Learning the low frequency earthquake daily intensity on the central San Andreas Fault

Christopher Johnson[1], Claudia Hulbert[2], Bertrand Rouet-Leduc[2], and Paul Johnson[2]

[1]Los Alamos National Laboratory,Los Alamos National Laboratory
[2]Los Alamos National Laboratory

November 23, 2022

Abstract

Low frequency earthquakes (LFEs) originating below the central San Andreas Fault are associated with slow-slip within the more ductile portion of the crust beneath the seismogenic zone. Monitoring efforts over 15 years recorded >1 million LFEs with >70 per day. We apply machine learning (ML) to statistical features describing the seismic waveforms and estimate the LFE daily intensity. Using 4 years of independent data, the ML model produces a 0.68 correlation. The burst-like LFE behavior is reproduced and the largest misfit occurs during the low-amplitude daily undulations. The ability to continuously monitor LFE activity provides insight to when geodetic measurements of slow slip are possible, without the need for developing a computational-intensive template-matching catalog. Similarities are found between detecting LFEs and tremors, which provides evidence tremors are composed of LFEs. The approach reveals by ML the rich information contained in the features of continuous seismic waveforms.

Learning the low frequency earthquake daily intensity on the central San Andreas Fault

Christopher W. Johnson[1], Claudia Hulbert[2], Bertrand Rouet-Leduc[1], Paul A. Johnson[1]

[1]Los Alamos National Laboratory, Geophysics Group, Los Alamos, N.M.
[2]Laboratoire de Géologie, Département de Géosciences, École Normale Supérieure, PSL Université, CNRS UMR 8538, Paris, France.

Key Points:

- Machine learning estimates daily LFE rate from statistical features of continuous seismic waveforms.
- Model estimates show high correlation with LFE bursts and long term trends in activity.
- Abundant information is available in seismic waveforms to characterize weak ground motions.

Corresponding author: Christopher W. Johnson, `cwj@lanl.gov`

Abstract

Low frequency earthquakes (LFEs) originating below the central San Andreas Fault are associated with slow-slip within the more ductile portion of the crust beneath the seismogenic zone. Monitoring efforts over 15 years recorded >1 million LFEs with >70 per day. We apply machine learning (ML) to statistical features describing the seismic waveforms and estimate the LFE daily intensity. Using 4 years of independent data, the ML model produces a 0.68 correlation. The burst-like LFE behavior is reproduced and the largest misfit occurs during the low-amplitude daily undulations. The ability to continuously monitor LFE activity provides insight to when geodetic measurements of slow slip are possible, without the need for developing a computational-intensive template-matching catalog. Similarities are found between detecting LFEs and tremors, which provides evidence tremors are composed of LFEs. The approach reveals by ML the rich information contained in the features of continuous seismic waveforms.

Plain Language Summary

Low frequency earthquakes (LFEs) are a class of events occurring beneath the section of a fault that produces strong ground shaking. This type of event has been observed along the central San Andrea Fault and occurs much more frequently than regular earthquakes. This study applies machine learning (ML) using statistical features derived from continuous daily seismic waveforms to train a ML model that is capable of estimating the daily LFE intensity. Inferring the daily rate of LFEs allows continuous monitoring of the fault zone using statistical features of daily seismic waveforms, without developing a computationally expensive LFE catalog. Bursts of these events are associated with deep slow-slip at the base of the fault that is integral to quantifying the entire slip budget. The ML model uses features that quantify the energy released and varying frequency content in daily seismic waveforms to estimate the LFE activity. Similarities are found between monitoring for LFEs and detecting tremors, providing evidence that tremors are composed of LFEs. The technique exemplifies the abundant information in seismic waveforms that is capable of training ML models to identify processes deep in the fault zone with the potential to extract more information related to slip events.

1 Introduction

Non-volcanic tremor is inferred to be the superposition of rapidly occurring low-frequency earthquakes (LFEs) that coincide with slow-slip on the lower-crustal fault interface where material behaves in a ductile-like manner (Shelly et al., 2007). Observations of this class of earthquake have provided insight to better understand how faults accommodate plate motions and allow discovery by informing physical models of the fault structure and frictional regime in the deep roots of a fault zone (Bürgmann, 2018; Rubinstein et al., 2009; Peng & Gomberg, 2010). The phenomenon was first observed in the Nankai trough subduction zone in Japan, downdip from the locked plate interface (Obara, 2002), and later in the Cascadia (Rogers & Dragert, 2003) and Mexican (Frank et al., 2013) subduction zones. Along the more shallow, crustal strike-slip (transform) San Andreas Fault (SAF) near Parkfield, California, Nadeau and Dolenc (2005) also observed non-volcanic tremor, which provided evidence of slow slip in tectonic environments other than subduction thrusts. Following these initial discoveries, slow slip is now observed at most major tectonic plate boundaries and is considered an significant percentage of the total slip budget (Jolivet & Frank, 2020).

Observational evidence suggests LFEs represent deep shear slip at the base of a fault zone and the continuous monitoring of LFE activity could serve as a proxy for slow slip (Shelly, 2017; Shelly et al., 2007). Non-volcanic tremors also originating from the deep fault are low amplitude seismic signals that contain bursts of energy in the 1-5 Hz range, but are depleted in higher frequencies and are believed to be composed of LFEs (Shelly et al., 2007). Tremor signals have been decomposed into individual LFEs using earthquake waveform techniques, e.g. template matching, to show the rapid succession of these events produce tremors (Shelly & Hardebeck, 2010). Near the Parkfield section of the SAF (Figure 1), the time and locations of LFEs are cataloged for 15 years of activity and provide a detailed record of deep crustal deformation (Shelly, 2017). The LFEs migrate along strike at rates up to 80 km/hr (Shelly, 2017; Shelly & Hardebeck, 2010; Shelly, 2010a), show episodic, near-continuous, and bimodal recurrent activity (Shelly, 2010b, 2010a), exhibit decoupled behaviour from the northern to southern sections of the fault (Trugman et al., 2015), and can be triggered by low amplitude stresses produced by tides and teleseismic earthquakes, suggesting a weak frictional environment (Thomas et al., 2012, 2009; Peng et al., 2009; van der Elst et al., 2016; Delorey et al., 2017). Observing complementary geodetic observations of deep slow-slip on the SAF is challenged by the low signal-to-noise ratio of GPS and InSAR measurements compared to the expected sub-millimeter displacements. At Parkfield, Rousset et al. (2019) quantify the average slow-slip moment release, equivalent to a M4.9 earthquake, by stacking all GPS measurements recorded during bursts of LFE activity with the highest daily rates. This geodetic observation does not quantify individual slow slip events, but does show bursts of LFE activity can be used as a proxy for deep slip on the SAF.

