forecasting

MDPI

Article

Machine Learning Forecasting of Strong Subsequent Events in New Zealand Using the NESTORE Algorithm

Letizia Caravella and Stefania Gentili *

National Institute of Oceanography and Applied Geophysics—OGS, Via Treviso 55, 33100 Udine, Italy; lcaravella@ogs.it

* Correspondence: sgentili@ogs.it

Highlights

What are the main findings?

- The machine learning algorithm NESTORE accurately forecasts the likelihood of strong aftershocks in New Zealand using seismicity recorded within hours after a mainshock.
- Across the 1988–2025 dataset, NESTORE correctly classified 88% of clusters, including the Canterbury–Christchurch 2010–2011 sequence.

What are the implications of the main findings?

- NESTORE provides a promising approach under retrospective testing for near-real-time assessment of strong aftershock potential, supporting rapid response in one of the world's most active seismic regions.
- The method can in future enhance operational post-earthquake forecasting and contribute to risk-mitigation strategies in New Zealand.

Abstract

New Zealand, located along the boundary between the Pacific and Australian plates, is among the most seismically active regions in the world. In such an area, reliable short-term forecasting of strong aftershocks is essential for seismic risk mitigation. In this study, we apply NESTORE (NExt STrOng Related Earthquake), a machine learning probabilistic forecasting algorithm, to the New Zealand earthquake catalogue to evaluate the probability that a mainshock of magnitude M_m will be followed by an event of magnitude $\geq M_m - 1$ within a defined space–time window. NESTORE uses nine features describing early post-mainshock seismicity and outputs the probability that a cluster is Type A (i.e., containing a strong aftershock) or not (Type B). We assess performance using two testing strategies: chronological training–testing splits and k-fold cross-validation and refine the training set using the REPENESE outlier-detection procedure. The k-fold approach proves more robust than the chronological one, despite changes in catalogue characteristics over time. Eighteen hours after the mainshock, NESTORE correctly classified 88% of clusters (75% for Type A and 92% for Type B; Precision = 0.75). Notably, the highly destructive 2010–2011 Canterbury–Christchurch sequence was correctly identified as Type A. These findings support the applicability of NESTORE for short-term aftershock forecasting in New Zealand.

Keywords: machine learning algorithm; NESTORE; k-fold validation; outlier detection; aftershocks; New Zealand seismicity; seismicity clusters; forecasting

Academic Editor: Sonia Leva

Received: 25 November 2025
Revised: 19 January 2026
Accepted: 2 February 2026
Published: 12 February 2026

1. Introduction

Aftershocks following large earthquakes can substantially amplify the overall impact of the mainshock by imposing additional loads on already damaged structures, potentially leading to further degradation or collapse, and by disrupting rescue and recovery operations. Large mainshocks are commonly followed by multiple aftershocks with magnitudes ≥ 5. Consequently, reliable estimates of aftershock occurrence probabilities are a critical component of post-seismic decision-support systems for emergency response and early recovery planning [1].

New Zealand lies along the boundary between the Pacific and Australian tectonic plates, making it one of the most seismically active regions in the world. Several highly destructive earthquakes have occurred during the past century. Among the most devastating were the 1931 Mw 7.8 Hawke's Bay earthquake, which struck the North Island near Napier and Hastings and caused 256 fatalities [2–4], and the Canterbury–Christchurch earthquake sequence, which resulted in 185 deaths and widespread destruction in Christchurch [4,5]. The Canterbury–Christchurch sequence provides a well-documented example of why forecasting strong subsequent events is a crucial issue for seismic risk mitigation: the September 2010 Mw 7.1 Darfield (Canterbury) earthquake was followed by the February 2011 Mw 6.1 Christchurch earthquake [6]. Although the latter event was of lower magnitude, it produced far more severe impacts due to its proximity to the city and the vulnerability of buildings and infrastructure already weakened by the preceding mainshock [7].

One of the earliest systematic investigations into the expected magnitude of the largest aftershock was conducted by Båth [8]. By analyzing the magnitude difference Δm between a mainshock and its largest aftershock, Båth showed that the mean Δm is approximately 1.2, leading to the formulation of the well-known Båth's law: the magnitude difference between a mainshock and its strongest aftershock is, to first approximation, independent of the mainshock magnitude. However, the variability of Δm is substantial, and subsequent studies have highlighted that more accurate estimates require consideration of additional sequence-specific parameters (e.g., [9]).

Following a damaging earthquake, both the public and emergency managers are primarily concerned with the probability of further significant seismic activity. For instance, in California, the likelihood that a moderate earthquake will be followed within five days and 10 km by a larger event is approximately 5% [10,11]. Operational Earthquake Forecasting (OEF) [12,13] aims to quantify these probabilities using statistical models grounded in empirical observations, including ETAS [14,15] and STEP [16,17].

Many OEF models incorporate three fundamental properties of aftershock sequences: (a) the modified Omori law which governs the temporal decay of aftershock rates [18]; (b) the Utsu relation, which links aftershock productivity to mainshock magnitude [19]; and (c) the Gutenberg–Richter frequency–magnitude distribution [20]. Reasenberg and Jones [21] combined these relations into a time- and magnitude-dependent rate model parameterized by a, b, c, and p, which describe productivity, magnitude scaling, and temporal decay. Region-specific parameterizations have since been developed (e.g., [22]), although the Reasenberg–Jones formulation does not include explicit spatial forecasting. The Short-term Aftershock Probability (STEP) model [16] extends the Reasenberg–Jones approach by incorporating a spatial decay term proportional to r^{-2} and by translating expected aftershock rates into short-term seismic hazard estimates. STEP also accounts for cascading triggering from larger aftershocks and multiple mainshocks. The Epidemic-Type Aftershock Sequence (ETAS) model [14] is the most widely used statistical framework for short-term seismicity. ETAS explicitly models secondary triggering, estimated to account for about 50% of aftershocks [23], and allows model parameters to be updated as an aftershock sequence evolve. When combined with long-term fault-based hazard models,

ETAS yields hybrid systems such as the Third Uniform California Earthquake Rupture Forecast (UCERF3-ETAS) model [24], which has been implemented operationally following major earthquakes in California, including the 2019 Ridgecrest sequence [25,26].

In New Zealand, significant effort has been dedicated to developing both short- and medium-term forecasting methods [27]. Notable contributions include the STEP model [16] and the Every Earthquake a Precursor According to Scale (EEPAS) model [28]. Forecasting capabilities also include an ETAS system, the national ETAS–Harte implementation [29], and more recent hybrid approaches such as the Hybrid Forecast Tool (HFT) [30].

An alternative to statistical seismicity-rate models is the use of pattern-recognition approaches for forecasting strong aftershocks. These methods evaluate one or more features of early aftershock activity to determine whether their values indicate that a large subsequent event is likely. Typically, a threshold on Δm is defined, and the predictive capability of selected features is assessed with respect to whether the largest aftershock will have a magnitude deficit smaller than this threshold.

Early pattern-recognition studies by Vorobieva and Panza [31] and Vorobieva [32] introduced a set of diagnostic functions designed to identify premonitory phenomena associated with $\Delta m \leq 1$. Within these frameworks, the strongest forthcoming aftershock is treated as a critical point (a singularity), and—consistent with nonlinear dynamical systems theory—the approach to this critical state is expected to manifest as marked variations in observable seismic parameters, such as elevated activity rates or increased irregularity. Physically, these changes can be interpreted as an enhanced sensitivity of the lithosphere to tectonic loading during the preparatory phase of a strong aftershock. Gentili and Bressan [33] identified a relationship between Δm and the radiated energy of early aftershocks using a dataset of sequences from northeastern Italy (1977–2007). This work was expanded in Bressan et al. [34], which further explored connections among Δm, stress drop, and released energy. Other studies have focused on the parameters a and b of the Gutenberg–Richter relation. Shcherbakov [35] proposed a modified form of Båth's law based on extrapolating the Gutenberg–Richter distribution, in which the magnitude of the largest aftershock is estimated from the ratio a/b. Chan and Wu [9] applied approaches based on both b alone and a–b jointly to four Taiwanese sequences, finding that the performance of each method varied across sequences. Gulia and Wiemer [17] developed a forecasting method based on b-value variations aimed at identifying cases where $\Delta m \leq 0$. However, this approach is limited by the high instability of b estimates when only a small number of aftershocks are available. To address this issue, Gulia et al. [36] proposed an enhanced method using the b-positive formulation [37] to distinguish between a typical decaying aftershock sequence—characterized by $b > 1$—and sequences likely to culminate in a strong aftershock, which exhibit negative b-values.

In this study, we apply the NESTORE (NExt STrOng Related Earthquake) algorithm—a data-driven, supervised-learning approach for identifying seismic clusters likely to generate strong aftershocks—to New Zealand seismicity. As in Vorobieva and Panza [31] and Vorobieva [32], we adopt a threshold of $\Delta m = 1$ to distinguish cluster types: clusters with $\Delta m \leq 1$ are classified as Type A, while those with $\Delta m > 1$ are classified as Type B. The algorithm has previously been deployed successfully in several regions worldwide, including California, Italy, western Slovenia, Greece, and Japan [38–46]. Over time, NESTORE has been progressively refined to enhance robustness and adaptability across different seismic environments. In this work, we introduce further improvements to the performance evaluation framework.

The paper is structured as follows. Section 2 provides an overview of the seismotectonic setting of New Zealand and the catalogue used. Section 3 describes the NESTORE methodology, highlighting the modifications introduced in this study. Section 4 presents

the results obtained using two distinct testing strategies. Finally, Section 5 discusses and interprets the differences between these results, compares them with outcomes from previous regional applications, and explores possible physical explanations for some of the observed behaviors.

2. Analyzed Region

2.1. Seismotectonic

New Zealand is a residual, surficial portion of Zealandia [47], a continent formed in the Mesozoic along the subduction margin of Gondwana [48], which is now crossed by the Pacific–Australian plate boundary (Figure 1). This complex plate boundary system interests the whole country, except for the far north of the North Island and the far southeast of the South Island and has generated 28 earthquakes of magnitude equal or above 7 since 1900 [49]. The general relative motion of the Pacific plate to the Australian plate is in the range of 32–49 mm/yr [50], changes from north to south, and generates frequent earthquakes with depths ranging between 0 and 600 km [51].

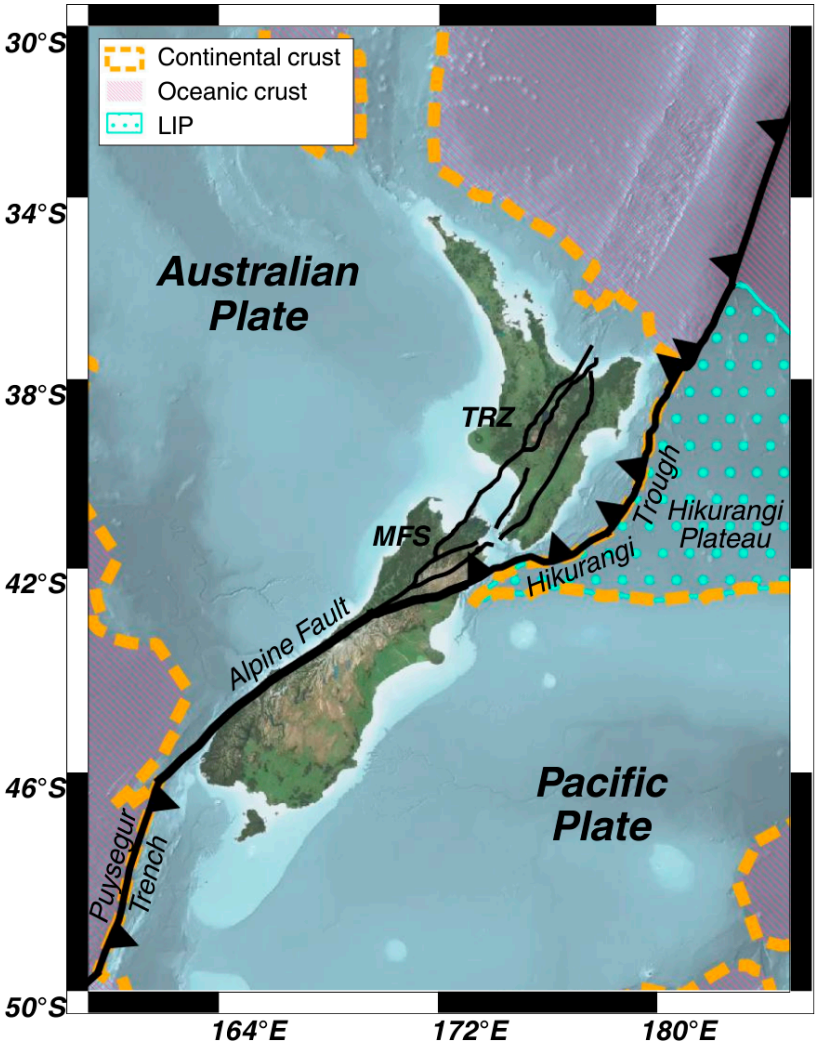

Figure 1. Present-day tectonic setting of New Zealand (Aotearoa) along the obliquely convergent Pacific–Australian plate boundary, highlighting the main tectonic provinces and structures discussed in the text. LIP = Large Igneous Province; TRZ = Taupō Rift Zone; MFS = Marlborough Fault System. The map was created using geospatial layers from the TRAMZ data © GNS Science 2020 [52]. Additional data sources listed in the Data Availability section.

