Earth, Planets and Space

FULL PAPER Open Access

Towards a deep learning approach for short-term data-driven spatiotemporal seismicity rate forecasting

Foteini Dervisi[1,2*], Margarita Segou[1], Piero Poli[3], Brian Baptie[1], Ian Main[2] and Andrew Curtis[2]

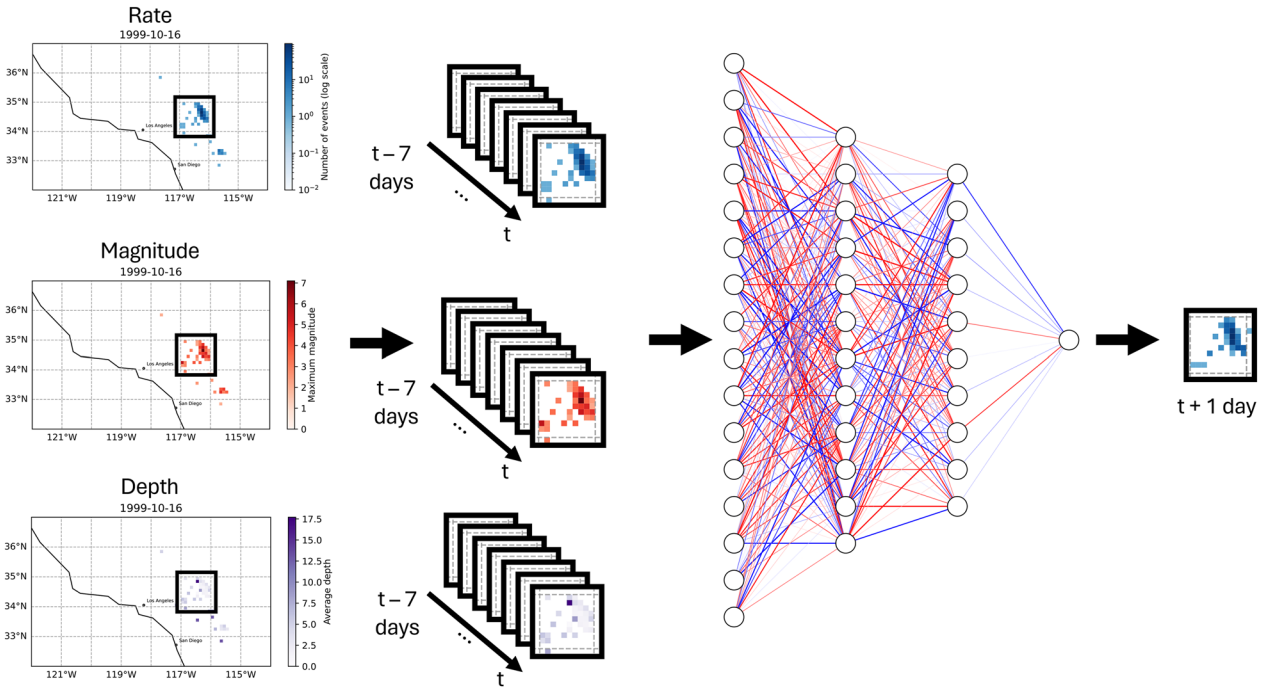
Abstract

Recent advances in earthquake monitoring have led to the development of methods for the automatic generation of high-resolution catalogues. These catalogues are created at considerably reduced processing times and contain significantly larger volumes of data concerning seismic activity compared to standard catalogues created by human analysts. Disciplinary statistics and physics-based earthquake forecasting models have shown improved performance when rich catalogues are used. The use of high-resolution catalogues paired with machine learning algorithms, which have recently evolved due to the rise in the availability of data and computational power, is therefore a promising approach to uncovering underlying patterns and hidden laws within earthquake sequences. This study focuses on the development of short-term data-driven spatiotemporal seismicity forecasting models with the help of deep learning and tests the hypothesis that deep neural networks can uncover complex patterns within earthquake catalogues. The performance of the forecasting models is assessed using metrics from the data science and earthquake forecasting communities. The results show that deep learning algorithms are a promising solution for generating short-term seismicity forecasts, provided that they are trained on a representative dataset that accurately captures the properties of earthquake sequences. Comparisons of machine learning-based forecasting models with an epidemic-type aftershock sequence benchmark show that both types of models outperform the persistence null hypothesis commonly used as a benchmark in forecasting the behaviour of other types of non-linear systems. Machine learning forecasting models achieve similar performance to that of an epidemic-type aftershock sequence benchmark on the Southern California and Italy test datasets at significantly reduced processing times - a major advantage in applications to short-term operational earthquake forecasting.

Keywords Seismicity, Earthquake catalogues, Spatiotemporal forecasting, Deep learning

*Correspondence:
Foteini Dervisi
f.dervisi@sms.ed.ac.uk; fdervisi@bgs.ac.uk
Full list of author information is available at the end of the article

Graphical Abstract



1 Introduction

The rising volumes of available data and computational power have recently led to the rapid development of the field of machine learning (ML) and the use of its power to address computational challenges in various scientific fields including seismology, where ML models are used to tackle different tasks, from Earth model inversions to seismic phase picking and event discrimination (Mousavi and Beroza 2022, 2023). The use of ML tools is also a promising development in the field of earthquake forecasting. Within this scope, Beroza et al (2021) described how multi-object deep learning catalogues will revolutionise earthquake forecasts and triggering studies. Segou (2020) posed the question of whether standard forecasts using catalogues generated with the help of ML or data-driven ML-based forecast models will present higher predictability. Mancini et al (2022) explored the predictability of physics-based and statistical models using standard and ML catalogues to find that forecasting models benefit from the use of high-resolution catalogues when advanced experimental setups, such as fine spatial grids, are adopted. In this paper, we investigate whether data-driven models using standard and high-resolution catalogues can robustly forecast short-term seismicity. The rapid evolution of artificial intelligence has revolutionized data assimilation to the point that ML-based catalogues include a factor of ten more events compared to standard catalogues produced by human analysts (e.g. Tan et al (2021)). The community is now starting to explore how data-driven ML models can contribute towards improving predictability and, perhaps, discovering currently unknown physical laws that govern earthquake occurrence (Mizrahi et al 2024).

Zlydenko et al (2023) introduced FERN, a neural encoder-decoder model for spatiotemporal earthquake rate forecasting using a point-process framework based on a multilayer perceptron, a simple neural network that consists of multiple layers of neurons, with each neuron using a non-linear activation function. FERN learns spatial and temporal embeddings that are able to capture complex correlations, thus succeeding in producing accurate spatiotemporal rate forecasts based on standard ML evaluation metrics (log-likelihood score, area under Receiver Operating Characteristic curve) as well as metrics that are specific to earthquake forecast evaluation (Average Information Gain Per Earthquake (IGPE), S-test). FERN is applied to the region of Japan, using data from the JMA catalogue (Japan Meteorological Agency 2024). Stockman et al (2023) developed a highly

flexible neural point process for short-term seismicity forecasting, which proved to be fast to train and robust to missing data. This is an important asset as earthquake catalogues are incomplete due to the fact that when large events occur, seismic stations receive waveforms that correspond to many events simultaneously. Many of these events are relatively small magnitude aftershocks whose waveforms overlap with the waveforms of larger events and are therefore not detected. This phenomenon is known as short-term aftershock incompleteness (STAI). Stockman et al (2023) tested their approach on the 2016–2017 Central Apennines high-resolution catalogue by Tan et al (2021) using the log-likelihood score and the Cumulative Information Gain (CIG) as evaluation metrics, leading to the conclusion that the model is able to make use of the wealth of information present in high-resolution catalogues due to its ability to handle incomplete data , i.e. to constrain the likelihood of future triggered events based on the information provided in incomplete catalogues. This is a major advance over physical and statistical models, which generally require complete data above a given magnitude threshold. Dascher-Cousineau et al (2023) introduced RECAST, a flexible recurrent neural network-based point process model, which was tested on Southern California earthquake catalogues using the log-likelihood score and proved to be efficient on large datasets, showing improved performance when provided with more training data. All of these approaches are based on point processes, which are the basis of statistical forecasting models. A different approach is to represent the seismicity recorded in earthquake catalogues using spatiotemporal series of seismic maps. Within this scope, Zhang and Wang (2023) used a convolutional long short-term memory (ConvLSTM) neural network to learn temporal and spatial correlations of global-scale seismicity data. They evaluated model performance using the precision, recall, accuracy, Critical Success Index (CSI), False Alarm Ratio (FAR) and R-score metrics. Their model achieves good performance at forecasting earthquakes with moment magnitude above 4, but struggles to forecast larger events (with magnitudes greater than 6) due to the fact that the amount of larger magnitude data available for training is very limited, therefore the magnitude distribution of examples is highly skewed (class imbalance).

