Exploring Ground Motion Databases with Benford's Law

A Jupyter Notebook Workflow Using NGA-West2 and SGMD

Fan Hu, Ph.D.

Postdoctoral Researcher, UC Berkeley – PEER

Supervisor: Prof. Khalid M. Mosalam

Why Benford's law for ground motion databases?

Motivation

- Simulated ground motions are increasingly important for engineering applications.
- A lightweight way to compare simulated and recorded datasets is still useful.
- Benford's law gives a simple view of scaling behavior through leading digits.
- This study stays at the intensity-measure level: PGA, PGV, and PGD.

Comparison frame

NGA-West2 **vs** **SGMD**

Recorded motions — Simulated motions

Key question

Do the leading-digit distributions of common intensity measures differ in a systematic way?

Objective: A first-pass screen implemented through a transparent Jupyter notebook workflow, rather than a full validation framework.

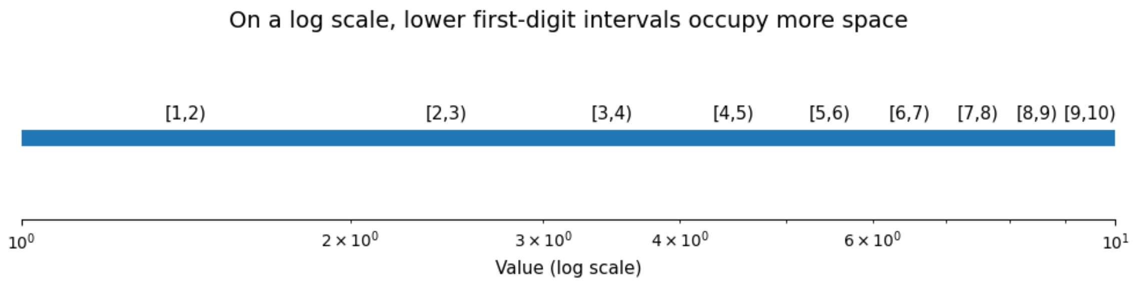
Building on prior PEER work showing that recorded ground-motion metadata follow a modified Benford distribution (Mosalam et al., unpublished manuscript), this study extends the same lens to SGMD.

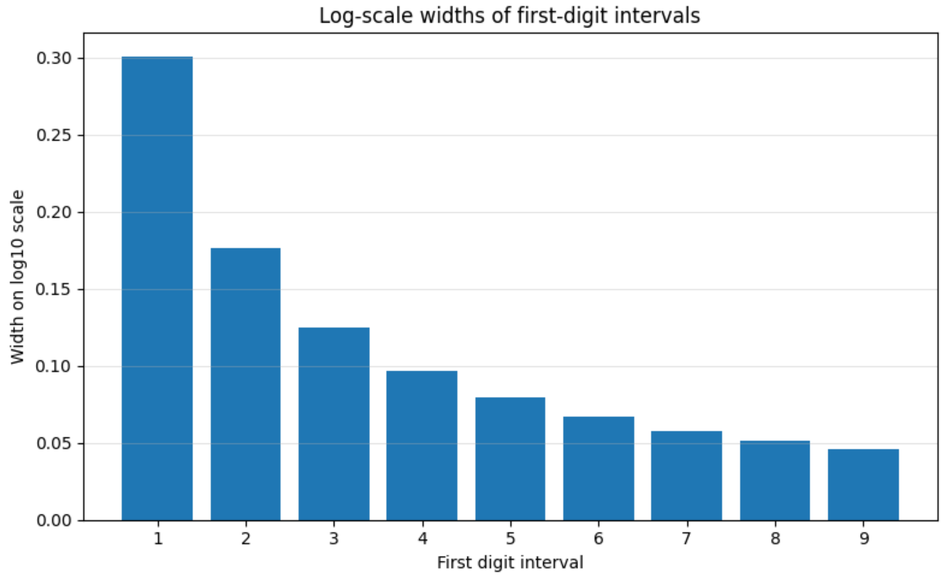
Benford’s law: background and intuition

Why are first digits not always uniform?

The story

- **1881:** Simon Newcomb noticed that early pages of logarithm tables were more worn.
- **1938:** Frank Benford tested many real-world datasets and formalized the pattern.
- **Since then:** Benford’s law has been used as a screening tool in accounting, fraud detection, economic data, and scientific datasets.
- **Key idea:** leading digits are related to multiplicative, not additive, changes.
- **Key caveat:** it is most meaningful when data span a broad dynamic range.





On a log scale, lower first-digit intervals occupy more space, so digit 1 appears more often than digit 9.

Benford metric and Jupyter workflow

Benford's law

$$P(d) = log_{10}\left(1 + \frac{1}{d}\right)$$

For many naturally occurring datasets that span multiple orders of magnitude, digit 1 is expected most often and digit 9 least often.

