Katsumata and Nakatani
Earth, Planets and Space (2021) 73:100
https://doi.org/10.1186/s40623-021-01418-z

Earth, Planets and Space

FULL PAPER

Open Access

Testing the seismic quiescence hypothesis through retrospective trials of alarm-based earthquake prediction in the Kurile–Japan subduction zone

Kei Katsumata[1]* and Masao Nakatani[2]

Abstract

We make trial binary forecasts for the Kurile–Japan subduction zone for the period 1988–2014 by hypothesizing that seismic quiescence (i.e., the absence of earthquakes of $M \geq 5$ for a minimum period of T_q) is a precursor of a large ($7.5 \leq M_w < 8.5$) earthquake in the coming period T_a within a radius R of the quiescence. We evaluate the receiver-operating-characteristic diagram constructed using a range of forecast models specified by (T_q, R, T_a). A forecast experiment targeting eight large earthquakes in the studied spacetime suggests that the risk of a large earthquake is modestly (probability gain $G \sim 2$) but significantly (p-value less than 5%) heightened for several years following a long quiescent period of $T_q \geq 9$ years, within several tens of kilometers of the quiescence. We then attempt cross-validation, where we use half the data for training [i.e., optimization of (T_q, R, T_a)] and the remaining half for evaluation. With only four target earthquakes available for evaluation of the forecasts in each of the learning and evaluation periods, our forecast scheme did not pass the cross-validation test (with a criterion that the p-value is less than 5%). Hence, we cannot formally deny the possibility that our positive results for the overall period are a ghost arising from over-fitting. However, through detailed comparison of optimal models in the overall test with those in the cross-validation tests, we argue that severe over-fitting is unlikely involved for the modest G of ~ 2 obtained in the overall test. There is thus a reasonable chance that the presently tested type of quiescence will pass the cross-validation test when more target earthquakes become available in the near future. In the meantime, we find that G improves to ~ 5 when target earthquakes are limited to $8 \leq M_w < 8.5$, though we cannot say anything about the possible involvement of over-fitting because we have only three such very large target earthquakes.

Keywords: Earthquake prediction, Seismic quiescence, Cross-validation, Kurile–Japan subduction zone

Introduction

Large earthquakes are often preceded by a period of *seismic quiescence*; that is, a remarkable drop in regional background seismicity lasting for several years. Drawing attention as a possible intermediate-term earthquake precursor (e.g., Scholz 2019), quiescence has a long history of research. Whereas earlier studies (e.g., Inouye 1965; Utsu 1968; Mogi 1969; Ohtake et al. 1977; Kanamori 1981) lacked an objective definition of quiescence, later studies have developed a variety of measures quantifying quiescence, including the ZMAP method (Wiemer and Wyss 1994; Katsumata 2011, 2017a), ETAS modeling (Ogata 1992), and RTL/RTM method (e.g., Sobolev and Tyupkin 1997; Nagao et al. 2011). Thus, seismic quiescence has been established as an objectively demonstrable anomalous incident, which

*Correspondence: kkatsu@sci.hokudai.ac.jp
[1] Institute of Seismology and Volcanology, Faculty of Science, Hokkaido University, Sapporo, Japan
Full list of author information is available at the end of the article

often precedes large (typically $M_w \geq 7.5$) earthquakes by several years to decades.

Finding more and more earthquakes preceded by quiescence, however, does not answer the real question of interest; does a statistically significant tendency exist for quiescence to precede large earthquakes? Specifically, we must ask if the observed frequency of the coincidence of the two occurrences (quiescence and a subsequent large earthquake) is beyond the level explainable as a mere matter of chance. To our knowledge, this point has not been tested, except for an attempt by Katsumata (2017a), who, unfortunately, failed to set an objective criterion for the coincidence as detailed later. Thus, while long having received research attention, quiescence has remained a well-known precursor *candidate* for half a century.

In the meantime, an *earthquake-preceding tendency* (e.g., Nakatani 2020) has been proven to exist for a special class of quiescence called 'relative quiescence' (Ogata 2001), a phenomenon that the aftershock seismicity following a large earthquake starts depleting at one point of time during the aftershock period, compared with extrapolation following Omori's law. However, there is no reason to presume that the affirmative conclusion about the precursory relative quiescence applies to the (general) quiescence seen in the non-aftershock period.

To judge if a certain phenomenon has an earthquake-preceding tendency, we need to scan the entire spacetime of study, not only the times preceding large earthquakes, for the same type of phenomenon. We refer to the phenomena as '*anomalies*' following the convention adopted in earthquake precursor research, but 'anomalies' in this context can be any incidents that are objectively definable; they do not have to be rare or anomalous in a statistical sense.

On the basis of the above concept, Katsumata (2017a) exhaustively scanned background seismicity in subduction zones around Japan in the period 1975–2012, for long-term quiescence anomalies (lasting longer than 9 years). He compared the detected anomalies with four earthquakes having $M_w \geq 8.25$ that occurred in the studied spacetime, using a contingency table. A Fisher's exact test found a p-value (i.e., the probability that the observed or higher extent of correlation emerges by chance under the null hypothesis of no correlation between the anomalies and earthquakes) of 0.021, suggesting that the quiescence has an earthquake-preceding tendency. Conceptually, this is a correct method of evaluation. However, he did not set a consistent, objective criterion with which to judge whether a large earthquake followed a quiescence anomaly, leaving room for doubt. More importantly, it seems that Katsumata (2017a) hesitated to set an objective criterion for the anomaly–earthquake association to get around a technical difficulty inherent to a contingency table as explained below.

In evaluating a contingency table for an earthquake-preceding tendency, the number of anomalies or alerts incurred by them needs to be counted. When multiple incidents of anomalies occur close to each other in space and time, incurred alerts largely overlap. In such cases, one physical incident of a correct (or incorrect) anomaly or alert is evaluated as many successful (or unsuccessful) incidents, while it should be ideally counted as just one incident of success (or failure) in the contingency table. Thus, in practice, number counting is not a good way to quantify anomalies or alerts. Probably as a subconscious workaround for this problem, Katsumata (2017a) subjectively combined clustered anomalies into one incident of quiescence for the contingency table. The evaluation conducted by Katsumata (2017a) thus lacked objectivity, even though his subjective grouping is probably reasonable from a physical point of view.

The above difficulties with the contingency table can be avoided by instead using an *alarm map*, with each spatiotemporal cell being given a binary value of either *alarm-on* or *alarm-off* according to any objective rule. A simple example may be alerting a region within a distance R from an anomaly for a time duration T_a following the anomaly. The total alerted volume, instead of the number count of alarms or anomalies, can be used for alarm quantification. In this way, largely overlapped alerts from clustered anomalies do not cause disproportionate weighting of their success or failure.

Under the null hypothesis of no correlation between the anomalies and earthquakes, expectancy for the prediction rate (i.e., the ratio of the number of alerted earthquakes to the total number of target earthquakes) is equal to the alarm fraction (i.e., the ratio of the alerted volume to the total spacetime of the study). On this basis, the statistical significance of the earthquake-preceding tendency and its strength can be evaluated using the p-value and the probability gain, respectively (e.g., Zechar and Jordan 2008; Nakatani 2020).

Note that the above statistical testing cannot tell if individual anomalies that preceded the target class of earthquakes were indeed related to the earthquake. Instead, the 'existence of an earthquake-preceding tendency' means that one or more of the anomalies followed by earthquakes were indeed related to the subsequent earthquake occurrence.

In the present paper, we evaluate the earthquake-preceding tendency of long-term (~ 10-year) quiescence in the Kurile–Japan subduction zone (Katsumata 2017a) adopting the alarm-map-based testing described above.

As the primary goal of the present paper is to demonstrate rigorous yet generic statistical procedures of alarm-map-based evaluation, we only consider quiescence to avoid distraction. However, we recognize that various, sometimes even opposite, senses of seismicity changes [e.g., quiescence and activation (Reasenberg and Matthews 1988) and acceleration and deceleration (Hardebeck et al. 2008)] might precede earthquakes. We emphasize that there is no logical or practical difficulty in producing and evaluating alarm maps based on such mixed behaviors as long as the alerting procedure is stated objectively; e.g., the M8 algorithm (Keilis-Borok and Kossobokov 1990).

The statistical power depends strongly on the number of target-class earthquakes available, and we thus make trial forecasts targeting earthquakes of $M_w \geq 7.5$ instead of $M_w \geq 8.25$ for which the precursory long-term quiescence has already been suggested (Katsumata 2017a). However, there are still only nine earthquakes of $M_w \geq 7.5$, and this small number of target earthquakes is certainly a weak point of the present study. Although more target earthquakes would be available if we add more study regions, we decided not to add other tectonic regions as our priority is to investigate the effects of retrospective optimization (e.g., Mulargia 1997). For that purpose, we examine a suite of forecast models, where we vary the model's three main adjustable parameters (i.e., the anomaly detection threshold and the temporal and spatial limits within which anomalies are associated with subsequent earthquakes) over a wide range. We also examine the likelihood of over-fitting by conducting additional cross-validation experiments. We therefore take a minimalistic approach in other regards.

Data

The present study area comprises subduction zones along the western margin of the Pacific Plate, spanning 25°–55° N and 138°–163° E and including the east coast of Kamchatka, the Kurile Islands, and the Izu-Bonin Islands. Forecast targets are shallow (having a centroid depth less than 70 km) earthquakes of $M_w \geq 7.5$ listed in the Global Centroid Moment Tensor catalog. The term 'centroid' means the gravity center of the seismic moment released by an earthquake, while the term 'hypocenter' means the location where the main shock rupture started. For large earthquakes, the centroid and hypocenter usually do not coincide. Nine earthquakes (Fig. 1 and Table 1) satisfy these conditions in the period from January 1, 1988 to December 31, 2014. These nine earthquakes are clearly main shocks, sufficiently isolated on a space–time plot of seismicity. Evaluation of a forecast is not straightforward if the forecast targets include aftershocks (Michael 1997).

We exclude EQ#9, the 2011 Tohoku Earthquake (M_w 9.1), from our forecast targets because it was preceded by super-long (i.e., super-strong) quiescence that we could not detect owing to a mere practical limitation. As detailed in "Methods" section, we detect changes in seismicity by examining 15-year sub-catalogs of the background seismicity, such that there is no way of detecting quiescence lasting 15 years or more. If we use sufficiently long sub-catalogs, we can correctly recognize the exceptionally long quiescence exceeding 20 years that preceded EQ#9 (Katsumata 2017b). Therefore, counting EQ#9 as a false negative would not be scientifically adequate. We therefore decided to exclude EQ#9. If one still finds it unfair to ignore EQ#9, whose precursor is missed by our presently adopted algorithm, we can modify our forecast target to $7.5 \leq M_w < 8.5$, instead of $M_w \geq 7.5$.

