Evaluation of short-term probabilistic eruption forecasting at Whakaari, New Zealand

D. E. Dempsey1 · A. W. Kempa-Liehr2 · A. Ardid1 · A. Li2 · S. Orenia2 · J. Singh2 · A. J. Tyler2 · S. J. Cronin2

Received: 18 February 2022 / Accepted: 31 August 2022 / Published online: 12 September 2022 © The Author(s) 2022

Abstract
Phreatic explosions at volcanoes are difficult to forecast but can be locally devastating, as illustrated by the deadly 2019 Whakaari (New Zealand) eruption. Quantifying eruption likelihood is essential for risk calculations that underpin volcano access decisions and disaster response. But estimating eruption probabilities is notoriously difficult for sudden onset eruptions. Here, we describe two retrospectively developed models for short-term (48 h) probabilistic forecasting of phreatic eruptions at Whakaari. The models are based on a pseudo-prospective analysis of seven Whakaari eruptions whose precursors were identified by time series feature engineering of continuous seismic data. The first model, an optimized warning system, could anticipate six out of seven eruptions at the cost of 14 warning days each year. While a warning is in effect, the probability of eruption is about 8% in 48 h, which is about 126 times higher than outside the warning. The second model used isotonic calibration to translate the output of the forecast model onto a probability scale. When applied pseudo-prospectively in the 48 h prior to the December 2019 eruption, it indicated an eruption probability up to 400 times higher than the background. Finally, we quantified the accuracy of these seismic data-driven forecasts, alongside an observatory expert elicitation that used multiple data sources. To do this, we used a forecast skill score that was benchmarked against the average rate of eruptions at Whakaari between 2011 and 2019. This exercise highlights the conditions under which the three different forecasting approaches perform well and where potential improvements could be made.

Keywords Phreatic eruption · Forecasting · Whakaari · Probability · Machine learning · Elicitation

Introduction
Active volcanoes are popular tourist attractions, but they can also pose a threat to visitors (Baxter & Gresham, 1997; Heggie, 2009; Bertolaso et al., 2009). At some volcanoes, steam-driven explosions can occur with little warning and deadly consequences (Auker et al., 2013; Brown et al., 2017). For example, at Whakaari (White Island), New Zealand, an eruption on 9 December 2019 at 01:11 (all times in UTC unless otherwise noted) killed 22 tourists and guides and severely injured 26 others who were visiting the volcanic island (MOJ, 2020). Similarly, in 2014, a phreatic eruption at Mt Ontake, Japan, killed ~60 hikers (Yamaoka et al., 2016; Mannen et al., 2019). Although the impacts of these eruptions are local, they can still be deadly if there are tourists gathered on-site.

As a result, volcano observatories mandated with monitoring such unpredictable volcanoes routinely track a wide range of data, e.g. seismic, geodetic, visual observations, geochemical and temperature. They also curate regular assessment metrics, reports and web-based services, while working closely with national and local government agencies (UNDRO, 1985; Harris et al., 2017; Pallister et al., 2019). Monitoring active volcanoes is even more complicated when there is an active hydrothermal system or other non-eruptive unrest processes that create a noisy and unpredictable environment (Gottsmann et al., 2007; Jolly et al., 2020). At such sites, particularly those prone to sudden hydrothermal explosive events, an essential field of study is the discrimination of pre-eruption signals (precursors) from other non-eruptive noise (Mannen et al., 2019; Ardid et al., 2022). This information can be used by forecast models whose goal is to quantify short-term eruption probabilities...
at hydrothermal systems. However, identifying these precursors in real-time and without the benefit of hindsight is extremely challenging.

Once a hazard forecast (e.g., eruption probability) has been constructed, it is communicated to the public or other stakeholders (Solana et al., 2008; Doyle et al., 2014; Wadge & Aspinall, 2014). One system for doing this is the Volcano Alert Level (VAL) framework (e.g., UNDRO, 1985; Fearnley et al., 2012; Potter et al., 2014). The alert levels are generally an increasing ordinal scale that, depending on the country of implementation, can categorize some pre-defined combination of current or future volcanic phenomena, hazard, risk, or response. Alert levels are typically set based on a scientific assessment of monitoring data and could vary from a low “background” level to a high “major eruption” level, although definitions will vary (De la Cruz-Reyna & Tilling, 2008; Fearnley et al., 2012; Kato & Yamasato, 2013).

A VAL system is used at New Zealand volcanoes (Potter et al., 2014), including Whakaari. Here, it communicates a collective assessment by volcano scientists of the current volcanic phenomena, i.e., future eruption likelihood is explicitly precluded. Monitoring scientists assess the broad range of available data (seismic, gas, deformation, etc.), alongside their prior experience, and pass this information along as a collective assessment by volcano scientists of the current volcanic phenomena, hazard, risk, or response. Alert levels are typically set based on a scientific assessment of monitoring data and could vary from a low “background” level to a high “major eruption” level, although definitions will vary (De la Cruz-Reyna & Tilling, 2008; Fearnley et al., 2012; Kato & Yamasato, 2013).

A VAL system is used at New Zealand volcanoes (Potter et al., 2014), including Whakaari. Here, it communicates a collective assessment by volcano scientists of the current volcanic phenomena, i.e., future eruption likelihood is explicitly precluded. Monitoring scientists assess the broad range of available data (seismic, gas, deformation, etc.), alongside their prior experience, and pass this information along as a discrete level of volcano activity, usually determined by a vote (Potter et al., 2014; Deligne et al., 2018). The possible levels are zero (no unrest), one (minor unrest), two (moderate to heightened unrest), three (minor eruption), four (moderate eruption) and five (major eruption).

At a minimum, VAL is reconsidered each week at a meeting of the monitoring scientists. However, any team member can call for an ad hoc monitoring meeting to reassess VAL at any time, for instance, when new signs of unrest are recognized (G Jolly, pers. comm., 2022). For example, 3 weeks before the 2019 eruption at Whakaari, amid a building tremor and gas signal (GeoNet, 2019a), the VAL was raised to two, which is the highest indication of non-eruptive volcanic unrest. VAL changes are frequently accompanied by published Volcano Alert Bulletins (since renamed Volcanic Activity Bulletins) that provide additional narrative text describing the volcanic activity and, sometimes, eruption likelihood.

Once the VAL is determined, a separate elicitation is conducted, surveying those same scientists for their independent judgements on the probability of a future eruption. This probability is associated with a defined time period (weeks to months) that depends on the current VAL (Deligne et al., 2018). These estimates are collated and a range of values are reported, typically the 50th to 84th percentile (Christophersen et al., 2022). For example, 1 week before the 2019 Whakaari eruption, this elicitation of probabilities judged there to be an 8–14% likelihood of eruption in the following 4 weeks (GeoNet, 2019b). Elicitations are also renewed if the prior forecast expires and the volcano is still at a VAL one or higher (Deligne et al., 2018).

Elicitation generally occurs when there is a change of VAL. This means that new eruption probabilities at volcanoes displaying unrest are calculated every 1 to 3 months, or on the request of a monitoring scientist. Fearnley and Beven (2018) have suggested that end-users of VAL systems might be better served by timely (more rapid) alert information. However, there are inevitably human, time and financial resource limits that constrain the frequency and detail of these communications. Poland and Anderson (2020) have suggested that VAL systems could be integrated with automated, real-time warning systems (cf. Thompson & West, 2010; Ripepe et al., 2018) and might benefit from recent machine-learning advances (Carniel & Guzmán, 2020). Machine learning has previously been applied to understand signals that occur before (Manley et al., 2021), during (Ren et al., 2020) and at the end of (Manley et al., 2020) of eruptive periods. In an earlier study of Whakaari (Dempsey et al., 2020), we used a machine-learning model to transform a continuous input seismic signal into a short-term eruption potential score. Documented appropriately, these models could provide an updating eruption likelihood at high frequency, while also filling gaps in human attentiveness (Thompson & West, 2010).

In this study, we describe the development, application and accuracy of short-term, automated eruption forecasting systems at Whakaari volcano, New Zealand. The models we have developed are based on a retrospective analysis of seven previous eruptions at the volcano. One of the forecast models has been deployed for an 18-month period of real-time prospective testing. This article describes the forecaster’s performance during the testing period, several improvements that arose from this exercise, and a reappraisal of its accuracy for future prospective use. Finally, we consider the 2019 eruption at Whakaari, and the associated science response by the monitoring agency, as a case study to illustrate how such automated systems could contribute during future volcanic crises.

### Short-term eruption forecasting

Systems for estimating near future eruption probabilities have several key requirements. First, they need access to information about the changing volcano state, i.e. monitoring data. Second, from the data, they need to be able to discriminate possible warning signs of eruption, i.e. precursors. Third, given a set of precursors, a method is needed to estimate the probability of an eruption, i.e. the forecast. Finally, being able to establish the accuracy of forecasts can help build confidence in their operational use.
Eruption precursors

Possible pre-eruption signals have been identified using satellite-based observations of geodetic or thermal anomalies (e.g. Oppenheimer et al., 1993; Biggs et al., 2014; Girona et al., 2021), from seismic networks (e.g. Chastin & Main, 2003; Roman and Cashman, 2018; Pesicek et al., 2018) and with gas-based measurements (e.g. Aiuppa et al., 2007; de Moor et al., 2016; Stix and de Moor, 2018). These signals can be detected over time scales of hours (Larsen et al., 2009; Dempsey et al., 2020) to weeks (Roman & Cashman, 2018; Pesicek et al., 2018; Ardid et al., 2022), and even years (Caudron et al., 2019; Girona et al., 2021). For example, at Mt Etna, infrasound is used to detect the signals of building Strombolian activity, and this has provided the basis of a successful early warning system for lava fountains since 2010 (Ripepe et al., 2018). However, this style of system is tailored to open-vent volcanoes where pre-eruption signals directly couple to the atmosphere.

To be useful for forecasting, eruption precursors should satisfy two criteria (Ardid et al., 2022): (1) they should reliably recur before most eruptions, and (2) they should be differentiable from non-eruptive unrest or repose by recurring at a reduced frequency during these periods. The latter property is important if the signal is later to be used to estimate a probability of eruption. For example, Chastin and Main (2003) noted building seismic levels at Kilauea in the days prior to effusive eruptions, but showed this was statistically indistinct from the repose seismicity. Pesicek et al. (2018) analysed 20 eruptions across eight Alaskan volcanoes and showed that only 30% had pre-eruptive seismic rate anomalies. Although rate anomalies were more frequent before larger VEI eruptions (43% for VEI ≥ 2 and 56% for VEI ≥ 3), they were less common (13%) at open-vent systems. In general, it is difficult to prove that a precursor is both recurring and discriminating unless monitoring records include multiple eruptions and long periods of repose. Furthermore, once a precursor is established and then observed, a method is needed to calculate the probability of eruption.