- Use the leading non-zero digit to handle values smaller than 1. E.g., 0.0035 → 3.
- Compute observed digit frequencies for each intensity measure.
- Quantify deviation with RMSE versus the Benford distribution.

Jupyter workflow

Read flatfiles — NGA-West2 and SGMD analysis-ready tables

Align IMs — PGA, PGV, and PGD with common units

Extract digits — Leading non-zero digits only; zeros and negatives excluded

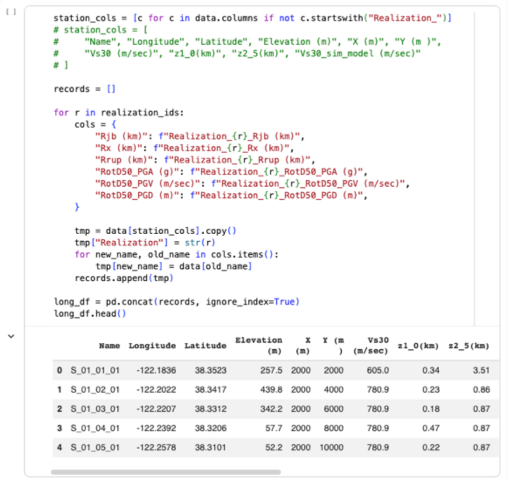
Compare distributions — Against Benford and across datasets

Export figures and RMSE — A transparent starting point that can later be extended to **more IMs** or **future SGMD releases**.

Jupyter notebook workflow for this analysis

1 SGMD Benford Analysis

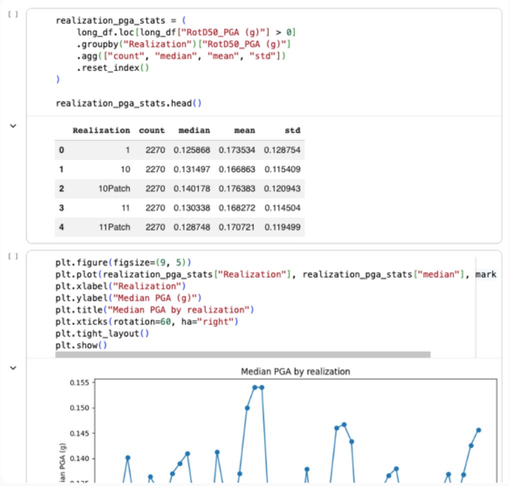
Read SGMD flatfile; clean headers; extract IMs.



1_sgmd_benford.ipynb

2 SGMD PGA diagnostics

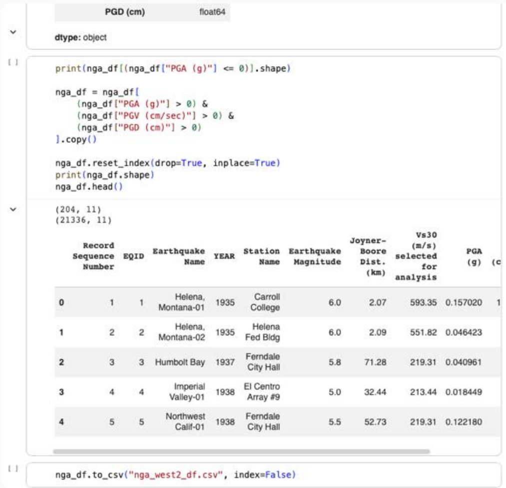
Inspect raw/log PGA distributions and realization behavior.



2_sgmd_pga_distribution.ipynb

3 NGA-West2 reference

Prepare recorded-motion IMs with consistent units.



3_nga_west2_benford.ipynb

4 SGMD vs NGA comparison

Align datasets; compute RMSE; export figures.

4_sgmd_vs_nga_comparison.ipynb

Takeaway: the Benford comparison is implemented as reusable notebooks, so researchers can start from the SGMD flatfile and adapt the workflow.

Datasets

SGMD

- Current SGMD release: Bay Area region only, provided through one metadata flatfile: *Hayward_M7_GMs_Flatfile.xlsx*
- Notebook 1 converts the raw flatfile into an analysis-ready site-realization table.
- First pass uses 2,270 sites × 50 realizations = **113.5k** site-realization samples.
- Comparable IMs: RotD50 PGA, RotD50 PGV, and RotD50 PGD.

