Application of Machine Learning for Identification of Hidden Rock Sites using Earthquake Records

Michael R. Dupuis, P.E., M.A.Sc.[1]

[1]Geosyntec Consultants Inc., 803 2nd Street, Suite D, Davis, CA 95616; Email: michael.dupuis@geosyntec.com

ABSTRACT

Earthquake ground motions recorded at rock sites can be used to reduce seismic hazard uncertainty for dams and other critical infrastructure on rock; however, few ground motion records exist from known rock sites because most seismic stations are on soil or lack measured shear wave velocities. In this study, a dataset of California ground motions is examined to identify additional instrumented "Very Dense Soil" and "Soft Rock" sites which may actually be "Rock". This approach offers a cost- and time-efficient alternative to installing new seismic stations. To identify possible rock sites, partially-crossed linear mixed-effects regression is used to compute site residuals from empirical ground-motion model predictions and recorded ground motions. Based on these site residuals and other site characteristics, machine learning is then applied to estimate shear wave velocities at sites without measured velocities. Using these estimated shear wave velocities, a ranked list of candidate rock sites is developed. Geophysical testing is proposed at a selection of these sites to verify the high estimated shear wave velocities, with a focus on sites with many recorded ground motions, as an efficient means to expand the catalogue of rock ground motions.

INTRODUCTION

From review of the PRJ-3031 ground motion database (Ji et al, 2025), it was observed that measured time-averaged shear-wave velocity in the top 30 meters, V_{S30}, for seismic stations in California can differ significantly from proxy V_{S30} values, which are inferred from available geological data. For example, certain sites have measured V_{S30} greater than 1000 m/s and have proxy V_{S30} of only 350 m/s. Furthermore, very few sites have proxy V_{S30} values greater than 900 m/s. Based on these observations, it was inferred that there may be stations with significantly higher V_{S30} than is currently recognized by the proxy values, i.e., unrecognized rock sites. Reclassification of these existing seismic stations offers a practical and efficient approach to expand the database of rock ground motions and would allow for improved seismic hazard estimates for rock sites. This work is particularly relevant for concrete dams, which are typically

founded on rock, have long service lives, and must be designed and assessed for rare, long return-period earthquake ground motions.

BACKGROUND

Dam owners typically rely on ergodic site adjustment factors to represent site conditions within ground motion models, including at rock sites. Commonly used approaches incorporate the NGA-West2 ground motion models (e.g., Campbell and Bozorgnia, 2014) and hard-rock adjustments (e.g., Ktenidou and Abrahamson, 2016). While these models represent the current state of practice, they are largely based on empirical data from relatively few rock sites in California.

Advancements in ground motion modeling are trending toward non-ergodic approaches that account for location-specific source, path, and site effects. One method is the partially non-ergodic model, where the site term is specific to the location (Stewart et al., 2017). Recent applications in dam engineering have demonstrated the feasibility of this approach using site-specific ground motion recordings at rock sites, where linear response allows for the use of weak ground motion records to make inferences for strong ground-motion response (e.g., Vecchietti et al., 2019; Hassani et al., 2023).

Despite these advances, it is rare for site-specific ground motion records to be available at a given dam site. For these cases, ground motion recordings from nearby rock sites, not necessarily at the dam site itself, can still support more accurate seismic hazard calculations by enabling estimation of non-ergodic site terms. Therefore, identifying additional earthquake records from rock sites has the potential to reduce the uncertainty in seismic hazard analysis for dam sites on rock in California.

METHODS

Observed earthquake ground motions and site data. Earthquake ground-motion records from version 2 of the PRJ-3031 ground motion database (PRJ-3031) (Ji et al, 2025) were used for this analysis. PRJ-3031 includes 269 300 records from 2 582 earthquakes which have occurred in California and neighbouring regions between 1999 and 2021. These events were recorded at over 1061 strong-motion sites distributed throughout California. Only events with moment magnitudes ranging from 3.2 to 7.2 and focal depths shallower than 20 km are included in the dataset.

Ground-motion records are paired with detailed source, path, and site metadata. The database incorporates a compilation of site-specific parameters such as V_{S30}. Critically, there is both measured and inferred proxy values for V_{S30}, with the proxy values having only limited accuracy. It was hypothesized that certain unmeasured sites may have higher V_{S30} than indicated by the proxy values and thus presents an untapped resource which could improve accuracy of seismic hazard analysis for concrete dam sites in California.

