Application of Data-Driven Approaches for Estimation of Site Response Utilizing mHVSR and Site Proxies in California

Francisco Javier G. Ornelas[1]; Christopher A. de la Torre[2]; Tristan E. Buckreis[1]; Chukwuebuka C. Nweke[3]; Scott J. Brandenberg[1]; and Jonathan P. Stewart[1]

[1]University of California, Los Angeles (jornela1@g.ucla.edu), [2]University of Canterbury, Christchurch, NZ, [3]University of Southern California

Introduction

Traditional ground motion models (GMMs) use proxies like V_{S30} (Borcherdt, 1994) and depths to shear-wave velocity isosurfaces ($z_{1.0}$ and $z_{2.5}$) to represent site effects (e.g., Boore et al., 2014; BSSA14). While effective for ergodic modeling, these proxies often fail to capture site-specific geological complexities, contributing to epistemic uncertainty in seismic hazard assessments.

This study explores the use of parameters derived from microtremor-based horizontal-to-vertical spectral ratio (mHVSR) curves—such as the predominant frequency (f_0), amplitude (a_0), and full spectral shape—in combination with traditional site parameters (i.e., V_{S30}) to evaluate which parameters have the strongest predictive power for site amplification. **Figure 1** presents an example of a peak in a mHVSR curve which is fit with a Gaussian pulse function (Wang et al., 2022; Buckreis et al., 2024) using an algorithm developed by Wang et al. (2023). While the pulse fit enables the extraction of various features, this study focuses on two key parameters—f_0 and a_0—which have previously been found to be related to site response characteristics (e.g., Kwak et al. 2017; Wang et al. 2022; Buckreis et al., 2024).

We employ a random forest (RF) regression model using data from 685 sites across California to evaluate the performance of various proxy combinations in predicting linear site response (F_{lin}). This work builds upon a previous study that focused exclusively on the use of the mHVSR spectral shape as a predictor variable (Ornelas et al., 2026). Our results indicate that the incorporation of mHVSR-derived features significantly improves predictive accuracy and reduces epistemic uncertainty compared to the benchmark site amplification model (Seyhan and Stewart, 2014; SS14). These findings underscore the value of various proxy combinations in enhancing site response predictions across California and provide insight into the specific features that influence key aspects of site response, such as resonance. However, further research is needed to understand the underlying mechanisms of this RF model, particularly how it is making predictions, before it can be applied in practical applications. Nonetheless, the results offer valuable insights into the potential advantages of incorporating mHVSR parameters into future site response models..

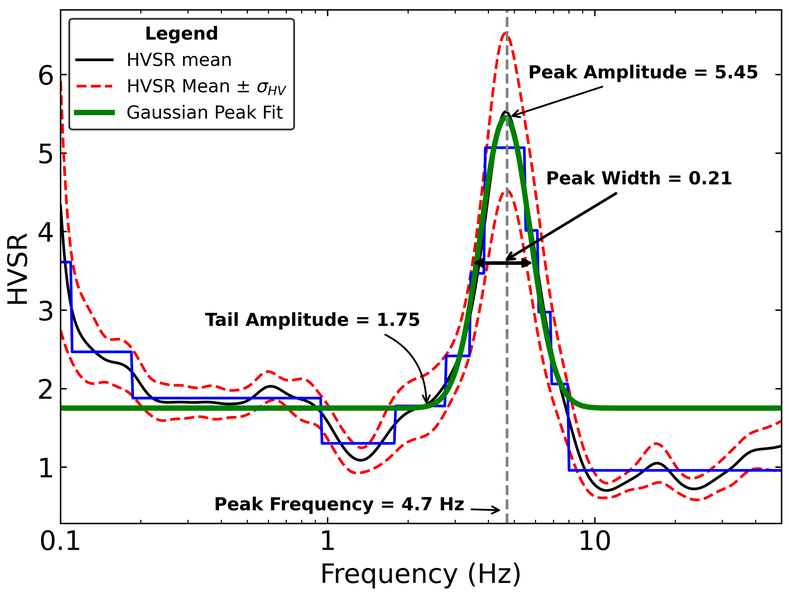

Figure 1: Illustration of parameters extracted from a Gaussian pulse fit applied to a mHVSR curve. These parameters can be used as input features for various predictive models. In this study, two key features utilized were the peak frequency (f_0) and the peak amplitude (a_0).

