

Generalized Eruption Forecasting Models using Machine Learning Trained on Seismic Data from 24 Volcanoes

Alberto Ardid

`alberto.ardid@canterbury.ac.nz`

University of Canterbury <https://orcid.org/0000-0001-8040-8193>

David Dempsey

University of Canterbury <https://orcid.org/0000-0003-2135-5129>

Corentin Caudron

Université libre de Bruxelles

Shane Cronin

University Auckland <https://orcid.org/0000-0001-7499-603X>

Ben Kennedy

University of Canterbury <https://orcid.org/0000-0001-7235-6493>

Társilo Girona

Geophysical Institute, University of Alaska Fairbanks

Sally Potter

Te Pū Ao | GNS Science

Oliver Lamb

Te Pū Ao | GNS Science

Craig Miller

Te Pū Ao | GNS Science

Leoncio Cabrera

Université Grenoble Alpes, CNRS <https://orcid.org/0000-0001-9401-2608>

Sergio Ruiz

Universidad de Chile

Javier Pacheco

National University of Costa Rica

Mauricio Mora

University of Costa Rica <https://orcid.org/0000-0001-7571-6749>

Anto Maranto

Center for Volcanology and Geological Hazard Mitigation

Yesim Cubuk Sabuncu

Icelandic Meteorological Office

Silvio De Angelis

University of Liverpool

Diana Roman

Carnegie Institution for Science <https://orcid.org/0000-0003-1282-5803>

Physical Sciences - Article

Keywords: Volcano eruption forecasting, transfer machine learning, generalized and tailored models

Posted Date: December 5th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3483573/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on February 25th, 2025. See the published version at <https://doi.org/10.1038/s41467-025-56689-x>.

Generalized Eruption Forecasting Models using Machine Learning Trained on Seismic Data from 24 Volcanoes

Alberto Ardid¹, David Dempsey¹, Corentin Caudron², Shane Cronin³, Ben Kennedy¹, Tàrsilo Girona^{4,5}, Sally Potter⁶, Oliver D. Lamb⁶, Craig Miller⁶, Leoncio Cabrera⁷, Sergio Ruiz⁷, Javier Pacheco⁸, Mauricio M. Mora⁹, Anto Martanto¹⁰, Yesim Cubuk-Sabuncu¹¹, Silvio De Angelis¹², Diana Roman¹³,

¹University of Canterbury, New Zealand

²Université libre de Bruxelles, Belgium

³University of Auckland, New Zealand

⁴Geophysical Institute, University of Alaska Fairbanks, Fairbanks, Alaska 99775, USA

⁵Alaska Volcano Observatory (AVO), Alaska, USA

⁶Te Pū Ao | GNS Science, New Zealand

⁷Departamento de Geofísica, Universidad de Chile, Chile

⁸National University of Costa Rica, Costa Rica

⁹Central American School of Geology, University of Costa Rica, Costa Rica

¹⁰Center for Volcanology and Geological Hazard Mitigation, Indonesia

¹¹Icelandic Met Office, Iceland

¹²University of Liverpool, UK

¹³Carnegie Institution of Washington DC, USA

Email corresponding author: aardids@gmail.com

Keywords: Volcano eruption forecasting, transfer machine learning, generalized and tailored models.

ABSTRACT

Seismic data collected before volcanic eruptions holds important clues for forecasting future eruptions. However, many volcanoes have short monitoring histories or have infrequent eruptions. Here, we aim to enhance volcanic eruption forecasting through machine learning, tackling challenges in generalizing eruption precursors across diverse volcanoes, including those with limited seismic data. Additionally, we investigate similarities and differences in pre-eruptive seismic activity. We combine records from 24 volcanoes to train machine learning (ML) models that recognize eruption precursors and tested the accuracy by forecasting on volcanoes whose data had been withheld during training. These forecasts had nearly the same accuracy as other models that were given the advantage of training on past eruptions at the target volcano. We find that numerous eruption precursors are applicable across various volcanoes, challenging the notion that volcanoes exhibit individualistic behavior. The accuracy of the ML models is about 50 to 70% higher than traditional methods based on averaged seismic amplitude. Our ML models should be a tool for volcano observatories to provide timely probabilistic forecasts suitable to address volcanic risk, especially where prior eruption data is limited. This methodology and workflow can be applied to other hazard types that suffer from regional data-scarcity, e.g., wildfires and floods.

INTRODUCTION

Volcanic eruptions are one of the most spectacular natural phenomena on Earth, but they can be deadly if not properly anticipated^{1,2}. Eruption forecasting is the practice of estimating the likelihood, location, and timing of eruptions^{3,4}; and the type of hazards they can present^{5,6}. Volcano observatories are a key line of defense to ensure the safety and welfare of the ~29 million individuals who reside within 10 km of active volcanoes⁷. Eruptions are also disruptive to air travel and transportation systems, regional air quality, food and drinking water systems, and even influence weather patterns^{8,9}. Several recent eruptions have been successfully forecasted including: 1991 Mount Pinatubo¹⁰ (Philippines, by PHIVOLCS and USGS), 2009 Redoubt¹¹ (Alaska, USA, by AVO), 2010 Merapi¹² (Indonesia, by CVGHM), 2014 Villarrica¹³ (Chile, by OVDAS) and 2014 Bárðarbunga¹⁴ (Iceland, by NCIP) eruptions. These cases exemplify the primacy of scientific monitoring and underscore the importance of historic data for effective eruption forecasting.

Forecasting often relies on the detection of meaningful changes in signals encoded in seismic data (i.e., seismic precursors). Seismic amplitude filtered to informative frequency bands is often used, e.g., real-time seismic amplitude measurement¹⁵ (RSAM) and displacement seismic amplitude ratio^{16,17} (DSAR). Precursor signals are thought to reflect fluid-rock interactions in magma plumbing and hydrothermal systems, induced oscillations in cracks and conduits^{18,19}, or accumulations of gas¹⁶ – all physical processes that may precede an eruption.

Generalized eruption forecasting models that are suitable for many volcanoes remain scarce^{20,21,22}. These models have the advantage of not requiring extensive customization using data from the target volcano^{23,24}. However, a difficulty with their implementation has been accounting for the wide variety of volcanic systems worldwide, which exhibit a range of sometimes unique, precursory behaviors. Unheralded eruptions that have either subtle or no known precursors^{25,2} is

another difficulty for generalized forecasters to address. These unpredictable eruptions are a feature of extremely complex volcanic systems and they expose the limitations of current forecasting approaches.

Machine learning (ML) is an emerging tool for eruption forecasting²⁶. ML approaches have the potential to automate some manual monitoring processes, freeing up volcano observatory personnel to spend additional time on interpretation and decision-making. They can also identify pre-eruption time-series patterns in large datasets²⁷ that may otherwise be unclear or not appropriately emphasized by human operators. ML can also help to mitigate human biases when weighing the statistical evidence of a precursor¹⁷, although it risks introducing its own biases, e.g., during sampling, labeling, data, and algorithms selection^{28,29}. ML can be used to analyze various types of data, including seismic^{30,31}, gas³², and satellite data³³, which supports an increasingly comprehensive characterization of volcanic activity.

Using data from multiple volcanoes for generalized forecasting

Our study identifies sets of precursor signals of eruptions that are shared amongst groups of volcanoes. Doing so supports the development of global forecasts that can be applied at analogue volcanoes with no historic eruption data. To achieve this, we developed a new type of machine learning pipeline that shares data from many volcanoes to identify common seismic patterns that recur preferentially before eruptions at different volcanoes. We tested the forecasting ability of these precursors by training global, multiple volcano forecast models and seeing how they performed compared to models for individual volcanoes. In particular, our machine learning pipeline consists of a random forest classifiers whose individual decision trees were trained on diverse, but balanced, subsets of the statistically significant precursors, which reduces overfitting to specific volcanoes or eruptions. The resulting model, when fed new seismic data, outputs a forecast number between 0 and 1 with higher numbers indicating a higher likelihood of eruption³⁴.

