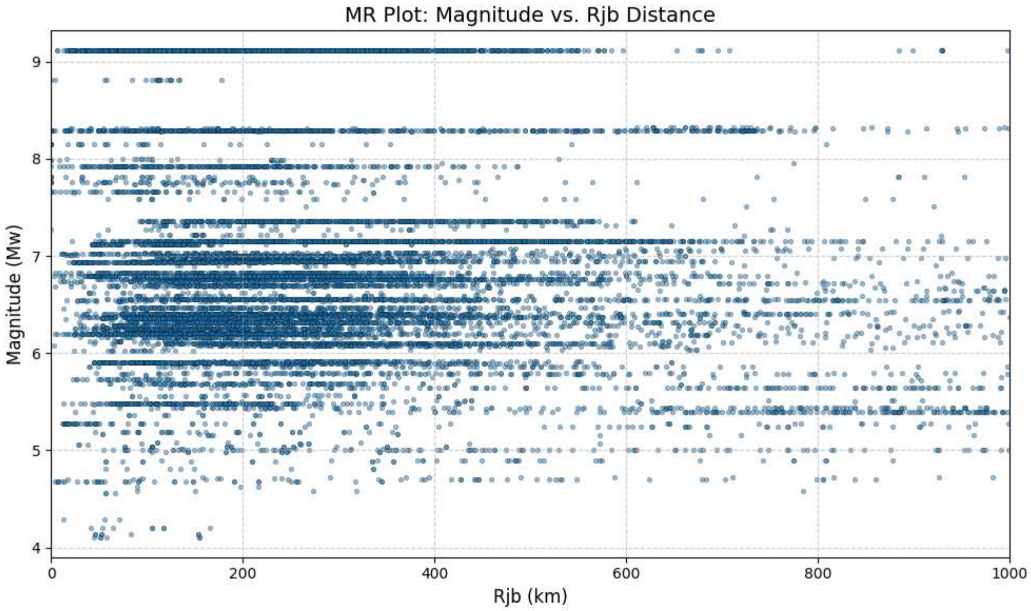
Prediction of Spectral Acceleration Using Random Forest

1. Introduction

This study develops a Random Forest model to predict 20 spectral acceleration (SA) values based on five input ground motion features: magnitude (mag), rupture distance (rjb), logrjb , logvs30 , and event type (inter-intra). The model includes a careful preprocessing pipeline, model training with early stopping, residual decomposition using mixed-effects modeling, Residual analysis,Ground motion physics,Importance,SHAP analysis for explainability.

2.Magnitude vs Rjb Scatter Plot:

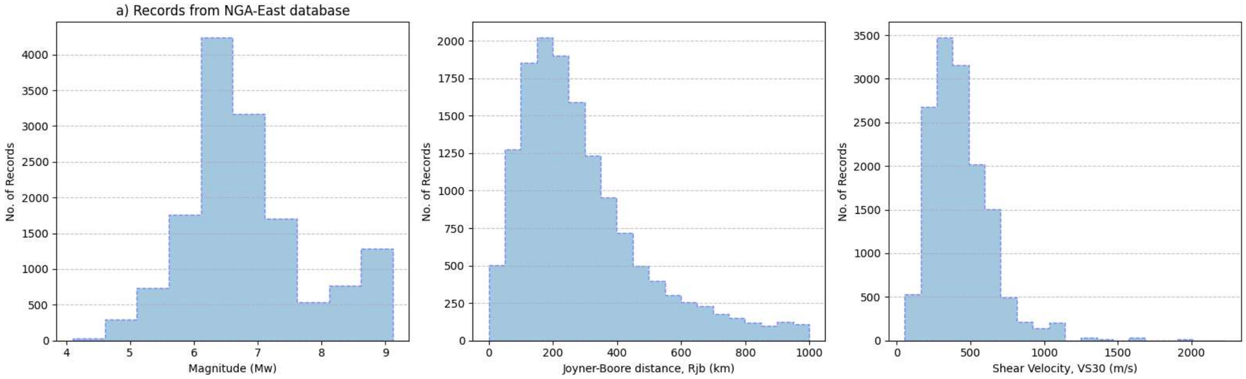
This scatter plot visualizes the distribution of events across different magnitude (mag) and Joyner-Boore distance (rjb) combinations in the dataset used for training and evaluation.



- The plot shows a dense cluster of data points for **moderate magnitudes (5.0–6.5)** and **short-to-moderate distances (0–100 km)**, which is typical of recorded ground motion datasets like NGA.
- Fewer data points appear at **larger distances (>200 km)** or for **larger magnitudes (>7.0)**, consistent with the relative rarity of such records.
- The coverage ensures that the model is well-trained across the critical near-field range but may have increased uncertainty for predictions at far distances or large magnitudes due to data sparsity.