Developing the LFE catalog for the SAF utilizes waveform template matching with a 6 second LFE example to scan the entire local seismic network and identify individual events (Shelly, 2017). The ability to quantify the daily LFE intensity without compiling a complete catalog has the potential to provide insight into the physics of fault mechanics and potentially help constrain the slip budget of large magnitude earthquakes. Machine learning (ML) has shown the ability to predict the timing of laboratory earthquakes (Rouet-Leduc et al., 2017) and quantify the physics prior to the slip event in these experiments (Hulbert et al., 2019; Rouet-Leduc et al., 2018). In the Cascadia subduction zone, ML models are able to increase the detection potential of tremors (Rouet-Leduc et al., 2020), estimate the GPS measured surface displacement (Hulbert et al., 2020), and identify the release of seismic energy before the slow-slip events (Hulbert et al., 2020). In this study, we show a ML model can estimate the daily LFE rate on the SAF. The ML model is trained with statistical features describing the continuous seismic waveforms from a subset of local borehole seismic sensors. The final ML model estimates the daily LFE intensity directly from features of the waveforms. The application demonstrated here provides new evidence of the ability of ML models to identify weak sources of ground motion associated with LFEs and the potential to extract more information related to slow slip events.

2 Data and Methods

2.1 LFE Catalog and Daily Rates

The LFE catalog developed by Shelly (2017) contains more than 1 million events that occur at >15 km depth in the lower crust near Parkfield, California, which includes the transition from the northern creeping to the southern locked regions of this $\sim$160 km section of the SAF (Figure 1a). The events are distributed throughout 88 families at discrete locations that produce nearly identical waveforms and enable the detection of repeating families with template matching. The Parkfield section of the SAF has hosted numerous M$\sim$6 earthquakes, with the most recent in September 2004 (Bakun et al., 2005). The seismicity data along a 160 km transect and within 7.5 km of the fault shows much more activity in the northern creeping section (Figure 1b). The brittle to ductile transition is estimated using the 95^{th} percentile of the seismicity depths along the fault and varies from about 9 km to the north and 15 km to the south. Similarly, the LFE families northwest of Parkfield are between 20-25 km depth and shallower when comparing to the 22-30 km depth to the southeast (Figure 1b).

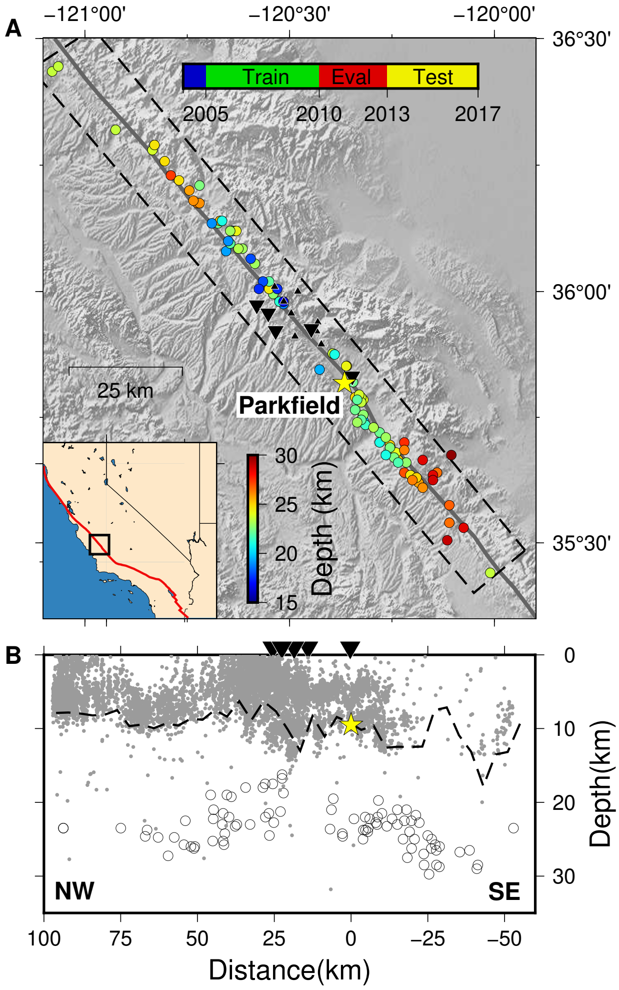

Figure 1. Central coastal range of California in map view. Inset shows the western U.S. with the San Andreas Fault in red and the study area indicated by the black box. (**A**) Creeping and locked section of the fault (gray line) near Parkfield, shown with a yellow star. Low frequency earthquake family locations are shown as circles with the depth indicated by color. Inverted black triangles are the HRSN seismic sensors used in the analysis; smaller black triangles show the entire HRSN network. The time periods of seismic data from the HRSN applied to model training, evaluating, and testing is shown between 2004-2017. (**B**) Depth profile showing seismicity (gray dots) within dashed box in **A** and low frequency earthquakes (open circles). The distance is relative to the Parkfield 2004 M6 hypocenter shown with a yellow star with northwest (NW) and southeast (SE) relative to map view. The dashed black line is the 95^{th} percentile of event depth along the fault indicating the transition to a more ductile environment.

We develop a daily LFE-intensity time-series that is compiled using all cataloged events between January 2004 and December 2016. The daily count ranges up to 2050 LFEs per day, with peak activity following the 2004 M6 Parkfield earthquake, and an average of 202 LFEs per day. Below the locked section south of Parkfield the daily average is 131 LFEs per day with 6% of the times exceeding twice the standard deviation from the mean. In this section of the SAF the LFEs waveforms have higher amplitudes and occur at a more steady rate (Nadeau & Dolenc, 2005; Shelly & Hardebeck, 2010). Below the creeping section to the north the average is 70 LFEs per day with 10% of the times exceeding twice the standard deviation from the mean. Here the LFEs exhibit more burst-like activity that was used to constrain the geodetic observations (Rousset et al., 2019).