We used a dataset comprising 41 eruptions across 24 volcanoes (Fig 1a) with a combined seismic record length of ~73 years. Based on eruption type, we defined three pools of volcanoes – magmatic (9 volcanoes, 16 eruptions), phreatic (6 volcanoes, 15 eruptions), and global (all volcanoes and eruptions) – and tested the similarity of their pre-eruption signals (Fig 1b; see Methods). We were interested mainly in short-term changes that are useful for responsive monitoring and to inform evacuation orders. Therefore, we focused on signals that occurred 48-hours prior to eruptions using a technique called time-series feature engineering^{27,17}, which applies brute-force pattern matching to the windowed data. The 48-hour window focuses on short-term changes in volcano state that may be relevant to short-term changes in hazard; for example, abrupt changes in fluid pressurization that may preface an eruption or fluid release in subsequent hours to days¹⁷.

For this study, the forecasting skill of generalized models on data-scarce volcanoes is paramount. To measure model performance, we use a cross-validation method that excludes all data from a target “unobserved” volcano during training, and then later checks the accuracy of the model’s forecast on the excluded volcano (referred to as cross-validation volcano, CVV). This pseudo-prospective forecast is not as good as true prospective forecasting, but it is superior to hindcasting models that train on data from the target volcano. An alternative training leaves out only one eruption but otherwise includes data from all volcanoes in a pool, including the target (cross-validation eruption, CVE; see Fig 1c & Methods). Both strategies are important to assess and understand how generalized forecasters would perform at well-monitored volcanoes with eruptive records (CVE) and those with no historic monitoring (CVV). We also compared generalized models to (1) tailored models trained exclusively on the data from the target volcano (for volcanoes with three or more recorded eruptions), and (2) an optimized seismic-amplitude (RSAM) model.

Forecast performance is assessed in terms of the reliability and discriminability of the models on test datasets. Reliability refers to a model’s capacity to anticipate eruptions and, hence minimizing missed eruptions. Discriminability measures the model’s ability to distinguish between eruptive and non-eruptive periods, hence minimizing false positives. Both aspects need to be considered when developing and implementing forecasters across many types of natural hazards.

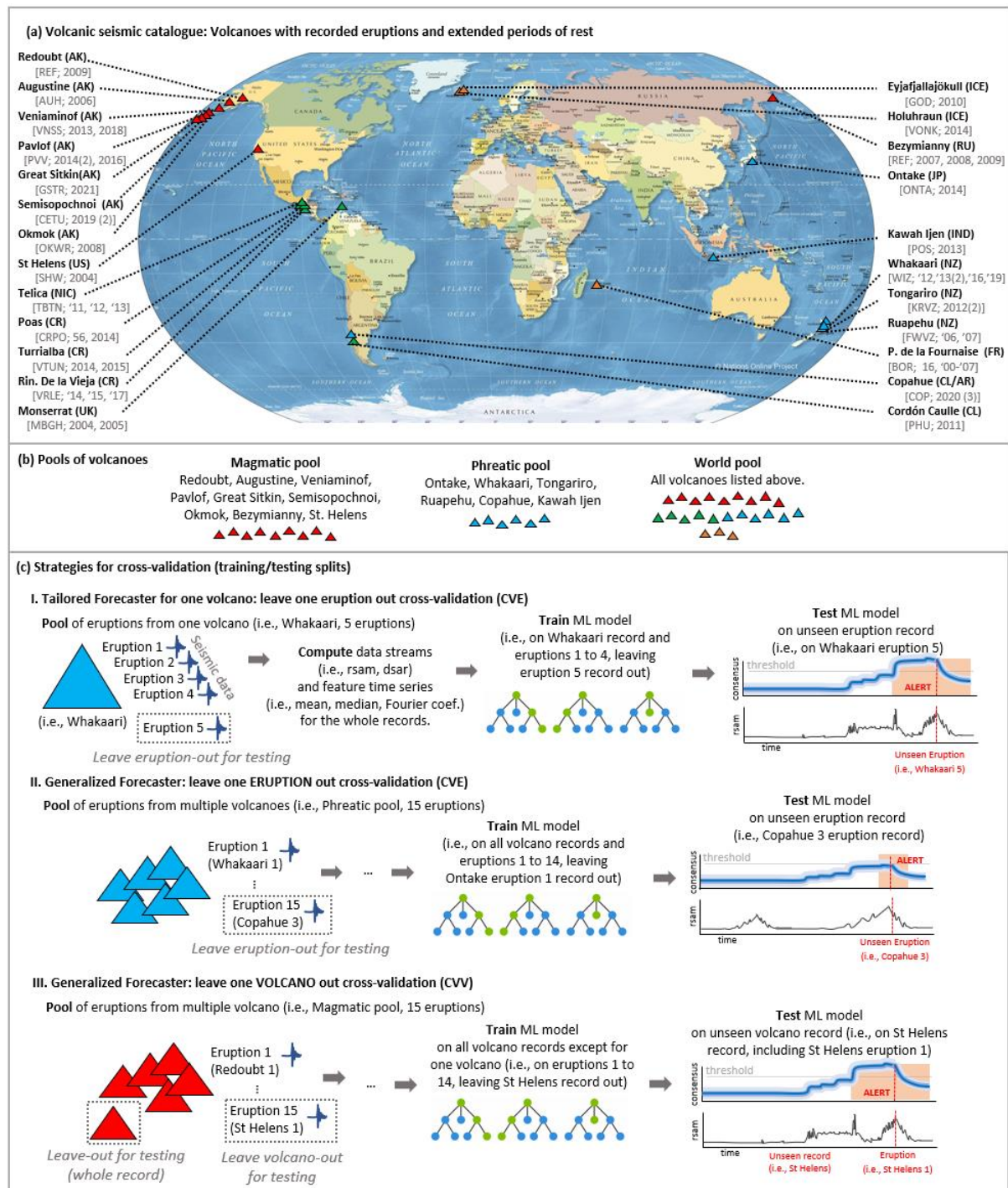


Figure 1. (a) Catalog of volcanoes and eruptions used in this study. (b) Sets of volcanoes grouped to train generalized eruption forecasting models (pools). (c) Cross-validation testing strategy for the forecasting models CVE (leave one eruption out). (I) CVE for tailored forecasters, (II) CVE for generalized forecasters, and (III) CVV for generalized forecasters. For details on cross-validation strategies see Methods section

RESULTS

Reliable forecasting of eruptions at unobserved volcanoes

A reliable forecast should be consistently accurate, without significant errors or biases^{15,35,36}, hence building confidence in forecasting processes and outcomes. Fig 2 shows the output of generalized forecast models on four target volcanoes: Bezymianny, Whakaari, Semisopchnoi, & St Helens. Models are anticipating eruptions well when their output increases

substantially prior to the eruption. In this study, if it exceeds a predetermined threshold, then the eruption is deemed to have been forecast: an arbitrary value of 0.6 has been used in Fig 2.

Using a 0.6 triggering threshold, the accuracy of generalized forecast models at (unseen) Whakaari, Bezymianny, and Copahue is high, anticipating 10 out of 11 eruptions (91%) (Fig S1). Tailored models for the individual volcanoes are also effective, anticipating 9 out of 11 eruptions (82%) (Fig S2). The main difference is that the generalized model identifies precursors to a 2016 eruption at Whakaari (Fig S3) that has previously been difficult to anticipate using tailored forecasting approach^{27,34}. This underscores the advantage of pooling eruptions precursors learned from many volcanoes.

For ten other volcanoes (Fig. 2d-e & Fig S4) that each have fewer than three eruptions, generalized models trained on phreatic and magmatic pools can anticipate 10 out of 14 eruptions at a threshold of 0.6. This highlights the potential of this approach at volcanoes with limited eruption records. The obvious caveat is that, for a sufficiently low threshold, all eruptions can eventually be anticipated. It must therefore be demonstrated that the associated false positive rate is acceptable for a chosen threshold, which is addressed below.