In addition, the 1993 Hokkaido Nansei-oki earthquake (M_w7.7, 42.71° N, 139.28° E, 16.5 km deep) is excluded from our forecast targets because the earthquake occurred in the "*alarm-undecidable*" spacetime defined in the step (2) of "Alarm map" section. This event is not included in Table 1 or Fig. 1.

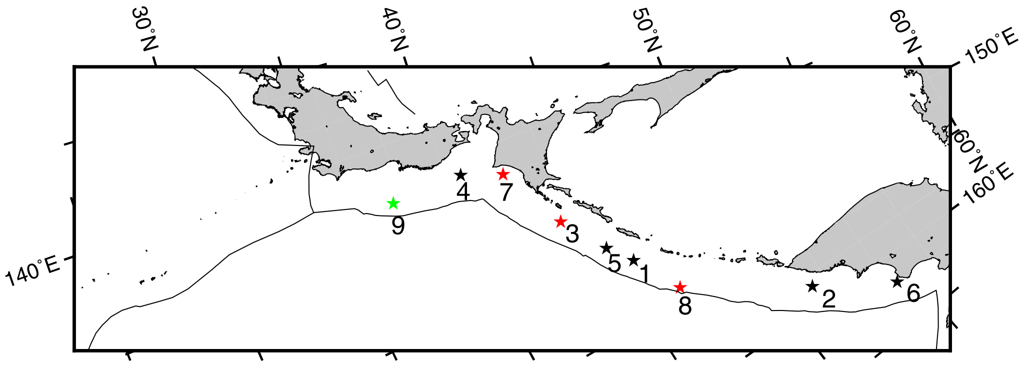


Fig. 1 Centroid location of each earthquake listed in Table 1. Red stars indicate earthquakes having $8.0 \leq M_w < 8.5$. Black and green stars, respectively, indicate earthquakes having $7.5 \leq M_w < 8.0$ and $M_w \geq 8.5$. The numerals denote the earthquake ID# in Table 1. Thin lines in the ocean indicate plate boundaries (Bird 2003)

Table 1 Main shocks with $M_w \geq 7.5$ from January 1, 1988 to December 31, 2014

EQ No. †	Y	M	D	Lon,°E	Lat,°N	Depth, km	M_w †
1	1991	12	22	151.55	45.58	31.2	7.6
2	1993	06	08	158.75	51.36	45.9	7.5
3	1994	10	04	147.63	43.60	68.2	8.3
4	1994	12	28	142.99	40.56	27.7	7.7
5	1995	12	03	150.17	44.82	25.9	7.9
6	1997	12	05	161.91	54.31	33.6	7.8
7	2003	09	25	143.84	42.21	28.2	8.3
8	2006	11	15	154.33	46.71	13.5	8.3
9	2011	03	11	143.05	37.52	20.0	9.1

† Shown in red for events with $8.0 \leq M_w < 8.5$ and in green for events with $M_w \geq 8.5$

To detect quiescence, we analyze seismicity (i.e., body-wave magnitude $m_b \geq 5.0$, depth ≤ 60 km, 5792 events in total) in the study area for January 1, 1964, through December 31, 2014. We downloaded data from *Reviewed ISC Bulletin* (ftp://isc-mirror.iris.washington.edu/pub/prerebuild/ffb/catalogue/). Although the International Seismological Centre (ISC) has finished rebuilding the bulletin from 1964 to 1979 (Storchak et al. 2017), we use old data stored in the "prerebuild" directory to ensure temporal homogeneity of source parameters, especially the magnitude.

Catalog completeness may vary with space and time. As an example, Michael (2014) found that the magnitude of completeness, Mc, of the global ISC-GEM catalog is 6.0 for shallow (depth ≤ 60 km) earthquakes that occurred from 1964 to 1975 around the world. However, in our study area, we find that Mc of the ISC catalog (depth ≤ 60 km) is much better, ranging from 3.5 to 4.7 throughout the study period (Additional file 1: Figure S1). Hence, our analysis using only events with $m_b \geq 5.0$ is prudent.

Methods

Sub-catalogs

We first make 16-year subsets of the 5792 earthquakes of $m_b \geq 5.0$. We make 351 sub-catalogs, with the starting date varying from the year 1964.0 to 1999.0, in 0.1-year steps, so that the first sub-catalog covers 1964.0–1980.0, the second sub-catalog covers 1964.1–1980.1, ..., and the 351st sub-catalog covers 1999.0–2015.0. We then remove aftershocks in each sub-catalog, using a declustering algorithm based on the eight-parameter ETAS model (Zhuang et al. 2002, 2005), where the observed seismicity is decomposed into the quasi-steady background seismicity and the temporary surge due to aftershock-type triggering following the Omori–Utsu law (Utsu 1957).

We decluster each sub-catalog independently. We first determine ETAS parameters by fitting only the seismicity in each sub-catalog. All eight ETAS parameters were reasonably stable among different sub-catalogs (Additional file 1: Figure S2). It is thus unlikely that uncertainties in the ETAS parameters affect the conclusions of the present study, though we do not attempt a quantitative assessment of the impact (e.g., Wang et al. 2010).

We then produce a declustered sub-catalog that only retains the events likely to belong to the background seismicity. Specifically, we judge that an earthquake belongs to the background seismicity if the ETAS-modeled probability P_{back} that an earthquake belongs to the background seismicity exceeds P_{rand}, a random number drawn from a uniform distribution between 0 and 1 (Zhuang et al. 2002).

We then discard the first 1-year portion of each declustered sub-catalog because the ETAS model starts working properly only after seeing a sufficient length of prior seismicity. Although the clustering parameters of Page et al. (2016) for subduction zones suggest the possibility of aftershocks having $m_b \geq 5.0$ in the second and later years, we find that aftershocks having $m_b \geq 5$ do not exceed the background rate in our study region for more than a year even after an earthquake as large as M8 class, except for the prolonged aftershock period following the 2011 M9 earthquake (Additional file 1: Figure S3). Therefore, discarding the initial year is sufficient.

We thus obtain 351 declustered sub-catalogs, each of which is 15 years long. The first sub-catalog covers 1965.0–1980.0, the second sub-catalog covers 1965.1–1980.1, …., and the last (351st) sub-catalog covers 2000.0–2015.0. Hereafter, the term 'sub-catalog' refers to these 15-year declustered catalogs, not the 16-year catalogs before declustering. Each sub-catalog contains 671–986 events, or ~ 850 on average.

ETAS parameters are also spatially variable. However, for simplicity and the stability of analysis, we assume ETAS parameters to be spatially uniform. Though not shown, we attempted declustering with spatially variable ETAS parameters, only to find that the alarm maps based on this version of declustered sub-catalogs are almost identical to the presently shown and evaluated maps made with spatially uniform ETAS parameters.

Trial forecast—anomaly detection and construction of the alarm map

In the present algorithm of trial forecasts, we first map the occurrence of quiescence anomalies throughout the studied spacetime ("Anomaly detection" section). Using the thus made *anomaly map*, we issue alarms according to the spatiotemporal distance from anomalies ("Alarm map" section). We update this alarm map every 0.1 years. Both anomaly criteria and alert criteria involve adjustable parameters. We will review the performance of the trial forecasts made with a wide range of parameter values in order to assess if the presently defined type of long-term quiescence has an earthquake-preceding tendency.

Anomaly detection

A variety of quiescence measures have been proposed as described in "Introduction". In the present study, we adopt the duration of the streak of no-earthquake days in the region, motivated by Katsumata (2017b), who investigated the seismicity preceding 23 earthquakes having $M_w \geq 8$ for the period 1990–2014. He found that a streak of no-earthquake days lasting longer than ~ 10 years preceded all but four earthquakes that occurred in regions where the background seismicity was already too low to detect any further drop.

Judgment of quiescence anomalies is made every 0.1 years, for each of the spatial grid points laid at intervals of 0.1° N × 0.1° E. The algorithm is as follows:

(1) Select a sub-catalog, which covers T_1 through T_2 $(= T_1 + 15$ years). $1965.0 \leq T_1 \leq 2000.0$ and $1980.0 \leq T_2 \leq 2015.0$. Each grid point shall be diagnosed for the occurrence of a quiescence anomaly as of T_2.

(2) Select a grid point for which the anomaly judgment is made. From the selected sub-catalog, find the six nearest earthquakes around the grid point.

(3) If the epicentral distance to the sixth nearest earthquake exceeds 100 km, conclude the grid point as *'anomaly-undetectable'* as of T_2 because an anomalous drop in seismicity is difficult to recognize for the spacetime that is already very quiet.

(4) Let dt be the time interval from the most recent of the six earthquakes to T_2. This dt represents the duration of quiescence (i.e., streak of no-earthquake ($m_b \geq 5$) days) around the grid point as of T_2. If $dt \geq T_q$, conclude the grid point as *'yes-anomaly'* as of T_2. The value of T_q, the threshold for dt to be regarded as anomalous, is one of the three adjustable parameters of the present forecast algorithm. It may represent the forecaster's idea about the duration of precursory quiescence. A higher T_q means a more stringent selection of quiescence anomalies.

(5) If neither (3) nor (4) applies to the grid point, conclude it as *'no-anomaly'.*

Alarm map

On the basis of the anomaly map obtained in "Anomaly detection" section, we now construct an alarm map, any spatiotemporal point of which is given one of the three forecast values: *'alarm-on'*, *'alarm-off'*, or *'alarm-undecidable'.* The algorithm is as follows.

(1) Label the spacetime as *'alarm-on'* if at least one 'yes-anomaly' grid point exists in the preceding period of length T_a within a distance R. R and T_a are two adjustable parameters of the present forecast scheme. They may represent the spatial and temporal ranges up to which the forecaster expects that quiescence anomalies are likely related to the subsequent earthquake.

(2) Label the spacetime as *'alarm-undecidable'* if the spacetime does not satisfy (1) and if at least one 'anomaly-undetectable' grid point exists in the preceding T_a within the distance R.

(3) Label all the remaining spacetime as *'alarm-off'.* Note that in the alarm-off spacetime, it is guaranteed that no quiescence anomaly occurred within R during the preceding T_a.

(4) Finally, we relabel any spacetime belonging to the aftershock region of earthquakes having $M_w \geq 7.5$ as 'alarm-undecidable'. This is for prudence; application to an aftershock period is beyond the scope of the traditional precursory quiescence hypoth-

esis. Using the ETAS parameters obtained in the declustering procedure, we calculate $P_{\mathrm{after}} = 1$ - P_{back} as a function of the time elapsed since the occurrence of the main shock. For the time period when $P_{\mathrm{after}} \geq 0.01$, the region within R of the grid point closest to the centroid of the main shock is designated as 'alarm-undecidable'. This step overrides decisions made in earlier steps. Spacetime labeled 'alarm-undecidable' shall not count in the evaluation (see "Evaluation of the trial forecasts" section).