In this study, we develop ML -based seismicity forecasting models based on architectures that have been shown to successfully address spatiotemporal time series (Yu et al 2024). We focus on building models that are able to produce spatiotemporal next-day forecasts of aftershocks following events of magnitude 4 and above. We address the challenging question of whether a larger magnitude earthquake is likely to follow a moderate-sized event, which is something that current statistics and physics-based forecasting methods struggle with. We explore whether recent advances in the fields of seismology and artificial intelligence, including high-resolution catalogues and machine learning techniques, can effectively model seismicity rates following an event of magnitude 4 and above based on recent preceding seismicity. This forms the basis of our testing strategy in this work. We focus on building ML forecasting models that are trained on a bulk seismic catalogue dataset that consists of catalogues from different tectonic regions, aiming to create models that are able to generalise well in a variety of different scenarios. We follow different paths to training those models in order to understand how existing seismic catalogues, standard and high-resolution, could play a critical role in predictability and influence the models' generalisability. Considering that earthquake forecasting research now involves both data scientists and seismologists, we evaluate the models using performance metrics from both communities. We find that ML models are a promising solution for producing short-term seismicity

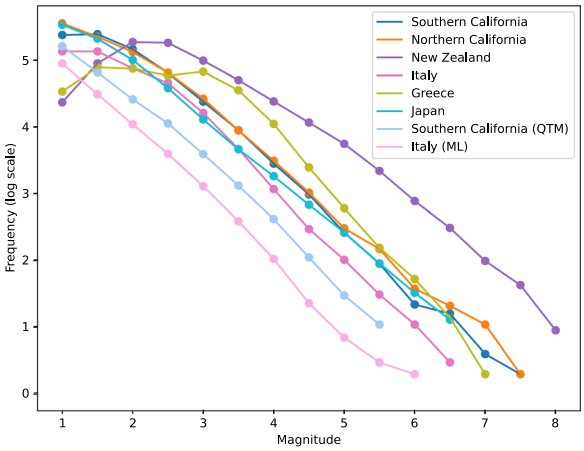


Fig. 1 Frequency-magnitude distributions of the catalogues used in this study

Table 1 Temporal extent and magnitude of completeness (M_c) of earthquake catalogues. The M_c was estimated based on the method described in Mizrahi et al (2021a)

Catalogue	Dates	# days	M_c
Southern California	01/01/1980-30/09/2023	15979	3.8
Northern California	01/01/1980-30/09/2023	15979	3.3
New Zealand	01/01/1980-30/09/2023	15979	4.0
Italy	02/01/1985-30/09/2023	14151	3.1
Greece	01/01/1980-30/09/2023	15979	4.0
Japan	01/10/2000-31/12/2012	4475	2.5
Southern California (QTM)	01/01/2008-31/12/2017	3653	2.4
Italy (ML)	15/08/2016-15/08/2017	366	2.8

Table 2 Number of events in earthquake catalogues

Catalogue	# earthquakes	# M2+ earthquakes (depth ≤ 40 km)	# M4+ earthquakes (depth ≤ 40 km)
Southern California	820787	137975	1300
Northern California	1094185	143327	2338
New Zealand	559221	272183	8892
Italy	446702	104100	866
Greece	359031	202056	5574
Japan	1091640	108030	1921
Southern California (QTM)	1811362	28187	303
Italy (ML)	900058	10128	65

forecasts given that they are exposed to a large enough high-quality dataset during training.

2 Data

We assemble a dataset containing publicly available earthquake catalogues from diverse tectonic regions: Southern California (SCEDC 2013), Northern California (NCEDC 2014), New Zealand (GNS Science 1970), Italy (ISIDe Working Group 2007), Greece (NOAIG-CATALOGUE 2024) and Japan (Yano et al 2017). The frequency-magnitude distributions of the catalogues can be seen in Fig. 1. Using the method described in Mizrahi et al (2021a) , we estimate the magnitude of completeness (M_c) of the catalogues, which can be seen in Table 1. As our target is to forecast what happens immediately following events of magnitude 4 and above, we only take into account events with a minimum magnitude (M_{min}) of two orders below that, i.e. events of magnitude 2 and above. Although this allows us to take into account all events that are felt by humans and have the potential to cause damage, it also means that our models inevitably learn the incompleteness of the catalogues. This is not a problem in our case, as ML models have been shown to perform well in incomplete data settings (Stockman et al

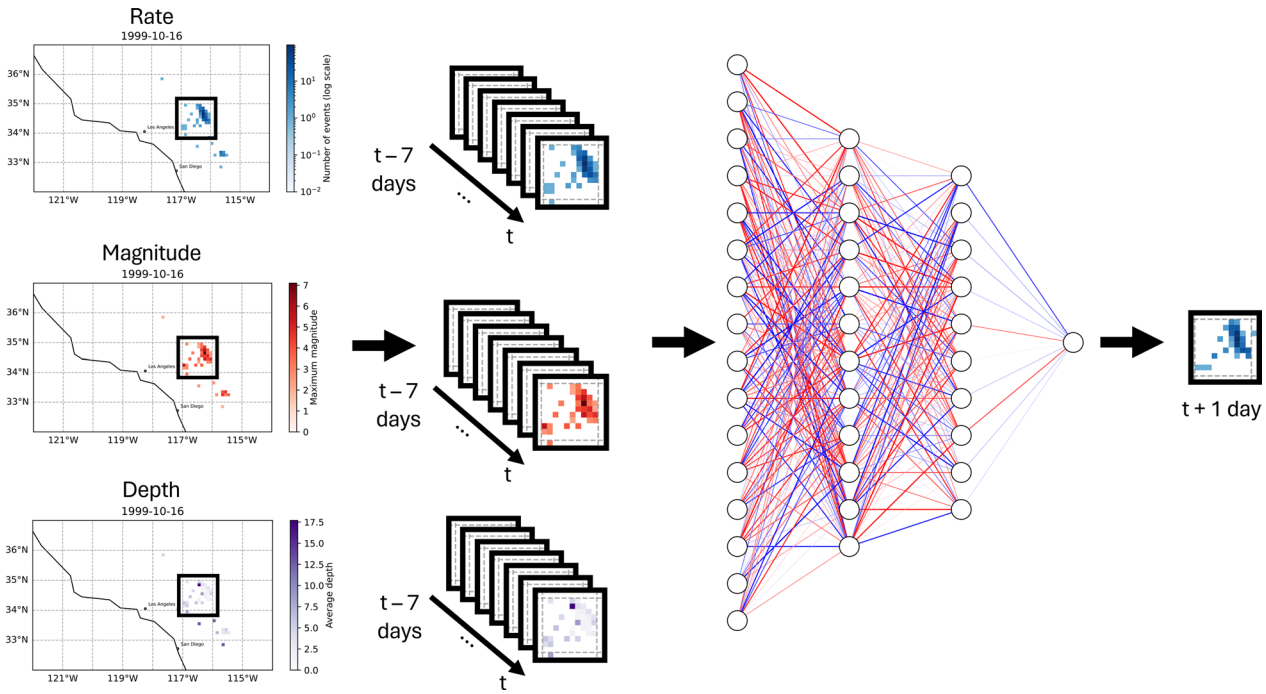


Fig. 2 Spatiotemporal rate, maximum magnitude and average depth sequences used to produce next-day rate forecasts. We identify events with magnitude 4 and above, create a square spatial grid around them and produce deep learning-based next-day rate forecasts using the rate, maximum magnitude and average depth maps of the previous 7 days as input. The neural network visualisation was created using http://alexlenail.me/NN-SVG/. BGS©UKRI 2025