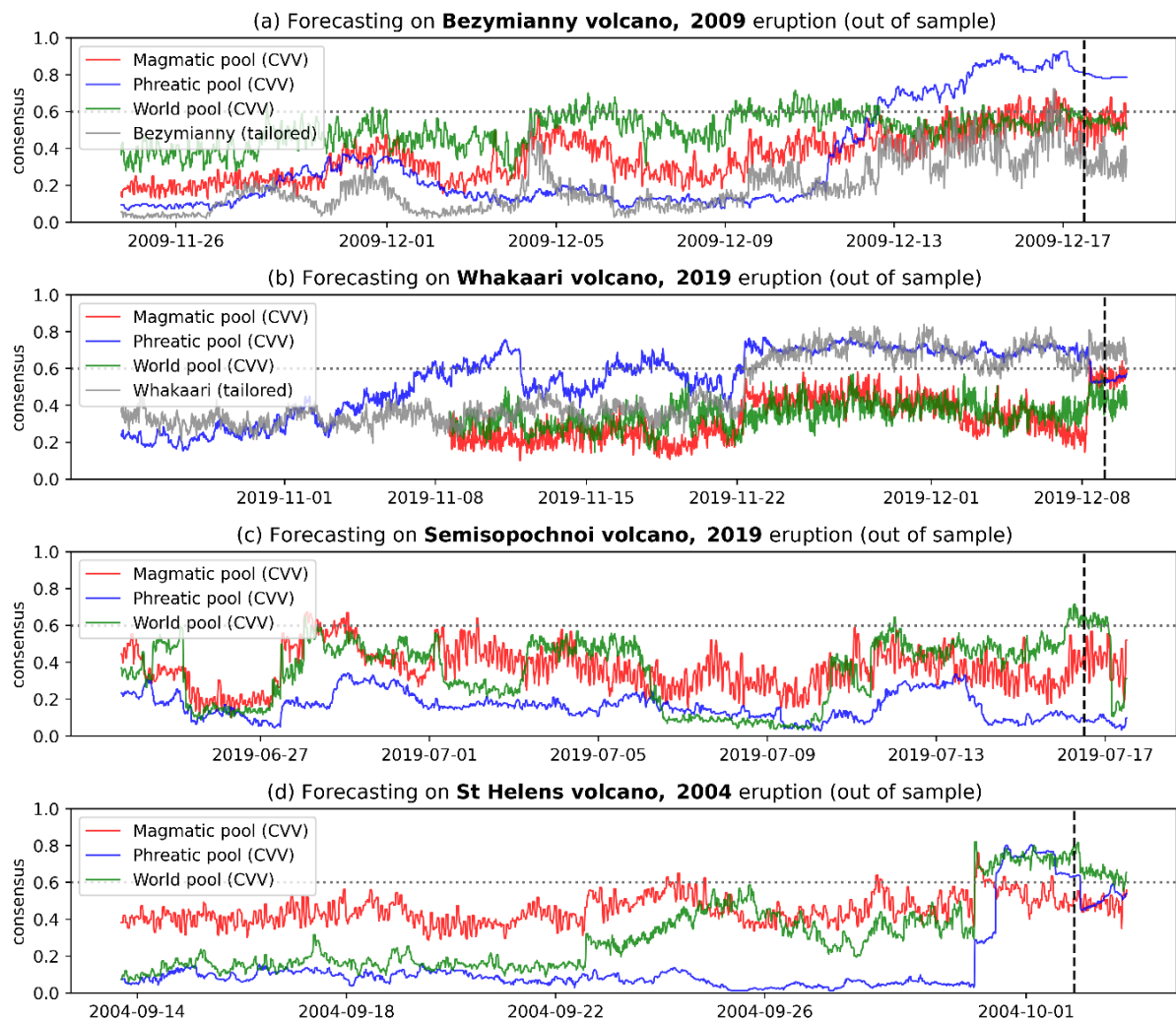


Figure 2. Forecast for different models prior to four eruptions considered in the catalog (pools are indicated in the legend, eruptions are indicated in the titles, and eruption times by the black dash line). The forecasts correspond to the models tested on eruptions that were not included in the training (out-of-sample). A threshold of 0.7 is indicated as a reference (dashed gray line). This value is arbitrary, and the performance metrics of the models (described in the next figures) are calculated for one hundred thresholds in the range [0, 1]. The consensus values, solid lines, correspond to a 2-day rolling 90th percentile for the consensus. This visual element is included in this plot specifically to highlight instances when the consensus exceeds the threshold and to improve the clarity of the plot. Same figure without the 2-day rolling 90th percentile for the consensus is shown in Figure S5.

Discriminability: Minimizing false positives on non-eruptive data

Discriminability measures how well forecasting models distinguish between pre-eruptive and non-eruptive, but potentially unrestful, data, which are common at volcanoes^{37,38}. A low discriminability is likely to lead to forecasts with unacceptably high rates of false positives.

To measure discriminability, using the model output over the whole records (e.g., Fig S6), we computed receiver operating characteristic (ROC) curves, which are a graphical measure of the performance of a classification model across all possible thresholds (Fig S7). As discriminability improves, a ROC curve will depart from the diagonal towards the top left corner, denoting a reduced false positive rate. ROC curves for Bezymianny, Whakaari and Copahue (Fig. S3a-c) show that tailored and generalized forecast models have comparable discriminability indicating that performance is not reduced when data from non-target volcanoes is included. This is further underscored by ROC curves for the three volcano pools (Fig. S3d-f) that show similar performance for CVV and CVE generalized models. All models have similar or better discriminability than seismic-amplitude approaches, particularly for magmatic eruptions.

Generalized, tailored and seismic-amplitude models can be directly compared using the Area Under the (ROC) Curve (AUC), which is a single numerical measure of average model performance. Using AUC to evaluate performance, we find that tailored and generalized models perform 27% and 22% better, respectively, than seismic amplitude approaches (Methods; Table S1). Alternatively, we use Area Over the Curve (AOC) as a measure of the remaining forecast error relative to a perfectly predicting model (AUC=1 or AOC=0). For this measure, we find that generalized forecasts have 50% less error than seismic amplitude models, and tailored models have a further 20% less error than generalized ones.

DISCUSSION

Improving decision support for eruption warning systems

The purpose of eruption forecasts is to provide decision-makers objective, preferably quantified information about the likelihood of an eruption in a specified warning period^{39,5}. Here, we consider that warnings could be triggered when the model output exceeds a threshold and end once it has dropped below the threshold for more than 48-hours. As an example, applying the phreatic pool model to the 11-year dataset at Whakaari, for a threshold of 0.8, results in 80% of eruptions (4 out of 5) occurring inside a warning period and a total warning period of 11% (i.e., 441 days). This is approximately a 2% chance of eruption in a 48-hour period when a warning is in effect and a 0.06% chance outside the warning period (see Methods).

These conditional probabilities have low absolute values and thus may be better expressed as hazard ratios, which quantify forecast skill in terms of the relative likelihood of an adverse event under different warning conditions. For the example above, eruptions are an estimated thirty-three times more likely when a warning is occurring than otherwise. See Methods and Fig S8 for further details on these calculations.

Conditional probabilities and hazard ratios derived from global ML forecasters could be a useful resource for volcano observatories, particularly where expert elicitation or Bayesian methods are being used for forecasting, or where scientific decision-making is required to set volcano alert levels^{5,40}. These activities necessarily involve human reasoning and judgement that, although a clear strength, leaves room for bias or human error. Data-driven forecasts can play a complementary role, by providing a sanity check on probability calculations²⁷ or spotlighting subtle signals for follow-up by human operators. In this way, they can help inform the issuance and content of forecasts and supporting information to key users, all of which enhances effective risk communication. This, in turn, supports decision-makers as they assess risk and tailor a mitigation response, for instance, defining evacuation and re-opening thresholds^{41,40} or deciding resource allocation.

Overcoming data scarcity via an ergodic paradigm

A key finding of this study is that data scarcity may no longer be a major impediment to eruption forecasting. The use of analogue volcanoes is a well-established practice within the volcanic crisis response community^{42,43}, however, analogues that provide quantitative eruption probabilities in response to short-term data have not been developed. In this sense, the comparable accuracy of global CVV models and tailored forecasts (Table 1) is particularly significant for observatories responsible for volcanoes with limited eruption records or for those that spread their resources across large and remote volcanic regions, potentially with only one seismic station per volcano. The expertise gap has sometimes been filled by international volcanic crisis-response teams, e.g., UN, and VDAC programs¹⁰. We suggest that in future, these might be augmented through the rapid deployment global forecasting ML systems or an online webservice that monitoring data from any station can be feed into.

