Geophysical Journal International

Geophys. J. Int. (2025) **240,** 81–95
Advance Access publication 2024 October 18
GJI Seismology

https://doi.org/10.1093/gji/ggae373

Forecasting future earthquakes with deep neural networks: application to California

Ying Zhang,[1,2,3] Chengxiang Zhan,[4] Qinghua Huang[2] and Didier Sornette[5]

[1] *School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing* 100083, *China*
[2] *Department of Geophysics, School of Earth and Space Sciences, Peking University, Beijing* 100871, *China. E-mail: huangq@pku.edu.cn*
[3] *Beijing Engineering Research Center of Industrial Spectrum Imaging, Beijing* 100083, *China*
[4] *School of Science, China University of Geosciences (Beijing), Beijing* 100083, *China*
[5] *Institute of Risk Analysis, Prediction and Management (Risks-X), Southern University of Science and Technology (SUSTech), Shenzhen* 518055, *China*

Accepted 2024 October 14. Received 2024 September 25; in original form 2024 June 1

SUMMARY

We use the spatial map of the logarithm of past estimated released earthquake energies as input of fully convolutional networks (FCN) to forecast future earthquakes. This model is applied to California and compared with an elaborated version of the epidemic type aftershock sequence (ETAS) model. Our long-term earthquake forecast simulations show that the FCN model is close to the ETAS model in forecasting earthquakes with $M \geq 3.0,\ 4.0,$ and 5.0 according to the Molchan diagram. Moreover, training and implementing the FCN model is 2000–4000 times faster than calibrating the ETAS model and generating its probabilistic forecasts. The FCN model is straightforward in terms of its neural network structure and feature engineering. It does not require extensive knowledge of statistical seismology or the analysis of earthquake catalogue completeness. Using the earthquake catalogue with $M \geq 0$ as FCN input can enhance the model's performance in some time–magnitude forecasting windows.

Key words: Machine learning; Probabilistic forecasting; Earthquake interaction, forecasting and prediction; Statistical seismology.

1 INTRODUCTION

Despite the persisting uncertainties concerning the feasibility of earthquake prediction, the pursuit of methods to forecast seismic events remains an active domain of research in the scientific community (Huang 2015; Kossobokov & Soloviev 2021). Given their performance in discovering hidden patterns in very large data sets, it is not surprising that machine learning or artificial intelligence techniques have become popular in seismology (Kong *et al.* 2018), with applications to earthquake detection and phase picking, earthquake early warning, ground-motion prediction, seismic tomography, earthquake geodesy, as well as earthquake prediction and forecasting.

Efforts to use machine learning for earthquake prediction or forecasting began as early as 30 yr ago. Among seismologists, there has been a growing debate in recent years over whether current machine learning models genuinely offer a breakthrough in earthquake prediction and forecasting. Let us review briefly some of the most prominent viewpoints and arguments in this debate. DeVries et al (2018) used a deep neural network with six hidden layers and 13 451 parameters to forecast the spatial distribution of aftershocks. However, Mignan & Broccardo (2019) reformulated these results using a two-parameter logistic regression (that is, one neuron) and obtained the same performance as that of reference (DeVries *et al.* 2018); moreover, they found that the accuracy of such a deep and complex neural network is no better than that provided by a simple empirical law.

Before systematically introducing AI-based research on earthquake prediction and earthquake forecasting, it is important to clearly distinguish between the concepts of 'earthquake prediction' and 'earthquake forecasting'. An example of an earthquake prediction published on a given day (say 2024 September 23) would be: 'A magnitude 7.0 earthquake will occur in the San Francisco Bay Area at 2:30 p.m. on 2024 October 12.' Earthquake prediction thus refers to the attempt to specify the exact time, location and magnitude of a future earthquake, with a high degree of precision. An example of an earthquake forecast published on a given day (say 2024 September 23) would be: 'there is a 5 per cent chance of a magnitude 5 or greater earthquake occurring in the San Francisco bay area within the next 30 d'. Earthquake forecasting focuses on providing probabilistic estimates that indicate the likelihood of earthquakes occurring in a given region within a specific period and in a given magnitude range.

After reviewing 77 articles from 1994 to 2019 on the application of artificial neural networks (ANN) to earthquake forecasting, Mignan & Broccardo (2020) found that these previous ANN models do not seem to provide new insight into earthquake predictability. They report that only 47 per cent of these works have been compared to a baseline. Among these models with a baseline, 22 per cent use a Poisson null hypothesis or randomized data as a baseline, while the remaining 78 per cent are compared to other machine-learning methods. To prove the superiority of a new earthquake prediction model, it should compete with powerful rivals, such as the epidemic type aftershock sequence (ETAS) model. Comparing performance against the simplistic Poisson null hypothesis or other basic machine-learning methods can create a misleading sense of achievement (Zhang *et al.* 2024). It is akin to believing a novice boxer could defeat Mike Tyson in boxing just because he has outmatched a toddler.

We collected relevant papers on earthquake prediction and forecasting using machine learning from 1994 to the present (Q4-2023), including conference papers, science citation index (SCI) papers, and other publications. 'Other publications' include papers in non-SCI indexed journals, a chapter in a monograph, and so on. The search method involved pairing keywords 'Deep Learning, Artificial Intelligence, Artificial Neural Network' and 'Earthquake Prediction, Earthquake Forecast, Earthquake Forecasting, Aftershock Prediction, Aftershock Forecast, Short-term prediction, Long-term prediction, Operational Earthquake Forecasting, Seismic Precursor' and conducting searches on Google Scholar. In addition, we supplemented our article database with the references listed in the three review papers (Banna *et al.* 2020; Mignan & Broccardo 2020; Ridzwan & Yusoff 2023). This comprehensive investigation led us to select 136 papers that utilized machine learning techniques to construct earthquake prediction or forecasting models, and the information about these papers is available at https://github.com/AI-earthquake/citations-for-Predicting-future-Californian-earthquakes-with-deep-neural-networks/tree/main. Fig. 1 shows the annual publication count of these articles, the impact factors of SCI papers, with the impact factor being the latest for 2022–2023. In terms of quantity, one can observe that the use of machine learning for earthquake prediction has seemingly been in a new active phase over the past three years.

In our survey extending that of Mignan and Broccardo, we arrive at similar conclusions. Only 66.9 per cent of works using AI to predict earthquakes were compared to a baseline, most of which are other machine-learning methods or simple Poisson null hypothesis. Four recent works standout by comparing their AI models with geophysical or statistical seismology models. Chen *et al.* (2020) developed a deep neural network that uses data from the Seismic Research Center Moment Tensor Catalogue finite fault data and the International Seismological Centre earthquake catalogue as input to forecast aftershocks of the 2008 Wenchuan earthquake and the 2011 Tohoku earthquake. They compared the AI model with the Coulomb failure stress change model. Their results showed that the AI model's performance is on par with that of the traditional physical model while being computationally more efficient. Dascher-Cousineau *et al.* (2023) developed a deep learning model to forecast only the occurrence time of earthquakes in small regions of Southern California, while magnitudes and spatial dependences are not considered. The benchmark is taken as a time-only ETAS-like model transformed with an exponential (rather than linear in the standard ETAS model; Ogata 1988) dependence on the past earthquake history. They claim a better performance than their 'minimum performance benchmark' (in their own words). Zlydenko *et al.* (2023) used an ANN encoder that imitates the mathematical structure of ETAS, with response functions learned by the ANN based on a number of features such as time intervals between present time and past earthquakes, distances, magnitudes and so on. This ANN model performs on par with, or even surpasses, a standard ETAS model in terms of average information gain per earthquake while achieving a 1000-fold reduction in computation time. Another valuable attribute of this ANN is that it is able to learn some anisotropic spatial dependence of the spatial kernel of earthquake triggering, in agreement with the known tendency of earthquakes to cluster along faults, thus improving on the ETAS isotropic spatial kernels. Stockman *et al.* (2023) used an extended temporal neural model to forecast short-term seismicity, and the neural point process is formulated similarly to the ETAS model but with a much more flexible way of representing the intensity function. They found that the neural model outperforms the ETAS model and is faster to train.

Additionally, in the professional earthquake forecasting community, it is essential to conduct both prospective and pseudo-prospective testing of the developed models. Prospective testing assesses forecasts against future events. For example, a forecast made and archived in real time is later evaluated by comparing it

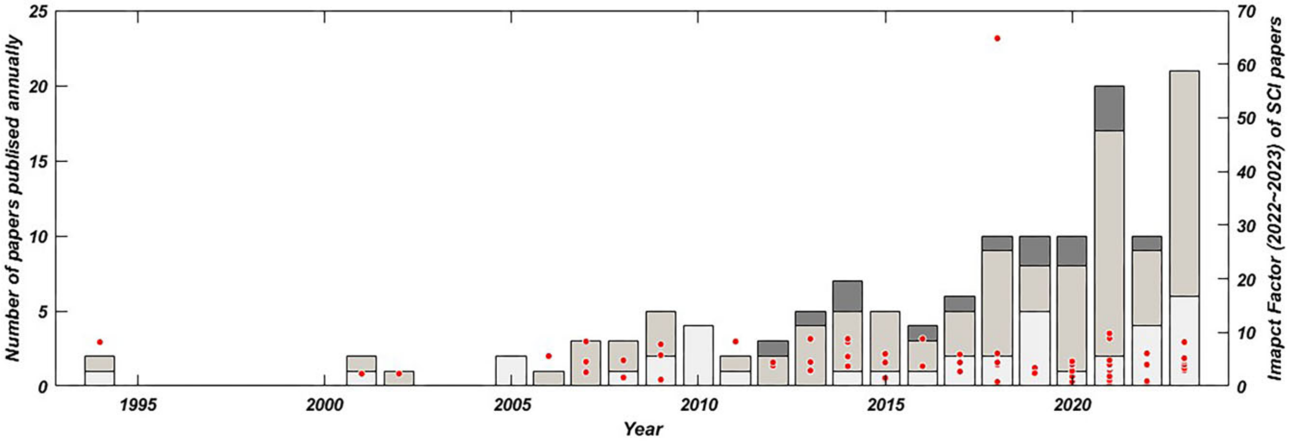


Figure 1. Annual rate of publications (histogram) and the impact factors of the published SCI papers over the period 1994–2023 obtained from a comprehensive survey of 136 papers on the topic of earthquake prediction using machine learning techniques. 'Other paper' refers to articles that are neither conference papers nor papers published in SCI journals, such as those published in journals not included in the SCI database. In total, there are 136 papers, comprising 37 conference papers, 84 SCI papers, and 15 other papers.

with earthquake data collected from the forecast issue date until the end of the testing period, also known as the forecast horizon. Prospective testing is thus the gold standard for assessing the performance of a forecasting system. However, prospective testing, while essential for fully validating the performance of a forecasting model, presents several challenges. It requires a well-organized framework for real-time forecasting and data collection, often involving collaboration between institutions and consistent monitoring over extended periods. The major difficulty is the inherent time-consuming nature of this approach—since one must wait for future earthquake events to occur before assessing the model's accuracy. This waiting period can span months or even years, making the process slow, resource-intensive and impractical for rapidly exploring and testing novel forecasting models that could advance the state of the art.

In contrast, pseudo-prospective testing simulates this process while maintaining time-dependent causality. In this approach, models are calibrated using data up to a specific time t_0 in the past and then used to generate forecasts for periods beyond t_0 as though the modelers have no knowledge of events after t_0. Pseudo-prospective testing offers the flexibility of 'moving through time.' Researchers can retrospectively select a point in the past, calibrate their models up to that point and immediately test their forecasts against known future data. This accelerates the evaluation process, allowing for quicker iterations and refinements. Of the four studies mentioned above, three (Dascher-Cousineau *et al.* 2023; Stockman *et al.* 2023; Zlydenko *et al.* 2023) explicitly indicate that their models underwent pseudo-prospective testing, which is a highly valuable step towards practical earthquake forecasting.

