Geophysical Research Letters

RESEARCH LETTER
10.1029/2019GL086615

Predicting Imminence of Analog Megathrust Earthquakes With Machine Learning: Implications for Monitoring Subduction Zones

F. Corbi[1,2,3], J. Bedford[3], L. Sandri[4], F. Funiciello[2], A. Gualandi[5], and M. Rosenau[3]

[1]Department of Earth Sciences, Institute of Geological Sciences, Freie Universität Berlin, Berlin, Germany, [2]Dip. Scienze, Laboratory of Experimental Tectonics, Università "Roma TRE", Rome, Italy, [3]Helmholtz Centre Potsdam - GFZ German Research Centre for Geosciences, Potsdam, Germany, [4]Istituto Nazionale di Geofisica e Vulcanologia, Bologna, Italy, [5]Department of Geology and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA

Abstract Subduction zones are monitored using space geodesy with increasing resolution, with the aim of better capturing the deformation accompanying the seismic cycle. Here, we investigate data characteristics that maximize the performance of a machine learning binary classifier predicting slip-event imminence. We overcome the scarcity of recorded instances from real subduction zones using data from a seismotectonic analog model monitored with a spatially dense, continuously recording onshore geodetic network. We show that a 70–85 km-wide coastal swath recording interseismic deformation gives the most important information on slip imminence. Prediction performances are mainly influenced by the alarm duration (amount of time that we consider an event as imminent), with density of stations and record length playing a secondary role. The techniques developed in this study are most likely applicable in regions of slow earthquakes, where stick-slip-like failures occur at time intervals of months to years.

Plain Language Summary Machine learning, a group of algorithms that produce predictions based on past "experience," has been successfully used to predict various aspects of the earthquake process, including slip imminence. The accuracy of those algorithms depends on a variety of data characteristics, for example, the amount of data used for building the "experience" of the model. We focus on this point using a scaled representation of a seismic subduction zone and a monitoring technique similar to Global Navigation Satellite System. We identify the most useful surface regions to be monitored and the parameter that most strongly influences prediction accuracy for the timing of upcoming laboratory earthquakes. The routine implemented in this study could be used to predict the onset and extent of slow earthquakes.

Key Points:
- We investigate the performances of a binary classifier predicting slip-event imminence in analog models of megathrust seismic cycling
- A 70–85 km-wide coastal swath is the region producing the most important information for the imminence classification
- Length of time that we consider an event imminent plays a primary role in tuning the performances of a binary classifier predicting the imminence of analog earthquakes

Supporting Information:
- Supporting Information S1

Correspondence to:
F. Corbi,
fabio.corbi3@gmail.com

Citation:
Corbi, F., Bedford, J., Sandri, L., Funiciello, F., Gualandi, A., & Rosenau, M. (2020). Predicting imminence of analog megathrust earthquakes with machine learning: Implications for monitoring subduction zones. *Geophysical Research Letters*, *47*, e2019GL086615. https://doi.org/10.1029/2019GL086615

Received 10 DEC 2019
Accepted 9 MAR 2020
Accepted article online 12 MAR 2020

1. Introduction

The preparatory phase of large subduction earthquakes can be depicted as a period of slow, continuous stress accumulation caused by the frictional interaction between converging plates (e.g., Hyndman et al., 1997). However, as geodetic (Global Navigation Satellite System [GNSS]) observation networks have matured, it has become apparent that there are significant variations of interplate locking before and/or after large earthquakes. These include transient slow slip events, indicating sub seismic-cycle scale, short (days to months) variations in the rates of stress accumulation/release (Heki & Mitsui, 2013; Loveless & Meade, 2016; Mavrommatis et al., 2014; Melnick et al., 2017). Such variations result in nonsteady (transient) surface motions measurable with space geodesy. Stress variations prior to large earthquakes may manifest as a series of foreshocks gradually unzipping the plate interface—as in the case of the 2014 Iquique M 8.1 earthquake (Schurr et al., 2014)—or as accelerating aseismic creep—as suggested for the 2011 Tohoku M 9.0 earthquake (Kato et al., 2012; Mavrommatis et al., 2014). The recognition of these pre-earthquake transients raises the potential for using them as a diagnostic tool for earthquake imminence. However, the scarcity of recorded instances hinders understanding whether and which transient signal may be used as a reliable indicator for earthquake prediction.

Machine learning (ML) represents a group of algorithms efficient in identifying indicators and not-so-obvious ("hidden") patterns in large data sets ("big data"). The possibility to train an algorithm and use it for making accurate predictions based on the "past experience" is one of the complex tasks that ML

can achieve (e.g., Bergen et al., 2019). Recently, the earthquake research community has demonstrated such capability of ML to draw inferences about fault physics: The acoustic signal emitted by rock samples sheared in a direct shear apparatus has been used for predicting the onset time of laboratory earthquakes (Rouet-Leduc et al., 2017), for estimating the instantaneous fault analog friction (Rouet-Leduc et al., 2018), and for predicting earthquake slip mode (Hulbert et al., 2019). Changing scale from laboratory to nature, ML has been used to identify a tremor-like signal emitted by Cascadia's megathrust that tracks the instantaneous displacement rate measured by a GNSS station (Rouet-Leduc et al., 2019).

Accordingly, we have possibly entered a new era of seismological discovery in which the full spectrum of transient signals from seismogenic faults is identified with ML and in which ML might diagnose imminence of events such as slow slip or even earthquakes. In other words, ML might be able to recognize and classify a set of processes or a characteristic pattern diagnostic of the imminence of fault failure. To gather the maximum benefit from such a novel approach, some technical points regarding the type and characteristics of data to use should be addressed in advance. For geodetic networks, the key questions would be as follows: Which region of the subduction margin is the most diagnostic? How important is the space-time data coverage? How far in advance can coseismic slip be predicted?