2023). For example, a cluster of small earthquakes, even of magnitude below the completeness threshold, could indicate a raised probability of a larger event at that location. We consider shallow crustal events with depths up to 40km, as these tend to be the most destructive events. Further details about the catalogues can be seen in Tables 1 and 2. We use 80% of the data for training, 10% for validation and 10% for testing and follow a temporal data splitting strategy, with older data used for training and newer data used for testing. We perform an iterative training procedure in order to select an appropriate set of hyperparameters, which are various configuration variables that are manually set and used to manage ML model training. Once the hyperparameter tuning process is complete, we retrain the ML models on the training and validation data and evaluate using the independent test set, which has been kept out of the whole process up to that point.

Motivated by the fact that the training process of deep learning models requires the use of large volumes of data, we also employ high-resolution catalogues to further train the models and improve their performance. For this purpose, we use the Southern California quake template matching (QTM) catalogue (Ross et al 2019), a long-term catalogue containing events that occurred in the 10 years between 2008 and 2017, and the short-term -based catalogue introduced by Tan et al (2021), a localised high-resolution catalogue covering the year-long 2016–2017 Central Italy earthquake sequence. The Southern California QTM catalogue reduces the minimum magnitude of completeness by more than a full magnitude unit over the 10-year period covered, whereas the Italy ML catalogue also contains considerably more events compared to the standard Italy catalogue for the same year.

3 Methods

3.1 Goal

The aim of this study is to model next-day seismicity following events with a magnitude of 4 and above. The spatial area that we consider is a square centred at the M4+ event, with sides equal to 2 longitude and latitude degrees. This is essentially an aftershock forecasting scenario and is highly relevant to local communities. Such information can aid authorities in decision-making and resource allocation during the earthquake response and recovery process, contribute to infrastructure risk assessment and guide the general public towards safety measures (Hardebeck et al 2024).

3.2 Features

We create daily maps of key metrics by splitting the spatial area covered by the catalogues into bins of 0.1 degrees of longitude and latitude. Three types of two-dimensional daily maps are used, containing: (i) the number of events (rate per unit time) in each grid cell, (ii) the maximum magnitude of events in each grid cell and iii) the average depth of events in each grid cell. We identify events with magnitude ≥ 4 and aim to forecast the next day's seismicity (rate) in the spatial area where those events have occurred. For each one of the M4+ earthquakes, we use the M2+ events in the catalogues to create 7-day rate, magnitude and depth spatiotemporal training sequences. These training sequences cover the seven non-overlapping 24-hour time intervals that end with and include the M4+ event that triggered their creation, whereas the target map covers the 24-hour period that starts directly after the M4+ event. We pass these 7-day sequences through a deep learning model and produce localised next-day spatiotemporal rate forecasts with a grid resolution of 0.1 longitude and latitude degrees. The developed forecasting models are testable following the guidelines and principles of Jordan et al (2011) for earthquake forecasting research. The whole workflow is shown in Fig. 2.

3.3 Models

3.3.1 Small attention UNet (SmaAt-UNet)

The first deep learning model used is a convolutional neural network (CNN) called UNet (Ronneberger et al 2015). UNets were first designed for biomedical image segmentation but have been shown to perform well in various tasks involving two-dimensional data, which makes them an appropriate choice for our task. They consist of two parts: the encoder and the decoder. The encoder consists of a series of convolutional operations followed by rectified linear unit (ReLU) activations and maximum pooling (MaxPool) operations used for downsampling. At each downsampling step, the number of feature channels is doubled. The decoder consists of consecutive upsampling steps, each of which halves the number of feature channels, and a series of convolutional operations followed by ReLU activations. UNets also include skip connections, which are shortcuts used to connect the output of each encoder layer to the corresponding decoder layer in order to minimise the loss of spatial information due to downsampling.

The small attention UNet used in this study was first introduced by Trebing et al (2021) for the task of precipitation nowcasting, yielding promising results despite its compact size. This neural network is a UNet that includes convolutional block attention modules (CBAM) (Woo et al 2018), which are mechanisms that apply attention to the channels and spatial dimensions of two-dimensional data. It also uses depthwise-separable convolutions (DSC) (Chollet 2017) instead of regular convolutions in order to reduce the number of

parameters without compromising on the network's performance. The SmaAt-UNet is therefore a high-performing convolutional neural network that can be trained in a relatively short amount of time, which is important for operational forecasting models.

3.3.2 Earthformer

This is an example of a sequence-to-sequence model (Sutskever et al 2014), a neural network that can be used to convert an input sequence into a target sequence. Sequence-to-sequence models have been introduced to tackle problems with a sequential nature, such as machine translation or time series forecasting, and are therefore a reasonable framework choice for our problem. Like UNets, these networks consist of two parts: the encoder and the decoder. The encoder is responsible for creating a representation to encode information about the source sequence, producing a final hidden state. The decoder then receives the encoder's final hidden state as input and uses it to generate the target sequence. These models were initially implemented with the use of recurrent neural networks (RNNs), but the emergence of the mechanism of attention and the transformer neural network architecture largely changed sequence-to-sequence model design.

The attention mechanism (Bahdanau et al 2014) was introduced to address a bottleneck in sequence-to-sequence encoder-decoder models. Up until that point, the whole input sequence was represented by a single hidden state, the encoder's final hidden state. This meant that the decoder often wasn't provided with sufficient information to generate the target sequence. The introduction of attention allows the decoder to select multiple hidden states of the encoder instead of only using the final hidden state. In other words, attention enables the decoder to *pay attention* to the most important elements of the input sequence. This is achieved by calculating the attention scores, which are based on matching each encoder's hidden state to every hidden state of the decoder. Each score is proportional to the relevance of each encoder's hidden state to the decoder, with higher scores indicating higher relevance. These scores add up to one, and can therefore be used to calculate a weighted average of the encoder's hidden states, which can be used as the decoder's input.

Transformers (Vaswani et al 2017) are sequence-to-sequence models that offer an alternative to CNNs and RNNs. They use an encoder neural network that is responsible for creating intermediate representations of input sequences and a decoder neural network that predicts output sequences based on the source sequences and the intermediate steps generated by the encoder. Transformers use the mechanism of self-attention to decide which parts of the input sequence are more relevant for generating the output sequence by projecting the inputs to three weight matrices, which are initialised randomly and optimised during the training process. Multiple parallel self-attention mechanisms called heads are often used to help with capturing different aspects of the inputs. Models that are based on this idea are becoming increasingly popular for tasks in the fields of natural language processing, computer vision and time series forecasting, as they can easily be parallelised and require considerably less computational power compared to CNNs and RNNs of similar size (Kamath et al 2022). The downside is that transformer-based models often need larger datasets to be effectively trained, as they do not have an inductive bias and thus tend to overfit small datasets more easily than CNNs and RNNs (Dosovitskiy 2020).

Earthformer (Gao et al 2022), the second deep learning model used in this study, is a space-time transformer for Earth system forecasting. In this case, the input data have both a spatial and a temporal dimension and can therefore be seen as cubes. These cubes are split into non-overlapping cuboids. The cuboid attention mechanism, an extension of the attention mechanism to spatiotemporal data, is then employed. Self-attention is applied to each cuboid, a calculation that can be done in parallel to speed up the process. A set of global vectors is also used, which attends to all cuboids and can therefore transmit information about the overall state of the cube to them. The Earthformer is an excellent architecture choice in our case, as it is a space-time transformer model specifically designed to handle spatiotemporal sequences with a particular focus on applications involving Earth system data.