The global forecasts developed here are exploiting ergodicity in volcanic systems. This is a property of physical systems whereby observations drawn from a large enough (volcano) ensemble contain enough information to approximate future behaviour at a target (volcano) outside the subset. ML models exploit this relationship at the time series feature level (Fig. 3). For example, HF seismic oscillations at 40-min and 3-hours are discriminated in pre-eruptive sequences from volcanoes in the training ensemble (e.g., Mt St Helens). Later, these same features are seen to strengthen prior to an eruption at the target volcano (Augustine). Furthermore, the pre-eruptive and non-eruptive feature distributions (Fig. 3d-e) are more

discriminable than the raw seismic amplitude values (Fig. 4f), which highlights how difficult it would be for a human operator to identify these precursors without assistance from automated algorithms.

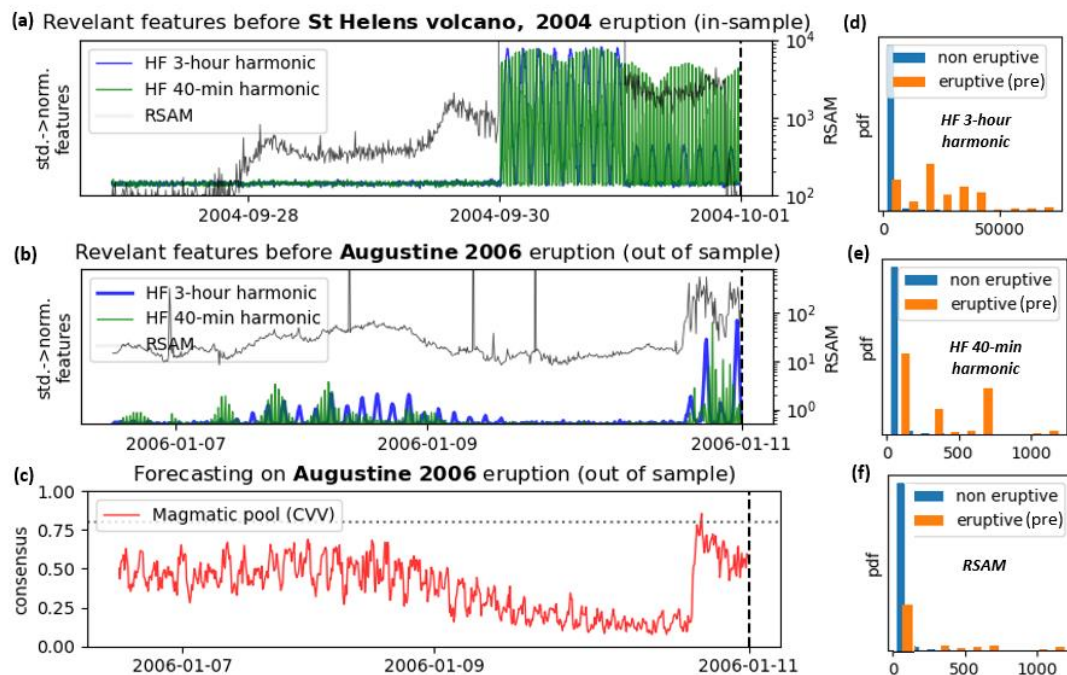


Figure 3. Illustration of precursor identification and transfer to out-of-sample target volcano. (a) Example of two features identified as significant during model training due to their elevated strength in the lead up to the 2004 eruption of St Helen's: 40-min (green) and 3-hour (blue) high-frequency (HF) oscillations. (b) The same features strengthening prior to the 2006 eruption of Augustine (out-of-sample target volcano), and the associated forecast (c). (d-e) Histograms of the two features distinguishing pre-eruptive and non-eruptive values. (f) The same distributions but for unprocessed seismic amplitude (RSAM) values.

The ergodic approach follows the long-standing paradigm of seismic hazard assessment, which relies on the compilation of large regional earthquake catalogues that can then be analysed for frequency-magnitude and ground motion trends^{44,45}. These seismic forecasting relations, when applied at a particular site, exploit an expanded spatial window to compensate a truncated time window that is limited to the instrumental era and, hence, retain accuracy. Although non-ergodic forecasts (tailored) are possible for a handful of privileged volcanoes, e.g., Whakaari with five eruptive episodes, many other volcanoes may benefit from an ergodic treatment (Fig S9).

An ergodic paradigm, facilitated by transfer machine learning, could improve forecasting for other hazards like floods and wildfires. For example, wildfire forecasting^{46,47} shares some aspects with volcano monitoring, including use of time series (meteorological) data that are proxies for system state, intermittent or infrequent hazard events, and an uneven distribution of instrumentation locations. The problem is also recognized in flood forecasting, with regionalization explored to establish a physical basis for generalizing across catchments⁴⁸. We hypothesise that even catchments or forested estate with enough data to train reliable forecasts now could nevertheless later suffer from data scarcity. We conclude that generalized or transferable forecast models that exploit ergodic features in the underlying data are ready to contribute to volcanic forecasting. In addition, the global move towards open hazard monitoring data will allow much easier aggregating of data from multiple regions to into comprehensive training datasets^{49,50} that will improve the accuracy and usefulness of forecasts across multiple hazards.

MAIN REFERENCES

- ¹Wright, T. L., & Pierson, T. C. (1992). Living with Volcanos: The US Geological Survey's Volcano Hazards Program. (No Title).
- ²Yamaoka, K., Geshi, N., Hashimoto, T., Ingebritsen, S. E. & Oikawa, T. Special issue "The phreatic eruption of Mt. Ontake volcano in 2014". Earth, Planets Space 68, 1–8 (2016).
- ³Chouet, B. Long-period volcano seismicity: its source and use in eruption forecasting. Nature 380, 309–316 (1996).

220 ⁴Marzocchi, W., Bebbington, M.S. Probabilistic eruption forecasting at short and long time scales. *Bull Volcanol* 74, 1777–
221 1805 (2012). <https://doi.org/10.1007/s00445-012-0633-x>.

222 ⁵Potter, S. H., Jolly, G. E., Neall, V. E., Johnston, D. M., & Scott, B. J. (2014). Communicating the status of volcanic activity:
223 revising New Zealand’s volcanic alert level system. *Journal of Applied Volcanology*, 3(1), 1-16.

224 ⁶Andredakis, I., & De Groeve, T. (2016). Towards a global humanitarian volcano impact alert model integrated into a multi-
225 hazard system. Geological Society, London, Special Publications, 426(1), 553-566.

226 ⁷Freire, S., Florczyk, A. J., Pesaresi, M., & Sliuzas, R. (2019). An improved global analysis of population distribution in
227 proximity to active volcanoes, 1975–2015. *ISPRS international journal of geo-information*, 8(8), 341.

228 ⁸Bolić, T., & Sivčev, Ž. (2011). Eruption of Eyjafjallajökull in Iceland: Experience of European air traffic management.
229 *Transportation research record*, 2214(1), 136-143.

230 ⁹Mastin, L., Pavolonis, M., Engwell, S., Clarkson, R., Witham, C., Brock, G., ... & Rennie, G. (2022). Progress in protecting air
231 travel from volcanic ash clouds. *Bulletin of Volcanology*, 84(1), 9.

232 ¹⁰Newhall, C. G., & Punongbayan, R. S. (1996). *Fire and mud: eruptions and lahars of Mount Pinatubo, Philippines*.
233 University of Washington Press.

234 ¹¹Schaefer, J. R. (2011). The 2009 eruption of Redoubt Volcano, Alaska. *Report of Investigations*, 5, 45.

235 ¹²Surono, Jousset, P., Pallister, J., Boichu, M., Buongiorno, M. F., Budisantoso, A., Costa, F., Andreastuti, S., Prata, F.,
236 Schneider, D., Clarisse, L., Humaida, H., Sumarti, S., Bignami, C., Griswold, J., Carn, S., Oppenheimer, C., & Lavigne, F.
237 (2012). The 2010 explosive eruption of Java’s Merapi volcano — a “100-year” event. *Journal of Volcanology and*
238 *Geothermal Research*, 241-242, 121–135. <https://doi.org/10.1016/j.jvolgeores.2012.06.018>

239 ¹³Johnson, J. B., Watson, L. M., Palma, J. L., Dunham, E. M., & Anderson, J. F. (2018). Forecasting the eruption of an open-
240 vent volcano using resonant infrasound tones. *Geophysical Research Letters*, 45(5), 2213-2220.