A precursor must also be practical to observe. Generally, this imposes a restriction on the type of data that can be used for automated, real-time eruption forecasting. At Whakaari, two telemetered seismic stations provide near real-time monitoring data at a high rate. The data they provide have been used in several studies of precursory activity at the volcano (Chardot et al., 2015; Jolly et al., 2018; Park et al., 2020; Dempsey et al., 2020; Ardid et al., 2022). For example, strong bursts in the seismic data recurred in the hours and days prior to four of the five most recent eruptive periods (Dempsey et al., 2020). This suggests there are repeatable signals that automated systems could be trained to monitor. There have also been indications of satellite-detectable deformation and degassing prior to the 2019 Whakaari eruption (Harvey, 2021; Hamling & Kilgour, 2021; Burton et al., 2021). However, these data types at Whakaari are less frequent, with irregular over-flights to sample CO2 and SO2, or daily-to-weekly satellite passes (Harvey, 2021; Burton et al., 2021). For these reasons, seismic is the sole data stream used to develop forecast models in this study.

Forecast types and testing

The term forecast has a range of meanings, which vary in their diversity of outcomes, implied precision and time/space windows. Here, we adopt a narrow definition of an eruption forecast that is suitable for the goals of this study. A forecast is defined as the estimated probability (e.g. 5%) that an eruption will occur in some pre-determined and specified future time window (e.g. 48 h). This differs from the presumption that an eruption is expected to occur, in which case the forecast is an estimate of the expected eruption start time (e.g. De la Cruz-Reyna & Reyes-Dávila, 2001; Gregg et al., 2022). The forecast can be superseded by a new one before the time window of the previous forecast expires.

Forecasts also differ in terms of the information that is available when they are constructed. A prospective forecast is generated using only geophysical information available at the present time, i.e. it is solely reliant on foresight. These are the forecasts that are constructed during times of volcanic unrest or crisis (e.g. De la Cruz-Reyna & Reyes-Dávila, 2001; Wright et al., 2013; Gregg et al., 2022). A retrospective forecast, also called a hindcast, is constructed after an eruption has occurred. It has the same geophysical information as a prospective forecast but is aided by knowing that an eruption did occur and some idea of its timing, i.e. the benefit of hindsight. Hindcasting is a good exercise for testing physical models of eruption precursors (e.g. Chardot et al., 2015; Crampin et al., 2015; Gregg et al., 2022).

There is a third type of forecast, termed pseudo-prospective, that mocks the conditions of a prospective forecast by minimizing as far as possible the benefits of hindsight (e.g. Bell et al., 2018; Dempsey et al., 2020). This can be achieved through a train-test split, where a model is developed using, and then calibrated to, a subset of data and then tested on held-out eruptions (cf. Dempsey et al., 2020). Of course, the forecast maker still knows about the “unseen” eruption. This is why it is essential to devolve, as much as is practical, the tasks of model development and tuning to automated algorithms that can be “tricked” into not knowing about the test data. For the models developed here, this includes using algorithms that select the most relevant time series features from training data (model development) and automated training of the classifier that will produce forecasts on the training data (tuning). Pseudo-prospective forecasting is useful for developing, testing and improving forecasting models as much as possible before embarking on lengthy periods of
prospective testing (e.g. Tashman, 2000; Steacy et al., 2014; Steyerberg & Harrell, 2016).

One goal of forecast testing is to establish accuracy. This could include measuring rates of false positives (an eruption is predicted, but does not occur), false negatives (an eruption occurs, but was not predicted) or the long-term correlation between probabilities and realized outcomes. Forecast accuracy can be partially assessed using pseudo-prospective testing. However, the highest confidence will always be attained when testing under prospective conditions (Schorlemmer et al., 2007; Strader et al., 2017; Whitehead and Bebbington, 2021). In a review of many types of eruption forecast model, Whitehead et al. (2022) argued that, following prospective testing, any iterative improvement to a forecasting model must necessarily restart prospective testing from scratch. This requirement is suggested only for machine-learning forecast models, sometimes referred to as data-driven due to their reliance on patterns recognized in data. Whitehead et al. (2022) did not provide guidance on whether other forecast models — elicitation, event trees, belief networks — should be held to a similar standard. This latter set is sometimes referred to as expert-driven models because of their reliance on human understanding of volcanic processes (e.g. Hincks et al., 2014; Christophersen et al., 2018; Wright et al., 2013).

Forecasting models developed in this study

Our study begins with a description of the Dempsey et al. (2020) forecasting system and a prospective analysis of its operation between February 2020 and August 2021. Three false positives were issued during this 18-month period and understanding their causes allows several model improvements, including regional earthquake filtering and data normalization.

Next, we introduce two data-driven methods for calculating short-term (48 h) probabilities of phreatic eruptions at Whakaari. The first method is based on a warning-style system, where periods of elevated eruption probability are declared when the output of a forecasting model exceeds a trigger value. A pseudo-prospective analysis of this system allows us to express the conditional eruption probability, which quantifies eruption likelihood during and outside a warning. The second method relies on a calibrating function that directly converts a forecast model output to a time-continuous probability. The improved models are tested in a pseudo-prospective sense to estimate their accuracy for use in prospective forecasting.

It is important to emphasize that this analysis is ultimately retrospective because we are learning from eruptions that have already occurred. However, when training and testing models, we adhere strictly to the pseudo-prospective standard. This means, unless explicitly indicated otherwise, we only report eruption probabilities that were calculated when the forecast model was presented a subset of data that had been held back during training. Replaying the model on past data in this way gives the best estimate of a forecaster’s future performance. However, when in the following sections we make statements like “the model anticipates eruption X” or “the model issued Y false positives”, these should all be read as strictly hypothetical situations.

Prospective testing of a forecasting model

Geological setting

Whakaari is a partially submerged andesite volcano located about 50 km offshore of New Zealand’s North Island (Fig. 1a). Eruption accounts by European explorers date back to 1769 (Luke, 1959), with the island experiencing a prolonged eruptive period between 1976 and 2000 (Clark and Cole, 1989; Kilgour et al., 2021). The latest phase of eruptive activity began in 2012 (Kilgour et al., 2021) with subsequent eruptions in 2013, 2016 and 2019 and extended periods of unrest through to the present (Fig. 1b–e). Whakaari is thus a good volcano for developing and refining forecasting methods because it has been well-studied from eruption forecasting (Chardot et al., 2015; Christophersen et al., 2018; Deligne et al., 2018; Dempsey et al., 2020) and volcano seismicity perspectives (Jolly et al., 2018; Yates et al., 2019; Park et al., 2020; Caudron et al., 2021; Ardid et al., 2022). Most importantly, it has a broadband seismic monitoring record covering seven phreatic eruptions across five distinct eruptive periods since 2010. Other forms of prior and current monitoring at Whakaari are reviewed in Kilgour et al. (2021).

In the case of a volcano in a semi-permanent state of unrest, defining an “eruption” requires thresholding of the size and severity of an event. We wish to concentrate on the hazard to human life in the proximity of the volcano. At Whakaari, several small-scale manifestations of steam plumes, geysering and changes in degassing rate may have led to observer reports of plumes or “minor ash venting”, especially if air humidity and visibility conditions were optimal. Here, we examined the reports of events during the study period (Chardot et al., 2015; Kilgour et al., 2019; Dempsey et al., 2020) and defined eruptions as events that were clearly impulsive explosions with hazardous ashfall, surge deposits, ash/steam plumes and/or SO{sub}2 emissions. This includes one event of known pure phreatic/hydrothermal origin in April 2016 (Kilgour et al., 2019), which generated ballistics and a pyroclastic surge. The five eruptions studied by Chardot et al. (2015) (August 2012, August 2013 and a sequence of three events in October 2013) and the fatal 2019 eruption are also included. Aside from the 2016 event, the origin of these eruptions remains unclear. Note: this classification excludes dome extrusion episodes (after the 2012...
Following Chardot et al. (2015), the August 2012 and August 2013 eruptions are considered separate events, although elevated unrest occurred throughout this period. Small impulsive steam venting events also occur occasionally such as a short sequence that occurred on 29 December 2020 (GeoNet, 2020a). In this case, the steam venting was considered to be in the normal range of fumarolic behaviour and was not described by monitoring scientists as an eruption and nor was VAL elevated from its previous level of one. Hence, these fumarolic-type steam venting episodes (often seen best with cold/clear atmospheric conditions) were not included as separate eruptions.

**Seismic input data**

For eruption forecasting, volcanic tremor is an excellent data stream for characterizing the current state of a volcano because it is information-rich, continuous and (depending on the network) reliable (e.g. Endo & Murray, 1991; McNutt, 1996; Chardot et al., 2015). Tremor is a continuous, long-period seismic signal between 1 and 15 Hz that frequently captures anomalous signals preceding eruptions (Konstantinou & Schlindwein, 2003). This is important for establishing the existence of precursors. Conceptual models for the generation of tremor include a range of rock-fluid interactions in the volcanic edifice: crack and conduit dynamics (e.g. Aki et al., 1977; Gottsmann et al., 2007; Ohminato, 2006) or fluid, bubble and gas dynamics (e.g. Ripepe & Gordeev, 1999; Caudron et al., 2019; Girona et al., 2019). These models can provide a window into subsurface dynamics of certain volcanoes. Complemented by other data sources, such as gas, geochemistry, deformation and visual observations (e.g. Baubron et al., 1991; Dzurisin, 2003; Rouwet et al., 2014), they assist volcanologists to translate observed signals into a conceptual understanding of processes that may or may not lead to an eruption. This task is complicated at volcanoes with active hydrothermal systems, where precursors are not as well characterized and where unrest does not always lead to eruptions.
At New Zealand volcanoes, station seismic velocity is band-pass filtered between 2 and 5 Hz and then averaged over 10-min periods (Chardot et al., 2015) to produce real-time seismic amplitude measure (RSAM; Endo & Murray, 1991). This minimizes the influence of ocean microseism, which is generally found at lower frequencies (Chardot et al., 2015; Caudron et al., 2021). At Whakaari, we have found other frequency bands (MF = medium frequency: 4.5–8 Hz, HF = high frequency: 8–16 Hz) and their ratios (displacement seismic amplitude ratio: DSAR = MF/HF; Caudron et al., 2019)) also provide useful information about the volcano state. In particular, week-long increases in DSAR are associated with all five Whakaari eruptions (Ardid et al., 2022) and are hypothesized to reflect sealing and pressurization of the shallow hydrothermal system.

This study uses data from the GeoNet-operated WIZ seismic station at Whakaari (Fig. 1a). WIZ was installed in 1976 and has operated with a broadband sensor since 2007 (Park et al., 2020). Kilgour et al. (2021) report that the station was upgraded in 2014, although this has not introduced an obvious bias into the distributional features of RSAM (see Online Resource 1). In this study, we have used data since the beginning of 2011. Data from a second station on the island, WSRZ, was considered but ultimately not used. This is because its recording period covers two fewer eruptions than WIZ, placing it at a disadvantage for model training.

