Forecasting Strong Subsequent Earthquakes in Japan Using NESTORE Machine Learning Algorithm: preliminary results

EGU General Assembly 2024
Vienna, Austria & Online | 14–19 April 2024

S. Gentili[1*], G. D. Chiappetta[1], G. Petrillo[2,3], P. Brondi[1], J. Zhuang[2], R. Di Giovambattista[4]
[1] National Institute of Oceanography and Applied Geophysics - OGS, Italy, *sgentili@ogs.it;
[2] The Institute of Statistical Mathematics – ISM, Japan;
[3] Scuola Superiore Meridionale – SSM, Italy;
[4] Istituto Nazionale de Geofisica e Vulcanologia - INGV, Roma, Italy;

Sharing is encouraged

ABSTRACT

NESTORE (Next STrOng Related Earthquake) is a machine learning algorithm for forecasting strong aftershocks during ongoing earthquake clusters. It has already been successfully applied to Italian, Greek and Californian seismicity in the past. A free version of the software in MATLAB (**NESTOREv1.0**) is available on **GitHub**. The method is trained on the region under investigation using seismicity characteristics. The obtained region-specific parameters are used to provide **the probability, for the ongoing clusters, that the strongest aftershock has a magnitude greater than or equal to that of the mainshock - 1**. If this probability is **greater than or equal to 0.5**, the cluster is labeled as **type A**, otherwise as **type B**. The current version of the code is modular and the cluster identification method is based on a window approach, where the size of the spatio-temporal window can be adjusted according to the characteristics of the analyzed region. In this study, we applied **NESTORE** to the **seismicity of Japan** using the Japan Meteorological Agency (JMA) catalogue from 1973 to 2022. To account for the highly complex seismicity of the region, we replaced the cluster identification module with software that uses a stochastic declustering approach based on the ETAS model. The analysis is performed in increasing time intervals after the mainshock, starting a few hours later, to simulate the evolution of knowledge over time. The analysis showed a high prevalence of clusters where there are no strong earthquakes, leading to an imbalance between type A and type B classes. NESTORE was **trained with data from 1973 to 2004** and **tested from 2005 onwards**. The large imbalance in the data was mitigated by carefully analyzing the training set and developing techniques to remove outliers. The **cluster type forecasting was correct in 93% of cases**.

