

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

Recorded motions

vs

SGMD

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.

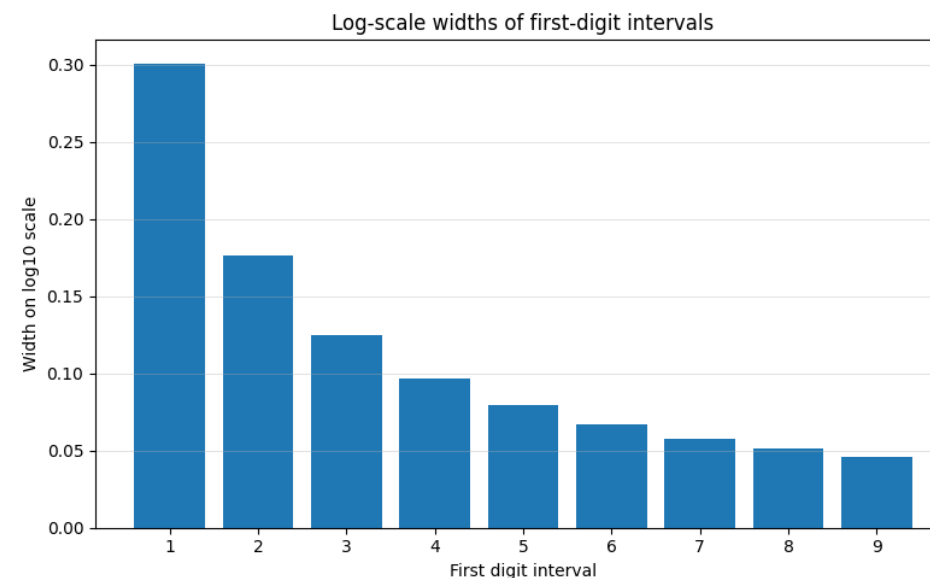
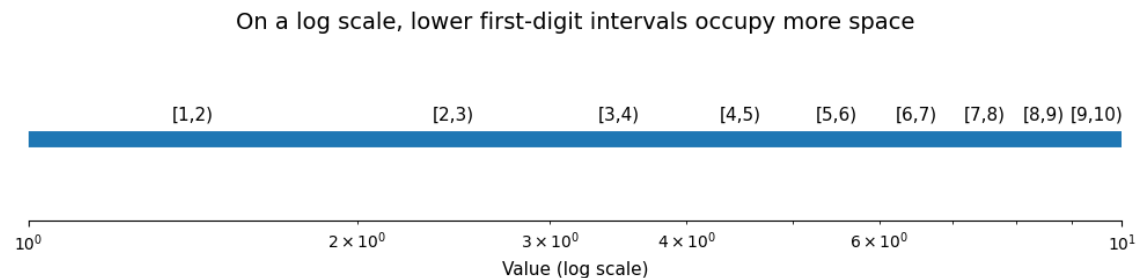
Building on prior PEER work showing that recorded ground-motion metadata follow a modified Benford distribution (Mosalam et al., unpublished manuscript), this study extends the same lens to SGMD.