241 ¹⁴Barsotti, S., Oddsson, B., Gudmundsson, M. T., Pfeffer, M. A., Parks, M. M., Ófeigsson, B. G., ... & Vogfjörð, K. (2020).
242 Operational response and hazards assessment during the 2014–2015 volcanic crisis at Bárðarbunga volcano and associated
243 eruption at Holuhraun, Iceland. *Journal of Volcanology and Geothermal Research*, 390, 106753.

244 ¹⁵Endo, E.T., Murray, T. Real-time Seismic Amplitude Measurement (RSAM): a volcano monitoring and prediction tool. *Bull*
245 *Volcanol* 53, 533–545 (1991). <https://doi.org/10.1007/BF00298154>

246 ¹⁶Caudron, C., Girona, T., Jolly, A., Christenson, B., Savage, M. K., Carniel, R., ... & Mazot, A. (2021). A quest for unrest in
247 multiparameter observations at Whakaari/White Island volcano, New Zealand 2007–2018. *Earth, Planets and Space*, 73(1),
248 1-21.

249 ¹⁷Ardid, A., Dempsey, D., Caudron, C. et al. Seismic precursors to the Whakaari 2019 phreatic eruption are transferable to
250 other eruptions and volcanoes. *Nat Commun* 13, 2002 (2022). <https://doi.org/10.1038/s41467-022-29681-y>

251 ¹⁸Chouet, B. Long-period volcano seismicity: its source and use in eruption forecasting. *Nature* 380, 309–316 (1996).
252 <https://doi.org/10.1038/380309a0>

253 ¹⁹Girona, T., Caudron, C., & Huber, C. (2019). Origin of shallow volcanic tremor: The dynamics of gas pockets trapped
254 beneath thin permeable media. *Journal of Geophysical Research: Solid Earth*, 124, 4831–4861.
255 <https://doi.org/10.1029/2019JB017482>

256 ²⁰Voight, B. (1988). A method for prediction of volcanic eruptions. *Nature* 332, 125–130. doi: 10.1038/332125a0

257 ²¹Sparks, R. S. J., & Aspinall, W. P. (2004). Volcanic activity: frontiers and challenges in forecasting, prediction and risk
258 assessment. *The state of the planet: Frontiers and challenges in geophysics*, 150, 359-371.

259 ²²Malfante, M., Dalla Mura, M., Métaixian, J. P., Mars, J. I., Macedo, O., & Inza, A. (2018). Machine learning for volcano-
260 seismic signals: Challenges and perspectives. *IEEE Signal Processing Magazine*, 35(2), 20-30. Carniel, R., & Guzmán, S.
261 (2020). Chapter Machine Learning in Volcanology: A Review. <https://library.oapen.org/handle/20.500.12657/49362>

262 ²³Bebbington, M. S. (2010), Trends and clustering in the onsets of volcanic eruptions, *J. Geophys. Res.*, 115, B01203,
263 doi:10.1029/2009JB006581.

264 ²⁴Wang, T., Bebbington, M., Cronin, S., & Carman, J. (2022). Forecasting eruptions at poorly known volcanoes using analogs
265 and multivariate renewal processes. *Geophysical Research Letters*, 49, e2021GL096715.
266 <https://doi.org/10.1029/2021GL096715>.

267 ²⁵Kilgour, G., Kennedy, B., Scott, B., Christenson, B., Jolly, A., Asher, C., ... & Saunders, K. (2021). Whakaari/White Island: a
268 review of New Zealand's most active volcano. *New Zealand Journal of Geology and Geophysics*, 64(2-3), 273-295.

269 ²⁶Carniel, R., & Guzmán, S. (2020). Chapter Machine Learning in Volcanology: A Review.
270 <https://library.oapen.org/handle/20.500.12657/49362>

271 ²⁷Dempsey, D.E., Cronin, S.J., Mei, S., & Kempa-Liehr, A.W. Automatic precursor recognition and real-time forecasting of
272 sudden explosive volcanic eruptions at Whakaari, New Zealand. *Nature Communications* 11, 3562
273 <https://doi.org/10.1038/s41467-020-17375-2> (2020).

274 ²⁸Berardi, V. L., & Zhang, G. P. (2003). An empirical investigation of bias and variance in time series forecasting: modeling
275 considerations and error evaluation. *IEEE Transactions on Neural Networks*, 14(3), 668-679.

276 ²⁹Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal*
277 *Society A*, 379(2194), 20200209.

278 ³⁰Rouet-Leduc, B., Hulbert, C. & Johnson, P. A. Continuous chatter of the Cascadia subduction zone revealed by machine
279 learning. *Nat. Geosci.* 12, 75–79 (2019).

280 ³¹Ren, C. X., Peltier, A., Ferrazzini, V., Rouet-Leduc, B., Johnson, P. A., & Brenguier, F. (2020). Machine learning reveals the
281 seismic signature of eruptive behavior at piton de la fournaise volcano. *Geophysical Research Letters*, 47(3),
282 e2019GL085523. <https://doi.org/10.1029/2019GL085523>.

283 ³²Manley, G. F., Mather, T. A., Pyle, D. M., Clifton, D. A., Rodgers, M., Thompson, G., & Roman, D. C. (2021). Machine
284 learning approaches to identifying changes in eruptive state using multi-parameter datasets from the 2006 eruption of
285 Augustine Volcano, Alaska. *Journal of Geophysical Research: Solid Earth*, 126, e2021JB022323.
286 <https://doi.org/10.1029/2021JB022323>

287 ³³Biggs, J., Anantrasirichai, N., Albino, F., Lazecky, M., & Maghsoudi, Y. (2022). Large-scale demonstration of machine
288 learning for the detection of volcanic deformation in Sentinel-1 satellite imagery. *Bulletin of Volcanology*, 84(12), 100.

289 ³⁴Dempsey, D.E., Kempa-Liehr, A.W., Ardid, A. et al. Evaluation of short-term probabilistic eruption forecasting at
290 Whakaari, New Zealand. *Bull Volcanol* 84, 91 (2022). <https://doi.org/10.1007/s00445-022-01600-5>

291 ³⁵McNutt, S. R. Volcanic tremor from around the world: 1992 update. *Acta Vulc.* 5, 197–200 (1994).

292 ³⁶Chardot, L., Jolly, A. D., Kennedy, B. M., Fournier, N., & Sherburn, S. (2015). Using volcanic tremor for eruption forecasting
293 at White Island volcano (Whakaari), New Zealand. *Journal of Volcanology and Geothermal Research*, 302, 11-23.

294 ³⁷Moran, S. C., Newhall, C., & Roman, D. C. (2011). Failed magmatic eruptions: late-stage cessation of magma ascent.
295 *Bulletin of Volcanology*, 73, 115-122.

296 ³⁸Pritchard, M. E., Biggs, J., Wauthier, C., Sansosti, E., Arnold, D. W., Delgado, F., ... & Zoffoli, S. (2018). Towards
297 coordinated regional multi-satellite InSAR volcano observations: results from the Latin America pilot project. *Journal of*
298 *Applied Volcanology*, 7(1), 1-28.

299 ³⁹Leonard, G. S., Johnston, D. M., Paton, D., Christianson, A., Becker, J., & Keys, H. (2008). Developing effective warning
300 systems: Ongoing research at Ruapehu volcano, New Zealand. *Journal of Volcanology and Geothermal Research*, 172(3-4),
301 199-215.

302 ⁴⁰Christophersen, A., Behr, Y., & Miller, C. (2022). Automated Eruption Forecasting at Frequently Active Volcanoes Using
303 Bayesian Networks Learned From Monitoring Data and Expert Elicitation: Application to Mt Ruapehu, Aotearoa, New
304 Zealand. *Front. Earth Sci*, 10, 905965.

