographical sensitivity

Repeated episodes at one volcano and geographically related records motivate two uncertainty calculations. We report sandwich covariance estimates clustered by volcano and, separately, by heuristic geographical region. Intervals use a t reference with G − 1 degrees of freedom, where G is the number of clusters, and delta-method gradients of the standardized contrasts. The primary sample contains 661 volcano clusters and 30 regional clusters. These intervals condition on the empirical covariate target. Five of the 30 model-region labels contain records in both hemispheres, including the residual Other category; the remaining 25 are one-sided. Standardization therefore relies partly on predictions across differing geographical support.

The archived regional algorithm assigns each location to the first matching latitude–longitude box. Some boxes overlap. A second route excludes entries matching more than one distinct regional label, without changing boundaries or choosing their order according to results. This is a support sensitivity, not a geological correction. We also omit each of the 30 regions in turn from the post-1500 model.

As a descriptive regional benchmark, we retain a Mantel–Haenszel comparison of four cross-equatorial regions with at least five scored entries in each hemisphere. We audit actual membership rather than assuming that all candidate geographical boxes define independent strata. This pooled contrast targets the eligible regions, not the global fixed sample. Its usual record-level interval is not a substitute for volcano-cluster inference.

Missing recorded VEI

We first summarize missing VEI by hemisphere and period. A separate sensitivity analysis uses all 9,913 included episodes. Among observed outcomes, a logistic model includes hemisphere, post-1500 period and their interaction. For a missing outcome, its fitted log odds of VEI ≥ 4 are shifted by a hemisphere-specific offset δ. This uses group-specific log-odds sensitivity parameters for incomplete binary outcomes (Tompsett et al. 2020). We use the five offsets log(0.25), log(0.5), 0, log(2) and log(4) in all 25 NH–SH combinations. Observed outcomes are retained and missing outcomes contribute their expected probabilities to hemisphere-specific fractions.

These offsets specify multipliers of the fitted odds of a large event among missing entries. Offset zero uses the observed-outcome model directly; the other offsets represent departures from that model. We also calculate extreme bounds by assigning all missing outcomes in each hemisphere to zero or one. The grid and bounds describe sensitivity among included episodes; scenario ranges are distinct from sampling confidence intervals. Missing eruptions require a separate recording model.

Large-language-model tools assisted analysis design, code development and review, substantive restructuring, and drafting and revision of the manuscript. Executed code applied to the archived data produced the statistical outputs. The author remains responsible for checking the analyses, references, interpretation and final manuscript.

Results

The crude contrast and missingness distribution

The NH sample contains 554 VEI ≥ 4 episodes among 4,761 scored entries; the SH sample contains 199 among 2,960. The recorded fractions are 11.64% and 6.72%, respectively, a difference of 4.91 percentage points. The proportion ratio is 1.731 and the odds ratio is 1.827.

VEI is missing for 2,192 of the 9,913 recorded episodes, or 22.11%. Missingness is concentrated in the earlier record (Table 1; Fig. 1). It affects 1,538 of 2,817 pre-1500 NH entries and 328 of 608 pre-1500 SH entries, compared with 212 of 3,694 and 114 of 2,794 post-1500 entries. Thus similar within-period missingness can coexist with unequal overall missingness when the hemispheres contain different mixtures of periods.

Table 1
Recorded episodes and missing VEI by hemisphere and period.

Hemisphere and period	All episodes	Missing VEI	Missing percent
NH before 1500 CE	2,817	1,538	54.60
SH before 1500 CE	608	328	53.95
NH from 1500 CE	3,694	212	5.74
SH from 1500 CE	2,794	114	4.08

Temporal specification changes the standardized contrast

Without a calendar term, the standardized NH and SH probabilities are 12.20% and 6.21%, giving a difference of 6.00 percentage points and a conditional odds ratio of 2.151. Adding the post-1500 term on exactly the same 7,595 episodes gives probabilities of 10.08% and 9.04%, a difference of 1.04 percentage points and a conditional odds ratio of 1.163 (Table 2; Fig. 2). The volcano-cluster interval for the probability difference is − 2.08 to 4.16 percentage points; the regional-cluster interval is − 1.35 to 3.43.