To address these questions, we experiment here with ML binary classifiers for predictions of earthquake imminence using GNSS-like surface deformation data from seismotectonic scale (analog) modeling (Rosenau et al., 2017). Analog modeling allows us to experimentally overcome the lack of long time series from real subduction zones, with a smaller physical model mimicking a multicentury history of seismic cycling in a few minutes in the lab (see Hubbert, 1937, for theory of scaling in analog modeling). Such analog model reproduces the basic features of earthquakes and seismic cycles, namely the elastic loading and release of a frictional fault embedded in an elastic medium. Indeed, our model is strongly simplified with respect to the natural prototype, where additional factors such as pore fluid pressure (e.g., Moreno et al., 2014; Moreno et al., 2018) or off-megathrust fault networks (e.g., Wang & Bilek, 2011) might control seismicity. Nevertheless, similar models have been shown to successfully reproduce complexity by means of the intrinsic variability of, for example, recurrence and slip patterns of natural systems to first order (e.g., Corbi et al., 2013; Corbi et al., 2017; Rosenau et al., 2010, 2019). As in Corbi et al. (2019a), we here use the surface deformation time series of such a model for ML-based analysis. ML provides a versatile tool for testing how various data characteristics influence prediction performances. With respect to a previous study by Corbi et al. (2019a), we here impose the following modifications:

1. Instead of framing the scientific problem with regression of time to failure, here we step toward binary classification of alarm state—that is, the ML predicts whether a given deformation field is characteristic of the few seconds that precede slip onset or not. Binary classification has the advantage of easy to interpret metrics for evaluating the prediction performances.
2. Instead of using an ideal, uniformly spaced GNSS network extending all the way to the trench, here we exclude data coming from above the analog offshore seismogenic zone to mimic limitations in geodetic coverage at subduction zones (i.e., we use only data from above the inland aseismic zone). Hence, we assume that the base of the seismogenic zone coincides with the coastline (Ruff & Tichelaar, 1996; Saillard et al., 2017; Figure 1a).

We show that ML can be used for identifying the most informative region of the convergent margin regarding imminent asperity failure and provides a useful indication of how many measurement points are needed on the surface as well as the optimal record length.

2. Data, Method, and Metrics

Data used in this study are derived from a seismotectonic analog model that represents a subduction zone characterized by two asperities of equal size and friction (Figure 1a; Corbi et al., 2017; details on the setup are in Supporting Information S1). The model produces analog earthquakes equivalent to magnitude Mw 6.2–8.3 when scaled to nature, with a coefficient of variation in recurrence intervals of 0.5, similar to real subduction zones (Williams et al., 2019). Here we use the data of Corbi et al. (2019a) that are available open access in Corbi et al. (2019b). Data consist of a 400 s recording of incremental surface displacement (trench parallel and orthogonal components; Figure 1b) capturing 40 seismic cycles. Displacement is measured with a precision of few tens of μm (i.e., few tens of m when scaled to nature) and at a resolution of few mm spacing