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

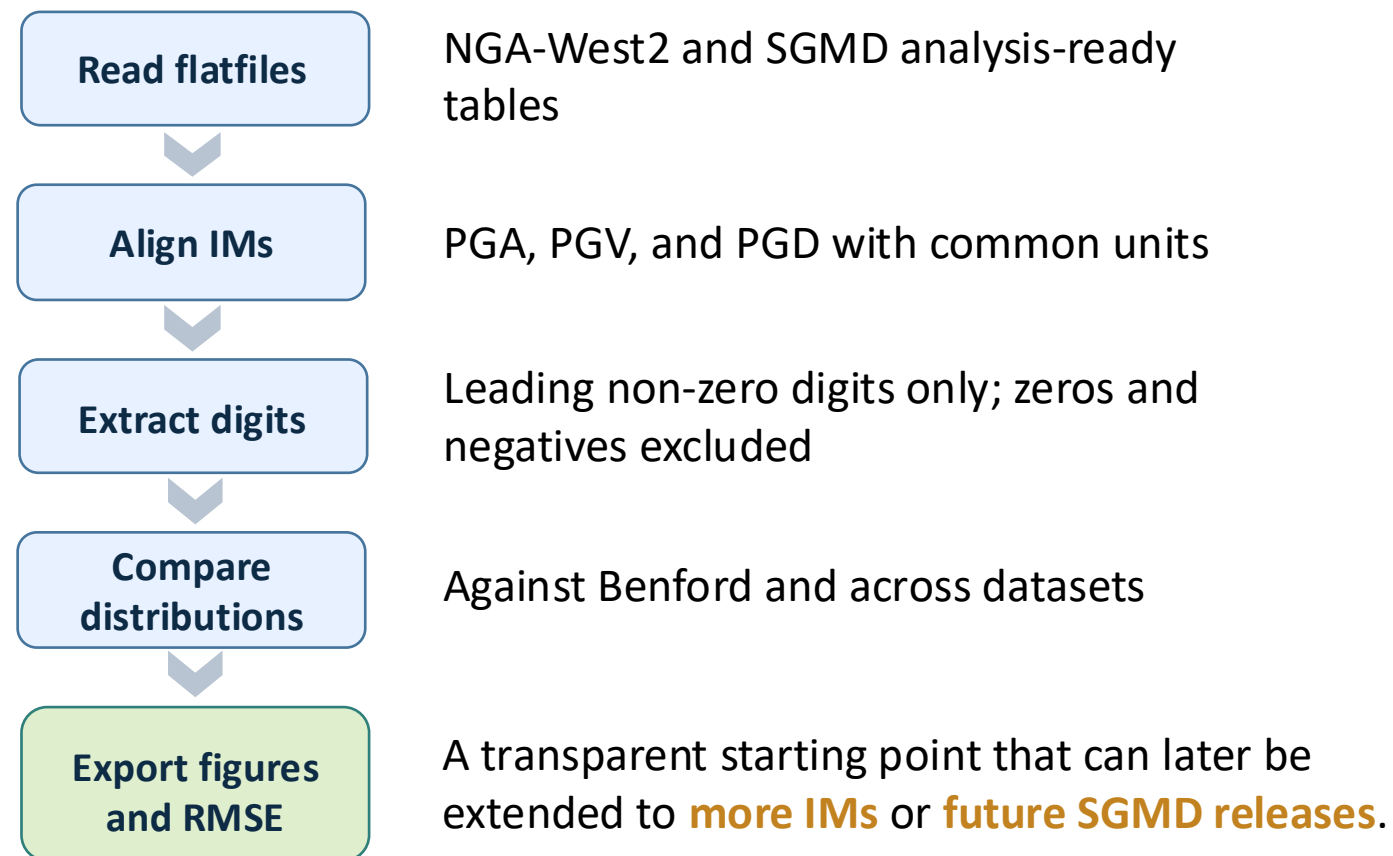
Benford's law

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

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



Jupyter notebook workflow for this analysis

1 SGMD Benford Analysis

Read SGMD flatfile; clean headers; extract IMs.

```
[ ] station_cols = [c for c in data.columns if not c.startswith("Realization_")]
# station_cols = [
#   "Name", "Longitude", "Latitude", "Elevation (m)", "X (m)", "Y (m)",
#   "Vs30 (m/sec)", "z1_0(km)", "z2_5(km)", "Vs30_sln_model (m/sec)"
# ]

records = []
for r in realization_ids:
    cols = {}
    "Rjb (km)": f"Realization_{r}_Rjb (km)",
    "Rx (km)": f"Realization_{r}_Rx (km)",
    "Rrup (km)": f"Realization_{r}_Rrup (km)",
    "RotD50_PGA (g)": f"Realization_{r}_RotD50_PGA (g)",
    "RotD50_PGV (m/sec)": f"Realization_{r}_RotD50_PGV (m/sec)",
    "RotD50_PGD (m)": f"Realization_{r}_RotD50_PGD (m)",
    }

    tmp = data[station_cols].copy()
    tmp["Realization"] = str(r)
    for new_name, old_name in cols.items():
        tmp[new_name] = data[old_name]
    records.append(tmp)

long_df = pd.concat(records, ignore_index=True)
long_df.head()
```

	Name	Longitude	Latitude	Elevation (m)	X (m)	Y (m)	Vs30 (m/sec)	z1_0(km)	z2_5(km)
0	S_01_01_01	-122.1836	38.3523	257.5	2000	2000	605.0	0.34	3.51
1	S_01_02_01	-122.2022	38.3417	439.8	2000	4000	780.9	0.23	0.86
2	S_01_03_01	-122.2207	38.3312	342.2	2000	6000	780.9	0.18	0.87
3	S_01_04_01	-122.2392	38.3206	57.7	2000	8000	780.9	0.47	0.87
4	S_01_05_01	-122.2578	38.3101	52.2	2000	10000	780.9	0.22	0.87

1_sgmd_benford.ipynb

2 SGMD PGA diagnostics

Inspect raw/log PGA distributions and realization behavior.

```
[ ] realization_pga_stats = (
    long_df.loc[long_df["RotD50_PGA (g)"] > 0]
    .groupby("Realization")["RotD50_PGA (g)"]
    .agg(["count", "median", "mean", "std"])
    .reset_index()

realization_pga_stats.head()
```

	Realization	count	median	mean	std
0	1	2270	0.125868	0.173534	0.128754
1	10	2270	0.131497	0.166863	0.115409
2	10Patch	2270	0.140178	0.176383	0.120943
3	11	2270	0.130338	0.168272	0.114504
4	11Patch	2270	0.128748	0.170721	0.119499

```
[ ] plt.figure(figsize=(9, 5))
plt.plot(realization_pga_stats["Realization"], realization_pga_stats["median"], mark
plt.xlabel("Realization")
plt.ylabel("Median PGA (g)")
plt.title("Median PGA by realization")
plt.xticks(rotation=60, ha="right")
plt.tight_layout()
plt.show()
```