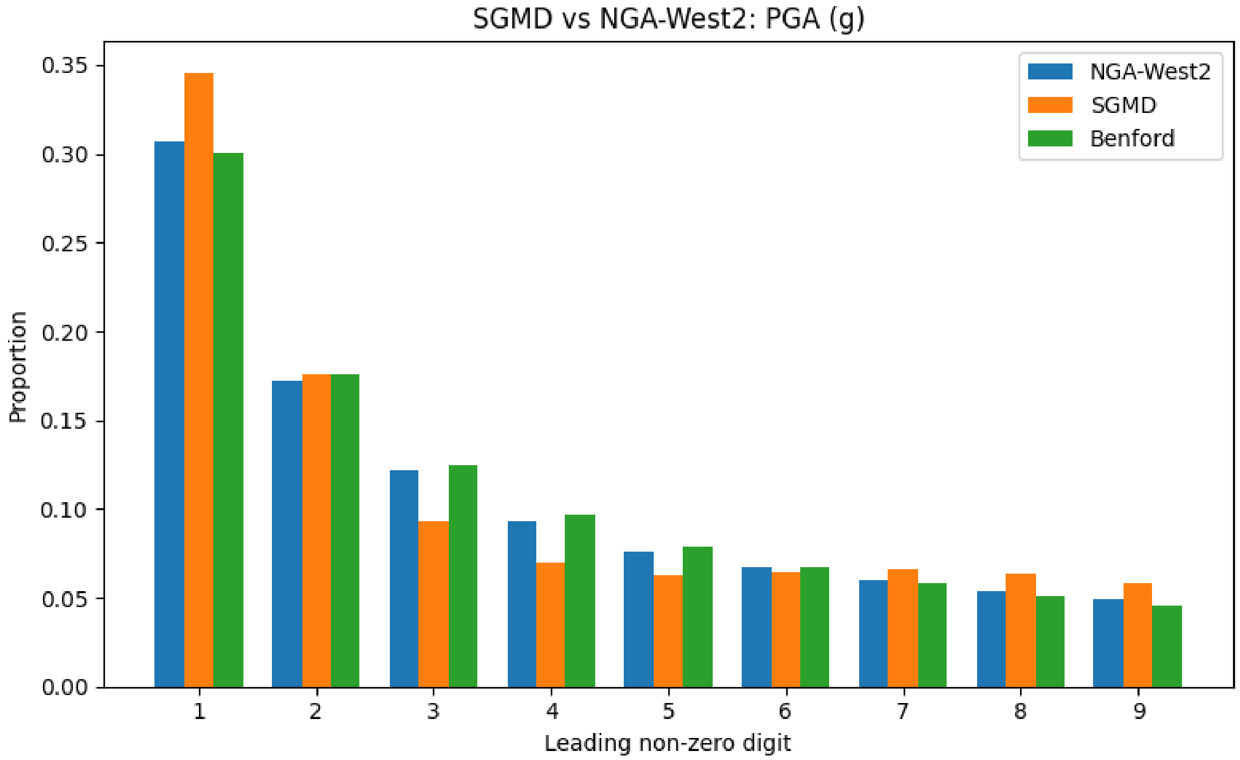
NGA-West2

- Notebook 3 processes the NGA-West2 flatfile and saves an analysis-ready reference table.
- **21,336** recorded ground motions are used in this first-pass comparison.
- The same three IMs were retained: PGA (g), PGV (cm/sec), and PGD (cm).
- Consistent units: PGA in g for both datasets; SGMD PGV converted from m/s to cm/sec; SGMD PGD converted from m to cm.

Caveat: SGMD here represents a single-scenario, many-realization simulated dataset, whereas NGA-West2 is a broad recorded-motion database. Part of the difference may reflect narrower scenario structure, not only simulation bias.