Table 3 Hyperparameters used for training the SmaAt-UNet and Earthformer neural networks. BGS©UKRI 2025

Hyperparameter	SmaAt-UNet	Earthformer
optimizer	Adam	Adam
Adam β_1	0.9	0.9
Adam β_2	0.99	0.99
starting learning rate	0.001	0.00001
learning rate scheduler	PyTorch StepLR	PyTorch StepLR
StepLR step size	30	30
StepLR gamma	0.1	0.1
early stopping	Yes	Yes
early stopping patience	20 epochs	20 epochs
maximum number of epochs	500	500
batch size	64	8
input map concatenation	True	False
logarithmic rate maps	True	True

3.4 Training

We use an adaptive moment estimation (Adam) optimizer (Kingma and Ba 2014) with initial decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.99$ and a step learning rate scheduler that reduces the learning rate by a factor of 10 every 30 iterations (epochs). We also use an early stopping mechanism, which monitors the validation loss, i.e. the misfit between the ground truth and the forecasts in the validation set, and stops training in case there is no improvement for a period of 20 epochs. The maximum number of epochs a model can be trained for is set to 500. In the case of the SmaAt-UNet model, we merge the weekly rate, magnitude and depth sequences into 3 single maps by calculating the pixel-wise sum of the 7 rate maps, the pixel-wise maximum of the 7 magnitude maps and the pixel-wise average of the 7 depth maps. In the case of the Earthformer model, each of the 7 rate, magnitude and depth maps in the sequence is used separately in the input data cube. In order to reduce the range of the rate data and ensure greater stability in the input time series, we use logarithmic rate maps, which leads to improved learning performance. The hyperparameters used were selected through a trial-and-error process and are summarised in Table 3. We use the mean squared error (MSE) between the ground truth, i.e. the true next-day rate maps, and the model output rate maps (or forecasts) as a loss function, which is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

where n is the number of cells in target rate maps, y_i is the ith cell of the ground truth rate map and $\hat{y}_i$ is the ith cell of the output or forecast rate map.

3.5 Evaluation

The model performance is evaluated by calculating a collection of metrics from the data science (Hewamalage et al 2023; Rainio et al 2024) and earthquake forecasting (Zechar et al 2010a; Schorlemmer et al 2018; Savran et al 2022a) communities.

3.5.1 Regression metrics

The problem that our models are addressing is a regression problem, as the output is a two-dimensional map of continuous values. We therefore calculate the mean absolute error (MAE) and the root mean squared error (RMSE), which show the difference between the forecast number of events and the observed number of events per spatial bin. The mean absolute error is given by

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{2}$$

and the root mean squared error is given by

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

with n representing the number of cells in target rate maps, y_i the ith cell of the ground truth rate map and $\hat{y}_i$ the ith cell of the output or forecast rate map.

3.5.2 Classification metrics

In addition, we calculate a collection of metrics that are used in classification tasks. To do this, we convert the forecast and ground truth maps to binary maps, with 0 representing grid cells where no events have occurred and 1 representing grid cells where at least one event has occurred. As the rate values in the forecast maps are continuous, we use a threshold of 0.5 events per day to distinguish grid cells that belong to the 0 and 1 classes. In our case, these metrics essentially show whether the models are able to forecast events in the correct spatial bins. We calculate the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) in rate maps that belong to the test set . We then calculate the accuracy, which measures how many observations were correctly classified and is given by

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{4}$$

the precision, which measures the proportion of predicted positives that were actual positives and is given by

$$Precision = \frac{TP}{TP + FP}, \tag{5}$$

the recall, which measures the proportion of actual positives that were classified correctly and is given by

$$Recall = \frac{TP}{TP + FN}, \tag{6}$$

the F1 score, which is the harmonic mean of precision and recall given by

$$F_1 = 2\frac{Precision \cdot Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}, \tag{7}$$

the critical success index (CSI), which is often used in binary forecasting and measures how well the predicted positives correspond to the actual positives (Ebert and Milne 2022), given by

$$CSI = \frac{TP}{TP + FP + FN} \tag{8}$$

and the false alarm ratio (FAR), also frequently used in binary forecasting and showing the number of false alarms per the total number of alarms, given by

$$FAR = \frac{FP}{TP + FP}. \tag{9}$$

Furthermore, we use the receiver operating characteristic (ROC) curve, which shows the trade-off between the true positive rate ($TPR = \frac{TP}{TP+FN}$) and the false positive rate ($FPR = \frac{FP}{FP+TN}$) for different cutoff thresholds. We calculate the area under the ROC curve (ROC AUC), which is a measure of the model's performance across all possible classification thresholds. We also calculate the area under the Precision-Recall curve (PRC AUC), which measures performance based on the combination of precision and recall scores for different classification thresholds.

As earthquakes show clustering in time and space, we observe a class imbalance issue (Guo et al 2008): the number of true negatives is much larger compared to the number of true positives , meaning that almost all target cells (97%) do not contain earthquakes. This indicates that the accuracy metric is not very informative on the likelihood of true positives alone. The large number of true negatives significantly influences the score, resulting in high accuracy values regardless of the model's ability to identify positive class events. For this reason, we also calculate the CSI, which gives a similar indication of the model's performance without taking into account the true negatives. The ROC curve is also greatly influenced by the large number of true negatives, therefore the Precision-Recall curve, which looks at the percentage of true positives in comparison to the percentage of predicted positives, is a more informative measure of model performance in this regard (Saito and Rehmsmeier 2015).

3.5.3 CSEP metrics

The Collaboratory for the Study of Earthquake Predictability (CSEP) is an international community aiming to support earthquake predictability research by defining the objectives of earthquake forecasting experiments and developing metrics that are suitable for their evaluation (Zechar et al 2010b). Our forecasts are grid-based, therefore the testing region $\boldsymbol{R}$ consists of the combination of magnitude bins $\boldsymbol{M}$ and spatial bins $\boldsymbol{S}$:

$$\boldsymbol{R} = \boldsymbol{M} \times \boldsymbol{S}. \tag{10}$$

In this study we don't include information about the magnitude in the model output, therefore we can assume that there is only one magnitude bin covering the whole magnitude range. The earthquake forecast can be seen as the expected number of events in each magnitude-space bin:

$$\boldsymbol{\Lambda} = \{\lambda(i,j) | i \in \boldsymbol{M}, j \in \boldsymbol{S}\} = \{\lambda(j) | j \in \boldsymbol{S}\}, \tag{11}$$

where $\lambda(i,j)$ is the forecast number of earthquakes in a specific magnitude-space bin (i,j). Similarly, the observed data can be written as

$$\boldsymbol{\Omega} = \{\omega(i,j) | i \in \boldsymbol{M}, j \in \boldsymbol{S}\} = \{\omega(j) | j \in \boldsymbol{S}\}, \tag{12}$$

where $\omega(i,j)$ is the observed number of earthquakes in a specific magnitude-space bin (i,j).