3.Histograms of Input Features:

This figure presents histograms of three key input parameters—Moment Magnitude (Mw), Joyner-Boore distance (Rjb), and Shear-wave velocity at 30 m depth (Vs30)—from the NGA-East database used in this study.



- **Magnitude (Mw)** is concentrated around 6.0–6.5, reflecting a dataset dominated by moderate earthquakes.
- **Rjb** is right-skewed, with most recordings within 0–300 km, ensuring good coverage of near-field motions.
- **Vs30** peaks around 300–500 m/s, indicating a prevalence of stiff soil and soft rock sites in the data.

4.Summary Statistics of Input and Output:

Input Parameters:

Parameter	mag	rjb	logrjb	logvs30	intra_inter
min	4.1	0.01	-2	1.7243	0
max	9.12	999.0898	2.9996	3.3483	1
mean	6.8318	289.7475	2.352	2.5906	0.4232
std	1.0028	196.9747	0.3695	0.2032	0.4941
skewness	0.7859	1.2926	-3.3307	-0.087	0.3107
kurtosis	0.3906	1.535	33.8885	0.1169	-1.9035

- **Magnitude (mag):** Ranges from 4.1 to 9.12, with a mean of 6.83, showing variability in seismic event intensity. Slight positive skew (0.79) and near-normal distribution.
- **Rupture Distance (rjb):** Varies widely from 0.01 to 999.09, with a mean of 289.75, showing high variability and positive skew (1.29).
- **Log of Rupture Distance (logrjb):** Range from -2.00 to 2.99, mean of 2.35, with a highly negative skew (-3.33) and heavy-tailed distribution (high kurtosis).
- **Log of Shear-Wave Velocity (logvs30):** Ranges from 1.72 to 3.35, with a mean of 2.59, close to normal distribution.

- **Intra-Inter Event Flag (intra_inter):** Ranges from 0.00 to 1.00, with a mean of 0.42, indicating mixed intra- and inter-event data, with light tails in distribution

Output Parameters:

Parameter	T0pt010S	T0pt020S	T0pt030S	T0pt050S	T0pt075S	T0pt100S	T0pt150S	T0pt200S	T0pt300S	T0pt400S	T0pt500S	T0pt750S	T1pt000S	T1pt500S	T2pt000S	T2pt500S	T3pt000S	T3pt500S	T4pt000S	T5pt000S
min	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
max	2.5801	2.7391	3.5567	4.9801	5.9791	3.6631	5.8752	6.2565	5.252	4.234	3.0608	2.27	1.2481	1.3501	1.2663	0.6708	0.3824	0.3857	0.2931	0.2265
mean	0.0304	0.0311	0.033	0.0396	0.0499	0.0608	0.0715	0.0738	0.0678	0.0591	0.0515	0.0382	0.03	0.0198	0.0143	0.0108	0.0084	0.0068	0.0055	0.0039
std	0.085	0.0884	0.099	0.128	0.1542	0.1829	0.2191	0.2259	0.2036	0.1681	0.1412	0.0969	0.074	0.0509	0.0373	0.0272	0.0213	0.0176	0.014	0.0098
skewness	8.2602	8.5575	10.0383	11.3257	10.0677	7.6269	8.5839	8.8185	8.9577	7.8704	6.9268	6.591	6.028	7.8387	8.7858	6.9784	6.2945	6.8688	6.221	5.9755
kurtosis	120.298	128.9357	190.275	242.2775	208.1898	82.9457	117.0851	124.5151	131.3643	99.5985	68.6172	68.6205	51.906	106.6552	156.0315	87.9723	60.6166	79.6645	60.9231	57.1025

Most parameters show high skewness (>7) and heavy kurtosis, suggesting significant outliers and concentrated distributions around low values. Parameters like **T0pt010S to T0pt100S** have lower mean values, while others (e.g., **T0pt150S to T0pt500S**) show increasing variability.

5.Plots of Actual vs Predicted log10(SA) Across Time Periods:

Time Period Index 0

- **High accuracy:** Points tightly cluster around the 1:1 line.
- **Low bias:** No clear over- or under-prediction trend.
- **Conclusion:** Excellent model performance at short periods.