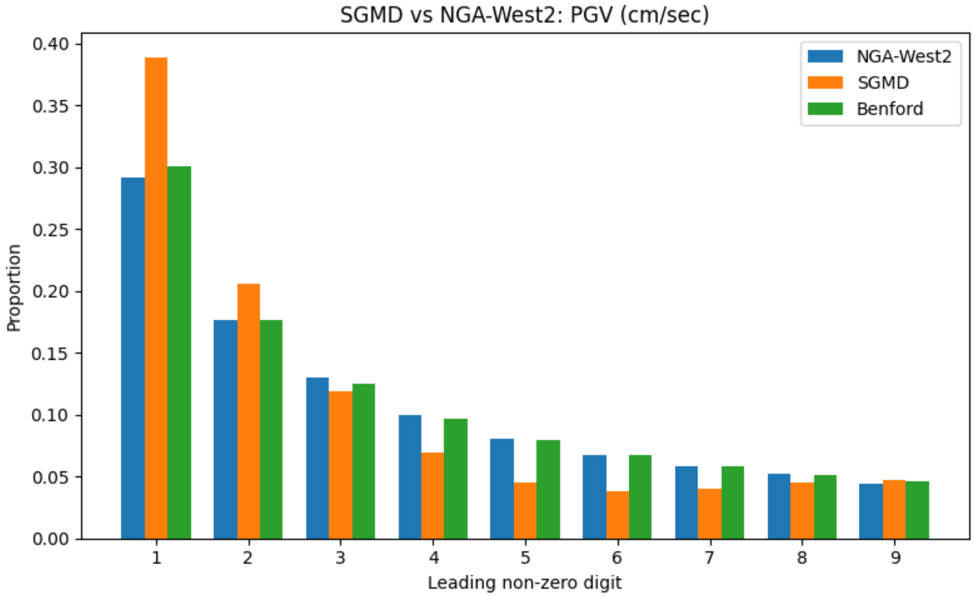
Main comparison: SGMD vs NGA-West2

PGA digit distribution

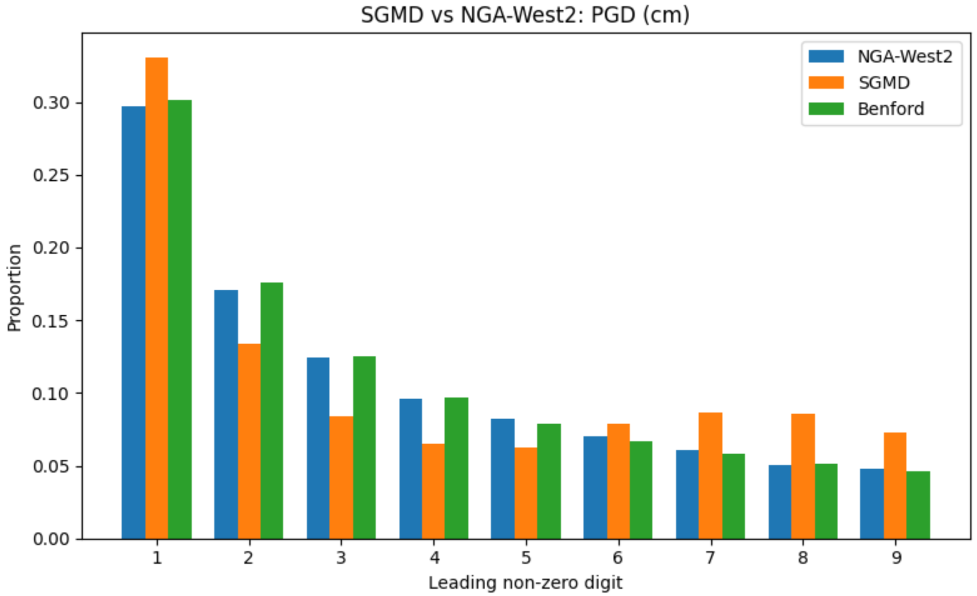


- Compared with NGA-West2, SGMD shows stronger concentration at digit 1 and lower frequencies through the middle digits.
- In this Benford lens, the current single-scenario SGMD dataset appears more constrained in its digit-scaling behavior.

Main comparison: SGMD vs NGA-West2



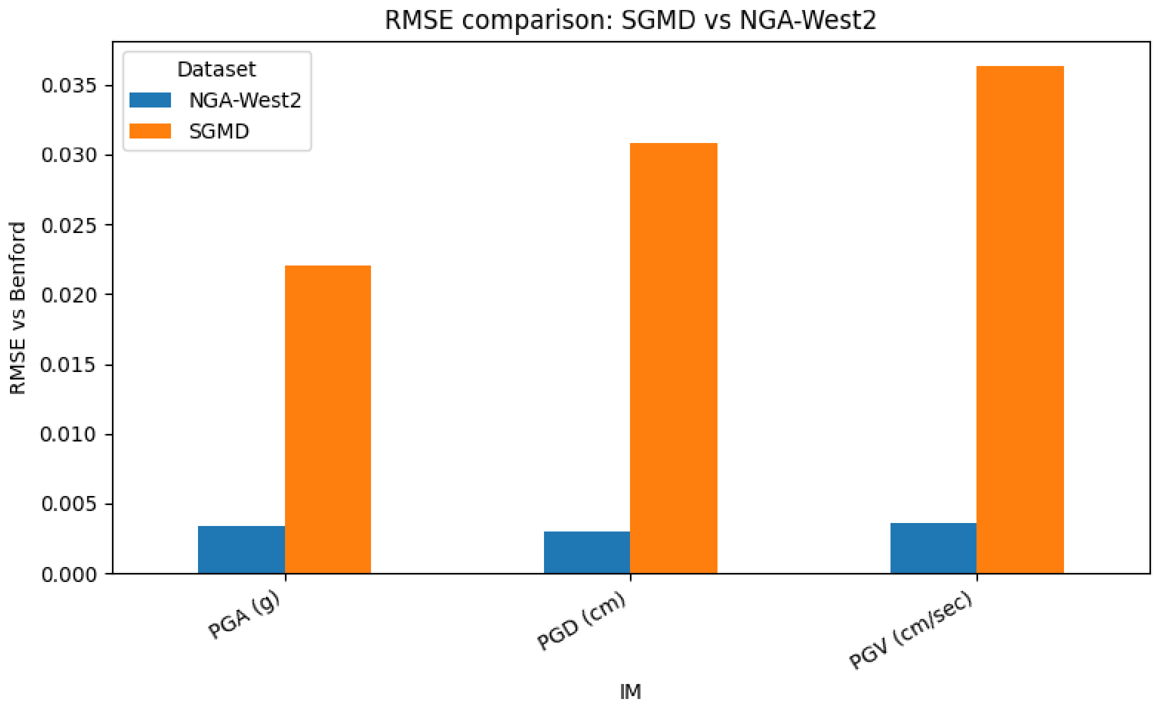
PGV



PGD

PGV and PGD reinforce the same qualitative separation seen in PGA, but more strongly.