PREVIOUS STUDY CASES

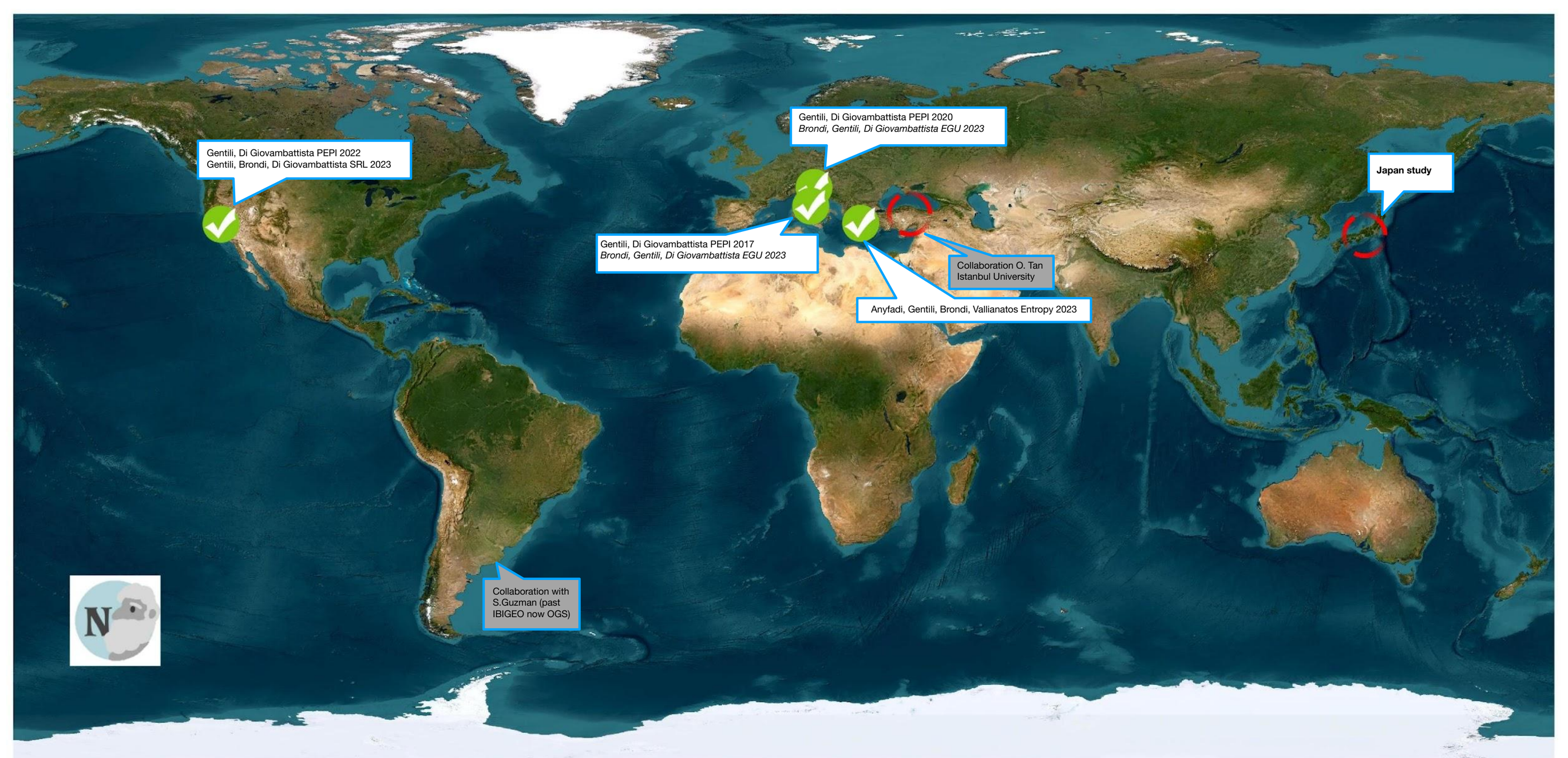
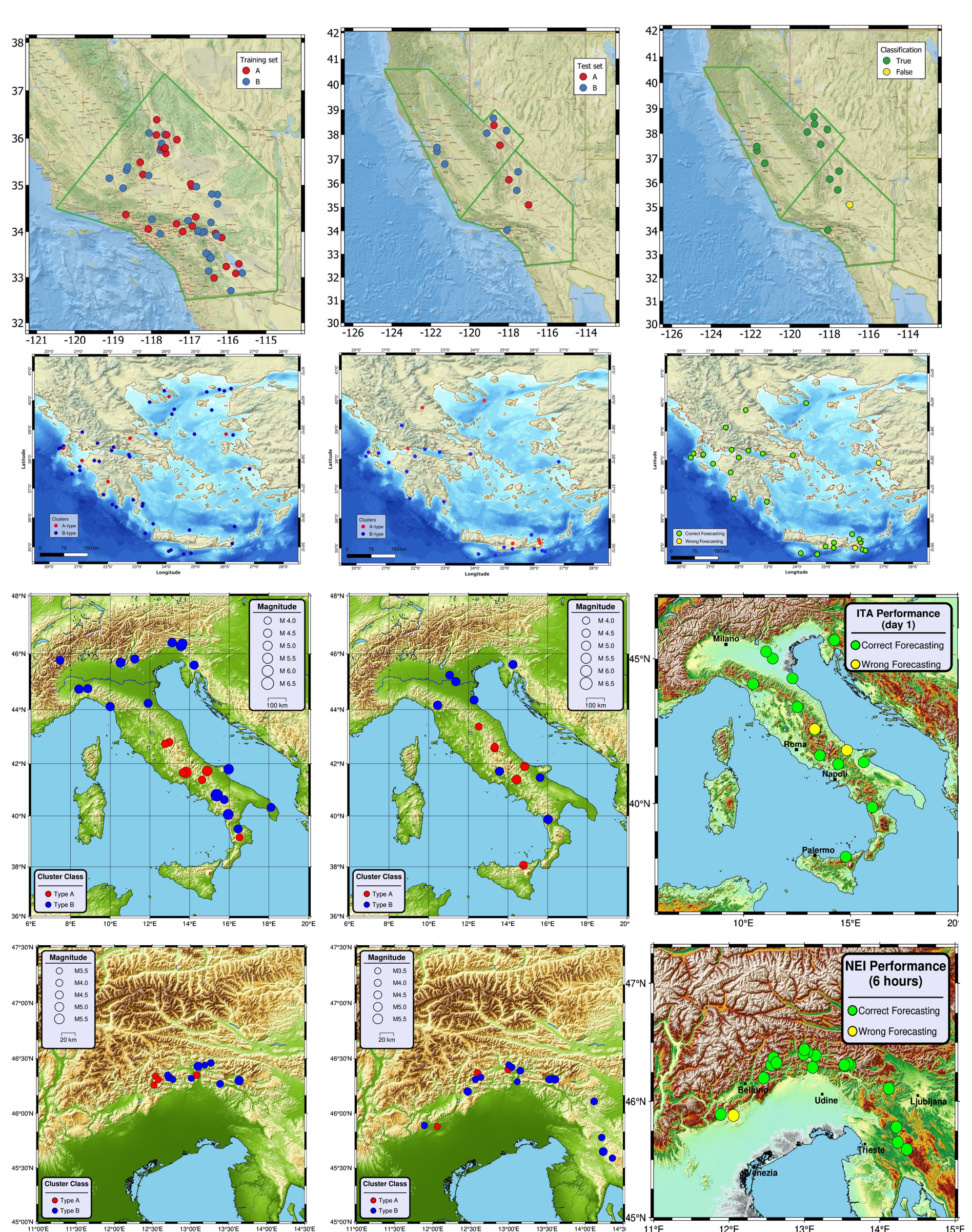


Fig. 1: World map showing the regions where NESTORE has already been applied.



California: SCEC + Comcat
Training: 50 clusters, **21 A** and **29 B**
Testing: 11 clusters, **4 A** and **9 B**
Best performances T=12h
Correct: 92%

Greece: AUTH catalogue
Training: 46 clusters, **6 A** and **40 B**.
Testing: 27 clusters, **6 A** and **21 B**
Best performances T=6h
Correct: 92%

Italy: Lolli and Gasperini (2006) + ISIDE catalogues
Training: 24 clusters, **7 A** and **17 B**
Testing: 14 clusters, **6 A** and **8 B**
Best performances: T=1 day
Correct: 86%

NE-Italy and Slovenia: OGS catalogue
Training: 13 clusters, **4 A** and **9 B**.
Testing: 18 clusters, **3 A** and **15 B**
Best performances: T=6h
Correct: 94%

Fig. 2: Maps of the training-set (left panels), test-set (middle panels) and classifications (right panels) for the previous study cases.