Forecast model

In Dempsey et al. (2020), we described how time series feature engineering can be used to extract eruption precursors from tremor data recorded by WIZ (Fig. 1a). Systematic time series feature engineering (Christ et al., 2018) decomposes a continuous signal into a finite number of windowed features — statistical parameters (e.g. mean, maximum), harmonics, energy content and many more — and selects features that are statistically significant for distinguishing periods of interest. The number of features depends on the size of the window but is usually around 700. Some features are highly correlated, e.g. median and mean. Time series features have been used to characterize seismic signals during eruptions at Piton de la Fournaise volcano (Ren et al., 2020) as well as slip precursors in laboratory earthquake studies (Rouet Leduc et al., 2017).

One of the identified precursors at Whakaari was a 4-h burst of seismicity, which is extracted by the Fourier transform coefficient at the corresponding period (Fig. 2b). This and other precursors were used to train a random forest that was made up of a number of decision trees that have seen different subsets of data. Forecast models, \( f(\cdot) \), are structured to accept a 48-h block of seismic data as input, \( X \), and to return an eruption potential score, \( y \), between 0 and 1 (Fig. 2c). Mathematically, the forecast model is expressed as \( y = f(X) \), with larger output values implying an increased likelihood of eruption over the following 48 h. Although \( y \) is not the same as a probability of eruption, it can be converted to one using a suitable function, \( p = P(y) \).

For each of the five eruptive periods recorded at Whakaari since 2012, we trained a separate forecast model using only data from the other four eruptive periods. Data is labelled as “pre-eruption” if it occurs 48 h before the first eruption in a period. All other data is considered “not pre-eruptive” including the data immediately following an eruption. Training a model includes both the selection of time series features associated with the pre-eruption period and the calibration of a random forest classification model. Therefore, there is no guarantee that precursor features selected from the training set will be useful when forecasting the held-out eruption in the testing set. This is important for maintaining pseudo-prospective conditions.

Although several thousand features are calculated, far fewer are used for model selection. After discarding statistically insignificant features (adjusting for false discovery), up to several hundred features may be used to train a given model. However, because the random forest is comprised of many decision trees, some features end up being used within the model much more often than others. For example, in one model, the top 30 out of 380 features accounted for 50% of the total weighting within the random forest.

After the model is trained, a pseudo-prospective forecast is produced using data from the withheld eruption: this simulates prospective forecasting conditions. Finally, appropriate trigger thresholds for raising eruption warnings were assessed on a retrospective basis (Fig. 2c, d). An eruption warning continues until it drops below the trigger level for at least 48 h. An eruption was deemed to have been anticipated if it occurred during a pseudo-prospectively constructed warning period. For an optimized trigger, the Dempsey et al. (2020) model could anticipate five of the past seven eruptions at a cost of 12 false positives for each true one. Specific methodological details, verification tests and out-of-sample performance on all eruptions are given in Dempsey et al. (2020) and the Online Resource 1 (Text S1-S2).

In February 2020, we deployed the system in continuous, real-time operation. The goal was to evaluate whether it could anticipate future eruptions and how many false positives it would raise and to check its robustness to data- and model-side outages. The operational system comprised a Python framework that (1) downloaded new WIZ data from GeoNet, (2) evaluated the new data using the pre-trained model, (3) exposed the model output at a private web address and (4) sent automatic emails to a private account if the warning trigger was exceeded. The cloud-based framework executed steps 1–4 every 10 min. Since the end of 2021, the operation of the system has been impacted by the intermittency of WIZ, whose performance has been degraded because GeoNet has been unable to access the volcano for maintenance. For the
evaluation period in this study, the station down time was 1.1%. Total down time since 2010 was 0.7%.

**Prospective testing in 2020–2021**

Here, we describe the performance of the forecasting system between February 2020 and August 2021 (Fig. 3a). There were no eruptions in this period, and therefore, the discussion is limited to the three false positives that occurred.

The first false positive was triggered on 29 August 2020, midway through an earthquake swarm located 85 km north of Whakaari (Fig. 3b). The swarm comprised 18 events with $M_W \geq 4$ over a 5-day period with the largest an $M_W 5.3$. The swarm started 12 h after an accelerating tremor spike whose signature matched an episode of small steam explosions at Whakaari in December 2020 (GeoNet, 2020a). The cumulative impact of the swarm and tremor spike (Fig. 3c) triggered the false positive. A second false positive occurred in
September 2020, which was triggered by a low-level tremor that followed a signal spike on 12 September (Fig. 3d). The spike was from an $M_W$ 4.8 earthquake that was a continuation of the swarm to the north earlier that month (Fig. 3b).

The third false positive occurred in March 2021 in response to a triplet of large earthquakes on the Kermadec subduction zone (Fig. 3f). The first event ($M_W$ 7.3) east of New Zealand was followed hours later by two more, 900 km away ($M_W$ 7.3 and 8.1), near Raoul Island (Fig. 3e). Contamination of the tremor signal by shaking from these events and their aftershocks led to a false positive 48 h after the first quake. All three false positives would have been avoided if regional earthquake signatures had been removed in real time from the tremor signal.

All three false positives were “anticipated” or at least not unexpected when they did occur. This is because the earthquakes did not immediately trigger the model, but rather started it on an upward trend over several hours, toward the trigger level. This provided time to confirm (by waveform inspection and from GeoNet event relocations) that it was regional earthquakes that were impacting the model.

**Improvements to the forecasting model**

**Pre-processing to filter earthquakes**

Wave trains produced by regional earthquakes were found to confuse the Dempsey et al. (2020) model. They did this by temporarily elevating the velocity record at the seismic station above the background signal. When averaged over a 10-min interval, this produced outlier spikes in the RSAM record (Fig. 3c, f). Although human operators can be quick to diagnose and dismiss this noise, it can confuse algorithms with no contextual understanding. A possible remedy is to use the official GeoNet catalog to identify and remove earthquakes in pre-processing. However, we chose not to do this because this
would introduce an uncertain latency and dependency on external systems. Alternatively, multi-station networks could rapidly classify events as within or outside the edifice; however, this increases instrumentation requirements and the potential for outages. Additional stations could provide useful redundancy for forecasting systems, especially if one of the stations has an outage or is damaged during a prolonged eruption sequence.

Instead, we introduced a pre-processing step to filter out earthquake signals. This involved, first, identifying anomalous velocity transients using outlier detection and, second, constructing an envelope that wraps the waveform so it may be ignored during RSAM calculation. The key steps, integrated with the standard RSAM processing workflow, are (Fig. 4a, b):

1. Obtain 24 h of raw seismic velocity record, \( X \), with instrument response removed
2. Bandpass filter the data at the desired frequency range, usually 2 to 5 Hz, producing \( X_{\text{RSAM}} \)
3. Compute the absolute value of the signal, \( |X_{\text{RSAM}}| \)
4. Subdivide the signal into 10-min intervals, denoting the \( i^{th} \) as \( |X_{\text{RSAM}},i| \)
5. Apply a z-score transformation to the subinterval, \( z_i = (|X_{\text{RSAM}},i| - \mu_i) / \sigma_i \), where \( \mu_i \) and \( \sigma_i \) are the mean and standard deviation of \( |X_{\text{RSAM}},i| \)
6. Check an outlier criterion, \( \max(z_i) > z_{\text{th}} \), and return location of \( \max(z_i) \) if satisfied
7. If an outlier was detected, construct a 150-s window that brackets it, beginning 15 s before the outlier. This is intended to capture the earlier p-wave and prolonged s-wave of the assumed earthquake
8. RSAM for the 10-min interval is computed as the average of \( |X_{\text{RSAM}},i| \), excluding values in the 150-s window (if constructed)

We tested a range of values for \( z_{\text{th}} \) and found 3.2 was a good trade-off between suppressing earthquakes and not discarding too
much signal. However, this value might be different for other stations. Furthermore, because it only focuses on a single feature of a complex signal, outlier detection is a simplistic method for earthquake detection compared to recently proposed alternatives (Mousavi et al., 2020). It could also obscure some non-earthquake signals or earthquakes located in the edifice that are informative of the volcano state. Finally, the method is not suitable to remove more than one earthquake in close succession (Fig. 4c, d). Ultimately, such unintended side effects will be acceptable if the result is an overall improvement in model forecasting performance.

The above method removed most short-duration events from the signal. However, there was still contamination by longer earthquakes, notably the $M_L \geq 7.3$ triplet of subduction events that occurred in March 2021 (Fig. 5e). Therefore, a two-window (20 min) moving-minimum was computed to suppress these. In real-time operation when data need to be updated every 10 min, steps 1–8 are applied to the previous 24 h of data, but only the final 10-min block is retained.

Log-scale Z-score normalization of input data

Data normalization is a pre-processing step designed to balance the relative influence of training inputs that span disparate value ranges. For example, our model uses tremor computed in different frequency bands (RSAM: 2–5 Hz, MF: 4.5–8 Hz, HF: 8–16 Hz) and ratios (DSAR) that have different intensity distributions (Fig. 5b).

Other monitoring systems apply data normalization by station-volcano distance, e.g. reduced displacement (Aki & Koyanagi, 1981; Thompson & West, 2010). Here, we applied a z-score normalization in log-space (Fig. 5), which has been shown to have good performance on classification problems (Singh & Singh, 2020). In log-space, the distributions all have central tendency and not excessively large skew. Therefore, we applied a z-score transformation to the logarithm of the data, which standardizes it to a distribution with zero mean and unit standard deviation. Then, we transformed it back to real space:

$$\widetilde{X}_S = 10^{\log_{10}X_S} \cdot z(X_S) = \log_{10}X_S - \overline{\mu} \overline{\sigma}$$

where $X_S$ is a data stream with $S$ one of RSAM, MF, HF or DSAR, and $\overline{\mu}$ and $\overline{\sigma}$ are the mean and standard deviation of the log-transformed data, $\log_{10}X_S$, in the training set.

Rapid model retraining during eruptive sequence

During an eruptive period, several eruptions may occur in close succession. For instance, at Whakaari, three eruptions occurred over 8 days in October 2013. In pseudo-prospective testing, the first and third events are anticipated by an improved model (at a $y_c = 0.81$ trigger; Fig. 6) that includes the aforementioned earthquake filtering and data normalization. However, the model does not anticipate the middle eruption except at very low trigger levels.

The model’s performance is improved when it is immediately retrained after the first eruption in the sequence. By incorporating additional precursor information, model confidence of an imminent eruption is greatly boosted prior to the middle event. The time between first and second eruptions is 4.5 days, whereas retraining and redeploying an updated model takes less than an hour. Although more eruptive sequences are needed to determine whether such retraining consistently improves performance, this early indication is encouraging. However, adopting this adaptive learning approach in an operational setting would mean forgoing a restarted round of prospective testing, as recommended by Whitehead et al. (2022).