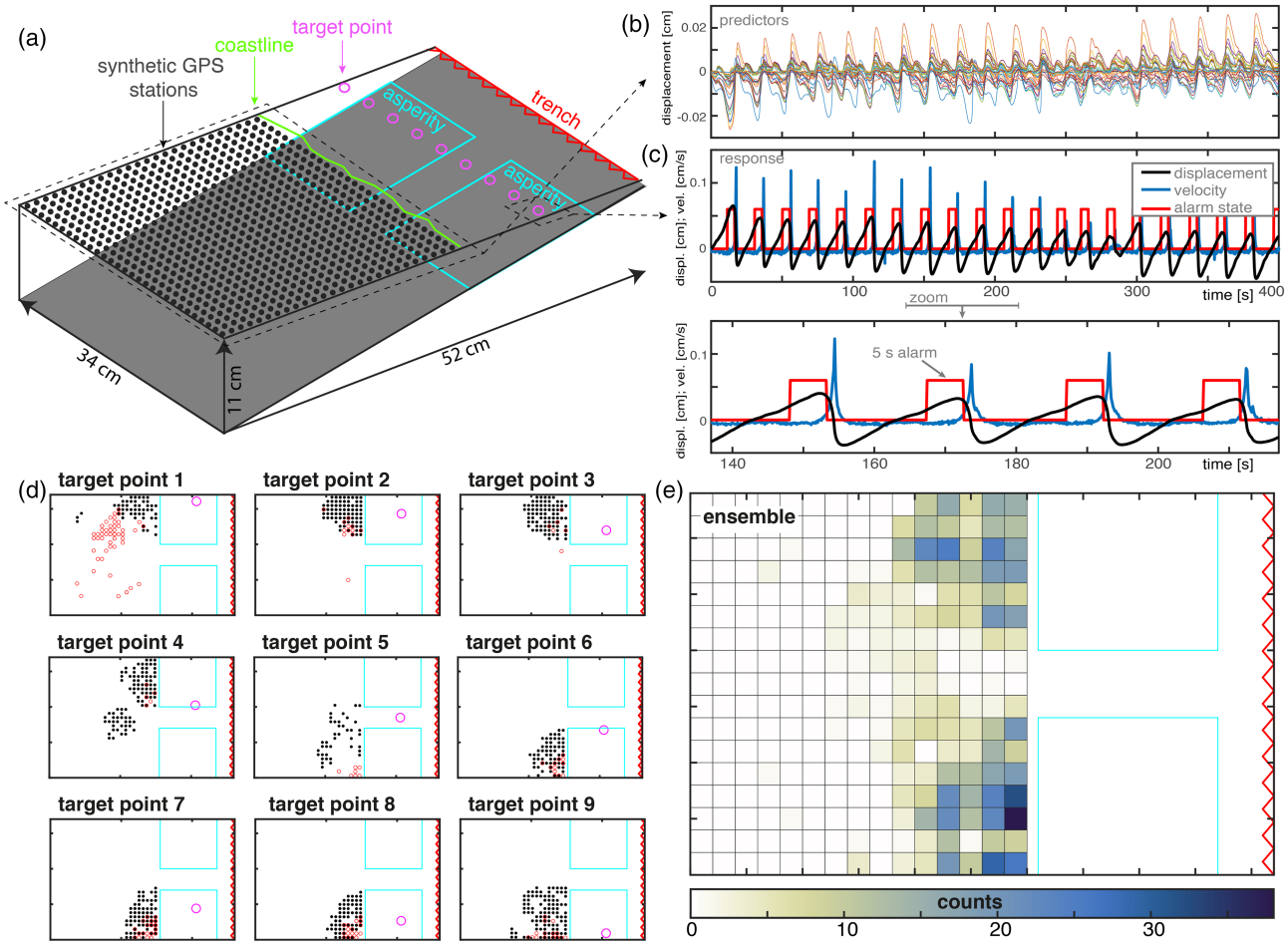

Figure 1. Three-dimensional sketch of the analog model (panel a). The deformation measured at several points on the model surface (black points) is used for building a predictor data set (colored lines in panel b). Trench orthogonal displacement, trench orthogonal component of velocity, and alarm state for one target point (panel c). Synthetic GNSS stations selected by the sequential feature selection algorithm (panel d). A model with 5 s alarm duration, 200 s training, and 72 stations per dm^2 is shown. Black points and red circles highlight whether the trench normal or trench parallel components of the velocity field are selected, respectively. Magenta circles highlight the position of the selected target points. Two-dimensional histogram with color coded number of stations counts (panel e).

between virtual GNSS stations (i.e., ~7 km when scaled to nature) using Particle Image Velocimetry PIV (Sveen, 2004). PIV data can be considered equivalent to a spatially dense, continuous GNSS network.

Data are organized in a matrix of predictors X (3,000 rows by 1,508 columns) where each column corresponds to "displacement measured at a synthetic GNSS station" and rows correspond to time steps (increments). We use exclusively the trench orthogonal and trench parallel components of the displacement field as input features because they have been shown to be the most informative (among surface deformation descriptors) for this model (Corbi et al., 2019a). For our target variables, we select nine target points distributed along the margin (Figure 1a) for which we identify the slipping time of analog earthquakes (i.e., experimental time at which displacement rate exceeds 0.01 cm/s) and the slip onset t_{so}. Then, for each event, we assign the label "alarm" at those time steps that are comprised between t_{so}-Δt and t_{so} (where Δt indicates alarm duration) and "no alarm" at the remaining ones (Figure 1c). This procedure is applied at each target point, so that the output is made of nine response vectors Y. Data are split into training and testing sets. We use the supervised learning Random Undersampling RUSBoost ensemble algorithm running under Matlab (Seiffert et al., 2008), which, in the training phase, "learns" the relationship between the displacement field and the alarm state of each target point. The algorithm is then fed with testing data and predicts if, at a specified time, a portion of the margin is in alarm given the current displacement field (Figure S1).

RUSBoost selects (subsamples) a random fraction of the most represented class (no alarm) in order to have a balanced data set. This allows RUSBoost to be particularly effective in classifying an imbalanced data set as

in our case, where alarms represent 3–27% of the observations, depending on alarm duration and selected target point. After this initial step, RUSBoost proceeds, as in the Adaptive Boosting approach (e.g., Freund & Schapire, 1997), with sequentially building an ensemble of binary decision trees where nodes are displacements measured at various points on the model surface. For each tree, the algorithm computes the weighted classification error and then increases weights for observations misclassified at a given step and reduces weights for observations correctly classified, so that the following tree is trained with updated weights. This procedure is repeated to progressively improve the classification performances (Supporting Information S2).

Binary classification algorithms produce "positive" or "negative" outcomes depending on the identified system state (i.e., alarm and no alarm). An event may occur or not (we truly have a slip event or not). Based on the correctness of the predicted outcome, four cases are possible, as summarized by the confusion matrix (Figure S2): true positive TP, true negative TN, false positive FP, and false negative FN. Precision and recall are two basic evaluation measures for binary classifiers. Precision is defined as TP/(TP + FP) and tells us how many times the raised alarm is correct over the total times we raised an alarm. Recall is defined as TP/(TP + FN) and represents the number of analog earthquakes that have been forecasted by alarms over the total number of earthquakes (both correctly forecasted and missed). The receiver operating curve ROC and the precision-recall curve PR are two other useful indicators of model performance. The ROC is a representation of recall against the false positive rate at various cutoff values used by the algorithm to separate between alarm and no alarm and informs on how well the classifier separates the two classes. The PR shows the trade-off between precision and recall and provides information about how effective a classifier is without raising too many false alarms. The areas under ROC and PR curves (AUC-ROC and AUC-PR) provide single measures of the classification performances ranging from 0 to 1, with 1 representing a perfect classifier. Here we report all above-mentioned metrics and highlight that, in our case, AUC-PR is more informative than the ROC, being independent from the larger fraction of no alarms of our time series (Saito & Rehmsmeier, 2015).

3. Results

3.1. Sequential Features Selection to Identify the Most Informative Region of the Margin

In ML, feature selection (FS) is an important step that precedes model training, especially when a data set has a number of features larger than observations. In our case, where the features to observations ratio can be >1, FS discards less useful and redundant features to improve the model performance and avoid overfitting. FS, by reducing the number of features to interpret, also increases the interpretability of the ML results. To perform FS, we used the Matlab algorithm *sequentialfs* on training data. *sequentialfs* considers the interaction of features and selects a subset of them by sequentially adding one feature while keeping track of the misclassification rate and features index. The minimum of the misclassification error is used as stopping criterion. This way, *sequentialfs* identifies the most relevant features for classification. Since the features that build our X are synthetic GNSS time series of known coordinates, *sequentialfs* highlights the most diagnostic region of the margin. This approach implies that additional stations located outside the region highlighted by *sequentialfs* would be surplus to requirements for predictions.