Time Period Index 10

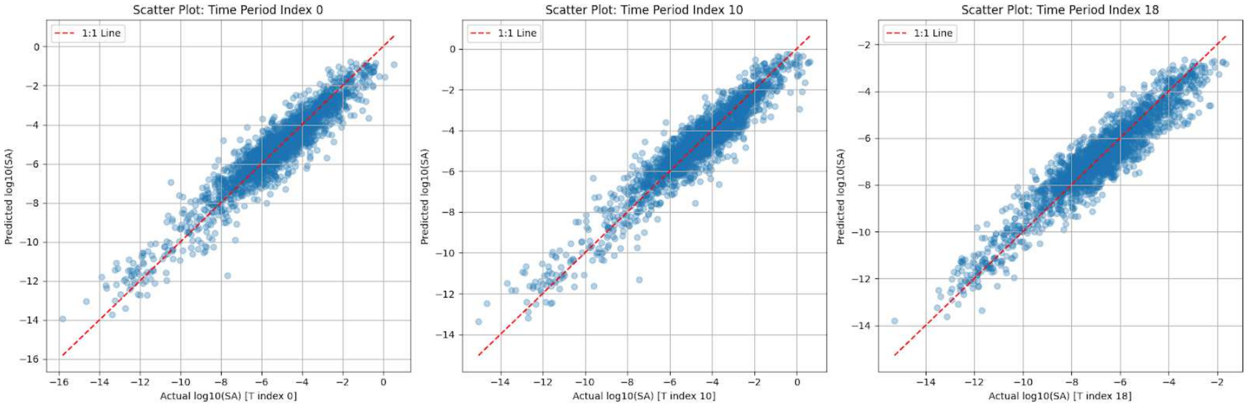
- **Moderate scatter:** Still well-aligned with the 1:1 line.
- **Mild under-prediction:** Slight deviation below the line at higher SA values.

- **Conclusion:** Good performance, but accuracy slightly drops.

Time Period Index 18

- **Increased scatter:** Wider spread around the 1:1 line.
- **Consistent under-prediction:** Noticeable for large actual values.
- **Conclusion:** Performance degrades at longer periods, with growing bias and variance.

Overall: The model predicts well across all periods, with highest accuracy at low periods and increasing error/bias at longer periods.



6.Model Architecture:

- Model predicts 20 log-transformed PSa values using 5 ground motion features
- Uses RandomForestRegressor with 300 trees and depth limit of 20
- Replaces 0 and inf in `rjb` and `vs30`, then applies log10 transform
- Applies natural log to PSa targets for numerical stability
- Standardizes inputs and targets using StandardScaler
- Splits data into train, validation, and test sets randomly
- Trains model iteratively up to 15 times, tracking validation MSE
- Saves best model based on lowest validation loss
- Computes residuals by predicting full dataset and subtracting predictions from true log-PSa
- Applies MixedLM to extract inter-event residuals using EqID as group
- Computes intra-event residuals as difference between total residual and inter-event
- Tracks per-period training loss (LH) and change in loss (LHR)
- Stops early if validation loss change is below 10% after two iterations

7.Model Performance Metrics for Target Variables:

- **R²**: Ranges from 0.8315 to 0.8795, indicating good predictive accuracy for all targets.
- **Inter-Std (τ)**: Shows moderate variability between groups, with values from 0.4989 to 0.7750.

- **Intra-Std (φ)**: Reflects variability within the same group, ranging from 0.6342 to 0.8195.
- **Total Std**: Total variability, which decreases from 1.1280 for "T0pt100S" to 0.8069 for "T5pt000S".

Target Variable	R^2	Inter-Std (τ)	Intra-Std (φ)	Total Std
T0pt010S	0.8666	0.6708	0.6867	0.96
T0pt020S	0.8658	0.6747	0.6897	0.9648
T0pt030S	0.8636	0.6878	0.6966	0.979
T0pt050S	0.8541	0.7215	0.7298	1.0262
T0pt075S	0.8394	0.7575	0.7851	1.091
T0pt100S	0.8315	0.775	0.8195	1.128
T0pt150S	0.8365	0.749	0.8168	1.1083
T0pt200S	0.8434	0.7258	0.7941	1.0758
T0pt300S	0.8591	0.6866	0.7351	1.0058
T0pt400S	0.8675	0.6673	0.6985	0.966
T0pt500S	0.87	0.6459	0.6825	0.9397
T0pt750S	0.8645	0.6164	0.6784	0.9166
T1pt000S	0.8587	0.598	0.6883	0.9118
T1pt500S	0.8487	0.5586	0.7086	0.9023
T2pt000S	0.8475	0.5479	0.7103	0.897
T2pt500S	0.8524	0.5298	0.7022	0.8797
T3pt000S	0.8583	0.5239	0.6907	0.8669
T3pt500S	0.8625	0.5219	0.6809	0.8579
T4pt000S	0.8678	0.5133	0.6655	0.8404
T5pt000S	0.8795	0.4989	0.6342	0.8069

Overall, the model shows consistent performance, with R^2 values improving slightly as the target variables increase. However, there remains variability within and between targets, suggesting potential areas for further refinement.

8.Residual Analysis:

Inter-event Residual vs Magnitude (Top Row)

- Across all periods (0.1s, 1.0s, 3.0s), the inter-event residuals show no strong trend with magnitude (Mw), indicating that the model captures magnitude scaling well.
- The mean residuals are generally close to zero with moderate spread, showing unbiased event-specific performance.