Estimating short-term eruption probabilities

There are several ways to calculate a probability of eruption depending on the time scale of interest and operational setting (Marzocchi & Bebbington, 2012; Whitehead & Bebbington, 2021). At long-lived volcanic centres with age-dating of past eruptions, long-term eruption probabilities can be estimated using a renewal model of the interevent time distribution (Wickman, 1966; Bebbington & Lai, 1996). A key step for these models is calibration, where parameters are tuned so there is a good match with prior data, ensuring that the output probabilities have reasonable magnitudes.

On shorter time scales, for instance, during an evolving volcanic crisis (Aspinall 2006) or at a volcano whose restlessness changes over months and years (Deligne et al., 2018), structured expert elicitation can be used to obtain eruption probabilities (e.g. Aspinall 2006; Selva et al. 2012; Christophersen et al. 2018). This method aggregates the views of multiple experts judging the available data, with mechanisms to solicit and distil uncertainty. In this case, the expert or expert group becomes “the model”, accepting monitoring data as inputs and outputting probabilities. Alternatively, Bayesian networks provide a meta-structure for integrating multiple models (physics-, data- and expert-driven) that inform probabilities through multiple event pathways (Aspinall and Woo 2014; Marzocchi et al. 2008; Wright et al. 2013; Christophersen et al. 2022).

Creating training sets for probability models

Data-driven forecasting models can also be improved by recasting the output as a probability of eruption. In the sections below, we detail two methods for calculating short-term probabilities, $p$: a binary warning system and a calibrated probability function. Both methods involve finding a suitable function $p = P(y)$ to transform the forecast model output,
Because there are two models to be trained ($f(\cdot)$ and $P(\cdot)$), to maintain pseudo-prospective conditions, we needed to use a nested train-test split of the data (Fig. 7). Furthermore, because we intend to test these models on the December 2019 eruption, only data recorded up to 1 month before the eruption is included in the training set. In the outer level, the data for eruptive period $j$ — the eruption itself plus 1 month either side — is withheld in the test set, $X_j$. The remaining four eruptive periods and all non-eruptive data are assigned to the training set, $X_{\sim j}$, where $\sim j$ denotes "not eruption $j$". A forecast model, $f_j(\cdot)$, is trained using $X_{\sim j}$, and this is used to generate in-sample, $y_0 = f_j(X_{\sim j})$, and out-of-sample forecasts, $y_j = f_j(X_j)$. The forecast model is a random forest comprising many decision trees. Every decision tree is trained on the same pre-eruption training data but a different, randomly selected subset of non-eruptive data. Downsampling the large amount of non-eruptive data is important for balanced training of a decision tree. Furthermore, any particular non-eruptive data

![Input data normalization. a Data time series prior and following the 2019 eruption (red dashed line): RSAM (black), MF (blue), HF (green), DSAR (magenta). b Histogram distributions of all data values in the 10.5-year study period. c Normalized data time series over the same time period as a. d Histogram distributions of data after applying z-score normalization.](image)

![Effect of model retraining on forecast performance during an eruptive sequence. Here, a pre-trained model (cyan) detects two out of three eruptions in a sequence. When the model is retrained after the first eruption using the recently acquired data, all subsequent eruptions are anticipated at high confidence.](image)
(a) Data windows and labels

1 2 3 pre-eruption
4 5

eruptions no eruption

(b) FIRST split into train and test sets

train, $X_{-j}$

test $X_j$

(c) Train forecast model, obtain prediction for unseen eruption AND training data.

1. training

2. out-of-sample prediction, $y_j$

forecast model, $f_j(\cdot)$

3. mixed-sample prediction, $y_0$

effective out-of-sample

in-sample

(d) SECOND split into train and test sets

test, $X_i$

train, $X_{-i,j}$

subset 1

subset 4

(e) For each subset in $D$, train a forecast model and obtain out-of-sample prediction, $y_i$, for unseen eruption, $i$.

1. training

2. input test data, $X_i$

forecast model, $f_{i,j}(\cdot)$

3. out-of-sample prediction, $y_i$

(f) Merge $y_i$ and $y_0$ to create $y_{i,agg}$

$y_0$

$y_i$

$y_{i,agg}$

(g) Fit probability model, $P_j(\cdot)$, and apply to $y_j$ from $C$

1. calibration

2. input prediction, $y_j$

probability model, $P_j(\cdot)$

3. output probability, $p_j$
window will have only been seen by a small fraction of the decision trees that make up the random forest. This means that the aggregate prediction of all trees (the model), $y_0$, on the non-eruptive data in $X_{-j}$, is effectively an out-of-sample forecast (Dempsey et al., 2020). This is only true for the non-eruptive data; pre-eruptive data is seen by all decision trees. This issue is addressed below.

The training set $X_{-j}$ is further subdivided by assigning eruptive period $i$ to the second test set $X_c$ with the remainder held in $X_{-c} = \{ i, j \}$ denotes “neither $i$ nor $j$”). The data $X_{-c}$ are used to train a forecast model $f(c)$ and an out-of-sample forecast obtained for the withheld eruption, $y_j = f_c(X_j)$. This eruption forecast is inserted into $y_0$, replacing the in-sample forecast obtained from $f(c)$. This is repeated for the other three eruptive periods in $X_{-j}$, yielding an aggregated out-of-sample forecast, $y_{agg}$.

The out-of-sample forecast, $y_{agg}$, is used to train a probability model, $P_j(\cdot)$, using one of the two methods described in the subsections above. The probability model can then be used to obtain an out-of-sample probabilistic forecast for the original unseen eruption, $p_j = P_j(y_j)$. Although this process is convoluted, it is necessary to guard against information leakage from the test eruption, $j$, into the training set.

**Constant eruption probability derived from eruption frequency**

A simple model for estimating eruption probabilities is to calculate the frequency of eruptions. For example, the training dataset that excludes the December 2019 eruption spans the period 1 January 2011 to 9 November 2019, a total of 3234 days or 1617 48-h periods. During this period, there were six eruptions in four eruptive periods. The probability that a randomly selected 48-h period contains an eruption is 0.0037 (0.37%). In the average model, this value is adopted as the estimated eruption probability for any 48-h period in the future. As this model does not make use of geophysical data, we refer to it as the “uninformed model”.

As the number of past eruptions is increased, there can be more certainty in the estimated eruption rate. If the six observed eruptions above were generated by a Poisson point process with an unknown rate parameter, the implied 95% confidence interval on that parameter (estimated from a $\chi^2$ distribution; Garwood (1936)) is rather large: [0.00077, 0.0089] ([0.077, 0.89]%).

The estimated eruption rate also depends on the definition of an eruption and the observation period within which eruptions are counted. For example, Christophersen et al. (2018) counted four eruptions impacting beyond the crater rim (thus excluding the first two events of the October 2013 sequence due to their smaller size) in the period 2001 to mid-2018. Thus, they arrive at a probability, $p_i = 0.13\%$, for an event in any given 48-h period (approximately 2% in 1 month).

In practice, the main value of the uninformed model is that it provides a benchmark against which to measure other models, in particular, how much value they extract from geophysical data. If they cannot perform better than a guess based on the past, then it is not clear that the model is extracting value from the sensor.

**Conditional probabilities of a warning system**

Here, we consider a basic warning system that triggers when the output of the improved forecast model, $y$, equals or exceeds a trigger threshold, $y_c$. Because our model is trained with 48 h of foresight, warnings are cancelled after a 48-h period in which $y < y_c$. For this system, the threshold $y_c$ is the only tuneable parameter and controls both the total duration of all warnings as well as the number of missed eruptions (those occurring outside a warning). For this analysis, we have used the aggregate out-of-sample forecast from January 2011 up to August 2021 that excludes the December 2019 eruption as we will be testing the method on this event. That is, the forecast model output is $y_{agg}$ where $j = 5$ denotes that the December 2019 eruption has been withheld from the training set.

Figure 8 illustrates the trade-offs of warning systems. For example, increasing $y_c$ above 0.88 causes more eruptive periods to be missed (falling outside a warning; Fig. 8a). Furthermore, the eruption in 2016 is missed for all except very low values of $y_c$, suggesting that little useful information about its precursors was extracted from the other eruptions. Jolly et al. (2018) analysed this eruption and identified very long period events (VLPs) up to 2 h before, which they linked to ascent of deep gas pulses. Arditi et al. (2022) have also shown that this eruption shares a common precursor with Whakaari eruptions in 2012, 2013 and 2019: a week-to-month long rise of 2-day smoothed DSAR that has been hypothesized to reflect sealing and pressurization of the
shallow hydrothermal system. However, as the present study only considers the past 2 days of data when calculating a forecast, this longer-term feature is not currently considered.

Although reducing $y_c$ increases the likelihood that eruptions are anticipated, it also increases the total duration of all issued warnings (Fig. 8b). For example, anticipating the eruption in 2016 would require a trigger threshold so low that warnings would be in effect over about 50% of the analysis period (Fig. 8b).

Graphing the missed eruption rate versus total warning duration for different selections of $y_c$ (Fig. 8c) produces a modified version of the receiver operating characteristic (ROC) curve, which is often used to assess a classifier’s performance. Selecting a preferred trigger is a balance between the rate of missed eruptions (false negatives) and false positives, with different users likely to make different selections based on their individual risk tolerance.

By comparing to the evaluation model, Fig. 8 highlights the incremental gains due to filtering regional earthquakes, normalizing input data and rapid model retraining after eruptions. The result is a leftward shift of the modified ROC curve (Fig. 8c) as well as a (retrospective) elimination of all false positives that occurred during the February 2020 to August 2021 evaluation period. Thus, this curve provides one measure of the performance or accuracy of a forecast model.

Based on the ROC curve of the improved model (Fig. 8c), a possible optimal choice to trigger a warning is $y_c = 0.88$. This results in 124 warning days in the nearly 9-year training period. Over the long term, the percentage of time spent under a warning is $124/3234 = 3.8\%$, which is about 14 days each year.

Out of the six eruptions covered by $y_{5_{agg}}$, five fall within warnings and only one falls outside (the 2016 eruption). This provides the basis for calculating a conditional probability

---

**Fig. 8** Performance of two probabilistic warning systems trained using data prior to December 2019: the Dempsey et al. (2020) evaluation model (dotted), and the improved model with earthquake filtering, data normalization and retraining (solid). **a** Rate of missed eruptions for different warning triggers, $y_c$. **b** Total duration of all warnings, out of a maximum possible 3234 days. Panels **a** and **b** share the same legend. **c** A modified receiver operating characteristic curve with inset showing a zoomed view for high values of $y_c$. **d** 48-h eruption probabilities inside (black) and outside warning (blue) for different triggers. Note: outside warning probabilities are 0 below ~0.2 because all eruptions fall within warnings at these trigger levels.
of eruption in a 48-h period given the current warning state. Denoting $E$ as an eruption, $W$ as a current warning and complementary events by overbars, then $P_{48hr}(E|\overline{W}) = 5/62 = 0.081 (8.1\%$, the quotient of 5 eruptions per 62 periods of 48 h each). In contrast, with one eruption occurring in the other 3109 days outside of warnings, we can compute a second, “low-likelihood” eruption probability, $P_{48hr}(E|\overline{W}) = 0.064\%$. Thus, there is a 126-times higher likelihood of an eruption inside a warning compared to outside.