The following filtering steps were applied to the PRJ-3031 ground motion database to develop a curated dataset suitable for the mixed-effects regression analysis:

1. Frequency content filters:
    a. Removed 82 139 records where the high-pass corner frequency (1.25 x high-pass cutoff) exceeded 0.5 Hz.
    b. Removed 172 926 records where the low-pass corner frequency (1.25 x low-pass cutoff) was less than 15 Hz.
2. Site-based filters:
    a. Removed 75 records with V_{S30} values less than 180 m/s.
    b. Removed 8 records from the following site IDs: YB.MOJA.HH, YB.MOJA.BH.
3. Event-based filters:
    a. Removed 17 records with moment magnitude less than 3.5.
    b. Removed 15 324 records with rupture distance greater than 100 km.
    c. Removed 136 records with hypocenter depth less than 0 km.
    d. Removed 312 records with hypocenter depth greater than 20 km.
4. Quality control flag filters:
    a. Removed 1715 records flagged as outliers.
    b. Removed 278 records flagged for including multiple earthquakes.
    c. Removed 953 records flagged for channel quality issues.
5. Instrument hierarchy filter:
    a. Removed 174 records from collocated instruments based on the following sensor hierarchy preference: HN, HH, EH, EN, SH, DH, BH, BN, CN, SN.
6. Minimum data requirement filter:
    a. Mixed-effects regression was done enforcing 3 high-quality ground motions per site and per earthquake. This removed 1755 records from events and sites with fewer than three records.

The filtered dataset contained 11 992 records from 782 events and 1060 unique sites; the geospatial extent of the dataset used in the mixed-effects regression is shown in Figure 1.