Figure 2 shows, in the form of yearly snapshots, an example ($T_q = 11$ years, $R = 60$ km, $T_a = 7$ years) of the spatiotemporal alarm map created adopting the above procedure. The time attached to each snapshot is that when the alarm status was decided and issued; that is, T_2 of the newest sub-catalog available then. For the anomaly map obtained in "Anomaly detection" section, T_2 ranges $1980.0 \leq T_2 \leq 2015.0$, but, for the alarm map, T_2 ranges $1988.0 \leq T_2 \leq 2015.0$ because T_a up to 8 years is considered in the present study.

Evaluation of the trial forecasts

For each forecast model, specified by a combination of (T_q, R, T_a), we thus obtain one spatiotemporal alarm map stating the model's forecasts for 1988.0 through 2015.0. We evaluate the performance of each model by looking at its prediction rate r, compared with the alarm fraction f invested by the model. Considering the existence of 'alarm-undecidable' spacetime, the exact formulae for f and r are

$$f = V_{\mathrm{on}}/(V_{\mathrm{on}} + V_{\mathrm{off}}), \tag{1}$$

and

$$r = N_{\mathrm{on}}/(N_{\mathrm{on}} + N_{\mathrm{off}}), \tag{2}$$

where V_{on} is the total spatiotemporal volume labeled 'alarm-on', V_{off} is that labeled 'alarm-off', N_{on} is the number of target (i.e., $7.5 \leq M_w < 8.5$) earthquakes that occurred in V_{on}, and N_{off} is the number of target earthquakes that occurred in V_{off}.

To obtain r, one needs to compare the forecast (alarm map) with the catalog of target-class earthquakes. By contrast, f is independent of target earthquakes; it depends solely on the forecast model (T_q, R, T_a) and the background seismicity up to the date of alarm issue T_2. Figure 3 shows f as a function of T_2, for representative models discussed in the present paper: $T_q = 9$, 10, 11, and 12 years, with T_a and R, respectively, fixed at 7 years and 60 km. We see f is smaller for higher T_q, where models become more selective in recognizing anomalies. Additionally, note that f is quite large ($>10\%$), meaning that the present paper deals with quite vague forecasts.

Following Zechar and Jordan (2008), we define the probability gain G as

$$G = r/f. \tag{3}$$

As mentioned earlier ("Introduction" section), f is the expectancy of r under the null hypothesis that the presently defined quiescence anomalies are not relevant to the subsequent occurrence of earthquakes having $M_w \geq 7.5$. Therefore, the right-hand side of Eq. (3) represents the improvement ratio of the prediction rate, whereas the standard definition of probability gain is the improvement ratio of the success rate (i.e., the enhancement of the probability density of earthquake occurrence in the alerted spacetime) (e.g., Aki 1981). However, it can be shown mathematically that the two improvement ratios necessarily coincide (Nakatani 2020). Incidentally, note that the theoretical upper limit for G is $1/f$. As seen in Fig. 3, f was $>10\%$ in the present study, such that G cannot exceed 10, even if none of the target earthquakes are missed.

The null hypothesis of no correlation is equivalent to the proposition that the true value of G is unity. As shown in "Results and discussion" section, many of our forecast models exhibited $G > 1$, suggesting an earthquake-preceding tendency. To check if the tendency is statistically significant, we will calculate the p-value, which is the probability that G (or r) equal to or higher than the observed occurs in the random forecasts having f equal to that of the forecast (i.e., alarm maps) being scrutinized. The formula of binomial probability giving the p-value (e.g., Zechar and Jordan 2008) is

$$p(f, N_{\mathrm{on}}, N_{\mathrm{off}}) = \sum_{n=N_{\mathrm{on}}}^{(N_{\mathrm{on}}+N_{\mathrm{off}})} \binom{N_{\mathrm{on}} + N_{\mathrm{off}}}{n} f^n (1-f)^{\{(N_{\mathrm{on}}+N_{\mathrm{off}})-n\}}. \tag{4}$$

In the present study, where we can use a maximum of eight target earthquakes for statistical testing, we tentatively adopt $p < 5\%$ as the criterion for statistical significance, though the choice of threshold is a subjective matter after all.

Results and discussion

Experiment 1 (main experiment)

As Mulargia (1997) pointed out, retrospective optimization of a forecasting method is prone to over-fitting, which would lead to overrating of the method. We therefore illustrate the method's robustness by showing the results of all 210 forecast models produced in retrospective optimization instead of judging the significance solely according to the lowest p-value among them. In our optimization, we search with respect to all three

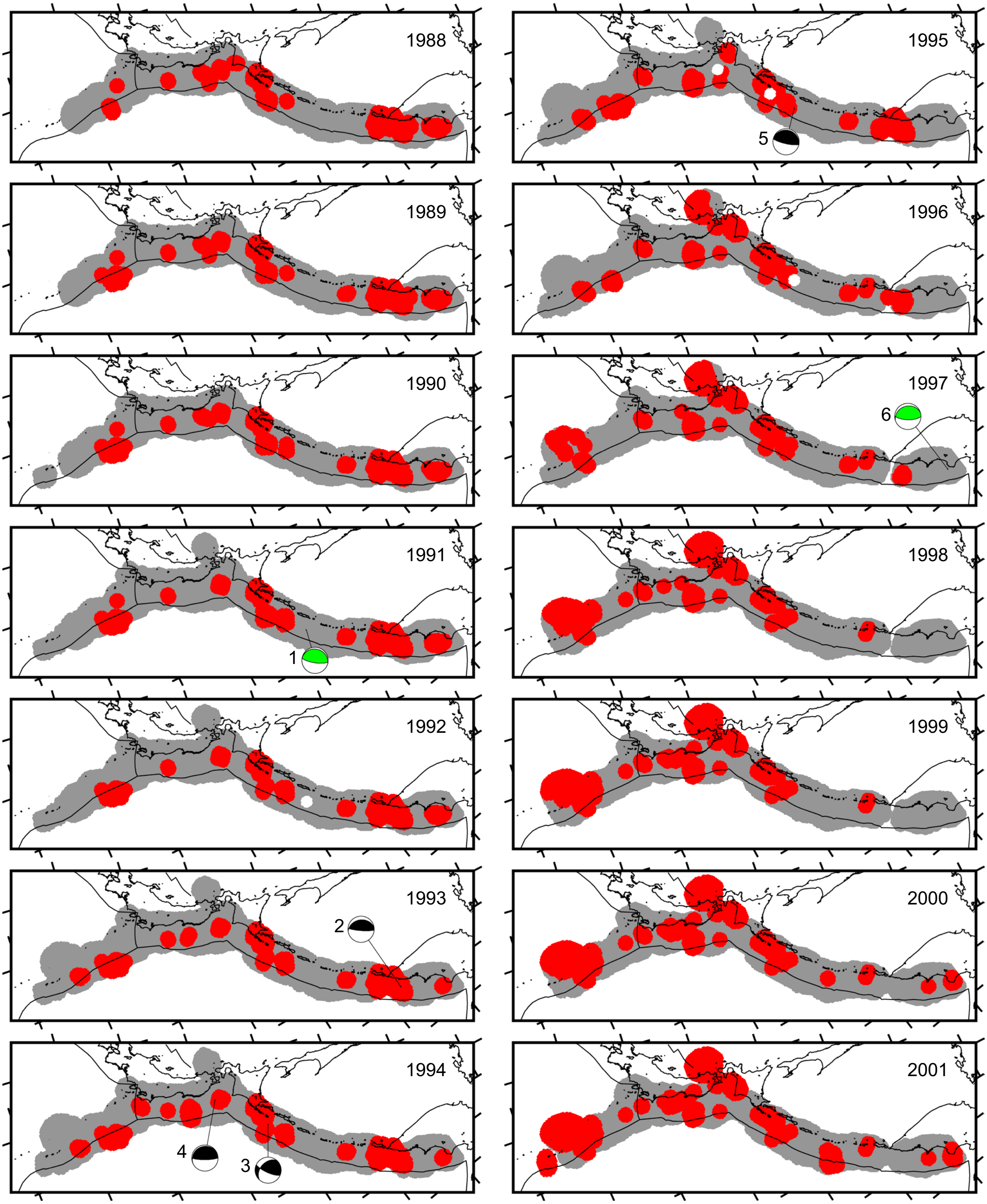


Fig. 2 An example alarm map. The spatiotemporal map is made using a forecast model (1–12 in Table 2) and is shown in yearly snapshots. Areas in red, gray, and white indicate *alarm-on*, *alarm-off*, and *alarm-undecidable* areas, respectively. The year for the forecast is shown at the top-right corner of each panel. The alarm status shown in each panel is decided at $T_2 = 0{:}00$ am, January 1 of the year, using the seismicity before. Focal mechanisms, numbered 1–8, represent the target-class ($7.5 \leq M_w < 8.5$) earthquakes that occurred in the year and correspond to EQ#1–8 in Table 1. Predicted earthquakes are shown in black, whereas missed earthquakes are shown in green. Thin lines in the ocean indicate plate boundaries (Bird 2003)