Figure 1d shows the location of the synthetic GNSS stations selected by the algorithm for each of the nine target points. Except for Target Point 1 where trench normal and trench parallel components are of equal importance (probably due to boundary effects), the trench normal component appears to be more informative than the trench parallel one. The most informative synthetic GNSS stations generally face the offshore asperity where the selected target point is located. We merged the nine groups of selected stations into a 2-D histogram counting the number of times a synthetic station is selected in a given portion of the margin. The histogram highlights a 10–12 cm-wide (or equivalently 70–85 km when scaled to nature) region parallel and adjacent to the analog coastline as the most diagnostic area to monitor for successful predictions, especially regions adjacent to the two asperities (Figure 1e).

3.2. The Role of Training Window Length, Density of Stations, and Alarm Duration

After FS, the ML algorithm is fed only with the most informative features. Our procedure then consists in building and updating a predictive model three times using a shifting training window (Figure S3). The number of updates has a minor influence on predictions (Figure S4). This procedure is repeated for each

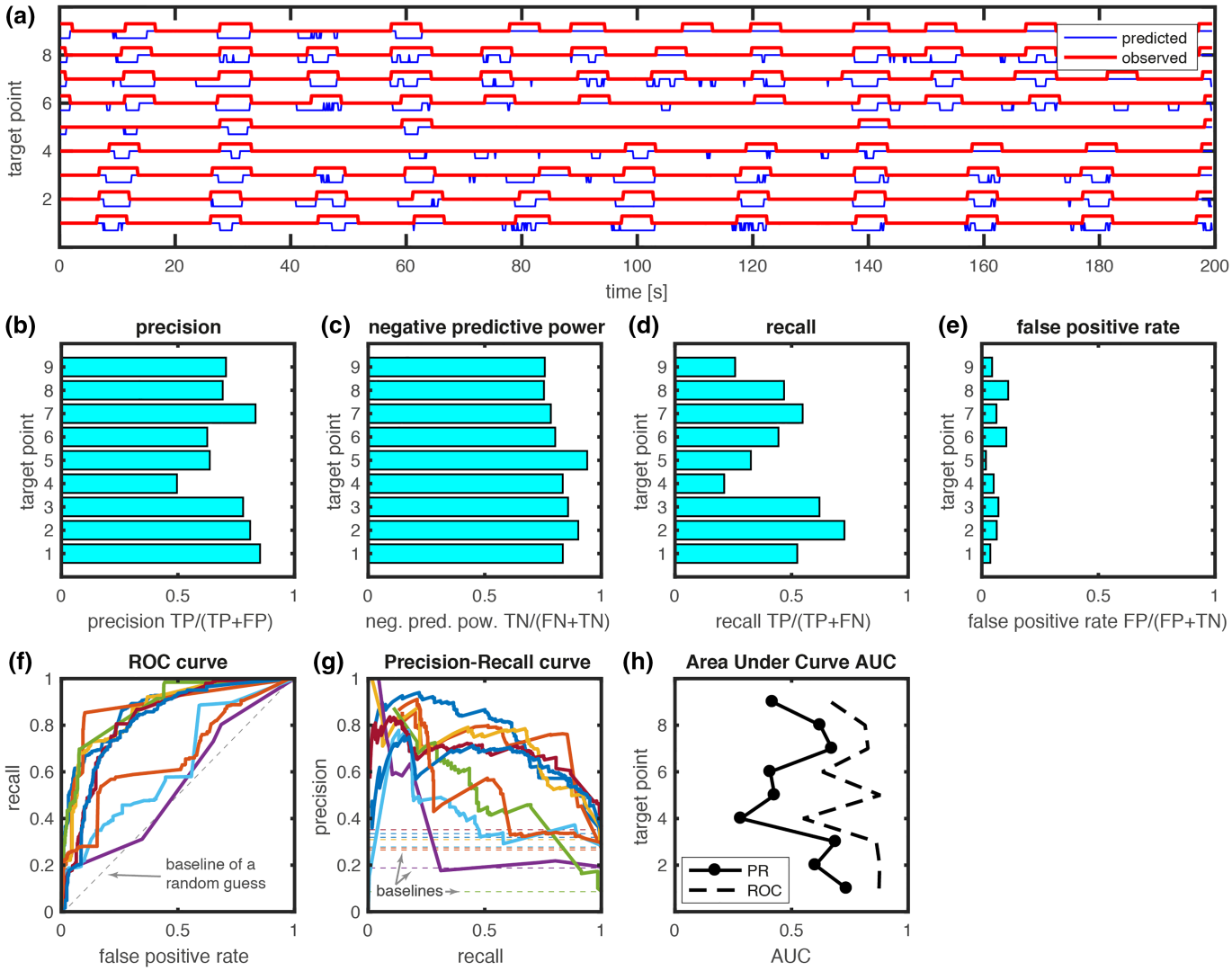

Figure 2. ML results. Comparison between observed and predicted alarm state as a function of time and space (i.e., for various target points; panel a). The model with best prediction averaged over the nine target points is shown (i.e., 5 s alarm duration, 200 s training, and six stations per dm^2). A red line moving up in panel a indicates an imminent analog earthquake occurring at that target point. Blue lines moving down indicate predicted alarms. The lack of blue lines indicate the correct classification of no-alarm periods. Observed alarms longer than 5 s indicate two events occurring within the alarm duration window. A variety of metrics describe the performances of the prediction (panels b–e). ROC and PR curves for each target point, respectively (panels f and g). Areas under PR and ROC curves (panel h).