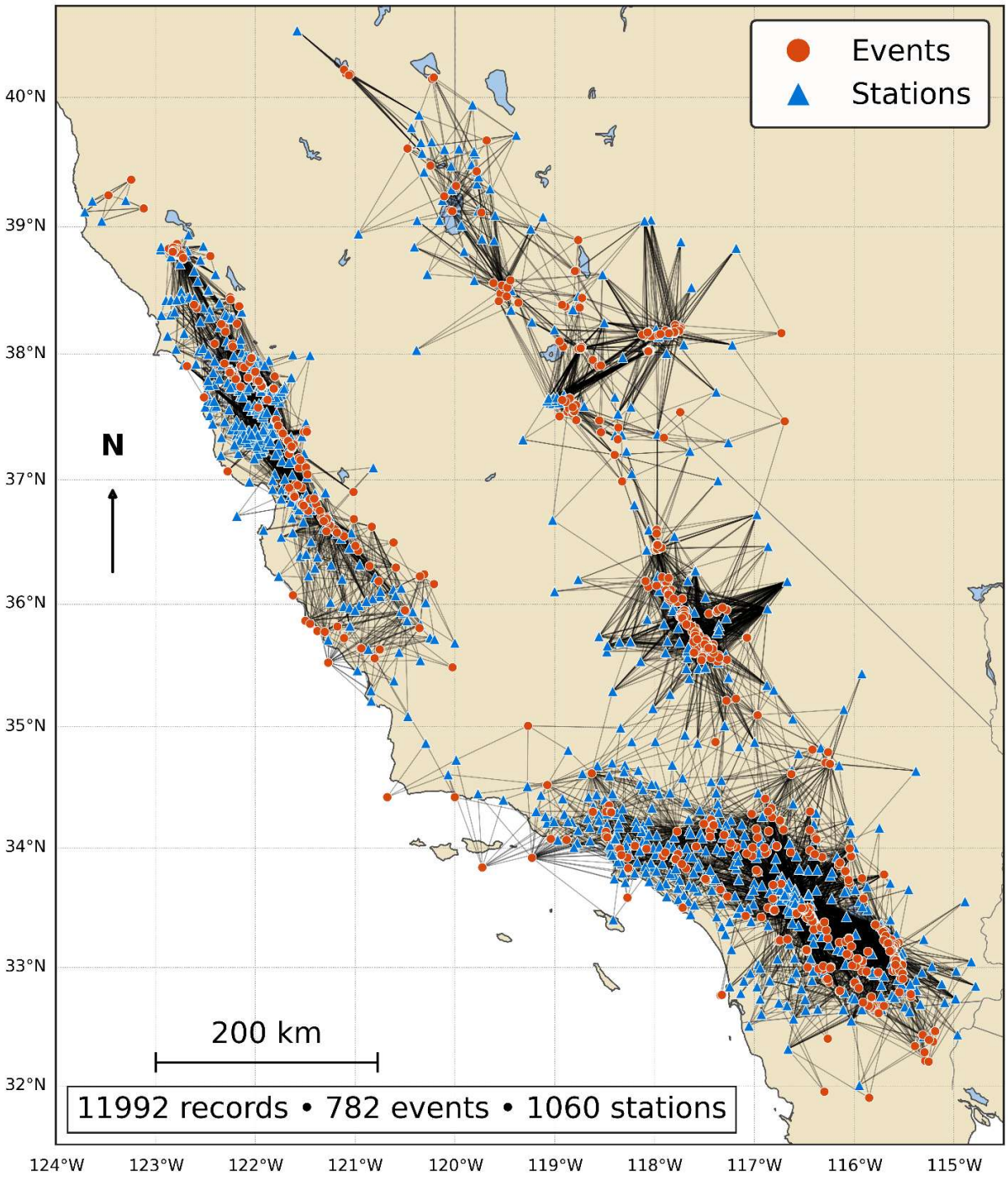

Figure 1. Ground motion data included in the partially-crossed linear mixed-effects regression analysis. Red circles: earthquake epicenters, blue triangles: seismic stations, black lines: ray paths.

Predicted motions from empirical ground-motion model. For each observed ground-motion record, corresponding ground-motion predictions were made using the empirical ground-motion model of Chiou and Youngs (2014). Predictions were made using source, path, and site metadata available from the PRJ-3031 database (Ji et al, 2025). A reference V_{S30} of 760 m/s was used to make predictions for all sites such that systematic site residuals would include the effect of misfit between the reference V_{S30} (760 m/s) and "true" site V_{S30}. It was hypothesized that these systematic site residuals would be informative features on which the machine learning model could be trained.

Partially-crossed linear mixed-effects regression. A partially-crossed linear mixed-effects regression was applied to partition prediction residuals into various components of variability (Bates et al 2014; Stafford 2014). The language and notation used is that of Atik et al (2010) for ground-motion prediction validation which follows the general form of a GMM for an event, *e*, and site, *s*, pairing, explained subsequently.

The general and expanded forms of the equation, presented with the notation of Atik et al (2010), are:

$$\Delta = ln(IM_{es}) - f_{es} = a + \delta B_e + \delta S2S_s + \delta W_{es}^0$$

where Δ is the total prediction residual; $ln(IM_{es})$ is the natural logarithm of the observed intensity measure (IM) for event *e* and site *s*; and f_{es} is the median (for the case without parameter uncertainty there is a single IM prediction) of the predicted logarithmic IM either from a simulation or empirical GMM. a is the model bias, δB_e is the between-event residual for event *e*, $\delta S2S_s$ is the systematic site-to-site residual for site *s*, and δW^0_{es} is the "remaining" within-event residual.

Trends of the misfit between the measured and proxy V_{S30} values with selected site properties are shown in Figure 2. Site properties considered included systematic site-to-site residual of pseudo-spectral acceleration, pSA, at T = 0.2 s, 0.5 s, and 1.5 s, as well as site elevation. As shown, the V_{S30} misfit exhibits correlations with these site parameters, which were included as predictor variables, i.e., features, in the machine learning model.