2_sgmd_pga_distribution.ipynb

3 NGA-West2 reference

Prepare recorded-motion IMs with consistent units.

```
PGA (cm) float64
dtype: object

[ ] print(nga_df[(nga_df["PGA (g)"] <= 0)].shape)

nga_df = nga_df[
    (nga_df["PGA (g)"] > 0) &
    (nga_df["PGV (cm/sec)"] > 0) &
    (nga_df["PGD (cm)"] > 0)
].copy()

nga_df.reset_index(drop=True, inplace=True)
print(nga_df.shape)
nga_df.head()
```

Record Sequence Number	SG12	Earthquake Name	YEAR	Station Name	Earthquake Magnitude	Joyner-Boore Dist. (km)	Vs30 (m/s) selected for analysis	PGA (g)
0	1	Helena, Montana-01	1935	Carroll College	6.0	2.07	593.35	0.157020
1	2	Helena, Montana-02	1935	Helena Fed Bldg	6.0	2.09	551.82	0.046423
2	3	Humboldt Bay	1937	Ferndale City Hall	5.8	71.28	219.31	0.040961
3	4	Imperial Valley-01	1938	El Centro Army Hq	5.0	32.44	213.44	0.018449
4	5	Northwest Calif-01	1938	Ferndale City Hall	5.5	52.73	219.31	0.122180

```
[ ] nga_df.to_csv("nga_west2_df.csv", index=False)
```

3_nga_west2_benford.ipynb

4 SGMD vs NGA comparison

Align datasets; compute RMSE; export figures.

```
[ ] pivot_rmse = summary.pivot(index="IM", columns="Dataset", values="RMSE_vs_Benford")
pivot_rmse
```

IM	Dataset	RMSE_vs_Benford
PGA (g)	SGMD	0.003436
PGA (g)	NGA-West2	0.022089
PGD (cm)	SGMD	0.002978
PGD (cm)	NGA-West2	0.030796
PGV (cm/sec)	SGMD	0.003668
PGV (cm/sec)	NGA-West2	0.038290

```
[ ] pivot_rmse.plot(kind="bar", figsize=(8, 5))
plt.ylabel("RMSE vs Benford")
plt.title("RMSE comparison: SGMD vs NGA-West2")
plt.xticks(rotation=30, ha="right")
plt.tight_layout()
plt.show()
```

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.