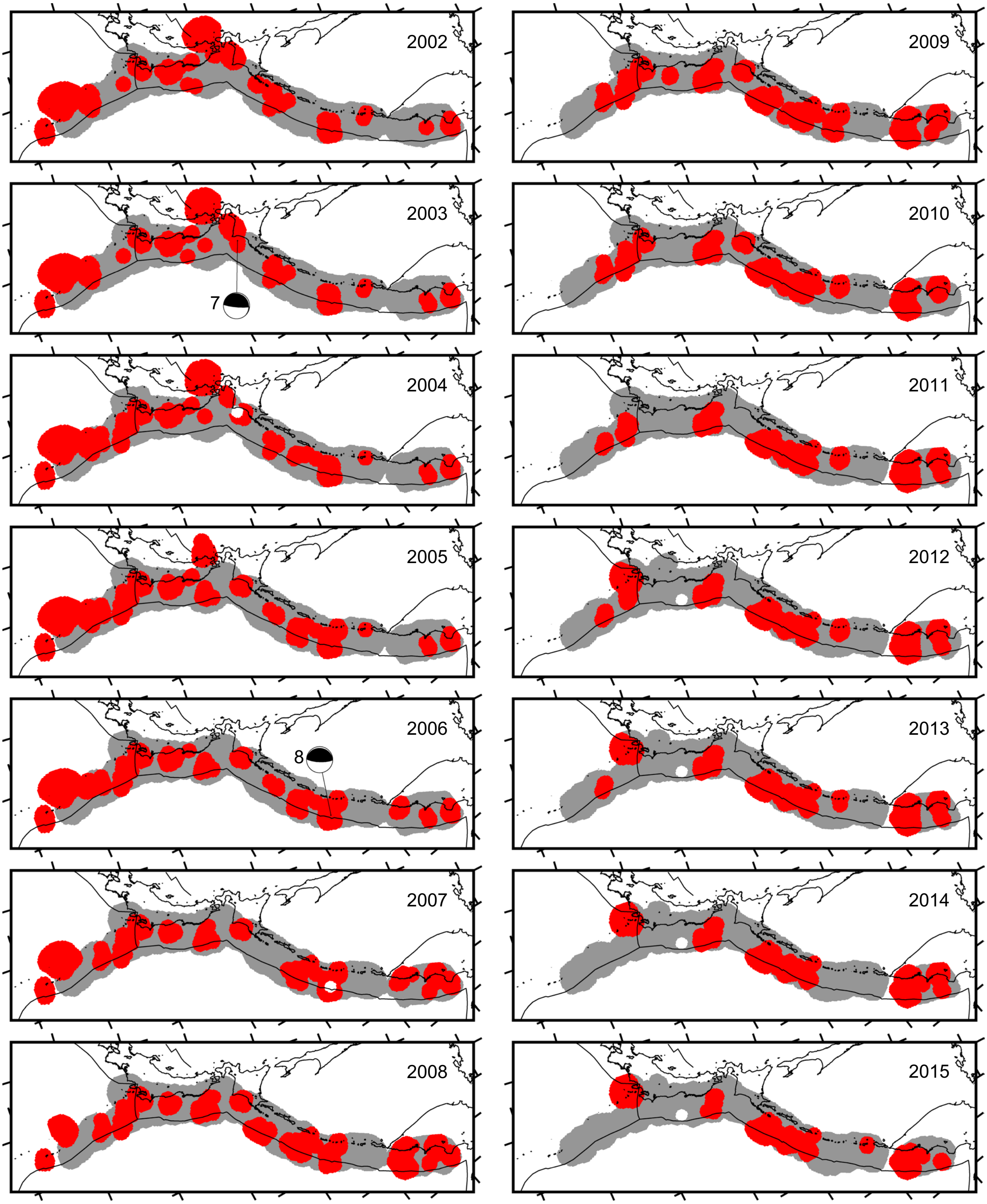


Fig. 2 continued

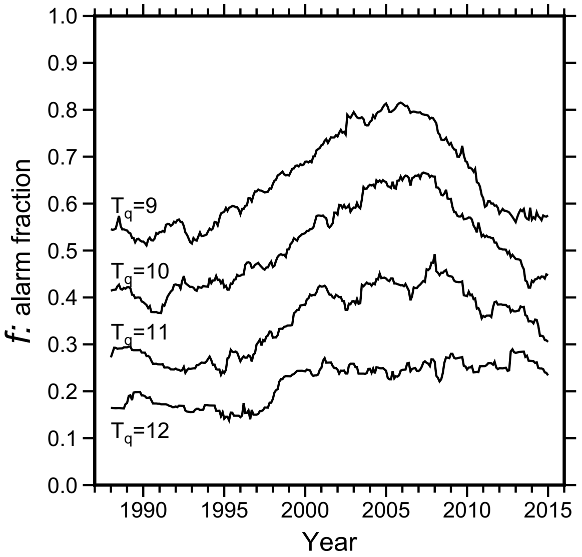


Fig. 3 Temporal variation of the alarm fraction *f*. Results of four representative models, with $T_q = 9$, 10, 11, and 12 years, are shown. $R = 60$ km and $T_a = 7$ years. Plotted against the date of alarm-map updates

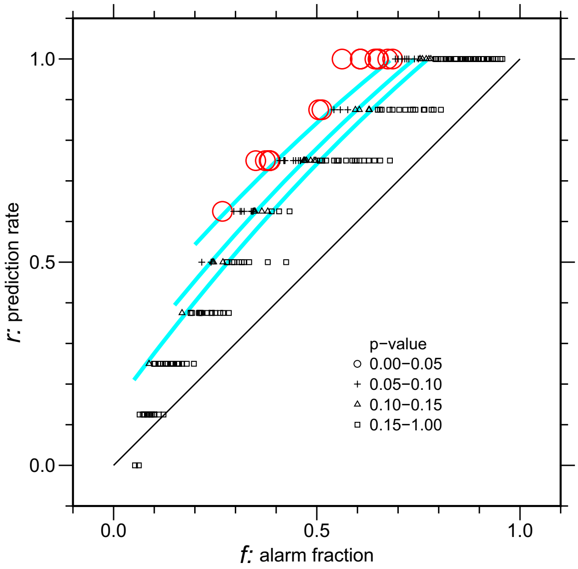


Fig. 4 ROC diagram for Experiment 1. All eight target earthquakes ($7.5 \leq M_w < 8.5$, 1988.0–2015.0, Table 1) were considered in calculating *r* for each of the 210 models tried. The marker shape denotes the *p*-value. Large red markers are used for optimal models ($p < 5\%$ and $r > 0.5$), listed in Table 2. The three blue curves are contours that correspond to $p = 5\%$, 10%, and 15%. The black diagonal line ($f = r$) corresponds to $G = 1$

main aspects that generally constitute a precursor-based forecast; T_q ($7 \leq T_q \leq 13$ years, at 1-year intervals) regulates the recognition of the suspected precursory phenomenon, while R ($50 \leq R \leq 100$ km, at 10-km intervals) and T_a ($4 \leq T_a \leq 8$ years, at 1-year intervals), respectively, regulate the spatial and temporal association between suspected precursors and subsequent earthquakes. Figure 4 is the receiver-operating-characteristic (ROC) diagram, constructed by plotting (*f*, *r*) for all 210 models. All eight target ($7.5 \leq M_w < 8.5$) earthquakes (EQ#1–8 in Table 1) are considered in calculating *r* for each model.

Figure 4 shows that *r* is higher than *f* for almost all models, suggesting that the presently defined ("Anomaly detection" section) long-term quiescence provides information that helps discriminate spacetime with heightened risk. At the same time, G ($=r/f$) is only ~ 2; the information provided by the quiescence is so weak that the risk in the alerted spacetime is only twice the secular level.

In Fig. 4, we use different marker shapes according to the model's *p*-value. Furthermore, we use large red markers for the optimal models yielding $p < 5\%$ and $r > 0.5$. Table 2 lists these optimal models. (In Experiment 1, all models that yielded $p < 5\%$ also yielded $r > 0.5$; however, the condition $r > 0.5$ matters for consistency with corresponding plots in other experiments shown later.) At face value, $p < 5\%$ for these models implies the statistical significance of the earthquake-preceding tendency. Below, we look into the characteristics of optimal models for Experiment 1 (Table 2), in an attempt to elucidate the characteristics of precursory long-term quiescence. Note that Experiment 1 is the main and base experiment of the present paper.

Ten of the 15 optimal models use $T_q = 9$ years (Table 2, models 1–1 through 1–10), while R ranges 50–80 km and T_a ranges 4–8 years. These 10 models yield high *r*; eight achieve $r = 8/8$ and two achieve $r = 7/8$. For the 10 models, *f* ranges 50%–70% and G ranges 1.5–1.8. The remaining five models (Table 2, models 1–11 through 1–15) use stricter criteria for quiescence anomalies, $T_q = 10$–12 years. As a result, EQ#1 and #6 become surprise events, lowering *r* to 6/8. For model 1–15, which adopts $T_q = 12$ years, EQ#5 also becomes a surprise event, further lowering *r* to 5/8. Nevertheless, these five models with higher T_q (i.e., 10–12 years) achieve higher G of 1.9–2.3 because *f* is much improved (i.e., slashed) to 27%–39% thanks to the stringent selection of quiescence anomalies. The size of the alerted spacetime (R, T_a) around an anomaly is generally larger than that for the 10 models with $T_q = 9$ years, but not by much.

As seen above, there is a trade-off between the two performance demands; that is, higher *r* and lower *f*. The *p*-value strongly depends on *r* and *f* in the relevant range (Fig. 5). The balance between the two seems to be regulated by (T_q, R, T_a) in a self-evident sense. It is

Table 2 Optimal ($p < 5\%$ and $r > 0.5$) forecast models from Experiment 1

Model	T_q, year	R, km	T_a, year	G	r	f	p	†Alerted=1; Missed=0
1-1	9	50	5	1.7	0.88	0.51	0.041	01111111-
1-2	9	50	6	1.8	1.00	0.56	0.010	11111111-
1-3	9	50	7	1.6	1.00	0.61	0.019	11111111-
1-4	9	50	8	1.5	1.00	0.65	0.032	11111111-
1-5	9	60	4	1.7	0.88	0.50	0.037	01111111-
1-6	9	60	6	1.6	1.00	0.61	0.018	11111111-
1-7	9	60	7	1.5	1.00	0.65	0.033	11111111-
1-8	9	70	6	1.6	1.00	0.64	0.029	11111111-
1-9	9	70	7	1.5	1.00	0.69	0.049	11111111-
1-10	9	80	6	1.5	1.00	0.67	0.043	11111111-
1-11	10	60	4	2.0	0.75	0.37	0.035	01111011-
1-12	11	60	7	2.1	0.75	0.35	0.025	01111011-
1-13	11	60	8	2.0	0.75	0.38	0.040	01111011-
1-14	11	70	7	1.9	0.75	0.39	0.041	01111011-
1-15	12	70	8	2.3	0.62	0.27	0.037	01110011-

† EQ#1 through 9 from left to right. Shown in red for events with $8.0 \leq M_w < 8.5$

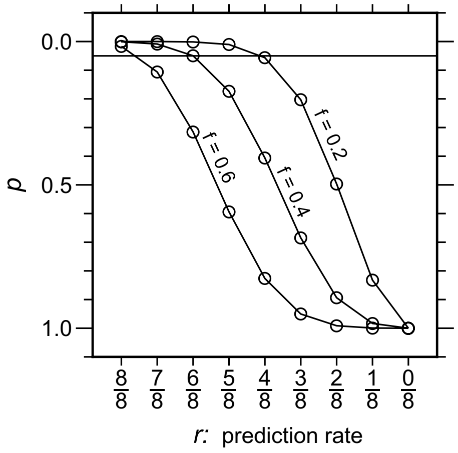


Fig. 5 *p*-value as a function of the prediction rate. The horizontal line of $p = 5\%$ is shown for reference

thus hardly meaningful to ask which model is the best. Instead, we emphasize that these well-performing models (Table 2) lie coherently in the model space (T_q, R, T_a). This implies that the favorable performance of the optimized models (Table 2) originates from decent optimization reflecting universal properties of mechanisms underlying the earthquake-preceding tendency of the long-term quiescence, rather than originating from the use of deliberate, complex algorithms that do whatever to score well. Such gerrymandering is unlikely because our forecast algorithm, including adjustment through (T_q, R, T_a), is straightforward. However, the present trial forecast is calibrated with only eight earthquakes. Hence, over-fitting, if not gerrymandering, can readily occur. We will check this point by reporting on cross-validation experiments in "Experiments 2 and 3 (cross-validation)" section.