Intra-event Residual vs Rjb (Middle Row)

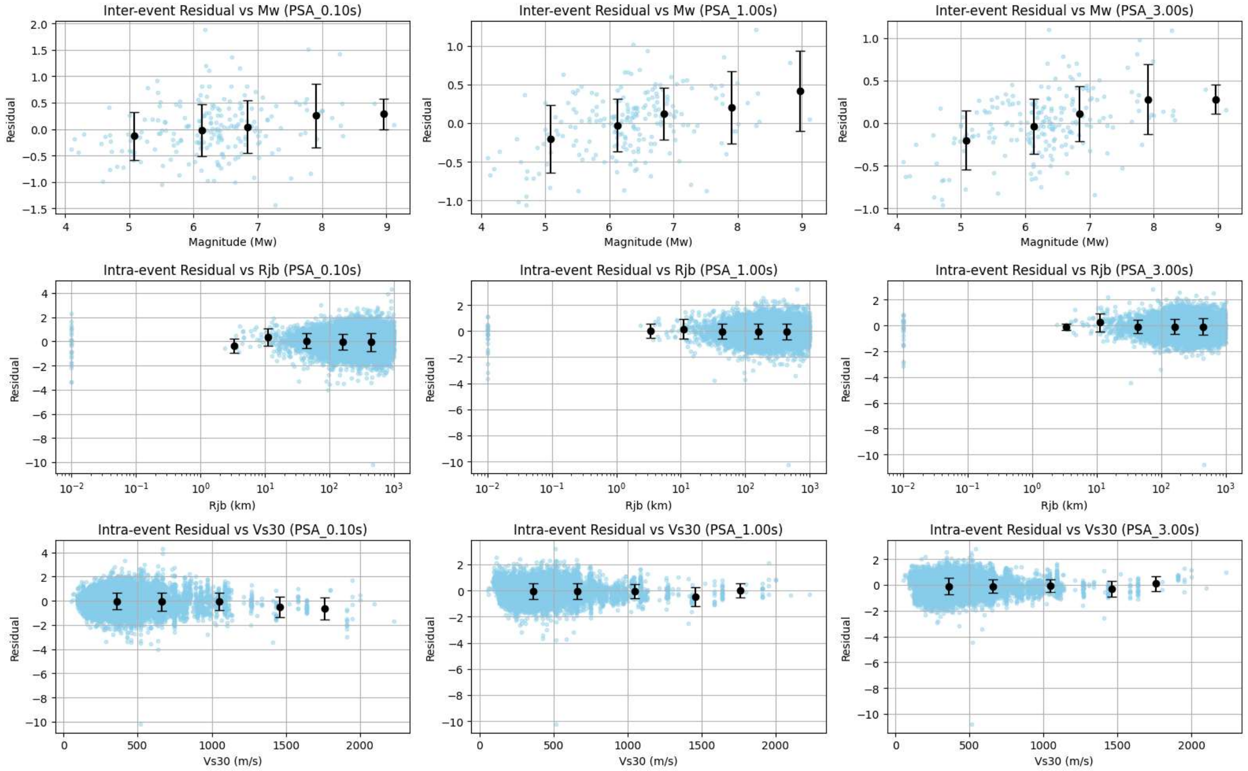
- Residuals slightly decrease with increasing distance (Rjb), especially beyond ~10 km, suggesting mild underprediction at farther distances.
- The variability is larger at shorter distances but reduces at greater distances, which is typical in ground motion models due to signal attenuation.

Intra-event Residual vs Vs30 (Bottom Row)

- Residuals show a negative trend with Vs30, particularly for lower Vs30 values (< 1000 m/s), indicating underprediction at soft sites.
- This trend weakens at higher Vs30 values, suggesting the model is more accurate for stiffer sites.

Summary

The random forest model performs robustly with respect to magnitude but shows minor biases with distance and site conditions, especially underpredicting for soft soils and at greater distances.



9.Magnitude Sensitivity Plot:

Inter-event (Left):