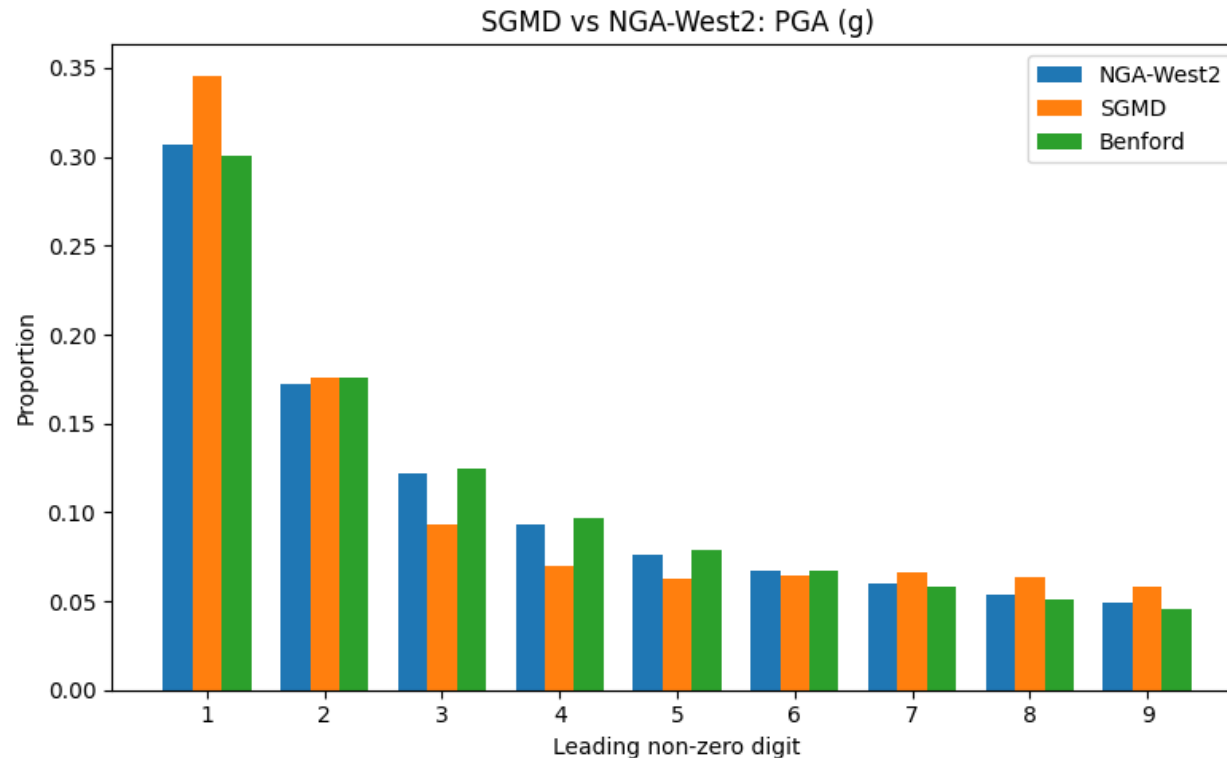
NGA-West2

- 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.

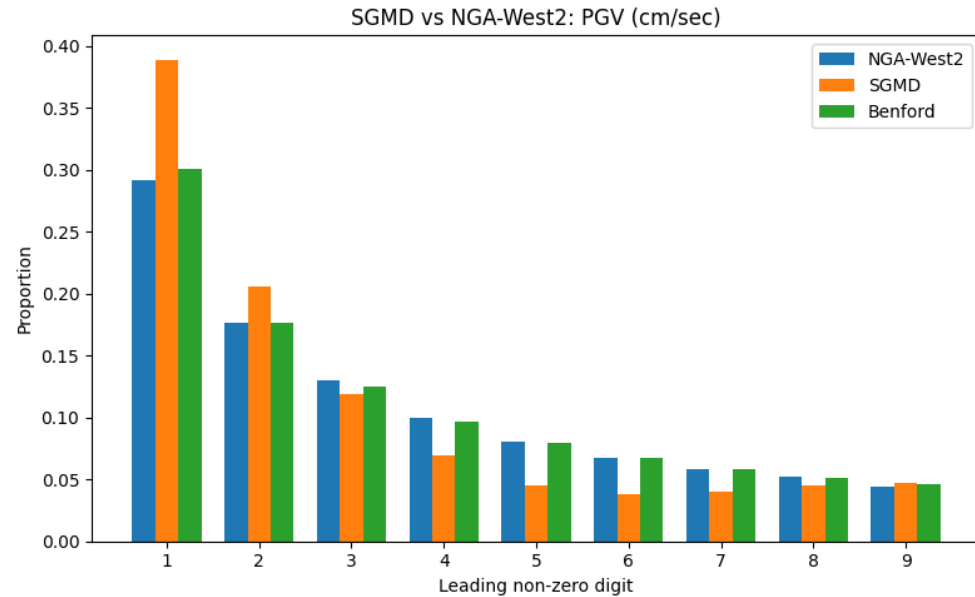
Main comparison: SGMD vs NGA-West2

PGA digit distribution

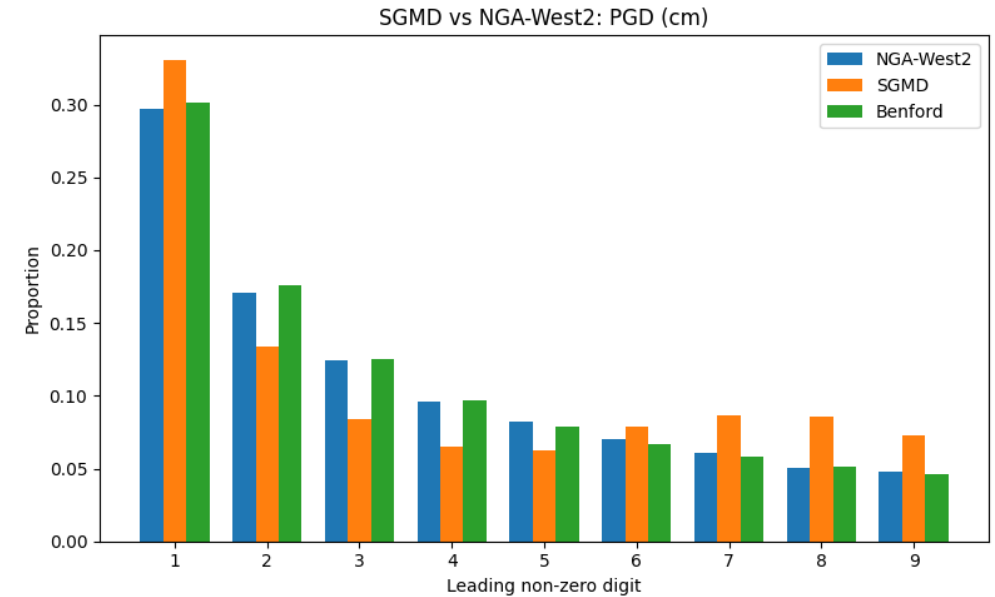


- Compared with NGA-West2, SGMD shows stronger concentration at digit 1 and lower frequencies through the middle digits.
- In this Benford lens, the current single-scenario SGMD dataset appears more constrained in its digit-scaling behavior.