In summary, this short review reveals that existing efforts to use machine learning for earthquake prediction have failed to convincingly demonstrate superiority, primarily because they have not rigorously compared themselves to the strongest established rival methods, such as the most advanced ETAS model. Indeed, machine learning methods need to recognize the existence of the mature field of statistical seismology that represents the state-of-the-art in accounting for spatio-temporal patterns of earthquakes. Statistical seismology builds first on the hypothesis that 'spatial location and magnitude distribution of the past earthquakes reveal the probable location and size of the future events' (Helmstetter *et al.* 2006; Nandan *et al.* 2019a). A second ingredient is the observation that earthquakes also cluster in time according to the Omori–Utsu law. Statistical seismicity models include (i) the ETAS model (Kagan & Knopoff 1981, 1987; Ogata 1988, 1998) sometimes integrating spatially variable (Nandan *et al.* 2017, 2021) and magnitude-dependent parameters (Nandan *et al.* 2019a, 2022), (ii) models based on the seismic quiescence hypothesis (Wesson & Ellsworth 1973; Kelleher & Savino 1975; Mogi 1979; Huang & Nagao 2002; Huang *et al.* 2002; Huang 2006, 2008), (iii) the forecasting model based on relative intensity index (Tiampo *et al.* 2002) and so on. Among these statistical seismological methods, ETAS is one of the most powerful models for describing the occurrence of earthquakes in space, time and magnitude. Moreover, it is also the best-performing model according to the current community-based prospective 'blind' testing protocols established by the collaboratory of the study of earthquake predictability. The ETAS model is a version of the self-excited conditional Hawkes point process (Hawkes 1971; Hawkes & Oakes 1974) adapted to earthquake modelling, in which the intensity of the seismicity at any location and time is determined by a weighted sum of contributions over all historical earthquakes. In the ETAS model, the Gutenberg–Richter law (Gutenberg & Richter 1944), Omori law (Utsu *et al.* 1995), the fertility law and spatial power-law kernel function, are used to simulate the spatio-temporal distribution of events triggered by background events and by previously triggered earthquakes.

In this study, we develop a fully convolutional network (FCN) that utilizes historical seismicity as input to forecast earthquakes with magnitudes of $M \geq 3.0,\ 4.0$, and 5.0 in the upcoming 15 to 90 d. We apply this model to California and conduct a comparative analysis with an enhanced version of the ETAS model, utilizing long-term pseudo-prospective earthquake forecasts from 2010 January 1 to 2020 December 31. Section 2 introduces the data set, comprising the earthquake catalogue spanning the entire California region. In Section 3, we provide an overview of the model architectures for both our FCN and the ETAS model. Comparison of the performances of the FCN and ETAS models are presented in Section 4. Section 5 shows how to transform the model output into an alert map. This approach enhances the suitability of our model for earthquake prediction applications. The discussion of our results and conclusion are presented in Sections 6 and 7, respectively.

2 DATA SET

The earthquake catalogue ($M \geq 0$, depth ≤ 40 km) used in this work is provided by the advanced national seismic system. We use the period from 1980 January 1 until 2020 December 31. The spatial distribution of earthquakes, frequency–magnitude relation, and cumulative number of events in time are presented in Fig. 2. The spatial extent of the study region is the regional earthquake likelihood models (RELM) polygon proposed by Schorlemmer & Gerstenberger (2007). The magnitude of completeness of this region varied with time, but is found never larger than 3.0 over the whole time interval considered here (Nandan *et al.* 2017).

The reason we chose California as the study region is that many earthquake prediction attempts have been conducted there (Gerstenberger *et al.* 2005; Helmstetter *et al.* 2006; Nandan *et al.* 2017, 2019b; Asencio–Cortés *et al.* 2018), providing informative comparisons.

3 MODEL ARCHITECTURE AND BENCHMARK

In this work, we use FCN to get the probability that earthquakes will occur in each grid cell ($0.1^\circ \times 0.1^\circ$) for different time and magnitude windows in California. In addition to the discrete space paved by squares of size $0.1^\circ \times 0.1^\circ$, We use discrete times with time step of 10 d. We put the FCN in competition with one of the strongest if not the strongest competitors, namely an advanced ETAS model. $(\vec{r}, t)$ are the location and time for the output of the time-space unit, and the output is $\Pr(\vec{r}, t, M, T)$, which is the probability that at least one earthquake with magnitude $\geq M$ will occur in the grid cell of $\vec{r}$ in the time window $[t+1,\ t+T]$. In this work, the threshold magnitude M of the target earthquakes is set to be in $\{3.0, 4.0$ and $5.0\}$ and the forecast time horizon T is set to be in $\{15, 30, 60$ and 90 d$\}$, therefore, we consider 12 different time–magnitude windows for forecasting earthquakes. To be clear, this does not correspond to target classes but rather to running 12 different experiments in parallel.

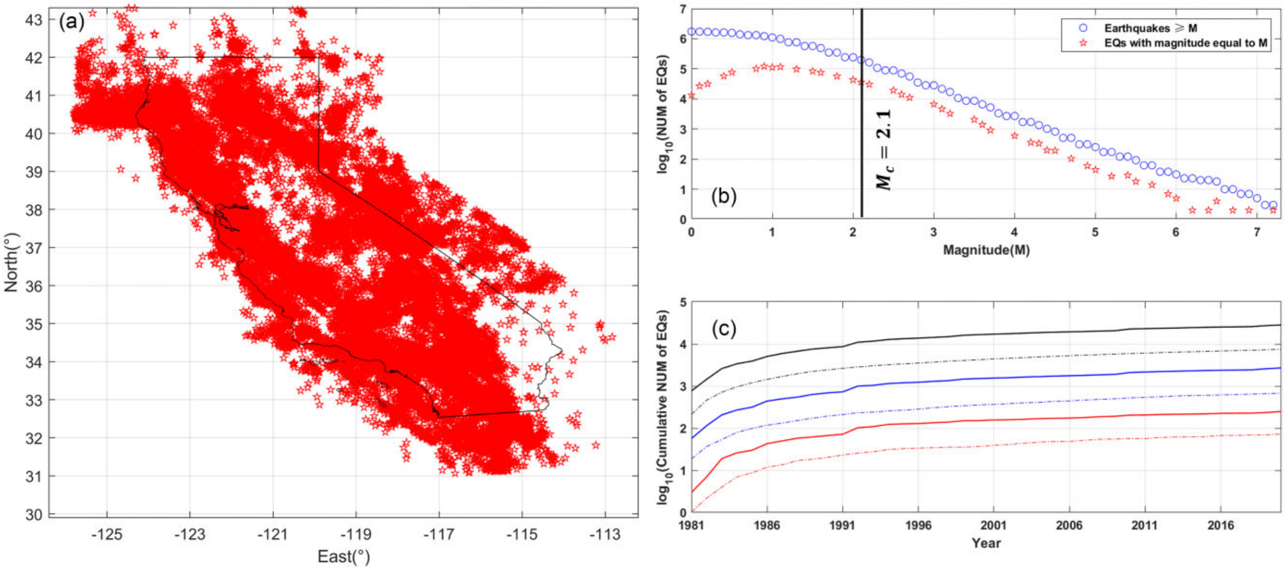


Figure 2. (a) Spatial distribution of earthquakes with magnitude larger than 0 that occurred within the time period from 1981 January 1 to 2020 December 31 in the regional earthquake likelihood models (RELM) polygon. (b) Frequency–magnitude distribution of earthquakes. The vertical solid line shows the overall magnitude of completeness ($M_{\rm c} = 2.1$) estimated using the method by Clauset *et al.* (2009). (c) Three pairs of lines are the logarithms of the cumulative numbers for events with $M \geq 3.0,\ 4.0,\ 5.0$, respectively. Solid lines are the numbers of the full sequence, while the dotted lines are the numbers of independent events.

3.1 Input data and label for neural network

In this work, we decompose the seismicity to construct the input of the neural network. We aim to test whether AI models can autonomously learn the true time–space distribution of earthquakes. Therefore, we do not provide the model with well-developed and powerful statistical seismology knowledge. We introduce a simple feature engineering construction to transform the earthquake catalogue into a physically based log-energy input to the FCN.

The input to the FCN is a log-energy measure having values on the space-time discretized with a spatio-temporal mesh (0.1° × 0.1° × 10 d). In other words, we replace the spatio-temporal sequence of earthquakes by the value of a log-energy measure in discrete time with time step of 10 d and discrete space paved by squares of size 0.1° × 0.1°. The log-energy measure at a given space–time grid point $(x,\ y,\ t)$, where x and y are in units of 0.1° and t is in unit of 10 d, is obtained as the sum of energy contributions provided by a certain number of earthquakes in the neighbourhood of the grid point $(x,\ y,\ t)$.

Let $EQ_i(\overrightarrow{r_i}, t_i, m_i)$ be an earthquake that occurred at location $\overrightarrow{r_i}$, on day t_i and with magnitude m_i. The empirical relation

$$E_i = 10^{1.96m_i + 9.05} \tag{1}$$

is used to estimate the total released energy E_i of the event EQ_i (Kanamori *et al.* 1993). The rupture length L_i of EQ_i is estimated with the empirical relation using the logarithm function in base 10 (Kasahara 1981),

$$\begin{cases} \log(L_i) = 0.5m_i - 1.8, \quad L_i\ in\ km \\ L_i\ \ (in\ km)\ is\ transformed\ into\ \ l_i\ \ (in\ units\ of\ 0.1^\circ) \end{cases} \tag{2}$$

Thus, l_i is the rupture length of earthquake EQ_i in the units of the discrete space lattice with mesh of 0.1°.

Following the idea of Region–Time–Length algorithm (Huang & Nagao 2002), we assume that an earthquake $\mathrm{EQ_i}$ provides a log-energy contribution only to spatio-temporal grid cells that are placed at a Chebyshev distance no larger than $2l_i$ from the earthquake epicentre. The Chebyshev distance between the two grid cells of coordinate (x_1, y_1) and (x_2, y_2) is by definition the maximum of $|x_1 - x_2|$ and $|y_1 - y_2|$ and is chosen here to facilitate the calculation of the distance between two grid cells. Thus, the set of grid cells at a given Chebyshev distance $r \geq 1$ from the earthquake EQ_i epicentre is on the perimeter of a square centred on the epicentre and of side length $2r$. The corresponding number N_r of grid cell is thus $N_r = 8r$, as a square has four sides.

Similarly to an elastic energy Green function, we postulate that the earthquake energy E_i of earthquake EQ_i can be decomposed into $2l_i + 1$ energy shells $e_i(r),\ \ i = 0, \ldots, 2l_i$ such that

$$E_i = \sum_{r=0}^{2l_i} e_i(r) \tag{3}$$

with

$$e_i(r) = \frac{E_i k_i}{r^2 + 1} \tag{4}$$

where $k_{\rm i}$ is a normalizing coefficient ensuring that the identity (3) is verified. This implies $1/k_i = \sum_{r=0}^{2l_i} \frac{1}{r^2+1}$. The energy $e_i(r)$ given by expression (4) in the square energy shell at Chebyshev distance r to the epicentre of earthquake EQ_i is further equally divided between the N_r grid cells along the perimeter of this square energy shell, so that each such grid cell receives the energy contribution $e_i(x, y)$ from earthquake EQ_i:

$$\begin{cases} e_i(x, y) = \dfrac{e_i(r)}{N_r} \\ N_r = \begin{cases} 1, & r = 0 \\ 8r, & r \geq 1 \end{cases} \end{cases} \tag{5}$$

Table 1. The number of positive (P) and negative (N) samples in the training, validation, testing data sets for 12 different time–magnitude windows.