target point; the prediction model is thus an aggregate composed of nine time series. We compare the predictions with observations as a function of space (i.e., at various target points) and time (Figure 2a). A red line moving up highlights an observed alarm. If the corresponding blue line moves downward simultaneously, it indicates the classifier correctly interpreted the current deformation field as leading to failure in the near future. Different scenarios appear. Looking at the first 50 s of Target Point 2, for example, we observe predicted alarms appearing with a delay with respect to observations, almost perfectly on time, or when there is no alarm at all. This variation in model performance can be quantified as follows: ML predicted correctly the 80% of alarms (i.e., precision), and 90% of no-alarm predictions were reliable (i.e., negative predictive power; Figures 2b– 2h). Predictions for target points located above the barrier (i.e., Target Points 4–6) display relatively worse performances, with precision decreasing to ≈50% due to the smaller number of instances available for the training if compared to target points located above the asperities.

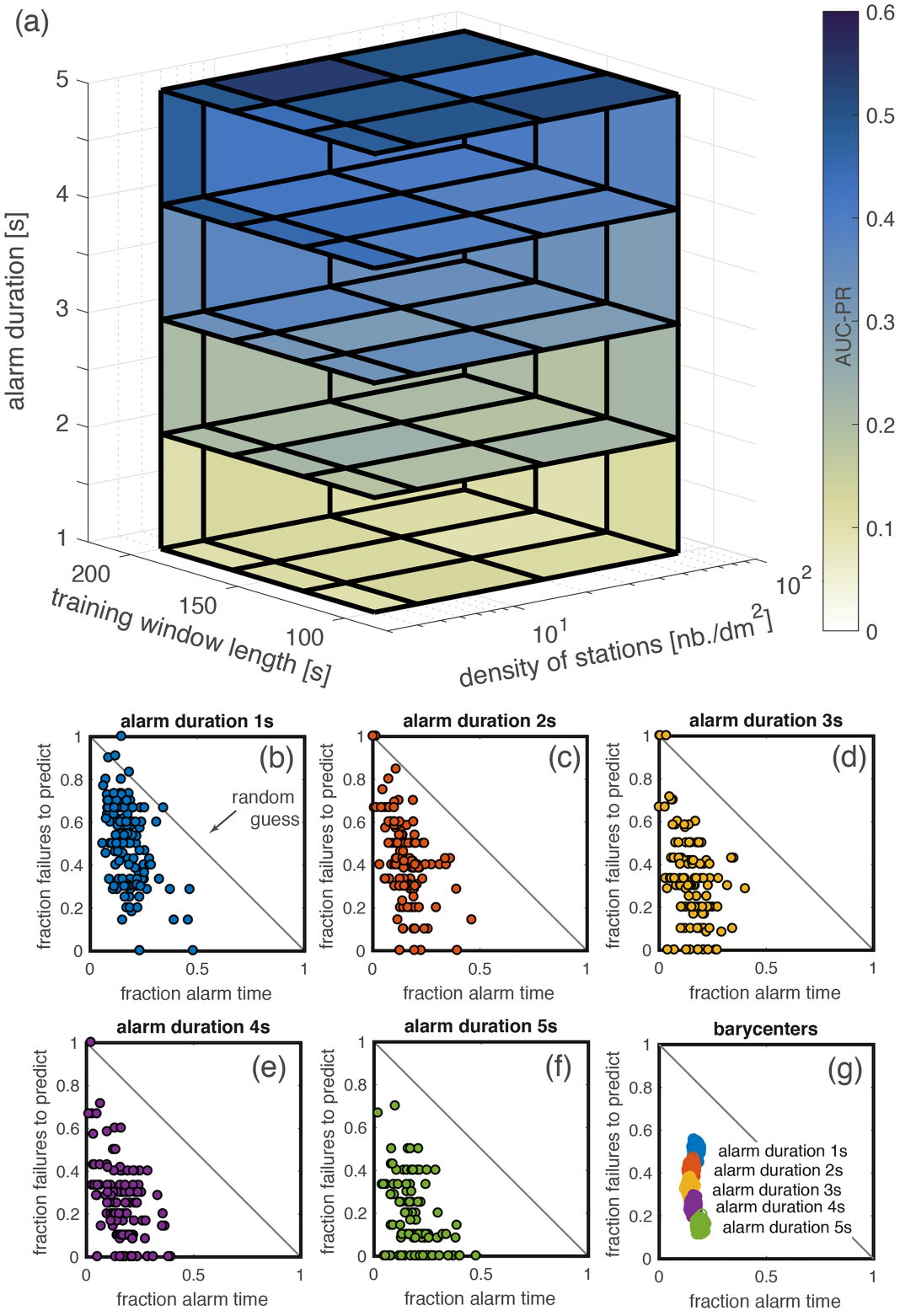

Figure 3. Results of the parametric study. AUC-PR for various training window lengths, density of stations, and alarm durations illustrated with slices for volumetric data (panel a). The corresponding data (144 time series per alarm duration obtained as follows: nine target points times four station densities times four training window lengths) are shown as error diagrams (panel b–f). Centroids of bootstrapped points in panels b–f are represented in panel g.