Main comparison: SGMD vs NGA-West2



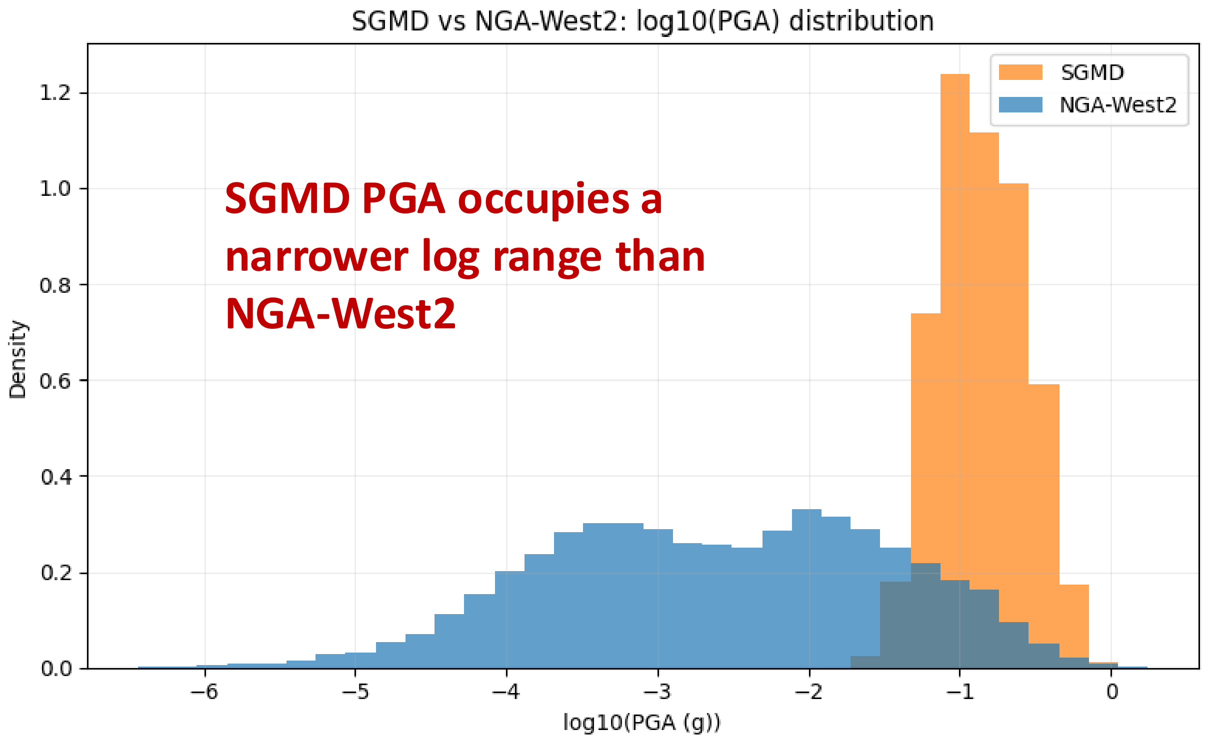
$$\mathrm{RMSE} = \sqrt{\frac{1}{9}\sum_{d=1}^{9}\left(P_{\mathrm{obs}}(d) - P_{\mathrm{Benford}}(d)\right)^2}$$

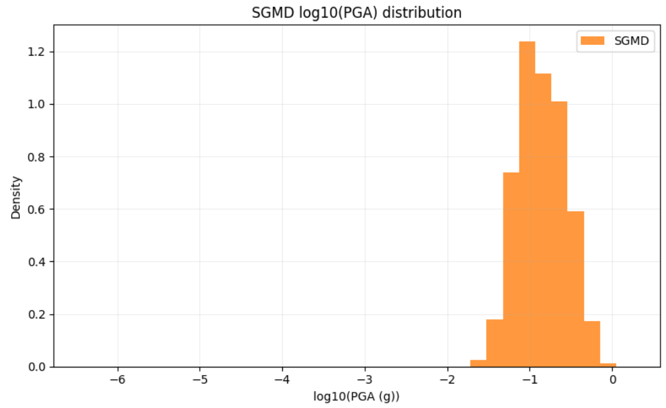
$P_{\mathrm{obs}}(d)$ is the observed proportion for leading digit d; $P_{\mathrm{Benford}}(d) = \log_{10}(1 + \frac{1}{d})$.