2.2 Seismic Waveforms

The High-Resolution Seismic Network (HRSN; BP network) is a permanent array of 13 closely spaced borehole seismometers located near Parkfield and operated by the Berkeley Seismological Laboratory (Figure 1). The network is designed to enhance microseismicity detection along the SAF and is used in the development of the LFE catalog (Shelly, 2017). We use 5 stations (EADB, FROB, SCYB, SMNB, and VCAB) that perform well when developing the LFE catalog (Shelly, 2017), and obtain all available 3-channel (DP; 500 sample per second) daily records between 2004 and 2016. We reverse the polarity and perform channel swaps following the corrections documented in Shelly (2017). The data is preprocessed by deconvolving the instrument response function to obtain waveforms in the native m/s units. From 2010 to 2013 some instruments were upgraded with gain amplifiers to improve small event detection, but not all instrument response files were correctly documented, which can produce inconsistent waveform amplitudes after deconvolving the instrument response function. Days containing multiple file segments for the entire day are used and any gaps between segments are filled with zeros. Days with inconsistent channel recordings or only partial waveforms records are discarded.

2.3 Data Features

Data features are calculated using the 3 channels of each sensor as follows. The waveforms are filtered with a 4^{th} order zero-phase Butterworth bandpass filter using corners of 1-4 Hz, 4-8 Hz, 8-12 Hz, and 12-16 Hz. For each filtered waveform the zero-crossing-rate, the 5-95%, 10-90%, 25-75%, 40-60% inter-quantile-range (IQR), the variance, the skew, the kurtosis, the min-max range, and the root-mean-squared are calculated. This produces 40 features for each channel, 120 features per day for each sensor, and 600 total for the 5 sensors (4 filters * 10 statistics * 3 channels * 5 sensors = 600 features). To develop a continuous time series with 600 features per day, sensors with missing daily waveforms are represented as a vector of 120 not-a-number (NaN) values when assembling the feature matrix. The values are scaled to unit variance using the standard deviation of the previous 15 days. Although the ML model is insensitive to scale differences between individual features, this technique scales the features consistently through time and removes amplitude variations from the equipment upgrades. Additionally, no future information is used to modify a point in the time series, unlike scaling by the standard deviation of the entire series. The short window length is selected to remove seasonality observed in the waveforms that could potentially bias the ML model. The scaled feature time series is split into training (N=1826), test (N=1096), and blind test (N=1461) data sets. Prior to splitting, shuffling is not applied to retain the temporal behavior inherent to the data.

2.4 Gradient Boosted Tree ML Model

We develop a ML model based on gradient boosted trees (XGBoost package; Chen & Guestrin, 2016) that is designed as a regression analysis to estimate the daily LFE intensity from 600 statistical features of the waveforms for that day. The ML model is trained using 5 years of data from January 2005 to December 2009 and the performance is evaluated throughout the training process using 3 years of test data from January 2010 to December 2012. We fit 9 hyperparameters (max_depth, learning_rate, n_estimators, gamma, min_child_weight, subsample, colsample_bytree, reg_alpha, and reg_lambda) using a Bayesian optimizer (scikit-optimize package; Head et al., 2018). Determining the best combination of hyperparameters is an iterative process and requires training thousands of ML models. The hyperparameter optimizer is updated using the average Pearson's cross correlation coefficient from the training data 5 fold cross validation. The best fit hyperparameters obtained from the cross validation are applied to the test data, which allows a quantitative metric to further constrain the search space during additional model training to converge at a global minimum. This procedure ensures an unbiased metric when reporting the performance, but produces data leakage since the best-fit parameters are unintentionally tuned to the test data. The final analysis uses the blind-test data set between 2013 and 2016 to evaluate the ML model.

3 Results

3.1 Model Training, Testing, and Blind-Test

The results for the 3 data sets are shown as the LFE intensity versus the model estimate, and quantified with the Pearson cross correlation and R^2 values (Figure 2). The correlation metric describes the similarity in the shape of the curves and the R^2 value describes the variance between the known values and model estimates, which is consistently lower compared to the correlation. The training data used in the 5 fold cross validation has a range between 0.75 and 0.80 correlation values for each year (Figure S1), and a 0.73 correlation value and 0.54 R^2 value for the entire 5 year period (Figure 2a). The test dataset results are consistent with the training and have a 0.72 correlation value and 0.52 R^2 value (Figure 2b). When viewing each year from 2010 to 2011 separately (Figure S2), the correlation value decreases annually from 0.77 to 0.68, which coincides with the network upgrades (Shelly, 2017). The training and test results show the longer wavelength undulation and LFE bursts are reproduced by the model, with the largest discrepancy observed in estimating the higher-frequency lower-amplitude variations (Figures S1 and S2). Consistent values are obtained when using 4 and 7 splits in the cross validation to vary the subsets of data used in each validation. For this data set, the hyperparameters (Test S1) are robust to develop a model that estimates the LFE intensity from the waveform statistical features.

After the training and testing is complete, the ML model is applied to the blind-test dataset and a 0.69 correlation value and 0.45 R^2 value are reported (Figure 2c). The correlation values range from 0.54-0.75 if the test data set is separated into the individual years, with 2013 and 2016 showing the lowest correlation (Figure S3). The time series of the first 180 days in 2015 show the model performance (Figure 3). Qualitatively it captures the multi-day rate changes and adequately estimates the bursts of LFE activity, but does not always correctly capture the higher frequency variations. The results indicate the model is over estimating the LFE intensity when the observed daily rate is $<$100 LFEs per day. This is shown in the density plot with the highest concentration of points below the $\pm$50 interval (Figure 2c) and observed in the blind-test time series (Figure S3). Some of the time intervals that are over estimated coincide with periods of missing seismic data. Station FROB and SMNB are problematic during days 1-20 in 2015 and the model estimate is above the near zero LFE intensity reported in the catalog (Figure 3). Beginning in 2010 long periods of network degradation occur more frequently (Figure