We repeated the above procedure 80 times, exploring the role of training window length (i.e., record length), density of stations, and alarm duration, following a 3-D grid search approach (Figure 3a). In particular, we varied the training window length in the 100–200 s range, or equivalently for 2–10 analog seismic cycles depending on the training window length and target point; density of stations from 4 to 72 stations per square decimeter, or equivalently 6 to 107 stations per 100 km^2 (see Supporting Information S1 for information about scaling); and alarm duration from 1 to 5 s, or equivalently 0.05 to 0.25 the average seismic cycle duration (a rough estimation of the ratio between slip and stick phases in nature is in the order of 10^{-8} for large subduction earthquakes and 10^{-1} for slow earthquakes; see section 4.2). For each aggregate we report AUC-PR averaged over the nine target points as a single parameter describing the prediction performance of the aggregate. Independently from alarm duration, we generally observe that models with longer training have higher AUC-PR than models with a short training. Such improvement appears more evident for models with higher alarm durations (Figure S5). The effect of station density in the explored range shows AUC-PR oscillating with unclear trends. In general, the effect of station density and training window length on prediction performances for a given alarm duration is smaller than 10%. The highest average AUC-PR value has been recorded for the longest record (i.e., 200 s) and a relatively sparse (i.e., six stations/dm^2) network. A clear increase of average AUC-PR emerges upon comparing models with increasing alarm durations (Figure 3a).

4. Discussion

4.1. Implications for Monitoring Subduction Zones

We have used GNSS-like data from an analog model reproducing multiple subduction megathrust seismic cycles to feed a ML algorithm that predicts the imminence of a slip event. We have identified the most important features to feed into this algorithm and have tested influence of training data size, imminence duration, and measurement spatial density on the performance of the binary classifier.

From feature selection, we found that the region bringing the most important information is located along a 10–12 cm-wide (or equivalently 70–85 km when scaled to nature) swath parallel and adjacent to the coastline, where the highest displacements are measured. This finding supports a scenario where signs of imminent failure come from the downdip edges of asperities, being the closest to the coastline, in agreement with regions of highest shear stresses found in mechanical modeling studies (Bürgmann et al., 2005; Moreno et al., 2018). The trench-normal component of displacement appears more informative than the trench parallel one, likely due to its larger signal caused by trench perpendicular convergence. Unfortunately, we had no access to vertical deformation, and it would be interesting to test whether these data could further improve the predictions. It would also be beneficial to test whether the width of the informative swath scales with the width of the seismogenic zone. The recent development of seafloor geodetic observations in some margins affords us better updip resolution of interplate coupling (e.g., Yokota et al., 2016) and coseismic slip (e.g, Romano et al., 2012). Because in our approach ML primarily tracks the loading and unloading history of the forearc (Corbi et al., 2019a), we tested to what extent offshore kinematic observation would improve the prediction. To do so, we first checked which region would be highlighted by *sequentialfs* assuming the availability of a dense network of stations both onshore and offshore. In this case, the vast majority of informative stations would be offshore, centered above the asperities along strike and with a preferential elongation of the network along the dip (Figure S6). Since implementing such a dense offshore network would be too expensive, we thus tested the effect of the availability of only two stations centered above the asperities or an array of nine offshore stations aligned along the margin, in addition to onshore stations. Taking an onshore network of roughly one station every 50 km as a reference, we observed an improvement (between 3% and 73% depending on alarm duration and network configuration) of the predictions performances for both tested configurations (Figure S7). In particular, we found that for the short alarm durations, the nine stations configuration provides the best result, while upon gradually increasing the alarm duration, the two stations configuration has the best performances. These findings strengthen the idea that having access to the outer wedge deformation would be advantageous for future discoveries while also showing that the optimal configuration to use depends on the investigated target.

We have also shown that, under the studied configuration and with the adopted technique, a short duration (roughly 0.05–0.10 the duration of analog seismic cycles) prediction is unfeasible—even if a dense network

with a record of up to 10 cycles is available. This is due to our framing: The large number of stations creates a wide X with many features (i.e., columns) and few observations (i.e., rows) that are controlled by number of events and alarm duration (at least in the investigated range of training window length and earthquake recurrence time). Therefore, ML has too few data for classification so that a given deformation can be interpreted either as an alarm or a no alarm. On the contrary, if we ask ML for a less focused prediction (e.g., 0.15–0.25 times the duration of analog seismic cycles), the training has a larger number of instances, because at longer alarm duration corresponds a longer X and the algorithm can better classify the deformation field. Increasing the alarm window also increases the chance of getting a slip event in the given window. Therefore, an improvement of prediction performances is expected. We show that the improvement we get is more significant than what would be expected by chance using error diagrams (Figures 3b– 3f). Error diagrams plot the fraction of alarm time versus the fraction of failures to predict (i.e., 1- fraction of events correctly predicted), and they are considered a useful earthquake predictability measurement (Kagan, 2007). In these diagrams an optimal prediction would correspond to a point near the diagram origin at (0, 0), while a random forecast would fall along the diagonal connecting (0, 1) and (1, 0). We labeled a predicted event if at least one declared alarm falls within the observed alarm window. Each point on the graph represents the prediction at a given target point for different training window lengths and density of stations. Comparing various models with increasing alarm durations, we observe that the cloud of points moves progressively toward smaller values of fraction of failure to predict with only a minor shift toward large fractions of alarm time, as highlighted by the downward shift of centroids (i.e., the arithmetic mean position of all the points in the figure) of the points in error diagrams. This indicates that models with longer alarm durations are more precise while requiring almost the same number of declared alarms as models with short alarm durations (Figure S8).

Also apparent from our results is that the availability of data with high spatial resolution is less important than having access to long time series in which the investigated phenomena repeat several times (e.g., 10 times). The observation that a better classification is achieved for models with longer alarm durations raises the additional argument of the impact of sampling rate, because sampling interval together with alarm duration contribute to the number of data with alarm labels. To test the role of sampling interval, we run seven additional models with fixed parameters (i.e., alarm duration of 5 s, training window length of 200 s, and density of stations equal to $6/dm^2$) and varying sampling intervals in the 0.13–3.25 s range. We found that, with exception of the smallest interval, classifications become progressively more reliable reducing the acquisition interval (Figure S9). This suggests that possibly, a better classification may be achieved also for short duration alarms by monitoring the experiment with higher frequency. This observation supports a scenario where GNSS, given the higher acquisition frequency, is more useful than InSAR when attempting to export the proof of concept tested in this study to natural subduction zones.