The conditional probability model, $P(\cdot)$, is a piece-wise constant function that returns $P_{48hr}(E|\overline{W})$ if $y \geq y_c$ and $P_{48hr}(E|\overline{W})$ otherwise. Figure 8d shows how the two probabilities change for different choices of the warning trigger and two different models.

**Time-continuous probability**

An alternative probability model is one where eruption likelihood varies dynamically with the geophysical quantities that inform it. When combined with an assessment of eruption hazards, this model would support more nuanced risk calculations by allowing levels of acceptable risk to vary for different users or activities (e.g. expert vs. tourist access).

Because our forecast model is a random forest, we can transform its output to a probability using a suitable regression approach. Suppose the model outputs a value $y = 0.85$ (e.g. the righthand side of the blue curve in Fig. 9a): how likely is an eruption to follow in the next 48 h? The best guide is the past behaviour of the model. On the aggregated test output, $y_{5, \text{agg}}$, a value of 0.85 occurred 203 times (out of 465,357 total outputs every 10 min in the training set). Of those, 11 instances were followed 48 h later by an eruption. Therefore, the next time the model outputs a value 0.85, we might expect the probability of eruption in the next 48 h to be 5.4%. In this manner, we can quantify a proportion, $p$, for any value of model output, $y$.

We tested logistic and isotonic regression as models to formalize the probability calibration (Niculescu-Mizil & Caruana, 2005). Both approaches search for the best-fit function through the compiled proportions, $p$, and outputs, $y$ (Fig. 9a). Logistic regression fits a sigmoid function whereas isotonic regression selects a piece-wise constant function. Both functions are monotonically increasing, which ensures that the eruption probability never drops as $y$ increases, even if the proportions in $y_{5, \text{agg}}$ sometimes fluctuate down. For an isotonic model calibrated on $y_{5, \text{agg}}$, the probability of eruption is 4.1% for $y > 0.75$ and 80% for $y > 0.98$. The lowest value at $y = 0$ is $P_0 = 0.01\%$. As this characterizes 23% of the study period, we have used it as the “low-likelihood” reference for this model.

The probabilistic forecast $P(\cdot)$ is a point assessment of eruption likelihood over the next 48 h given data from the previous 48 h. However, under operational conditions, new data are received and a new probability is issued every 10 min. Should the forecast from 10 min ago be superseded and discarded? Would forecast users tolerate or trust a probability model that whiplashes back and forth as different levels of the isotonic regression are accessed? To address this issue, we introduced a time-averaged measure of probability that accounts for recent forecasts.

Suppose two experts are issuing eruption forecasts for the next 48 h. Expert A judges the available data and estimates a probability, $p_1$. Expert B waits 24 h and, with the benefit of additional data, estimates a probability, $p_2$. How should the two judgements be combined? Expert B has more recent function with width 350 samples, using a moving window of 288 samples. The time-averaged probability (black) has memory of prior forecasts.
data to inform their judgement, but Expert A still has 24 h left to run on their 48-h forecast. One way to combine the two probabilities is by assigning them different weights in a belief network, say, 0.33 and 0.67, respectively, reflecting the relative time remaining for each forecast. This is the approach we used when considering all still-valid forecasts issued by our model over a 48-h period. Figure 9b shows how the time-averaged probability responds with a degree of inertia to an increase of the instantaneous probability, but also retains some “memory” of past values.

**Expert elicitation**

Expert elicitation (Cooke et al., 1991) has previously been used to estimate eruption probabilities at Whakaari (Christophersen et al., 2018; Deligne et al., 2018; GeoNet, 2020b). For example, for the 4 weeks beginning 2 December 2019, the method was used to estimate an 8 to 14% probability (GeoNet, 2019b) of an eruption impacting the area beyond the vent (Fig. 1a). Note: this definition is similar to Christophersen et al. (2018) and therefore likely excludes events similar in magnitude to the first two events of the October 2013 eruption period. This forecast was constructed 1 week before the 9 December eruption. The range of values was the 50th to 84th percentile range of unweighted estimates from members of the volcano monitoring team.

Presenting the probability as a range recognizes that there are varied opinions within the expert group. Communicating high and low values is important if it helps to guard against any temptation to conservatively inflate eruption probabilities in matters of life safety. This is because conservatism can be incorporated in the high estimate while realism can be retained at the low end. To assist with the post-eruption disaster response in December 2019, an elicitation was undertaken every 24 h to estimate an updated probability (Online Resource, Table S1).

To compare the 48-h probabilities described in the previous two subsections with elicitation probabilities over different time periods, we needed to convert the latter to an equivalent 48-h probability. Formally, if an eruption will occur with probability, $p_x$, in the next $Y$ days, we seek the probability, $p_y$, that one will occur in the next $X$ days. Here, $X$ is an integer factor of $Y$, and $N = Y/X$ is the integer number of periods of length $X$ that $Y$ can be subdivided into, with $p_x$ constant in each of those subdivisions. The probability that the eruption does not occur in $Y$ days provides a basis to link $p_x$ and $p_y$, i.e.

$$\left(1 - p_y\right) = \left(1 - p_x\right)^{Y/X} \quad (2)$$

where the RHS is a multiplication of successive subdivision events in which the eruption did not occur. This expression is rewritten for $p_x$ as

$$p_x = 1 - \exp\left[\frac{\ln(1 - p_y)}{Y/X}\right] = 1 - (1 - p_y)^{Y/X} \quad (3)$$

For example, if the probability of an eruption is estimated to be 8% in 4 weeks, then calculation of the equivalent probability in a 48-h period is $p_y=0.08$ (8%), $Y/X=28/2=14$ and $p_x=0.0059$ (0.59%). The same method can be used to convert the average eruption probability (0.37% in 48 h) to reflect 24-h and 4-week windows used in different elicitations, respectively, 0.0019 and 0.051 (0.19% and 5.1%).

**Comparison of probability models applied to the December 2019 eruption**

On 9 December 2019, at 01:11 UTC (14:11 local time), an eruption occurred at Whakaari with 22 casualties. The eruption was the culmination of weeks of reported unrest (GeoNet, 2019a; Ardid et al., 2022), including elevated tremor, gas output (Burton et al., 2021) and vigorous activity through the crater lake (gas jetting and mud fountaining, GeoNet, 2019c). Approximately 16 h before the eruption, there was a strong tremor spike, which we refer to here as the 16-h precursor. Dempsey et al. (2020) have suggested this could have been an upward moving fluid pulse that later triggered the eruption. Following the 16-h precursor, RSAM exhibited an accelerating pattern characteristic of cascading material failure (Ardid et al., 2022; Chardot et al., 2015) until the main explosion occurred.

In the aftermath of the eruption, search-and-rescue operations were conducted on the island, supported by a science response (GeoNet, 2020b). Following the initial explosion, VAL was raised to level 3 to indicate an eruption in progress. Volcanic tremor decreased over the next day, before changing trend and rapidly climbing to very high levels for about 4 days. This was accompanied by ongoing steam and mud-jetting (GeoNet, 2019d) with the consensus view that magma was degassing at shallow depths. VAL was lowered to level 2 on 12 December (GeoNet, 2019d). On 22 January 2020, observations confirmed lava in the vent created by the eruption (GeoNet, 2020c).

Figure 10a shows how RSAM varied 4 days before and 8 days after the eruption. Also shown is the pseudo-prospective performance of the improved forecasting model trained using the other four eruptive periods on record. This model recognizes the 16-h precursor as a feature of concern and elevates its forecast accordingly. About 6 days after the eruption, and after 2 days of low RSAM, the forecast model drops back to a low value.

We first consider the performance of a warning system (Fig. 10b) that evaluates the output of the forecast model. We set the trigger value to $y_c=0.88$, identified as an optimal value in pseudo-prospective analysis (Fig. 8c). At this
threshold, a warning is issued about the time of the 16-h prior precursor and remains active through to the eruption. The associated within-warning eruption probability is 8.1% in 48 h, which corresponds to a relative likelihood 126 times higher than outside alarm (0.064%). The warning would be rescinded about 4 days after the eruption.

In the days before the 16-h precursor, the instantaneous probability model (Fig. 10c) judges a varying eruption likelihood between 2 and 4%. After the 16-h precursor, time-averaged probability tracks steadily upward and is 3% at the time of the eruption. Because the absolute probabilities are quite small, it could be more useful to quote the likelihood relative to the model background (0.01%). In this case, relative likelihoods prior to the 16-h precursor were 200 to 400 times higher than normal, and up to 300 times higher after the 16-h precursor.

Unlike the data-driven forecasting model, which only has access to seismic data at a single station, elicitation-derived probabilities can be informed by a broader range of monitoring data (e.g. gas, deformation, observations) and expert knowledge. Elicitation probabilities estimated on 2 December, prior to the 2019 eruption (Fig. 10d), ranged from 0.6 to 1.1% in 48 h. This is quite similar in size to the calibrated probability model at that time (~2.5%) as well as in the 4 days before the eruption (2 to 4%). Elicitation was more frequent following the initial explosion — every 24 h, beginning 48 h post-explosion — where it was used to estimate the probability of a subsequent explosive eruption impacting beyond the vent area. As a sequence of three closely spaced eruptions had previously occurred in October 2013, this scenario deserved genuine assessment. Elicited 24-h probabilities between 11 and 14 December were between 40 and 60%, corresponding to a period of extremely strong tremor and a consensus view that magma was degassing at shallow depths. Daily eruption probabilities were progressively lowered in the days that followed (Fig. 10e, and Online Resource, Table S1).

For comparison, pseudo-prospectively assessed post-explosion probabilities climbed as high as 4% in 48 h (approximately, 2% in 24 h), before dropping to the

---

**Fig. 10** Probabilistic forecasting of the December 2019 Whakaari eruption. a Tremor (black) and a time series of improved pseudo-prospective forecast model (cyan) prior to and following the eruption (red dashed). b Pseudo-prospective probability of eruption in 48 h calculated using a conditional probability model (Fig. 8) for a $y_c=0.88$ value to trigger a warning. c Pseudo-prospective probability of eruption in 48 h calculated using a calibrated probability model (Fig. 9). Blue = instantaneous, black = time-averaged probability. d Prospective probability of eruption in 24 h assessed by expert elicitation. Note, values prior to the December 2019 eruption have been converted from a 4-week forecast. e Comparison of the three probabilistic forecasts, where elicitation values have been converted to equivalent 48-h windows. The probability of eruption by an uninformed model is also shown (0.37% in 48 h, black dotted line).
background value 4 to 6 days after the eruption. This was a delayed response to (i) the 16-h precursor and (ii) certain seismic patterns associated with the eruption. We know this because the instantaneous probability drops sharply when those two events no longer fall within the 48-h backward-looking window used to construct the forecast (Fig. 10c). There is no indication that the model is responding to the 4 days of strong post-explosion tremor, with the assessed eruption probability trending downward during this period.