Main comparison: SGMD vs NGA-West2



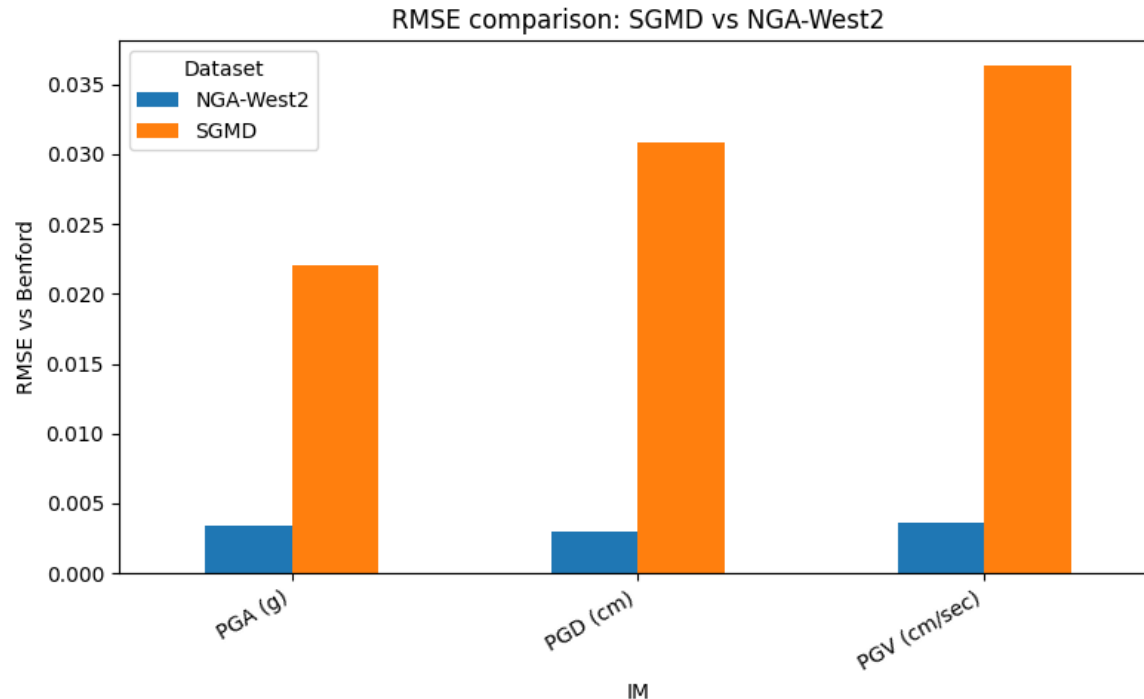
PGV



PGD

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

Main comparison: SGMD vs NGA-West2



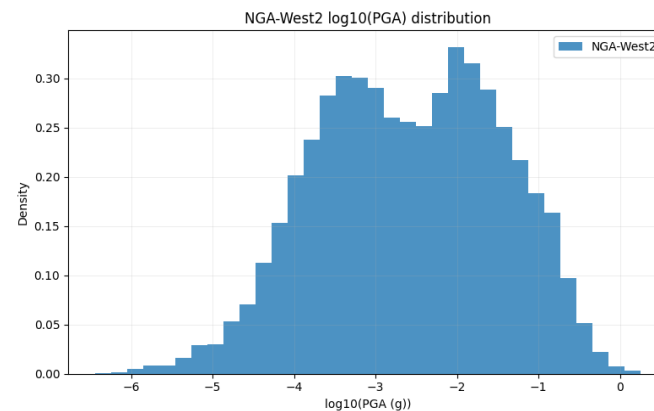
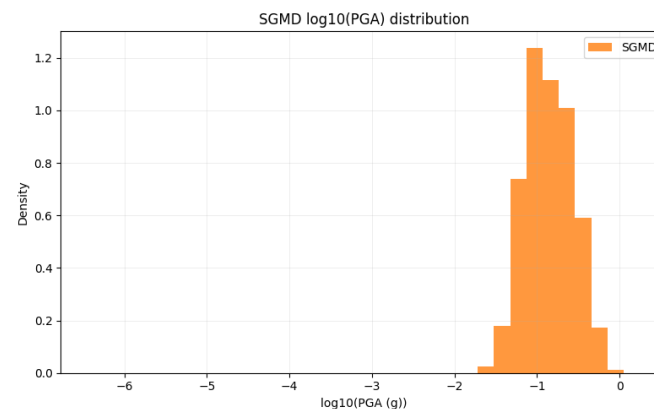
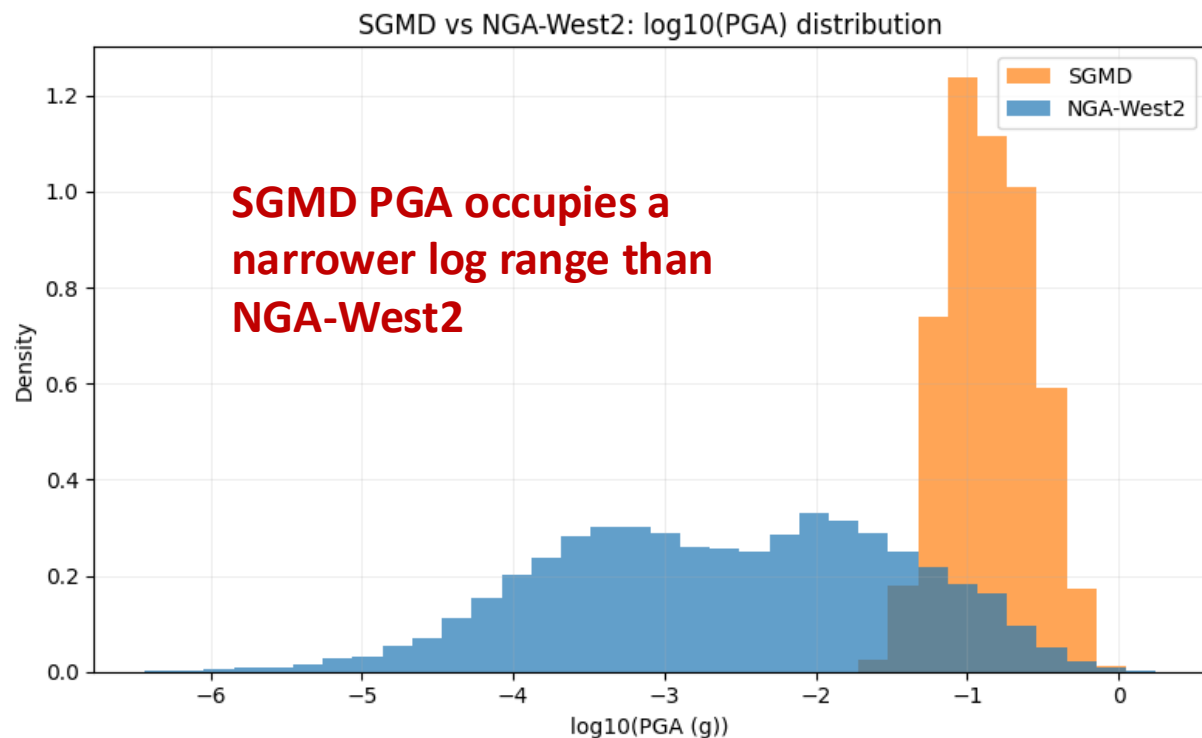
- For all three IMs, NGA-West2 RMSE is only about 0.003-0.004.
- SGMD RMSE is much larger: ~0.022 for PGA, ~0.031 for PGD, and ~0.036 for PGV.