Experiments 2 and 3 (cross-validation)

Although they look good, our results so far could benefit from over-fitting as cautioned already; good performance exhibited by the retrospectively optimized forecast models does not assure good performance for future data (e.g., Mulargia 1997). We thus conduct cross-validation experiments by dividing the study period of Experiment 1 into two, using one period to train models and the other period to evaluate elite models in the learning period. Experiment 2 uses the later period for training and the earlier period for evaluation, whereas Experiment 3 uses the earlier period for training and the later period for evaluation. In both experiments, we explore the same range of the parameter space (T_q, R, T_a) as in Experiment 1, amounting to 210 models. The target class of forecast remains $7.5 \le M_w < 8.5$, the same as in Experiment 1. Unfortunately, our low number of target earthquakes does not allow us to conduct cross-validation experiments separately for earthquakes having $M_w < 8.0$ and $M_w \ge 8.0$, though this would be desirable given the plausible dependence of precursory quiescence on the main shock magnitude (Additional file 2).

Experiment 2

In Experiment 2, models are first optimized through trial forecasts based on sub-catalogs in the second half ($1995.0 \le T_2 < 2015.0$), for which four earthquakes having $7.5 \le M_w < 8.5$ (EQ#5–8) are available as forecast targets. Figure 6 shows the performance in this learning period for all 210 models. We draw common reference contours (blue curves) to directly compare the strength of the apparent correlations seen in Figs. 4 and 6. The three contours correspond to contours of (f, r) that would yield $p = 5\%$, 10%, and 15% provided that eight target earthquakes were available as in Experiment 1. The marker shape convention is the same as that in Fig. 4. This ROC diagram, mostly implying $G > 1$, is generally similar to that of Experiment 1, but no model achieves $p < 5\%$ owing to the limited number of target earthquakes (i.e., four target earthquakes). Thus, formally, our cross-validation attempt has failed already in the learning period; the number of available earthquakes (four in the learning period, four in the evaluation period) is too few to attempt cross-validation of the weak ($G < 10$) earthquake-preceding tendency.

Nonetheless, we proceed by choosing eight models that achieve $p < 15\%$ and $r > 0.5$ in the learning period (indicated by large red symbols in Fig. 6) as optimal models (Table 3). Performances of these learning-period elites are then evaluated by making forecasts based on the sub-catalogs in the first half ($1988.0 \le T_2 < 1995.0$), the evaluation period of Experiment 2. All four target earthquakes (EQ#1–4) in the first half are used for evaluation. Figure 7 shows the ROC diagram for this evaluation period. Note that marker shapes in Fig. 7 reflect the p-values of the respective models in the learning period, and large red symbols represent the learning-period elites listed in Table 3. Table 4 shows how these learning-period elites perform in the evaluation period.

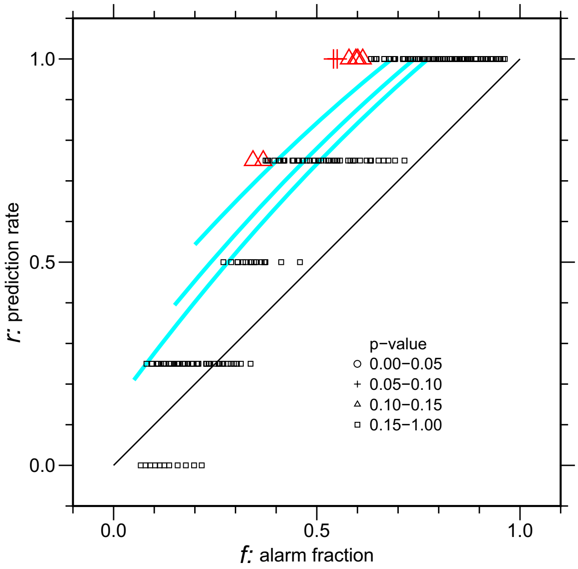


Fig. 6 ROC diagram for the learning period of Experiment 2. Four earthquakes having $7.5 \le M_w < 8.5$ that occurred in 1995.0–2015.0 were used in calculating r for each of the 210 models tried. The marker shape denotes the p-value. Large red markers are used for optimal models ($p < 15\%$ and $r > 0.5$), listed in Table 3. The three blue curves are identical to those in Fig. 4; that is, they correspond to (f, r) that would yield $p = 5\%$, 10%, and 15% if there were eight target earthquakes

In both the learning (Fig. 6) and evaluation (Fig. 7) periods, r is higher than f for most models, implying skill better than random forecasts. We thus examine the performance in the two periods in detail to seek signatures of over-fitting.

A comparison of Tables 3 and 4 shows that N_{on} is lower by 1 in the evaluation period for most models. In closer examination, we see that almost all the models miss EQ#1 in the evaluation period, worsening the p-value. We now explain what happened concretely here. All four target earthquakes in the learning period occurred within 4 years of the quiescence of $T_q = 9$ years. Hence, most of the optimized models adopt relatively short T_a of 4 or 5 years to avoid excessive f. In contrast, EQ#1, in the evaluation period, occurred 6 years after a 9-year quiescence and is missed by all the learning-period elites except model 2–2, which adopts $T_a = 6$ years. This can be said to be over-fitting, if mild.

Table 3 Learning-period performance for the learning-period elite models ($p < 15\%$ and $r > 0.5$ in the learning period) in Experiment 2

Model	T_q,	R,	T_a,	G	r	f	p	Alerted=1;
	year	km	year					Missed=0
2-1	9	50	5	1.8	1.00	0.55	0.092	----1111-
2-2	9	50	6	1.7	1.00	0.60	0.130	----1111-
2-3	9	60	4	1.8	1.00	0.54	0.086	----1111-
2-4	9	60	5	1.7	1.00	0.60	0.126	----1111-
2-5	9	70	4	1.7	1.00	0.58	0.113	----1111-
2-6	9	80	4	1.6	1.00	0.61	0.141	----1111-
2-7	11	60	6	2.2	0.75	0.34	0.120	----1011-
2-8	11	80	5	2.0	0.75	0.37	0.144	----1011-

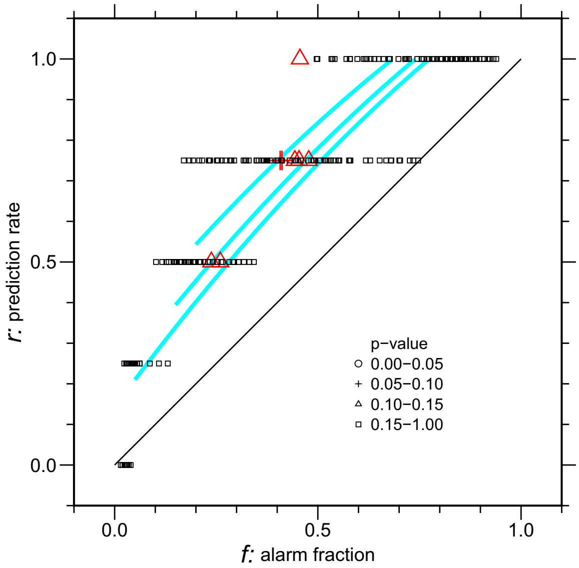


Fig. 7 ROC diagram for the evaluation period of Experiment 2. Four earthquakes having $7.5 \leq M_w < 8.5$ that occurred in 1988.0–1995.0 were used in calculating *r* for each of the 210 models tried. The marker shape denotes the *p*-value in the learning period. Large red markers are used for learning-period elites ($p < 15\%$ and $r > 0.5$ in the learning period), listed in Tables 3 and 4. The three blue curves are identical to those in Figs. 4 and 6; that is, they correspond to (f, r) that would yield $p = 5\%$, 10%, and 15% if there were eight target earthquakes

However, when we look at *G*, learning-period elites perform equally well ($G \sim 2$) in the evaluation period (red markers in Fig. 8a). This is largely explained by the fact that *f* was generally lower in the evaluation period, being 70–80% of the value in the learning period (Tables 3 and 4). This cancels the adverse effect of the ~ 25% drop in *r*. Overall, we may say that there is no severe over-fitting in Experiment 2.

Furthermore, the (T_q, R, T_a) range of the optimal models in Experiment 1, the main experiment of the present paper, resembles that in Experiment 2 (Table 3). As described above, the optimal models in Experiment 2 do not seem to involve severe over-fitting like in Experiment 3 ("Experiment 3" section). Hence, the favorable results ($G \sim 2$) of Experiment 1, even though the same data are used for learning and evaluation, are probably not due to over-fitting. We thus surmise (but not prove, unfortunately) that the favorable performance (G ~ 2) of Experiment 1 is likely achieved by decent optimization and it may be taken as an encouraging result.

Experiment 3

We conduct another cross-validation experiment, Experiment 3, where the learning period and evaluation period are flipped relative to those in Experiment 2. Figure 9 is the ROC diagram for the learning period (i.e., the first half) of Experiment 3. For consistency with Experiment 2, we regard models that achieve $p < 15\%$ and $r > 0.5$ (Table 5) as optimal and highlight them using large red symbols, though quite a few (nine) of them achieve $p < 5\%$. Moreover, the ROC for the learning period of Experiment 3 (Fig. 9) looks far better than those for Experiments 1 and 2, yielding higher *G* often exceeding 3. Table 5 shows that models

Table 4 Evaluation-period performance for the learning-period elite models ($p < 15\%$ and $r > 0.5$ in the learning period) in Experiment 2

Model	T_q, year	R, year	T_a, year	G	r	f	p	Alerted=1; Missed=0
2-1	9	50	5	1.8	0.75	0.41	0.193	0111-----
2-2	9	50	6	2.2	1.00	0.46	0.043	1111-----
2-3	9	60	4	1.8	0.75	0.41	0.188	0111-----
2-4	9	60	5	1.7	0.75	0.45	0.247	0111-----
2-5	9	70	4	1.7	0.75	0.44	0.233	0111-----
2-6	9	80	4	1.6	0.75	0.48	0.280	0111-----
2-7	11	60	6	2.1	0.50	0.24	0.242	0110-----
2-8	11	80	5	1.9	0.50	0.26	0.278	0110-----