The post-1826 specification gives probabilities of 10.55% and 8.06% and a difference of 2.50 percentage points. Its volcano-cluster interval includes zero, whereas its regional-cluster interval excludes zero. The two interval calculations show how the uncertainty assessment changes with the grouping used to represent dependence.

Table 2
Standardized recorded probabilities and differences on the fixed sample of 7,595 episodes. Differences and interval endpoints are percentage points. Differences are computed from unrounded probabilities before two-decimal display.

Calendar term	NH percent	SH percent	NH minus SH	Volcano cluster 95% CI	Region cluster 95% CI
None	12.20	6.21	6.00	2.38 to 9.62	3.29 to 8.70
Post 1500	10.08	9.04	1.04	−2.08 to 4.16	−1.35 to 3.43
Post 1826	10.55	8.06	2.50	−0.90 to 5.89	0.48 to 4.51

A high-latitude sensitivity comparison replaces the primary rule (latitude > 55° or latitude < − 40°) with |latitude|≥45° on the same 7,595 episodes. The replacement changes 758 indicators and increases the post-1500 conditional odds ratio from 1.163 to 1.244. All other model terms remain fixed.

Geographical support sensitivity

Excluding 858 entries with ambiguous regional assignments leaves 6,737 episodes, including 674 with VEI ≥ 4. Under the post-1500 model, the standardized probabilities are 10.51% and 9.00%, a difference of 1.50 percentage points. Its volcano-cluster interval is − 1.63 to 4.64 percentage points; its regional-cluster interval is − 0.85 to 3.86. These estimates describe the restricted sample and its empirical covariate target.

All 30 leave-one-region-out fits completed. Their standardized differences range from 0.40 percentage points when Sunda–Banda is removed to 1.63 when Kamchatka–Kurile is removed. This range summarizes the influence of individual regions on the point estimate.

The four-region benchmark uses Sunda–Banda, East Africa Rift, Northern Andes and Galápagos. No episode is duplicated among the strata meeting the bilateral sample threshold. Sulawesi–Halmahera has only three scored SH entries, and Cameroon has none, so neither enters the pool. Online Resource 1, Supplementary Methods S8 and Table S2, reports the regional counts and pooled benchmark, which describes the eligible regional strata using a record-level interval.

Missingness assumptions can change the direction

Under zero offsets, the expected fractions among all recorded episodes are 16.09% for NH and 10.21% for SH, a difference of 5.88 percentage points. This all-record calculation retains each hemisphere's observed period composition and has a different denominator from the fixed-sample standardized estimates in Table 2.

Across the 25 specified scenarios, the NH–SH difference ranges from − 2.55 to 16.84 percentage points, with a negative value in three scenarios (Fig. 3; Online Resource 1, Table S3). For example, multiplying the odds of VEI ≥ 4 among missing NH entries by one quarter and among missing SH entries by four gives expected fractions of 10.98% and 13.53%. This combination illustrates how different assumptions for the two hemispheres can reverse the all-record contrast.

The extreme missing-outcome bounds are 8.51–35.39% for NH and 5.85–18.84% for SH. Their overlap summarizes the range permitted by assigning the missing outcomes to their extremes.

Discussion

Temporal composition of the recorded contrast

The full scored catalogue contains a greater fraction of VEI ≥ 4 entries in the NH, but a single pooled ratio conceals several choices. The model-standardized difference changes substantially when calendar composition is represented on a fixed row set. Missingness is also concentrated in older entries, and assumptions about those missing outcomes can change the hemispheric direction. These results make the catalogue denominator and temporal support central to interpretation.