- In this first pass, the SGMD deviation is largest for PGV, followed by PGD and PGA.

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

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

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

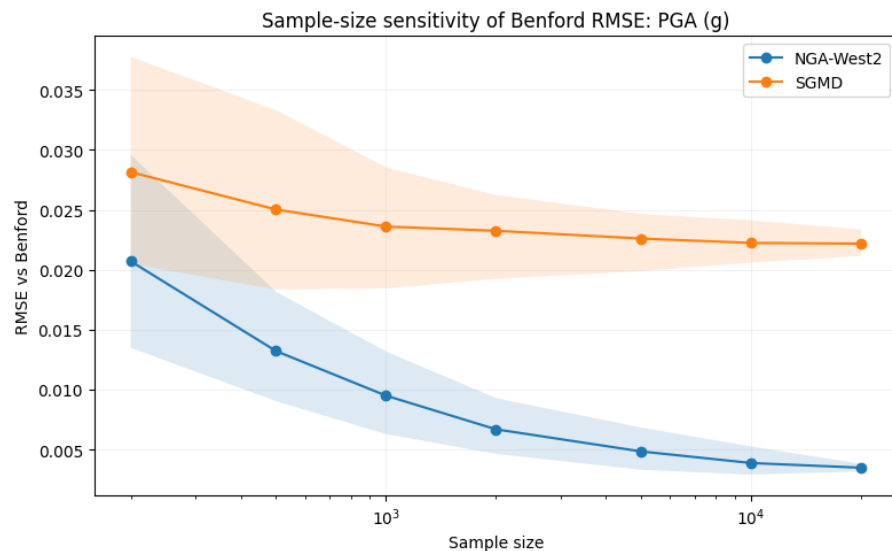
Why does SGMD show larger Benford deviation?



- 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.