Time–magnitude window		Training data set		Validation data set		Testing data set	
M	T (*in* d)	P	N	P	N	P	N
≥ 3.0	15	4788	2 973 012	2624	2 280 356	2980	2 230 370
	30	4404	1 464 644	2568	1 178 626	2722	1 099 064
	60	3958	710 714	2612	622 726	2396	533 608
	90	3746	472 702	2480	414 412	2280	355 056
≥ 4.0	15	553	2 977 247	356	2 282 624	404	2 232 946
	30	516	1 468 532	354	1 180 840	381	1 101 405
	60	493	714 179	378	624 960	343	535 661
	90	482	475 966	363	416 529	336	357 000
≥ 5.0	15	78	2 977 722	35	2 282 945	42	2 233 308
	30	77	1 468 971	35	1 181 159	41	1 101 745
	60	72	714 600	37	625 301	42	535 962
	90	71	476 377	37	416 855	42	357 294

Recall that N_r is the total number of grid cells at Chebyshev distance r from the epicentre of EQ_i.

We then define $CulE(x, y, t)$ as the cumulative energy released by all N earthquakes in the time-space grid cell (x, y, t):

$$CulE(x, y, t) = \sum_{i=1}^{N} e_i(x, y)\, I(r \leq 2l_i) \times I(m_i \geq M_{\text{cutoff}})\, I(0 \leq t - t_i \leq 10\ \text{d}) \tag{6}$$

in which $I(\zeta)$ is a logical function given by

$$I(\zeta) = \begin{cases} 1, \zeta \ is\ true \\ 0,\ otherwise \end{cases} \tag{7}$$

N is the total number of earthquakes and M_{cutoff} is the magnitude cut-off. Expression (6) constructs the cumulative energy in grid cell (x, y, t) as the sum of all $e_i(x, y)$'s defined by equation (5) over all earthquakes that occurred in the last 10 d and at a Chebyshev distance smaller than or equal to twice the earthquake rupture length from the grid cell.

We take the logarithm of $CulE(x, y, t)$ as the input of FCN.

For the labels, if at least one earthquake occurs in the given time–space–magnitude window, the label given that time–space–magnitude window will be set to '1′, otherwise, it will be set to '0′. We consider the task of earthquake prediction as a binary classification, i.e. the occurrence (1)/non-occurrence (0) of earthquakes in the given time–space–magnitude volume. In this work, mean square error (MSE) will be considered as the loss function of neural networks, and the MSE is defined as

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2, \tag{8}$$

where y_i is the label for the ith time–space bin. If at least one earthquake occurs in the ith time–space bin, $y_i = 1$, indicating a positive sample; otherwise, $y_i = 0$, representing a negative sample. n is the total number of time–space bins. $\widehat{y_i} \in [0, 1]$ is the probability provided by the neural network that at least one earthquake will occur in the ith time–space bin. This definition (8) makes MSE identical to the well-known Brier score (Brier 1950) that measures the accuracy of probabilistic forecasting.

The earthquake catalogues with $M \geq 3.0$ from 1990 January 1 to 1999 December 31, from 2000 January 1 to 2009 December 31, and from 2010 January 1 to 2020 December 31 are taken as training, validation and testing data sets, respectively. The number of positive and negative samples in the training, validation and testing data sets for 12 different time–magnitude windows is presented in Table 1.

As shown in Table 1, there is a significant imbalance between positive and negative samples, with negative samples far outnumbering positive ones. If we directly use MSE as the loss function to train the network, the model would likely output zero for all samples, resulting in a clearly flawed model. This issue is common across many classification tasks in the field of artificial intelligence field as well. A common solution is to re-weight the loss for positive samples by multiplying it by a factor α, where α is defined as the inverse of the proportion of positive samples (Yang *et al.* 2021), ensuring the model gives more importance to the minority class. Therefore, the MSE is modified to be

$$\text{MSE}^* = \frac{1}{n} \sum_{i=1}^{n} w_i(y_i) * (y_i - \widehat{y_i})^2, \tag{9}$$

where the value of $w_i(y_i)$ depends on y_i. If $y_i = 1$, $w_i(y_i) = \alpha$; otherwise, $w_i(y_i) = 1$.

However, the re-weighting strategy can also have some negative effects on the model. Specifically, although it leads to a smaller error for positive samples, it results in a larger error for negative samples. We will discuss the impact of re-weighting in detail in Section 5.1 and attempt to propose a solution to mitigate this negative effect in Section 5.2.

In this work, the training and validation data sets are used to train and determine the optimal parameters of the FCN, as well as to prevent overfitting. The optimal model is the one that achieves the highest area skill score on the validation data set. The details of the area skill score will be introduced in Section 3.4. Subsequently, the performance of the model is evaluated using the testing data set.

3.2 Fully convolutional networks

Proposed by Shelhamer *et al.* (2017), FCN are types of convolutional neural networks with efficient inference and learning, which are composed solely of convolutional layers designed to process inputs of varying sizes and to produce outputs of varying sizes. Unlike the traditional convolutional neural network which only provides one prediction for the whole image, FCN is capable of providing a prediction for each pixel, which makes it a more convenient tool to forecast earthquakes in different space bins. It is why we chose FCN to construct earthquake forecasting models in this work.

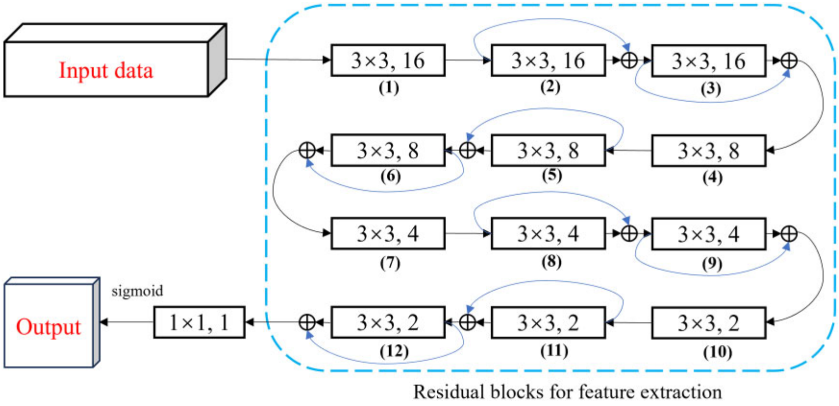


Figure 3. The structure of fully convolutional network used in this work. The blocks in the figure, as indicated by labelled blocks (1)–(12), represent convolutional layers. The "3 × 3" in the convolutional layer indicates the size of the convolutional kernel, while the numbers after the "3 × 3", such as 16, 8, 4 and 2, represent the number of kernels in the convolutional layer. The arrows and symbols '⊕' represent the summation of the input and output of the current convolutional layer, which serves as the input for the next convolutional layer. Every three convolutional layers form a residual block. For instance, the first residual block consists of convolutional layers (1), (2) and (3). In this model, there a total of four residual blocks. The second, third and fourth residual blocks, respectively, contain convolutional layers from (4) to (6), (7) to (9) and (10) to (12).

In this work, we build a fully convolutional network model from a convolutional layer with a residual structure, and the network structure is presented in Fig. 3. The FCN simply consisted of four residual blocks and one output layer, each residual block contains three convolutional layers and their number of filters are 16, 8, 4, and 2, all of which have a size of 3 × 3. The output layer includes a 1 × 1 filter. This architecture follows the principles of residual networks (ResNets), where residual blocks with skip connections are stacked to create deep networks while mitigating the vanishing gradient problem. The decreasing number of filters in each block is a common design choice to reduce the computational complexity of the network while preserving important features. The 1 × 1 filter in the output layer is typically used to aggregate information from the feature maps generated by the preceding layers and produce the final output of the network.

The spatial size of the study region is 125 × 133 (0.1° × 0.1°). To forecast future earthquakes, we will utilize 700 d (70 × 10 d) of historical seismicity data across the entire study region (125, 133). Here, '70' denotes the number of spatio-temporal meshes employed on the timeline. Consequently, the input size is (70, 125 and 133). It is worth noting that the number of meshes, set to '70' on the timeline, is constrained by the computational capabilities at our disposal. Given access to more powerful computing devices, researchers could potentially increase this value, thereby integrating additional historical earthquake data. The output of the FCN represents the probability of at least one earthquake occurring within the specified time–magnitude window for each spatial bin (0.1° × 0.1°) across the entire study region. Therefore, the output size is (125, 133). The codes of FCN are written in Python using the TensorFlow (version = 1.13.1) libraries (https://www.tensorflow.org). We employed an RTX2080Ti GPU to train, validate and test the FCN model in this experiment.

3.3 Epidemic type aftershock sequence model

The player competing with our neural network is the ETAS model augmented to account for the spatial variability of the background rates, which was developed by Nandan et al (2017, 2021). The earthquake catalogue with $M \geq 3.0$ from 1980 January 1 to 1989 December 31 is used to calibrate the initial model. Then, the parameters are updated every 15 d, which is also our minimum horizon time for predicting earthquakes. Using a total number of 60 000 simulations, we get the ETAS earthquake predictions from the years 1990 to 2020 in California, using the same time–magnitude prediction windows as for the FCN models. Details about the ETAS model used in this work can be found in references (Nandan *et al.* 2017, 2021).

3.4 Evaluation metric

We assess the power of the earthquake forecasting models using the Molchan diagram (Molchan 1990, 1991). The Molchan diagram is a popular evaluation metric in the seismological community, and it compares the rate υ of missed events or false negatives (earthquakes that occurred and were not predicted) to the expected number of target events in the alerted time–space volume under that hypothesis of pure randomness. In this work, choosing the spatial invariant Poisson model as the reference model, the x-axis of the Molchan diagram is the fraction τ of space–time occupied by alarms. In this study, τ also represents the proportion of samples classified as positive by the model relative to the total number of samples. Random predictions correspond to $\upsilon = 1 - \tau$. The goal of a good forecasting model is to obtain the smallest υ for a given τ, while scanning all τs allows one to adapt to different possible costs of alarms versus harm from missing events. By setting different thresholds for the probability provided by the earthquake forecasting model, we will also get a trajectory in the Molchan error diagram, and the area above its Molchan trajectory and below the $\upsilon = 1$ line is the area skill score (Zechar & Jordan 2008). The closer it is to 1, the better the model in forecasting. More details about the Molchan diagram can be found in the reference (Molchan & Keilis-Borok 2008; Zechar & Jordan 2008).

It should be noted that the reference model used in this study for the Molchan diagram is the spatially uniform Poisson model, which is what we referred to earlier as the 'novice boxer.' A high score given by the Molchan diagram only demonstrates that the tested model is superior to the reference model and does not necessarily prove that the model is good. More related discussions about the reference model of evaluation metrics can be found in (Zhang *et al.* 2023, 2024). Therefore, we emphasize that we use the Molchan

Table 2. The numbers shown in this table are area skill scores obtained for the ETAS and FCN models to predict full sequence of events with different time–magnitude windows. Time windows correspond to the four columns with heading 15, 30, 60 and 90 (d). Magnitude windows correspond to the three rows with headings 3, 4 and 5. Area skill scores are defined as the area between the Molchan trajectory $v(\tau)$ and the $v = 1 - \tau$ line. The scores of the FCN model that are larger than or equal to the scores of the ETAS model are marked in bold. FCN (Input $\geq$ 3.0) is the FCN model with earthquakes with $M \geq 3.0$ as input. FCN (Input $\geq$ 0) is the FCN model with earthquakes with $M \geq 0$ as input.

Magnitude\Time	Model	15	30	60	90
3	ETAS	0.891	0.890	0.891	0.882
	FCN (Input $\geq$ 0)	**0.912**	**0.899**	**0.900**	**0.894**
	FCN (Input $\geq$ 3)	0.881	0.888	0.882	**0.890**
4	ETAS	0.881	0.885	0.889	0.880
	FCN (Input $\geq$ 0)	**0.883**	0.871	0.875	0.866
	FCN (Input $\geq$ 3)	**0.884**	**0.885**	0.887	0.872
5	ETAS	0.830	0.856	0.844	0.869
	FCN (Input $\geq$ 0)	**0.858**	0.842	**0.853**	0.854
	FCN (Input $\geq$ 3)	**0.864**	**0.865**	**0.882**	**0.874**

diagram only to compare the performance of FCN with ETAS as a null hypothesis.