305 ⁴¹Jolly, G. E., Keys, H. J. R., Procter, J. N., & Deligne, N. I. (2014). Overview of the co-ordinated risk-based approach to
306 science and management response and recovery for the 2012 eruptions of Tongariro volcano, New Zealand. *Journal of*
307 *Volcanology and Geothermal Research*, 286, 184-207. doi:<https://doi.org/10.1016/j.jvolgeores.2014.08.028>

308 ⁴²Sheldrake, T. (2014). Long-term forecasting of eruption hazards: A hierarchical approach to merge analogous eruptive
309 histories. *Journal of Volcanology and Geothermal Research*, 286, 15-23.

- ⁴³Burgos, V., Jenkins, S. F., Bono Troncoso, L., Perales Moya, C. V., Bebbington, M., Newhall, C., ... & Taisne, B. (2023). Identifying analogues for data-limited volcanoes using hierarchical clustering and expert knowledge: a case study of Melimoyu (Chile). *Frontiers in Earth Science*, 11, 1144386
- ⁴⁴Tiampo, K. F., Rundle, J. B., Klein, W., Holliday, J., Martins, J. S., & Ferguson, C. D. (2007). Ergodicity in natural earthquake fault networks. *Physical Review E*, 75(6), 066107.
- ⁴⁵Kotha, S. R., Bindi, D., & Cotton, F. (2017). From ergodic to region-and site-specific probabilistic seismic hazard assessment: Method development and application at European and Middle Eastern sites. *Earthquake spectra*, 33(4), 1433-1453
- ⁴⁶Jain et al., (2020). A review of machine learning applications in wildfire science and management. *Environ. Rev.* 28: 478–505 (2020) dx.doi.org/10.1139/er-2020-0019^a
- ⁴⁷Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., et al. (2022). Wildfire danger prediction and understanding with Deep Learning. *Geophysical Research Letters*, 49, e2022GL099368. <https://doi.org/10.1029/2022GL099368>
- ⁴⁸Muñoz-Sabater et al., (2021). *Earth Syst. Sci. Data*, 13, 4349–4383, 2021. <https://doi.org/10.5194/essd-13-4349-2021>
- ⁴⁹Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., ... & Duan, Q. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, 19(1), 209-223.
- ⁵⁰Jain, S. K., Mani, P., Jain, S. K., Prakash, P., Singh, V. P., Tullos, D., ... & Dimri, A. P. (2018). A Brief review of flood forecasting techniques and their applications. *International Journal of River Basin Management*, 16(3), 329-344.

METHODS

Machine learning pipeline (workflow)

We used continuous seismic data from 24 volcanoes, with one station per volcano (as shown in Fig 1a and listed in Supplement Table S2). Where multiple stations were available, we selected the station that had recorded the most eruptions and was closest to the eruptive vent. The dataset includes more than 40 major eruptive episodes. The majority have been classified by the Global Volcanism Program, 2023. We excluded minor eruptive activity, such as geysering and passive ash emissions, as our focus was on the initial abrupt, impulsive (phreatic and magmatic) onset of eruption. The machine learning pipeline^{27,34} is summarized below.

Data preprocessing. We use the vertical velocity component time series at each station with the instrument response removed. We processed each station's time series to produce four data streams, each its own time series with a sampling interval of 10 minutes^{17,34}. Three data streams are generated by bandpass filtering to three frequency ranges, which captures different parts of the volcano-seismic signal. Filtering between 2 to 5 Hz focuses on a tremor signal of frequent volcanic origin while excluding ocean noise at lower frequencies (mainly < 1 Hz). This filtered time series was incremented into 10-min, non-overlapping windows with the average absolute velocity computed for the window. We have referred to this as Real-time Seismic Amplitude Measurement (RSAM) although we note that RSAM is traditionally computed on the unfiltered or very loosely filtered trace, e.g., 0.5-20 Hz to exclude microseism/anthropogenic contributions. We compute Median Frequency (MF) and High Frequency (HF) data streams in the same way as RSAM but instead filtering in the range of 4.5 to 8 Hz and 8 to 16 Hz, respectively. These two data streams focus on signal attenuation effects above the frequency range tremor energy is commonly radiated¹⁶. Finally, Displacement Seismic Amplitude Ratio (DSAR) is calculated as the ratio of the integrals of the MF and HF signals¹⁷. High values of DSAR have been inferred to correlate with high gas levels in the edifice, suggesting either reduced fluid motion and/or trapping that has led to a gas-accumulation^{17,51} (Ardid et al., 2022-23). Time series gaps were imputed by linear interpolation and because these gaps occur in noneruptive periods, a later step of down sampling reduces their impact on the models²⁷.

Regional earthquake filtering. To isolate continuous volcanic signals, we removed signatures from regional and volcano-tectonic (VT) earthquakes. Wave trains from earthquakes perturb time-averaged station velocity above the background, which then confuses ML algorithms that don't have the context of human operators. To identify and remove earthquake effects, we applied outlier detection to the raw waveform data before bandpass filtering³⁴. We excised a 2-minute envelope from the raw velocity trace whenever its value exceeded three standard deviations above the mean of the data in the 10-minute window. This removed signal contamination from most short duration events. Contamination from long-duration teleseismic waves from large-magnitude subduction earthquakes still occurred and was removed by applying a two-window (20-minute) moving minimum to the data stream³⁴.

Data normalization: We used z-score normalization applied in log-space to eliminate disparate magnitude and ranges effects arising between the different data streams³⁴. This produces normalized distributions of data stream values, which prevents one data type from dominating feature selection. It also improves comparisons between volcanoes where stations are located at different distances from the craters. In log-space, distributions exhibit central tendencies and do not display excessively large asymmetries (Fig S10). Normalization uses the entire signal record and is hence a potential source of information leakage. However, we checked the effect of removing the small amount of pre-eruption data from normalization and the effects are imperceptible. An alternative normalization approach could be to use reduced displacement^{52,53}.

Feature calculation. Data time series are divided into 48-hour windows, with each window consisting of 288 10-minute samples. During model training, we allow adjacent windows to overlap by 36 hours (75% of their length) which increases the likelihood that a precursor pattern is fully captured within at least one window rather than split by window boundary. During model forecasting, adjacent windows overlap almost entirely, except for a one sample (10-minute) shift forward in time. For each data stream in each window, we calculate over 700 time series features using the Python package "tsfresh"⁵⁴. These features include distribution measures (mean, standard deviation, number of peaks), correlation measures (lags, Fourier coefficients), linearity (slope and intercept of linear regressors), and other information measures (entropy, energy, nonlinear scores). All feature values are stored in a matrix, where individual columns represent different features, and rows represent the end time index of each window. Our choice to use a 48-hour window focuses on detecting short-term fluctuations in volcano state. This ultimately derives from the intent of this analysis, which is to provide information on short-term changes in hazard that would be useful for responsive volcanic monitoring. We affirm that shorter or longer time windows might also be useful and should be explored thoroughly in future work. However, previous studies^{27,34} explored a range of windows between 12 hours and 5 days at Whakaari and found that the choice did not greatly impact forecast accuracy at that volcano. We also acknowledge that some volcanoes show meaningful unrest on longer timescales⁵⁵.

Labelling. We define a classification problem that prioritizes the algorithm attention on signals occurring in the 48-hours before eruptions, called the "look forward" window. All feature windows whose end time occurs 48 hours or less before the date of an eruption onset are labeled '1', while the remaining windows are labeled as '0'. For training windows that overlap 75%, this results in 160 label-1 windows (pre-eruptive; four for each of the 40 eruptions in the dataset) and more than 50 000 label-0 windows (non-eruptive). The large imbalance is handled by randomly discarding non-eruptive '0' windows until the dataset is balanced. This is repeated many times to reduce the effects of randomness on down-sampling.

Feature ranking. For a given down-sampled window subset, feature values are subdivided into groups corresponding to label-0 and label-1. The importance of each feature is then determined by whether there is a statistically significant difference between the distribution of feature values. Significance is assessed using the Mann-Whitney U test, which evaluates for a possible difference in median between two distributions. Features are retained as significant if the output p-value from the test falls below an adjusted Benjamini-Yekutieli threshold^{56,27}, which controls for a fixed false discovery rate of 5% when conducting many statistical tests. Features are then ranked (by lowest p-value) and passed on for model training.