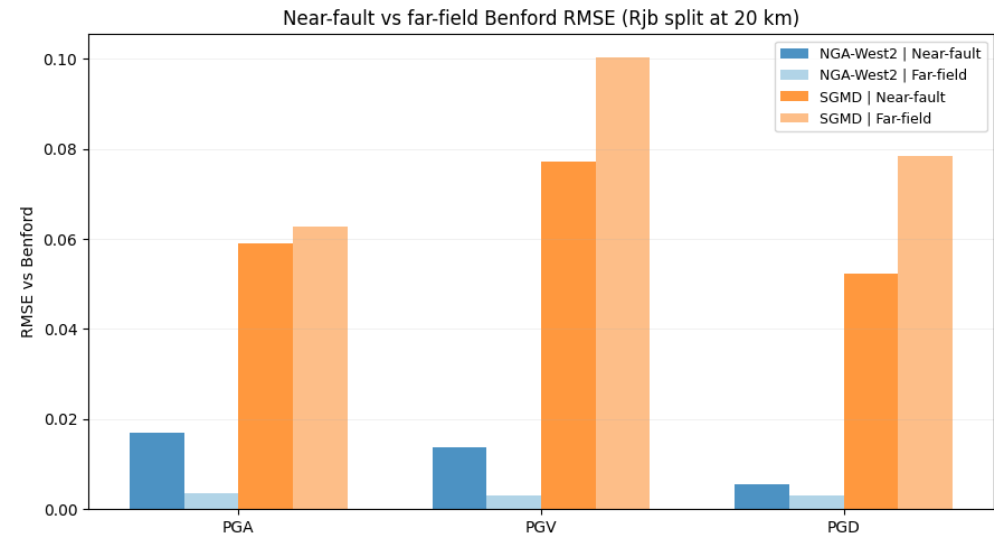
Preliminary robustness checks

Sample-size sensitivity (PGA)



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

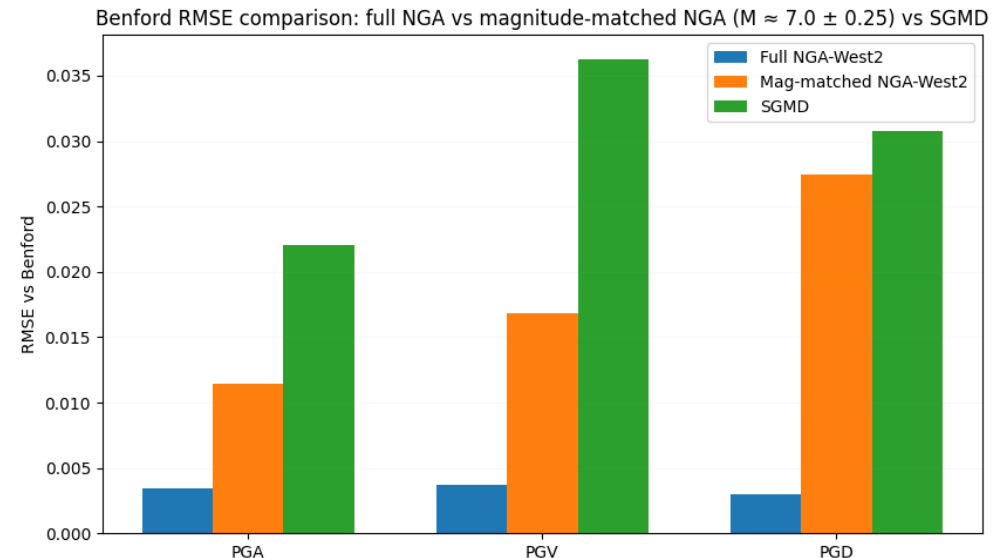
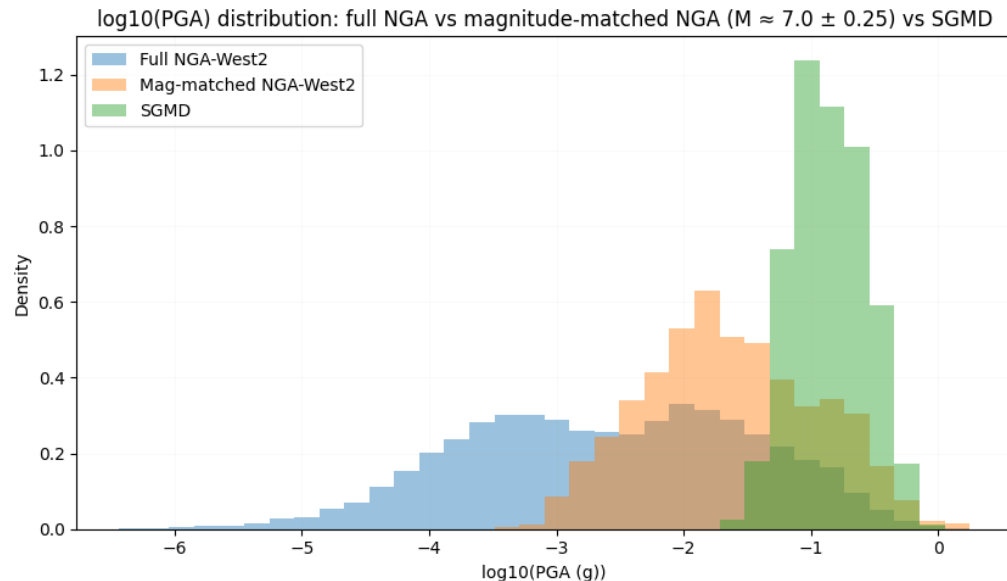
Distance split check ($R_{jb} \leq 20$ km vs > 20 km)