As the ETAS model formulates forecasts in terms of probabilities, one can turn these probabilistic forecasts into binary alarms by setting a probability cut-off. Varying the cut-off from 0 to 1 corresponds to going from $\tau = 0,\ v = 1$ to $\tau = 1,\ v = 0$ in the Molchan diagram.

4 PERFORMANCE COMPARISON BETWEEN FCN AND ETAS MODELS

Table 2 compiles the area skill scores of ETAS and FCN (Input $\geq$ 3.0) for predicting the full sequence earthquakes with 12 different time–magnitude windows in the testing dataset, where FCN (Input $\geq$ 3.0) is the FCN model with earthquakes with $M \geq 3.0$ as input. The Molchan diagrams for these results are shown in Fig. 4. For target earthquake magnitudes $M \geq 3.0$ and 4.0, the area skill scores of FCN (Input $\geq$ 3.0) and ETAS for forecasting the full sequence are roughly the same, and the difference between the scores is smaller than 0.01 in most time-magnitude windows. For $M = 5.0$, in terms of area skill score, FCN (Input $\geq$ 3.0) outperforms ETAS for forecasting the full sequence events.

5 CONVERTING TO ALERT LEVEL MAP

Data imbalance is ubiquitous and inherent in the real world. The power of current machine learning or deep learning techniques is still limited in the presence of imbalanced data sets. In this work, we adopt a standard strategy to balance the influence of positive and negative samples on the gradient of the network by re-weighting the loss function. However, the re-weighting method involves a tradeoff between the accuracy obtained for negative and positive samples. That is, increasing the weights of positive samples leads to a smaller error of the model for positive samples and a larger error for negative samples (Yang *et al.* 2021).

5.1 The influence of the re-weighting strategy on the output of FCN

The dependence of the area skill score, mean squared error for negative and positive samples, and distribution of probabilities $\Pr(\vec{r}, t, M, T)$ that at least one earthquake with magnitude $\geq$ M will occur in the grid cell of $\vec{r}$ in the time window $[t + 1,\ t + T]$ in the testing data set are shown in Fig. 5 as a function of the weight applied to positive samples (measured in terms of multiples of α, itself defined as the inverse of the proportion of positive samples), for the FCN (Input $\geq$ 3.0) model with $M = \{3.0,\ 4.0,\ 5.0\}$ and $T = 30$ d. The area skill score and mean squared error for ETAS model are also presented in the last column of each subplot in Fig. 5. As the weight applied to positive samples is increased, the influence of positive samples on the gradient of FCN also increases, leading to an increase in the $\Pr(\vec{r}, t, M, T)$, an increase in MSE for negative samples, and a decrease in MSE for positive samples. Since the whole sample is dominated by negative samples, the MSE for the whole sample also increases with the weight applied to positive samples. Such a result is common in binary classification for machine learning. However, as shown in Fig. 5, when the weight applied to the positive sample is smaller than 0.15 α, an increase in the weight applied to positive samples significantly improves the area skill score of FCN; when the weight applied to positive samples is larger than 0.15 α, a change in the weight applied to positive samples does not change much the area skill score. The area skill score and Molchan diagram are typical rank metrics that measure how well the ordering ranks positive examples above negative samples. This provides a summary of the performance of a model across all possible thresholds.

Unsurprisingly, the re-weighting strategy that amplifies the role of positive samples (i.e. occurrences of earthquakes) in the training of the FCN model has the consequence of forecasting too many earthquakes. In other words, the FCN model is overpessimistic by proposing probabilities for earthquakes to occur that are larger than their empirical frequencies. As the re-weighting is a meta-parameter, the FCN does not learn the real imbalanced distribution of occurrence and non-occurrence of earthquakes.

Currently, we are in a dilemma. Without re-weighting, the model will classify all samples as negative, resulting in a useless model that assumes no earthquakes will occur in the future. On the other hand, with re-weighting, the FCN fails to learn the true distribution of positive and negative earthquake samples. This is a common issue encountered by other AI models that simplify earthquake prediction into a binary classification problem. We need a practical and effective solution to address this issue.

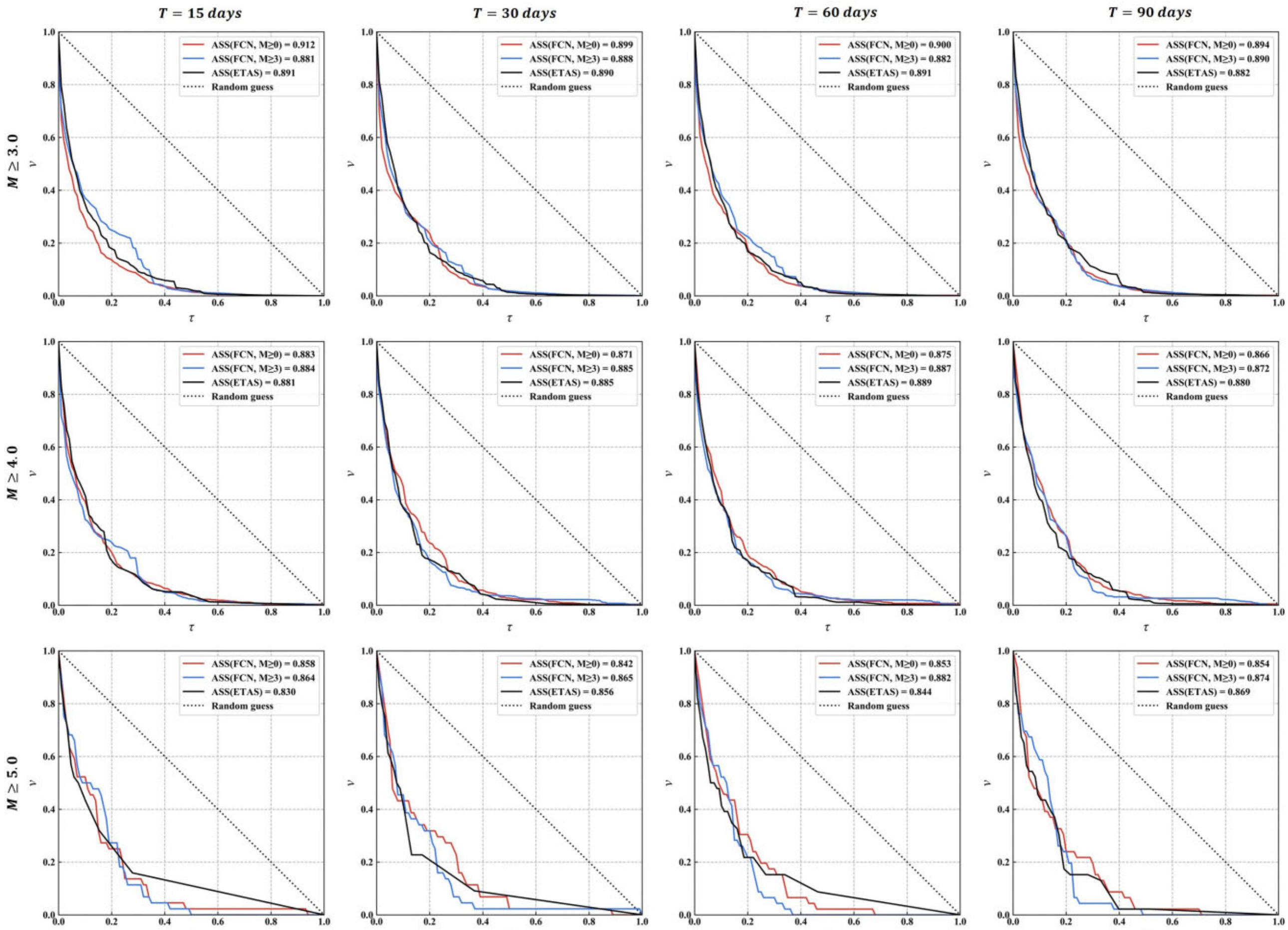


Figure 4. Molchan diagram and area skill score for the ETAS model, FCN (Input ≥ 0), and FCN (Input ≥ 3.0) with different time–magnitude earthquake prediction windows.

5.2 Performing coarse-graining operations to convert the output into alert level

To tackle this issue, we suggest interpreting the probability $\Pr(\vec{r}, t, M, T)$ of the FCN that at least one earthquake with magnitude $\geq M$ will occur in the grid cell of $\vec{r}$ in the time window $[t+1,\ t+T]$, not as a direct probability of an earthquake occurrence, but rather as an indicator signifying the high- or low-alert level of earthquakes in various locations and times. To make this information readily available to the public, we need to refine the network's output, minimizing the adverse impacts of the re-weighting method. As such, we propose to present the model's $\Pr(\vec{r}, t, M, T)$ as an alert level instead of a mere probability.

We categorize alert levels into nine tiers, ranging from Level I to Level IX, with the alert level of earthquake occurrence increasing with each subsequent level. The alert level assessment for the testing data set is derived from the output probabilities of both training and validation data sets. These probabilities are arranged in ascending order. We designate nine distinct probability ranges, each corresponding to a specific earthquake alert level. The assignment of alert levels based on these probability ranges follows the criteria laid out in Table 3. This corresponds to a nonlinear conversion of probabilities into alert levels, aiming to mitigate the FCN model's tendency to overpredict earthquakes.

For the ETAS model, the absolute probability of significant earthquake occurrence remains low (typically less than 1 per cent per day). Converting these low-probability forecasts into effective decision-making is a complex and difficult task (Jordan *et al.* 2011). Our strategy of converting probabilities into alert levels enhances the practicality of using short-term large earthquake forecasts from the ETAS model. In this work, we get the alert levels for the ETAS model of the testing data set with the same method. In the ETAS model, the probability intervals are determined by the probabilities provided by ETAS for the training and validation data sets.

It is important to note that the determination of alert level is subjective. Earthquake forecasters can adjust the rules for assigning alert levels based on their own earthquake prediction strategies. For instance, if we require a pessimistic model with fewer missed events, we can accordingly widen the output probability range for high-alert levels. Conversely, if we aim for a model with fewer false alarms, we can provide a narrower output probability range for high-alert levels.