METHOD

- **NESTOREv1.0** software package consists of four main parts:
  - **cluster identification module**
  - **training module**
  - **testing module**
  - **NRT classification (1 cluster)**
- All modules can be launched separately by **command line**;
- **The software is available on GitHub** and the corresponding paper has been published in SRL;
- Link: https://github.com/StefaniaGentili/NESTORE

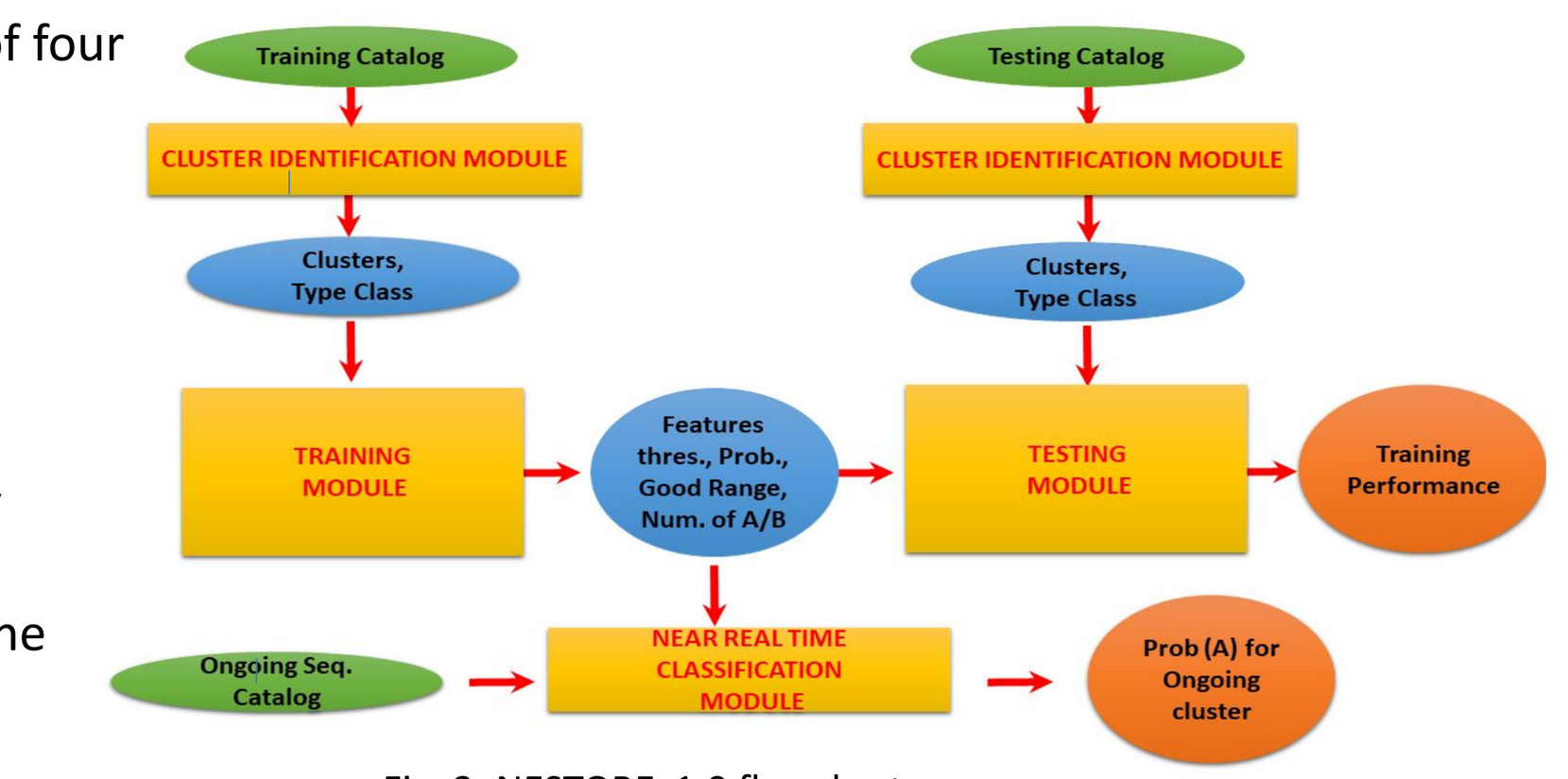


Fig. 3: NESTOREv1.0 flowchart.

Table 1 - The nine features considered by NESTOREv1.0.

Number	Feature Name	Definition	Type
1	Number of events	$N_2(i) = \sum_i n_i$ if $m_i \geq M_m - 2$ where m_i = magnitude of the i^{th} event, M_m = mainshock magnitude	Number and Spatial distribution of events
2	Linear Concentration of events	$Z(i) = \frac{mean(10^{0.69m_i-3.22})}{mean(r_{ij})}$ where r_{ij} = distance between the generic i^{th} and j^{th} event	
3	Normalized event Source area	$S(i) = \sum_i 10^{(m_i-M_m)}$ where m_i = magnitude of the i^{th} event, M_m = mainshock magnitude	Source area and Magnitude trend along time
4	Cumulative deviation of S from the long-term trend on increasing length windows	$SLCum(i) = \sum_i abs[S(t_i) - S(t_{i-1})\frac{i \cdot dt}{(i-1) \cdot dt}$ Where S calculated on $[s_1,t_i]$, $t_i = s_1 + i \cdot dt$, dt = 6 hours	
5	Cumulative deviation of S from the long-term trend on sliding windows	$SLCum2(i) = \sum_i abs[S([s_1 + (i-1) \cdot dt, s_1 + i \cdot dt]) - S([s_1 + (i-1) \cdot dt, s_1 + (i-1) \cdot dt + d\tau])\frac{dt}{d\tau}]$ Where S calculated over the generic time interval [a, b]	
6	Cumulative variation of magnitude from event to event	$V_m(i) = \sum_i \lvert m_i - m_{i-1} \rvert$ where m_i = magnitude of the i^{th} event, m_{i-1} = magnitude of previous one	
7	Normalized radiated Energy	$Q(i) = \frac{\sum_i E_i}{E_m}$ where E_m = energy of the mainshock, E_i = energy of the i^{th} event	Energy trend along time
8	Cumulative deviation of Q from the long-term trend on increasing length windows	$QLCum(i) = \sum_i abs[Q(t_i) - Q(t_{i-1})\frac{i \cdot dt}{(i-1) \cdot dt}$ Where Q calculated on $[s_1,t_i]$, $t_i = s_1 + i \cdot dt$, dt = 6 hours	
9	Cumulative deviation of Q from the long-term trend on sliding windows	$QLCum2(i) = \sum_i abs[Q([s_1 + (i-1) \cdot dt, s_1 + i \cdot dt]) - Q([s_1 + (i-1) \cdot dt, s_1 + (i-1) \cdot dt + d\tau])\frac{dt}{d\tau}]$ Where Q calculated over the generic time interval [a, b]	