Discussion

Relative performance of different probability models

A common way to assess a forecast’s quality is to calculate its Brier Score (Brier et al., 1950), which is the average deviation between the issued probability and the binary observed outcome, $e_k$, over $n$ forecasts:

$$ BS = \frac{1}{n} \sum_{k=1}^{n} (p_k - e_k)^2, \quad (4) $$

where $p_k$ is the forecast probability that an event will occur, $e_k=1$ if the event happens and 0 otherwise. However, the Brier Score is not suitable for highly infrequent events like volcanic eruptions. For example, a probability model that always selects $p_k=0$ (no eruption will ever happen) will be right nearly all the time and achieve a high score. It is also functionally useless for forecasting.

Here, we use a version of the logarithmic score of Benedetti (2010), $FS$, expressed below:

$$ FS = \frac{1}{n} \sum_{k=1}^{n} [e_k \ln p_k + (1 - e_k) \ln(1 - p_k)] \quad \quad (5) $$

This has a maximum value of 0 for perfect forecast skill, with more negative scores indicating worse skill. The formulation penalizes trivial models like $p_k=0$ because these return a $\infty$ score if the forecast is wrong even once.

Our use of the forecast skill score here is restricted to a binary event class: eruption or no eruption. However, in other situations, it could be modified to score event trees that have multiple parallel outcomes (e.g. different hazard types; Wright et al., 2013) with attached probabilities.

For the uninformed model, $p_k=0.0037$ (0.37% likelihood of eruption in 48 h). If the goal is to compare the accuracy of uninformed and data-driven models, then $n=465,357$, which is the number of issued forecasts, one every 10 min in the training period January 2011 to November 2019. Finally, $e_k$ is a binary label with “1” denoting that an eruption will occur in the next 48 h and “0” otherwise. The calculated score for the uninformed model is $FS_0=-0.024$, which we used as a benchmark to assess other models:

$$ \widetilde{FS} = \frac{FS - FS_0}{FS_0} \quad \quad (6) $$

where $\widetilde{FS}$ is the scaled forecast skill. It has a maximum value of 1 for perfect prediction ($FS = 0$), a reference of 0 for models that perform as well as one uninformed by data ($FS = FS_0$) and negative values for models that perform worse than the uninformed. Forecasting scores for models considered here are reported in Table 1 and the Online Resource (Table S2).

For the conditional (warning system) and calibrated probability models, $p_k$ is the output of $P_y(\cdot)$ over the effective out-of-sample $y_{s, agg}$. The other parameters, $n$ and $\epsilon_k$, are the same as for the uninformed model. The corresponding scores are 0.12 and 0.21, respectively. This shows that both models are gaining an advantage from the data to improve their forecasting ability. Future refinements to the forecasting system can be benchmarked against this performance.

If restricted to the period around the December 2019 eruption (2 to 18 December, to coincide with elicitation), then $FS_0 = -0.70$, and scaled forecast skills for conditional and calibrated probabilities are $-0.02$ and $0.33$, respectively. Thus, forecast skill measured over the short term, e.g. a single eruptive period, can be higher or lower than the long-term score. For example, the performance of the warning system is about the same as the uninformed model during the December 2019 eruption. This is because the warning was late — triggered only 16 h before the eruption, instead of the 48-h target — and then extended a further 4 days post-explosion. In contrast, the calibrated probability model performed better than the uninformed model.

For the six other eruptions in the training set, the calibrated model outperformed the uninformed model for five of them ($\widetilde{FS}$ between 0.27 and 0.42) but was slightly worse for the 2016 eruption ($\widetilde{FS} = -0.13$) for which its poor performance has previously been noted. A full summary of the different forecast accuracies over these eruptions is given in Online Resource 1, Table S2.

Performance scores can also be computed for elicitation for the period 2 to 18 December 2019. A direct comparison with either the conditional or calibrated 48-h forecasts is not possible, because the elicitations were conducted for different windows: a 4-week forecast on 2 December and 24-h forecasts from 11 December. Nor would such a comparison be appropriate, given that the models developed here are pseudo-prospective and therefore have some (limited) benefit of hindsight. Elicitation relied solely on foresight.

Nevertheless, we can measure the performance of elicitation against an uninformed benchmark model, which also has no hindsight. The latter has equivalent eruption
Probabilities of 5.1% and 0.19%, respectively, for 4-week and 24-h windows. Scaled performance scores for elicitation in the post-explosion period are quite low, with $FS$ scores of $−1.0$ or $−1.5$ depending on whether the low or high estimates are used. This mostly reflects the high probabilities assigned to the outcome that a second explosive eruption would follow the 9 December event.

**Limitations and potential improvements of data-driven models**

Binary warning systems are simple to communicate but they have several drawbacks. First, they require preselection of a trigger threshold, which is a trade-off between sensitivity and accuracy (Fig. 8c). In choosing this trigger, different users would have to weigh the impacts of warnings (e.g. the economic effects of any access restrictions) against their own risk tolerance. Pseudo-prospective testing is helpful here because it provides the raw inputs to make this decision: relative rates of false positives and negatives. This also means that the same core forecasting model could feed different branching decision models depending on the relevant agency or activity, e.g. tourist access, essential maintenance of monitoring equipment, or search-and-rescue operations. Future work could investigate algorithms that optimally select the trigger threshold. However, ultimately, the decision should lie with emergency management authorities themselves where there is a political or societal mandate (Papale, 2017).

A second drawback of the warning system is the risk of overlooking or miscommunicating a nuanced probability calculation (Doyle et al., 2014; Harris, 2017). That is, the statement “a warning is in effect during which the probability of an eruption is 8.1%” may become abbreviated as “a warning is in effect”. In the latter statement, if the word “warning” is (mis)interpreted to mean a high likelihood of eruption, then a “false alarms” narrative may quickly develop that undermines the forecasting system. Indeed, in Dempsey et al. (2020), warnings are described as “alarms” but the terminology has been intentionally changed here to avoid the implication that eruption is the most likely outcome. In any event, discussion and consultation between forecast-makers and end-users (Fearnley and Beaven, 2018) will be important for ensuring that forecast information is relayed without any loss of legitimacy.

Without the concepts of “warning” and “trigger”, continuous probability models have fewer drawbacks, but may be harder to communicate. With a continuous model, in theory, it may learn to anticipate eruptions with very high confidence: 50, 90 or even 99%. However, at Whakaari, we found this style of model achieved eruption probabilities of, at most, a few percent in 48 h. The output is thus more readily communicated when presented in terms of relative eruption likelihood, which was up to several hundred times higher than the baseline. In part, these low absolute probabilities reflect noise and uncertainty inherent in the seismic data used to train the model. Future investigations may uncover different features, time windows or data transformations that are better at eliminating noise and thus improve model forecasting. The approach could also be extended to include multiple data types (e.g. seismic and gas; Manley et al., 2021).

A key shortcoming of the data-driven models in this study is their inability to anticipate the 2016 eruption. This event was not preceded by any changes in standard monitoring parameters, although a 2-m drop in lake level was noted the weeks prior (GeoNet, 2016). At the time of the eruption, VAL was at level 1. More recently, Ardid et al. (2022) have described a time series precursor to this and other eruptions at Whakaari. The precursor is a 4-week-long change in DSAR that potentially indicates elevated gas content prior to eruptions. This in turn can be linked to a conceptual model of a shallow seal in the hydrothermal system that causes pressurized conditions before an eruption. The discovery of this precursor suggests that future improvements of the data-driven forecast model could benefit from a mix of short-term (48 h) and longer (4 weeks or more) time series features. Ultimately, this failure is encouraging because it shows that pseudo-prospective testing is indeed removing the benefits of hindsight. This increases our confidence that conditions of foresight are being properly approximated and that models are not being overfit.
The data-driven models, trained on a different subset of eruptions, miss the 2016 eruption, which is a good example of failure to generalize to unseen data. However, similarly trained models were able to generalize to the other six eruptions in the dataset. This suggests it is a limitation of the (amount of) data rather than the modelling procedures themselves. We can therefore have some confidence that this forecast model will generalize well to future eruptions sharing similar characteristics to eruptions in 2012, 2013 and 2019. Having now also learned from the 2016 eruption, we might expect some ability to anticipate a similar eruption in the future, although this hypothesis cannot be tested under pseudo-prospective conditions.

At Whakaari, there are two instances of dome extrusion following eruptions in 2012 and 2019 (Jolly et al., 2020; GeoNet, 2021). If the precise timing of these events is determined, then a separate forecast model could be trained to identify when extrusion is occurring. If it is assumed that the 4 days of strong post-explosion tremor in December 2019 was associated with dome extrusion, then there is no evidence that a modelled trained on explosion precursors responds well to co-extrusion signatures. Nor should it be expected to be the case due to the scope of the current model.

A reasonable concern is whether there are other explosive eruption styles at Whakaari that are not yet documented in the short seismic record and thus are unknown to a trained model. For example, phreatomagmatic and Strombolian eruptions occurred at Whakaari in the 1970s, 1980s and 1990s (Kilgour et al., 2021), which pre-dates the continuous seismic record available for this study. Of course, recognizing unknown eruption precursors is a challenge for all forecasting systems. Methods that rely on expert input have tackled this problem through analogs: volcanoes whose unrest signals can be abstracted or generalized with the assistance of conceptual models (e.g. Wright et al., 2013). Such an approach could also be fruitful for data-driven models. Signals extracted from seismic data prior to eruptions at other volcanoes could be used to train a general forecasting model or to make up for a data shortfall in a model specialized to a particular volcano. Pseudo-prospective testing could be used to test how transferable these precursors are and whether they convey any benefit to forecasting models. Early steps in the direction of such transfer machine learning have been taken by Ardid et al. (2022) who demonstrated that the 4-week DSAR precursor at Whakaari was also present before eruptions at Ruapehu and Tongariro volcanoes, also in New Zealand.

Another application of transfer learning could be to extend the length of seismic time series at a volcano by pooling precursors identified across a range of permanent stations, temporary deployments and older stations that have been decommissioned or damaged in past eruptions. Finally, future work could consider the quantification and representation of further sources of uncertainty in the data-driven forecasts. By outputting a probability, all forecasts considered here are embedding an element of uncertainty associated with historic noisiness in the data. But elicitation approaches go beyond this to pool the assessments of many experts (Deligne et al., 2018) and thus construct a range of probabilities. By analogy, each decision tree in the random forests used here could be considered an “individual expert”, in which case variability amongst the “experts” can be quantified through bootstrap resampling. An example of this is given in the Online Resource Text S1. More robust approaches could include comparing separate forecast models that are developed by independent teams. This would allow the impact of certain structural decisions in the design and training of the model to be assessed.