Holding the episode sample and empirical target fixed makes the temporal comparison directly interpretable as a change in specification. Calendar period can represent genuine temporal variation, regional composition, preservation and documentation together. The reduction in the standardized difference measures their combined association in the fitted models. Rougier et al. (2018) address a related question by estimating regional recording probabilities for large eruptions. Here the outcome includes both smaller and larger scored episodes, and the separate missingness calculation concerns VEI attributes of episodes already present in the catalogue.

Regional support and uncertainty

The ambiguity-exclusion estimate and all leave-region-out point estimates remain positive. Removing Sunda–Banda lowers the standardized difference from 1.04 to 0.40 percentage points, showing appreciable regional influence despite preservation of direction. The 25 one-sided regional labels also show why the model-standardized comparison differs from a matched within-region design. Its interpretation depends on both the fitted covariate relationships and the combinations of covariates represented in the catalogue.

The post-1500 volcano-cluster interval, − 2.08 to 4.16 percentage points, includes differences in either direction and remains compatible with materially important contrasts. The post-1826 estimate further shows that calendar specification and clustering can change the interval’s relationship to zero. These results favor reporting the estimates and both uncertainty calculations together.

Catalogue alignment and volcanological interpretation

Catalogue comparisons require alignment of their time domains and denominators. The archived LaMEVE extract spans ages up to 2.56 million years before present and is selected for large eruptions, whereas the present GVP outcome compares large and smaller scored episodes at catalogue-listed volcanoes. A cross-catalogue comparison therefore requires explicit age filtering, volcano mapping and a common denominator, while accounting for shared regional source literature (Brown et al. 2014).

Physical interpretation requires measurements matched to the proposed process. The high-latitude rule is a geographical covariate; testing unloading responses calls for reconstructed ice cover, regional timing and lagged responses (Watt et al. 2013). Similarly, geological distributions and interactions matter beyond a comparison of median crustal properties. The present catalogue diagnostics help identify which population and time support such process-specific analyses would need.

VEI describes explosive magnitude, while sulfur injection and climate forcing require additional eruption-specific information. Using the present catalogue fractions in those applications therefore calls for aligning the physical quantity and its denominator.

Conclusions

The scored eruption catalogue contains a greater fraction of VEI≥4 episodes in the NH. On the same 7,595 model episodes and empirical covariate target, adding a post-1500 term changes the standardized NH–SH difference from 6.00 to 1.04 percentage points, with a volcano-cluster 95% interval of −2.08 to 4.16. Calendar specification and geographical support alter its magnitude and uncertainty, and three specified missing-VEI scenarios reverse the all-record contrast. These results provide a reproducible hemispheric comparison whose denominator, temporal coverage and missing-outcome assumptions are explicit.

Declarations

Author approval and responsibility: Rongzhen Dai is the sole author, has approved the manuscript and accepts responsibility for its content.

Funding: This research received no external funding.

Competing interests: The author declares no competing interests.

Author Contribution

R.D., the sole author, was responsible for the study design, data analysis, interpretation of the results, visualization, and preparation and revision of the manuscript. R.D. approved the final manuscript and accepts responsibility for its content.

Data Availability

The frozen input data, including the derived analysis tables and archived GVP source exports, are provided in Online Resource 3. The executable analysis code, environment information and reference results are provided in Online Resource 2. Source versions and sample construction are documented in Online Resource 1.

Data and code availability

GVP source catalogues are available at https://volcano.si.edu. Online Resources 1–3 provide the supplementary methods, executable code and frozen input data described below. Online Resource 1, Supplementary Methods S10, documents the archived data-header versions and embedded citation. The

manuscript and supplementary methods retain source attribution and describe the construction and resolution of the assigned geological covariates.