Converting the FCN's $\Pr(\vec{r}, t, M, T)$ and the probabilities generated by the ETAS model into nine alert levels serves two purposes: it compensates for overprediction by adjusting the categorical

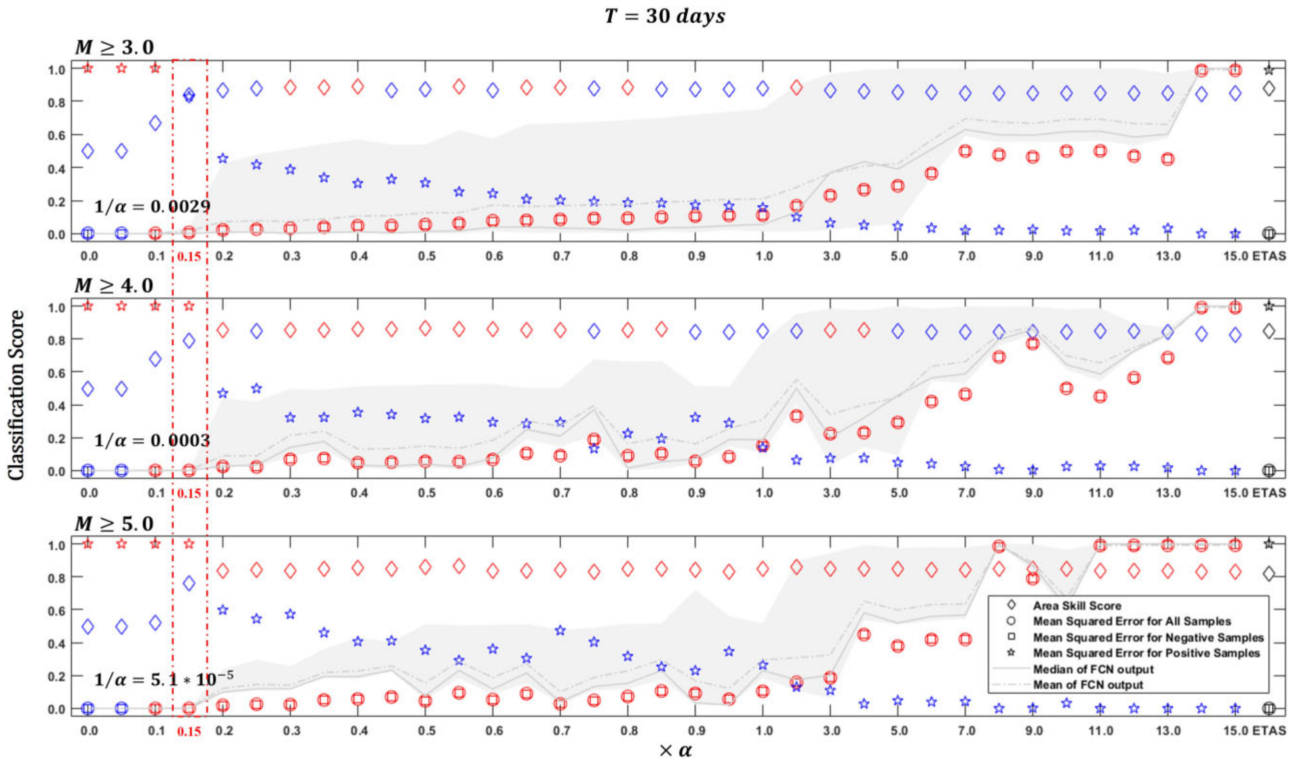
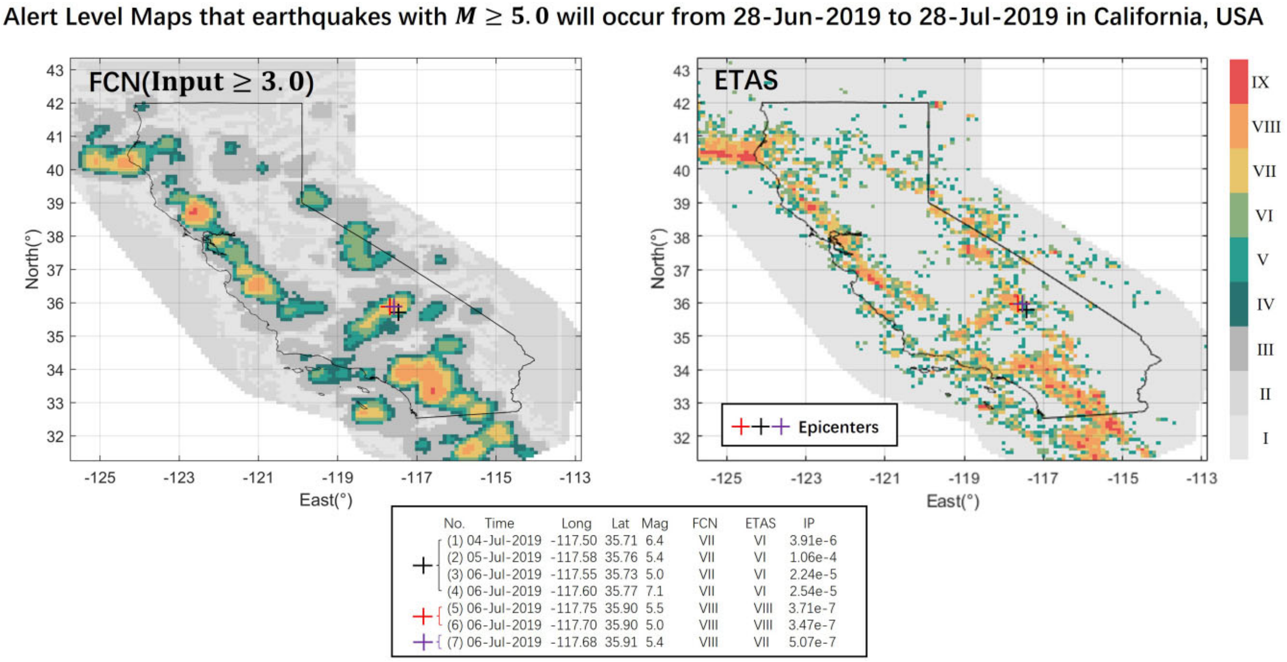


Figure 5. Dependence of the classification scores as a function of the weight applied to positive samples (measured in terms of multiples of α shown along the x-axis, itself defined as the inverse of the proportion of positive samples). Different symbols represent the average and median of the probability $\Pr(\vec{r}, t, M, T)$ determined by the FCN that at least one earthquake with magnitude $\geq$ M will occur in the grid cell of $\vec{r}$ in the time window $[t+1,\ t+T]$, mean squared error for positive, negative, and all samples and area skill score. We also present results of ETAS models at the end of the x-axis of each subplot. In each subplot, the ETAS results are marked in black and the FCN results are marked in blue or red. If the value of FCN is larger than or equal to the corresponding value of the ETAS model, the symbol is marked in red, otherwise, it is marked in blue. For example, the area skill score of the ETAS model is marked by a black rhombus; and if the area skill score of the FCN is higher than or equal to the ETAS model, it will be marked by a red rhombus, otherwise, it will be marked by a blue rhombus. The perfect model should be with an area skill score equaling 1, and the mean squared error for positive, negative as well as all samples equaling 0. The three panels correspond respectively to target magnitudes $M \geq 3,\ \ M \geq 4$ and $M \geq 5$, as indicated along the left y-axis, with the corresponding proportions of positive samples indicated as $1/\alpha = 0.0029,\ 0.0003$ and 0.000051. The grey area is the range between the 5th and 95th percentiles of the $\Pr(\vec{r}, t, M, T)$ of FCN as a function of the weight applied to positive samples for the three different magnitude targets. The red rectangle shows the results for $0.15\ \alpha$.

Table 3. Proposed rules for constructing alert levels for an earthquake forecasting model. p_x is defined as the xth percentile of the probability $\Pr(\vec{r}, t, M, T)$ calculated by the FCN that at least one earthquake with magnitude $\geq M$ will occur in the grid cell of $\vec{r}$ in the time window $[t+1,\ t+T]$, obtained from training and validation data sets. Specifically, from the distribution $F(p)$ of earthquake probability outputs of the FCN model and given a probability level x, p_x is solution of $F(p) = x$, i.e. $p_x = \mathrm{F}^{-1}(x)$, where $\mathrm{F}^{-1}(.)$ is the inverse function of $F(p)$.

Probability	Alert level	Alarm type
$(p_{99}, p_{100}]$	IX	
$(p_{95}, p_{99}]$	VIII	High alert
$(p_{90}, p_{95}]$	VII	
$(p_{85}, p_{90}]$	VI	
$(p_{80}, p_{85}]$	V	Medium alert
$(p_{75}, p_{80}]$	IV	
$(p_{50}, p_{75}]$	III	
$(p_{25}, p_{50}]$	II	Low alert
$[p_0, p_{25}]$ Or $\Pr(\vec{r}, t, M, T) = 0$	I	

boundaries, and it functions as a coarse-graining procedure simplifying the forecast output for practical use. Coarse-graining is a fundamental concept across multiple scientific disciplines. At its core, it refers to the process of simplifying a detailed, often microscopic, description of a system to gain insight into its macroscopic behaviour. By doing so, one effectively blurs or average out the finest details to focus on the larger, more collective, patterns or behaviours. For our applications, we abandon the notion of precise probabilities that give a false sense of precision, given all the modelling uncertainties and data limitations. Coarse-graining into nine levels provides a more realistic and useful assessment of the alert magnitude in each time–magnitude–space unit. Our motivation echoes the development of the 12 levels of the Modified Mercalli intensity scale developed to measure the intensity of shaking produced by an earthquake. It is also reminiscent of the seven levels of the International Nuclear Event Scale (INES) to enable prompt communication of safety significant information in case of nuclear accidents. Each level assesses the impact on people, the environment and facility safety by examining factors like radiation release, exposure and the effectiveness of safety protocols, suggesting that the INES may be overly qualitative and poorly related



	No.	Time	Long	Lat	Mag	FCN	ETAS	IP
+	(1)	04-Jul-2019	-117.50	35.71	6.4	VII	VI	3.91e-6
	(2)	05-Jul-2019	-117.58	35.76	5.4	VII	VI	1.06e-4
	(3)	06-Jul-2019	-117.55	35.73	5.0	VII	VI	2.24e-5
	(4)	06-Jul-2019	-117.60	35.77	7.1	VII	VI	2.54e-5
+	(5)	06-Jul-2019	-117.75	35.90	5.5	VIII	VIII	3.71e-7
	(6)	06-Jul-2019	-117.70	35.90	5.0	VIII	VIII	3.47e-7
+	(7)	06-Jul-2019	-117.68	35.91	5.4	VIII	VII	5.07e-7

Figure 6. Alert level maps constructed by the FCN (Input ≥ 3.0) model and ETAS model that target earthquakes with $M \geq 5.0$ will occur from 2019 June 28 to 2019 July 28. Symbols '+' indicate the epicentres of the earthquakes that occurred in this period. The earthquakes occurring in the same time–space unit are presented in one '+' symbol. The table at the bottom presents the information about these earthquakes, as well as the corresponding alert level provided by the FCN (Input ≥ 3.0) model and the ETAS model for the earthquakes to occur in a grid cell of size 0.1° × 0.1° centred on the realized epicentre.

to physical measures of severity. However, Wheatley *et al.* (2017) quantified the existence of an approximate proportionality between the INES score and the nuclear accident magnitude scale defined as NAMS $= log_{10}(20 \times R)$ with R being the radioactivity being released in terabecquerels, calculated as the equivalent dose of iodine-131. It is in the sense that our nine level scale is similar to that INES system.

5.3 The performance of alert level map provided by the FCN and the ETAS model

The earthquake with the maximum magnitude in the testing data set was the 2019 July 6th, Mw 7.1 Ridgecrest earthquake in eastern California. In this study, we use this event as a case study to demonstrate the alert level prediction capabilities of our model and the ETAS model. Fig. 6 presents the alert level maps constructed by the FCN (Input ≥ 3.0) model and the ETAS model for earthquakes with $M \geq 5.0$ in the time window from 2019 June 28 to 2019 July 28, and the time–magnitude prediction window of the given FCN model and ETAS model is $\{M \geq 5.0,\ T = 90\text{ d}\}$. The low-independent probabilities (IP) of these seven events suggest that all of these are triggered events. The IP is provided by the ETAS model (Nandan *et al.* 2021), and it is the probability that the event is an independent (also called background event), while 1-IP is the probability that the event is a triggered event [often called an aftershock if the triggered event has a smaller magnitude than that of the triggering event and a foreshock otherwise (Helmstetter & Sornette 2003; Helmstetter *et al.* 2003)]. The ETAS model also provides the probability that event A is an offspring of event B [the method to get this probability can be found in references (Nandan *et al.* 2017, 2021)]. The ETAS model shows that event (1) in Fig. 6 was triggered by an independent event with IP = 93.53 per cent and $M = 4.0$, occurring at (35.71°N, 117.50°W) on 2019 June 4, followed by a series of aftershocks including the events (2)–(7) shown in Fig. 6, and the probability that event (1) is the offspring of the 2019 June 4 $M = 4.0$ event is 99.999 92 per cent. FCN grades these events at VII or VIII alert levels, while ETAS grades these events at VI, VII or VIII alert levels.