**Strengths and weaknesses of elicitation**

When commenting on the performance of expert-driven models, it is important to distinguish the manner in which they are being used. For example, one common application is as a method of setting alert levels, that is, a decision to raise, lower or maintain a qualitative category of phenomena, hazard, risk or response. Whitehead and Bebbington (2021) suggested that the success rate of VAL systems containing a forecasting component, e.g. the Alaska Volcano Observatory’s (AVO) system, could be evaluated by counting the proportion of alert level increases that precede an eruption (Cameron et al., 2018). However, this approach is not strictly applicable to the New Zealand VAL system, which is not intended to communicate a forecast aspect.

The New Zealand VAL system is designed to inform stakeholders about current phenomena at a volcano. Within this remit, elevating Whakaari to VAL two 3 weeks prior to the eruption is the strongest “warning” that could be implied by the system. Furthermore, the Volcano Alert Bulletin (GeoNet, 2019a) that accompanied the VAL increase makes explicit mention that.

“The patterns of signals are similar to those through the 2011–2016 period [of elevated eruptive activity] and suggest that Whakaari/White Island may be entering a period where eruptive activity is more likely than normal…”.

A second kind of the expert-driven model is elicitation of probabilities, used to produce eruption likelihoods over well-defined forecast timeframes (Fig. 10d). Restricted to the period before the 2019 eruption at Whakaari, we note good agreement between elicitation probabilities (0.6 to 1.1% in 48 h, converted from 8 to 14% over 4 weeks) and those obtained by data-driven modelling (2 to 4%), keeping in mind that estimated probabilities in this study vary over three orders of magnitude. This suggests that, for this period, the elicitation method was yielding realistic probabilities that were consistent with the data.

An alternative way to validate elicitation is through a comparison with empirical observations (Colson and Cooke,
One way to quantify this comparison is through a forecast skill score (Eqs. 5 and 6). For the period 2 to 18 December 2019, elicitation yielded eruption probabilities that were objectively less accurate than the benchmark, which was a fixed probability model based on the rate of eruptions between 2011 and November 2019.

It is possible that elicited eruption probabilities at Whakaari are accurate in general but were inaccurate only in this specific instance. This was the case for the warning system model we developed, which outperformed the uninformed model over the long term ($\bar{FS} = 0.12$) but had worsened performance during the 2019 eruption ($\bar{FS} = -0.02$). The best way to evaluate this hypothesis is to calculate forecast scores for prior elicited probabilities at Whakaari over a longer period. As these probabilities are not routinely made public, we have not done that here. One outcome of this exercise could be to identify any systematic biases that exist with elicitation, which could then be corrected through a performance weighting approach (Colson & Cooke, 2018) or through formal probability calibration.

Assuming probability elicitation was only temporarily inaccurate following the 2019 eruption, it is worth examining the possible reasons for this. The first is that follow-on eruptions are genuinely probable events and the absence of one in this specific case was simply an unlikely outcome. At Whakaari, follow-on eruptions have been reported in one out of the last four eruptive periods (the October 2013 sequence; Fig. 6). For this sequence, the rate of follow-on eruptions was two events in 8 days. A Poisson model with rate parameter 0.25 events/day would thus assign a 22% daily probability of one or more eruptions. Based on these prior observations, a naïve event tree would assign a 5.5% daily probability for follow-on eruptions, well below the expert elicited values. Understanding eruption clustering dynamics (e.g. Bevilacqua et al. 2016) at Whakaari could provide further refinement and representation of post-eruption scenarios, insofar as the available records allow.

Another possible source of inaccuracy is that elicitation was applied in circumstances to which it was ill-suited. At Whakaari, forecast windows for prior elicitation were weeks to months long (Deligne et al., 2018) and were undertaken in the context of volcanic unrest. In contrast, post-explosion elicitation used a much-shortened 24-h forecasting window. During this period, there were very high levels of tremor. Scientists also noted ongoing steam and mud-jetting and elevated gas emission leading to a combined interpretation of magma degassing close to the surface (GeoNet, 2019e). All of this contributed to elicitation being undertaken in a less-familiar situation. Ordinarily, this should suggest a greater degree of uncertainty (wider probability bands). In the previous section, we described how data-driven models generalized poorly to the 2016 eruption because its seismic precursors were so different to other learning experiences. A similar issue may have affected this elicitation.

The third reason is that follow-on eruption probabilities may have been conservatively inflated, subconsciously or otherwise. This is understandable when those numbers are used to calculate the hazard exposure of disaster response and recovery teams who are putting their lives at risk. However, this conservativism would be better confined to the high estimate (upper bound) of eruption probability. This ensures that the low estimate can fulfil a critical role, which is to communicate the uncertainty of the expert panel and the limit of their foresight.

Retrospective evaluation of the post-eruption elicitation may seem an unfair use of hindsight. However, our analysis should not be mistaken as criticism of the decision to use elicitation before or after the 2019 Whakaari eruption. After all, it was the established method at Whakaari (Deligne et al., 2018). Indeed, our post hoc analysis shows that elicitation probabilities in the pre-eruption period are consistent with the data-driven approaches developed here. Rather, our intention has been to highlight that both elicitation and data-driven methods can be temporarily inaccurate while having acceptable performance over the long term. Such retrospective reflection is essential if shortcomings in either method are to be identified and corrected in future use. For elicitation, this could involve revision of how lower bounds are estimated or calibration against benchmark models. For data-driven approaches, it includes the discrimination and removal of noise (regional earthquakes), adaptive assimilation of new data (early events in an eruption sequence) and future work to estimate other sources of uncertainty.

**Generalization of forecasting methods to other volcanoes**

The data-driven methods we present for calculating probabilistic eruption forecasts could be extended to other volcanoes with different eruption styles or available data. The major requirements are (1) that those volcanoes have records of past eruptions and (2) the available data contain recurrent patterns prior to those eruptions that are differentiable from non-eruptive periods (Ardid et al., 2022). The appropriate amount and type of data required may need to be determined on a volcano-by-volcano basis but having a high-rate time series of some geophysical parameter (tremor, infrasound, geodesy, electromagnetic signal) is a clear advantage. Furthermore, confidence should be developed through a mix of pseudo-prospective and prospective testing.

Here, the warning system for calculating conditional eruption probabilities focused on comparing the forecast model output, $y$, to a trigger threshold, $y_c$. But we could have used a different forecast model (indeed, we tested two in Fig. 8) or, more simply, a processed geophysical proxy of volcano state like RSAM. Different trigger values will be required depending on the choice of model or geophysical proxy. However,
the selection of this trigger will continue to create a trade-off of different performance characteristics (Fig. 8c).

Similarly, for the probability calibration method, the main input is a continuous indicator of the volcano state. Here, we used our forecast model output, \( y \), but this could also have been the output of a competing model, a single feature time series, or even the RSAM time series. The important step is construction of the calibrating function \( \hat{P}(\cdot) \) (Fig. 9a), which compares the chosen indicator to a set of binary labels that highlight eruption imminence.

A key aspect of both methods is the attempt to minimize overfitting by emphasizing out-of-sample training of the probability model. With increasingly complex models, it is tempting to overly customize to the sparse eruption data. In these cases, cross-validation and construction of out-of-sample training sets become important steps. Simple geophysical proxies like RSAM are likely less susceptible to customization because of community-held standards or because they need to generalize across many volcanoes. For the warning system, a further check is whether optimally determined trigger thresholds on an eruption subset (Fig. 8b) perform well for a hold-out event (Fig. 10b).

With suitable scoring metrics, forecasting skill can be quantified and compared across different methods. The logarithmic forecast skill score (Eqs. 5 and 6) can be applied to any model that outputs a binary prediction probability, and for which there exist corroborating observations. Finally, assessing model performance relative to simple benchmarks can help identify the value extracted from geophysical data or highlight periods of underperformance for further investigation.

**Challenges reconciling short-term forecasting with Volcano Alert Levels**

A potential use for data-driven forecast models is to augment existing volcano monitoring practices and communication frameworks, particularly Volcanic Alert Level systems. VAL systems differ in terms of the information they provide (volcanic phenomena, hazard, risk or disaster response) and the foresight they imply (current or future). Our consideration is restricted to the current New Zealand VAL system, whose remit is to communicate a current appraisal of volcanic phenomena (Potter et al., 2014). We also consider associated monitoring and communication processes including expert elicitation of eruption probability and narrative text in the form of Volcanic Activity Bulletins (VAB).

The frequency that the VAL is reassessed at a given volcano is set by the following rules: (i) at a minimum each week, or (ii) whenever a member of the monitoring group calls an ad hoc meeting, or (iii) under urgency by the Volcano Duty Officer and/or Volcano Science Advisor when an eruption is suspected or confirmed (G. Jolly, pers. comm., 2022). The second rule readily accommodates a real-time forecasting system because any unusual or rapid departure of probability from the baseline could justify an ad hoc meeting and for further assessment. Possible challenges with this approach include whether ad hoc meetings can be convened and concluded quickly enough to respond to rapidly changing precursors and situations and whether there is a maximum tolerance by scientists for the frequency and number of ad hoc meetings.

VAL is formulated as a discrete ordinal scale. By design, this limits the granularity of information that can be communicated. For example, suppose VAL is at level 2 (reflecting moderate volcanic unrest) and a forecasting system identifies a precursor of concern and responds by increasing the eruption probability. The VAL cannot be raised to level 3 because this can only happen once an eruption begins. Nor is there a level 2.5 or 2.9 in the present system. In these cases, any heightened level of concern could be communicated through the narrative text of a VAB. These provide an opportunity to detail the nature of concerns, e.g. increased likelihood of eruption, and their observational basis, e.g. a description of the precursor and inferred volcanic processes.

In principle, VAL does not communicate the likelihood of future eruption. Arguably then, the output of a data-driven forecasting model should have no bearing on the decision of scientists to maintain or change alert level. In practice, Christoffersen et al. (2022) note that, although VAL is not directly determined this way, its changes often reflect the monitoring group’s judgement of eruption likelihood. In this case, real-time forecasting systems can contribute by adding to the information used by monitoring scientists to form their VAL assessment. However, if it is desirable that VAL remains impartial to eruption likelihood, then eruption forecast models might better be withheld from monitoring scientists until VAL is determined. Afterward, forecasts could be made available to assist the expert elicitation of eruption probability.

**Conclusion**

We here evaluate the prospective operation of a data-driven forecasting system for explosive phreatic eruptions at Whakaari volcano deployed between February 2020 and August 2021. No eruptions occurred during this period, although three false positives were triggered, all in response to regional earthquakes (Fig. 3). Based on this experience, we implement a pre-processing filter to remove earthquakes from the tremor calculation (Fig. 4). We also apply data normalization (Fig. 5) to rebalance contributions from different frequency bands, and rapid model retraining (Fig. 6) in the case of sequences of multiple eruptions. These changes improve the pseudo-prospective performance of the model between 2011 and 2021 (Fig. 8) by reducing false positives and identifying eruptions with higher confidence.
We introduce two methods for calculating short-term eruption probabilities and compare these to elicitation-derived probabilities around the December 2019 eruption. The first method is an optimized warning system that anticipated six out of seven eruptions under pseudo-prospective conditions (Figs. 8c and 10b). The probability of eruption when a warning was in effect was 8.1% in 48 h and 0.064% when it was not. On average, there were 14 warning days per year. A drawback of such systems is that they require someone to choose an appropriate trigger level, and this may not capture the diverse risk tolerances of all stakeholders. On the other hand, warning systems can set clear expectations on the frequency of false positives and allow communication protocols to be agreed ahead of a volcanic crisis.