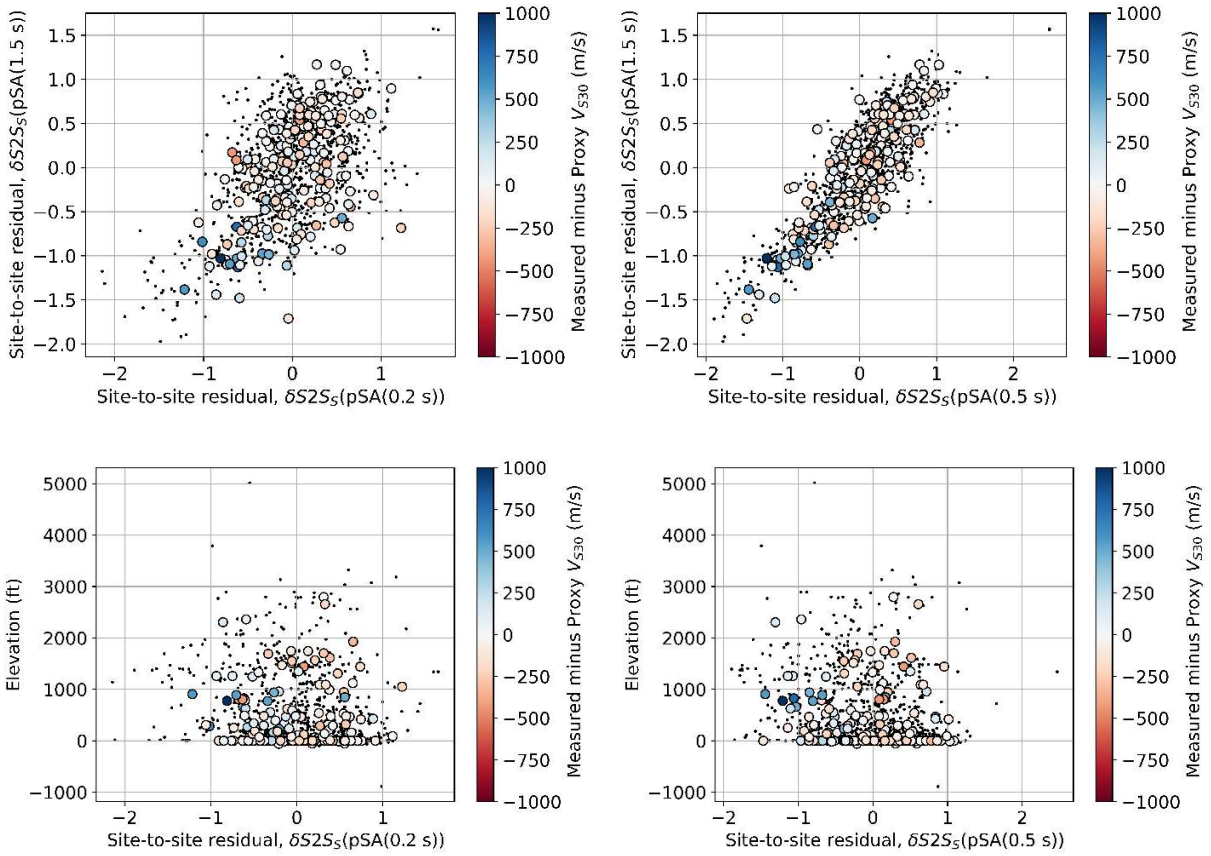

Figure 2. Measured minus proxy V_{S30} for all 196 seismic stations with measured V_{S30} values. The 864 sites without measured V_{S30} are shown as points. Top left: $\delta S2S_s$(pSA(1.5 s)) vs $\delta S2S_s$(pSA(0.2 s)), top right: $\delta S2S_s$(pSA(1.5 s)) vs $\delta S2S_s$(pSA(0.5 s)), bottom left: elevation vs $\delta S2S_s$(pSA(0.2 s)), and bottom right: elevation vs $\delta S2S_s$(pSA(0.5 s)).

MACHINE LEARNING MODEL

Selection of model features

To estimate V_{S30} at stations without measured V_{S30}, a machine learning regression model was developed using available site data and the systematic site-to-site residuals. Measured V_{S30} values were used as targets for the model. The method employed an XGBoost (Chen and Guestrin, 2016) regression model trained on selected features derived from the ground motion metadata and residuals from the mixed-effects regression:

- Proxy V_{S30} value
- Site latitude, longitude, and elevation
- Systematic site-to-site residuals of pSA between 0.2 s and 1.5 s

Model training and validation.

The XGBoost model was trained with a parameter grid search for optimization. A three-fold cross-validation scheme was used to identify the best-performing hyperparameters, including tree depth, learning rate, and column and row subsampling rates. The final model was trained with the optimized parameters using early stopping to prevent overfitting.

Training and testing datasets were constructed by randomly shuffling the measured dataset and partitioning it into a training set (150 samples) and a test set (46 remaining measured samples). The quantity of training, testing, and unmeasured data (i.e., sites) is shown in Table 1. Model performance was assessed, as shown in Figure 3, by comparing estimated V_{S30} values to measured V_{S30} for the testing dataset using root mean squared error, mean absolute error, and bias, among other metrics. It was observed that there is relatively high coefficient of determination, indicating good performance for the testing dataset, and that the model tends to underestimate V_{S30} in the testing dataset, as indicated by the negative mean bias.

Table 1. Training, testing, and unmeasured data.