4.2. Unlocking the Possibility to Predict the Onset of Slow Earthquakes

Our analysis showed that, in reconstructing the spatiotemporally complex forearc loading history, ML can predict the timing and size (tracking simultaneous alarms at various target points) of analog megathrust earthquakes under given circumstances of relatively long alarms and a relatively long observation record. Laboratory models represent the ideal candidates for attempting this type of prediction given their ability to produce the necessary observational data and to repeat a given process several times (e.g., Corbi et al., 2019a; Rouet-Leduc, 2018). Here we have investigated the impact of data space-time distributions on the ability of a ML-based technique to predict the onset of analog earthquakes using geodetic-like observations. We found that, having access to the deformation history of about 10 cycles at a limited number of stations, ML can reach a precision generally larger than 0.7 (0.5 minimum value over the nine target points; Figure 2b) and with very few false alarms (false positive rates <0.1; Figure 2e). In nature we do not have access to geodetic time series including multiple large subduction earthquakes to test this method. However, so-called slow events, with slip durations of few tens of days (Obara, 2002), are potential candidates to test our approach on. Cascadia is one of those subduction zones where roughly once per year (depending on the latitude) the megathrust hosts slow slip episodes and where the geodetic record is more than 10 years long (e.g., Michel et al., 2019). Given the similarity in space-time recurrence behavior (with partial ruptures alternating with larger events) and deformation pattern (with alternating trenchward and landward surface velocities), we suggest that ML could be applied to predict the onset and the extent of slow slip in this area.

5. Conclusions

We investigated the role of space-time coverage and alarm duration on the performance of analog earthquake prediction. We found that alarm duration plays a primary role in tuning the performances of a binary classifier. A sharp, accurate analog earthquake prediction is unfeasible with the algorithm used in this study, even in a simplified system with a perfectly designed monitoring network. But alarm periods become in reasonably good agreement with observed earthquakes when tens of seismic cycles have been recorded and when the alarm duration is longer. These results can be further improved by tuning the network design and acquisition rates. Given these findings, we propose that it might be possible to predict the onset and extent of slow earthquakes in real subduction zones by applying similar ML approaches to those developed in this study on GNSS time series.

5.1. Data and Material Availability

All data and materials used in the analysis are available through GFZ Data Services and published open access in Corbi et al. (2019b).

Acknowledgments
We thank two anonymous reviewers and the editor G. Hayes for their constructive comments. In Figures 1 and 3 we used the perceptually uniform colormap Davos by F. Crameri. The work leading to this publication was supported by the PRIME program of the German Academic Exchange Service (DAAD) with funds from the German Federal Ministry of Education and Research (BMBF). M. R. has benefitted from the inspiring environment of the CRC 1114 "Scaling Cascades in Complex Systems" (funded by Deutsche Forschungsgemeinschaft [DFG]). The Grant to Department of Science, Roma Tre University (MIUR-Italy Dipartimenti di Eccellenza, ARTICOLO 1, COMMI 314 – 337 LEGGE 232/2016) is gratefully acknowledged.

References

Bergen, K. J., Johnson, P. A., De Hoop, M. V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, *363*(6433). https://doi.org/10.1126/science.aau0323

Bürgmann, R., Kogan, M. G., Steblov, G. M., Hilley, G., Levin, V. E., & Apel, E. (2005). Interseismic coupling and asperity distribution along the Kamchatka subduction zone. *Journal of Geophysical Research*, *110*, B07405. https://doi.org/10.1029/2005JB003648

Corbi, F., Funiciello, F., Brizzi, S., Lallemand, S., & Rosenau, M. (2017). Control of asperities size and spacing on seismic behavior of subduction megathrusts. *Geophysical Research Letters*, *44*, 8227–8235. https://doi.org/10.1002/2017GL074182

Corbi, F., Funiciello, F., Moroni, M., Van Dinther, Y., Mai, P. M., Dalguer, L. A., & Faccenna, C. (2013). The seismic cycle at subduction thrusts: 1. Insights from laboratory models. *Journal of Geophysical Research: Solid Earth*, *118*, 1483–1501. https://doi.org/10.1029/2012JB009481

Corbi, F., Sandri, L., Bedford, J., Funiciello, F., Brizzi, S., Rosenau, M., & Lallemand, S. (2019a). Machine learning can predict the timing and size of analog earthquakes. *Geophysical Research Letters*, *46*, 1303–1311. https://doi.org/10.1029/2018GL081251

Corbi, F., Sandri, L., Bedford, J., Funiciello, F., Brizzi, S., Rosenau, M., & Lallemand, S. (2019b) Supplementary material to "machine learning can predict the timing and size of analog earthquakes". GFZ Data Services. https://doi.org/10.5880/fidgeo.2018.071.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

Heki, K., & Mitsui, Y. (2013). Accelerated pacific plate subduction following interplate thrust earthquakes at the Japan trench. *Earth and Planetary Science Letters*, *363*, 44–49. https://doi.org/10.1016/j.epsl.2012.12.031

Hubbert, M. K. (1937). Theory of scale models as applied to the study of geological structures. *Geological Society of America Bulletin*, *48*, 1459–1520.