One of the two existing subduction interfaces between these tectonic plates is the west-dipping Hikurangi subduction zone, located off the East Coast of the Northern Island, where the Hikurangi Plateau, an Early Cretaceous, intra-oceanic Large Igneous Province (LIP), subducts beneath the Australian plate. Along the interface, the Pacific slab generates earthquakes up to 300 km [53] but extends up to 900 km of depth [54]. The back-arc rifting deformation related to this margin is the Taupō Rift Zone (TRZ), a forearc opening basin, that in the Northern Island led to the formation since Quaternary of the active Taupō Volcanic Zone [55].

This boundary segment is followed in the northern part of the South Island by the intra-continental Marlborough transfer zone, a plate corner hosting the subparallel strike-slip Marlborough Fault System (MFS), which in November 2016 was capable of generating a Mw 7.8 earthquake in Kaikōura. This transfer zone accommodates subduction to the strike-slip Alpine fault, a 460 km long southeast-dipping continental transform located in the South Island that formed in the Late Oligocene; along with the Hikurangi Thrust, the Alpine Fault accommodates for the 80% of the relative plate motion which occurred during the Quaternary [56] and according to Sutherland et al. [57], it is realistic to expect Mw $\geq$ 8 earthquakes. Looking at a closer scale, the Alpine Fault consists of both thrust and strike-slip segments [54], of which the latter are parallel to the plate motion vector of the two tectonic plates [58].

Offshore from the South Island, the Alpine Fault becomes vertical and is followed by the 400 km long, southeast-dipping Puysegur subduction zone. Here, the subduction phase of the Australian plate beneath the Pacific place is still at an early stage, which began in Miocene–Quaternary [59], and generated only one Quaternary volcano (Solander Island) [60]. The earthquakes hypocenters can reach up to 150 km of depth [61].

2.2. Catalog

In the 1970s, the seismic network in New Zealand already comprised 26 stations [51]. The first broadband digital seismic stations were installed in the late 1990s, preceding the start of the GeoNet initiative in 2001 [62]. This program provides for the production and maintenance of New Zealand's national operational catalog (Earthquake Information Database, EID), which currently operates 120 short-period and 65 permanent broadband stations. Along with network improvements, different location methods and magnitude scales have been proposed over time (for more details, see [51]).

To analyze New Zealand seismicity, we concatenated data from two seismic catalogues (see Data Availability section for further details):

1. The revised earthquake catalogue for New Zealand, compiled in 2022 [63] as part of the revision of the New Zealand National Seismic Hazard Model [64] that comprehends events between 1492 and 2020 (we will refer it as "NZNSHM catalog"). In the NZNSHM catalog, we detected some gaps in the magnitude range before 1988, probably related to the merging of different catalogues; therefore, for our analysis we selected events from this year onwards (1988–2020).
2. The New Zealand earthquake catalogue available through the GeoNet Quake Search system [47] (henceforth "GeoNet catalog") that we fetched by using a Python code we developed called NZQuakeParser [65], available on GitHub: we obtained a list of earthquakes for the period from 2021 to 15 May 2025.

We focused on shallow earthquakes with a maximum hypocentral depth of 50 km, due to possible viscoelastic effects: deeper events may result in different seismicity characteristics for clusters (see, e.g., [42,43,45]). With this selection, the two catalogs count 295,376 and 65,897 events, for a total of about 362,000 events used for our analysis.

We then selected clusters whose mainshocks are onshore, to avoid errors in location and magnitude related to inaccuracies in event location and (consequently) in magnitude assessment. The analysis area we selected is highlighted in red in Figure 2.

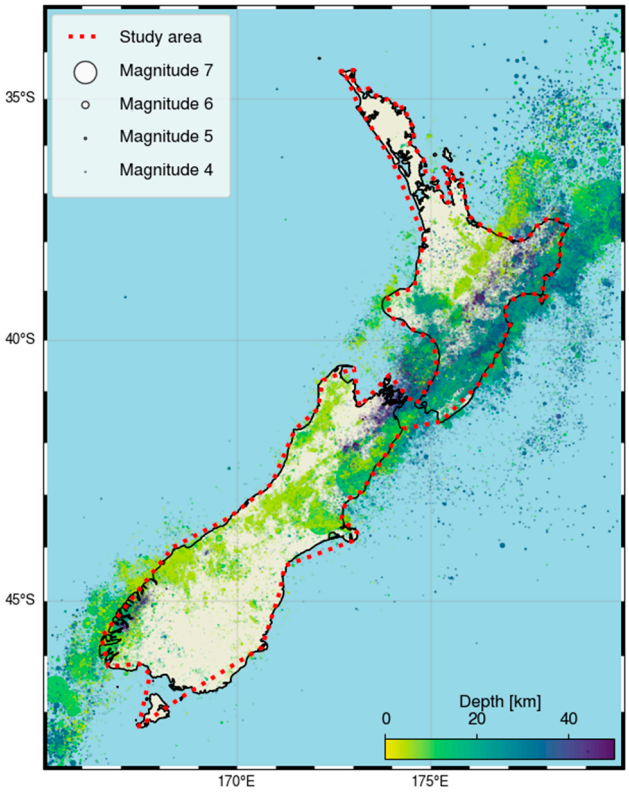

Figure 2. Earthquake locations by magnitude and depth of the events used in this study (1988–2025). The dotted line marks the NESTORE study area.

Regarding the magnitude scale adopted, both catalogues report earthquake magnitudes in various units. The magnitude (Mstd) proposed as preferred by Christophersen et al. [51] approximates Mw but is not yet available in GeoNet Quake Search for recent events. Since NESTORE has mainly been applied to local magnitudes, and to maintain consistency within the dataset, we used ML.

The NZNSHM catalog lists ML magnitude for almost all events; starting with 1988, the value of ML is missing in a very small percentage of cases (0.002%), mainly corresponding to offshore events. For the GeoNet catalog, we derived ML where missing using orthogonal regressions on other magnitude scales. In order to have a large and revised dataset for the regression, we used NZNSHM data. Figure 3 shows the data we used and their fitting line; Table 1 shows the corresponding regression coefficients. The residuals of each regression are shown in Figure S1 in the Electronic Supplement. Overall, the data of the GeoNet catalog from 2021 to 2025 comprised approximately 66,000 earthquakes, with ML available in 96% of cases. When ML was not available, we used the preferred magnitude suggested therein. For about 1000 events, the preferred magnitude was MLv, the local magnitude estimated from the vertical component of the signal instead of the horizontal components. For approximately 1800 events, the preferred magnitude was what the catalogue lists as "M", which is a combination of MLv, moment magnitude Mw(mB) estimated from body wave magnitude, and the number of stations supplying Mw(mB) (see https://www.geonet.org.nz/data/supplementary/earthquake_location, accessed on 20 November 2025, for further details). Both M and MLv fit the data approximately 1:1. In only two cases was the preferred magnitude in the GeoNet catalog Mw(mB), and in one

case it was Mw. However, these three events are offshore, far from the coast, and are not considered in our cluster analysis.

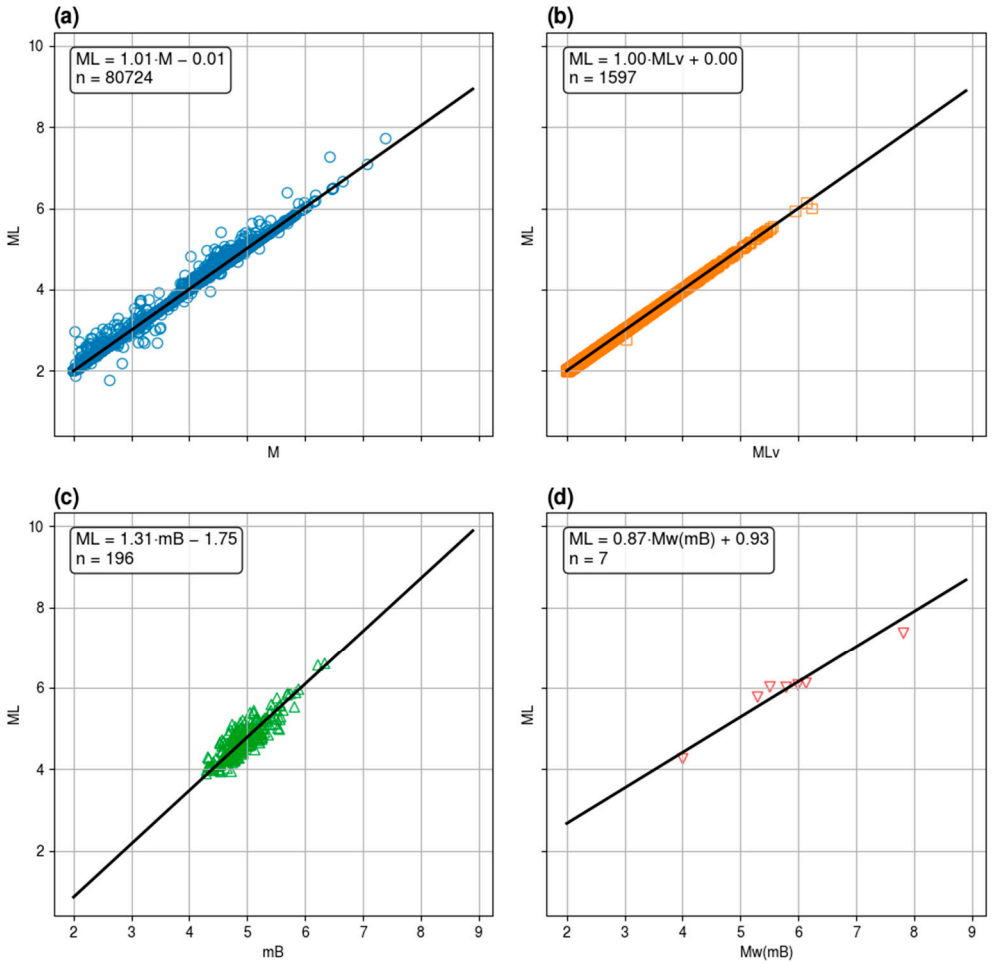

Figure 3. (**a**–**d**) Orthogonal regression between magnitude scales listed in GeoNet Quake Search system and ML from the NZNSHM catalogue for New Zealand [63]. The magnitude scales are ML = local magnitude; Mw(mB) = moment magnitude derived from body waves; MLv = local magnitude calculated over vertical components; M = combination of MLv, Mw(mB) estimated from body wave magnitude, and the number of stations supplying Mw(mB). For each regression, the linear fit and the number of events is listed; for more details about the residuals see Figure S1 in the Electronic Supplement. More details about the magnitude scales used in the GeoNet catalog are available at https://www.geonet.org.nz/data/supplementary/earthquake_location, accessed on 20 November 2025.

Table 1. Orthogonal regressions parameters.

Magnitude	Slope	Std Error	Intercept	Std Error
M	1.006	0.0001	−0.0147	0.0003
MLv	0.9994	0.0003	0.001	0.001
mB	1.30	0.05	−1.7	0.2
Mw(mB)	0.87	0.08	0.9	0.4

To obtain reliable results that are not affected by artefacts from the network's inability to detect smaller events, it is necessary to use data above the completeness magnitude, which is the magnitude above which all earthquakes are detected. Figure 4 shows how completeness magnitude varies spatially. The completeness magnitude was estimated using the maximum curvature method [66] plus 0.2 to account for the method's tendency

to underestimate completeness magnitude (e.g., [45]). The analysis was performed using ZMAP software [67]. After applying a window-based declustering (radius as proposed by Christophersen [68] and time window as proposed by the Gardner and Knopoff law [69]; see Section 3 for more details). Within the region highlighted in Figure 2 by the dotted red line, the completeness magnitude Mc is generally $\leq$3; the two exceptions are due to border effects: in the south, because the cell contains offshore events, in the north, because seismicity is low and the algorithm automatically extends the search area to offshore events. We therefore selected Mc = 3 as the minimum magnitude for our analysis across the entire area. Our results are in good agreement with Zaccagnino et al. [70]. Within this area, only one event was missing ML in the NZNSHM catalogue: an earthquake with Mw = 7.8 on 15 July 2009 in the southwestern part of the analyzed area, which had a too close in time strongest aftershock (~20 min) to be analyzed by NESTORE (NESTORE performs the first analysis 6 h after the mainshock to ensure reliable statistical information).

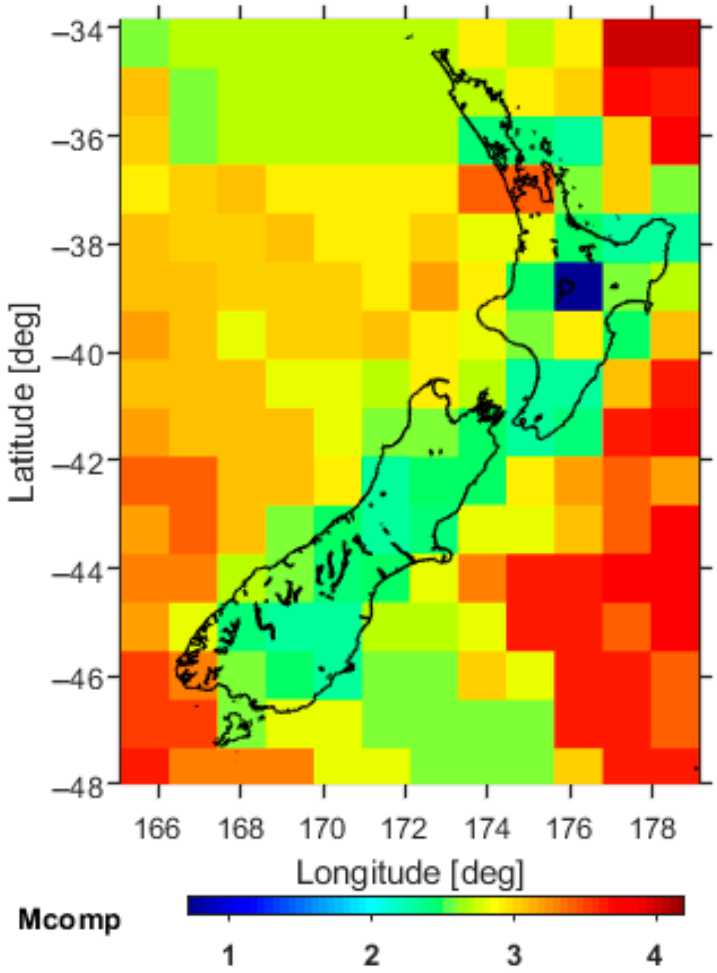

Figure 4. Completeness magnitude of the declustered catalogue in space. The map was made using ZMAP software [67].