References

1. Brown SK, Crosweller HS, Sparks RSJ, Cottrell E, Deligne NI, Guerrero NO, Hobbs L, Kiyosugi K, Loughlin SC, Siebert L, Takarada S (2014) Characterisation of the Quaternary eruption record: analysis of the Large Magnitude Explosive Volcanic Eruptions (LaMEVE) database. J Appl Volcanology 3:5. https://doi.org/10.1186/2191-5040-3-5
2. Crosweller HS, Arora B, Brown SK, Cottrell E, Deligne NI, Guerrero NO, Hobbs L, Kiyosugi K, Loughlin SC, Lowndes J, Nayembil M, Siebert L, Sparks RSJ, Takarada S, Venzke E (2012) Global database on large magnitude explosive volcanic eruptions (LaMEVE). J Appl Volcanology 1:4. https://doi.org/10.1186/2191-5040-1-4
3. Deligne NI, Sparks RSJ, Brown SK (2017) Report on potential sampling biases in the LaMEVE database of global volcanism. J Appl Volcanology 6:7. https://doi.org/10.1186/s13617-017-0058-5
4. EarthByte (undated) Seafloor age data and models. https://www.earthbyte.org/category/resources/data-models/seafloor-age/ Accessed 3 November 2026
5. Global Volcanism Program (2026) (undated) What volcanoes have had the longest eruptions? Smithsonian Institution. https://volcano.si.edu/faq/Longest_Eruptions.cfm Accessed 3 November
6. Hasterok D, Halpin JA, Collins AS, Hand M, Kreemer C, Gard MG, Glorie S (2022) New maps of global geological provinces and tectonic plates. Earth Sci Rev 231:104069. https://doi.org/10.1016/j.earscirev.2022.104069
7. Laske G, Masters G, Ma Z, Pasyanos M (2013) Update on CRUST1.0 — A 1-degree global model of Earth's crust. Geophys Res Abstracts 15:EGU2013-2658. https://igppweb.ucsd.edu/~gabi/crust1.html
8. Mead S, Magill C (2014) Determining change points in data completeness for the Holocene eruption record. Bull Volcanol 76:874. https://doi.org/10.1007/s00445-014-0874-y
9. Muller CJ, MacLehose RF (2014) Estimating predicted probabilities from logistic regression: different methods correspond to different target populations. Int J Epidemiol 43:962–970. https://doi.org/10.1093/ije/dyu029
10. Newhall CG, Self S (1982) The Volcanic Explosivity Index (VEI): an estimate of explosive magnitude for historical volcanism. J Phys Res 87(C2):1231–1238. https://doi.org/10.1029/JC087iC02p01231
11. Rougier J, Sparks RSJ, Cashman KV (2018) Regional and global under-recording of large explosive eruptions in the last 1000 years. J Appl Volcanology 7:1. https://doi.org/10.1186/s13617-017-0070-9
12. Schuster NA, Twisk JWR, ter Riet G, Heymans MW, Rijnhart JJM (2021) Noncollapsibility and its role in quantifying confounding bias in logistic regression. BMC Med Res Methodol 21:136.

https://doi.org/10.1186/s12874-021-01316-8
13. Sheldrake T, Caricchi L, Scutari M (2020) Tectonic Controls on Global Variations of Large-Magnitude Explosive Eruptions in Volcanic Arcs. Front Earth Sci 8:127. https://doi.org/10.3389/feart.2020.00127
14. Tompsett D, Sutton S, Seaman SR, White IR (2020) A general method for elicitation, imputation, and sensitivity analysis for incomplete repeated binary data. Stat Med 39:2921–2935. https://doi.org/10.1002/sim.8584
15. Watt SFL, Pyle DM, Mather TA (2013) The volcanic response to deglaciation: evidence from glaciated arcs and a reassessment of global eruption records. Earth Sci Rev 122:77–102. https://doi.org/10.1016/j.earscirev.2013.03.007