Starting six hours after the occurrence of a strong mainshock of magnitude M_m, NESTOREv1.0, computes nine features related to the evolution of the number of events with $M \geq M_m - 2$ and their spatial-magnitude distribution over time in the first hours after the mainshock. The values of these features are used to train nine independent one-node decision trees in order to forecast the probability that the clusters are type A. In this study we introduced $N2_S$ which has higher performance than N2 itself. It is defined as:

$$N2_S = 110 \cdot S + N2$$

NESTORE automatically selects the best classifiers, and combines the results using Bayes' theorem.

APPLICATION TO JAPAN

We selected the study region based on the completeness magnitude and the coverage of the seismic network. Studying Japan seismicity has advantages/disadvantages:

Advantages:

- The very good quality of the JMA catalogue especially for inland area;
- High level of seismicity.

Disadvantages:

- Simple cluster identification methods do not work properly;
- Clusters with a small magnitude mainshock can not be easily considered isolated;
- Volcanic and tectonic seismicity close in space;
- Strong unbalancing of A-type and B-type classes;
- Very deep intraslab clusters with different behavior respect to shallower ones.

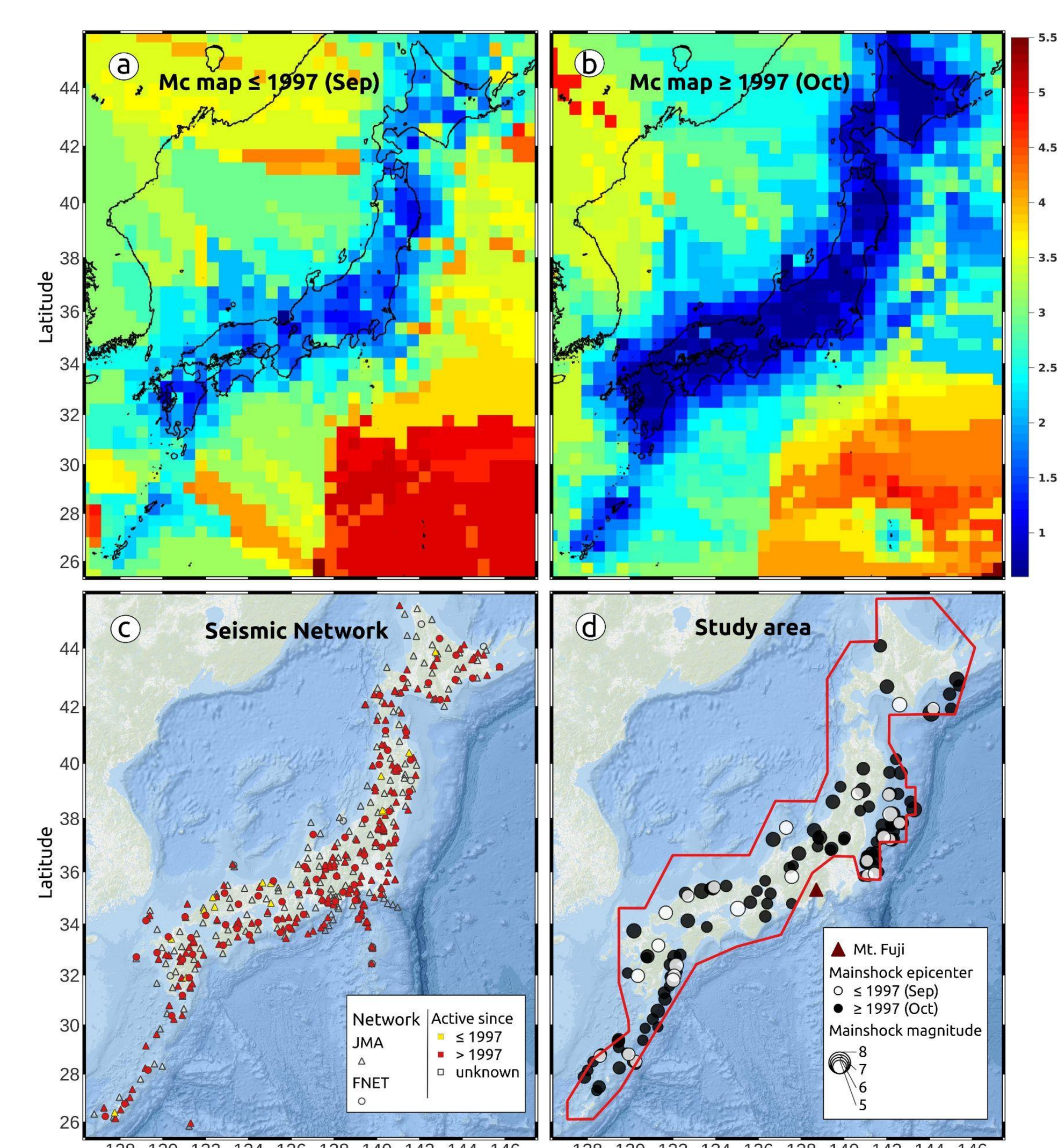


Fig. 4: Mc map computed on the available catalog before (a) and after (b) October, 1997. c) Map of the seismic stations of the JMA and the F-NET networks. d) Study area (red line) selected for the NESTORE application showing the epicenter of the mainshock of the initial set of clusters and the location of the Mt. Fuji.