3. Methods

NESTORE is a supervised machine learning based algorithm that aims to forecast the class of clusters of seismicity defined by the difference in magnitude between the mainshock and the strongest aftershock: if this difference is equal to or less than 1, i.e., $M_m - M_{aft} \leq 1$, the cluster is defined as "Type A", otherwise as "Type B".

NESTORE algorithm evolved during time. NESTOREv1.0, the MATLAB R2018a code available on GitHub (see Data Availability section) was described in Gentili et al. [41].

The code is divided into four modules that can be executed independently of each other: cluster identification module, training module, testing module, near-real-time module. In this paper, we describe the first three modules that were used in this work (Sections 3.1, 3.2 and 3.4). Gentili et al. [45] proposed a further algorithm, named REPENESE, added to the training module to remove the outliers from the training set and obtain more stable results (see Section 3.3). The last subsection (Section 3.5) describes the last innovations we introduced in this paper for a more robust estimate of the performance.

The cluster identification module extracts from the seismic catalogue earthquake clusters for use by training and testing modules. Both modules extract features from a selected subset of clusters, along with the corresponding class information ("Type A" or "Type B"). The features are unnormalized numerical values, ordered in vectors according to the clusters they reference. The training module uses the feature values and class information to determine output training parameters, which are stored in files. The testing module uses these output parameters to classify the assigned clusters, outputting a list of clusters and their classifications. The performance of the testing is evaluated by comparing the classifications to the actual classes.

3.1. Cluster Identification

Cluster identification is the first module of NESTORE, in which the seismic clusters are selected and stored to be available for the following steps. Choosing the best method for this task is not trivial, especially when the clusters are close to each other both in time and space, which is very often the case for earthquakes. One of the simplest yet widely used methods, which has been successfully applied in most of the previous works on NESTORE [38–40,42,43], is the so-called window-based method. In this approach, earthquakes belonging to clusters are selected within circular areas centered on the largest events. The radius of each circle depends on the magnitude M_m of the corresponding large event. Similarly, the temporal extent of the cluster depends on the magnitude M_m of the corresponding large event. If one of these events is already within the spatiotemporal domain of a previous event, the clusters are merged. Several authors have proposed laws for evaluating region-specific radius and time window length (e.g., [33,69,71–74]). For our application in New Zealand, we have found that the law for the radius proposed by Christophersen [68], which has a dependence on magnitude similar to that proposed by Utsu and Seki [75] and Kagan [72], best define the clusters; similarly, the best time extent is obtained with the Gardner and Knopoff law [69]. Following the approach of Anyfadi et al. [42], we selected these laws by comparing the duration and extent of manually estimated clusters with functions proposed in the literature. The equations we adopted are therefore the following:

$$r = 10^{0.51 * M_m - 1.65} \text{ [km]} \tag{1}$$

$$\begin{aligned} t &= 10^{0.032 * M_m + 2.7389}, \quad if\ M_m \geq 6.5 \text{ [days]} \\ t &= 10^{0.5409 * M_m - 0.547}, \quad else \end{aligned} \tag{2}$$

where M_m is the magnitude of the strong event generating the cluster, r is the radius of the cluster in kilometers, and t is its duration in days.

To ensure the functionality of NESTORE, the completeness magnitude within each cluster must be greater than $M_m - 2$. Having selected a completeness magnitude of 3 for the entire region, we have chosen 5 as the magnitude threshold for M_m. Note that, since our method is developed to work in semi-real time, the strong earthquake generating the cluster may not always be the mainshock, but it may be a strong foreshock of a stronger following mainshock. To avoid confusion, in NESTORE, the strong event is called "operative mainshock" (o-mainshock for simplicity). For further details on the cluster

identification code, the software is available online as part of the NESTORE package (see Data Availability Statement), derived from the declustering component of the ZMAP code [67]).

3.2. Training

In the training module, an input dataset with known classes of seismic clusters is used to train the classifier. Unlike versions of NESTORE prior to 2025, the dataset selected for training is preliminarily filtered to remove outliers using the REPENESE algorithm (see Section 3.3). The Training module is divided into four sub-blocks.

3.2.1. Feature Extraction

Nine seismic features are calculated within each cluster of seismicity, based on the cumulative source area, the energy released, the magnitude, and the number and spatial distribution of events (see Table 2). The features are computed at increasing time intervals T_i starting from one minute after the mainshock (for an analysis of the reason of this interval see [40]) up to 0.25, 0.50, 0.75, 1, 2, 3, 4, 5, and 7 days after. For simplicity, we will refer to T_i = x h as the interval ending x hours after the mainshock.

Table 2. Seismicity features adopted by NESTORE.

Feature	Definition	Type
Number of events (N_2)	$N_2(i) = \sum_i H[m_i - (M_m - 2)]$ (1)	Space–time distribution of events
Linear Concentration of Events (Z)	$Z(i) = \frac{mean\left(10^{0.69m_i - 3.22}\right)}{mean\left(r_{ij}\right)}$ (2)	
Normalized event source area (S)	$S(i) = \sum_i 10^{(m_i - M_m)}$	Source area and Magnitude trends over time
Cumulative deviation of S from the long-term trend on increasing length windows (SLCum)	$SLCum(i) = \sum_i abs[S(t_i) - S(t_{i-1})]\frac{i \cdot dt}{(i-1)\cdot dt}$	
Cumulative deviation of S from the long-term trend on sliding windows (SLCum2)	$\mathrm{SLCum2(i)} = \sum_i \mathrm{abs[S([s_1 + (i-1)\cdot dt, s_1 + i\cdot dt]) - S([s_1 + (i-1)\cdot dt, s_1 + (i-1)\cdot dt + d\tau])}\frac{dt}{d\tau}]$	
Cumulative variation of magnitude from event to event (Vm)	$V_m(i) = \sum_i \lvert m_i - m_{i-1} \rvert$	
Normalized radiated Energy (Q)	$Q(i) = \frac{\sum_i E_i}{E_m}$ (3)	Energy trends over time
Cumulative deviation of Q from the long-term trend on increasing length windows (QLcum)	$QLCum(i) = \sum_i abs[Q(t_i) - Q(t_{i-1})]\frac{i \cdot dt}{(i-1)\cdot dt}$	
Cumulative deviation of Q from the long-term trend on sliding windows (QLCum2)	$QLCum2(i) = \sum_i abs[Q([s_1 + (i-1)\cdot dt, s_1 + i\cdot dt]) - Q([s_1 + (i-1)\cdot dt, s_1 + (i-1)\cdot dt + d\tau])\frac{dt}{d\tau}]$	

(1) H = heaviside step function; m_i = magnitude of the ith event. (2) r_{ij} = distance between the ith and jth events. (3) E_m = mainshock energy; E_i = the energy of the ith event.

Note that we analyze Type A clusters until the first strong aftershock with $M \geq M_m - 1$ occurs, because after that the class is already defined (so further testing is not useful), and the subsequent seismicity results from the complex superposition of the aftershocks from both the mainshock and the strongest aftershock, potentially causing anomalous behavior. Therefore, the cardinality of class A changes depending on the interval T_i and decreases over longer intervals.

3.2.2. One-Node Decision Tree

Once the features are extracted, a decision tree is trained for each feature at every time interval to discriminate between the A and B class clusters. Small databases, such as in our case, are likely to experience overfitting; therefore, the number of parameters should be kept small. NESTORE uses a one-node tree so that a feature is considered successful in class discrimination if most Type A cluster feature values are greater than the threshold and most of Type B are lower. If no suitable value can be found, the threshold is set to NaN and the corresponding feature for that time interval is discarded.

3.2.3. Selection of Good Interval

The threshold performances (and the corresponding time intervals) are evaluated using the Leave One Out (LOO) method. This special case of k-fold cross-validation is suitable for small datasets, as each sample is tested with a classifier trained on all the other samples in the dataset. The reliability of the classifiers is then expressed in terms of Accuracy, Precision, Recall, and Informedness. The first three parameters range from 0 to 1, and the last one from -1 to 1. A feature calculated over a given time interval T_i is considered relevant for classification if (1) Accuracy, Precision, and Recall are greater than 0.5; (2) Accuracy is greater than or equal to the accuracy achieved by always predicting the most frequent class; (3) Informedness is greater than 0. The class distribution in our case study is consistent with previous applications of NESTORE: class B clusters are more frequent than class A [38,39,42,43,45,46], and therefore the second condition is crucial to account for class imbalance. The interval in which the feature can be used (the "good interval") begins at T_i = s1, the first interval where the feature is considered relevant, and ends at T_i = s2, where s2 is the time interval with the highest Informedness. The performances are illustrated using Precision–Recall and Receiver Operating Characteristic (ROC) diagrams.

3.2.4. Inheritance, Validation and Probability Estimation

For time intervals larger than s2, the feature is not removed from the classification, but both the feature values and the corresponding thresholds are inherited from the s2 time interval. To verify that the inheritance process does not add errors to the classification, the LOO method is applied once more for time intervals larger than s2 to compute the percentage of Type A correctly classified (hit rate) and the percentage of misclassified Type B (false alarm rate). The intervals T_i in which hit rate < false alarm rate are discarded, the corresponding thresholds are set to NaN, and the feature is not considered for the classification. The training ends when the time intervals of usability are validated. Then, for each feature and time interval, the probability of being a Type A cluster above and below the threshold is estimated. The training output consists of four vectors: (1) the threshold vector for each good interval; (2) the validity vector, providing the time ranges of validity for each feature; (3) the probability vector, which gives, for each feature of every seismic cluster, the probability of being Type A above and below the threshold; (4) the NAB vector, i.e., the number of class A and class B clusters for each T_i.

3.3. REPENESE

As previously mentioned, our dataset is limited in the number of samples and imbalanced in class distribution, as Type B clusters occur much more frequently than Type A clusters. The same issue was identified in earlier applications of NESTORE, prompting Gentili et al. [45] to develop a specific algorithm for outlier detection. The goal is to better refine the training dataset to ensure more reliable thresholds and a clearer distinction between the classes. The algorithm was developed based on the following considerations

for outlier selection: (1) it uses only relevant features; (2) it accounts for the class imbalance; (3) it replaces the commonly used concept of "cluster centroid distance" in outlier detection algorithms with a simpler and more appropriate distinction between "under" and "above" the threshold. In this way, features with values far from the centroid, but correctly assigned, are still considered. The resulting algorithm was named REPENESE and proceeds through the following steps:

- RE: RElevant features. Features are relevant if there exists a threshold for which (1) the true-positive rate is greater than 0.5, (2) the false-positive rate is less than 0.5, and (3) Precision is greater than 0.5.
- PE: class imbalance PErcentage. The probabilities of being Type A (PA) or Type B (PB) for a cluster are calculated from the percentages of Type A and Type B samples in the training set, thus providing an assessment of the asymmetry between the classes.
- NE: NEighborhood detection. For each relevant feature, the values are sorted. To determine whether a value is an outlier, its neighborhood is defined by considering the closest N1 larger and smaller values for the Type A and Type B class, respectively. The dimension of the neighborhood, Sn, is the total number of observations within it, while N is the number of samples of the class considered.
- SE: SElection. The sample is a possible outlier if $N \leq P \cdot Sn$, with P being PA or PB depending on the sample class. Samples are considered to be true outliers if they are consistently outliers across all relevant features and time periods for which they are analyzed.

For this work, REPENESE has been automatized through the implementation of an independent module.

3.4. Testing

The aim of the testing module is to forecast the class of clusters from an independent test set by using the information gained in the training phase. The new catalog undergoes a similar procedure to the one described so far. After the cluster identification phase, the features for each cluster are extracted at each time step T_i specified in the validity vector. The threshold vector obtained after the outlier removal made by REPENESE is compared with the calculated features. In accordance with the existing thresholds, the probability vector (referred instead to the whole training dataset, outliers included) is used to evaluate the probability that the cluster class is A if the nth feature is present at time Ti:

$$p_{n,i} = P(A|F_{n,i}) \tag{3}$$

Gentili and Di Giovambattista [39] have proposed an approach based on Bayes' Theorem to combine the feature-based probabilities and estimate for each given cluster the probability at each time T_i that it is of Type A:

$$P_i\Big(A \mid F_{1,i} \ldots F_{N,i}\Big) = \frac{[N(B)_i]^{N-1}\prod_{n=1}^{N} p_{n,i}}{[N(B)_i]^{N-1}\prod_{n=1}^{N} p_{n,i} + [N(A)_i]^{N-1}\prod_{n=1}^{N}(1-p_{n,i})} \tag{4}$$

$N(A)_i$ and $N(B)_i$ are retrieved from the NAB vector. By including the cardinality of each class in the dataset, the imbalance between the classes is taken into account. For the final evaluation of the quality of the algorithm result, the actual known class of the clusters is compared with the classification output. The ROC and Precision–Recall diagrams are re-generated to better visualize the performance of the algorithm. Like in other NESTORE applications, we set the class A if $Prob(A) \geq 0.5$.