Methodology and Database

mHVSR Data

- From VSPDB (~1,400 sites; Kwak et al., 2021; https://vspdb.org)
- Broadband seismometers only (permanent and temporary)
- Derived from ambient noise which can be natural or human induced sources
- Processed per Wang et al. (2022); Ornelas et al. (2024)

Ground Motion Data

- From GMDB (Buckreis et al., 2025): 62,506 records, 4,469 stations
- 2,453 stations retained after screening
- 685 stations co-located with mHVSR (≤150 m) – see **Figure 2**

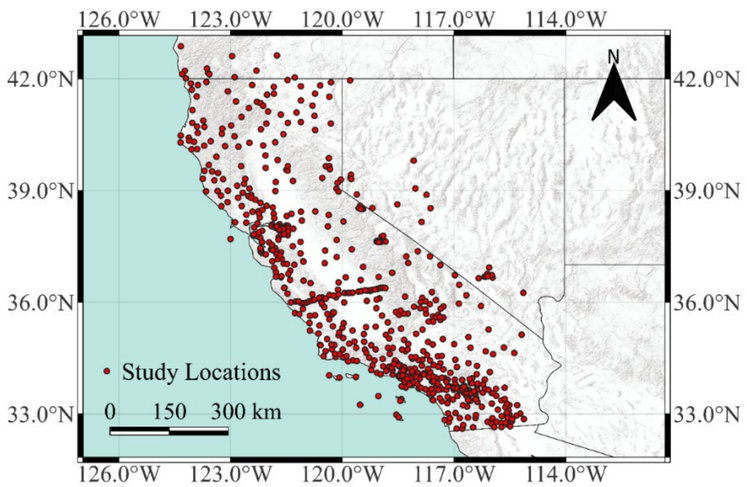

Figure 2: Location of mHVSR tests co-located with strong motion sensors.

- Site response computed from site terms derived from mixed-effects regression and linear component (F_{lin}) of the BSSA14 (i.e., SS14) ergodic model.
- Random forest (RF) model trained on 80 % of site response and input data. The other 20 % was used for model validation.
- **Figure 3** shows correlations between input and output variables using Pearson's *r*, indicating a strong correlation between mHVSR and site response at low frequencies, negative correlation with V_{S30} (i.e., increasing V_{S30} causes decreased site response), positive correlations with $z_{1.0}$, $z_{2.5}$, and weaker correlations for slope, terrain, f_0, a_0 and geologic units.
- **Figures 4** and **6** present overall model performance and **Figure 5** provides example results

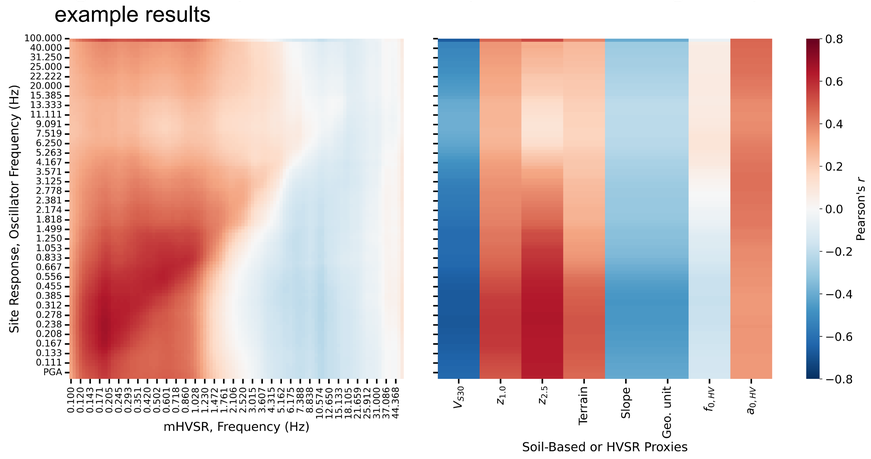

Figure 3: Comparison between soil-based (e.g., V_{S30}) and mHVSR-related proxies (e.g., f_0) using Pearson's *r* to evaluate correlation. The frequency labels on the left plot are related to the ordinate position relative to frequency.