Figures

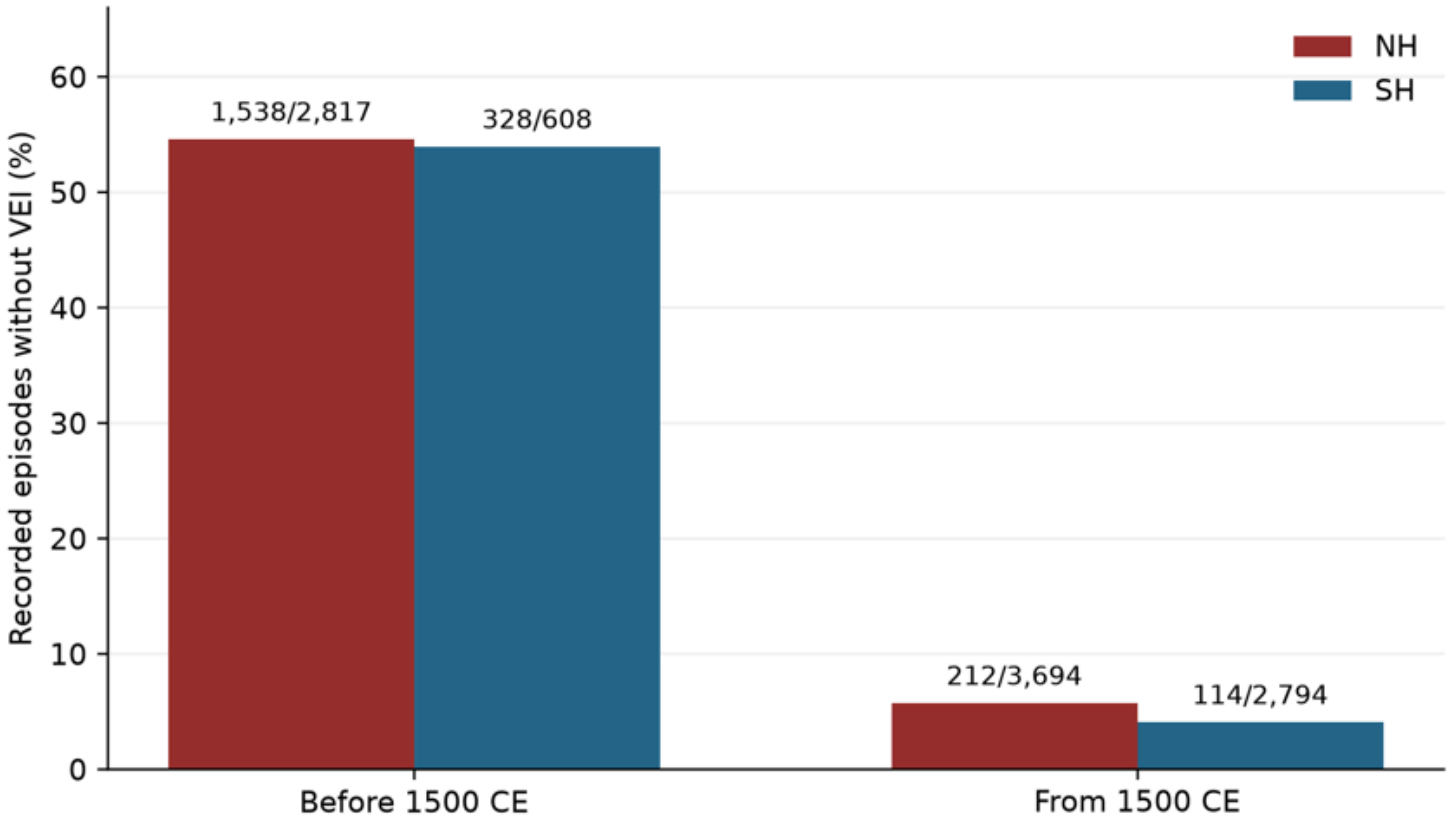


Figure 1

Missing VEI and temporal composition among recorded episodes. Bars give the percentage of included episodes without VEI in each hemisphere and period. Labels show missing and total records. Overall missingness reflects both the within-period percentages and each hemisphere's period composition