3.5. NESTORE Performances Validation

In previous versions of NESTORE, the dataset was divided into two distinct subsets—a training set and a test set—to enable independent evaluation of model performance. The division was based on the year of occurrence of each cluster's mainshock. This setup allowed for a retrospective forecasting procedure: the model was tested as if performing a real-time forecast of the cluster class, and, since the actual class was known afterwards, its predictive performance could be quantitatively assessed.

In this study, we adopt an additional evaluation strategy. Rather than assigning only the most recent data to the test set, we use a stratified k-fold cross-validation approach [76]. In k-fold cross-validation, the dataset instances (in this case, the clusters) are randomly divided into k groups, or folds, of approximately equal size. Each fold is used once as a validation set, while the remaining $k - 1$ folds are used for training. The overall performance of the method is then calculated as the mean performance across all k validation folds.

Compared to a simple train–test split, this approach provides a more reliable performance estimate, as it uses all available clusters for both training and validation, and avoids overfitting, since the training and test sets are disjointed for each iteration. The leave-one-out (LOO) method discussed in Section 3.2.3 is a special case of k-fold cross-validation where k equals the total number of instances in the dataset.

The stratified variant of k-fold cross-validation further improves the scheme of the k-fold cross-validation by preserving the original class distribution within each fold. This is particularly advantageous for classification problems involving imbalanced datasets, as it reduces the risk of biased performance estimates due to underrepresented classes.

Overall, this cross-validation approach yields a more robust and comprehensive assessment of model performance than a simple time-based data split, as it leverages the entire dataset—provided that the time of cluster occurrence does not affect the data quality.

A further analysis has been introduced in this latest version of NESTORE (see also Gentili et al. [44]), namely the probability α of obtaining h or more hits (Type A clusters correctly classified) by chance. This probability was calculated according to Zechar [77] as

$$\alpha = \sum_{i=h}^{N} \binom{N}{i} \tau^i (1-\tau)^{N-i} \quad (5)$$

where τ represents the fraction of the "space" occupied by alarms within the testing region. Here, "space" should not be interpreted in a purely geometrical sense (e.g., in kilometers or degrees). In the analyses conducted within the framework of the Collaboratory for the Study of Earthquake Predictability (CSEP) testing centers [78,79], the "space" coincides with the testing region that is defined as the Cartesian product of the binned magnitude range of interest and the binned spatial domain of interest.

In this study, τ is defined as the ratio between the number of clusters classified as Type A and the total number of clusters. Although in some applications related to mainshock forecasting the measure of time–space may not be uniquely defined [76], in this context, τ is uniquely determined, thus avoiding ambiguous results. In this formulation, N denotes the number of observed Type A clusters. A smaller value of α corresponds to an alarm with higher skill.

4. Results

To analyze NESTORE on New Zealand seismicity, we adopted three approaches:

- "Chronological", i.e., time-based analysis: training was performed using clusters from 1988 to 2013, and testing using clusters from 2014 to 2025.
- "K-fold", i.e., based on stratified k-fold analysis. We used k = 3.
- "Autotest", i.e., an analysis in which the same dataset is used for training and testing.

Table 3 shows the number of training and testing instances in the different approaches. The dataset appears unbalanced, as the cardinality of class B is approximately three times that of class A.

Table 3. Training and testing datasets selected following the chronological, the k-fold cross-validation method, and the autotest approach. The actual training sets are the ones without outliers.

Method	Dataset							
	Training				Testing			
	Subset	N Clus	N A	N B	Subset	N Clus	N A	N B
Chronological	1988–2013	37	10	27	2014–2025	15	4	11
	Without outliers	33	6	27				
k-fold, k = 3	$k\text{-}f_2 + k\text{-}f_3$	35	9	26	$k\text{-}f_1$	17	5	12
	Without outliers	32	6	26				
	$k\text{-}f_1 + k\text{-}f_3$	34	9	25	$k\text{-}f_2$	18	5	13
	Without outliers	31	6	25				
	$k\text{-}f_1 + k\text{-}f_2$	35	10	25	$k\text{-}f_3$	17	4	13
	Without outliers	33	8	25				
Autotest	All	52	14	38	All	52	14	38
	Without outliers	48	10	38				

Regarding the chronological and k-fold methods, free parameters (the starting year of testing and the number of folds k) need to be chosen to maximize the reliability of results. For small datasets, the quality of performance evaluation improves as the number of available instances of the two classes in the test set increases, because the data distribution is more accurately represented. However, for the same reason, the classifier's performance also improves as the number of instances of the two classes in the training set increases. Since the number of instances is low and the training and test sets are disjointed, there is a trade-off between the size of the training and test sets. NESTORE performs testing if at least three instances of the class with lower cardinality are available. With a total of 14 Type A clusters, we selected a three-fold stratified analysis. In this way, each fold contains four or five Type A instances. Similarly, in the chronological approach, we chose the testing starting year to ensure at least four Type A instances.

The autotest was not considered for assessing classification performance, as using the same set for training and testing may conceal potential overfitting due to the small training set. In other words, if parameter evaluation is overly influenced by noisy or anomalous data, testing on the same dataset would not reveal the issue, because the same noisy instances are present in the test set as well. In contrast, the other proposed approaches (k-fold, chronological) use independent test sets, so potential overfitting effects—which are even more likely with smaller training sets—are revealed during performance evaluation. Therefore, autotest should not be considered, as it may provide an overly "optimistic" estimate of the algorithm's performance. The outlier filtering performed by REPENESE partially mitigates this effect, because outliers do not contribute to threshold evaluation in training, while they are in the test set. The autotest testing should be regarded as an analysis of the dataset's internal consistency, because if it fails to predict a cluster class; this indicates that this cluster is an outlier whose behavior is so anomalous compared to the other clusters of its class in the training set that it cannot be assimilated to the others, even by an ad hoc choice of features.

Figure 5 shows the map of the epicenters of o-mainshocks of the clusters (red for Type A clusters, blue for Type B clusters). Crosses indicate the clusters that are misclassified in all time periods of the analysis. While the chronological analysis does not reveal any

misclassification, the k-fold analysis shows six clusters misclassified in all time intervals, of which four are Type A and two are Type B. It is interesting to compare this result with that of the autotest. In this case, five clusters remain misclassified, highlighting how different they are from the corresponding Type A and Type B populations, such that NESTORE cannot classify them in the corresponding class, even when specifically trained on it. Four of these five clusters were detected as outliers by REPENESE and not used in the following training. The 23 June 2005 and 26 July 2012 clusters, misclassified by the k-fold approach, were not recognized as outliers by REPENESE and were included in the training with all available data; the autotest training included the clusters and attempted to adjust the thresholds according to those cluster classifications. The testing shows that the obtained thresholds were able to correctly classify the 2012 cluster but not the 2005 one. This correct result may be due to overfitting, which tunes the threshold to include that data, or to sparse data near the threshold that does not allow for precise threshold determination in the smaller k-fold training set. In order to understand which of these hypotheses is correct, a large independent dataset will be necessary, analyzing the seismicity in the following years.

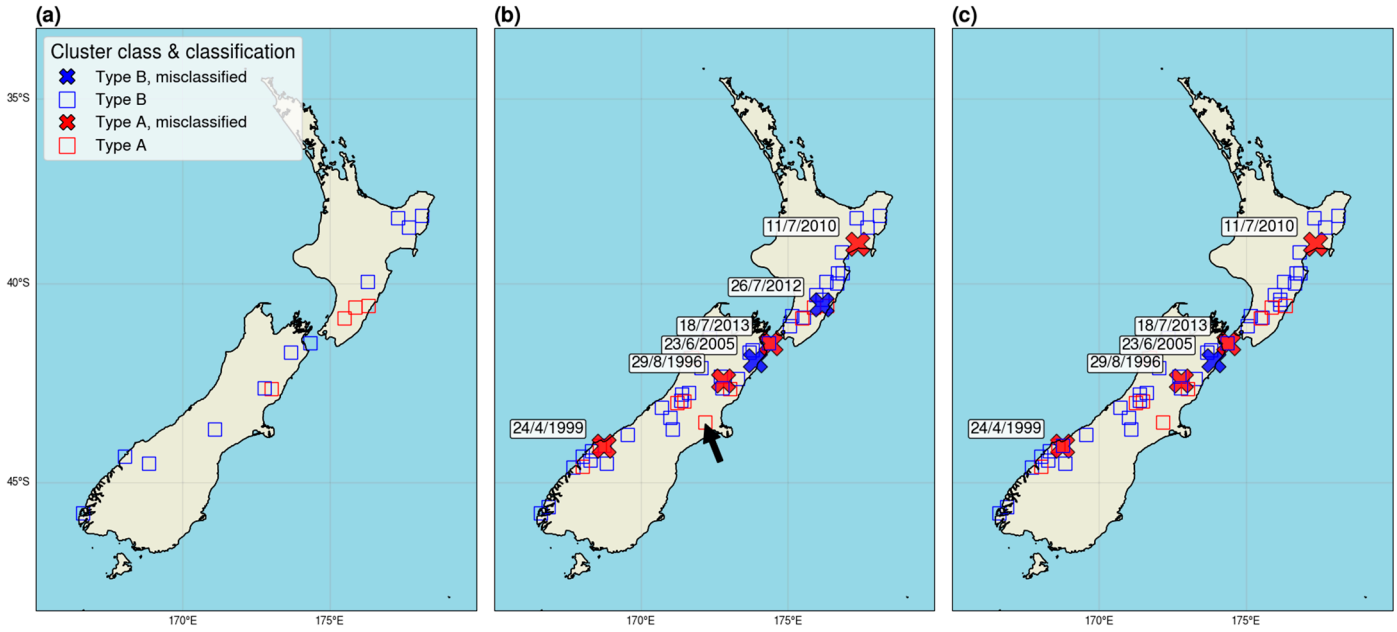

Figure 5. Test set in the three analyzed cases: (**a**) chronological; (**b**) k-fold; (**c**) autotest. Each point represents the cluster mainshock and is colored according to the NESTORE classification (Type A: red; Type B: blue). Squares indicate clusters correctly classified in at least one time period; filled crosses indicate clusters misclassified in all time periods. For the misclassified clusters, the mainshock date is also indicated. The black arrow in (**b**) marks the September 2010 Canterbury earthquake.

The black arrow shows the 2010–2011 Canterbury–Christchurch cluster, that was correctly classified by NESTORE as a Type A cluster.

Figure 6 shows the performance of the method in the three cases, using the ROC diagram. Binary classifier analyses are based on two classes: the positive and the negative class. In our analysis, we define class A as positive and class B as negative. The ROC graph displays on the horizontal axis the normalized (to 1) percentage of negative class instances wrongly classified (False Positive Rate—FPR); in our case, this corresponds to the normalized percentage of B clusters wrongly classified. The y axis shows the normalized percentage of positive instances correctly classified. This normalized percentage is called true-positive rate (TPR), or Recall, or hit rate. In our case, this is the normalized percentage of A clusters correctly classified. The random classifier corresponds to the diagonal dashed line. The ideal classifier corresponds to TPR = 1 and FPR = 0 (upper

left corner) and the closer the classifier is to the point (0,1), the higher its performance. Unsurprisingly, the autotest (Figure 6c) supplies generally better results than the k-fold, with recall in the range [0.64–0.75] depending on the time period, and a very low value of false-positive rate [0.03, 0.08]. The best results in the autotest were obtained for T_i = 18 h or 1 day (superimposed symbols). These results should be considered as an upper limit for correct performance estimation, as they are likely too optimistic due to possible overfitting. However, even with the k-fold analysis (Figure 6b), the performance remains high, well above the dashed line, corresponding to a random response. In the k-fold case, the best performance was obtained for time interval T_i = 0.5 days (12 h). Specifically, for T_i = 12 h, the Recall is 0.77 and the FPR is 0.08, due to a correct classification of one more Type A cluster respect to the autotest.

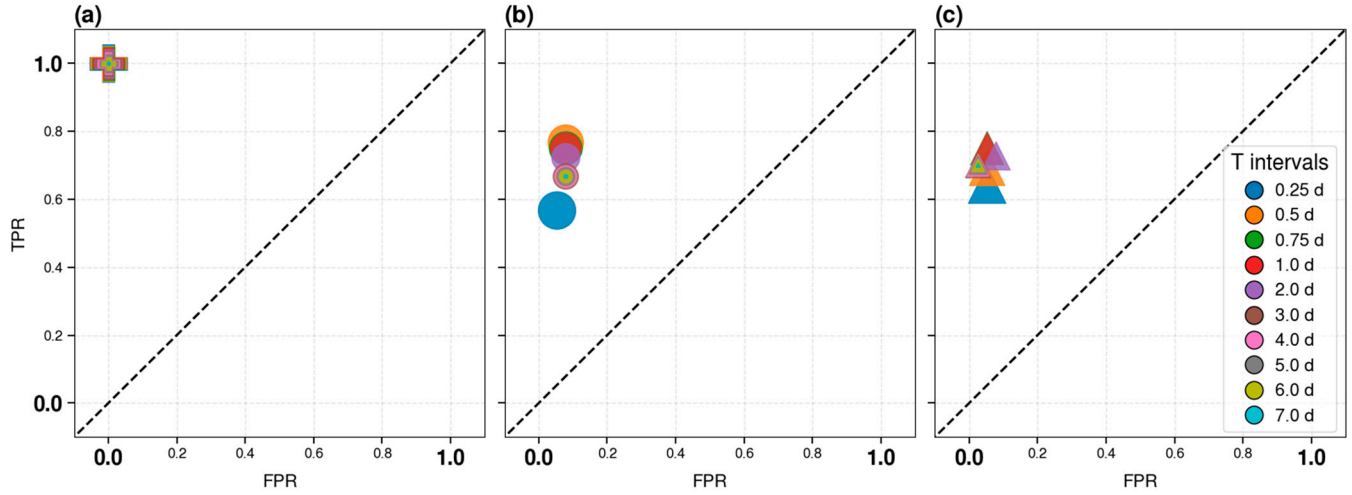

Figure 6. ROC diagram in the (**a**) chronological, (**b**) k-fold, and (**c**) autotest approaches. The time interval end is expressed as a fraction of a day. For the k-fold approach (**b**), we plotted the average value across the folds.