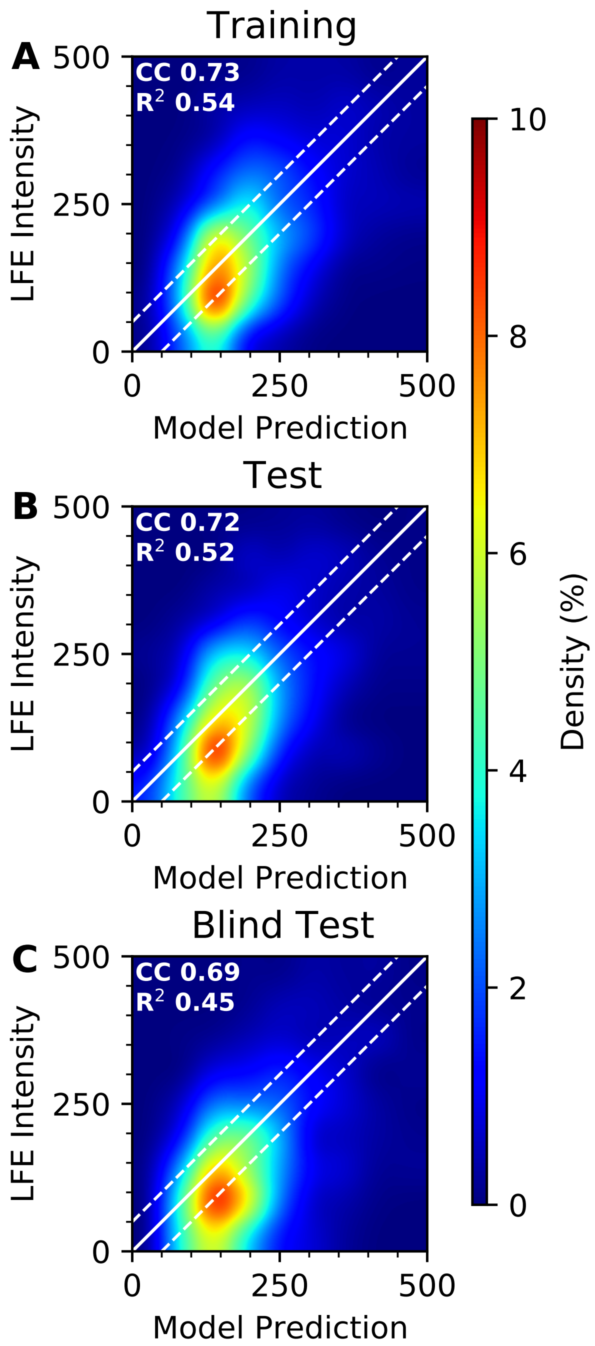

Figure 2. Density plot showing the model estimate versus the cataloged number of LFEs per day for the (**A**) training, (**B**) test, and (**C**) blind test datasets. The white line shows the 1:1 correlation and is bounded by ±50 shown as the white dashed line. The Pearson's cross correlation and R^2 values are listed for each in the upper left.

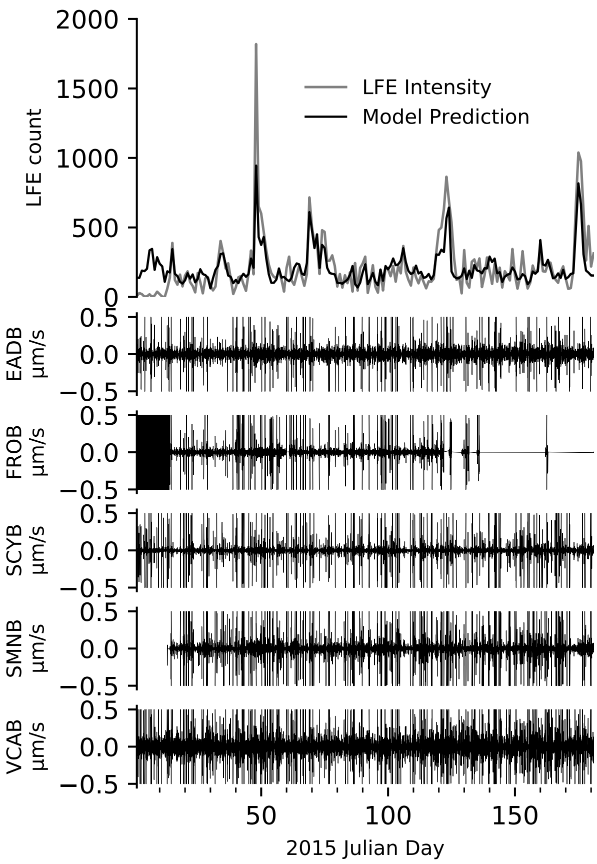

Figure 3. LFE intensity shown with 5 seismic waveform examples. The top curve is the LFE daily intensity from the catalog (grey line) shown with the model estimate (black line) for days 1-180 of 2015. The 5 waveform traces shown below are the horizontal channel (DP2) for each sensor used to calculate the statistical features. The vertical axis is clipped to highlight the amplitude variations in noise.

3 in Shelly, 2017), suggesting the ML model is correctly estimating the LFE activity when the template matching was unable to detect all events. To test if the model is over estimating because is it trained using data when the network is performing best, we do the opposite and train a model with the same hyperparameters using data from 2013.5 to 2017 and use 2013.0 to 2013.5 to estimate the LFE intensity. The results show an increase in correlation value from 0.55 to 0.66 for this time interval (Figure S4) indicating the model trained with the best data is most likely estimating an accurate LFE intensity, even if the network performance decreases.

3.2 Feature Importance

Tree based ML model architectures have the benefit of quantifying the feature importance to interpret which information is most influential in the model output. The feature importance is quantified using the SHAP summary value (Lundberg & Lee, 2017) to report the contribution and a positive or negative correlation with the target variable (Figure 4). The SHAP values for the 2 most influential features are more than 2 times greater than the others, but that does not indicate causality, only how the model is obtaining information to perform best. The most influential feature is the 5-95% IQR from

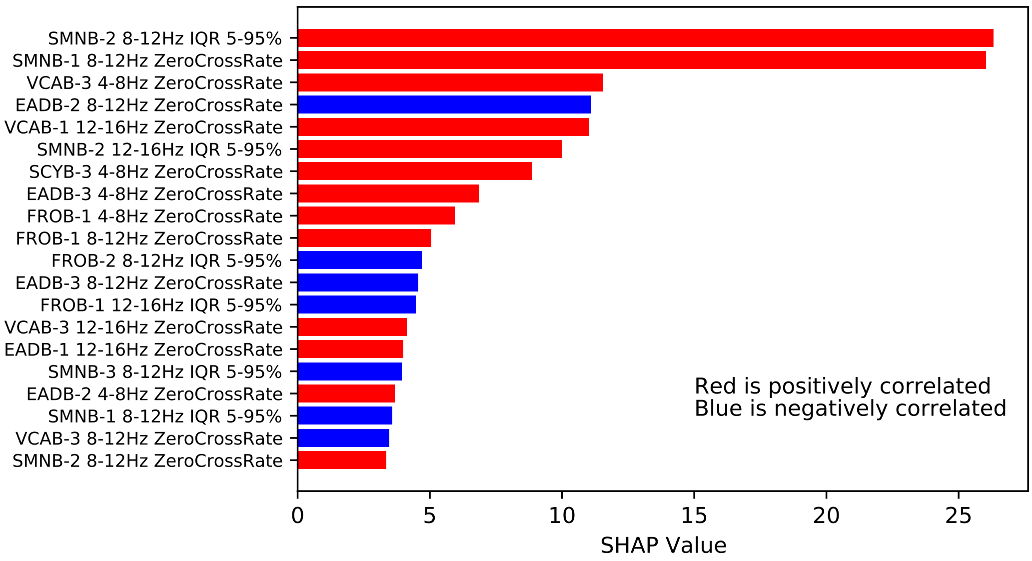

Figure 4. SHAP metric showing feature importance with the more influential features having a greater value. The 20 that contribute most to the model are shown in descending order with the feature name listed on the vertical axis. Features positively correlating are in red and negatively correlating in blue.