Moreover, we present the normalized frequency of alert levels provided by FCN and ETAS model for all positive and negative time–space bins of the testing data sets, and the results are shown in Table 4, where the positive time–space bin refers to at least one earthquake occurring in it and the negative time–space bin refers to no earthquakes occurring in it. In Table 4, the random model randomly assigns the alert level for each time–space bin with the prior probabilities for assigning alert levels being the same as the normalized distributions of level I to level IX presented in Table 3. The expected alert levels provided by the non-skillful random model are 3.01 both for positive and negative samples. As shown in Table 4, in all time–magnitude windows, the ETAS model and our FCNs offer much higher average alert levels for positive samples. Moreover, the average alert levels provided by the ETAS model and FCNs for negative samples range from 2.43 to 2.92, which are lower than the 3.01 given by the random model. Table 4 shows the potential of our FCN and the ETAS model for forecasting earthquakes in California.

We turn these alert level maps into binary alarms with different cut-offs for alert levels and then present the performance of the FCN (Input ≥ 3.0) and the ETAS model in terms of the positive predictive value (PPV), true positive rate (TPR) of events, and space–time correlation window (STCW), and the results are shown in Fig. 7. In Fig. 7, the x-axis is the cut-off R_x of alert levels. If the alert level

Table 4. Normalized frequency of alert levels provided by the FCN and ETAS models for the positive and negative samples in the testing dataset. *T–M* window refers to the time-–magnitude earthquake prediction window. P/N refers to the positive/negative samples. Avg. is the expected alert level or the average alert level for positive or negative samples provided by the different models. α^{-1} is the proportion of positive samples. In the same time–magnitude window, for the two average alert levels for positive or negative samples provided by FCNs and the ETAS model, the highest average alert level for positive samples is marked in bold, while the lowest average alert level for negative samples is marked in italic.

*T–M*window	α^{-1}(per cent)	Model	P/N	Alert levels (Normalized frequency/%)									
				I	II	III	IV	V	VI	VII	VIII	IX	Avg.
/	/	Random	P	25	25	25	5	5	5	5	4	1	3.01
		Model	N	25	25	25	5	5	5	5	4	1	3.01
$M \geq 3$	0.115	FCN	P	0.72	2.97	9.73	4.45	5.63	8.29	15.7	30.3	22.2	**6.84**
$T = 15$			N	28.5	25.6	24.4	4.52	4.40	4.36	4.46	3.27	0.57	*2.82*
		ETAS	P	0.46	2.01	10.6	4.29	5.66	11.5	16.4	30.5	18.5	6.76
			N	23.2	28.3	26.4	4.59	4.44	4.33	4.32	3.80	0.66	2.92
$M \geq 3$	0.288	FCN	P	0.97	3.27	10.7	4.86	5.48	9.64	18.2	21.2	25.6	**6.72**
$T = 30$			N	28.4	25.5	24.4	4.53	4.51	4.46	4.45	2.62	1.06	*2.82*
		ETAS	P	0.58	2.10	11.4	4.39	6.07	11.2	17.1	30.5	16.7	6.67
			N	25.6	26.9	25.8	4.47	4.46	4.22	4.23	3.74	0.62	2.87
$M \geq 3$	0.417	FCN	P	0.84	3.60	11.9	4.49	6.37	9.13	16.4	28.0	19.3	**6.61**
$T = 60$			N	27.9	25.7	24.4	4.51	4.42	4.38	4.47	3.51	0.66	*2.84*
		ETAS	P	0.65	2.19	12.5	4.38	5.65	12.0	16.4	29.8	16.4	6.61
			N	26.6	26.0	25.9	4.44	4.40	4.19	4.15	3.69	0.60	2.85
$M \geq 3$	0.596	FCN	P	0.72	4.35	12.1	4.31	7.37	10.6	17.4	25.7	17.5	6.48
$T = 90$			N	28.0	25.9	24.1	4.79	4.71	4.54	4.40	2.82	0.79	*2.82*
		ETAS	P	0.61	2.26	13.0	4.69	5.74	12.4	16.6	29.4	15.4	**6.56**
			N	26.9	25.8	25.9	4.46	4.38	4.18	4.11	3.65	0.57	2.84
$M \geq 4$	0.016	FCN	P	1.12	3.65	13.2	3.37	5.90	9.27	18.3	17.7	27.5	6.63
$T = 15$			N	28.5	25.5	24.3	4.72	4.66	4.53	4.28	2.51	1.01	*2.81*
		ETAS	P	1.13	0	11.5	4.51	7.61	9.86	17.5	29.0	18.9	**6.75**
			N	43.2	0.00	32.5	6.07	4.45	4.26	4.75	3.98	0.72	2.84
$M \geq 4$	0.030	FCN	P	3.11	2.54	8.47	6.78	8.76	11.6	21.5	19.2	18.1	6.41
$T = 30$			N	25.0	28.9	24.5	4.67	4.63	4.60	4.32	2.78	0.63	2.84
		ETAS	P	0.85	1.13	12.1	3.11	8.47	9.89	20.3	28.5	15.5	**6.64**
			N	36.2	15.0	26.2	4.36	4.59	4.65	4.42	3.93	0.70	*2.82*
$M \geq 4$	0.060	FCN	P	2.12	2.12	10.8	4.76	8.99	13.8	15.3	19.6	22.5	6.52
$T = 60$			N	24.2	29.1	24.4	4.72	4.60	4.63	4.40	3.12	0.80	2.88
		ETAS	P	0.53	1.86	10.6	3.98	6.90	11.7	20.4	25.5	18.6	**6.70**
			N	29.4	22.2	26.1	4.47	4.53	4.36	4.33	3.91	0.69	*2.87*
$M \geq 4$	0.087	FCN	P	4.13	0.28	10.5	6.89	7.16	13.2	19.6	21.8	16.5	6.39
$T = 90$			N	26.0	27.7	24.2	4.65	4.76	4.76	4.44	2.82	0.70	*2.85*
		ETAS	P	0.83	1.38	11.9	3.31	7.18	11.6	20.7	25.1	18.0	**6.66**
			N	25.6	26.1	25.9	4.60	4.63	4.30	4.31	3.92	0.68	2.90
$M \geq 5$	0.002	FCN	P	0.00	2.86	20.0	8.57	11.4	5.71	20.0	20.0	11.4	5.94
$T = 15$			N	28.3	26.3	24.1	4.61	4.74	4.74	4.33	2.31	0.59	2.78
		ETAS	P	14.3	0.00	0.00	0.00	8.57	8.57	11.4	42.9	14.3	**6.60**
			N	72.2	0.00	0.00	0.00	12.4	5.25	4.35	4.95	0.84	*2.43*
$M \geq 5$	0.005	FCN	P	2.86	0.00	25.7	8.57	5.71	5.71	17.1	22.9	11.4	5.83
$T = 30$			N	26.3	28.3	24.1	4.62	4.59	4.60	4.24	2.57	0.64	2.80
		ETAS	P	8.57	0.00	8.57	0.00	5.71	8.57	20.0	31.4	17.1	**6.60**
			N	63.2	0.00	13.1	0.00	6.58	6.52	5.39	4.44	0.78	*2.55*
$M \geq 5$	0.007	FCN	P	0.00	0.00	21.6	8.11	10.8	10.8	18.9	16.2	13.5	6.20
$T = 60$			N	26.0	28.3	24.3	4.59	4.68	4.51	4.19	2.67	0.76	2.82
		ETAS	P	8.11	0.00	5.41	8.11	8.11	5.41	16.2	32.4	16.2	**6.49**
			N	53.6	0.00	19.6	8.20	4.57	3.92	5.22	4.10	0.78	*2.68*
$M \geq 5$	0.009	FCN	P	0.00	2.70	10.8	5.41	27.0	13.5	8.11	18.9	13.5	6.05
$T = 90$			N	26.0	28.3	24.1	4.55	4.64	4.66	4.40	2.68	0.70	2.83
		ETAS	P	2.70	0.00	10.8	2.70	16.2	5.41	16.2	29.7	16.2	**6.57**
			N	48.5	0.00	27.0	5.29	4.85	4.41	4.99	4.14	0.77	*2.76*

for a specific time–space unit meets or exceeds R_x, an alert will be issued; otherwise, no alert will be generated. The PPV is the ratio of alerted time–space units that successfully hit events to the total number of alerted time–space units. The PPV thus represents the spatial-temporal accuracy of the model. The TPR is the ratio of successfully predicted earthquakes to the total number of earthquakes. The STCW is the fraction of the union of all alerted time–space units divided by the whole time–space domain. In earthquake prediction, the primary concern is the high-alert levels. Similar to the results displayed in Table 4, the FCN exhibits close agreement with the ETAS model in terms of STCW, TPR and PPV for the first eight time–magnitude windows with $M \geq 3.0$ and 4.0. However, for the time–magnitude windows with $M \geq 5.0$, the ETAS model surpasses FCNs in TPR while maintaining STCW levels that are either comparable or smaller than FCNs. In most instances involving high-alert cut-off levels (VII, VIII and IX), both FCN (Input ≥ 3.0) and the ETAS model have effectively predicted more than 50 per cent of earthquakes with STCW or alerted time–space volume of less than 10 per cent. Nonetheless, the PPV of the FCN and ETAS earthquake prediction model is currently low and requires enhancement in the future. The red symbols shown in Fig. 7 can serve as a reference for our earthquake prediction model in California, providing users with insights into the successful alarm rate, missing rate of events, and spatial-temporal accuracy of the model.

As shown in Table 4, in the time–magnitude windows with $M \geq 3.0$ and 4.0, the average alert levels produced by FCNs are close to those generated by the ETAS model. Among the positive samples, in six out of eight time–magnitude windows, the differences in average alert levels between FCNs and ETAS are less than 0.2. Moreover, for the negative samples, the disparities in average alert levels in seven of eight time–magnitude windows are equal to or less than 0.05. Furthermore, in the first three time–magnitude windows with $M \geq 3.0$, FCNs exhibit higher/lower average alert levels for positive/negative samples compared to the ETAS model. In contrast, in all time–magnitude windows with $M \geq 5.0$, the ETAS model produces higher/lower average alert levels for positive/negative samples than FCNs, with differences ranging from 0.29 to 0.77 for positive samples and from 0.07 to 0.35 for negative samples.

This prompts the question of why the alert level maps provided by the ETAS model surpass FCNs for both positive and negative samples in predicting earthquakes with $M \geq 5.0$. In Table 4, for time–magnitude windows with $M \geq 5.0$, the ETAS model allocates a greater proportion of Levels VII, VIII and IX to positive samples compared to FCNs. This observation is also evident in the Molchan diagram, where, theoretically, Alert Levels IX, VIII and VII correspond to intervals with τ values of $[0, 0.01)$, $[0.01, 0.05)$ and $[0.05, 0.10)$, respectively. While the probability distribution across training, validation, and testing datasets may vary, it is important to note that the boundaries of τ for high-alert levels may not be entirely precise. As depicted in Fig. 4, within the time–magnitude windows featuring $M \geq 5.0$ and their $\tau \leq 0.1$ segments, the black lines corresponding to the ETAS model consistently fall below the blue lines for FCNs. This indicates a lower rate of missed events with $M \geq 5.0$ by the ETAS model, with more positive samples assigned to high-alert levels. In the case of negative samples within time–magnitude windows with $M \geq 5.0$, the provision of zero probabilities by the ETAS model for numerous time–space units results in more than 45 per cent of time–space units being categorized as Level I, possibly contributing to the ETAS model's effectiveness in handling negative samples.