We employ the number test (N-test), which can be used to draw conclusions as to whether the number of forecast earthquakes is consistent with the number of observed earthquakes. The total number of forecast events can be written as the sum of $\lambda(i,j)$ over all spatial bins:

Table 4 Training and fine-tuning CNNs and transformers: Model parameters, training time on a single NVIDIA Quadro RTX 4000 GPU, number of epochs and learning rate at the start of the training process. In the case of SmaAt-UNet+ and Earthformer+, the times and epochs reported are the additional ones after training on the standard catalogues. BGS©UKRI 2025

Model	# parameters	Training time (min)	# epochs	Learning rate
SmaAt-UNet	4032205	27.82	28	0.001
SmaAt-UNet+	4032205	2.50	103	0.00001
Earthformer	106573193	776.86	41	0.00001
Earthformer+	106573193	38.20	171	0.00001
Persistence (day before)	0	0	0	-
Persistence (7-day avg.)	0	0	0	-
ETAS*	10	172.00	22	-

*In the case of ETAS, the time and number of epochs needed to estimate the initial global parameters are reported. However, we note that for each forecast, additional computational time is required for parameter calibration and simulations. The additional time ranges from a few minutes to several hours or even days in cases of high seismicity. On the other hand, the trained ML models and the persistence baseline are able to generate forecasts within seconds

$$N_{fore} = \sum_{(i,j)\in \boldsymbol{R}} \lambda(i,j) = \sum_{j\in \boldsymbol{S}} \lambda(j). \tag{13}$$

Similarly, the number of observed events can be written as the sum of $\omega(i,j)$ over all bins:

$$N_{obs} = \sum_{(i,j)\in \boldsymbol{R}} \omega(i,j) = \sum_{j\in \boldsymbol{S}} \omega(j). \tag{14}$$

We aim to understand if the number of observed earthquakes is consistent with the number of forecast earthquakes. In other words, assuming that the forecast rate distribution is correct, we need to determine where the number of observed events falls within this distribution. Initially, this was approached by generating a set of simulated rates and calculating the probabilities of observing at most and at least N_{obs} events (Kagan and Jackson 1995; Schorlemmer et al 2007). However, if an analytical form of forecast uncertainty is available, the corresponding cumulative distribution can be used instead. The N-test metrics, which are the probabilities of observing at least and at most N_{obs} earthquakes given that N_{fore} earthquakes are expected, can then be written as

$$\delta_1 = 1 - F((N_{obs} - 1)|N_{fore}), \tag{15}$$

and

$$\delta_2 = F(N_{obs}|N_{fore}), \tag{16}$$

where $F(x|\mu)$ is a Poisson cumulative distribution with $F(x|\mu) = 0$ for $x < 0$. This is a one-sided test, in which case the alternative hypothesis (the hypothesis we want to prove) states that the parameter value is either bigger or smaller compared to the parameter value specified in the null hypothesis (the hypothesis we want to disprove). To decide whether a forecast is considered consistent with the observation, we need to specify a significance level α. This indicates the risk of making a Type I error, i.e. the risk of rejecting the null hypothesis when it is in fact true. For an intended significance level α, i.e. for maintaining a Type I error rate of α, a forecast is considered consistent if both δ_1 and δ_2 are greater than the effective significance value $\alpha_{eff} = \frac{\alpha}{2}$. Therefore, if $\alpha = 5\%$ then $\alpha_{eff} = 0.025$, hence a forecast is consistent if $\delta_1 > 0.025$ and $\delta_2 > 0.025$. Too small δ_1 values signify underprediction, which means that the forecast rate is too low to be consistent with the observation, whereas too small δ_2 values signify overprediction, which means that the forecast rate is too high to be consistent with the observation (Zechar et al 2010a).

3.5.4 Epidemic-type aftershock sequence model

The Epidemic-Type Aftershock Sequence (ETAS) model, introduced by Ogata (1988) , is now considered to be the state-of-the-art seismicity forecasting model. It is used for operational forecasting in the United States, New Zealand and Italy (Mizrahi et al 2024). The ETAS model treats seismicity as an epidemic, where earthquakes trigger subsequent earthquakes. It divides earthquakes into two categories: background events, which are those that are not triggered by previous events and occur uniformly, and triggered events. The expected number of aftershocks is determined by empirical laws related to aftershock productivity based on the magnitude of the parent event, the spatial distribution of aftershocks and the decrease of aftershock rate over time. This approach allows the ETAS model to capture the space-time clustering of aftershocks immediately following an initial event and to reflect the Omori-Utsu aftershock decay (Omori 1894; Utsu 1961). However, the ETAS approach is computationally expensive due to the fact that it requires an inversion procedure to estimate the parameters and a large number of simulations to generate each forecast (Harte 2017; Kamranzad et al 2025).

As the ML models were trained on a bulk catalogue dataset and are not optimised for specific regions, we estimate some initial global ETAS parameters based on the training part of the bulk catalogue dataset using events above the completeness threshold. To ensure a fair comparison against ML models, we then calibrate the ETAS parameters for each test instance using the events of magnitude 2 and above that occurred within the first seven days of the spatiotemporal sequence. This is done in order for the ETAS forecasts to include events of magnitude 2 and above, which is consistent with the threshold used in the ML forecasts. We use the ETAS implementation from the GitHub repository by Mizrahi et al (2023) and estimate the ETAS parameters based on the expectation maximisation (EM) algorithm (Mizrahi et al 2021b). It is worth noting that the EM algorithm relies on a complete dataset for training, hence the resulting ETAS parameters may be biased due to the use of incomplete data in the parameter calibration step. Once the final parameters have been estimated, we generate M2+ forecasts for the eighth day by performing 100 simulations and calculating the mode of the number of events per grid cell across all simulations. Using the mode results in more reliable forecasts compared to the mean value, as the mean is highly influenced by outliers.

3.5.5 Persistence model

A common benchmark that is used for evaluating forecasting approaches is the persistence model (Hewamalage et al 2023). This is a simple baseline that doesn't require any computations, as it assumes that no change occurs between consecutive time steps. It therefore uses the previous day's map as the next day's forecast, which