Note that for long time intervals, greater than or equal to 3 days, the FPR is minimal in the autotest case but not in the k-fold case. We hypothesize that this is due to the inability of autotest to detect overfitting. For time periods longer than 3 days, all thresholds are inherited, and differences in performance may be due only to the decrease in Type A clusters in the dataset.

In more detail, comparing Figures 5 and 6 and the output classification of individual clusters for T_i = 18 h, we can see that k-fold and autotest provide similar results, due to the correct classification of the same 9 out of 12 A clusters. The only difference is that in the autotest case, 2 B clusters were classified as A, while in the k-fold case, 3 were classified as Type A, resulting in FPRs of 0.05 and 0.08, respectively. Comparing this result with Figure 5, the four misclassified Type A clusters are the ones identified as outliers for all time periods (red crosses in both Figure 5b,c). Conversely, the B clusters misclassified for T_i = 0.75 days are not all shown in Figure 5c, because one of the two (with o-mainshock on 26 July 2012, ML = 5) was correctly classified at Ti $\geq$ 3 days.

Figure 6a shows the performance of chronological analysis. In this case, the performances are better than the others. The chronological approach achieves ideal classifier performance 6 h (0.25 days) after the o-mainshock and maintains the same classification for all the following time intervals. The reason for such good performance is that the testing phase starts in 2014, while the last outlier cluster in chronological order is in 2013. We will discuss this result in more detail in Sections 5 and 6.

To validate these analyses, and to have a quantitative measure of the statistical significance of the observed performance, we evaluate the probability α of obtaining h-hits by chance, according to Equation (5). Table 4 shows the probability in the three cases for $T_i \leq 0.75$ days, which is the shortest time interval for which the three approaches provide the best results (as in the previous analysis, autotest results should be considered as a lower limit for α). In all cases, this probability is less than 1%, showing the good performance of the method. Note that the chronological approach α has higher values due to the smaller dataset involved.

Table 4. Values of α for each of the three approaches at T_i = 6,12,18 h.

Method	$\alpha \times 100\%$		
	6 h	**12 h**	**18 h**
Chronological	0.5057	0.5057	0.9840
K-fold	0.1839	0.0150	0.0280
Autotest	0.0594	0.0306	0.0137

The Precision–Recall graphs in Figure 7 show Recall on the horizontal axis, which coincides with the TPR shown in Figure 6, and Precision on the vertical axis; Precision, in our case, is the normalized percentage of clusters classified as A that are indeed A. Precision is a very important evaluation parameter in an unbalanced dataset because, if the negative class is larger than the positive one, even a small percentage of misclassified negative instances may result in a large percentage of events wrongly classified as positive. The random classifier corresponds to the normalized percentage of positives in the dataset. As the percentage of A in our dataset decreases over time, we show the line for the smallest time interval and the largest percentage, as this is the most conservative approach. The ideal performance for the Precision–Recall graph is Precision = Recall = 1, so the closer the symbol is to the upper right corner (1,1), the better the classifier. Again, the autotest and the k-fold confirm that the best performances are obtained for 0.75–1 and 0.5 day time intervals, respectively. K-fold analysis yields a Precision of 0.74 (compared to 0.82 in the autotest case). In terms of Precision, the best performances are obtained in autotest for intervals equal to or longer than 3 days, due to the lower FPR outlined in the ROC graph. These performances are not confirmed by k-fold analysis, in which, conversely, Precision decreases for larger time intervals. This result is possibly related to overfitting in the autotest approach like in the ROC diagram case. The Accuracy (i.e., the normalized percentage of correctly classified clusters) of the three methods for T_i = 18 h is 1 for the Chronological approach, 0.88 for the k-fold analysis, and 0.90 for the autotest.

Note that, while in testing analysis autotest results should be regarded as an (optimistic) upper limit on possible performance, for training and threshold determination, the more reliable thresholds are those obtained by the autotest approach, as they use the entire available dataset rather than being derived from subsets, which provide less complete information and lead to less stable estimates. These are therefore the thresholds that should be used when applying the method to future clusters of seismicity. The thresholds obtained from each method are shown in the Electronic Supplement Material (Table S1) along with the classification of each test set with T_i = 18 h (Tables S2–S5).

Table 5 shows the obtained thresholds. Values in bold are calculated for the corresponding time interval, while the others are inherited values.

All features except Vm and N2 contribute to the classification at T_i = 0.75 days according to the thresholds shown in the table. Vm and N2, which supply the best results 6 h after the o-mainshock, also contribute to the classification, but by using inherited values.

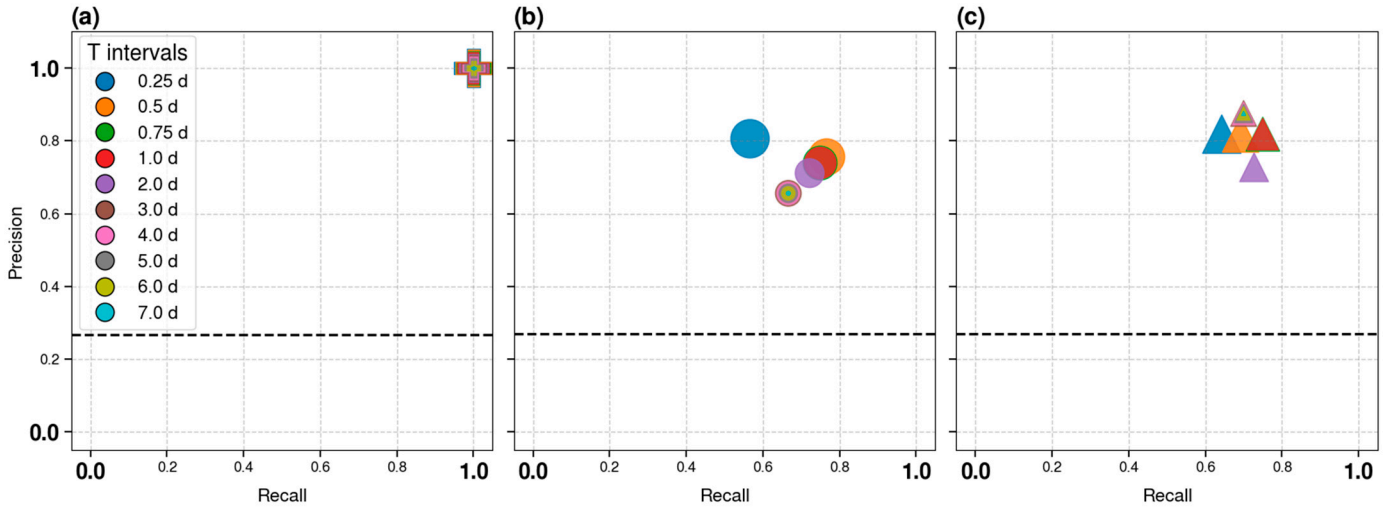

Figure 7. Precision–Recall diagrams for the (**a**) chronological, (**b**) k-fold, and (**c**) autotest approaches. The time interval end is expressed as a fraction of a day. For the k-fold approach (**b**), we plotted the average value across the folds.

Table 5. Thresholds obtained by training with the entire dataset. The bold values are calculated for the corresponding time interval, while the others are inherited from the previous intervals.

Feature	6 h	12 h	18 h	1 d	2 d	3 d
S	**0.028**	**0.028**	**0.0355**	**0.0355**	**0.0355**	**0.096**
Z	**0.053**	**0.062**	**0.062**	**0.0635**	0.0635	0.0635
SLcum	-	**0.027**	**0.027**	**0.027**	**0.055**	**0.1695**
QLcum	-	**2.468**	**2.468**	**3.702**	3.702	3.702
SLcum2	-	-	**0.017901**	**0.017901**	**0.027565**	0.027565
QLcum2	-	-	**2.468**	2.468	2.468	2.468
Q	**0.0035**	**0.0035**	**0.0035**	**0.0035**	**0.004**	0.004
Vm	**0.15**	0.15	0.15	0.15	0.15	0.15
N2	**3.5**	3.5	3.5	3.5	3.5	3.5

5. Discussion

This work should be considered in the context of the broad application of the NESTORE method in several countries beyond New Zealand, including Italy, Slovenia, California, Greece, and Japan [36–38,40,41,43]. The performance evaluation approach used in this paper, which employed the stratified k-fold method rather than the chronological approach, should be seen as part of the ongoing improvement of NESTORE across successive applications, with the aim of making the algorithm more stable and the performance assessment more accurate.

As in most applications of NESTORE to other regions [36–38,40,41,43], the dataset is small and unbalanced, with the cardinality of class B approximately three times that of class A (see Table 3). A review of the results obtained in previous applications of NESTORE is currently under revision [44]. In other applications, NESTORE successfully classified 66–100% of Type A clusters and 90–100% of Type B clusters. The model's Precision, representing the proportion of correctly identified Type A clusters among all those predicted as Type A, ranged from 0.75 to 1, with lower values observed under strong class imbalance favoring Type B clusters. In the application to New Zealand, in the case of the chronological approach, the Precision was 1 for both classes, with 100% of clusters correctly classified. In the k-fold analysis, 77% of Type A and 92% of Type B clusters were correctly classified, and the Precision was 0.76 for the time interval of 12 h after the mainshock. Therefore, both results are in good agreement with other applications of the method.

To account for the small dataset, we verified the probability α of obtaining the observed h-hits (Type A correctly forecasted) by chance. This probability is less than 1% in both the k-fold and chronological approaches, although it is higher for the chronological approach because the dataset is smaller. This result is consistent with applications to other countries, where α ranges from 0.1% to 4% [46].

An important point to discuss is whether, for future classifications, the more reliable performance estimate for New Zealand is provided by the chronological approach or the k-fold approach. One reason to consider the chronological approach carefully is that, while the k-fold results are generally worse than the autotest results, the chronological approach yields better results. As the autotest results should be regarded as an upper bound (an optimistic estimate of performance) due to possible overfitting, we expected its performance to be the best. If the quality of the catalogue in terms of location and magnitude assessment, at least for earthquakes with magnitude greater than 3, had remained unchanged over time, the k-fold approach would be the most reliable approach, as it analyses a larger dataset, i.e., the entire available database. However, it is well known that the quality of seismic catalogues generally changes over time, due to the increased number of stations, higher quality sensors, improved location algorithms, et cetera. As mentioned in Section 2.2, in New Zealand the seismic network and the data processing evolved with time [51]. A major change since the beginning of the instrumental era has been the switching to the open-source earthquake analysis program SeisComP (SC) [80] in 2012. Notably, most outliers occurred before 2012. We consider it unlikely that these anomalies are related to a common physical cause, as anomalous clusters are generally well separated in both space and time (see Figure 5). A possible interpretation is that most outliers are associated with catalogue inaccuracies during the period 1988–2011. To explore this hypothesis, we assume that such anomalous clusters (1) occur randomly and independently, (2) have a constant probability over time, and (3) do not influence one another. Under these assumptions, the occurrence of anomalous clusters can be described by a Poisson distribution. We therefore test the null hypothesis H0 that this timing of outliers is due to chance, against the alternative hypothesis H1, that the absence of outliers is not random (possibly related to independent causes, such as catalogue quality). From 1988 to the end of 2011, there are 24 years in which we detected four outliers according to autotest, while from 2012 to 15 May 2025, we have 13.37 years with one outlier. The annual mean of outliers in the catalogue is therefore 0.17 outliers per year, corresponding to $\mu = 2.23$ outliers in 13.39 years. According to the Poisson distribution, the probability of having $x = 1$ outliers is therefore

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!} = e^{-\mu} = 0.24 \quad (6)$$

In this case, the hypothesis H0 that this timing of outliers is due to chance cannot be rejected at 5% level; in other words, the occurrence of all outliers in the first time period is statistically compatible with the observed results, without requiring changes to the catalog that would affect the NESTORE analysis.

Given the presence of two performance estimates—one more optimistic and one more conservative—and the limited amount of data available for robust discrimination between them, we adopt the more conservative estimate to ensure a cautious interpretation of the results.

For the outliers identified in our performance evaluation, a more detailed analysis of these clusters will be necessary to determine whether their anomaly is genuine (i.e., the model cannot be applied to a small percentage of New Zealand clusters), or an artefact resulting from issues with cluster identification—a similar problem was observed in southern Italy for a sequence in 2018, whose anomalous behavior was related to fluid [44], due

to inconsistencies in magnitude definition, or some unknown physical cause. This issue will be addressed in a future paper (in preparation).