6 DISCUSSION

In the past three decades, AI technology has rapidly advanced. Data mining and AI experts have designed increasingly complex neurons and deeper network structures. These advanced artificial neural networks have brought about significant and impressive revolutionary breakthroughs in applications such as computer vision, image processing and text processing. Research in machine learning typically consists in introducing a newly developed machine learning network and comparing it with other existing state-of-the-art machine learning models. Machine learning techniques are also applied in many physics, engineering and geophysics research endeavours. But, influenced by the mainstream of the machine learning community, some experts in machine learning, particularly those with backgrounds in physics or geophysics, tend to prioritize the sophistication and performance of the networks they employ. This focus sometimes overlooks the effectiveness of well-established and powerful physical or geophysical models.

This calls for revisiting the longstanding debate concerning earthquake forecasting. We mentioned in our introduction that DeVries *et al.* (2018) developed an aftershock prediction model using a cutting-edge deep learning network, achieving an area under curve of 0.847 and claiming a significant success in predicting the spatial distribution of aftershocks. However, the performance of this sophisticated neural network has been found comparable to that of a straightforward empirical law (Mignan & Broccardo 2019). This highlights the importance of benchmarking AI models against the most advanced physics or geophysics models available. Unfortunately, many prior AI-based earthquake prediction studies have overlooked this critical comparison. Some tout 'success' based solely on the superiority of their AI models over others or simple benchmarks. However, these claims often lack depth and fail to provide genuine insights. For further insights on the necessity of comparing AI models with top-tier physical or geophysical models, we refer to Zhang *et al.* (2024).

In this context, our study gains significant relevance by demonstrating that the deep learning network we utilized, specifically the fully convolutional network, exhibits prediction accuracy comparable to that of existing advanced spatiotemporal ETAS models. This unequivocally validates the potential of AI technology in the realms of statistical seismology and earthquake prediction.

In this work, we also take the earthquakes $M \geq 0$ as the input of FCN, notwithstanding the issues of magnitude incompleteness. Table 2 compares the area skill scores obtained for the ETAS and FCN models with input magnitude (Input ≥ 3.0) to forecast full sequence events with different time–magnitude windows. These results show that the completeness of the earthquake catalogue does not weaken the power of FCN for forecasting earthquakes and taking events with $M < 3.0$ as input even improves the power of FCN in predicting earthquakes with small magnitudes. Stockman *et al.* (2023) observed a similar phenomenon in their work, and they argued that the largest gains made by their neural model are due to its ability to handle the incomplete data immediately following large earthquakes. In contrast, the completeness of the training catalogue is a fundamental requirement for the ETAS model. The reason why FCN and ETAS have different requirements for the completeness of the earthquake catalogue is that they work in different ways. The ETAS model has several assumptions for the spatial-temporal distribution of earthquakes, and the incompleteness of the earthquake catalogue will introduce a bias to ETAS parameter estimates

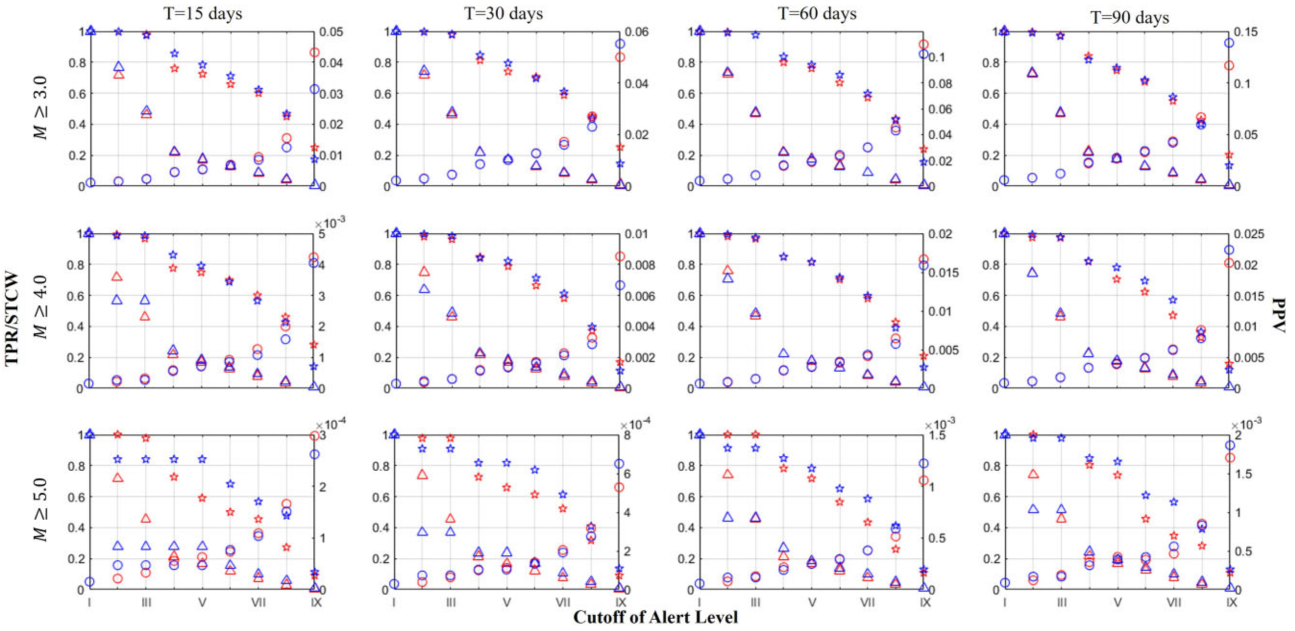


Figure 7. Positive predictive value (PPV; circles), true positive rate (TPR; stars), and space–time correlation window (STCW; triangles) of alert map provided by FCN (Input ≥ 3.0) and the ETAS model in the testing data set. Results formatted in (PPV, TPR and STCW) for different time–magnitude prediction windows and cut-off of alert level are presented in this plot. The left *y*-axis is for the values of TPR and STCW, while the right *y*-axis is for PPV. The blue symbols are for the ETAS model and the red symbols are for the FCNs.

and weaken the power of the ETAS model for predicting future earthquakes (Mizrahi *et al.* 2021). In FCN, the probability of future earthquakes is a function of the logarithm of estimated historical earthquake-released energy spread over a domain proportional to the earthquake rupture size, and the higher the input, the higher the probability that earthquakes will occur. The earthquakes with $M < 3.0$ also contain information about future earthquakes, therefore, it is unsurprised that the input including earthquakes with $M < 3.0$ will improve the power for predicting earthquakes in some time–magnitude windows.

Rather than directly taking the information on the presence or absence of earthquakes as the input of FCN, we calculated the logarithm of cumulative earthquake-released energy in each grid bin and then provided this information as input to the FCN. Such process of deciding about some transformation of the data to feed the FCN is called feature engineering in machine learning techniques. In this work, besides the logarithm of cumulative earthquake-released energy, we have also taken the cumulative earthquake-released energy, cumulative Benioff strain (Sornette & Sammis 1995; Bowman *et al.* 1998; Sammis & Sornette 2002) and its logarithmic value as the input of FCN, and we found that the logarithm of earthquake-released energy performs the best in most time–magnitude windows. More attention is needed to design better feature engineering in order to improve the performance of FCN.

With the application of machine learning to create next-generation earthquake catalogues, the volume of seismic data and the information it contains has been increasing exponentially. Mancini *et al.* (2022) has used the ETAS model together with the high-resolution seismic catalogues obtained using deep-learning for short-term earthquake forecasting, finding that the new ‘deeper’ catalogue significantly improves the forecasting performance of the ETAS model. However, as pointed out by Beroza *et al.* (2021), the significantly increased size and complexities in these next-generation earthquake catalogues pose a challenge to seismologists when it comes to exploration. Moreover, traditional approaches, it appears, are not capitalizing sufficiently on the abundance of valuable information present in these deeper next-generation catalogues (Beroza *et al.* 2021). Beroza *et al.* (2021) further suggest that machine learning technology has become the most promising approach for uncovering new relationships encoded in the seismicity.

In previous findings where the ETAS model was outperformed (Dascher-Cousineau *et al.* 2023; Stockman *et al.* 2023; Zlydenko *et al.* 2023), all these successful alternatives emulated the mathematical framework of the ETAS model. For instance, Zlydenko et al.’s model is designed with the ETAS model guiding its structure. Leveraging its own multilayer structural features and a high degree of nonlinearity, it does improve over the isotropic spatial kernel assumption of the ETAS model. As a result, its performance is close or even superior to that of its ETAS model benchmark. In contrast, our FCN is a pure deep learning network that is free from any advanced statistical seismology knowledge. This implies that our model emulates an exceptionally adept scientist who learns independently from pre-existing seismological knowledge, thereby circumventing potential biases and misconceptions entrenched in the specialized field that could hinder progress. Breaking free from the constraints of prior knowledge can sometimes facilitate the discovery of unknown knowledge, and this is precisely what future earthquake prediction models based on machine learning technology could establish further.

7 CONCLUSIONS

For earthquake probability forecasts, our deep learning model has three distinct advantages over the ETAS model. First, as shown in Table 2 and Fig. 4, the Molchan diagram indicates that the power of FCN is close to that of the ETAS model. Secondly, FCN is much faster to implement than the ETAS model. In this work, it only takes approximately 3 min to train this FCN model based on an RTX2080Ti GPU, and it only takes 4 to 8 seconds for FCN to produce a prediction. In contrast, tens of thousands of simulations for ETAS are needed to generate probabilistic forecasts, which are very time-consuming. In this work, the inversion of ETAS model parameters takes about 0.5 hours, and 60 000 simulations take about 4 hr, amounting to a total of 4.5 hr for ETAS to produce one prediction. Finally, this deep learning model is straightforward in terms of its neural network structure and feature engineering. It does not require extensive knowledge of statistical seismology, nor does it necessitate comprehensive analysis of earthquake catalog magnitudes. Using the earthquake catalogue with $M \geq 0$ as input to the network achieves the same results.

However, due to the re-weighting strategy, the overly pessimistic predictions of the FCN model (that proposes earthquake occurrence probabilities that are larger than their empirical frequencies) illustrate the existence of an inherent limitation when dealing with imbalanced classification using machine learning techniques. Converting the FCN output—$\Pr(\vec{r}, t, M, T)$ into an alert level becomes beneficial as a robustifying coarse-graining procedure. As machine learning advances in handling imbalanced data, we hope in the future to design successful schemes to convert FCN's predictions into genuine probabilities of earthquake occurrences.

Finally, we would like to make a call to both the AI community and the physics or geophysics community. We should not assess these works solely from the perspective of AI research, merely considering on whether they utilize state-of-the-art AI technology or propose better artificial neural networks. When AI technology is applied to physics or geophysics scenarios, what deserves more attention is whether it effectively addresses these physics or geophysics issues. In other words, our focus should be on comparing AI models with existing excellent physical or geophysical models. If less advanced networks manage to surpass existing excellent models or even unveil previously undiscovered patterns, it undoubtedly represents a progress for physics or geophysics. Therefore, let's refrain from overly prioritizing the design of deeper and more complex artificial neural networks, and let's avoid mechanically transplanting the latest technologies from the AI community into physics and geophysics problems in a fast-food style. Instead, let's devote more attention to physics or geophysics problems!

ACKNOWLEDGMENTS

Many thanks to Prof. Ian Main from the University of Edinburgh for his valuable suggestions on our paper, which helped us improve its quality.

This work was supported by the National Natural Science Foundation of China (Grant no. U2239201). The work of YZ was partially supported by the National Natural Science Foundation of China (Grant no. 42201363). The work of DS was partially supported by the National Natural Science Foundation of China (Grant no. U2039202), Shenzhen Science and Technology Innovation Commission (Grant no. GJHZ20210705141805017), and the Center for Computational Science and Engineering at Southern University of Science and Technology.

DATA AVAILABILITY

The earthquake catalogue within the RELM polygon is provided by Nandan *et al.* (2017, 2021). Readers can also obtain the earthquake catalogue used in this study by setting filtering criteria through the 'Search Earthquake Catalog' tool provided by U.S. Geological Survey. The link to this tool is: https://earthquake.usgs.gov/earthquakes/search/.