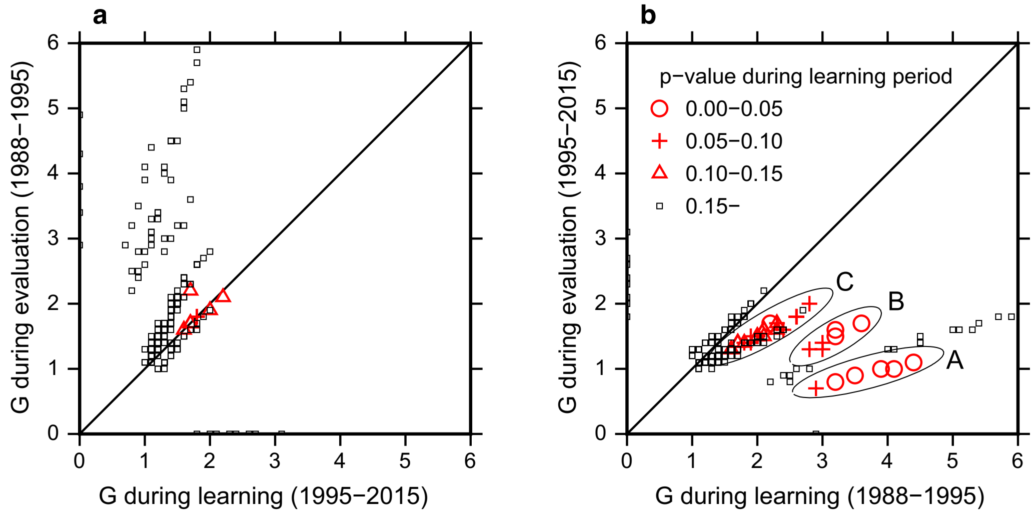


Fig. 8 Probability gain of respective models during learning vs. evaluation periods. Red markers denote learning-period elites. **a** Experiment 2. **b** Experiment 3. See the main text for the groupings of A, B, and C

with higher T_q perform particularly well in Experiment 3. In other words, this superb learning-period performance of Experiment 3 is due to many of the target earthquakes being preceded by particularly prolonged quiescence, allowing the slashing of f through stringent anomaly selection. This strategy does not sacrifice r in the learning period because as many as three earthquakes (EQ#2, 3, 4) were preceded by quiescence anomalies as strong as $T_q \geq 12$ years (see models 3–30 through 3–39).

Figure 10 is the ROC diagram of Experiment 3 for the evaluation period. Large red symbols highlight the learning-period elites (Table 5). Table 6 lists their performances in the evaluation period, which are considerably worse than those in the leaning period. Figure 8b (red markers) compares G of respective learning-period elites between learning and evaluation periods; all these models show a drop, often severe, in G during the evaluation period, implying serious over-fitting in Experiment 3, in contrast with the case of Experiment 2 (Fig. 8a).

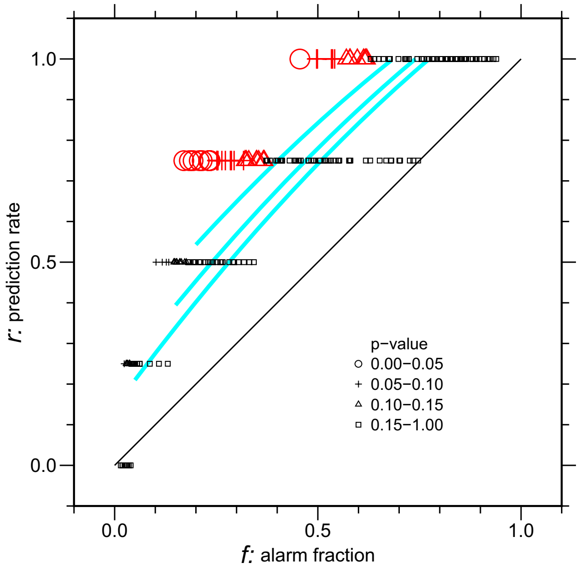


Fig. 9 ROC diagram for the learning period of Experiment 3. Four earthquakes having $7.5 \le M_w < 8.5$ that occurred in 1988.0–1995.0 were used in calculating *r* for each of the 210 models tried. The marker shape denotes the *p*-value. Large red markers are used for optimal models ($p < 15\%$ and $r > 0.5$), listed in Table 5. The three blue curves are identical to those in Fig. 4; that is, they correspond to (*f*, *r*) that would yield $p = 5\%$, 10%, and 15% if there were eight target earthquakes

Let us take a closer look. Results (red markers) in Fig. 8b may be divided into three clusters (groups A, B, and C in Fig. 8b). Group-A models have the most severe performance drop; *G* in the evaluation period is often below unity. In terms of the group-averaged *G*, Group A is the best in the learning period but is the worst in the evaluation period. In contrast, Group C is the worst in the learning period but is the best in the evaluation period. We examine Tables 5 and 6 in detail below to find specific mechanisms of over-fitting in Experiment 3.

All Group-A models adopt a high value of T_q, as high as 12 years, meaning stringent anomaly selection, which helps slash *f*. Additionally, values of T_a are mostly 7 years for Group-A models and shorter than T_a of 8 years adopted by the other models with $T_q = 12$ years, which belong to Group B. Shorter T_a also helps suppress *f*. In the learning period, three (out of four) earthquakes occurred within 7 years of strong quiescence anomalies lasting at least 12 years. Hence, Group-A models could achieve good *r* ($3/4 = 75\%$) despite their restrictive alerting policy. However, in the evaluation period, only one earthquake (out of four) was preceded by such strong quiescence, resulting in $r = 1/4$ and $G \sim 1$. Similar over-fitting occurred in Group-B models, which mostly adopt the same stringent T_q of 12 years as do Group-A models. The performance drop is less severe because, thanks to the adoption of $T_a = 8$ years instead of 7 years, Group-B models did not miss EQ#7 that occurred 7.5 years after the recognition of the quiescence of $T_q = 12$ years, so that *r* in the evaluation period is 2/4 instead of 1/4 for Group A.

Note the above difference in quiescence behavior between the learning and evaluation periods cannot be ascribed to the plausible dependence of the quiescence duration on the main shock magnitude (Additional file 2), because only one (EQ#3) of the three earthquakes having $M_w \ge 8.0$ occurred in the learning period. Instead, the superb performance in the learning period of Experiment 3 arises from pure luck that the quiescence anomalies preceding two (EQ#2 and #4) of the three earthquakes having $M_w < 8$ in the learning period happened to have particularly prolonged durations of at least 12 years. We thus conclude that most (i.e., groups A and B) learning-period elites commit a classical over-fitting. They became too picky in anomaly detection because they mistook mere apparent features introduced by natural fluctuation as a universal property of the precursory quiescence.

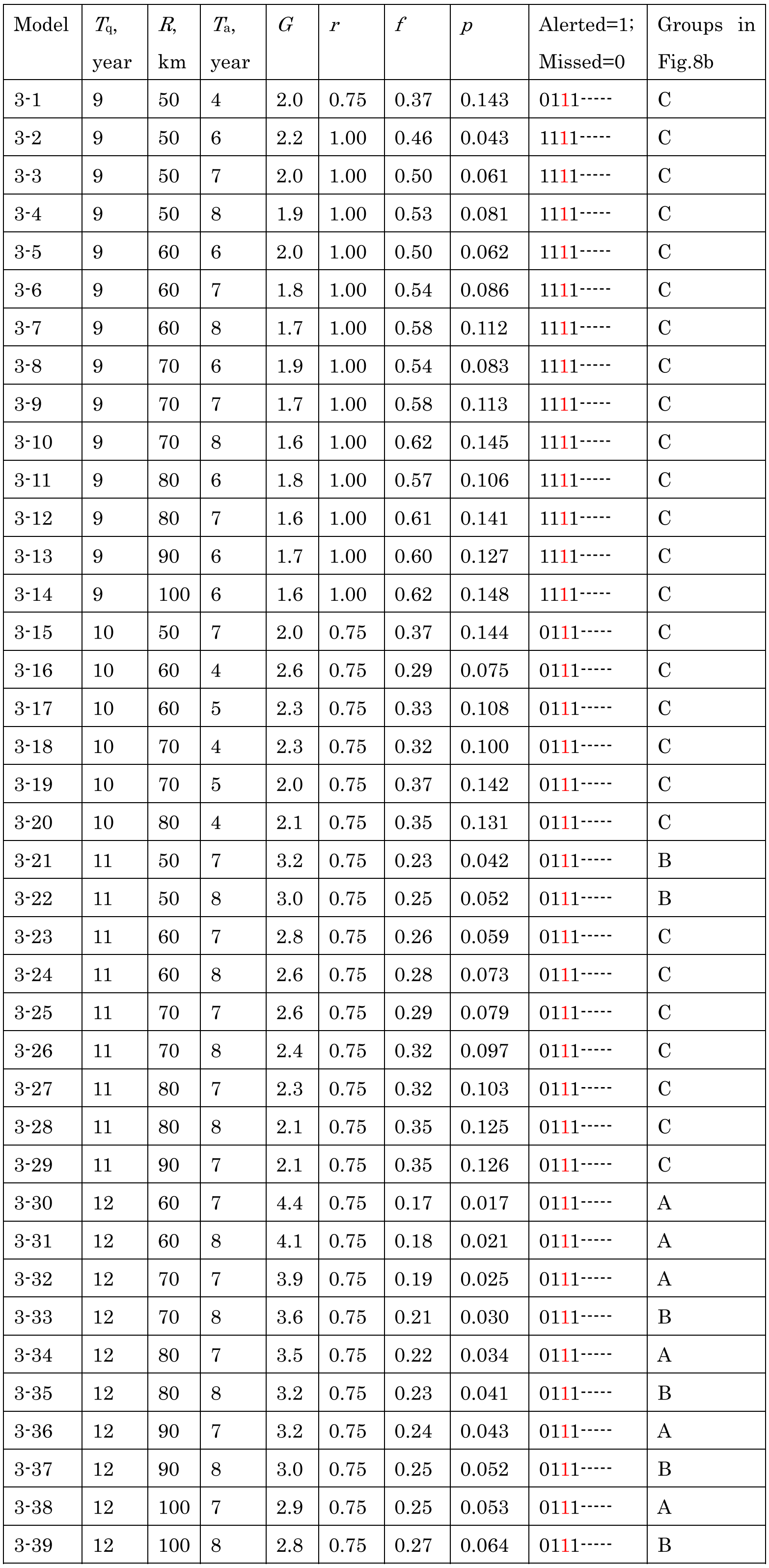
Although *G* drops in the evaluation period also for Group-C models, we do not recognize obvious over-fitting. Group-C models, being the least selective among the learning-period elites of Experiment 3, have $r = 3/4$ or 4/4 in both learning and evaluation periods. The observed mild drop in *G* is mostly attributable to higher *f* in the evaluation period (Fig. 3).

In conclusion, Experiment 3 involves severe over-fitting. This is pretty much expected, given the limited number of learning data (i.e., only four target earthquakes in the learning period). Specifically, over-fitting occurred because our optimization tried to slash *f* too aggressively, using apparent features that arose by pure luck. Of course, there is no way to tell that these features are not a real property until cases lacking these features are learned; optimization using a sufficient number of target earthquakes would ease the problem.