Results

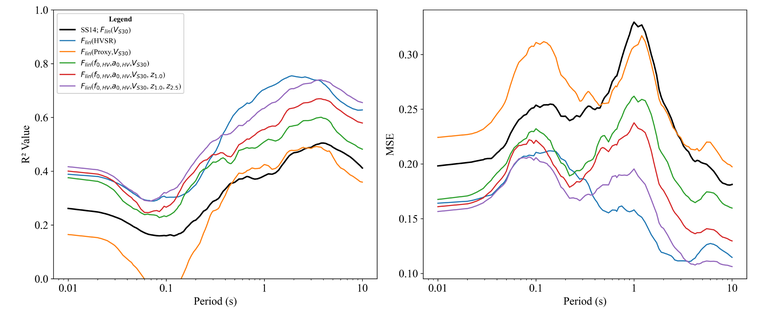

Figure 4: Comparison of model performance for different input parameter combinations. (a) Coefficient-of-Determination (R^2) (b) Mean Squared Error (MSE). Proxy in the legend indicates a combination of slope, geologic unit and terrain class. Results show improved model performance (higher R^2 and lower MSE) for models that include the f_0, a_0, V_{S30}, $z_{1.0}$, and $z_{2.5}$ parameters relative to models that use only V_{S30}-based models such as SS14.

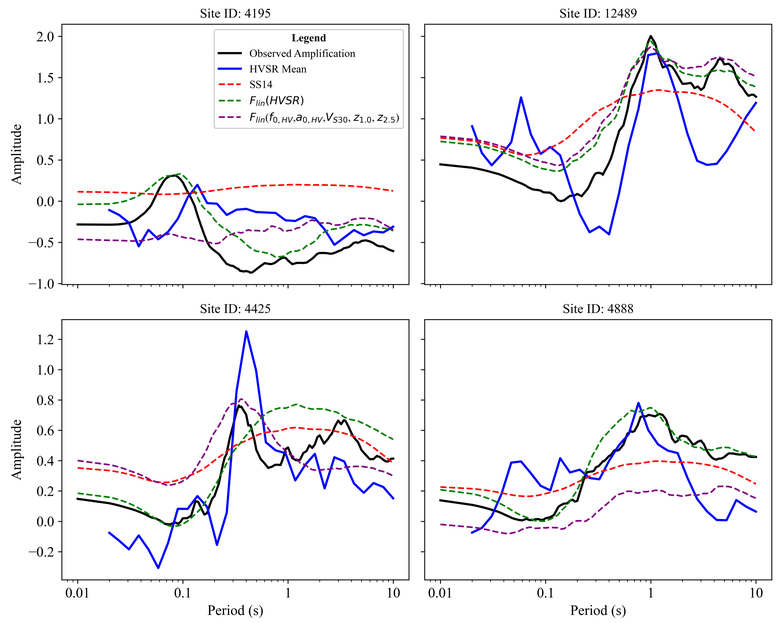

Figure 5: Comparison of site responses at four example sites and how different models predict it. The green dashed line represents a RF model using 15 ordinates from a HVSR mean curve to predict site response. The red dashed line represents a prediction from SS14. The purple dashed line represents a RF model using a combination of scalar-proxies. The results highlight scenarios in which one ML model outperforms the other, as well as cases where both models demonstrate improved performance, depending on the input parameters.

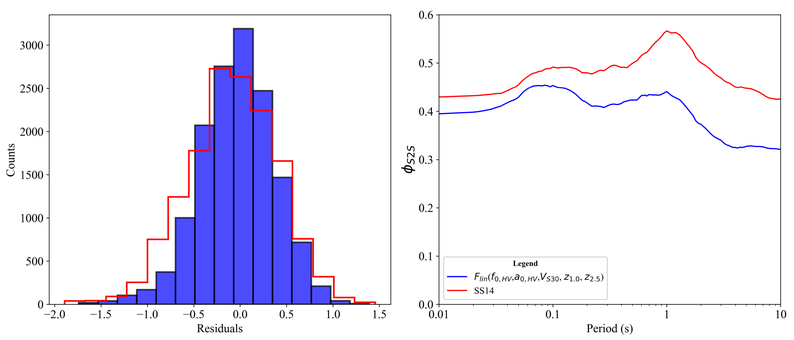

Figure 6: Comparison of model performance using SS14 V_{S30}-based model and RF model that incorporates f_0, a_0, V_{S30}, $z_{1.0}$, and $z_{2.5}$: (a) Within-event residuals in natural log units, (b) Site-to-site variability. The results indicate a reduction in variability, with residuals more tightly clustered.

Acknowledgements

This study was funded under CSMIP Agreement #1023-018, and we sincerely acknowledge their generous support. Additionally, we would like to express our gratitude to Pacific Gas and Electric (PG&E) for their partial funding contribution to this research.