CLUSTER IDENTIFICATION

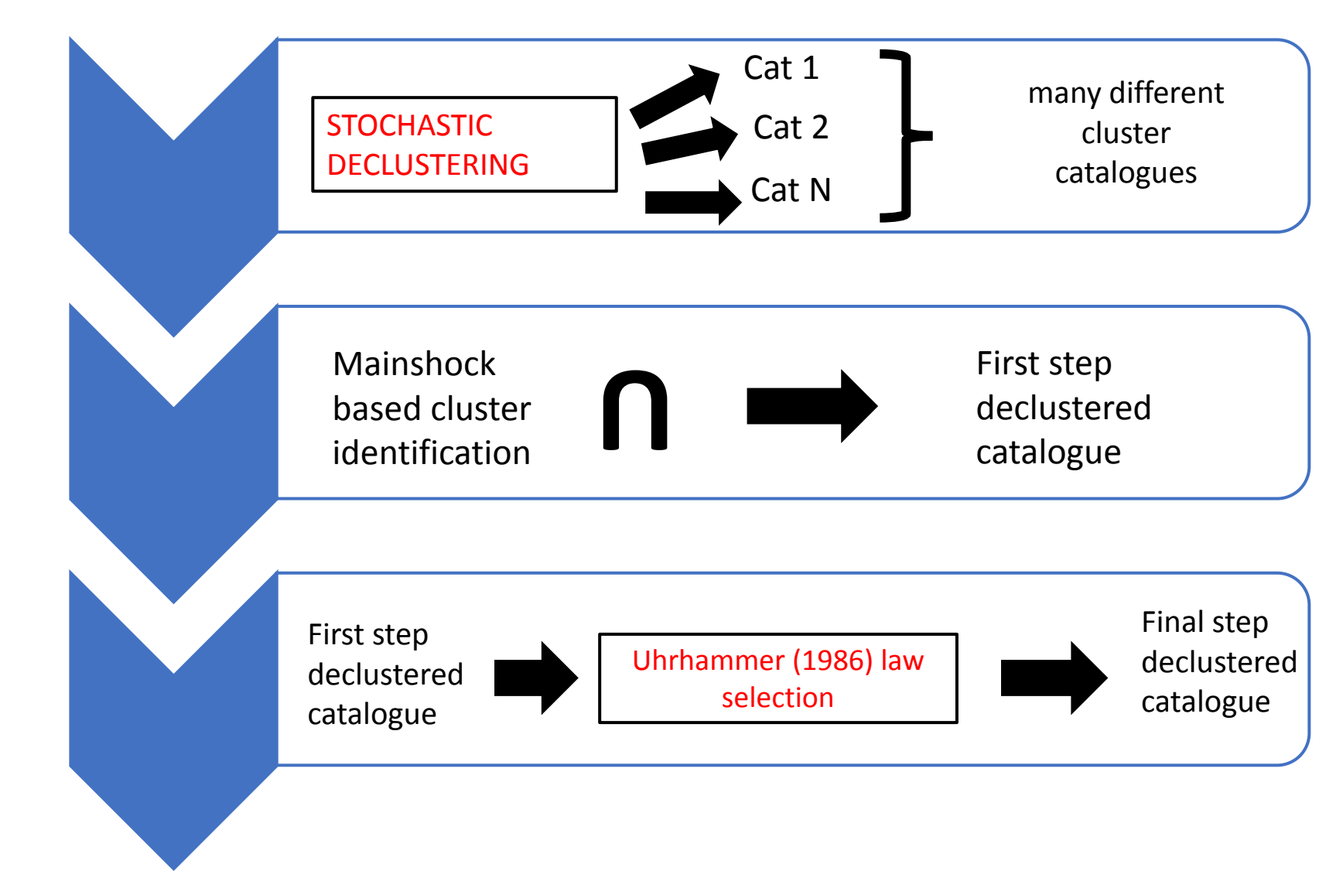
In this work we applied a combined cluster identification approach using the Stochastic Declustering method (Zhuang et al., 2002) based on ETAS:

STOCHASTIC DECLUSTERING

1. For each event j estimate the probability of the event j being triggered by a previous event i and the probability to be an offspring $\varphi_j = 1 - \sum_{i=1}^{j-1} \rho_{ij}$
2. For each event j:

   i) Generate a uniform random number U_j in [0,1]
   ii) If $U_j < \varphi_j$ => the j-th event is considered a background event
   iii) Otherwise, the j-th event is considered to be a descendant of the I-th event where I is the smallest number such that $U_j < \varphi_j + \sum_{i=1}^{I} \rho_{ij}$

COMBINED CLUSTER IDENTIFICATION APPROACH



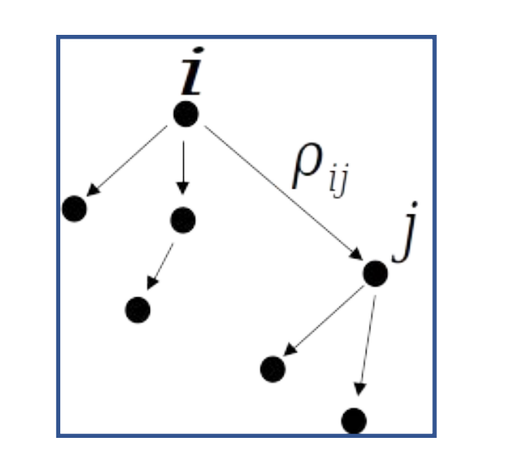


Fig. 5: Graph of the triggering probability.