NESTORE uses the feature classifiers together (see Section 3.4); therefore, each cluster classification depends not only on the overall importance of a single feature, but also on the level of impurity (class mix) in the training set distribution within the half-space (above or below the threshold) where that cluster feature falls. The final classification depends on the level of impurity of each feature in the corresponding half-space. To formalize this, we estimated for each feature both the Gini Impurity [81] in the two half-spaces and the overall Gini Gain. The Gini Gain is defined as the difference between the total Gini Impurity (i.e., the probability of incorrectly classifying a randomly chosen element in the dataset if it were randomly labelled according to the class distribution in the dataset) and the sum of the impurities of each branch, weighted by the number of elements in each. The Impurity can be expressed in terms of $p_{n,i}$ from Equation (4) as

$$I = 2p_{n,i}(1 - p_{n,i}) \tag{7}$$

Table 6 shows the results obtained at 18 h, the time at which we achieved the best performance. The Gini Gain ranges from 0.15 to 0.24, with a mean of 0.20. Gini Impurity for the left branch ranges from 0.14 to 0.24, with a mean of 0.18, and for the right branch from 0 to 0.46, with a mean of 0.25.

Table 6. The Gini Impurity of the selected features at T_i = 18 h in the half-spaces below and above the thresholds for the entire dataset, along with the overall Gini Gain relative to a random choice.

Feature	Gini Impurity $\leq$ Th	Gini Impurity > Th	Gini Gain
S	0.21	0.22	0.21
Z	0.21	0.22	0.21
SLcum	0.18	0.32	0.21
QLcum	0.14	0.18	0.15
SLcum2	0.15	0.46	0.24
QLcum2	0.14	0.18	0.15
Q	0.18	0.40	0.23
Vm	0.18	0.32	0.21
N2	0.24	0.00	0.21

Another interesting result, also in comparison with other applications, is that all the features contribute to classification at T_i = 18 h. Features N2 and Vm provide the best results shortly after the mainshock, and at T_i = 18 h, they are inherited from previous ones; the other seven features are estimated directly at 18 h.

Figure 8 shows a comparison with features obtained in other countries, adapted from [46]. A correlation between the seismotectonic characteristics of a region and the feature threshold has not yet been established, except for a similar trend between seismic productivity and N2 in the available data [46]. New Zealand exhibits thresholds similar to Italy for the features S, SLCum, SLCum2, Q, and Vm, while for other features, the threshold is higher, more closely resembling those of Greece. An interesting result from Gentili et al. [46] is that in most regions analyzed, the S feature achieves good results in class discrimination during training, which leads to its inclusion in the set of features used for cluster classification. This result is also confirmed for New Zealand, where the feature is adopted in time intervals from 0.25 to 3 days and inherited for longer intervals. In testing using the k-fold approach for T_i = 0.75 h, the S feature alone achieves Precision = 0.75, Recall = 0.58, and Accuracy = 0.84, showing good performance in correctly classifying B-Type clusters; furthermore, its Gini Gain (0.21) is very close to the mean in all cases.

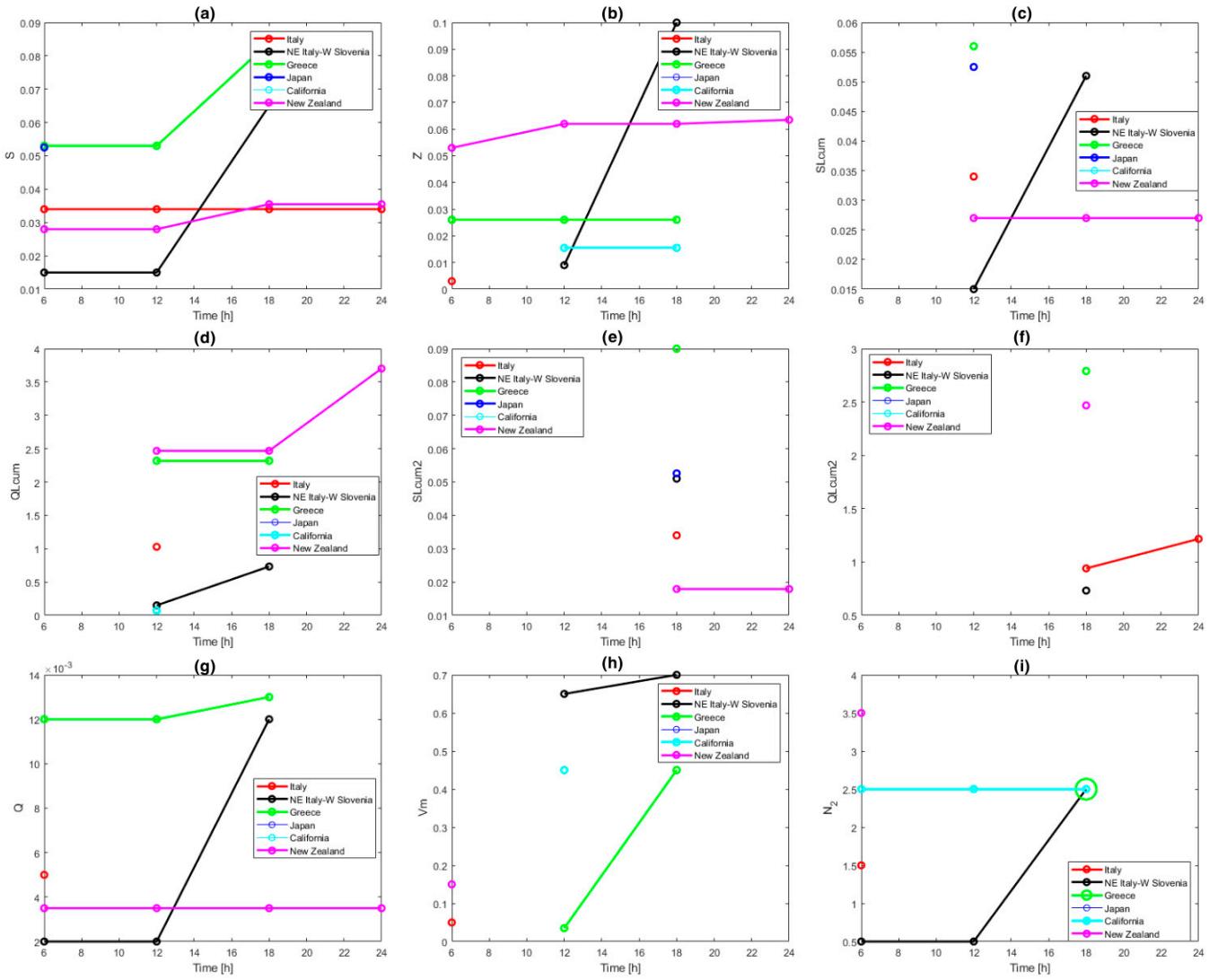

Figure 8. Comparison of feature thresholds in the different regions where NESTORE has been applied: (**a**) S; (**b**) Z; (**c**) SLCum; (**d**) QLCum; (**e**) SLCum2; (**f**) QLCum2; (**g**) Q; (**h**) Vm; (**i**) N2. Adapted from Gentili et al. [46]. When data is not available for a given feature, the corresponding line in the legend is thinner.

The ability of feature S to distinguish between Type A and Type B clusters shortly after an o-mainshock can be interpreted from a physical perspective [46]. Following the reasoning of Smirnov and Petruchov [82], regarding precursors of large earthquakes, it is notable that the value of the normalized cumulative source area (S) is calculated in NESTORE from the magnitudes of the aftershocks. From a physical point of view, as the source size is related to the scale of lithospheric heterogeneities involved in fracturing [83], the size distribution of these heterogeneities plays a key role in determining the distribution of aftershock magnitudes [84].

The progression of fracturing and the associated stress redistribution favor the enlargement and coalescence of cracks. The tendency of fractures to merge into larger ones, thereby generating stronger aftershocks, depends on factors such as the fracturing state of the medium, the environmental stress field, the orientation of the fractures relative to the stress field, and the presence of fluids. These parameters vary in both space and time. The clusters where this trend is most evident are those in which strong aftershocks (i.e., with high S) are most likely to occur, with magnitudes comparable to that of the o-mainshock, corresponding to Type A clusters. This finding is valid regardless of the region analyzed.

Type A clusters, on the other hand, are characterized by strong fluctuations in S (as shown by SLCum and SLCum2) and greater variation in magnitude from one event to another (as indicated by Vm). Such behavior may be interpreted as a sign of instability within the nonlinear fault system responsible for earthquake generation [32], consistent with patterns observed prior to major earthquakes and stronger events within seismic clusters [31,85].

The main advantage of our method over others in the literature is that it provides information on the expected occurrence of a strong aftershock within a few hours after the mainshock, even if the aftershock may occur several months later.

To compare our method with others, we evaluated the performance of alternative approaches over a similarly long time span. The comparison with Båth's law [8] is straightforward. This law assumes a fixed Δm of 1.2; in other words, all clusters should be classified as Type B. An analysis of our dataset shows that, for New Zealand seismicity, the mean value of this difference is 1.4, with a standard deviation of 0.7.

Well-known Operational Earthquake Forecasting (OEF) methods based on the Epidemic-Type Aftershock Sequence (ETAS) model [14,15], the ETES model [86], and the Short-Term Earthquake Probability (STEP) model [87] provide information on a day-by-day basis [88], allowing model parameters to adapt to ongoing seismicity.

A simpler approach uses fixed parameters, as in the Reasenberg and Jones (1990) formulation [10]. We tested this approach using parameters estimated by [89]: $a = -1.66$, $b = 1.03$, $c = 0.03$, $p = 1.02$, where a represents the productivity of the sequence [90], c and p are the parameters of the modified Omori law [91], and b is the parameter of the Gutenberg–Richter law [92].

Using this approach, we estimated the probability of observing an aftershock of magnitude $M_m - 1$ as a function of time. According to our results, the probability of such an event occurring within a time interval starting 6 h and ending one week after is approximately 53%, increasing to 66% if the ending time is 30 days, and reaching 75% when a four-month window is considered. If the analysis starts 18 h after the mainshock, the corresponding probabilities are 41%, 57%, and 68%, respectively. Since, according to Equation (2), all analyzed cluster durations exceed four months, all sequences show a high probability of being classified as type A clusters.

However, it is well known that this approach is affected by the large variability of parameter values from one sequence to another (see, e.g., [33]). Eberhart-Phillips also highlights that this variability is particularly pronounced among New Zealand sequences. Therefore, even in this case, the parameters of the seismic sequence under analysis should be estimated repeatedly over a short time window to obtain more reliable values.

While NESTORE achieves very encouraging results shortly after the mainshock, it does not provide information on the timing of the forecasted strong aftershock within the cluster. In light of these results, it appears worthwhile to integrate NESTORE with OEF methods, combining NESTORE's ability to rapidly issue an alert shortly after the mainshock with a refined estimation of the temporal occurrence of strong aftershocks during the sequence's evolution provided by other methods

6. Conclusions

New Zealand is highly prone to large earthquakes, and an event with a magnitude greater than 7—and possibly greater than 8—is considered likely in the coming years [57]. Because strong aftershocks in this region have, in several cases, caused severe damage, the ability to rapidly assess the likelihood of additional strong events following a mainshock is essential for effective emergency response and risk mitigation.

A textbook example is the 2010–2011 Canterbury–Christchurch sequence, in which the ML 6.3 Christchurch aftershock caused substantially greater damage than the preceding ML 7.1 Canterbury mainshock due to its location and the prior weakening of the built environment. In this study, we applied the latest version of the NESTORE algorithm [45] to New Zealand seismicity. NESTORE, a machine learning method that uses seismicity recorded in the first hours after a mainshock to forecast the likelihood of strong subsequent events, correctly classified the Canterbury–Christchurch sequence as a Type A cluster (strongest aftershock with magnitude $\geq M_m - 1$). This classification was achieved using data within hours of the September 2010 mainshock, more than five months before the damaging Christchurch earthquake.

These results were derived as part of a comprehensive analysis of all New Zealand seismicity clusters between 1988 and 2025 presented in this paper, using the New Zealand Earthquake Catalogue up to 2020 and data automatically retrieved from the GeoNet Quake Search system thereafter. Some constraints exist and need to be considered for future applications and developments of our method: we focused on shallow, onshore crustal seismicity ($\leq$50 km depth), resulting in a relatively small sample size (52 clusters); the network has undergone significant modifications which likely affected the catalog and the available magnitudes, which in some cases had to be calculated with orthogonal regression. Nevertheless, across the full dataset, NESTORE correctly classified 77% of Type A clusters and 92% of Type B clusters, with a precision of 0.76.

Overall, the findings demonstrate the potential of NESTORE for near-real-time application in New Zealand and support its use as a tool to enhance rapid post-earthquake forecasting capabilities.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/forecast8010016/s1, Figure S1: Residuals of ML with respect to the orthogonal regressions.; Table S1: Feature thresholds obtained by training among the different methods; Table S2. Cluster classification for each feature and their Bayes theorem-based combination at T_i = 18 h for the chronological approach (2014–2025); Table S3. Cluster classification for each feature and their Bayes theorem-based combination at T_i = 18 h for the k-fold approach (k = 1); Table S4. Cluster classification for each feature and their Bayes theorem-based combination at T_i 18 h for the k-fold approach (k = 2); Table S5. Cluster classification for each feature and their Bayes theorem-based combination at T_i = 18 h for the k-fold approach (k = 3); Table S6. Cluster classification for each feature and their Bayes theorem-based combination at T_i = 18 h for the autotest approach.