Training. First, out-of-sample volcano data are set aside for testing and not used during model training. We use a random forest algorithm to solve the classification problem. Each model consists of up to 200 individual decision trees, with each tree having been trained on a different subset of down-sampled windows. Thus, all non-eruptive data appears, at a very low frequency, across the training set²⁷. Models are trained using the scikit-learn Python library⁵⁷. The random forest classification model we implement has been preferred as it was previously found to outperform other classifiers on this problem²⁷.

Pseudo-prospective Forecasting. Forecasts are only constructed for data withheld during model training, which approximates but does not perfectly replicate prospective forecasting conditions. However, this approach provides a better assessment of performance than a hindcast, which include test data in the training set. For the test data, feature time series are calculated using overlapping 48-hour windows that each advance the previous by 10-minutes and hence overlap at 287 out of 288 points. This is intended to imitate real-time volcano monitoring where data is interpreted as it is received. Features from each subsequent window are passed to the trained random forest model. The binary outputs of each decision tree are averaged to produce a value ranging from 0 to 1, referred to as the consensus of the forecast model.

Warning model. For illustration purposes, we have considered a simple warning model, however, any future implementation will vary depending on the needs of the forecast user. The consensus output is compared to a pre-determined threshold between 0 and 1. If the consensus exceeds the threshold, then an eruption warning is created for the 48 hours that follow (the forecast's look forward period). The warning period restarts every time the consensus goes

back above the threshold. Therefore, a warning is only rescinded after a consecutive 48-hours of the forecast consensus remaining at or below the threshold. Warnings may last considerably longer than 48-hours for periods of high model consensus or if a low threshold value is used.

Volcano Training Pools. Volcanoes were divided into distinct groups or "pools" based on their reported eruption types. We used three specific pools in this study (Fig 1b). The magmatic pool comprised volcanoes in Alaska (Redoubt, Augustine, Veniaminof, Pavlof, Great Sitkin, Semisopochnoi, Okmok), Bezymianny in Kamchatka, Russia, and Mt. St. Helens in Washington, USA. These volcanoes were included due to their shared tectonic origin, geographic proximity, and tendency to produce magmatic eruptions. The phreatic pool included New Zealand volcanoes (Whakaari, Ruapehu, Tongariro), as well as Mt. Ontake in Japan (Yamaoka et al., 2016), Copahue in Chile/Argentina⁵¹, and Kawah Ijen in Indonesia⁵¹. These volcanoes all have well developed hydrothermal systems. Finally, a world pool comprising all the volcanoes was included.

Our experience has been that the best pool for out-of-sample forecasting on a target volcano needs to be determined on a case by case basis. The three pools we have studied are not the only possible choice for blending volcano data and any future operational implementation should involve careful consideration of other combinations that might have higher predictive skill.

Forecast Evaluation. The purpose of forecast testing is to determine overall accuracy, including the false positive rate (incorrectly forecasted that an eruption will occur), false negative rate (failing to forecast an eruption), and the correlation between forecasted probabilities and actual outcomes over long time periods. Accuracy can be tested under pseudo-prospective and prospective conditions with the latter needing real-time or future data to assess a frozen model³⁴. Pseudo-prospective forecast performance is not guaranteed to replicate real-time prospective performance.

In machine learning, cross-validation is used to evaluate model performance as well as its ability to generalize to new data or contexts. Leave-One-Out Cross-Validation (LOOCV) trains a model using all samples (volcanoes or eruptions) in a dataset except for one, and then using the excluded sample as a validation set to measure model performance. This is repeated for each sample in the dataset and the performance of all models is combined for an overall estimate of model performance. Here, we applied LOOCV at both the volcano (CVO) level to construct global ML forecasts and at the individual eruption (CVE) level for the comparison to tailored models (Fig 1c). For the latter, a single out-of-sample forecast at a volcano is assembled from multiple CVE forecasts of its individual eruptions. For the remaining non-eruptive data, a single model trained with all eruptions at the volcano is used, which, due to under-sampling of the non-eruptive data, is approximately out of sample.

Operating Characteristic (ROC) curves. Receiver ROC curves are used to measure the discriminability of classification models, that is, its ability to correctly classify that an eruption will or will not occur. ROC curves plot the true positive rate (sensitivity) against the false positive rate (1-specificity) for different choices of a threshold value. Hence, they evaluate the forecast across a range of threshold values with the trade-off between sensitivity and specificity allowing for an optimal threshold selected. The area under the (ROC) curve (AUC) is a single measure of model performance at all thresholds, with values approaching 1 indicating perfect performance.

Here, we constructed ROC curves according to the following steps: (1) a threshold is selected, (2) the threshold is applied to a forecast to produce a warning model; (3) a true positive (TP) or true negative (TN) is scored in each window that the warning model agrees with the label: both '1' for TP or both '0' for TN; (4) a false positive (FP) is scored when the warning model outputs a '1' but this does not match a corresponding '0' of the label vector; a false negative (FN) is scored for all other mismatches; (5) the true positive rate (TPR) and false positive rate (FPR) are calculated as $TPR = TP / (TP + FN)$ and $FPR = FP / (FP + TN)$; (6) steps (1-5) are repeated using 100 evenly spaced thresholds between 0 and 1 with the set of matching pairs forming the ROC curve; finally (7) AUC is calculated as the integral of the ROC curve.

Seismic Amplitude Forecast. To compare against the global ML models, we construct a simple forecast based on seismic amplitude. An RSAM forecast consensus is calculated by transforming the raw RSAM record (calculated here) at each volcano to an equivalent percentile value, e.g., an RSAM value at the median would correspond to an RSAM forecast consensus of 0.5. The percentile evaluation is done on a volcano-by-volcano basis and, as it assumes knowledge of future extreme maximum values, it is not pseudo-prospective. This means the RSAM model has an information advantage over global ML models however we deemed this acceptable as it is used here only as a reference case. The RSAM model was further advantaged by testing two variations of it, using a rolling median over six hours or two days. This was to avoid the high rate of false alarms that can be triggered by stochastic but unsustained high values of RSAM. We selected the version with the highest accuracy (6-hour rolling median) for comparison to the global ML models.

Conditional probability of eruption. For a warning model that has been constructed for M eruptions in a volcano pool, the events are subsetting into eruptions occurring inside a warning, M_{in} , and the rest outside, M_{out} , i.e., $M_{in} + M_{out} = M$. The rate of eruptions when a warning is in effect is $\lambda_{in} = M_{in}\tau/T_{in}$ where τ is the forecast look forward period (2 days) and T_{in} is the total length of all warnings. Similarly, the rate of eruptions outside a warning is $\lambda_{out} = M_{out}\tau/T_{out}$ where T_{out} is the total length of time not under warning. Assuming a Poisson model of eruptions, the probability of one or more eruptions in the forecast period, τ , when a warning is in effect is $P_{in} = 1 - \exp(-\lambda_{in})$ and when a warning is not in effect is $P_{out} =$

$1 - \exp(-\lambda_{out})$. This assumes that each eruptive event is independent of others, and the eruption rate is constant inside and outside of warnings. This necessarily simplified approach is suitable for comparing probabilities in our study. However, more advanced techniques like nonhomogeneous Poisson processes^{34,58} could be used to incorporate greater complexity into the probability calculations.

Hazard ratios. The hazard ratio is the relative likelihood of an eruption in the look-forward period given that a warning is in effect. It is calculated as $HR = P_{in}/P_{out}$ and it is a measure of a forecast's ability to discriminate pre-eruptive and non-eruptive conditions. For example, the CVV phreatic pool model forecasting on Whakaari eruptions with a threshold of 0.8 has $M_{in}=4$, $M_{out}=1$, $T_{in}=3678 \times 0.9$ days, $T_{out}=3678 \times 0.1$ days. The calculated eruptions rates (per 48-hours) are $\lambda_{in}=0.0197$ and $\lambda_{out}=0.00061$, and the probabilities are $P_{in}=0.0196$ and $P_{out}=0.000611$. Finally, the hazard ratio is 32. In comparison, a tailored Whakaari model using a threshold of 0.88 has $M_{in}=4$, $M_{out}=1$, $T_{in}=3678 \times 0.04$, $T_{out}=3678 \times 0.96$. The corresponding eruption rates (per 48-hours) are $\lambda_{in}=0.0544$ and $\lambda_{out}=0.000566$, with probabilities of $P_{in}=0.0529$ and $P_{out}=0.000566$. Finally, the hazard ratio is 93. See Fig S8 for further details on these calculations.