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. https://doi.org/10.1038/s41561-018-0272-8

Hyndman, R. D., Yamano, M., & Oleskevich, D. A. (1997). The seismogenic zone of subduction thrust faults. *Island Arc*, *6*(3), 244–260. https://doi.org/10.1111/j.1440-1738.1997.tb00175.x

Kagan, Y. Y. (2007). On earthquake predictability measurement: Information score and error diagram. *Pure and Applied Geophysics*, *164*(10), 1947–1962. https://doi.org/10.1007/s00024-007-0260-1

Kato, A., Obara, K., Igarashi, T., Tsuruoka, H., Nakagawa, S., & Hirata, N. (2012). Propagation of slow slip leading up to the 2011 Mw 9.0 Tohoku-Oki earthquake. *Science*, *335*(6069), 705–708. https://doi.org/10.1126/science.1215141

Loveless, J. P., & Meade, B. J. (2016). Two decades of spatiotemporal variations in subduction zone coupling offshore Japan. *Earth and Planetary Science Letters*, *436*, 19–30. https://doi.org/10.1016/j.epsl.2015.12.033

Mavrommatis, A. P., Segall, P., & Johnson, K. M. (2014). A decadal-scale deformation transient prior to the 2011 M w 9.0 Tohoku-oki earthquake. *Geophysical Research Letters*, *41*, 4486–4494. https://doi.org/10.1002/2014GL060139

Melnick, D., Moreno, M., Quinteros, J., Baez, J. C., Deng, Z., Li, S., & Oncken, O. (2017). The super-interseismic phase of the megathrust earthquake cycle in Chile. *Geophysical Research Letters*, *44*, 784–791. https://doi.org/10.1002/2016GL071845

Michel, S., Gualandi, A., & Avouac, J.-P. (2019). Interseismic coupling and slow slip events on the Cascadia Megathrust. *Pure and Applied Geophysics*, *176*(9), 3867–3891. https://doi.org/10.1007/s00024-018-1991-x

Moreno, M., Haberland, C., Oncken, O., Rietbrock, A., Angiboust, S., & Heidbach, O. (2014). Locking of the Chile subduction zone controlled by fluid pressure before the 2010 earthquake. *Nature Geoscience*, *7*(4), 292–296. https://doi.org/10.1038/ngeo2102

Moreno, M., Li, S., Melnick, D., Bedford, J. R., Baez, J. C., Motagh, M., et al. (2018). Chilean megathrust earthquake recurrence linked to frictional contrast at depth. *Nature Geoscience*, *11*(4), 285–290. https://doi.org/10.1038/s41561-018-0089-5

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

Romano, F., Piatanesi, A., Lorito, S., D'Agostino, N., Hirata, K., Atzori, S., et al. (2012). Clues from joint inversion of tsunami and geodetic data of the 2011 Tohoku-oki earthquake. *Scientific Reports*, *2*(1), 1–8. https://doi.org/10.1038/srep00385

Rosenau, M., Corbi, F., & Dominguez, S. (2017). Analogue earthquakes and seismic cycles: Experimental modelling across timescales. *Solid Earth*, *8*(3), 597–635. https://doi.org/10.5194/se-8-597-2017

Rosenau, M., Horenko, I., Corbi, F., Rudolf, M., Kornhuber, R., & Oncken, O. (2019). Synchronization of great subduction megathrust earthquakes: Insights from scale model analysis. *Journal of Geophysical Research: Solid Earth*, *124*, 3646–3661. https://doi.org/10.1029/2018JB016597

Rosenau, M., Nerlich, R., Brune, S., & Oncken, O. (2010). Experimental insights into the scaling and variability of local tsunamis triggered by giant subduction megathrust earthquakes. *Journal of Geophysical Research*, *115*, B09314. https://doi.org/10.1029/2009JB007100

Rouet-Leduc, B., Hulbert, C., Bolton, D. C., Ren, C. X., Riviere, J., Marone, C., et al. (2018). Estimating fault friction from seismic signals in the laboratory. *Geophysical Research Letters*, *45*, 1321–1329. https://doi.org/10.1002/2017GL076708

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. https://doi.org/10.1038/s41561-018-0274-6

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*, 9276–9282. https://doi.org/10.1002/2017GL074677

Ruff, J. R., & Tichelaar, B. W. (1996). What controls the seismogenic plate interface in subduction zones? In G. E. Bebout, et al. (Eds.), *Subduction: Top to bottom*, *Geophysical Monograph Series* (Vol. 96, pp. 105–111). Washington, DC: AGU.

Saillard, M., Audin, L., Rousset, B., Avouac, J. P., Chlieh, M., Hall, S. R., et al. (2017). From the seismic cycle to long-term deformation: Linking seismic coupling and quaternary coastal geomorphology along the Andean megathrust. *Tectonics*, *36*, 241–256. https://doi.org/10.1002/2016TC004156

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, *10*(3), 1, e0118432–21. https://doi.org/10.1371/journal.pone.0118432

Schurr, B., Asch, G., Hainzl, S., Bedford, J., Hoechner, A., Palo, M., et al. (2014). Gradual unlocking of plate boundary controlled initiation of the 2014 Iquique earthquake. *Nature*, *512*(7514), 299–302. https://doi.org/10.1038/nature13681

Seiffert, C., Khoshgoftaar, T., Hulse, J., & Napolitano, A. (2008). RUSBoost: Improving clasification performance when training data is skewed. 19th International Conference on Pattern Recognition (pp. 1–4).

Sveen, J. K. (2004). An introduction to MatPIV v.1.6.1, Eprint no. 2. ISSN 0809–4403, Department of Mathematics, University of Oslo. Retrieved from http://urn.nb.no/URN:NBN:no-27806

Wang, K., & Bilek, S. L. (2011). Do subducting seamounts generate or stop large earthquakes? *Geology*, *39*(9), 819–822. https://doi.org/10.1130/G31856.1

Williams, R. T., Davis, J. R., & Goodwin, L. B. (2019). Do large earthquakes occur at regular intervals through time? A perspective from the geologic record. *Geophysical Research Letters*, *46*, 8074–8081. https://doi.org/10.1029/2019GL083291

Yokota, Y., Ishikawa, T., Watanabe, S., Tashiro, T., & Asada, A. (2016). Seafloor geodetic constraints on interplate coupling of the Nankai Trough megathrust zone. *Nature*, *534*(7607), 4–6. https://doi.org/10.1038/nature17632