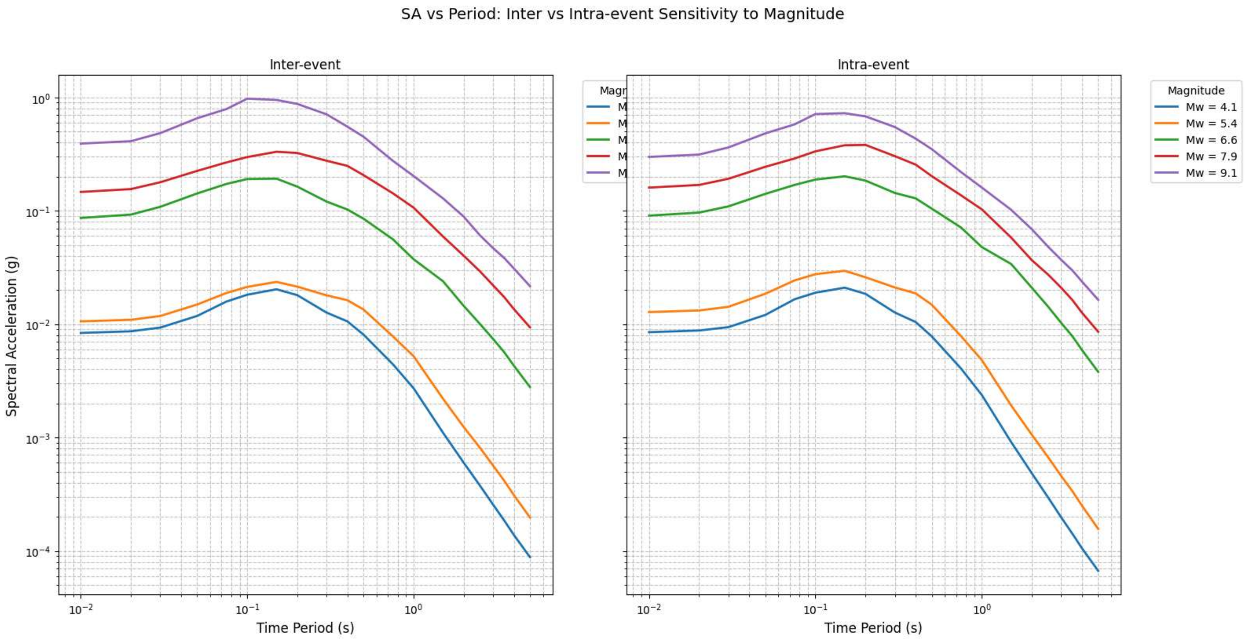
- **SA increases with magnitude** at all periods.
- **Peak SA** around 0.2–0.4s.
- **Magnitude sensitivity grows** at longer periods — higher separation between magnitude curves.

Intra-event (Right):

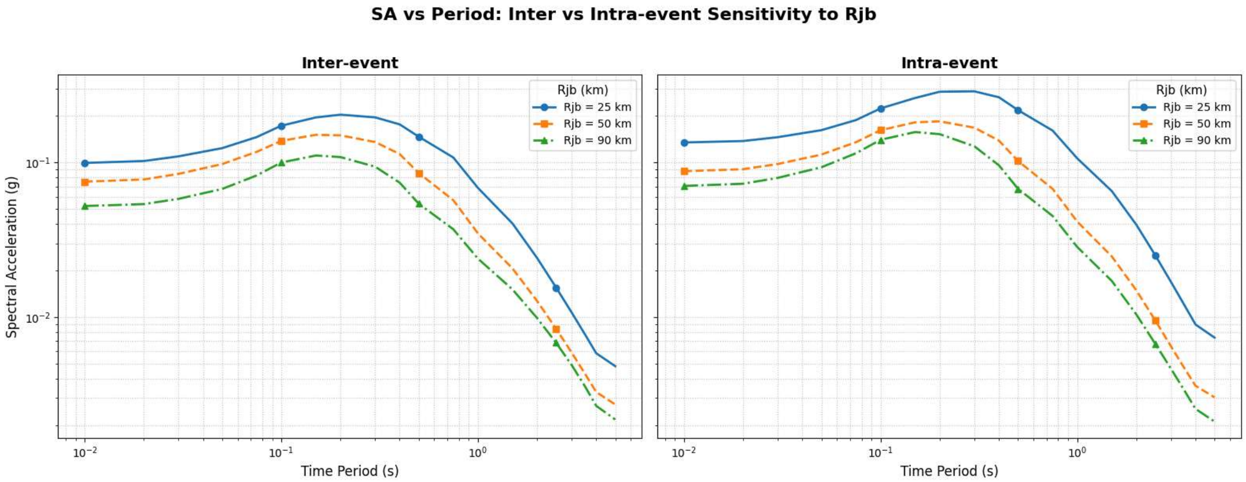
- Similar SA trends, but **curves are closer**.
- **Lower variability** and **weaker magnitude sensitivity**, especially at short periods.

Conclusion:

Magnitude has a **stronger effect on inter-event** variations, especially at long periods. **Intra-event variability** is more stable across magnitudes.



10.Rjb Sensitivity Plot



Inter-event (Left):

- SA decreases with increasing Rjb across all periods.
- Maximum SA occurs at ~0.3s, with greater separation at longer periods.
- Distance sensitivity (difference between 25 km and 90 km) becomes more prominent at periods > 0.5s.

Intra-event (Right):

- Similar trend: closer distances yield higher SA.
- The curves are closer together, indicating lower sensitivity to Rjb compared to inter-event.
- Still, some spread at long periods suggests moderate intra-event distance dependence.

Conclusion:

SA decreases with distance (Rjb), more strongly in inter-event variations. Intra-event variability is less sensitive but still shows distance dependence, especially at longer periods.