We, however, emphasize that the overly stringent policy in anomaly detection adopted by Experiment 3 was not preferred in Experiment 1, the main experiment of the present paper. Its optimal models (Table 2) mostly use $T_q = 9$ years, and only one of the 15 adopts $T_q = 12$ years. Given the insights from Experiment 3, we may say high T_q was not much preferred in Experiment 1 because securing a high *r* is more important than suppressing *f* to achieve a sufficiently small *p*, which is our primary criterion in choosing optimal models. We therefore believe that the modest ($G \sim 2$) but still favorable performance of Experiment 1 is unlikely to be a ghost arising from over-fitting.

Although the present cross-validation tests failed to confirm it formally, we would say, as an optimistic

Table 5 Learning-period performance for the learning-period elite models ($p < 15\%$ and $r > 0.5$ in the learning period) in Experiment 3



Model	T_q, year	R, km	T_a, year	G	r	f	p	Alerted=1; Missed=0	Groups in Fig.8b
3-1	9	50	4	2.0	0.75	0.37	0.143	0111-----	C
3-2	9	50	6	2.2	1.00	0.46	0.043	1111-----	C
3-3	9	50	7	2.0	1.00	0.50	0.061	1111-----	C
3-4	9	50	8	1.9	1.00	0.53	0.081	1111-----	C
3-5	9	60	6	2.0	1.00	0.50	0.062	1111-----	C
3-6	9	60	7	1.8	1.00	0.54	0.086	1111-----	C
3-7	9	60	8	1.7	1.00	0.58	0.112	1111-----	C
3-8	9	70	6	1.9	1.00	0.54	0.083	1111-----	C
3-9	9	70	7	1.7	1.00	0.58	0.113	1111-----	C
3-10	9	70	8	1.6	1.00	0.62	0.145	1111-----	C
3-11	9	80	6	1.8	1.00	0.57	0.106	1111-----	C
3-12	9	80	7	1.6	1.00	0.61	0.141	1111-----	C
3-13	9	90	6	1.7	1.00	0.60	0.127	1111-----	C
3-14	9	100	6	1.6	1.00	0.62	0.148	1111-----	C
3-15	10	50	7	2.0	0.75	0.37	0.144	0111-----	C
3-16	10	60	4	2.6	0.75	0.29	0.075	0111-----	C
3-17	10	60	5	2.3	0.75	0.33	0.108	0111-----	C
3-18	10	70	4	2.3	0.75	0.32	0.100	0111-----	C
3-19	10	70	5	2.0	0.75	0.37	0.142	0111-----	C
3-20	10	80	4	2.1	0.75	0.35	0.131	0111-----	C
3-21	11	50	7	3.2	0.75	0.23	0.042	0111-----	B
3-22	11	50	8	3.0	0.75	0.25	0.052	0111-----	B
3-23	11	60	7	2.8	0.75	0.26	0.059	0111-----	C
3-24	11	60	8	2.6	0.75	0.28	0.073	0111-----	C
3-25	11	70	7	2.6	0.75	0.29	0.079	0111-----	C
3-26	11	70	8	2.4	0.75	0.32	0.097	0111-----	C
3-27	11	80	7	2.3	0.75	0.32	0.103	0111-----	C
3-28	11	80	8	2.1	0.75	0.35	0.125	0111-----	C
3-29	11	90	7	2.1	0.75	0.35	0.126	0111-----	C
3-30	12	60	7	4.4	0.75	0.17	0.017	0111-----	A
3-31	12	60	8	4.1	0.75	0.18	0.021	0111-----	A
3-32	12	70	7	3.9	0.75	0.19	0.025	0111-----	A
3-33	12	70	8	3.6	0.75	0.21	0.030	0111-----	B
3-34	12	80	7	3.5	0.75	0.22	0.034	0111-----	A
3-35	12	80	8	3.2	0.75	0.23	0.041	0111-----	B
3-36	12	90	7	3.2	0.75	0.24	0.043	0111-----	A
3-37	12	90	8	3.0	0.75	0.25	0.052	0111-----	B
3-38	12	100	7	2.9	0.75	0.25	0.053	0111-----	A
3-39	12	100	8	2.8	0.75	0.27	0.064	0111-----	B

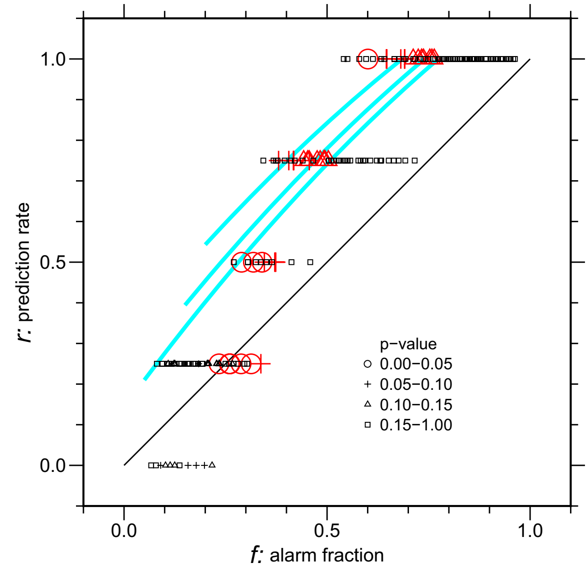


Fig. 10 ROC diagram for the evaluation period of Experiment 3. Four earthquakes having $7.5 \leq M_w < 8.5$ that occurred in 1995.0–2015.0 were used in calculating *r* for each of the 210 models tried. The marker shape denotes the *p*-value in the learning period. Large red markers are used for learning-period elites ($p < 15\%$ and $r > 0.5$ in the learning period), listed in Tables 5 and 6. The three blue curves are identical to those in Figs. 4 and 9; that is, they correspond to (*f*, *r*) that would yield $p = 5\%$, 10%, and 15% if there were eight target earthquakes

conjecture, that there is a good chance that the earthquake-preceding tendency of long-term quiescence will be proven one day by having a higher number of target earthquakes for trial forecasts, maybe several times more. As an example, if the number of target earthquakes doubles in the future (i.e., 16 in total), the learning and evaluation periods can use eight earthquakes each. With eight earthquakes, a *p*-value less than 5% can be shown for the modest ($G \sim 2$) performance we saw in Experiments 1 and 2 and for group C in Experiment 3. To give some numbers, $p < 5\%$ is attained by $G \geq 1.75$ ($r \geq 7/8$) if *f* is 50%, and by $G \geq 2.5$ ($r \geq 5/8$) if *f* is 25%.

Summary and conclusions

We made simple trial forecast experiments where we issue alarms valid for time T_a to regions within a distance *R* of the location where long-term quiescence is detected. We defined a quiescence anomaly as the absence of nearby earthquakes having $m_b \geq 5$ for a time duration of at least T_q. On the basis of these anomalies, alarm maps were made throughout the study area (i.e., the northwestern margin of the Pacific Plate) by exhaustively scanning the studied spacetime. In each of our forecast experiments (Experiments 1–3), alarm maps, made with a range of (T_q, *R*, T_a) values, were evaluated using a ROC diagram. Our main experiment (Experiment 1), targeting all eight earthquakes having $7.5 \leq \mathrm{M_w} < 8.5$, found a range of forecast models exhibiting *p*-values less than 5%, and *G* of ~ 2, supporting the existence of an earthquake-preceding tendency of long-term quiescence.

These favorable results of Experiment 1, however, could be a ghost owing to over-fitting because we used the same data for training and evaluation. We therefore attempted cross-validation using four of the eight earthquakes to train forecast models and the other four earthquakes to evaluate the optimal models obtained from the learning period. We conducted two cross-validation experiments (Experiments 2 and 3) by flipping the datasets for training and evaluation. In Experiment 2, the performance was similar for learning and evaluation periods. Additionally, optimal models and their *G* values were similar to those in Experiment 1. The above implies that the apparent success ($G \sim 2$) in Experiment 1 is unlikely attributable to over-fitting. However, we failed in formal cross-validation as none of the models achieved $p < 5\%$ in the learning period of Experiment 2. This is not surprising, given that statistical power with only four target earthquakes available is likely insufficient to demonstrate the significance of the modest earthquake-preceding tendency of $G \sim 2$, if any, exhibited in Experiment 1.

In the other cross-validation experiment (Experiment 3), many models, in the learning period, yielded $p < 5\%$ and *G* much higher than that in our base experiment (Experiment 1). However, this was a typical over-fitting effect; the models fared miserably in the evaluation period. Close examination of Experiment 3 has revealed that the over-fitting occurred because the models became too picky in anomaly detection because they mistook fortuitous features only seen in the learning period as real properties of the precursory quiescence. Fortunately, that excessive selectiveness was not preferred in the optimal models (Table 2) in our main experiment (Experiment 1), again supporting the likelihood that the favorable results of Experiment 1 are real, though we admit that we failed in formal cross-validation.

In the course of the present study, we have also found that *G* is much higher if the forecasts target only earthquakes having $M_w \geq 8$ (Experiment 5 in Additional file 2); $G > 5$ and $p < 5\%$ were easily achieved. However, this was a result obtained with only three target earthquakes, and we have nothing to say against the possibility of over-fitting. Meanwhile, forecasts (Experiment 4 in Additional file 2) targeting only five smaller earthquakes ($7.5 \leq M_w < 8$) found a correlation weaker than that in Experiment 1, which targeted all earthquakes of $7.5 \leq M_w < 8.5$, and a *p*-value less than 5% was not obtained. Hence, we have no direct basis for claiming that the quiescence's earthquake-preceding tendency

Table 6 Evaluation-period performance for the learning-period elite models ($p < 15\%$ and $r > 0.5$ in the learning period) in Experiment 3