station SMNB on channel DP2 in the 8-12 Hz bandpass with a SHAP value of 27 that is positively correlated. Inspection of the feature time-series indicates similarities with the LFE intensity but it does not, as expected, match peak-for-peak because of the non-linear relationship the model develops using information from the entire dataset. The features ranked 2 through 5 are all zero crossing rate in the 4-8 Hz, 8-12 Hz, and 12-16 Hz bandpass and 3 of them are positively correlated. The 4^{th} ranked feature is the zero crossing rate in the 8-12 Hz bandpass and correlates negatively with the LFE rate, indicating this value is informing the model of when not to expect LFEs. Regardless of station or channel, the top 20 features are 5-95% IQR or zero crossing rate, 13 correlate positively and 7 negatively, and 5 are in the 4-8 Hz bandpass without any in the 1-4 Hz bandpass. None of the central-moment statistics appear in the top 20 best features listed.

4 Discussion

The ability of ML when applied to seismic data from laboratory shear experiments to infer instantaneous and future behavior demonstrates that signals are emitted throughout the stress loading cycle (Rouet-Leduc et al., 2017; Lubbers et al., 2018; Rouet-Leduc et al., 2018). Features of the seismic signals illuminate pre-failure slip characteristics by identifying continuous micro-failures (Hulbert et al., 2019). As shown here, the statistical representation of seismic waveforms at Parkfield contains rich information regarding daily LFE intensity. The LFE intensity is thought to be a manifestation of micro-failure evolution on the deep portion of the slowly-slipping fault, similar to laboratory studies. Sensitivity tests incorporating station dropout or using a single station (the HRSN station VCAB and broadband station PKD were modeled for single station analysis) produce similar results, but manifest a reduction in LFE burst intensity. This is logical considering the sources are distributed along a 160 km section of the fault, and illustrates that a larger spatial sampling of features is required to capture the diverse LFE activity (Figure 1).

The less impactful features are in the 1-4 Hz bandpass, suggesting the ML model is identifying information to quantify the LFE intensity outside the spectral range typically associated with LFEs. This observation is supported by the zero-crossing rate in the 8-12 Hz bandpass as being an important feature since LFEs are depleted in energy above 10 Hz (Obara, 2002). The best features reported from ML models analyzing laboratory shear data are similar to those found here (Rouet-Leduc et al., 2018), suggesting that fault frictional characteristics are similar across multiple scales. Additionally, applying a variation of this method to slow slip in Cascadia shows that tremor activity is best characterized by the IQR in the 8-13 Hz bandpass (Rouet-Leduc et al., 2019; Hulbert et al., 2020). Further, a deep learning model trained using the frequency content of tremor with seismic data from Cascadia is able to identify tremor on the SAF near Parkfield (Rouet-Leduc et al., 2020). The deep learning model does not provide a specific best-feature due to the different model design, but does highlight the strong similarities between detecting tremors and LFEs. This collection of results suggests that a characteristic acoustic release of energy across multiple scales and tectonic environments is responsible for both tremors and LFEs. Indeed, the results show IQRs from 4-16 Hz map to the LFE intensity, identifying a statistical relationship between LFEs and tremor, providing new evidence for tremor being comprised of LFEs.

The technique presented here quantifies the LFE daily intensity with the goal of learning what information contained in seismic waveforms is relevant to forecasting instantaneous seismic activity. The statistical features applied to the ML model provide a snapshot of the physics recorded in the waveforms that are emitted in this low frictional environment. The daily sampling applied here filters the information into a 24-hour windows using all LFE families along the fault and possibly obscures useful characteristics contained in the $<$10 s LFE waveforms.

Applying instantaneous features, we also attempted to forecast the future LFE intensity. The results produced poor predictions, especially during the LFE bursts. This suggests that we must isolate LFE sources along the 160 km fault segment to test whether or not future behavior can be forecast for single source locations.

A limitation to the model was degraded performance during the aftershock sequence of the 2004 M6 Parkfield earthquake. The ML model captures the LFE increase, but underestimates the multi-month elevated activity (Figure S5). Since the ML model was not trained using a data set containing LFEs triggered from a large magnitude event, the waveform statistical properties of this type of activity will not be learned by the model.

The problem we describe is challenging because of the spatially synchronous behavior of LFE families that can produce simultaneous emissions at source locations spatially unrelated (Trugman et al., 2015), and the frequent earthquakes occurring along the creeping section of the fault. For these reasons the central SAF presents unique conditions in contrast to other regions where related problems are explored, e.g., tremor and slow-slip in Cascadia (Rouet-Leduc et al., 2019, 2020; Hulbert et al., 2020). Nevertheless, the ML model extracts the LFE intensity with a high correlation to the known rate and suggests the engineered features utilized are sufficient to characterize the slip behavior of the evolving fault system. It will be interesting to apply the trained ML model to other tectonic environments and learn if it generalizes to an efficient approach for monitoring LFE activity without retraining, or utilizing template based signal processing. Similar LFE analyses across different tectonic regions and faulting styles may provide additional insight into consistent and varying LFE, tremor, and slow-slip characteristics. Our results underscore the power of ML in seismic signal analysis, complimenting previous studies extracting new information from seismic waveforms (Rouet-Leduc et al., 2017, 2018; Lubbers et al., 2018; Rouet-Leduc et al., 2019, 2020; Hulbert et al., 2020).

5 Conclusion

We develop a ML model to estimate the daily LFE intensity on the central SAF using statistical features of seismic waveforms. The model is trained using the LFE catalog containing >1 million events to develop a daily rate and 5 borehole seismometers to calculate features representing characteristics of the waveforms. The ML model gets a correlation of 0.68 when applied to a blind-test data set. The largest misfit is observed when the cataloged LFE rate is <100 per day. Tests during periods of seismic station malfunction indicate the ML model is reporting an increased rate more consistent with long term activity. Similarities with the statistical features that best describe the LFE intensity are observed between other ML models that identify tremors and provide evidence tremors are composed of LFEs.