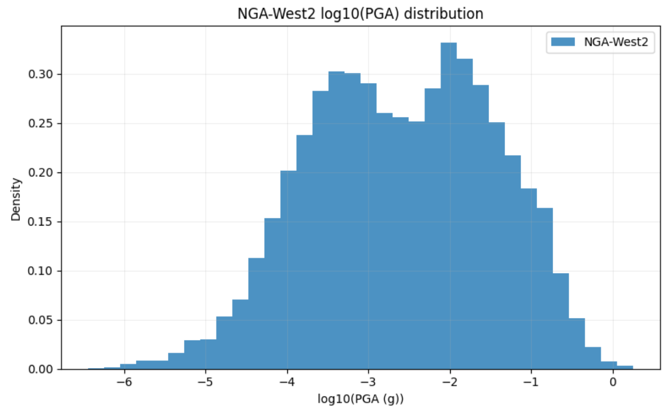
- For all three IMs, NGA-West2 RMSE is only about 0.003-0.004.
- SGMD RMSE is much larger: ~0.022 for PGA, ~0.031 for PGD, and ~0.036 for PGV.
- In this first pass, the SGMD deviation is largest for PGV, followed by PGD and PGA.

Recorded motions are much closer to the Benford pattern than the single-scenario simulated dataset in this first-pass comparison.

Why does SGMD show larger Benford deviation?



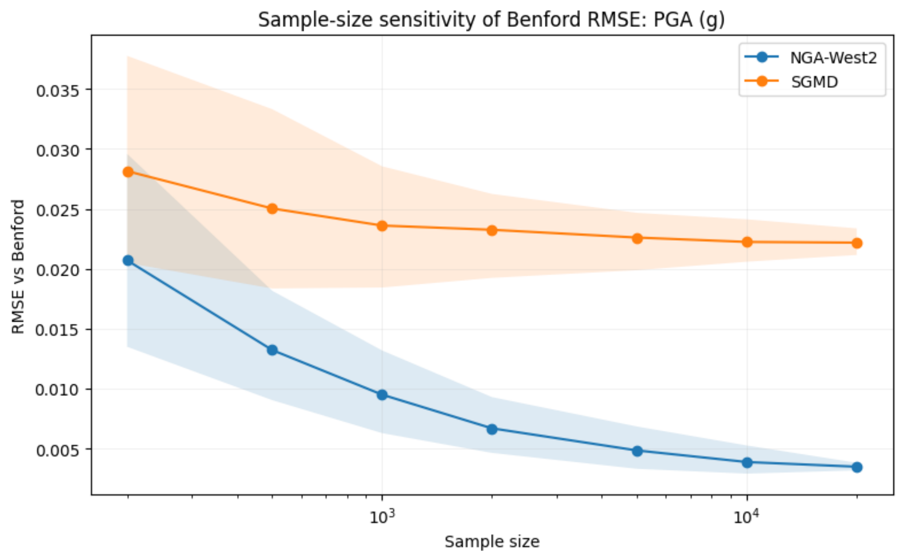




- SGMD log10(PGA) values are concentrated in a relatively narrow range.
- NGA-West2 spans a much broader dynamic range.
- This narrower SGMD range is consistent with its stronger digit-1 concentration and larger Benford RMSE.

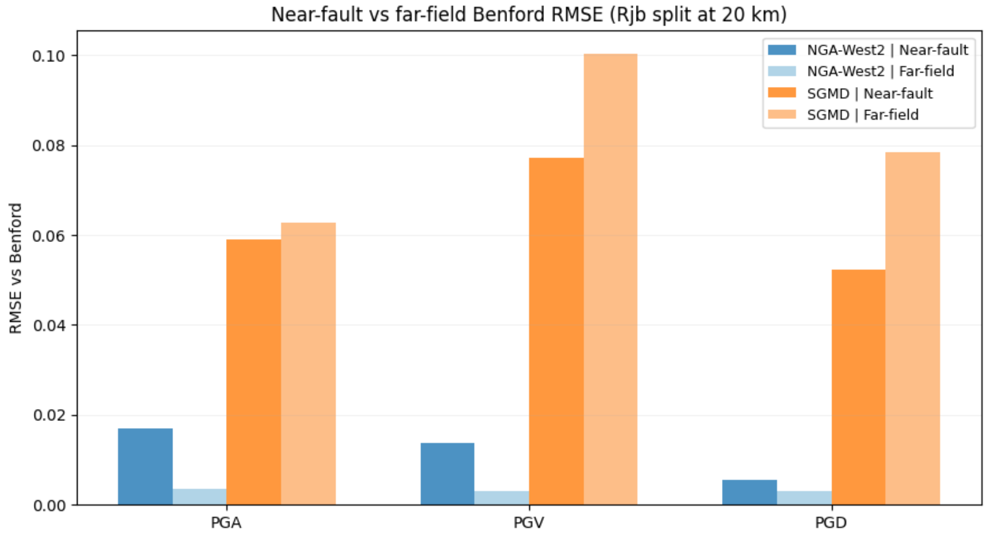
Preliminary robustness checks

Sample-size sensitivity (PGA)



Shaded bands indicate the 10^{th}– 90^{th} percentile range across 200 repeated random subsamples at each sample size.