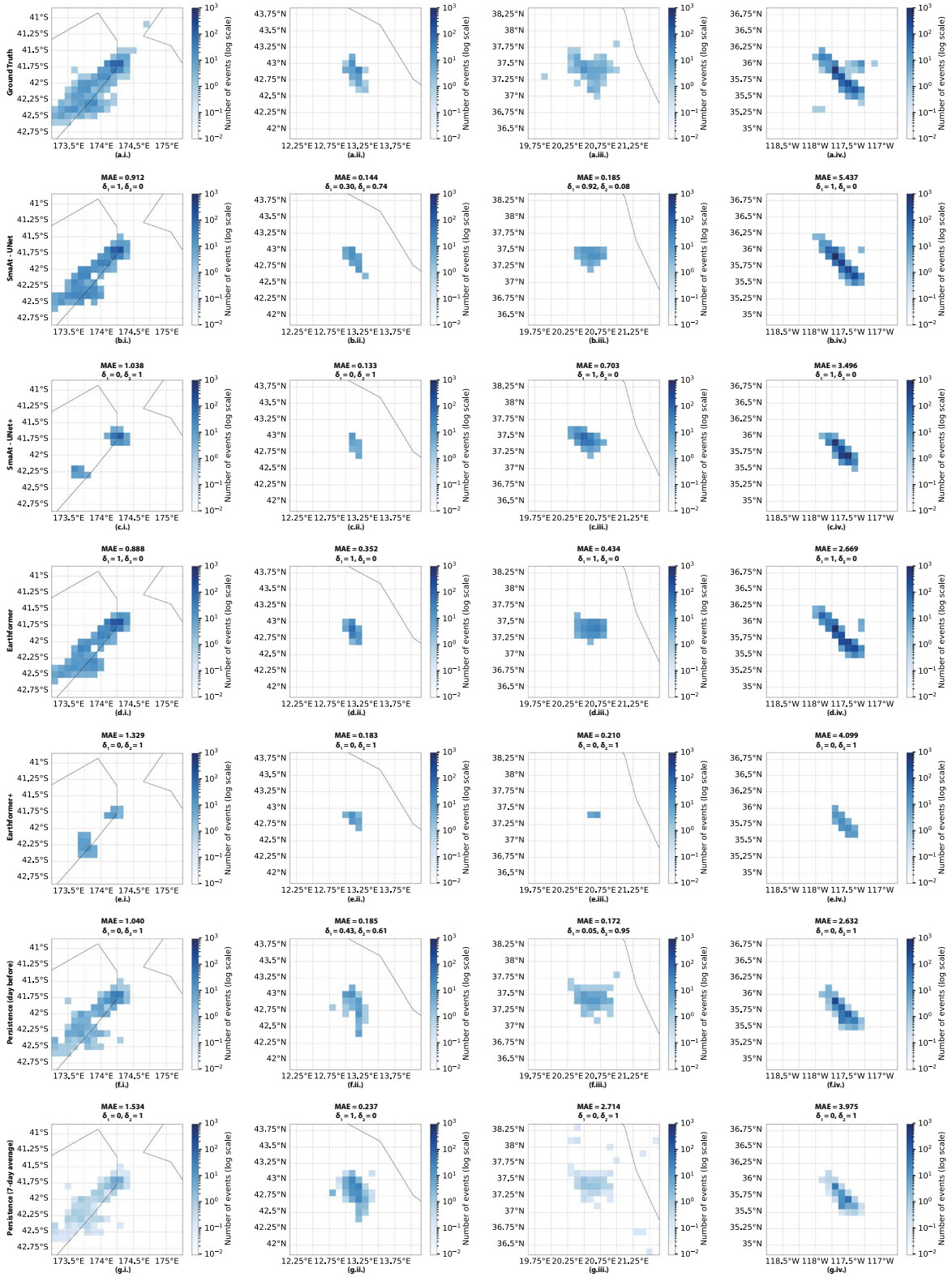


Fig. 3 Randomly selected examples of data-driven deep learning-based forecasts for datapoints that belong to the bulk catalogue data test set. Comparison between ground truth and the forecasts generated by SmaAt-UNet, SmaAt-UNet+, Earthformer, Earthformer+, persistence (day before) and persistence (average of previous 7 days). The columns correspond to the following events: (i) 13/11/2016, M4.2, New Zealand, (ii) 14/11/2016, M4.1, Italy, (iii) 26/10/2018, M4.1, Greece, (iv) 06/07/2019, M4.2, Southern California. BGS©UKRI 2025

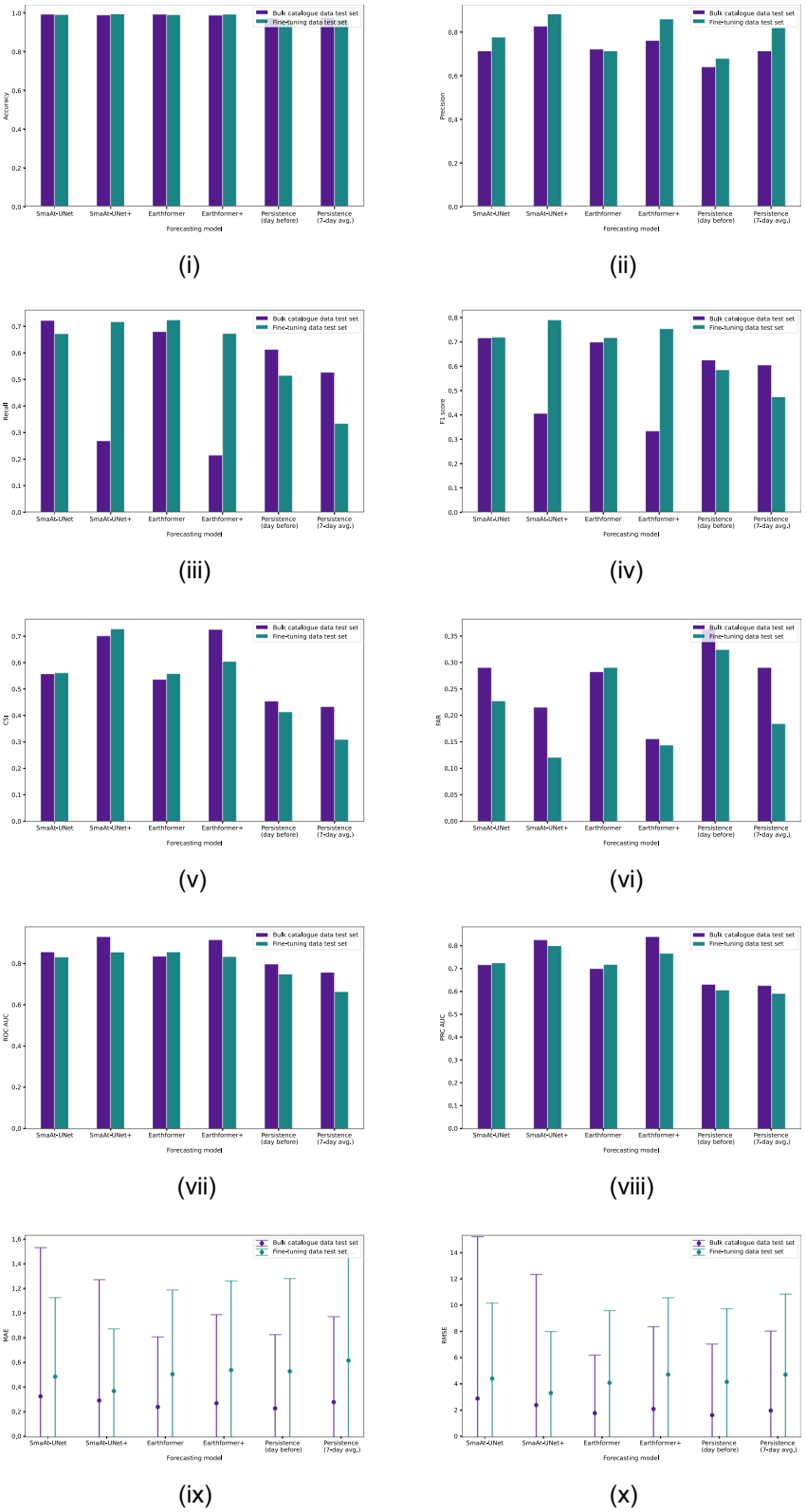


Fig. 4 Evaluation metrics (accuracy, precision, recall, F1 score, CSI, FAR, ROC AUC, PRC AUC, MAE and RMSE) on the bulk catalogue and fine-tuning data test sets. Comparison between SmaAt-UNet, SmaAt-UNet+, Earthformer, Earthformer+, persistence (day before) and persistence (average of previous 7 days). BGS©UKRI 2025

Table 5 Training and fine-tuning CNNs and transformers using a temporal data split: Classification and regression evaluation metrics (bulk catalogue data test set). For regression metrics (MAE and RMSE), the mean and the standard deviation of the errors are reported. BGS©UKRI 2025

Evaluation metric	SmaAt-UNet	SmaAt-UNet+	Earthformer	Earthformer+	Persistence (day before)	Persistence (7-day avg.)
Accuracy ↑	0.990	0.986	0.990	0.985	0.971	0.973
Precision ↑	0.710	0.823	0.718	0.758	0.637	0.710
Recall ↑	0.721	0.268	0.679	0.214	0.612	0.526
F1 score ↑	0.715	0.405	0.698	0.333	0.624	0.604
CSI ↑	0.557	0.254	0.536	0.200	0.454	0.433
FAR ↓	0.290	0.177	0.282	0.242	0.363	0.290
ROC AUC ↑	0.858	0.634	0.837	0.606	0.799	0.759
PRC AUC ↑	0.718	0.552	0.701	0.492	0.632	0.627
MAE ($\boldsymbol{\mu}, \boldsymbol{\sigma}$) ↓	0.330,1.206	0.296,0.980	0.244,0.568	0.274, 0.719	0.232,0.599	0.283,0.693
RMSE ($\boldsymbol{\mu}, \boldsymbol{\sigma}$) ↓	2.933, 12.331	2.430,9.950	1.822,4.427	2.137,6.265	1.666,5.409	2.016,6.058

Table 6 Training and fine-tuning CNNs and transformers using a temporal data split: Classification and regression evaluation metrics (high-resolution catalogue test set). For regression metrics (MAE and RMSE), the mean and the standard deviation of the errors are reported. BGS©UKRI 2025