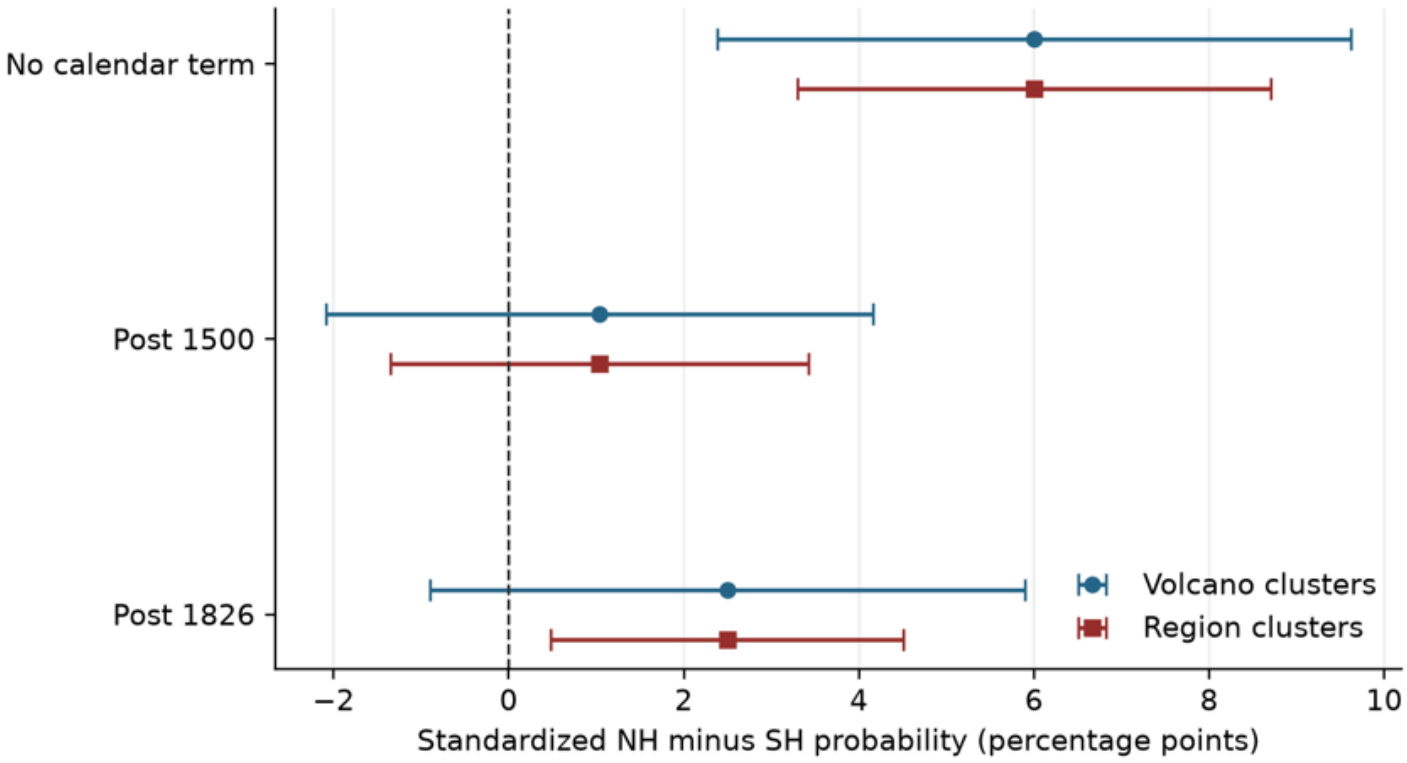


Figure 2

Model-standardized NH–SH probability differences. All three calendar specifications use the same 7,595 episodes and empirical covariate target. Intervals are 95% delta-method intervals using separate volcano- and region-cluster covariance estimates with t(G−1) reference. The dashed zero line denotes equal modeled recorded probabilities