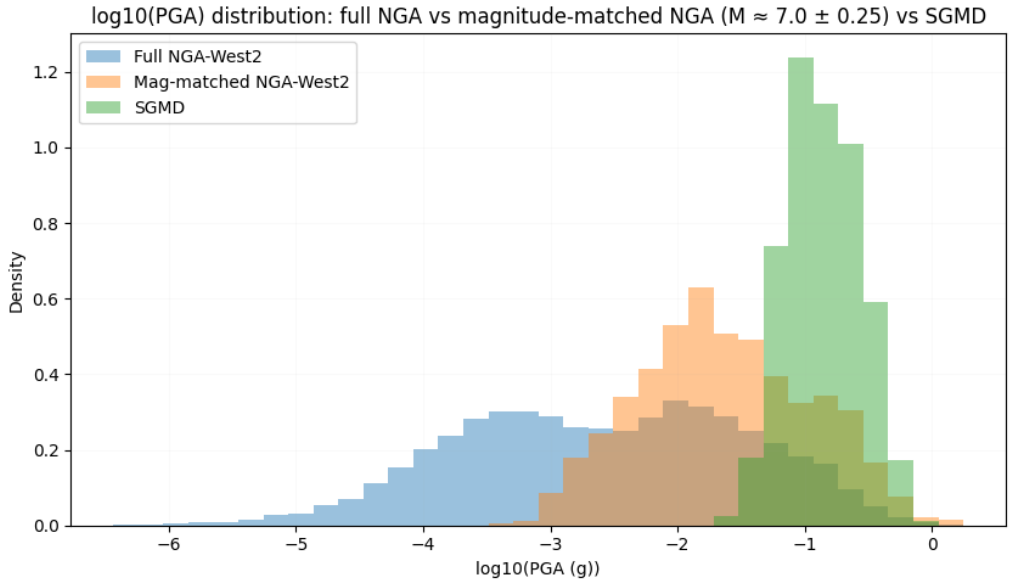
Distance split check ($R_{jb} \leq 20$ km vs > 20 km)

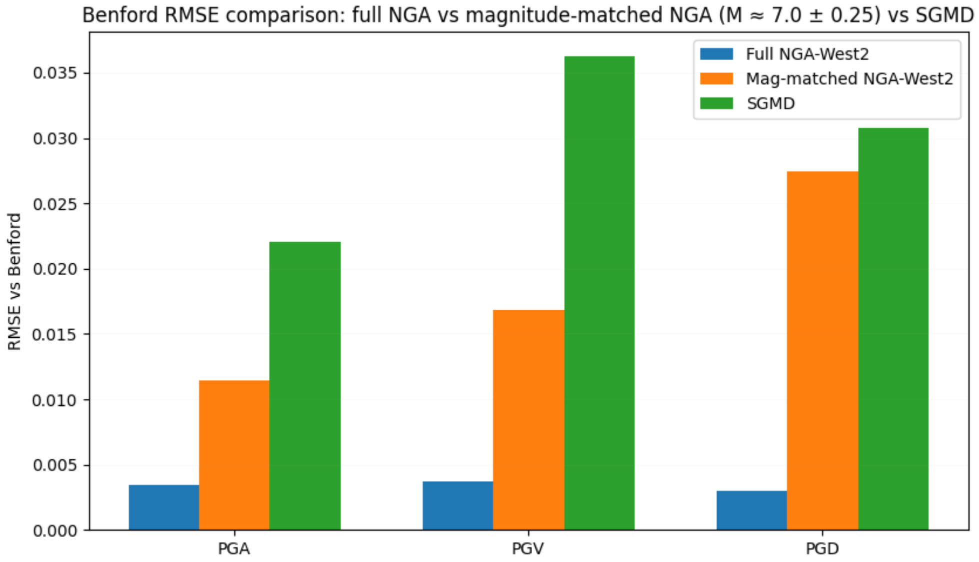


R_{jb} is the Joyner–Boore distance, defined as the shortest horizontal distance from a site to the surface projection of the fault rupture.

- Sample size affects stability, but does not explain the SGMD–NGA gap.
- Distance split affects RMSE, but does not remove the separation.
- These checks support the interpretation of a more constrained SGMD distribution.