Author Contributions: L.C.: software, validation, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, visualization; S.G.: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft, writing—review and editing, visualization, supervision, project administration, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This work is co-funded within the RETURN Extended Partnership and received funding from the European Union Next-GenerationEU (National Recovery and Resilience Plan—NRRP, Mission 4, Component 2, Investment 1.3—D.D. 1243 2/8/2022, PE0000005) and by the grant "Progetto INGV Pianeta Dinamico: Near real-time results of Physical and Statistical Seismology for earthquakes observations, modelling and forecasting (NEMESIS)"—code CUP D53J19000170001—funded by Italian Ministry MIUR ("Fondo Finalizzato al rilancio degli investimenti delle amministrazioni centrali dello Stato e allo sviluppo del Paese", legge 145/2018).

Data Availability Statement: The NExt STrOng Related Earthquake (NESTOREv1.0) toolbox is available for free download from GitHub at https://github.com/StefaniaGentili/NESTORE (last accessed on 24 November 2024), and the reproducibility package is available in Zenodo at: https://zenodo.org/account/settings/github/repository/StefaniaGentili/NESTORE (last accessed on 24 November 2024). At these links, the code is proposed along with a detailed readme for software us-

age. The software NZQuakeParser is available for free download from GitHub at https://github.com/LetiziaCaravella/NZQuakeParser (last accessed on 7 November 2025) and the reproducibility package is available in Zenodo at https://doi.org/10.5281/zenodo.17552472 (last accessed on 7 November 2025). The tests shown in this article were obtained using data from (1) the New Zealand Earthquake Catalogue for the revision of the 2022 National Seismic Hazard Model (NSHM) downloaded at https://geodata.nz/geonetwork/srv/api/records/daf6b614-9d0a-45e1-8fa6-df8e2afd7d4d; (last accessed on 18 December 2024); (2) the GeoNet Aotearoa New Zealand Earthquake Catalogue (GNS Science, 1970 47) downloaded at: https://quakesearch.geonet.org.nz/ (last accessed on 15 May 2025). The geospatial layers used in Figure 1 derive from the TRAMZ data package, downloaded from https://data.gns.cri.nz/mapservice/Content/Zealandia/downloads/TRAMZ_dataPackage.zip (last accessed on 1 October 2025).

Acknowledgments: Map Figure 1 was created using QGIS 3.44.7 (http://www.qgis.org), while Map Figure 4 was produced using ZMAP 6 (no longer available online for download). The other figures and evaluations are performed in MATLAB and Python (version 3.12.3) programming languages. We wish to thank Piero Brondi, Giuseppe Davide Chiappetta and Luigi Sante Zampa for their useful suggestions.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CSEP	Collaboratory for the Study of Earthquake Predictability
EID	Earthquake Information Database
EEPAS	Every Earthquake a Precursor According to Scale
ETAS	Epidemic-Type Aftershock Sequence
FPR	False-Positive Rate
HFT	Hybrid Forecast Tool
LIP	Large Igneous Province
MFS	Marlborough Fault System
NESTORE	NExt STrOng Related Earthquake
OEF	Operational Earthquake Forecasting
ROC	Receiver Operating Characteristic
REPENESE	RElevant features, class imbalance PErcentage, NEighborhood detection, SElection
STEP	Short-term Aftershock Probability
TRZ	Taupō Rift Zone
TPR	True-Positive Rate
UCERF3-ETAS	Third Uniform California Earthquake Rupture Forecast ETAS model

References

1. Iervolino, I.; Giorgio, M.; Chioccarelli, E. Closed-form Aftershock Reliability of Damage-cumulating Elastic-perfectly-plastic Systems. *Earthq. Eng. Struct. Dyn.* **2013**, *43*, 613–625. [CrossRef]
2. Davison, C. The Hawke's Bay Earthquake of February 3, 1931. *Nature* **1934**, *133*, 841–842. [CrossRef]
3. Dowrick, D.J. Damage and Intensities in the Magnitude 7.8 1931 Hawke's Bay, New Zealand, Earthquake. *Bull. N. Z. Soc. Earthq. Eng.* **1998**, *31*, 139–163. [CrossRef]
4. Rollins, C.; Gerstenberger, M.C.; Rhoades, D.A.; Rastin, S.J.; Christophersen, A.; Thingbaijam, K.K.S.; Van Dissen, R.J.; Graham, K.; DiCaprio, C.; Fraser, J. The Magnitude–Frequency Distributions of Earthquakes in Aotearoa New Zealand and on Adjoining Subduction Zones, Using a New Integrated Earthquake Catalog. *Bull. Seismol. Soc. Am.* **2024**, *114*, 150–181. [CrossRef]
5. Cattania, C.; Werner, M.J.; Marzocchi, W.; Hainzl, S.; Rhoades, D.; Gerstenberger, M.; Liukis, M.; Savran, W.; Christophersen, A.; Helmstetter, A.; et al. The Forecasting Skill of Physics-Based Seismicity Models during the 2010–2012 Canterbury, New Zealand, Earthquake Sequence. *Seismol. Res. Lett.* **2018**, *89*, 1238–1250. [CrossRef]

6. Herman, M.W.; Herrmann, R.B.; Benz, H.M.; Furlong, K.P. Using Regional Moment Tensors to Constrain the Kinematics and Stress Evolution of the 2010–2013 Canterbury Earthquake Sequence, South Island, New Zealand. *Tectonophysics* **2014**, *633*, 1–15. [CrossRef]
7. Bannister, S.; Gledhill, K. Evolution of the 2010–2012 Canterbury Earthquake Sequence. *N. Z. J. Geol. Geophys.* **2012**, *55*, 295–304. [CrossRef]
8. Båth, M. Lateral Inhomogeneities of the Upper Mantle. *Tectonophysics* **1965**, *2*, 483–514. [CrossRef]
9. Chan, C.-H.; Wu, Y.-M. Maximum Magnitudes in Aftershock Sequences in Taiwan. *J. Asian Earth Sci.* **2013**, *73*, 409–418. [CrossRef]
10. Reasenberg, P.A.; Jones, L.M. California Aftershock Hazard Forecasts. *Science* **1990**, *247*, 345–346. [CrossRef]
11. Roeloffs, E.; Goltz, J. The California Earthquake Advisory Plan: A History. *Seismol. Res. Lett.* **2017**, *88*, 784–797. [CrossRef]
12. Field, E.H.; Jordan, T.H.; Jones, L.M.; Michael, A.J.; Blanpied, M.L. The Potential Uses of Operational Earthquake Forecasting: Table 1. *Seismol. Res. Lett.* **2016**, *87*, 313–322. [CrossRef]
13. Zechar, J.D.; Marzocchi, W.; Wiemer, S. Operational Earthquake Forecasting in Europe: Progress, despite Challenges. *Bull. Earthq. Eng.* **2016**, *14*, 2459–2469. [CrossRef]
14. Ogata, Y. Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *J. Am. Stat. Assoc.* **1988**, *83*, 9–27. [CrossRef]
15. Ogata, Y. Space-Time Point-Process Models for Earthquake Occurrences. *Ann. Inst. Stat. Math.* **1998**, *50*, 379–402. [CrossRef]
16. Gerstenberger, M.C.; Wiemer, S.; Jones, L.M.; Reasenberg, P.A. Real-Time Forecasts of Tomorrow's Earthquakes in California. *Nature* **2005**, *435*, 328–331. [CrossRef]
17. Gulia, L.; Wiemer, S. Real-Time Discrimination of Earthquake Foreshocks and Aftershocks. *Nature* **2019**, *574*, 193–199. [CrossRef]
18. Utsu, T.; Ogata, Y.; Matsu'ura, R.S. The Centenary of the Omori Formula for a Decay Law of Aftershock Activity. *J. Phys. Earth* **1995**, *43*, 1–33. [CrossRef]
19. Utsu, T. Aftershocks and Earthquake Statistics(2): Further Investigation of Aftershocks and Other Earthquake Sequences Based on a New Classification of Earthquake Sequences. *Hokkaido Univ. Collect. Sch. Acad. Pap. (Hokkaido Univ.)* **1971**, *3*, 197–266.
20. Gutenberg, B.; Richter, C.F. Frequency of Earthquakes in California. *Bull. Seismol. Soc. Am.* **1944**, *34*, 185–188. [CrossRef]
21. Reasenberg, P.A.; Jones, L.M. Earthquake Hazard after a Mainshock in California. *Science* **1989**, *243*, 1173–1176. [CrossRef] [PubMed]
22. Page, M.T.; Van Der Elst, N.; Hardebeck, J.; Felzer, K.; Michael, A.J. Three Ingredients for Improved Global Aftershock Forecasts: Tectonic Region, Time-Dependent Catalog Incompleteness, and Intersequence Variability. *Bull. Seismol. Soc. Am.* **2016**, *106*, 2290–2301. [CrossRef]
23. Felzer, K.R. Secondary Aftershocks and Their Importance for Aftershock Forecasting. *Bull. Seismol. Soc. Am.* **2003**, *93*, 1433–1448. [CrossRef]
24. Field, E.H.; Milner, K.R.; Hardebeck, J.L.; Page, M.T.; Van Der Elst, N.; Jordan, T.H.; Michael, A.J.; Shaw, B.E.; Werner, M.J. A Spatiotemporal Clustering Model for the Third Uniform California Earthquake Rupture Forecast (UCERF3-ETAS): Toward an Operational Earthquake Forecast. *Bull. Seismol. Soc. Am.* **2017**, *107*, 1049–1081. [CrossRef]
25. Milner, K.R.; Field, E.H.; Savran, W.H.; Page, M.T.; Jordan, T.H. Operational Earthquake Forecasting during the 2019 Ridgecrest, California, Earthquake Sequence with the UCERF3-ETAS Model. *Seismol. Res. Lett.* **2020**, *91*, 1567–1578. [CrossRef]
26. Savran, W.H.; Werner, M.J.; Marzocchi, W.; Rhoades, D.A.; Jackson, D.D.; Milner, K.; Field, E.; Michael, A. Pseudoprospective Evaluation of UCERF3-ETAS Forecasts during the 2019 Ridgecrest Sequence. *Bull. Seismol. Soc. Am.* **2020**, *110*, 1799–1817. [CrossRef]
27. Mizrahi, L.; Dallo, I.; Van Der Elst, N.J.; Christophersen, A.; Spassiani, I.; Werner, M.J.; Iturrieta, P.; Bayona, J.A.; Iervolino, I.; Schneider, M.; et al. Developing, Testing, and Communicating Earthquake Forecasts: Current Practices and Future Directions. *Rev. Geophys.* **2024**, *62*, e2023RG000823. [CrossRef]
28. Rhoades, D.A.; Evison, F.F. Long-Range Earthquake Forecasting with Every Earthquake a Precursor According to Scale. *Pure Appl. Geophys.* **2004**, *161*, 47–72. [CrossRef]
29. Harte, D.S. Bias in Fitting the ETAS Model: A Case Study Based on New Zealand Seismicity. *Geophys. J. Int.* **2012**, *192*, 390–412. [CrossRef]
30. Graham, K.M.; Christophersen, A.; Rhoades, D.A.; Gerstenberger, M.C.; Jacobs, K.M.; Huso, R.; Canessa, S.; Zweck, C. A Software Tool for Hybrid Earthquake Forecasting in New Zealand. *Seismol. Res. Lett.* **2024**, *95*, 3250–3263. [CrossRef]
31. Vorobieva, I.A.; Panza, G.F. Prediction of the Occurrence of Related Strong Earthquakes in Italy. *Pure Appl. Geophys.* **1993**, *141*, 25–41. [CrossRef]
32. Vorobieva, I.A. Prediction of a Subsequent Large Earthquake. *Phys. Earth Planet. Inter.* **1999**, *111*, 197–206. [CrossRef]
33. Gentili, S.; Bressan, G. The Partitioning of Radiated Energy and the Largest Aftershock of Seismic Sequences Occurred in the Northeastern Italy and Western Slovenia. *J. Seismol.* **2007**, *12*, 343–354. [CrossRef]
34. Bressan, G.; Barnaba, C.; Gentili, S.; Rossi, G. Information Entropy of Earthquake Populations in Northeastern Italy and Western Slovenia. *Phys. Earth Planet. Inter.* **2017**, *271*, 29–46. [CrossRef]