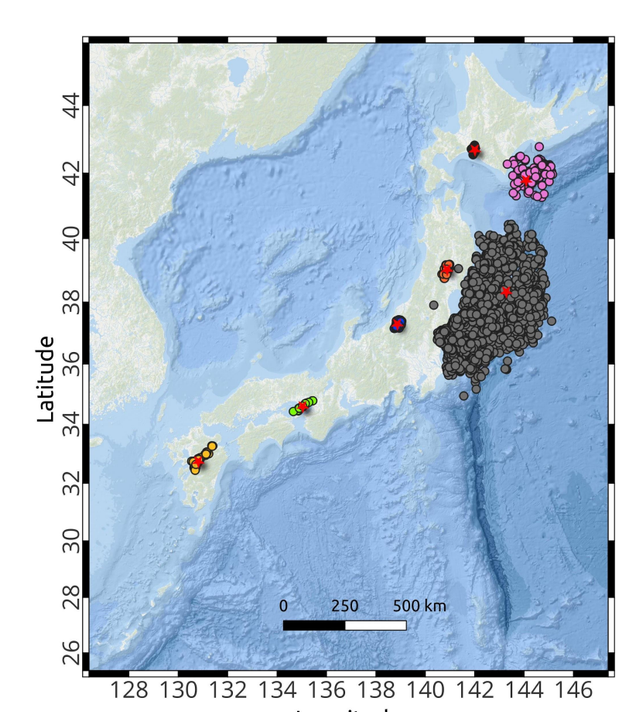


Fig. 6: Examples of clusters. Grey circles: Tōhoku seismic sequence of 2011.

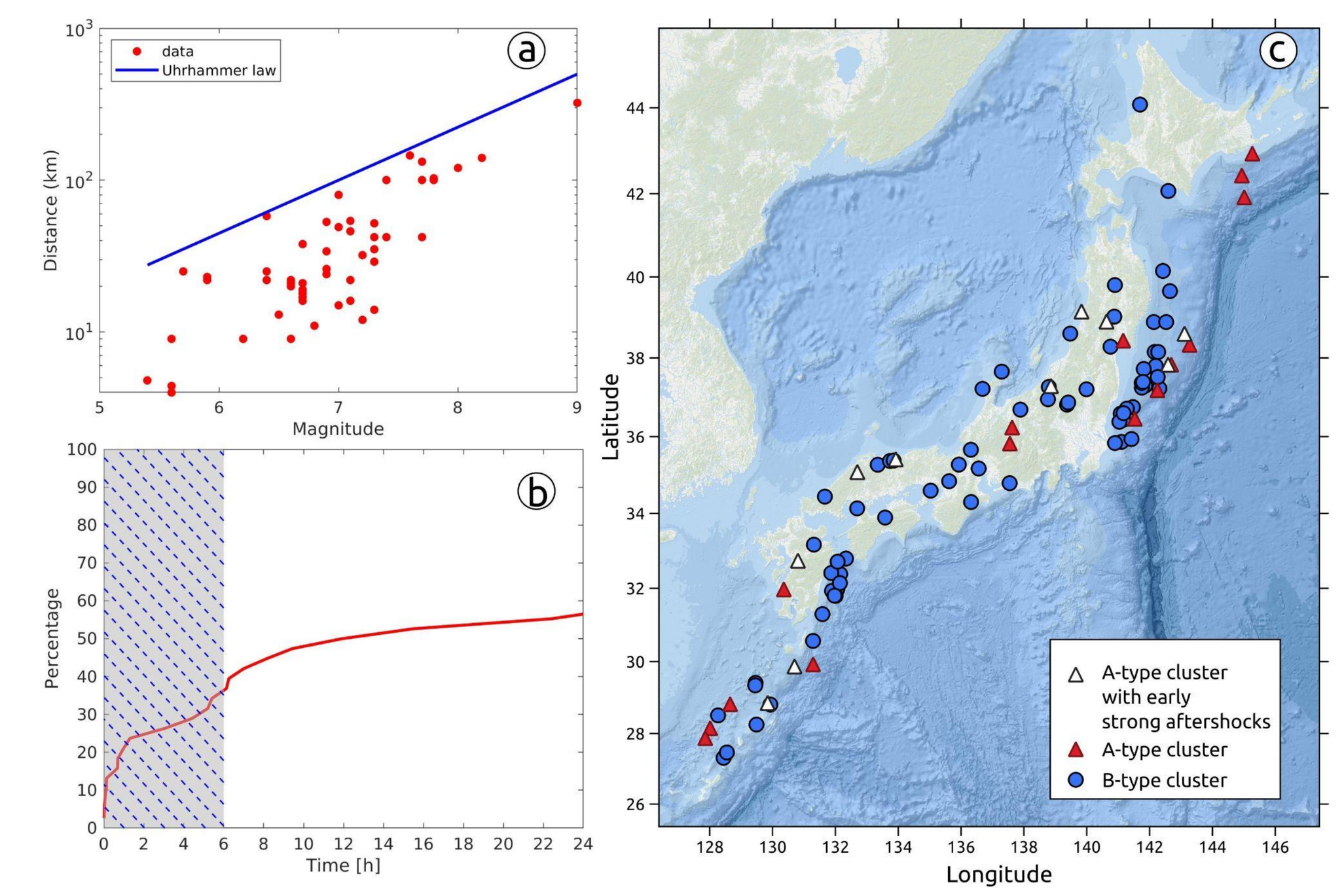


Fig. 7: a) Magnitude vs maximum radius in kilometers of the most representative clusters of our database and comparison with Uhrhammer (1986) space-law; b) Percentage of the A-type clusters having an early strong aftershock during the first day, the shadowed area highlights the percentage in the first 6 hours. c) Position of the mainshocks of the A- and B-type clusters. The A-type clusters with a strong early aftershock (before 6 hours) are represented by white triangles.