11.Vs30 Sensitivity Plot:

Inter-event (Left):

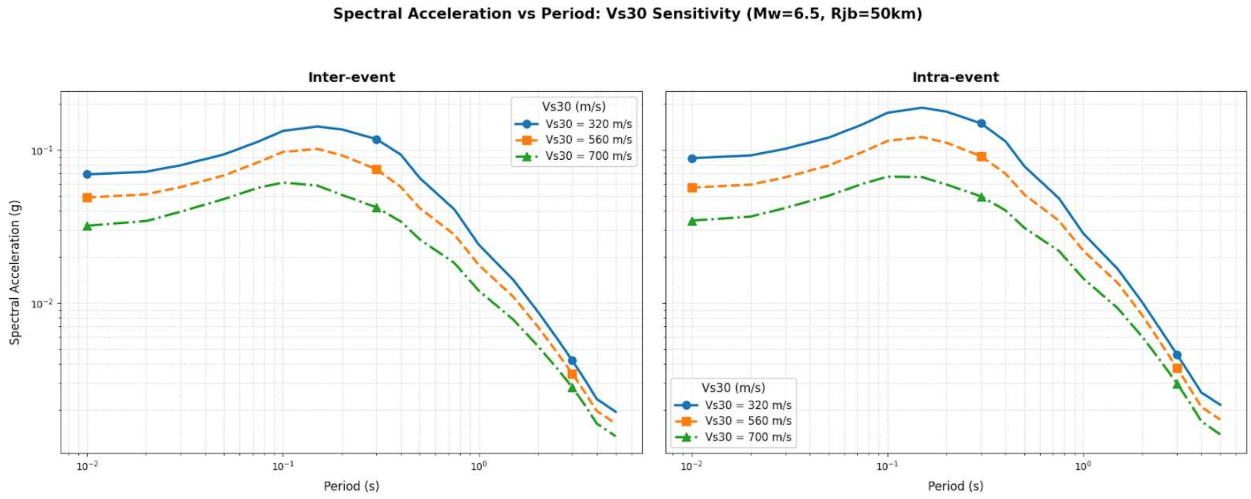
- SA decreases with increasing Vs30 (i.e., stiffer soils yield lower ground motion).
- Max SA around 0.3s, especially for Vs30 = 320 m/s.
- Difference between curves is significant across periods—inter-event residuals are sensitive to site conditions (Vs30).
- The gap narrows slightly at long periods (>1s), but remains evident.

Intra-event (Right):

- Same overall trend: softer sites (lower Vs30) produce higher SA.

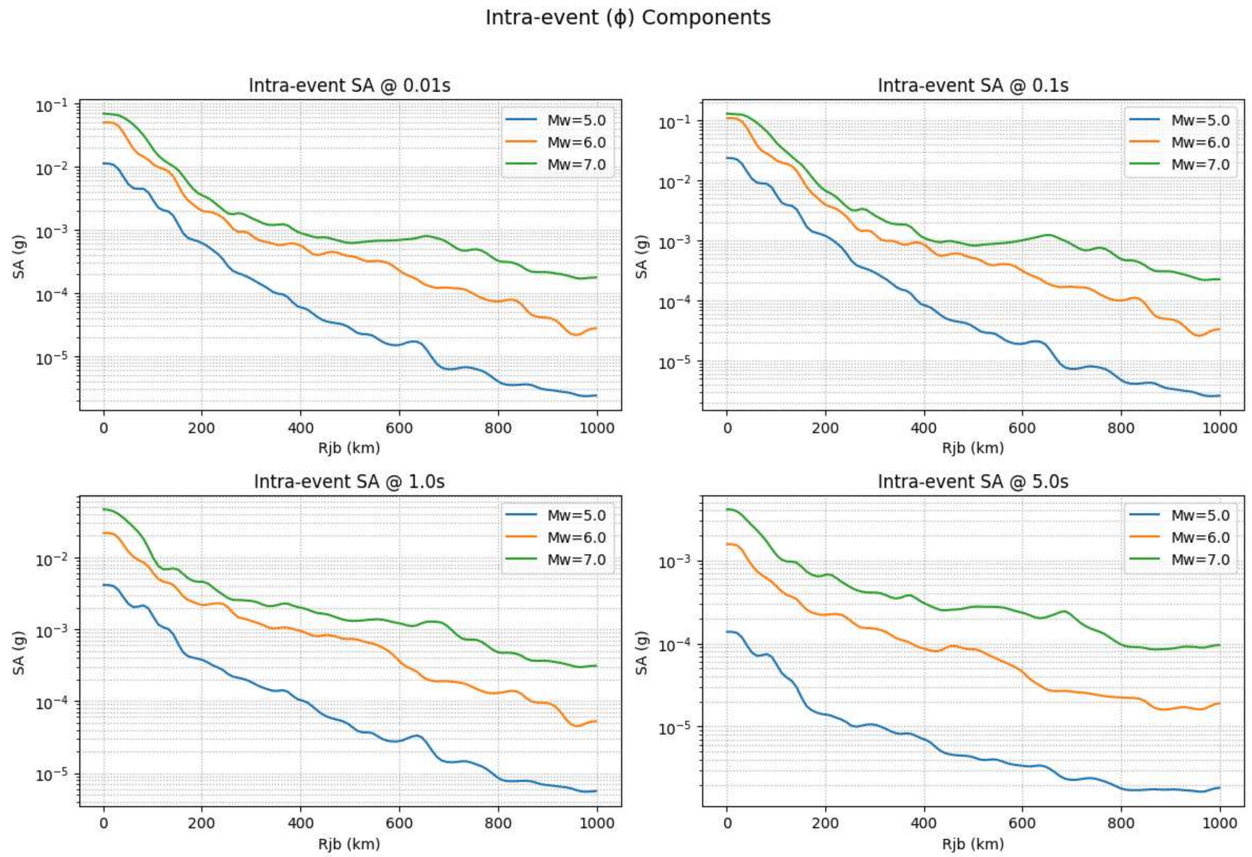
- Intra-event SA values are consistently higher for Vs30 = 320 m/s, indicating strong local site amplification.
- Sensitivity to Vs30 is evident at all periods, though slightly reduced at the longest periods.