Avenues for enhancing methods. To enhance accuracy, avenues for improvement include expanding precursory analysis by considering longer time periods (multi-time scale approaches; Ardid et al., 2023) and incorporating additional data types such as gas emission rate^{59,60}, thermal anomalies⁶¹, and magnetotelluric data^{62,63,64}. Also, application to geysers monitored with seismic data^{65,66} given their frequent eruptive nature and lower time between eruption offers an opportunity to improve these methods with more balanced datasets between eruptive and non-eruptive data for the ML classification problem.

DATA AND CODE AVAILABILITY

The codes required to replicate the results of this study are available at <https://github.com/ddempsey/puia>, released under the MIT license. Codes are provided in the Supplemental Material to compute the ROC curves for all three types of foresters. Raw waveform data for New Zealand volcanoes can be downloaded from GEONET (<https://service.geonet.org.nz>); and for Alaskan volcanos from IRIS (<http://service.iris.edu>). Both are operable as clients though the FDSN webservice (<https://www.fdsn.org/networks/>). The seismic data streams used in this study (rsam, mf, hf, dsar) for the New Zealand and Alaskan volcanoes are available as supplementary material in the form of .csv files. If you utilize the provided seismic data, please cite this article accordingly.

METHODS REFERENCES

⁵¹Ardid, A., Dempsey, D., Caudron, C., Cronin, S. J., Miller, C. A., Melchor, I., et al. (2023). Using template matching to detect hidden fluid release episodes beneath crater lakes in Ruapehu, Copahue, and Kawah Ijen volcanoes. *Journal of Geophysical Research: Solid Earth*, 128, e2023JB026729. <https://doi.org/10.1029/2023JB026729>

⁵²Aki, K., and Koyanagi, R. (1981). Deep volcanic tremor and magma ascent mechanism under Kilauea, Hawaii, *J. Geophys. Res.*, 86(B8), 7095–7109, doi:10.1029/JB086iB08p07095.

⁵³Sherburn, S., Bryan, C. J., Hurst, A. W., Latter, J. H., & Scott, B. J. (1999). Seismicity of Ruapehu volcano, New Zealand, 1971–1996: a review. *Journal of volcanology and geothermal research*, 88(4), 255–278.

⁵⁴Christ, M., Braun, N., Neuffer, J. & Kempa-Liehr, A. W. Time series feature extraction on basis of scalable hypothesis tests (tsfresh—a python package). *Neurocomputing* 307, 72–77 (2018).

⁵⁵Roman, D. C., & Gardine, M. D. (2013). Seismological evidence for long-term and rapidly accelerating magma pressurization preceding the 2009 eruption of Redoubt Volcano, Alaska. *Earth and Planetary Science Letters*, 371, 226–234

⁵⁶Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

⁵⁷Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).

⁵⁸Bebbington, M.S., Lai, C.D. On nonhomogeneous models for volcanic eruptions. *Math Geol* 28, 585–600 (1996). <https://doi.org/10.1007/BF02066102>.

⁵⁹De Moor, J. M., Stix, J., Avar, G., Muller, C., Corrales, E., Diaz, J. A., ... & Fischer, T. P. (2019). Insights on hydrothermal-magmatic interactions and eruptive processes at Poás Volcano (Costa Rica) from high-frequency gas monitoring and drone measurements. *Geophysical Research Letters*, 46(3), 1293–1302.

⁶⁰Burton, M., Hayer, C., Miller, C., & Christenson, B. (2021). Insights into the 9 December 2019 eruption of Whakaari/White Island from analysis of TROPOMI SO₂ imagery. *Science Advances*, 7(25), eabg1218.

⁶¹Girona, T., Realmuto, V. & Lundgren, P. Large-scale thermal unrest of volcanoes for years prior to eruption. *Nat. Geosci.* 14, 238–241 (2021). <https://doi.org/10.1038/s41561-021-00705-4>

⁶²Aizawa, K., Kanda, W., Ogawa, Y., Iguchi, M., Yokoo, A., Yakiwara, H., & Sugano, T. (2011). Temporal changes in electrical resistivity at Sakurajima volcano from continuous magnetotelluric observations. *Journal of Volcanology and Geothermal Research*, 199(1-2), 165-175

⁶³Ardid, A., Archer, R., Bertrand, E., Sepulveda, F., Tarits, P., & Dempsey, D. (2021a). Heat transfer through the Wairakei-Tauhara geothermal system quantified by multi-channel data modeling. *Geophysical Research Letters*, 48(8), e2020GL092056. <https://doi.org/10.1029/2020GL092056>

⁶⁴Ardid, A., Dempsey, D., Bertrand, E., Sepulveda, F., Tarits, P., Solon, F., & Archer, R. (2021b). Bayesian magnetotelluric inversion using methylene blue structural priors for imaging shallow conductors in geothermal fields. *Geophysics*, 86(3), E171–E183. <https://doi.org/10.1190/geo2020-0226.1>

⁶⁵Ardid, A., Vera, E., Kelly, C., Manga, M., Munoz-Saez, C., Maksymowicz, A., & Ortega-Culaciati, F. (2019). Geometry of geyser plumbing inferred from ground deformation. *Journal of Geophysical Research: Solid Earth*, 124(1), 1072-1083. <https://doi.org/10.1029/2018JB016454>

⁶⁶Sudibyo, M. R., Eibl, E. P., Hainzl, S., & Hersir, G. P. (2022). Eruption Forecasting of Strokkur Geyser, Iceland, Using Permutation Entropy. *Journal of Geophysical Research: Solid Earth*, 127(10), e2022JB024840.

ACKNOWLEDGEMENTS

We acknowledge GEONET (New Zealand) and AVO (Alaskan Volcano Observatory, USA) seismic monitoring system for supplying free access to their volcanic seismic data. We wish to thank NZ MBIE UOAX1913 (Transitioning Taranaki to a Volcanic Future) for support of AA, DD, and SJC during this study. We wish to thank NZ MBIE E7774 (Adapting to climate change through stronger geothermal enterprises') for support of AA, DD, and SJC during this study. LC thanks the support of the 'Programa de Riesgo Sísmico' (PRS, Actividades de Interés Nacional, Universidad de Chile). SP was supported by the New Zealand Ministry of Business, Innovation & Employment (MBIE) through the GNS Science Hazard & Risk Management programme (Strategic Science Investment Fund, contract C05X1702). We thank the Icelandic Meteorological Office (IMO) and Dr Kristín Jónsdóttir for providing data for the Eyja. volcano. We thank the Observatorio Volcanológico de los Andes del Sur (OVDAS-SERNAGEOMIN) and Dr Francisco Delgado for providing the seismological data of the Caulle volcano through the Portal de Transparencia system, which is freely accessible through the web site <https://www.portaltransparencia.cl/PortalPdT/>. We thank the Dr Ivan Melchor and 'Instituto de Investigación en Paleobiología y Geología' (UNRN-CONICET), Argentina, for providing the seismological data of the Copahue volcano. We thank Dr Yuta Maeda and the Nagoya University, Japan, for providing data of the Mt Ontake. We also thank Dr Yasua Ogawa for its support for collecting this data. We present and analyze the pre-eruptive performance of models exclusively for volcanoes where data were collected through open access, which are 12 volcanoes in the US and New Zealand. This is due to observatories sensitivities and shared data agreements. However, it's important to note that the models are trained using all available data for each respective pool.

AUTHOR CONTRIBUTIONS

A.A. and D.D. developed the machine learning pipeline, including the design and codes provided. C.C, T.G., B.K., and S.C. contribute to the analysis, interpretation and contributed to manuscript development. C.M, L.C., S.R., S.P., J.P., M.M., A.M., Y.C., S.A., O.L. and D.R. contributed to the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ED.pdf](#)