FEATURES

Features	Time intervals (days)		
	0.25	0.5	0.75
S	dark green	light green	light green
SLcum	white	dark green	light green
QLcum	white	dark green	light green
SLcum2	white	white	dark green
QLcum2	white	white	dark green
$N2_S$	dark green	dark green	light green

In Japan **6 features supplied reliable results in discriminating A and B classes**. They are the cumulative source area (S), two features depending on the deviation of **S** from long-term trend (**SLCum, SLCum2**) two features depending on the deviation of cumulative energy from long-term trend (**QLCum, QLCum2**) and the new feature $N2_S$ depending both on S and on the events number.

In **dark green** we outline the time intervals in which they supply reliable results. For longer time intervals, both the features and the corresponding thresholds are inherited from previous ones (**light green**). White color means not used time intervals.

REFERENCES

Anyfadi E.-A., Gentili S., Brondi P., Vallianatos F. Forecasting Strong Subsequent Earthquakes in Greece with the Machine Learning Algorithm NESTORE. Entropy, **25**, 797.
Brondi P., Gentili S. and Di Giovambattista R.; 2023: Forecasting strong aftershocks in the Italian territory: a National and Regional application for NESTOREv1.0 EGU 2023.
Gentili S. and Di Giovambattista R.; 2017: Pattern recognition approach to the subsequent event of damaging earthquakes in Italy. PEPI, **266**, pp. 1-17.
Gentili S. and Di Giovambattista R.; 2020: Forecasting strong aftershocks in earthquake clusters from northeastern Italy and western Slovenia. PEPI, **303**, 106483.
Gentili S. and Di Giovambattista, R.; 2022: Forecasting strong subsequent earthquakes in California clusters by machine learning, PEPI, **327**, 106879.
Gentili S., Brondi, P. and Di Giovambattista, R.; 2023: NESTOREv1.0: A MATLAB Package for Strong Forthcoming Earthquake Forecasting, SRL, **94**, 2003–2013.
Uhrhammer, R. (1986), Characteristics of Northern and Central California Seismicity, Earthquake Notes, **57**, 21-37.
Zhuang J., Ogata Y. and Vere-Jones D.; 2002: Stochastic Declustering of Space-Time Earthquake Occurrences, Journal of the American Statistical Association, **97**, 369-380.