Partitioned Dataset	Number of Sites
Measured V_{S30} used for Training	150
Measured V_{S30} used for Testing	46
Unmeasured V_{S30} (candidate rock sites)	864

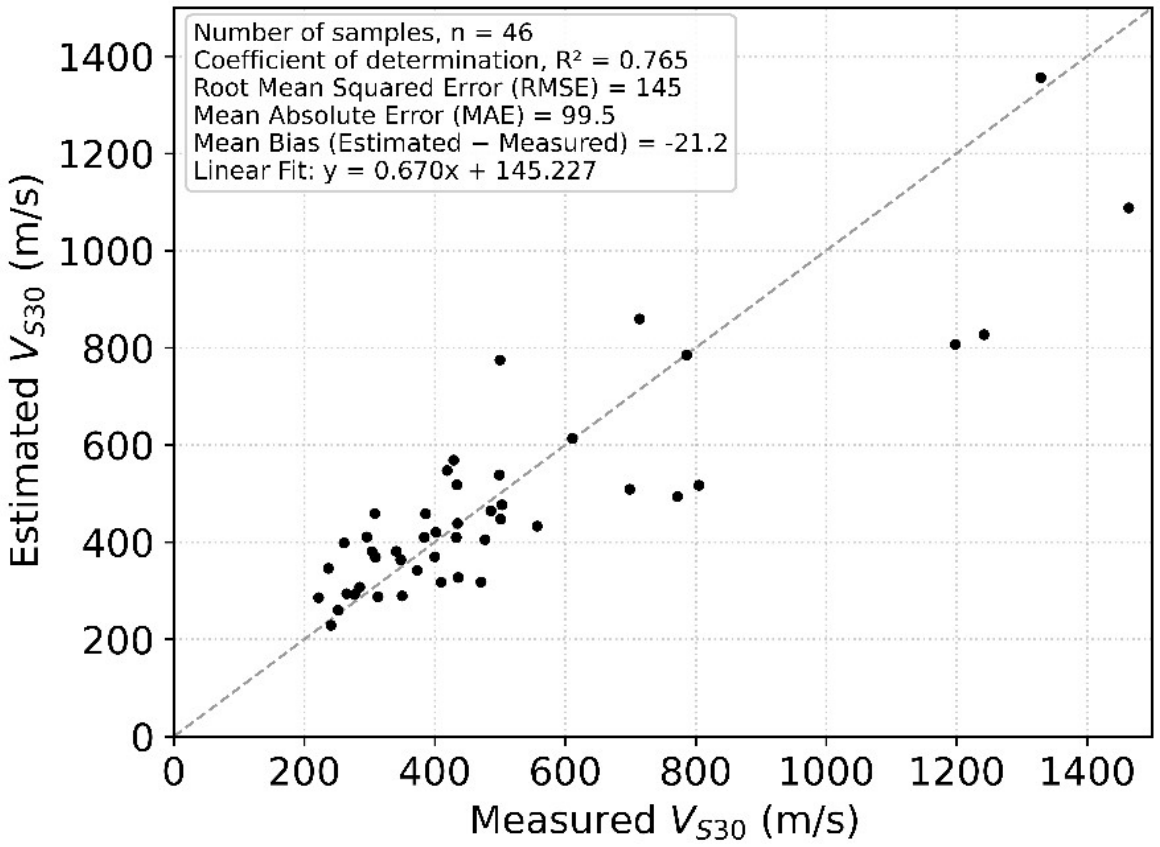

Figure 3. Model performance on the testing dataset.

RESULTS

Estimated V_{S30} values for unmeasured sites were examined and are compared with measured V_{S30} values in Figure 4. For sites with measured V_{S30}, the measured values are shown. For sites without measured data, the estimated values are compared with the proxy values and plotted as a function of systematic site-to-site residuals of pSA at 0.2 s and 1.5 s. As shown, the estimated values for certain sites significantly exceed the available proxy values. A similar observation is made for measured values, which can significantly exceed proxy values at certain measured sites. The general trends in estimated and measured V_{S30} with site parameters, e.g., δS2Ss(pSA(1.5 s)), are very similar, which supports the credibility of the model performance and estimated V_{S30}.