Model	T_q, year	R, km	T_a, year	G	r	f	p	Alerted=1; Missed=0	Groups in Fig.8b
3-1	9	50	4	1.5	0.75	0.50	0.305	----0111-	C
3-2	9	50	6	1.7	1.00	0.60	0.130	----1111-	C
3-3	9	50	7	1.5	1.00	0.65	0.176	----1111-	C
3-4	9	50	8	1.4	1.00	0.69	0.228	----1111-	C
3-5	9	60	6	1.5	1.00	0.65	0.174	----1111-	C
3-6	9	60	7	1.4	1.00	0.69	0.228	----1111-	C
3-7	9	60	8	1.4	1.00	0.73	0.289	----1111-	C
3-8	9	70	6	1.5	1.00	0.68	0.216	----1111-	C
3-9	9	70	7	1.4	1.00	0.72	0.276	----1111-	C
3-10	9	70	8	1.3	1.00	0.76	0.341	----1111-	C
3-11	9	80	6	1.4	1.00	0.71	0.257	----1111-	C
3-12	9	80	7	1.3	1.00	0.75	0.321	----1111-	C
3-13	9	90	6	1.4	1.00	0.74	0.296	----1111-	C
3-14	9	100	6	1.3	1.00	0.76	0.330	----1111-	C
3-15	10	50	7	1.5	0.75	0.50	0.317	----1011-	C
3-16	10	60	4	1.8	0.75	0.41	0.186	----1011-	C
3-17	10	60	5	1.6	0.75	0.46	0.250	----1011-	C
3-18	10	70	4	1.7	0.75	0.44	0.232	----1011-	C
3-19	10	70	5	1.5	0.75	0.49	0.303	----1011-	C
3-20	10	80	4	1.6	0.75	0.48	0.277	----1011-	C
3-21	11	50	7	1.5	0.50	0.34	0.419	----0011-	B
3-22	11	50	8	1.3	0.50	0.37	0.479	----0011-	B
3-23	11	60	7	2.0	0.75	0.38	0.158	----1011-	C
3-24	11	60	8	1.8	0.75	0.42	0.199	----1011-	C
3-25	11	70	7	1.8	0.75	0.42	0.201	----1011-	C
3-26	11	70	8	1.6	0.75	0.46	0.250	----1011-	C
3-27	11	80	7	1.7	0.75	0.45	0.245	----1011-	C
3-28	11	80	8	1.5	0.75	0.49	0.301	----1011-	C
3-29	11	90	7	1.6	0.75	0.48	0.287	----1011-	C
3-30	12	60	7	1.1	0.25	0.23	0.656	----0001-	A
3-31	12	60	8	1.0	0.25	0.26	0.700	----0001-	A
3-32	12	70	7	1.0	0.25	0.26	0.702	----0001-	A
3-33	12	70	8	1.7	0.50	0.29	0.330	----0011-	B
3-34	12	80	7	0.9	0.25	0.29	0.744	----0001-	A
3-35	12	80	8	1.6	0.50	0.32	0.381	----0011-	B
3-36	12	90	7	0.8	0.25	0.31	0.778	----0001-	A
3-37	12	90	8	1.4	0.50	0.35	0.429	----0011-	B
3-38	12	100	7	0.7	0.25	0.34	0.807	----0001-	A
3-39	12	100	8	1.3	0.50	0.37	0.475	----0011-	B

holds for earthquakes of $7.5 \le M_w < 8$ as well. However, the (T_q, R, T_a) range for optimal models in Experiment 4 was similar to that in Experiment 1. Hence, there remains a reasonable chance that long-term quiescence has a tendency, if weaker, to precede even these smaller earthquakes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40623-021-01418-z.

Additional file 1: Figure S1. Temporal change in the minimum magnitude of completeness (M_c). **Figure S2.** ETAS parameters of each sub-catalog. **Figure S3**. Seismicity with $m_b \ge 5.0$ before and after great earthquakes with $M_w \ge 8.0$.

Additional file 2. It describes the section "Experiments 4 and 5 (dependence on the main shock magnitude)", Tables S1 and S2, and Figures S4 and S5.

Acknowledgements

We thank Naoki Uchida, Takeshi Sagiya, Andrew Michael, and an anonymous reviewer for making valuable comments and Glenn Pennycook, MSc, from Edanz Group (https://en-author-services.edanz.com/ac) for editing a draft of this manuscript. The source code of the program used for the spacetime ETAS model and stochastic declustering is available at http://bemlar.ism.ac.jp/zhuang/software.html. GMT-SYSTEM (Wessel and Smith 1991) was used for mapping data.

Authors' contributions

KK developed the detection method for seismic quiescence, created the program for analysis, and performed calculations. MN developed the statistical evaluation method for the seismic quiescence and was a major contributor in writing the manuscript. Both authors read and approved the final manuscript.

Funding

This study was partly supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, under "The Second Earthquake and Volcano Hazards Observation and Research Program (Earthquake and Volcano Hazard Reduction Research)". This study was also partly supported by the Tokio Marine Kagami Memorial Foundation.

Availability of data and materials

The datasets used and/or analyzed in the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Institute of Seismology and Volcanology, Faculty of Science, Hokkaido University, Sapporo, Japan. [2] Earthquake Research Institute, The University of Tokyo, Tokyo, Japan.

Received: 5 October 2020 Accepted: 10 April 2021
Published online: 27 April 2021

References

Aki K (1981) A probabilistic synthesis of precursory phenomena. In: Simpson DW, Richards PG (eds) Earthquake Prediction (Maurice Ewing Series 4). American Geophysical Union, Washington, D.C., pp 566–574

Bird P (2003) An updated digital model of plate boundaries. Geochem Geophys Geosyst 4(3):1027. https://doi.org/10.1029/2001GC000252

Hardebeck JL, Felzer KR, Michael AJ (2008) Improved tests reveal that the accelerating moment release hypothesis is statistically insignificant. J Geophys Res 113:B08310. https://doi.org/10.1029/2007JB005410

Inouye W (1965) On the seismicity in the epicentral region and its neighborhood before the Niigata earthquake. Kenshin-jiho (Quarterly Journal of Seismology) 29:139–144 (**in Japanese**)

Kanamori H (1981) The nature of seismicity patterns before large earthquakes. In: Simpson DW, Richards PG (eds) Earthquake Prediction (Maurice Ewing Series 4). American Geophysical Union, Washington D. C., pp 1–19

Katsumata K (2011) Precursory seismic quiescence before the $M_w = 8.3$ Tokachi-oki, Japan earthquake on 26 September 2003 revealed by a re-examined earthquake catalog. J Geophys Res 116:B10307. https://doi.org/10.1029/2010JB007964

Katsumata K (2017a) Long-term seismic quiescences and great earthquakes in and around the Japan subduction zone between 1975 and 2012. Pure Appl Geophys 174:2427–2442. https://doi.org/10.1007/s00024-016-1415-8

Katsumata K (2017b) Long-term seismic quiescence before shallow great earthquakes with M_w8.0 or larger between 1990 and 2014. In: Abstracts of JpGU-AGU Joint Meeting 2017, Makuhari Messe, Japan, 20–25 May 2017. https://confit.atlas.jp/guide/event/jpguagu2017/subject/SSS14-P06/date?cryptoId=

Keilis-Borok VI, Kossobokov VG (1990) Premonitory activation of earthquake flow: algorithm M8. Phys Earth Planet Inter 61:73–83. https://doi.org/10.1016/0031-9201(90)90096-G

Michael AJ (1997) Testing prediction methods: Earthquake clustering versus the Poisson model. Geophys Res Lett 24:1891–1894. https://doi.org/10.1029/97GL01928

Michael AJ (2014) How complete is the ISC-GEM global earthquake catalog? Bull Seismol Soc Am 104:1829–1837. https://doi.org/10.1785/0120130227

Mogi K (1969) Some features of recent seismic activity in and near Japan (2), Activity before and after great earthquakes. Bull Earthquake Res Inst Tokyo Univ 47:395–417

Mulargia F (1997) Retrospective validation of the time association of precursors. Geophys J Int 131:500–504. https://doi.org/10.1111/j.1365-246X.1997.tb06594.x

Nagao T, Takeuchi A, Nakamura K (2011) A new algorithm for the detection of seismic quiescence: introduction of the RTM algorithm, a modified RTL algorithm. Earth Planets Space 63:315–324. https://doi.org/10.5047/eps.2010.12.007

Nakatani M (2020) Evaluation of phenomena preceding earthquakes and earthquake predictability. J Disaster Res 15:112–143. https://doi.org/10.20965/jdr.2020.p0112

Ogata Y (1992) Detection of precursory relative quiescence before great earthquakes through a statistical model. J Geophys Res 97:19845–19871. https://doi.org/10.1029/92JB00708

Ogata Y (2001) Increased probability of large earthquakes near aftershock regions with relative quiescence. J Geophys Res 106:8729–8744. https://doi.org/10.1029/2000JB900400

Ohtake M, Matsumoto T, Latham GV (1977) Seismicity gap near Oaxaca, Southern Mexico as a probable precursor to a large earthquake. Pure Appl Geophys 115:375–386. https://doi.org/10.1007/BF01637115

Page MT, van der Elst N, Hardebeck J, Felzer K, Michael AJ (2016) Three ingredients for improved global aftershock forecasts: tectonic region, time-dependent catalog incompleteness, and intersequence variability. Bull Seismol Soc Am 106:2290–2301. https://doi.org/10.1785/0120160073

Reasenberg PA, Matthews MV (1988) Precursory seismic quiescence: a preliminary assessment of the hypothesis. Pure Appl Geophys 126:373–406. https://doi.org/10.1007/BF00879004

Scholz CH (2019) The mechanics of earthquakes and faulting, 3rd edn. Cambridge University Press, Cambridge

Sobolev GA, Tyupkin YS (1997) Low-seismicity precursors of large earthquakes in Kamchatka. Volcanol Seismol 18:433–446

Storchak DA, Harris J, Brown L, Lieser K, Shumba B, Verney R et al (2017) Rebuild of the Bulletin of the International Seismological Centre (ISC), part 1: 1964–1979. Geosci Lett 4:32. https://doi.org/10.1186/s40562-017-0098-z

Utsu T (1957) Magnitude of earthquakes and occurrence of their aftershocks. Zisin J Seismol Soc Jpn 10:35–45. https://doi.org/10.4294/zisin1948.10.1_35 (**in Japanese**)
Utsu T (1968) Seismic activity in Hokkaido and its vicinity. Geophys Bull Hokkaido Univ 20:51–75. https://doi.org/10.14943/gbhu.20.51 (**in Japanese**)
Wang Q, Schoenberg FP, Jackson DD (2010) Standard errors of parameter estimates in the ETAS model. Bull Seismol Soc Am 100:1989–2001. https://doi.org/10.1785/0120100001
Wessel P, Smith WHF (1991) Free software helps map and display data. Eos Trans AGU 72:445–446. https://doi.org/10.1029/90EO00319
Wiemer S, Wyss M (1994) Seismic quiescence before the Landers (M=7.5) and Big Bare (M=6.5) 1992 earthquakes. Bull Seismol Soc Am 84:900–916
Wiemer S, Wyss M (2000) Minimum magnitude of completeness in earthquake catalogs: examples from Alaska, the western United States, and Japan. Bull Seismol Soc Am 90:859–869. https://doi.org/10.1785/0119990114
Zechar JD, Jordan TH (2008) Testing alarm-based earthquake predictions. Geophys J Int 172:715–724. https://doi.org/10.1111/j.1365-246X.2007.03676.x
Zhuang J, Ogata Y, Vere-Jones D (2002) Stochastic declustering of space-time earthquake occurrences. J Am Stat Assoc 97:369–380. https://doi.org/10.1198/016214502760046925
Zhuang J, Chang CP, Ogata Y, Chen YI (2005) A study on the background and clustering seismicity in the Taiwan region by using point process models. J Geophys Res 110:B05S18. https://doi.org/10.1029/2004JB003157

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.