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

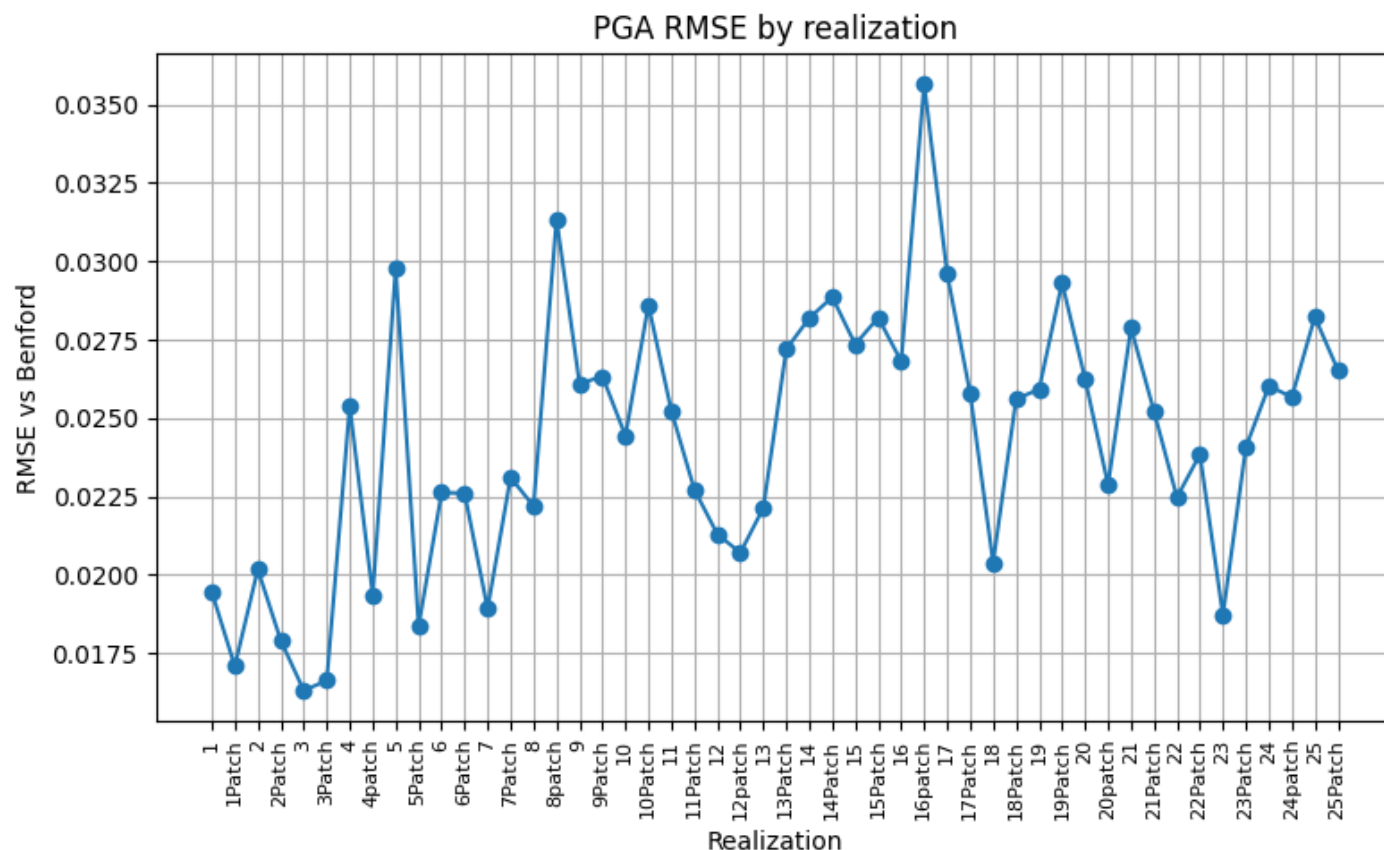
- Sample size affects stability, but does not explain the SGMD–NGA gap.
- Distance split affects RMSE, but does not remove the separation.
- These checks support the interpretation of a more constrained SGMD distribution.

Preliminary check using NGA records with $M7.0 \pm 0.25$



- Restricting NGA-West2 to $M7.0 \pm 0.25$ leaves 1,765 records from 20 earthquakes.
- Magnitude matching narrows the NGA-West2 $\log_{10}(\text{PGA})$ 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.