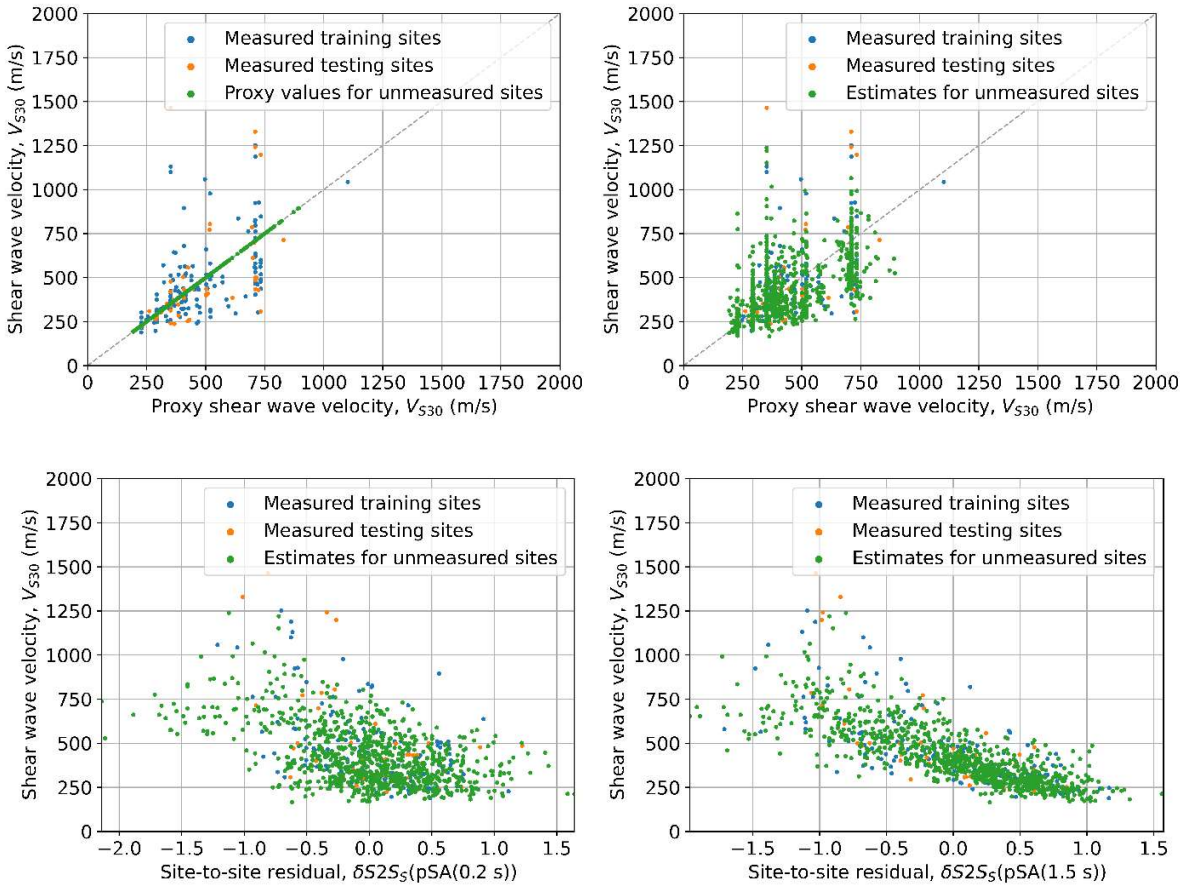

Figure 4. Top left: available V_{S30} data in the PRJ-3031 database vs proxy V_{S30}, top right: model results showing measured (training and validation data) and estimated V_{S30} vs proxy V_{S30}, bottom left: measured and estimated V_{S30} vs δS2Ss(pSA(0.2 s)), bottom right: measured and estimated V_{S30} vs δS2Ss(pSA(1.5 s)).

Candidate rock sites.

Stations with estimated V_{S30} exceeding 850 m/s were screened as potential "rock" sites. A subset of these screened stations was then prioritized based on number of recorded ground motions, geographic coverage, and estimated V_{S30}. A list of candidate rock sites is provided in Table 2 which is intended to guide prioritization of potential geophysical site investigations to identify additional rock sites in California. Figure 5 provides the National Earthquake Hazards Reduction Program (NEHRP) site classes, based on the measured and estimated V_{S30}, for all stations considered in this study. As Figure 5 illustrates, the site V_{S30} estimates, if confirmed through field testing, have the potential to significantly expand the catalogue of available rock ground motions in California.

Table 2. Unmeasured sites with estimated V_{S30} greater than 850 m/s listed in descending order of estimated V_{S30}. Sites proposed as priorities for geophysical testing are shown in shaded background.