Evaluation metric	SmaAt-UNet	SmaAt-UNet+	Earthformer	Earthformer+	Persistence (day before)	Persistence (7-day avg.)
Accuracy ↑	0.988	0.991	0.987	0.990	0.963	0.963
Precision ↑	0.773	0.879	0.710	0.856	0.676	0.816
Recall ↑	0.671	0.716	0.723	0.672	0.514	0.333
F1 score ↑	0.718	0.789	0.716	0.753	0.584	0.473
CSI ↑	0.561	0.727	0.558	0.604	0.413	0.310
FAR ↓	0.227	0.121	0.290	0.144	0.324	0.184
ROC AUC ↑	0.833	0.857	0.858	0.835	0.750	0.665
PRC AUC ↑	0.726	0.801	0.719	0.768	0.607	0.592
MAE ($\boldsymbol{\mu}, \boldsymbol{\sigma}$) ↓	0.490,0.641	0.373,0.504	0.510,0.682	0.543, 0.724	0.533,0.752	0.621, 0.868
RMSE ($\boldsymbol{\mu}, \boldsymbol{\sigma}$) ↓	4.455,5.754	3.532,4.691	4.131,5.500	4.762, 5.850	4.203,5.569	4.752,6.139

is often a highly accurate estimate due to the fact that the maps of consecutive days are usually highly correlated (Armstrong 2001). While the use of persistence as a baseline is common in other fields, like for example weather forecasting (Mittermaier 2008), it is not part of current practice in earthquake forecasting. Nevertheless, it may be a suitable null hypothesis that is not yet part of the CSEP protocols. In the context of seismicity forecasting, persistence is able to capture spatiotemporal clustering once it has started since the input data is from one or several days of prior seismicity in the target area, which is already clustered spatially and in time during a seismic sequence. However, it is not expected to be able to forecast the onset of a mainshock at the start of a sequence (also a property of the ETAS model) or the Omori-Utsu law decay in longer time windows (Omori 1894; Utsu 1961), which can be captured by the ETAS model. In this study, we use two different versions of persistence against which we compare our models: the previous day's map and the average of the daily maps of the previous 7 days. The first version is the standard baseline that is commonly used in forecasting, whereas the second version is a variation that incorporates information from the whole data sequence used as input to the ML forecasting models.

4 Experiments

4.1 Training and fine-tuning CNNs and transformers

We aim to assess different deep learning models' ability to produce data-driven earthquake forecasts using earthquake catalogue data. To do this, we train different models using a bulk catalogue training set mostly comprised of standard manually-derived catalogues.We then use the trained weights and fine-tune the models with the use of high-resolution catalogues. Specifically, we train the two deep learning models described above, the SmaAt-UNet and the Earthformer, using a bulk catalogue training set that consists of training examples from the standard

catalogues of Southern California, Northern California, New Zealand, Italy and Greece and the unified high-resolution relocated catalogue for Japan. The data used are split sequentially, with older data used for training and newer data used for testing, in line with the guidelines for pseudo-prospective testing (Mizrahi et al 2024). We use 80% of the data used for training, 10% for validation and 10% for testing. We then use the initial trained model weights and continue training the models with a new training set built solely from high-resolution catalogues: the Southern California QTM catalogue and the Italy ML catalogue. To avoid data leakage, we don't use the parts of the Southern California and Italy standard catalogues that cover the same time interval as the Southern California QTM catalogue and the Italy ML catalogue for training. The fine-tuned models are henceforth referred to as SmaAt-UNet+ and Earthformer+. We explored the option of freezing part of the trained network and updating the last few layers based on the new data, a practice commonly used in fine-tuning. However, this did not improve performance; hence the results reported here correspond to the case where all weights are updated. The number of parameters, the training time on a single NVIDIA Quadro RTX 4000 GPU , the learning rate at the start of the training process and the number of epochs each model was trained for can be seen in Table 4.

4.2 Comparison between ML-based and ETAS forecasts

In order to assess the forecasting potential of ML models, we need to compare them in terms of performance against an ETAS model, the most widely-used seismicity forecasting model. We quantitatively compare the forecasts generated by SmaAt-UNet and Earthformer against those generated by a baseline ETAS model (Mizrahi et al 2023) with initial global parameters estimated by the EM algorithm using the training part of our bulk catalogue dataset, taking into account events above catalogue completeness. The ETAS parameters were calibrated for each test instance based on the M2+ events that occurred within the first seven days of the spatiotemporal sequence, which is the part that is given as input to the ML models. We use two study regions, Southern California and Italy, as our testing ground. This was done due to the high computational cost of generating ETAS forecasts, which meant that performing a comparative study for all the catalogues in the bulk catalogue dataset was not feasible within a reasonable timeframe with the computational infrastructure that was used for ML model training and inference. We calculate the classification, regression and CSEP metrics for all three models (SmaAt-UNet, Earthformer and ETAS) and compare them against each other as well as against the persistence baseline.

4.3 Investigating the impact of individual catalogues and different data splitting strategies

We seek to understand the impact of individual catalogues and different train-test splitting strategies in the training process. We therefore investigate the possibility of train ing SmaAt-UNet and Earthformer on the bulk catalogue dataset using randomly selected training, validation and test instances. This means that instances across the whole time interval that we use in this study are used both for training and testing, which introduces look-ahead bias (Peixeiro 2022) but also allows the model to be trained on more recent catalogue data, which generally have lower magnitudes of completeness compared to older parts of the catalogues. We also train

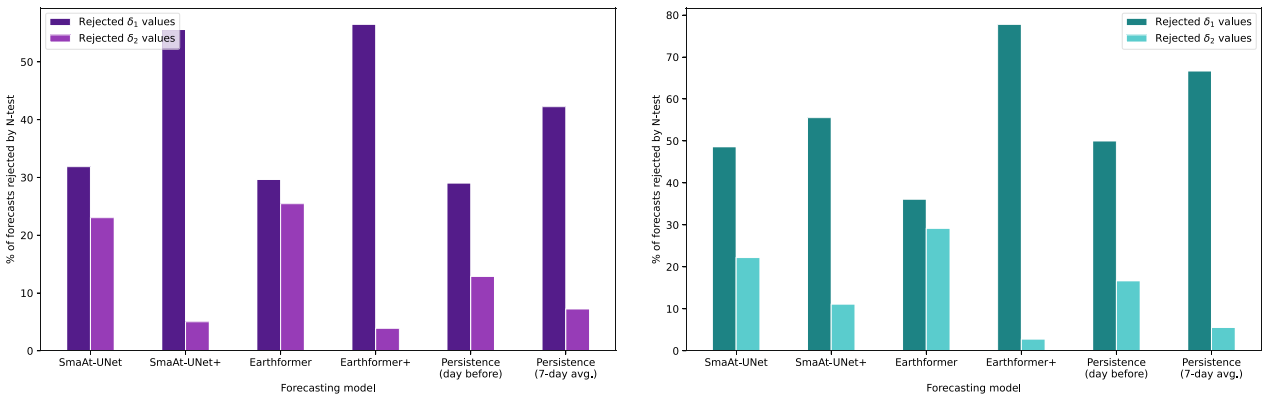


Fig. 5 Percentage of underpredictions (rejected δ_1 values) and overpredictions (rejected δ_2 values). Comparison between SmaAt-UNet, SmaAt-UNet+, Earthformer, Earthformer+, persistence (day before) and persistence (average of previous 7 days): (**i**) bulk catalogue data test set, (**ii**) fine-tuning data test set. BGS©UKRI 2025