35. Shcherbakov, R. A Modified Form of Bath's Law. *Bull. Seismol. Soc. Am.* **2004**, *94*, 1968–1975. [CrossRef]
36. Gulia, L.; Wiemer, S.; Biondini, E.; Enescu, B.; Vannucci, G. Improving the Foreshock Traffic Light Systems for Real-Time Discrimination between Foreshocks and Aftershocks. *Seismol. Res. Lett.* **2024**, *95*, 3579–3592. [CrossRef]
37. Van Der Elst, N.J. B-Positive: A Robust Estimator of Aftershock Magnitude Distribution in Transiently Incomplete Catalogs. *J. Geophys. Res. Solid Earth* **2021**, *126*, e2020JB021027. [CrossRef]
38. Gentili, S.; Di Giovambattista, R. Pattern Recognition Approach to the Subsequent Event of Damaging Earthquakes in Italy. *Phys. Earth Planet. Inter.* **2017**, *266*, 1–17. [CrossRef]
39. Gentili, S.; Di Giovambattista, R. Forecasting Strong Aftershocks in Earthquake Clusters from Northeastern Italy and Western Slovenia. *Phys. Earth Planet. Inter.* **2020**, *303*, 106483. [CrossRef]
40. Gentili, S.; Di Giovambattista, R. Forecasting Strong Subsequent Earthquakes in California Clusters by Machine Learning. *Phys. Earth Planet. Inter.* **2022**, *327*, 106879. [CrossRef]
41. Gentili, S.; Brondi, P.; Di Giovambattista, R. NESTOREV1.0: A MATLAB Package for Strong Forthcoming Earthquake Forecasting. *Seismol. Res. Lett.* **2023**, *94*, 2003–2013. [CrossRef]
42. Anyfadi, E.-A.; Gentili, S.; Brondi, P.; Vallianatos, F. Forecasting Strong Subsequent Earthquakes in Greece with the Machine Learning Algorithm NESTORE. *Entropy* **2023**, *25*, 797. [CrossRef]
43. Brondi, P.; Gentili, S.; Di Giovambattista, R. Forecasting Strong Subsequent Events in the Italian Territory: A National and Regional Application for NESTOREv1.0. *Nat. Hazards* **2024**, *121*, 3499–3531. [CrossRef]
44. Gentili, S.; Brondi, P.; Rossi, G.; Sugan, M.; Petrillo, G.; Zhuang, J.; Campanella, S. Seismic Clusters and Fluids Diffusion: A Lesson from the 2018 Molise (Southern Italy) Earthquake Sequence. *Earth Planets Space* **2024**, *76*, 157. [CrossRef]
45. Gentili, S.; Chiappetta, G.D.; Petrillo, G.; Brondi, P.; Zhuang, J. Forecasting Strong Subsequent Earthquakes in Japan Using an Improved Version of NESTORE Machine Learning Algorithm. *Geosci. Front.* **2025**, *16*, 102016. [CrossRef]
46. Gentili, S.; Brondi, P.; Chiappetta, G.D.; Petrillo, G.; Zhuang, J.; Anyfadi, E.-A.; Vallianatos, F.; Caravella, L.; Magrin, E.; Comelli, P.; et al. NESTORE Algorithm: A Machine Learning Approach for Strong Aftershock Forecasting. Comparison of California, Italy, Western Slovenia, Greece and Japan Results. *Bull. Geophys. Oceanogr.* **2025**, *accepted*.
47. Luyendyk, B.P. Hypothesis for Cretaceous Rifting of East Gondwana Caused by Subducted Slab Capture. *Geology* **1995**, *23*, 373–376. [CrossRef]
48. Mortimer, N.; Campbell, H.J.; Tulloch, A.J.; King, P.R.; Stagpoole, V.M.; Wood, R.A.; Rattenbury, M.S.; Sutherland, R.; Adams, C.J.; Collot, J.; et al. Zealandia: Earth's Hidden Continent. *GSA Today* **2017**, *27*, 27–35. [CrossRef]
49. Science, GNS. *New Zealand Earthquake Catalogue [Data Set]*; GNS Science: Lower Hutt, New Zealand, 1970. [CrossRef]
50. Hirschberg, H.; Sutherland, R. A Kinematic Model of Quaternary Fault Slip Rates and Distributed Deformation at the New Zealand Plate Boundary. *J. Geophys. Res. Solid Earth* **2022**, *127*, e2022JB024828. [CrossRef]
51. Christophersen, A.; Bourguignon, S.; Rhoades, D.A.; Allen, T.I.; Ristau, J.; Salichon, J.; Rollins, J.C.; Townend, J.; Gerstenberger, M.C. Standardizing Earthquake Magnitudes for the 2022 Revision of the Aotearoa New Zealand National Seismic Hazard Model. *Bull. Seismol. Soc. Am.* **2023**, *114*, 111–136. [CrossRef]
52. Mortimer, N.; Smith Lyttle, B.; Black, J. *Te Riu-a-Māui/Zealandia Digital Geoscience Data Compilation, Scale 1:8 500 000. GNS Science Geological Map 11*; GNS Science: Lower Hutt, New Zealand, 2020. [CrossRef]
53. Williams, C.A.; Eberhart-Phillips, D.; Bannister, S.; Barker, D.H.N.; Henrys, S.; Reyners, M.; Sutherland, R. Revised Interface Geometry for the Hikurangi Subduction Zone, New Zealand. *Seismol. Res. Lett.* **2013**, *84*, 1066–1073. [CrossRef]
54. Mortimer, N. Tectonics, Geology and Origins of Te Riu-a-Māui/Zealandia. *N. Z. J. Geol. Geophys.* **2025**, *68*, 531–567. [CrossRef]
55. Wilson, C.J.N.; Houghton, B.F.; McWilliams, M.O.; Lanphere, M.A.; Weaver, S.D.; Briggs, R.M. Volcanic and Structural Evolution of Taupo Volcanic Zone, New Zealand: A Review. *J. Volcanol. Geotherm. Res.* **1995**, *68*, 1–28. [CrossRef]
56. Seebeck, H.; Strogen, D.P.; Nicol, A.; Hines, B.R.; Bland, K.J. A Tectonic Reconstruction Model for Aotearoa-New Zealand from the Mid-Late Cretaceous to the Present Day. *N. Z. J. Geol. Geophys.* **2023**, *67*, 527–550. [CrossRef]
57. Sutherland, R.; Eberhart-Phillips, D.; Harris, R.A.; Stern, T.; Beavan, J.; Ellis, S.; Henrys, S.; Cox, S.; Norris, R.J.; Berryman, K.R.; et al. Do Great Earthquakes Occur on the Alpine Fault in Central South Island, New Zealand? *Geophys. Monogr.* **2007**, *175*, 235–251.
58. Norris, R.J.; Cooper, A.F. Origin of Small-Scale Segmentation and Transpressional Thrusting along the Alpine Fault, New Zealand. *Geol. Soc. Am. Bull.* **1995**, *107*, 231–240. [CrossRef]
59. Melhuish, A.; Sutherland, R.; Davey, F.J.; Lamarche, G. Crustal Structure and Neotectonics of the Puysegur Oblique Subduction Zone, New Zealand. *Tectonophysics* **1999**, *313*, 335–362. [CrossRef]
60. Mortimer, N.; Gans, P.B.; Foley, F.V.; Turner, M.B.; Daczko, N.; Robertson, M.; Turnbull, I.M. Geology and Age of Solander Volcano, Fiordland, New Zealand. *J. Geol.* **2013**, *121*, 475–487. [CrossRef]
61. Eberhart-Phillips, D.; Reyners, M. A Complex, Young Subduction Zone Imaged by Three-Dimensional Seismic Velocity, Fiordland, New Zealand. *Geophys. J. Int.* **2001**, *146*, 731–746. [CrossRef]

62. Petersen, T.; Gledhill, K.; Chadwick, M.; Gale, N.H.; Ristau, J. The New Zealand National Seismograph Network. *Seismol. Res. Lett.* **2011**, *82*, 9–20. [CrossRef]
63. Science, GNS. *New Zealand Earthquake Catalogue for the Revision of the 2022 National Seismic Hazard Model (NSHM) [Data Set]*; GNS Science: Lower Hutt, New Zealand, 2022. [CrossRef]
64. Gerstenberger, M.C.; Bora, S.; Bradley, B.A.; DiCaprio, C.; Kaiser, A.; Manea, E.F.; Nicol, A.; Rollins, C.; Stirling, M.W.; Thingbaijam, K.K.S.; et al. The 2022 Aotearoa New Zealand National Seismic Hazard Model: Process, Overview, and Results. *Bull. Seismol. Soc. Am.* **2023**, *114*, 7–36. [CrossRef]
65. Caravella, L. *NZQuakeParser [Software]*; Zenodo, 2025. Available online: https://zenodo.org/records/17552472 (accessed on 1 February 2026).
66. Wiemer, S. Minimum Magnitude of Completeness in Earthquake Catalogs: Examples from Alaska, the Western United States, and Japan. *Bull. Seismol. Soc. Am.* **2000**, *90*, 859–869. [CrossRef]
67. Wiemer, S. A Software Package to Analyze Seismicity: ZMAP. *Seismol. Res. Lett.* **2001**, *72*, 373–382. [CrossRef]
68. Christophersen, A. *Towards a New Zealand Model for Short-Term Earthquake Probability: Aftershock Productivity and Parameters from Global Catalogue Analysis» Natural Hazards Commission Toka Tū Ake*; Natural Hazards Commission Toka Tū Ake: Wellington, New Zealand, 2006.
69. Gardner, J.K.; Knopoff, L. Is the Sequence of Earthquakes in Southern California, with Aftershocks Removed, Poissonian? *Bull. Seismol. Soc. Am.* **1974**, *64*, 1363–1367. [CrossRef]
70. Zaccagnino, D.; Telesca, L.; Doglioni, C. Global versus Local Clustering of Seismicity: Implications with Earthquake Prediction. *Chaos Solitons Fractals* **2023**, *170*, 113419. [CrossRef]
71. Uhrhammer, R.A. Characteristics of Northern and Central California Seismicity. *Earthq. Notes* **1986**, *57*, 21.
72. Kagan, Y.Y. Seismic Moment Distribution Revisited: I. Statistical Results. *Geophys. J. Int.* **2002**, *148*, 520–541. [CrossRef]
73. Lolli, B.; Gasperini, P. Aftershocks Hazard in Italy Part I: Estimation of Time-Magnitude Distribution Model Parameters and Computation of Probabilities of Occurrence. *J. Seismol.* **2003**, *7*, 235–257. [CrossRef]
74. Spassiani, I.; Gentili, S.; Console, R.; Murru, M.; Taroni, M.; Falcone, G. Reconciling the Irreconcilable: Window-Based versus Stochastic Declustering Algorithms. *Geophys. J. Int.* **2024**, *240*, 1009–1027. [CrossRef]
75. Utsu, T. Relation between the Area of Afterhock Region and the Energy of the Main Shock. *J. Seismol. Soc. Jpn.* **1955**, *2*, 233–240.
76. Zeng, X.; Martinez, T.R. Distribution-Balanced Stratified Cross-Validation for Accuracy Estimation. *J. Exp. Theor. Artif. Intell.* **2000**, *12*, 1–12. [CrossRef]
77. Zechar, J.D. Evaluating Earthquake Predictions and Earthquake Forecasts: A Guide for Students and New Researchers. CORSSA-Community Online Resource for Statistical Seismicity Analysis. 2010. Available online: https://www.corssa.org/export/sites/corssa/.galleries/articles-pdf/zechar.pdf_2063069299.pdf (accessed on 1 February 2026).
78. Jordan, T.H. Earthquake Predictability, Brick by Brick. *Seismol. Res. Lett.* **2006**, *77*, 3–6. [CrossRef]
79. Zechar, J.D.; Schorlemmer, D.; Liukis, M.; Yu, J.; Euchner, F.; Maechling, P.J.; Jordan, T.H. The Collaboratory for the Study of Earthquake Predictability Perspective on Computational Earthquake Science. *Concurr. Comput. Pract. Exp.* **2009**, *22*, 1836–1847. [CrossRef]
80. Helmholtz-Centre Potsdam-GFZ German Research Centre for Geoscience and gempa GmbH. The SeisComP Seismological Software Package. *GFZ Data Services*. 2008. Available online: https://www.seiscomp.de/ (accessed on 1 February 2026).
81. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1984.
82. Smirnov, V.B.; Petrushov, A.A. On the Relationship between RTL and B-Value Anomalies of Seismicity. *Izv. Phys. Solid Earth* **2025**, *61*, 539–552. [CrossRef]
83. Scholz, C.H. *The Mechanics of Earthquakes and Faulting*; Cambridge University Press: Cambridge, UK, 2019.
84. Aki, K. A Probabilistic Synthesis of Precursory Phenomena. In *Earthquake Prediction: An International Review*; American Geophysical Union: Washington, DC, USA, 1981; Volume 4, pp. 566–574.
85. Keilis-Borok, V.I.; Rotwain, I.M. Diagnosis of Time of Increased Probability of Large Earthquakes in Different Regions of the World: Algorithm CN. *Phys. Earth Planet. Inter.* **1990**, *61*, 57–72. [CrossRef]
86. Falcone, G.; Console, R.; Murru, M. Short-Term and Long-Term Earthquake Occurrence Models for Italy: ETES, ERS and LTST. *Ann. Geophys.* **2010**, *53*, 41–50. [CrossRef]
87. Woessner, J.; Christophersen, A.; Zechar, J.D.; Monelli, D. Building Self-Consistent, Short-Term Earthquake Probability (STEP) Models: Improved Strategies and Calibration Procedures. *Repos. Publ. Res. Data (ETH Zur.)* **2010**, *53*, 141–154. [CrossRef]
88. Spassiani, I.; Falcone, G.; Murru, M.; Marzocchi, W. Operational Earthquake Forecasting in Italy: Validation after 10 Yr of Operativity. *Geophys. J. Int.* **2023**, *234*, 2501–2518. [CrossRef]
89. Eberhart-Phillips, D. Aftershock Sequence Parameters in New Zealand. *Bull. Seismol. Soc. Am.* **1998**, *88*, 1095–1097. [CrossRef]
90. Lolli, B.; Gasperini, P. Comparing Different Models of Aftershock Rate Decay: The Role of Catalog Incompleteness in the First Times after Main Shock. *Tectonophysics* **2006**, *423*, 43–59. [CrossRef]

91. Utsu, T. A Statistical Study on the Occurrence of Aftershocks. *Geophys. Mag.* **1961**, *30*, 521–605.
92. Gutenberg, B.; Richter, C. *Seismicity of the Earth and Associated Phenomena*; Princeton University Press: Princeton, NJ, USA, 1954.