Conclusion:
Spectral acceleration decreases with increasing Vs30. Site effects are clearly captured in both inter- and intra-event components, with inter-event showing slightly stronger sensitivity. Softer soils (lower Vs30) significantly amplify ground motions, especially around short to intermediate periods (~0.1–0.5s).

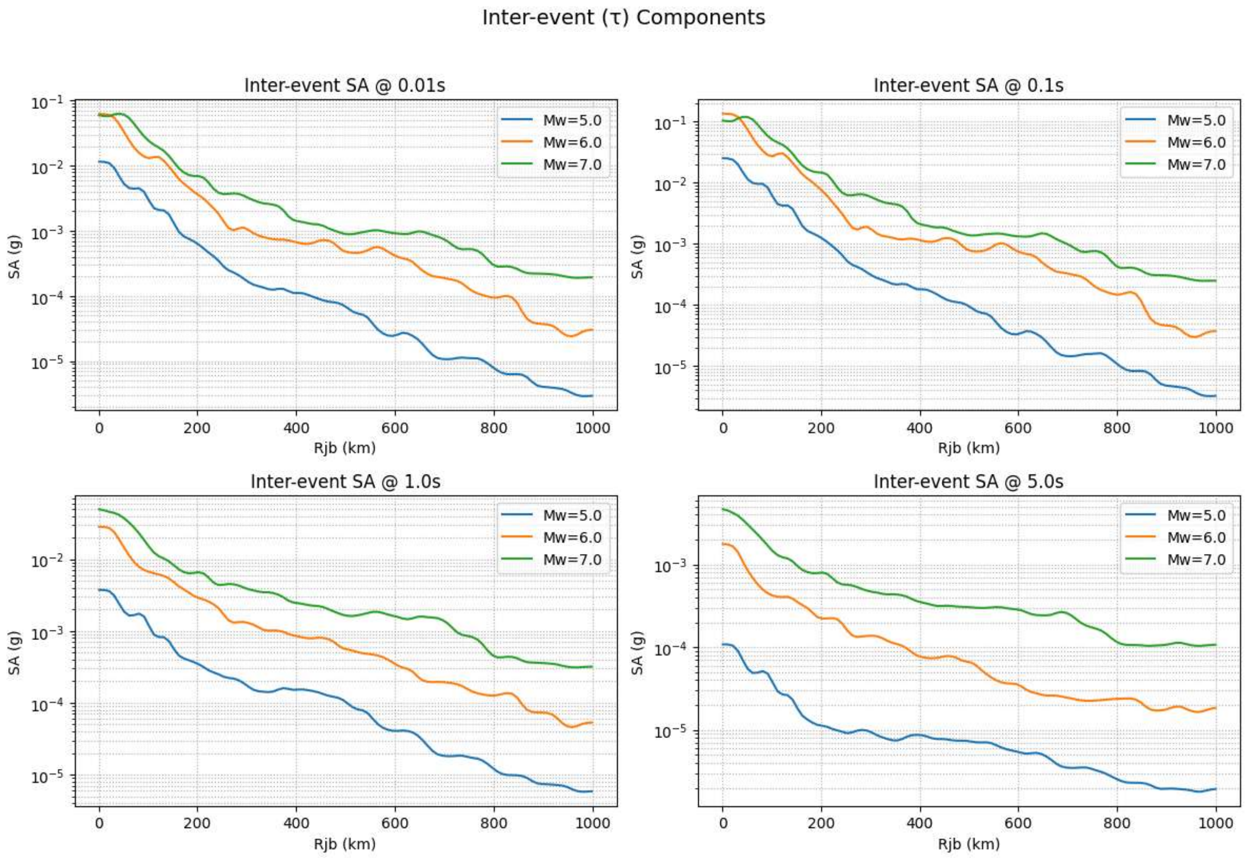


12.SA @ T vs Rjb:

- **Intra-event (ϕ) Components: SA vs Rjb**
  - **Intra-event variability (ϕ)** captures how SA varies across recording stations for the **same event** — and this variability is influenced by both **distance attenuation** and **event magnitude**.
  - The plots reflect:
    - **Higher site-to-site variation** for stronger and closer events.
    - **Period dependence**, with high variability at short periods (dominant for rigid structures) and sustained long-period variability for large events (critical for tall/flexible structures).
  - **Magnitude scaling and distance attenuation are both evident** and consistent with empirical ground motion models.

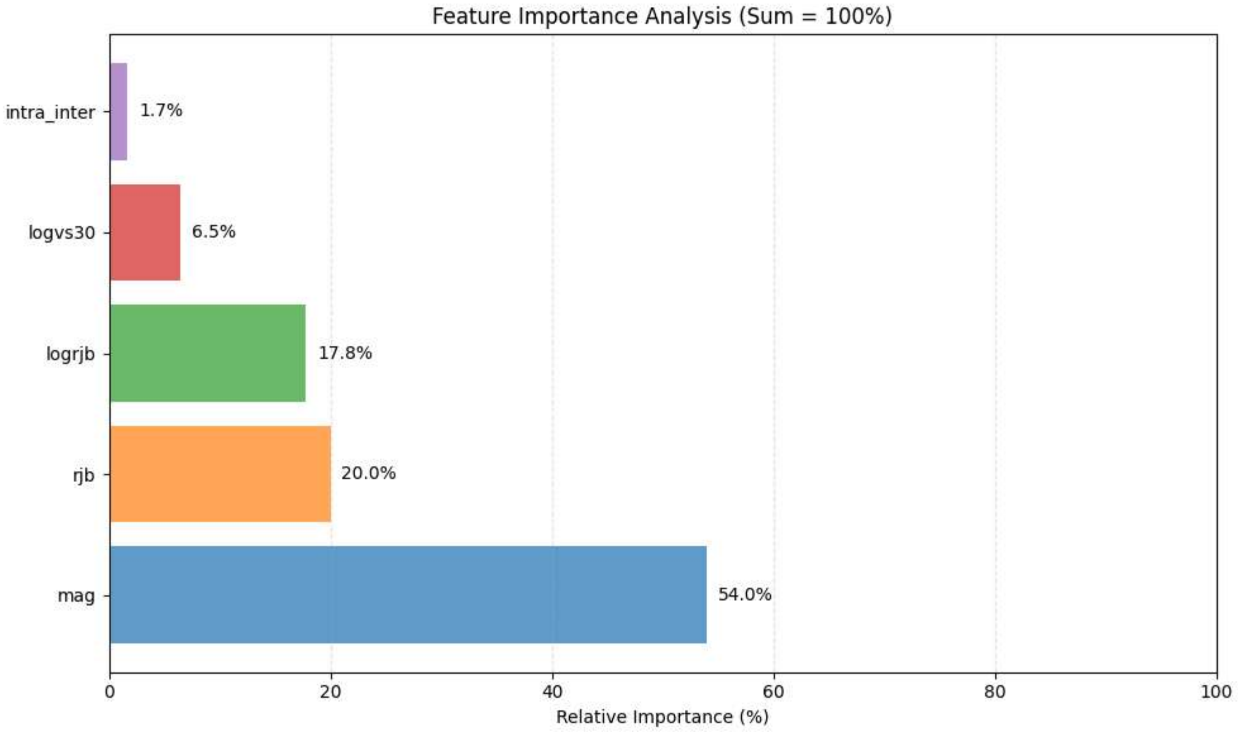


- **Inter-event (τ) Components: SA vs Rjb**
  - **Distance attenuation and magnitude scaling** are both captured well by the inter-event residuals.
  - **Larger earthquakes generate higher SA** and **maintain energy over longer distances**.
  - Period influences both the **rate of decay** and **amplitude of inter-event SA**, with **longer periods showing broader spacing between magnitudes**.



Inter-event (τ) Components

14.Feature Importance Summary:



Feature Importance Analysis (Sum = 100%)

- **Magnitude and distance** dominate SA prediction in your model.
- **Site effects (logvs30)** and **event type flags** are much less influential — though still non-zero.
- The model benefits from using **both linear and log distance terms**, which improves physical realism and flexibility.

Code: Random Forest Model