Site Code	Latitude	Longitude	Proxy V_{S30} (m/s)	Estimated V_{S30} (m/s)	Number of Records
SLV	33.23854	-115.304	352	1238	3
HAY	33.70734	-115.639	352	1220	7
B073	35.9467	-120.472	352	1152	8
ELS2	33.64907	-117.426	710	1065	6
B075	35.9292	-120.515	372	1016	8
B076	35.9398	-120.425	514	993	8
CAO	37.3485	-121.535	733	993	3
B921	35.5865	-117.462	710	991	5
DTP	35.26742	-117.846	710	974	8
SAO	36.76403	-121.447	352	945	7
DAW	36.271	-117.592	352	906	19
KYV	33.92545	-116.173	710	893	11
TUBB	33.2101	-116.409	710	887	44
ADS2	38.77446	-122.7	733	880	4
MONP2	32.892	-116.422	710	869	12
GVAR1	33.6663	-116.707	751	866	23
CVS	38.34526	-122.458	519	864	10
MP21	37.44159	-121.9	229	863	7

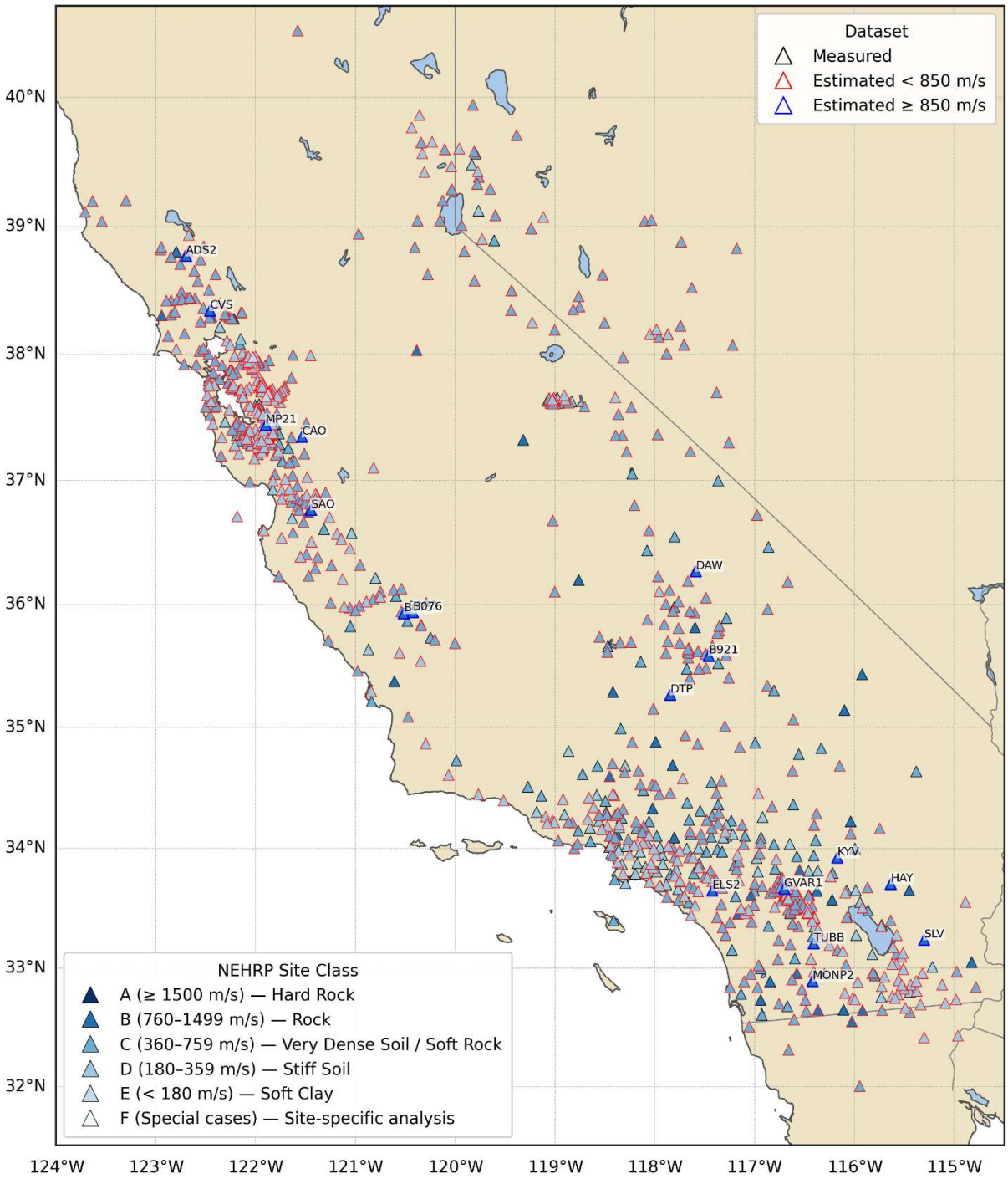

Figure 5. Measured and estimated NEHRP site classes of all seismic stations considered in this study. Sites with estimated V_{S30} greater than 850 m/s are labelled.