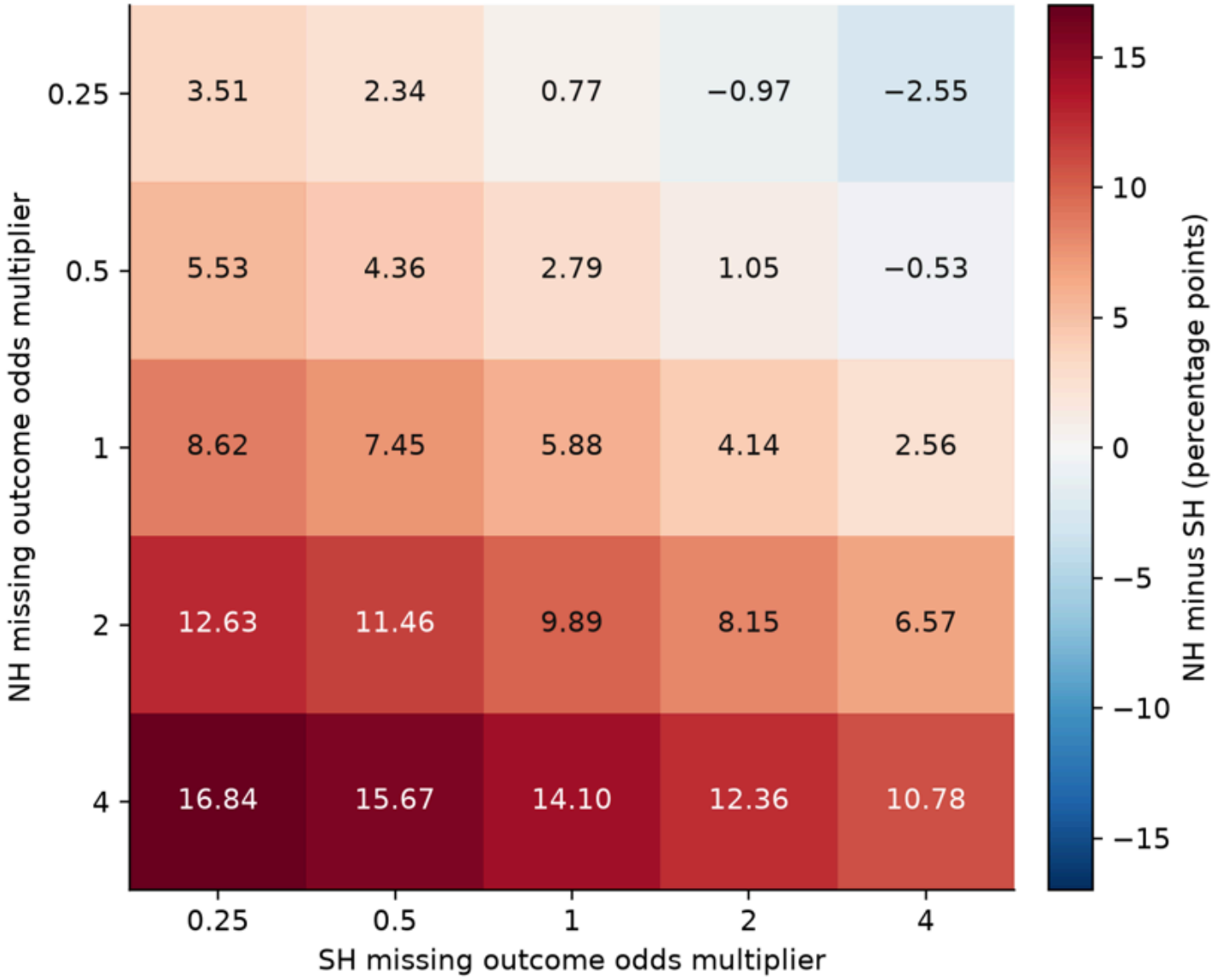


Figure 3

Sensitivity to missing VEI among included episodes. Cells give the NH–SH difference in percentage points after shifting the fitted odds for missing outcomes by separate hemisphere-specific multipliers. Observed outcomes are retained. The central cell uses the observed-outcome model directly. The grid represents specified missing-attribute assumptions among recorded episodes

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource1.pdf
- OnlineResource2.zip
- OnlineResource3.zip