Acknowledgments

We thank Nickolas Lubbers for insightful discussions and Taka'aki Taira for information about the HRSN waveform data. All data is publicly available through the Northern California Earthquake Data Center (http://ncedc.org/). CWJ and BR-L acknowledge support from Institutional Support (LDRD) at Los Alamos. CH acknowledges support from the joint research laboratory effort in the framework of the CEA-ENS Yves Rocard LRC (France), and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Geo-4D project, grant agreement 758210). PAJ acknowledges support of this work by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division under grant 89233218CNA000001.

References

Bakun, W., Aagaard, B., Dost, B., Ellsworth, W., Hardebeck, J., Harris, R., ... others (2005). Implications for prediction and hazard assessment from the 2004 parkfield earthquake. *Nature*, *437*(7061), 969–974.

Bürgmann, R. (2018). The geophysics, geology and mechanics of slow fault slip. *Earth and Planetary Science Letters*, *495*, 112–134.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Delorey, A. A., van der Elst, N. J., & Johnson, P. A. (2017). Tidal triggering of earthquakes suggests poroelastic behavior on the san andreas fault. *Earth and Planetary Science Letters*, *460*, 164 - 170. doi: 10.1016/j.epsl.2016.12.014

Frank, W. B., Shapiro, N. M., Kostoglodov, V., Husker, A. L., Campillo, M., Payero, J. S., & Prieto, G. A. (2013). Low-frequency earthquakes in the mexican sweet spot. *Geophysical Research Letters*, *40*(11), 2661–2666.

Head, T., MechCoder, Louppe, G., Shcherbatyi, I., fcharras, Vinícius, Z., ... Fabisch, A. (2018). *scikit-optimize/scikit-optimize: v0.5.2.* Zenodo. doi: 10.5281/zenodo.1207017

Hulbert, C., Rouet-Leduc, B., Johnson, P. A., Ren, C. X., Rivière, J., Bolton, D. C., & Marone, C. (2019). Similarity of fast and slow earthquakes illuminated by machine learning. *Nature Geoscience*, *12*(1), 69–74.

Hulbert, C., Rouet-Leduc, B., Jolivet, R., & Johnson, P. A. (2020). An exponential build-up in seismic energy suggests a months-long nucleation of slow slip in cascadia. *Nature Communications*, *11*(1), 4139. doi: 10.1038/s41467-020-17754-9

Jolivet, R., & Frank, W. B. (2020). The transient and intermittent nature of slow slip. *AGU Advances*, *1*(1). doi: 10.1029/2019AV000126

Lubbers, N., Bolton, D., Mohd-Yusof, J., Marone, C., Barros, K., & Johnson, P. (2018). Earthquake catalog-based machine learning identification of laboratory fault states and the effects of magnitude of completeness. *Geophysical Research Letters*, *45*. doi: 10.1029/2018GL079712

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc.

Nadeau, R. M., & Dolenc, D. (2005). Nonvolcanic tremors deep beneath the san andreas fault. *Science*, *307*(5708), 389–389. doi: 10.1126/science.1107142

Obara, K. (2002). Nonvolcanic deep tremor associated with subduction in southwest japan. *Science*, *296*(5573), 1679–1681. doi: 10.1126/science.1070378

Peng, Z., & Gomberg, J. (2010). An integrated perspective of the continuum between earthquakes and slow-slip phenomena. *Nature geoscience*, *3*(9), 599–607.

Peng, Z., Vidale, J., Wech, A., Nadeau, R., & Creager, K. (2009). Remote triggering of tremor around the parkfield section of the san andreas fault. *J. Geophys. Res*, *114*, B00A06.

Rogers, G., & Dragert, H. (2003). Episodic tremor and slip on the cascadia subduction zone: The chatter of silent slip. *Science*, *300*(5627), 1942–1943. doi: 10.1126/science.1084783

Rouet-Leduc, B., Hulbert, C., Bolton, D. C., Ren, C. X., Riviere, J., Marone, C., ... Johnson, P. A. (2018). Estimating fault friction from seismic signals in the laboratory. *Geophysical Research Letters*, *45*(3), 1321–1329.

Rouet-Leduc, B., Hulbert, C., & Johnson, P. A. (2019). Continuous chatter of the cascadia subduction zone revealed by machine learning. *Nature Geoscience*, *12*(1), 75–79.

Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C. J., & Johnson, P. A. (2017). Machine learning predicts laboratory earthquakes. *Geophysical Research Letters*, *44*(18), 9276–9282.

Rouet-Leduc, B., Hulbert, C., McBrearty, I. W., & Johnson, P. A. (2020). Probing slow earthquakes with deep learning. *Geophysical Research Letters*, *47*(4), e2019GL085870.

Rousset, B., Bürgmann, R., & Campillo, M. (2019). Slow slip events in the roots of the san andreas fault. *Science advances*, *5*(2).

Rubinstein, J. L., Shelly, D. R., & Ellsworth, W. L. (2009). Non-volcanic tremor: A window into the roots of fault zones. In *New frontiers in integrated solid earth sciences* (pp. 287–314). Springer.

Shelly, D. R. (2010a). Migrating tremors illuminate complex deformation beneath the seismogenic san andreas fault. *Nature*, *463*(7281), 648–652.

Shelly, D. R. (2010b). Periodic, chaotic, and doubled earthquake recurrence intervals on the deep san andreas fault. *Science*, *328*(5984), 1385–1388.

Shelly, D. R. (2017). A 15 year catalog of more than 1 million low-frequency earthquakes: Tracking tremor and slip along the deep san andreas fault. *Journal of Geophysical Research: Solid Earth*, *122*(5), 3739–3753.

Shelly, D. R., Beroza, G. C., & Ide, S. (2007). Non-volcanic tremor and low-frequency earthquake swarms. *Nature*, *446*(7133), 305–307.

Shelly, D. R., & Hardebeck, J. L. (2010). Precise tremor source locations and amplitude variations along the lower-crustal central san andreas fault. *Geophysical Research Letters*, *37*(14).

Thomas, A. M., Burgmann, R., Shelly, D., Beeler, N., & Rudolph, M. (2012). Tidal sensitivity of low frequency earthquakes near parkfield, ca: Implications for fault mechanics within the brittle-ductile transition. *J. Geophys. Res*, *117*, B05301.

Thomas, A. M., Nadeau, R. M., & Bürgmann, R. (2009). Tremor-tide correlations and near-lithostatic pore pressure on the deep san andreas fault. *Nature*, *462*(7276), 1048–1051.