The second method uses a calibrating function to convert the forecast model output into a continuously varying eruption probability. The result is a 48-h likelihood of eruption that ranges between a minimum of 0.01% and a maximum of 80% (Fig. 9a). This model indicated eruption probabilities (per 48 h) up to 4% before the December 2019 eruption at Whakaari (Fig. 10c), which is 400 times higher than its normal background level. This model also has the best forecasting skill of all methods tested (Table 1, Online Resource 1, Table S2). However, absolute probabilities are quite low, which suggests persistent challenges with noisy data. Furthermore, representations of structural uncertainty in the model have yet to be developed.

Finally, we compare the two data-driven methods for estimating eruption probabilities with values derived from published expert elicitation produced by volcano monitoring scientists before and after the 2019 Whakaari eruption (GeoNet, 2020b). Elicited probabilities prior to the eruption (up to 1.1% per 48 h) were in general agreement with the calibration method when considering that values from all methods could vary over three orders of magnitude. It is also notable that a separate elicitation 3 weeks before the eruption raised the volcano alert level and highlighted the possibility that the likelihood of an eruption may have increased. These attest to good performance of elicitation in the pre-eruption period. However, we show that probability elicitation in the post-explosion period had worse performance than a static model uninformed by data. We propose that retrospective analysis of probability elicitation over longer periods could be a fruitful approach for eliminating any systematic biases.

There are several opportunities for future studies to improve the accuracy or extend the application of data-driven forecasting models. First, extending precursor analysis to include longer time windows (Ardid et al., 2022), different time series features or other types of data (Manley et al., 2021) is one way to broaden the library of possible eruption precursors. These are the key inputs for design, testing and implementation of forecast models. A broader base of precursors would allow forecast models to respond to, say, seismic and gas data simultaneously (multi-data); or, to reflect on the previous 48 h of activity in the context of a longer 4-week assessment (multi-time scale); or, by closely matching data to past eruption templates. Second, invoking the concept of analogs, it could be useful to understand how precursors can be abstracted and generalized to other volcanoes (Ardid et al., 2022). This is one way to address issues of data availability, for instance, if a particular volcano has a short monitoring record and is thus deficient in either total eruptions or a particular eruption style. This could help to quantify key uncertainties, such as the possibility of unobserved eruption phenomena, while also extending applicability to volcanoes with newer monitoring networks.

Tourist fatalities at active volcanoes like Whakaari are tragic but, sadly, not unprecedented (cf. Etna in 1987; Yasur in 1995; Kilauea in 2000; Stromboli in 2001 & 2019; Mayon in 2013; Ontake in 2014: Brown et al., 2017). Improving our ability to forewarn of volcanic hazards is one way we can continue to reduce the risk of death or injury. In this spirit, we hope that our study provides the scientific community with some useful advances in the quantification and evaluation of eruption forecasts at hydrothermally active volcanoes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00445-022-01600-5.

Acknowledgements We are grateful to the Centre for e-Research at the University of Auckland, and for the use of RCC facilities at the University of Canterbury, for operational running of our forecasting system. Martin Feller and François Bissey are specifically acknowledged for technical support. All data and codes necessary to reproduce this research are available at https://github.com/ddempsey/whakaari/tree/real-time. We acknowledge the New Zealand GeoNet project and its sponsors the Earthquake Commission (EQC), the Institute of Geological and Nuclear Sciences (GNS Science), Land Information NZ (LINZ), the National Emergency Management Agency (NEMA) and the Ministry of Business, Innovation & Employment (MBIE) for providing data used in this study. We thank G Jolly for comments that improved our understanding of GNS monitoring protocols. Finally, we are grateful to two anonymous reviewers and editors Hannah Dietterich and Andrew Harris whose detailed comments improved this manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions The authors acknowledge the financial support for this research by the New Zealand government MBIE Endeavour research grants UOAX11807 “Experimental approach for reducing eruption risks in New Zealand Geothermal Fields” and UOA1913 “Transitioning Taranaki to a Volcanic Future”.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will
References

Aiuppa A, Moretti R, Federico C, Giudice G, Gurrieri S, Liuzzo M, Papale P, Shinozawa H, Valenza M (2007) Forecasting Etna eruptions by real-time observation of volcanic gas composition. Geology 35(12):1115–1118

Aki K, Fehler M, Das S (1977) Source mechanism of volcanic tremor: fluid-driven crack models and their application to the 1963 Kilauea eruption. J Volcanol Geoth Res 2(3):259–287

Aki K, Koyanagi R (1981) Deep volcanic tremor and magma ascent mechanism under Kilauea Hawaii. J Geophys Res Solid Earth 86(B8):7095–7109

Ardid A, Dempsey D, Caudron C, Cronin S (2022) Seismic precursors to the Whakaari 2019 phreatic eruption are transferable to other eruptions and volcanoes. Nat Commun 13(1):1–9

Aspinall WP (2006) Structured elicitation of expert judgment for probabilistic hazard and risk assessment in volcanic eruptions. SIV 1:15–30

Aspinall W, Woo G (2014) Santorini unrest 2011–2012: an immediate Bayesian belief network analysis of eruption scenario probabilities for urgent decision support under uncertainty. J Appl Volcanol 3(1):1–12

Auker MR, Sparks RSJ, Siebert L, Crosswell HS, Ewert J (2013) A statistical analysis of the global historical volcanic fatalities record. J Appl Volcanol 2(1):1–24

Baumbroon JC, Allard P, Sabroux JC, Tedesco D, Toutain JP (1991) Soil gas emissions as precursory indicators of volcanic eruptions. J Geol Soc 148(3):571–576

Baxter PJ, Gresham A (1997) Deaths and injuries in the eruption of Galeras Volcano, Colombia, 14 January 1993. J Volcanol Geoth Res 77(1–4):325–338

Bevington M, Lai C (1996) On non-homogeneous models for volcanic eruptions. Math Geol 28(5):585–600

Bell AF, Naylor M, Hernandez S, Main IG, Gaunt HE, Mothes P, Bebbington M, Lai C (1996) On nonhomogeneous models for volcanic eruptions and volcanic hazards. Nat Commun 11(1):1–9

Biggs J, Ebmeier S, Aspinall W, Lu Z, Pritchard M, Sparks R, Mather T (2014) Global link between deformation and volcanic eruption quantified by satellite imagery. Nat Commun 5(1):1–7

Brier GW et al (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78(1):1–3

Brown SK, Jenkins SF, Sparks RSJ, Odbert H, Auker MR (2017) Volcanic fatalities database: analysis of volcanic threat with distance and victim classification. J Appl Volcanol 6(1):1–20

Burton 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

Cameron CE, Prejean SG, Coombs ML, Wallace KL, Power JA, Roman DC (2018) Alaska volcano observatory alert and forecasting timelines: 1989–2017. Front Earth Sci 6:86

Carniel R, Guzmán SR (2020) Machine learning in volcanology: a review. Updates in Volcanology-Transdisciplinary Nature of Volcano Science

Caudron C, Girona T, Jolly A, Christenson B, Savage MK, Carmiel R, Lecocq T, Kennedy B, Lokmer I, Yates A, Hamling I, Park I, Kilgour G, Mazot A (2021) A quest for unrest in multiparameter observations at Whakaari/White Island volcano, New Zealand 2007–2018. EPS 73(1):1–21

Caudron C, Girona T, Taisne B, Gunawan H et al (2019) Change in seismic attenuation as a long-term precursor of gas-driven eruptions. Geology 47(7):632–636

Chastin SF, Main IG (2003) Statistical analysis of daily seismic event rate as a precursor to volcanic eruptions. Geophys Res Lett 30(13)

Christ M, Braun N, Neuffer J, Kempa-Liehr AW (2018) Time series feature extraction on basis of scalable hypothesis tests (tsfresh—a Python package). Neurocomputing 307:72–77

Christoffersen A, Deligne NI, Hanea AM, Chardot L, Fournier N, Aspinall WP (2018) Bayesian modeling and expert elicitation for probabilistic eruption forecasting: Pilot study for Whakaari/White Island New Zealand. Front Earth Sci 6:211

Christoffersen A, Behr Y, and Miller C (2022) Automated eruption forecasting at frequently active volcanoes using Bayesian networks learned from monitoring data and expert elicitation: application to Mt Ruapehu, Aotearoa, New Zealand. Front Earth Sci 10:905965. https://doi.org/10.3389/feart.2022.905965

Clark RH, and Cole JW (1989) Volcanic monitoring and surveillance at White Island before the 1976–82 eruption sequence. NZ Geological Survey Bulletin 103. Eds: Houghton BF, and Nairn IA

Colson AR, Cooke RM (2018) Expert elicitation: using the classical model to validate experts’ judgments. Rev Environ Econ Policy 12(1):113–132

Cooke R (1991) Experts in uncertainty: opinion and subjective probability in science. Oxford University Press, Oxford, UK

Crampon S, Gao Y, Bukits J (2015) A review of retrospective stress forecasts of earthquakes and eruptions. Phys Earth Planet Inter 245:76–87

de Moor JM, Aiuppa A, Pacheco J, Avard G, Kern C, Liuzzo M, Fischer TP (2016) Short-period volcanic gas precursors to phreatic eruptions: insights from Poás Volcano, Costa Rica. Earth Planet Sci Lett 442:218–227

De la Cruz-Reyna S, Reyes-Dávila GA (2001) A model to describe civil protection preparedness and responses to the ongoing volcanic crisis at Popocatépetl volcano (Whakaari), New Zealand. J Volcanol Geoth Res 2007–2018. EPS 73(1):1–21

De la Cruz-Reyna S, Tilling RI (2008) Scientific and public responses to the ongoing volcanic crisis at Popocatépetl Volcano, Mexico: importance of an effective hazards-warning system. J Volcanol Geoth Res 170(1–2):121–134

Deligne NI, Jolly GE, Taig T, Webb TH (2018) Evaluating life-safety risk for fieldwork on active volcanoes: the volcano life risk estimator (VoLReSt), a volcano observatory’s decision-support tool. J Appl Volcanol 7(1):1–19

Dempsey D, Cronin SJ, Mei S, Kempa-Liehr AW (2020) Automatic precursor recognition and real-time forecasting of sudden explosive volcanic eruptions at Whakaari New Zealand. Nat Commun 11(1):1–8