REFERENCES

Asencio–Cortés, G., Morales–Esteban, A., Shang, X. & Martínez–Álvarez, F., 2018. Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure, *Comput. Geosci.,* **115,** 198–210.

Banna, M.H.A., Taher, K.A., Kaiser, M.S., Mahmud, M., Rahman, M.S., Hosen, A.S.M.S. & Cho, G.H., 2020. Application of artificial intelligence in predicting earthquakes: state-of-the-art and future challenges, *IEEE Access,* **8,** 192 880–192 923.

Beroza, G.C., Segou, M. & Mostafa Mousavi, S., 2021. Machine learning and earthquake forecasting-next steps, *Nat. Commun.,* **12**(1), 4761, doi:10.1038/s41467-021-24952-6.

Bowman, D.D., Ouillon, G., Sammis, C.G., Sornette, A. & Sornette, D., 1998. An observational test of the critical earthquake concept, *J. Geophys. Res.: Solid Earth,* **103**(B10), 24 359–24 372.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability, *Mon. Weather Rev.,* **78**(1), 1–3.

Chen, J., Tang, H. & Chen, W., 2020. Deep learning of the aftershock hysteresis effect based on the elastic dislocation theory, *Nat. Hazards Earth Syst. Sci.,* **20**(11), 3117–3134.

Clauset, A., Shalizi, C.R. & Newman, M.E.J., 2009. Power-law distributions in empirical data, *SIAM Rev.,* **51**(4), 661–703.

Dascher-Cousineau, K., Shchur, O., Brodsky, E.E. & Günnemann, S., 2023. Using deep learning for flexible and scalable earthquake forecasting, *Geophys. Res. Lett.,* **50**(17), e2023GL103909, doi:10.1029/2023GL103909.

DeVries, P.M.R., Viégas, F., Wattenberg, M. & Meade, B.J., 2018. Deep learning of aftershock patterns following large earthquakes, *Nature,* **560**(7720), 632–634.

Gerstenberger, M.C., Wiemer, S., Jones, L.M. & Reasenberg, P.A., 2005. Real-time forecasts of tomorrow's earthquakes in California, *Nature,* **435**(7040), 328–331.

Gutenberg, B. & Richter, C.F., 1944. Frequency of earthquakes in California, *Bull. seism. Soc. Am.,* **34**(4), 185–188.

Hawkes, A.G., 1971. Point spectra of some mutually exciting point processes, *J. R. Stat. Soc. Ser. B: Methodol.,* **33**(3), 438–443.

Hawkes, A.G. & Oakes, D., 1974. A cluster process representation of a self-exciting process, *J. Appl. Probab.,* **11**(03), 493–503.

Helmstetter, A., Kagan, Y. & Jackson, D., 2006. Comparison of short-term and time-independent earthquake forecast models for southern California, *Bull. seism. Soc. Am.,* **96,** 90–106.

Helmstetter, A. & Sornette, D., 2003. Foreshocks explained by cascades of triggered seismicity, *J. Geophys. Res.: Solid Earth,* **108**(B10), 2457, doi:10.1029/2003JB002409.

Helmstetter, A., Sornette, D. & Grasso, J.-R., 2003. Mainshocks are aftershocks of conditional foreshocks: how do foreshock statistical properties emerge from aftershock laws, *J. Geophys. Res.: Solid Earth,* **108**(B1), 2046, doi:10.1029/2002JB001991.

Huang, Q., 2006. Search for reliable precursors: a case study of the seismic quiescence of the 2000 western Tottori prefecture earthquake, *J. Geophys. Res.: Solid Earth,* **111**(B4), B04301, doi:10.1029/2005JB003982.

Huang, Q., 2008. Seismicity changes prior to the Ms8.0 Wenchuan earthquake in Sichuan, China, *Geophys. Res. Lett.,* **35**(23), L23308, doi:10.1029/2008GL036270.

Huang, Q., 2015. Forecasting the epicenter of a future major earthquake, *Proc. Natl. Acad. Sci. USA,* **112**(4), 944–945.

Huang, Q. & Nagao, T., 2002. Seismic quiescence before the 2000 M = 7.3 Tottori earthquake, *Geophys. Res. Lett.,* **29**(12), 1578, doi:10.1029/2001GL013835.

Huang, Q., Öncel, A.O. & Sobolev, G.A., 2002. Precursory seismicity changes associated with the Mw=7.4 1999 August 17 Izmit (Turkey) earthquake, *Geophys. J. Int.,* **151**(1), 235–242.

Jordan, T.H. *et al.* 2011. Operational earthquake forecasting: state of knowledge and guidelines for utilization, *Ann. Geophys.,* **54**(4), 315–391.

Kagan, Y.Y. & Knopoff, L., 1981. Stochastic synthesis of earthquake catalogs, *J. Geophys. Res.: Solid Earth,* **86**(B4), 2853–2862.

Kagan, Y.Y. & Knopoff, L., 1987. Statistical short-term earthquake prediction, *Science,* **236**(4808), 1563–1567.

Kanamori, H., Mori, J., Hauksson, E., Heaton, T.H., Hutton, L.K. & Jones, L.M., 1993. Determination of earthquake energy release and ML using TERRA scope, *Bull. seism. Soc. Am.,* **83**(2), 330–346.

Kasahara, K., 1981. *Earthquake Mechanics,* Cambridge University Press, Cambridge

Kelleher, J. & Savino, J., 1975. Distribution of seismicity before large strike slip and thrust-type earthquakes, *J. Geophys. Res.,* **80**(2), 260–271.

Kong, Q., Trugman, D.T., Ross, Z.E., Bianco, M.J., Meade, B.J. & Gerstoft, P., 2018. Machine learning in seismology: turning data into insights, *Seismol. Res. Lett.,* **90**(1), 3–14.

Kossobokov, V.G. & Soloviev, A.A., 2021. Testing earthquake prediction algorithms, *J. Geol. Soc. India,* **97**(12), 1514–1519.

Mancini, S., Segou, M., Werner, M.J., Parsons, T., Beroza, G. & Chiaraluce, L., 2022. On the use of high-resolution and deep-learning deismic catalogs for short-term earthquake forecasts: potential benefits and current limitations, *J. Geophys. Res.: Solid Earth,* **127**(11), e2022JB025202, doi:10.1029/2022JB025202.

Mignan, A. & Broccardo, M., 2019. One neuron versus deep learning in aftershock prediction, *Nature,* **574**(7776), E1–E3.

Mignan, A. & Broccardo, M., 2020. Neural network applications in earthquake prediction (1994–2019): meta-analytic and statistical insights on their limitations, *Seismol. Res. Lett.,* **91**(4), 2330–2342.

Mizrahi, L., Nandan, S. & Wiemer, S., 2021. Embracing data incompleteness for better earthquake forecasting, *J. Geophys. Res.: Solid Earth,* **126**(12), e2021JB022379, doi:10.1029/2021JB022379.

Mogi, K., 1979. Two kinds of seismic gaps, *Pure Appl. Geophys.,* **117**(6), 1172–1186.

Molchan, G. & Keilis-Borok, V., 2008. Earthquake prediction: probabilistic aspect, *Geophys. J. Int.,* **173**(3), 1012–1017.

Molchan, G.M., 1990. Strategies in strong earthquake prediction, *Phys. Earth Planet. Inter.,* **61**(1), 84–98.

Molchan, G.M., 1991. Structure of optimal strategies in earthquake prediction, *Tectonophysics,* **193**(4), 267–276.

Nandan, S., Ouillon, G. & Sornette, D., 2019a. Magnitude of earthquakes controls the size distribution of their triggered events, *J. Geophys. Res.: Solid Earth,* **124**(3), 2762–2780.

Nandan, S., Ouillon, G., Sornette, D. & Wiemer, S., 2019b. Forecasting the rates of future aftershocks of all generations is essential to develop better earthquake forecast models, *J. Geophys. Res.: Solid Earth,* **124**(8), 8404–8425.

Nandan, S., Ouillon, G. & Sornette, D., 2022. Are large earthquakes preferentially triggered by other large events?, *J. Geophys. Res.: Solid Earth,* **127**(8), e2022JB024380, doi:10.1029/2022JB024380.

Nandan, S., Ouillon, G., Wiemer, S. & Sornette, D., 2017. Objective estimation of spatially variable parameters of epidemic type aftershock sequence model: application to California, *J. Geophys. Res.: Solid Earth,* **122**(7), 5118–5143.

Nandan, S., Ram, S.K., Ouillon, G. & Sornette, D., 2021. Is seismicity operating at a critical point?, *Phys. R0ev. Lett.,* **126**(12), 128 501.

Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes, *J. Am. Stat. Assoc.,* **83**(401), 9–27.

Ogata, Y., 1998. Space-time point-process models for earthquake occurrences, *Ann. Inst. Stat. Math.,* **50**(2), 379–402.

Ridzwan, N.S.M. & Yusoff, S.H.M., 2023. Machine learning for earthquake prediction: a review (2017–2021), *Earth Sci. Inform.,* **16**(2), 1133–1149.

Sammis, C.G. & Sornette, D., 2002. Positive feedback, memory, and the predictability of earthquakes, *Proc. Natl. Acad. Sci. USA,* **99,** 2501–2508.

Schorlemmer, D. & Gerstenberger, M.C., 2007. RELM testing center, *Seismol. Res. Lett.,* **78**(1), 30–36.

Shelhamer, E., Long, J. & Darrell, T., 2017. Fully convolutional networks for semantic segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.,* **39**(4), 640–651.

Sornette, D. & Sammis, C.G., 1995. Complex critical exponents from renormalization group theory of earthquakes: implications for earthquake predictions, *J. Phys. I,* **5,** 607–619.

Stockman, S., Lawson, D.J. & Werner, M.J., 2023. Forecasting the 2016–2017 Central Apennines earthquake sequence with a neural point process, *Earth's Future,* **11,** e2023EF003777, doi:10.1029/2023EF003777.

Tiampo, K.F., Rundle, J.B., McGinnis, S., Gross, S.J. & Klein, W., 2002. Mean-field threshold systems and phase dynamics: an application to earthquake fault systems, *Europhys. Lett.,* **60**(3), 481–488.

Utsu, T., Ogata, Y. & Matsu'Ura, R.S., 1995. The centenary of the Omori formula for a decay law of aftershock activity, *J. Phys. Earth,* **43**(1), 1–33.

Wesson, R.L. & Ellsworth, W.L., 1973. Seismicity preceding moderate earthquakes in California, *J. Geophys. Res.,* **78**(35), 8527–8546.

Wheatley, S., Sovacool, B. & Sornette, D., 2017. Of disasters and dragon kings: a statistical analysis of nuclear power incidents and accidents, *Risk Anal.,* **37**(1), 99–115.

Yang, Y., Zha, K., Chen, Y.-C., Wang, H. & Katabi, D., 2021. Delving into deep imbalanced regression, arXiv:2102.09554. Retrieved February 01, 2021, from, https://ui.adsabs.harvard.edu/abs/2021arXiv210209554Y.

Zechar, J.D. & Jordan, T.H., 2008. Testing alarm-based earthquake predictions, *Geophys. J. Int.,* **172**(2), 715–724.

Zhang, Y., Sornette, D. & Meng, Q., 2023. A new 3-D error diagram for a more balanced assessment of binary alarms for predicting earthquakes: application to TIR anomalies in Sichuan area, China, *IEEE Trans. Geosci. Remote Sens.,* **61,** 1–11.

Zhang, Y., Zhan, C., Huang, Q. & Sornette, D., 2024. Seismically informed reference models enhance AI-based earthquake prediction systems, *J. Geophys. Res.: Solid Earth,* **129**(3), e2023JB028037, doi:10.1029/2023JB028037.

Zlydenko, O. *et al.*, 2023. A neural encoder for earthquake rate forecasting, *Sci. Rep.,* **13**(1), 12350, doi:10.1038/s41598-023-38033-9.