Trugman, D. T., Wu, C., Guyer, R. A., & Johnson, P. A. (2015). Synchronous low frequency earthquakes and implications for deep san andreas fault slip. *Earth and Planetary Science Letters*, *424*, 132–139.

van der Elst, N. J., Delorey, A. A., Shelly, D. R., & Johnson, P. A. (2016). Fortnightly modulation of san andreas tremor and low-frequency earthquakes. *Proceedings of the National Academy of Sciences*, *113*(31), 8601–8605. doi: 10.1073/pnas.1524316113

Supporting Information for "Learning the low frequency earthquake daily intensity on the central San Andreas Fault"

Christopher W. Johnson[1], Claudia Hulbert[1,2], Bertrand Rouet-Leduc[1], Paul A. Johnson[1]

[1]Los Alamos National Laboratory, Geophysics Group, Los Alamos, N.M.

[2]Laboratoire de Géologie, Département de Géosciences, École Normale Supérieure, PSL Université, CNRS UMR 8538, Paris, France.

Supporting information includes the final model hyperparametets and the full time series plots of LFE intensity from catalog and ML model.

Contents of this file

1. Text S1
2. Figures S1 to S5

Text S1. The hyperparameter search space was refined to allow an initial large range, then systematically narrowed to avoid overfitting the training data. The final model is selected after the Gaussian optimizer converges on a set of hyperparameters and the average correlation value from the 5 fold cross validation stabilizes. The best-fit model hyperparameters are max_depth = 4, learning rate = 0.039, n_estimators = 688, gamma = 0, min_child_weight = 28.73, subsample = 0.764, colsample_bytree = 0.9, reg_alpha =

150, and reg_lambda = 59.668 which produce the highest correlation with the training data.

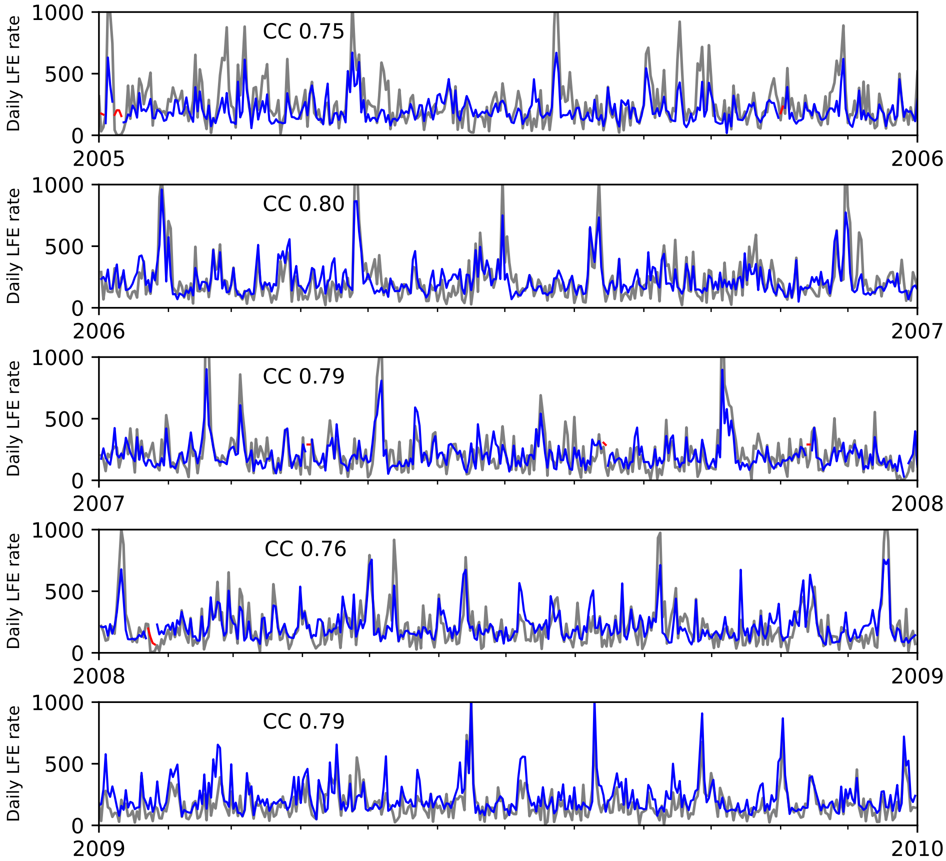

Figure S1. Training cross validation results. Each 1 year period is held out and the model is trained using the remaining 4 year. Shown are the model prediction in blue with the LFE intensity shown in black. Predictions in red indicate >50% of the data features are missing. The Pearson's cross correlation is shown for each time window

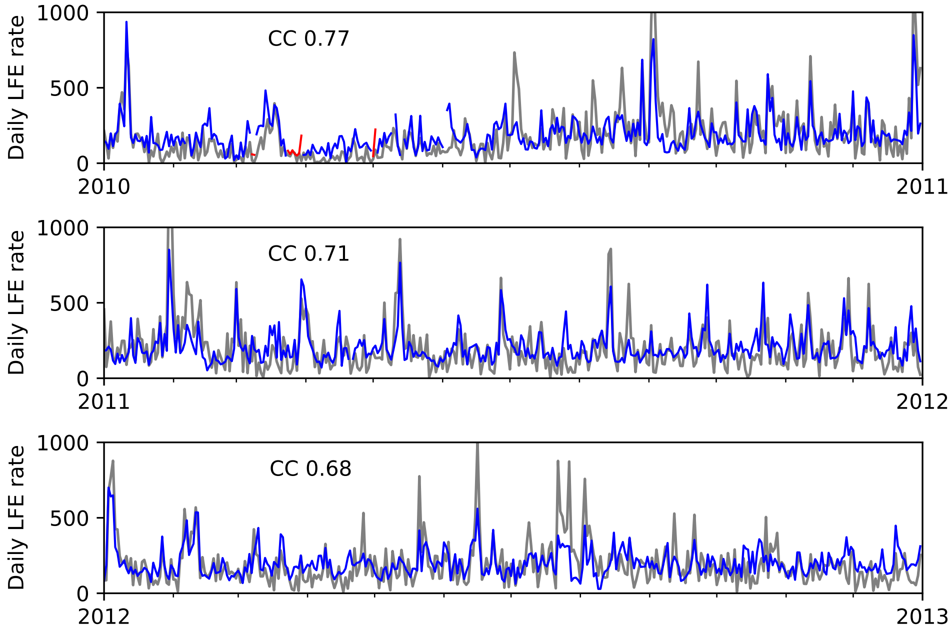

Figure S2. Test data results shown for each 1 year period. Shown are the model prediction in blue with the LFE intensity shown in black. Predictions in red indicate >50% of the data features are missing. The Pearson's cross correlation is shown for each time window