ANALYSIS AND RESULTS

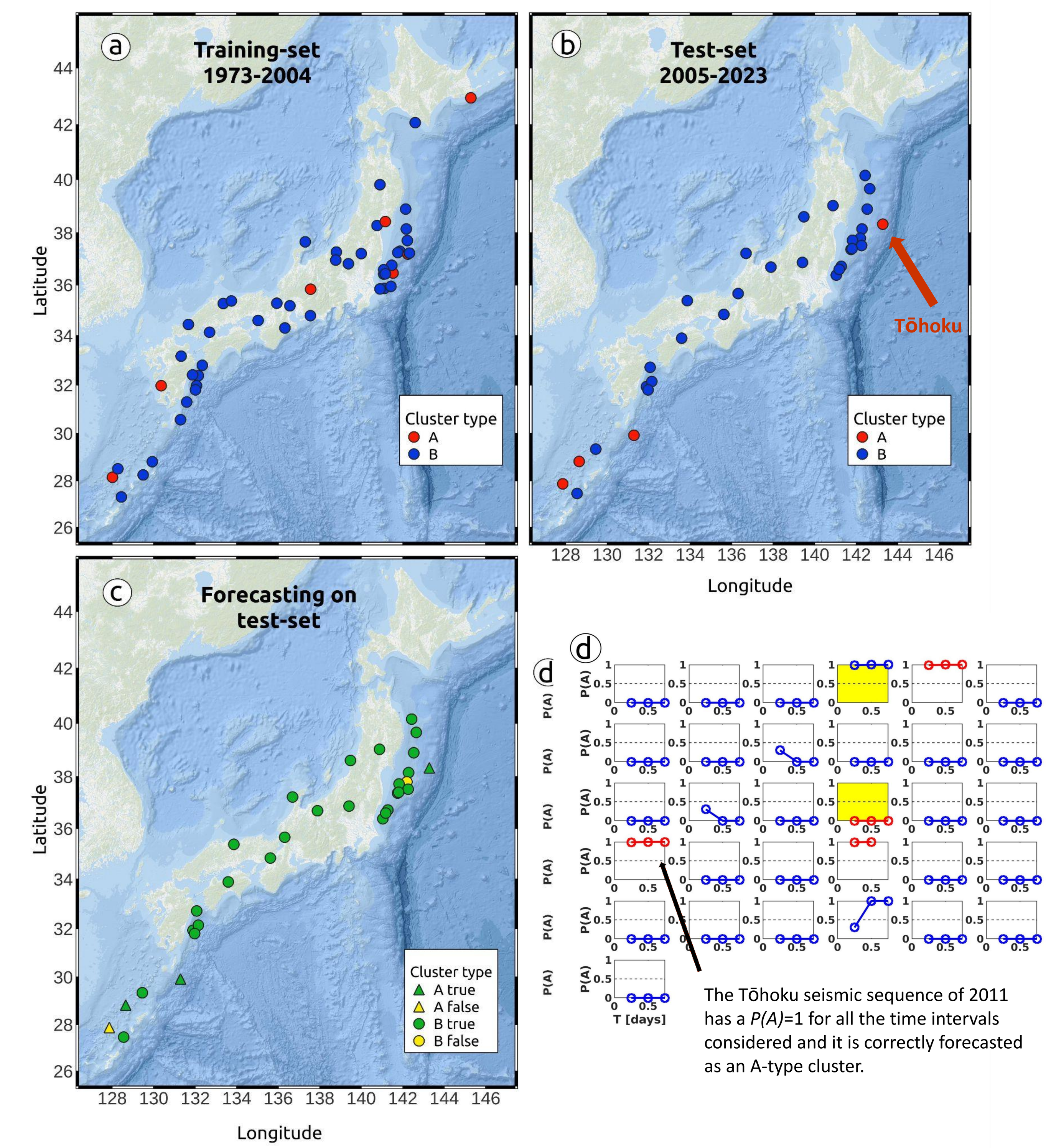


Fig. 8: a) Training-set, b) test-set, c) results of the forecasting. All these plots refer to T_i=6h. d) Probability of being an A-type cluster *vs* time.

- *Training*: 50 clusters, **7 A** and **43 B**, from 1973 to 2004.
- *Testing*: 31 clusters, **4 A** and **27 B**, from 2005 to 2023.
- **The best performances** have been obtained for $N2_S$ and **S** at the time interval of **6h**. At this time interval we also have the highest value of Precision, Recall, True Positive Rate and the lowest value of False Positive Rate.
- The accuracy of the forecasting on the test-set slightly decreases with increasing time intervals but still with high values. **The overall forecasting is successful in 93% of cases at 6h** and 90% of cases at 12 and 18h.

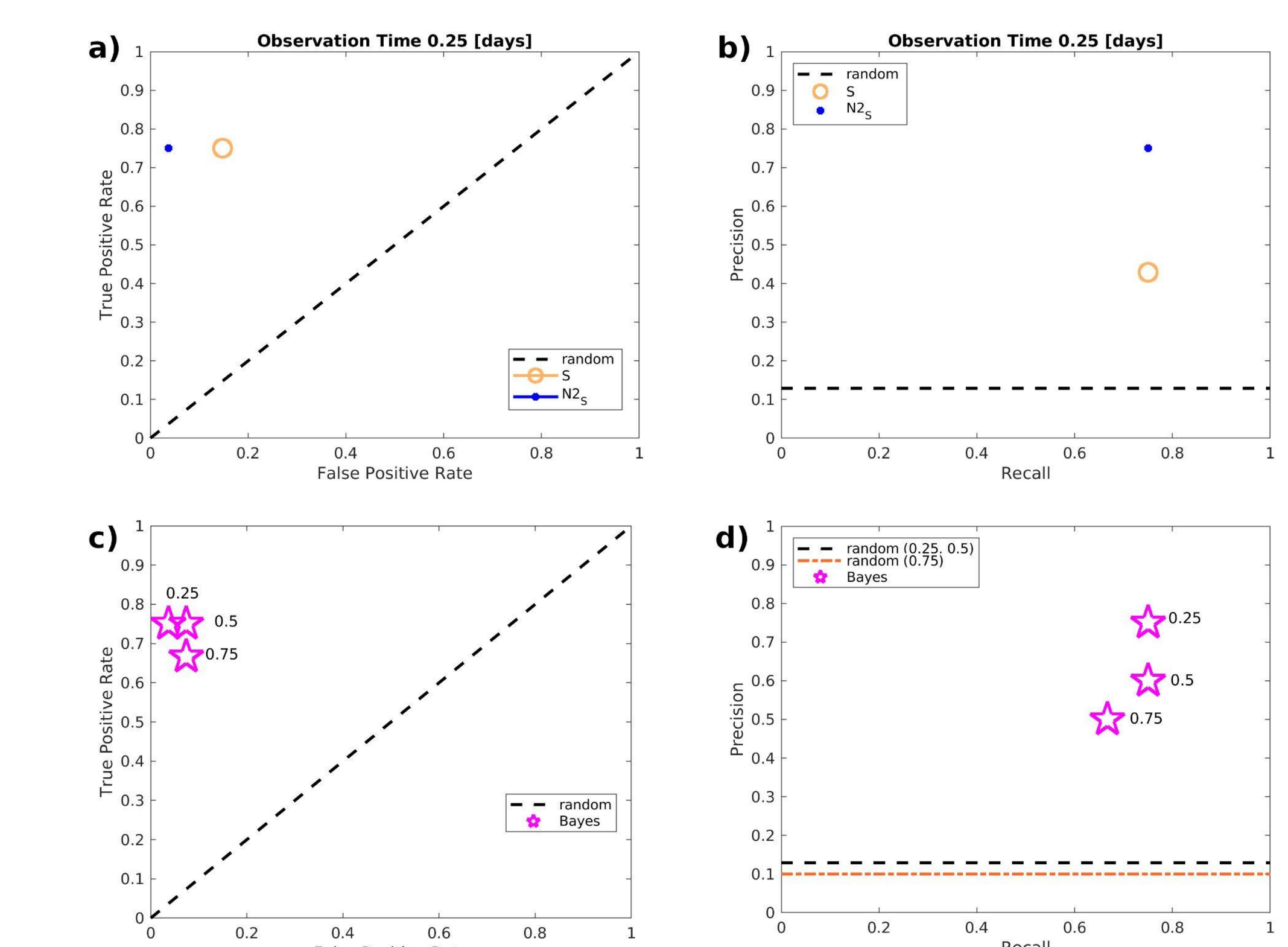


Fig. 9: Most representative results of the testing procedure showing the performances of the best features $N2_S$ and S at 6h after the mainshock and of the overall Bayesian classification in terms of ROC (a, c) and Precision-Recall (b, d).

Funded by: NEMESIS; Ministry of Foreign Affairs and International Cooperation; Return