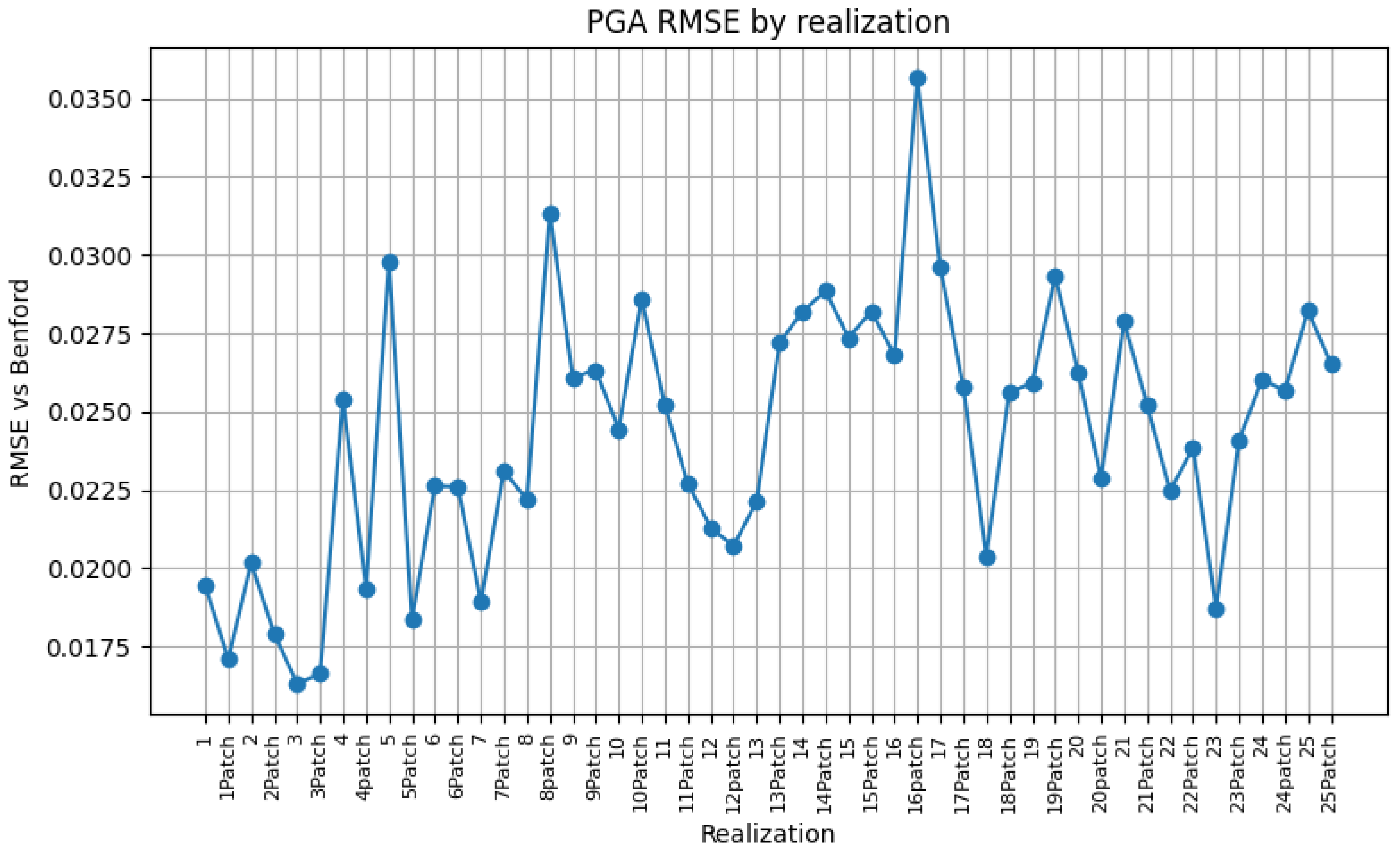
Preliminary check using NGA records with M7.0 ± 0.25





- Restricting NGA-West2 to M7.0 ± 0.25 leaves 1,765 records from 20 earthquakes.
- Magnitude matching narrows the NGA-West2 log10(PGA) distribution and increases its Benford RMSE.
- This suggests that part of the original SGMD–NGA separation reflects NGA-West2's broader magnitude mixture.
- However, SGMD remains more concentrated, so magnitude alone does not fully explain the difference.

Within SGMD: realization-level behavior



- Realization-to-realization variability is visible even within a single Hayward M7 scenario.
- Patch and no-patch averages are nearly identical for PGA RMSE.
- For this first-digit metric, patching is not the main driver of the observed deviation.

Mean RMSE:

- **no patch = 0.024435**
- **patch = 0.024106**

Interpretation, limits, and next steps

Interpretation

- The notebook workflow provides a reproducible way to use Benford’s law as a screening lens for ground-motion databases.
- SGMD reproduces the overall trend, but with systematic departures from the recorded reference.
- These results suggest a more constrained IM distribution for the single-scenario simulated dataset.

Limits

- Only the current Bay Area SGMD release was available.
- Benford works best for data spanning multiple orders of magnitude, whereas some seismic IMs may be more constrained.
- Benford is a screening lens, not a physical validation metric by itself.

Next steps

- Extend to more available IMs and spectral measures.
- Extend the preliminary split checks by distance, site condition, or hypocenter group.
- Adapt the notebooks to future SGMD releases and user-defined IMs or screening metrics.