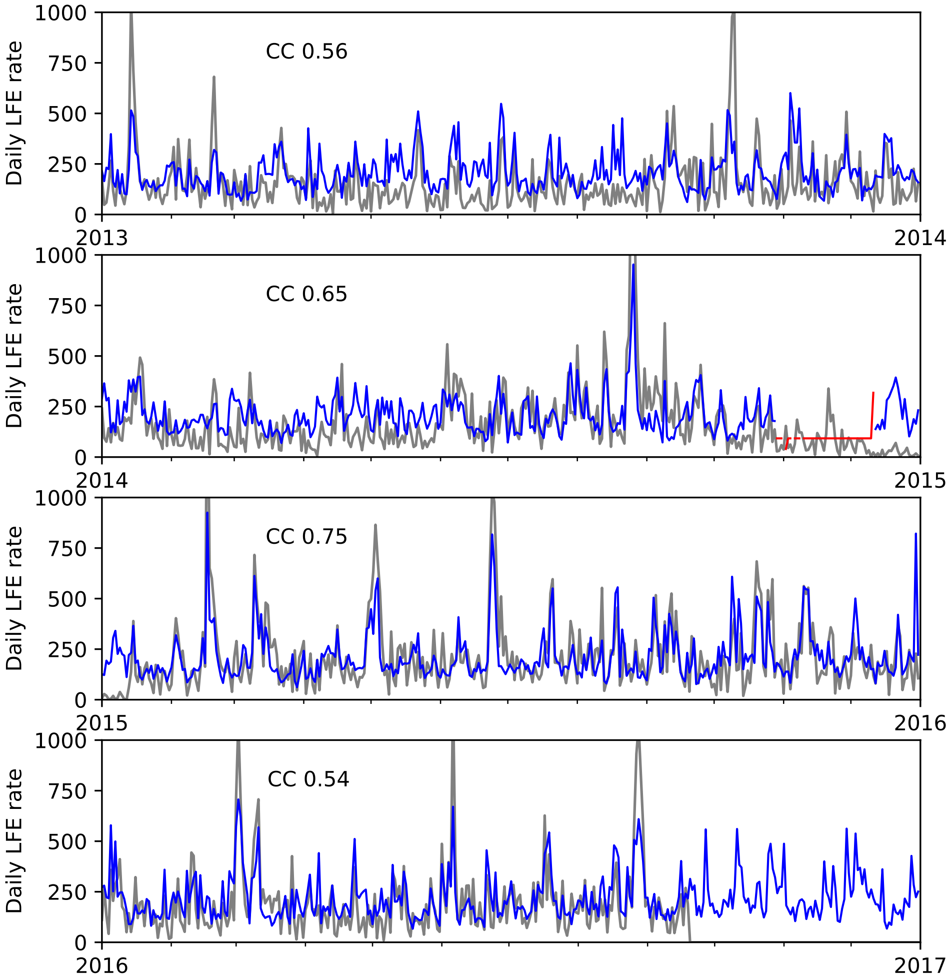

Figure S3. Blind-test results shown for each 1 year period. Shown are the model prediction in blue with the LFE intensity shown in black. Predictions in red indicate >50% of the data features are missing. The Pearson's cross correlation is shown for each time window

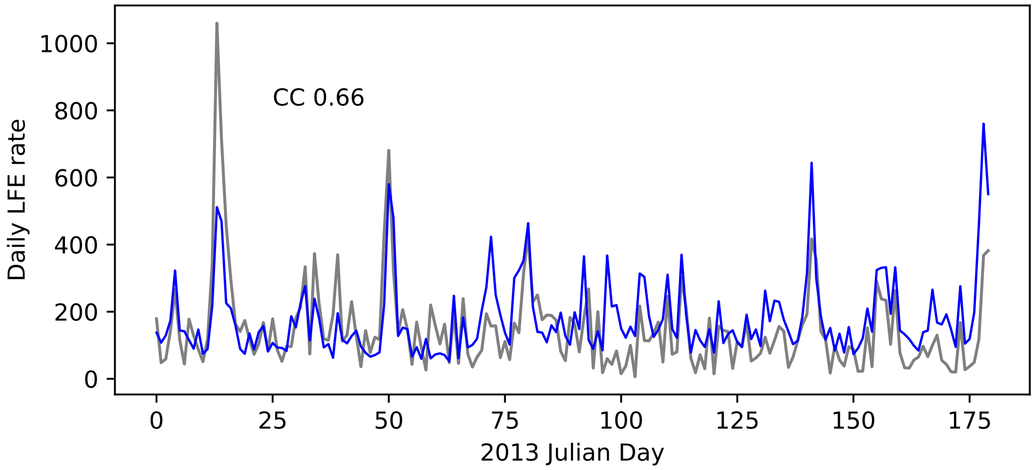

Figure S4. Evaluating the model performance when trained using data from periods when the seismic network is degrading. Shown are the model prediction in blue with the LFE intensity shown in black for the first 180 days for 2013. The Pearson's cross correlation is shown for each time window

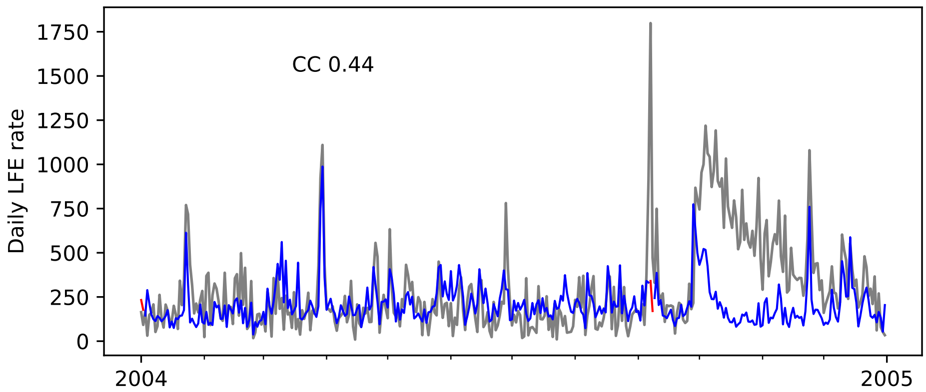

Figure S5. Results shown for 2004. Shown are the model prediction in blue with the LFE intensity shown in black. Predictions in red indicate >50% of the data features are missing. The Pearson's cross correlation is shown for each time window