LIMITATIONS

The estimates of V_{S30} provided in this study are based on limited site metadata and earthquake ground-motion prediction residuals and did not consider regional geology. Cross-validation of the estimates from this study with regional geological maps would likely provide further insights and may motivate prioritization of different sites than have been identified in Table 2.

CONCLUSIONS AND RECOMMENDATIONS

This study hypothesizes that certain California seismic stations with unmeasured V_{S30} and relatively low proxy V_{S30} may in fact be rock sites with greater V_{S30} than is currently reported. These sites may present a significant untapped resource to constrain seismic hazard uncertainty at rock sites in California. A curated earthquake ground motion dataset and an XGBoost regression model were used to identify candidate rock sites by estimating V_{S30} using site metadata and systematic site-to-site residuals from empirical ground motion model predictions. It is proposed that targeted geophysical site testing at a subset of prioritized sites be conducted, particularly those with numerous high-quality recorded earthquake ground motions and high estimated V_{S30}.

ACKNOWLEDGEMENTS

Observed ground-motion records were obtained from the PRJ-3031 database (https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published/PRJ-3031). Linear mixed-effects regression was done using code provided by Peter Stafford which implements the lme4 package in RStudio. Figures were created using the Matplotlib (https://matplotlib.org/) package in Python (https://www.python.org/). Thanks to Dr. Christie Hale for providing technical insights which supported this work and to Katherine Cheng for directing me to XGBoost. Finally, thanks to three anonymous reviewers who provided comments which substantially improved the quality of this paper.

REFERENCES

Atik, L. A., Abrahamson, N., Bommer, J. J., Scherbaum, F., Cotton, F., and Kuehn, N. (2010). "The variability of ground-motion prediction models and its components." Seismological Research Letters, 81(5), 794–801.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). "Fitting linear mixed-effects models using lme4." arXiv preprint, arXiv:1406.5823.

Campbell, K. W., & Bozorgnia, Y. (2014). NGA-West2 ground motion model for the average horizontal components of PGA, PGV, and 5% damped linear acceleration response spectra. Earthquake Spectra, 30(3), 1087-1115.

Chen, T., and Guestrin, C. (2016). "XGBoost: A scalable tree boosting system." Proc., 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, ACM, New York, NY, 785–794.

Chiou, B., and Youngs, R. R. (2014). "Update of the Chiou and Youngs NGA model for the average horizontal component of peak ground motion and response spectra." Earthquake Spectra, 30(3), 1117–1153.

Chiou, B., Youngs, R., Abrahamson, N., and Addo, K. (2010). "Ground-motion attenuation model for small-to-moderate shallow crustal earthquakes in California and its implications on regionalization of ground-motion prediction models." Earthquake Spectra, 26(4), 907–926.

Hassani, B., Fairhurst, M., Sheffer, M., & Yan, L. (2023). Non-ergodic site response for hard rock correction at a BC Hydro dam site. In Proceedings of the 2023 United States society on dams (USSD) annual conference (pp. 17-21).

Ji, C., Cabas, A., Kottke, A., Pilz, M., Macedo, J., and Liu, C. (2022). A DesignSafe Ground Motion Database: Time Series, Engineering Metrics, and Site Metadata, DesignSafe-CI, https://doi.org/10.17603/DS2-SYC5-NK92 (Apr. 25, 2025).

Ktenidou, O. J., & Abrahamson, N. A. (2016). Empirical estimation of high-frequency ground motion on hard rock. Seismological Research Letters, 87(6), 1465-1478.

Stafford, P. J. (2014). "Crossed and nested mixed-effects approaches for enhanced model development and removal of the ergodic assumption in empirical ground-motion models." Bull. Seismol. Soc. Am., 104(2), 702–719.

Stewart, J. P., Afshari, K., & Goulet, C. A. (2017). Non-ergodic site response in seismic hazard analysis. Earthquake Spectra, 33(4), 1385-1414.

Vecchietti, A., Stewart, J. P., Cecconi, M., Pane, V., & Russo, G. (2019). Non-ergodic site response model based on local recordings for Menta Dam site. In Earthquake Geotechnical Engineering for Protection and Development of Environment and Constructions (pp. 5513-5521). CRC Press.