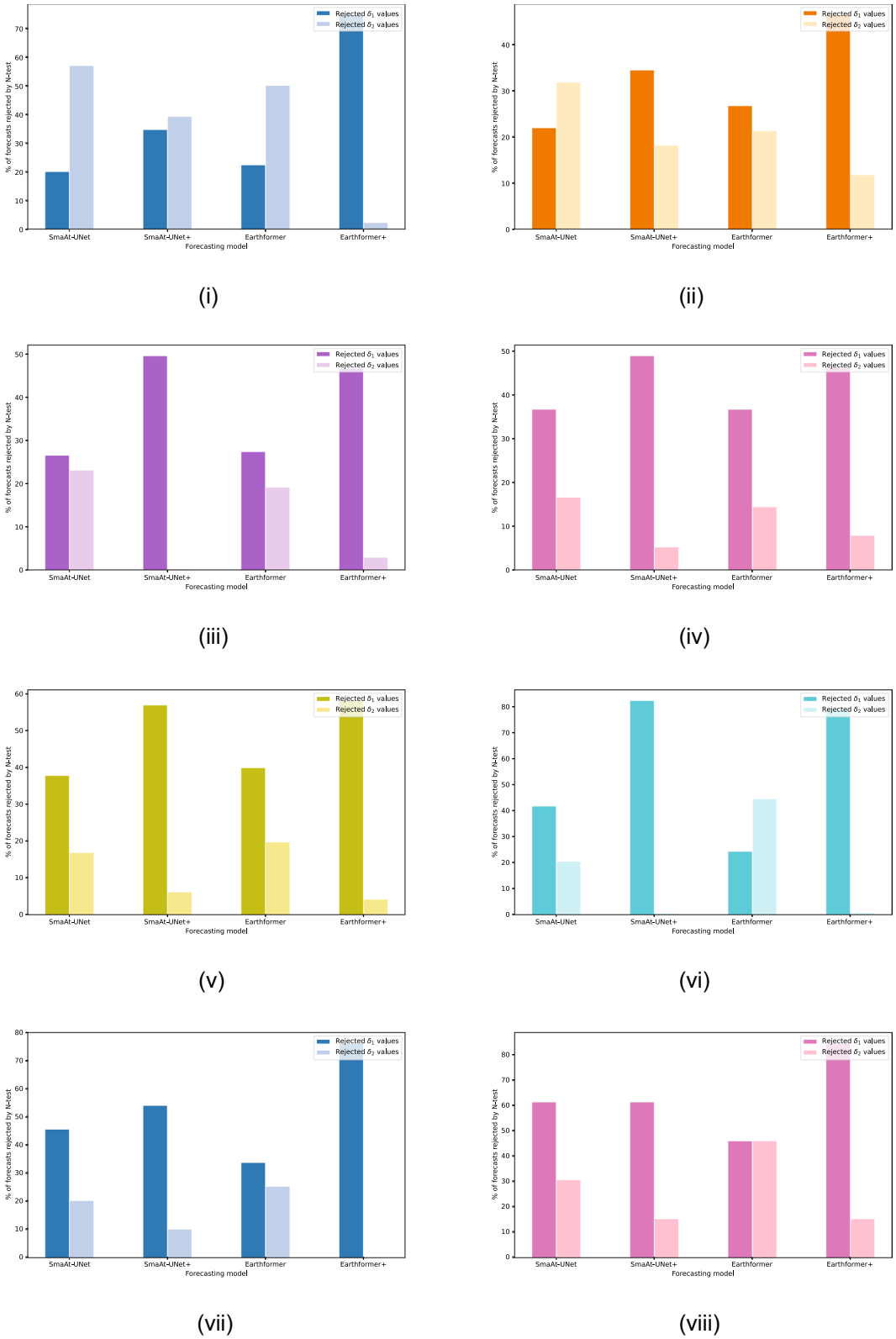


Fig. 6 Percentage of underpredictions (rejected δ_1 values) and overpredictions (rejected δ_2 values). Comparison between SmaAt-UNet, SmaAt-UNet+, Earthformer and Earthformer+: (**i**) Southern California catalogue, (**ii**) Northern California catalogue, (**iii**) New Zealand catalogue, (**iv**) Italy catalogue, (**v**) Greece catalogue, (**vi**) Japan catalogue, (**vii**) Southern California QTM catalogue, (**viii**) Italy ML catalogue. BGS©UKRI 2025

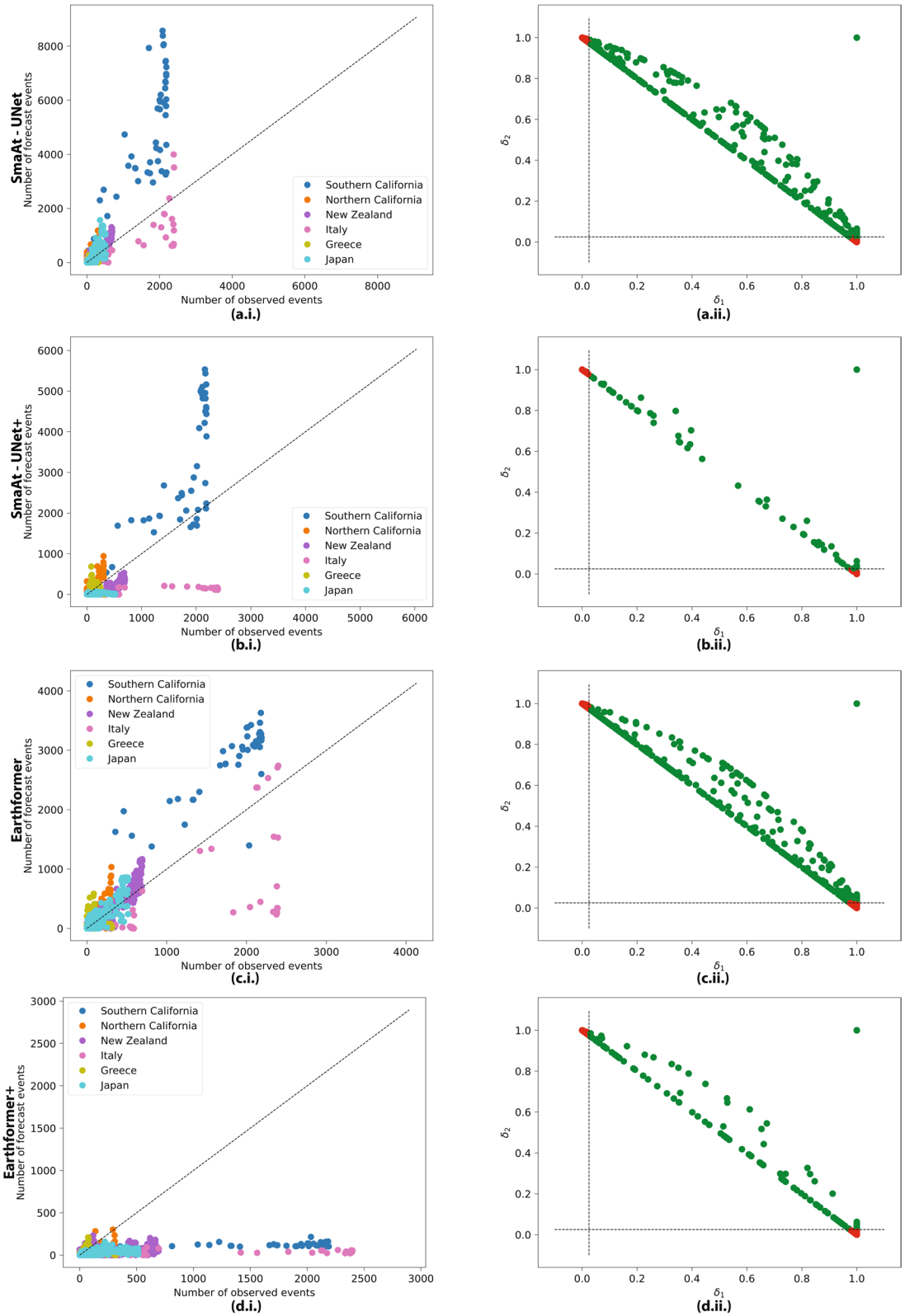


Fig. 7 N-test results on bulk catalogue data test set: underforecasting and overforecasting. (**i**) Observed versus forecast number of events. Points on or close to the $y = x$ line represent forecasts that are consistent with the observations. (**ii**) N-test δ_1 and δ_2 values. Forecasts with $\delta_1 > 0.025$ and $\delta_2 > 0.025$ are consistent with the observations. BGS©UKRI 2025

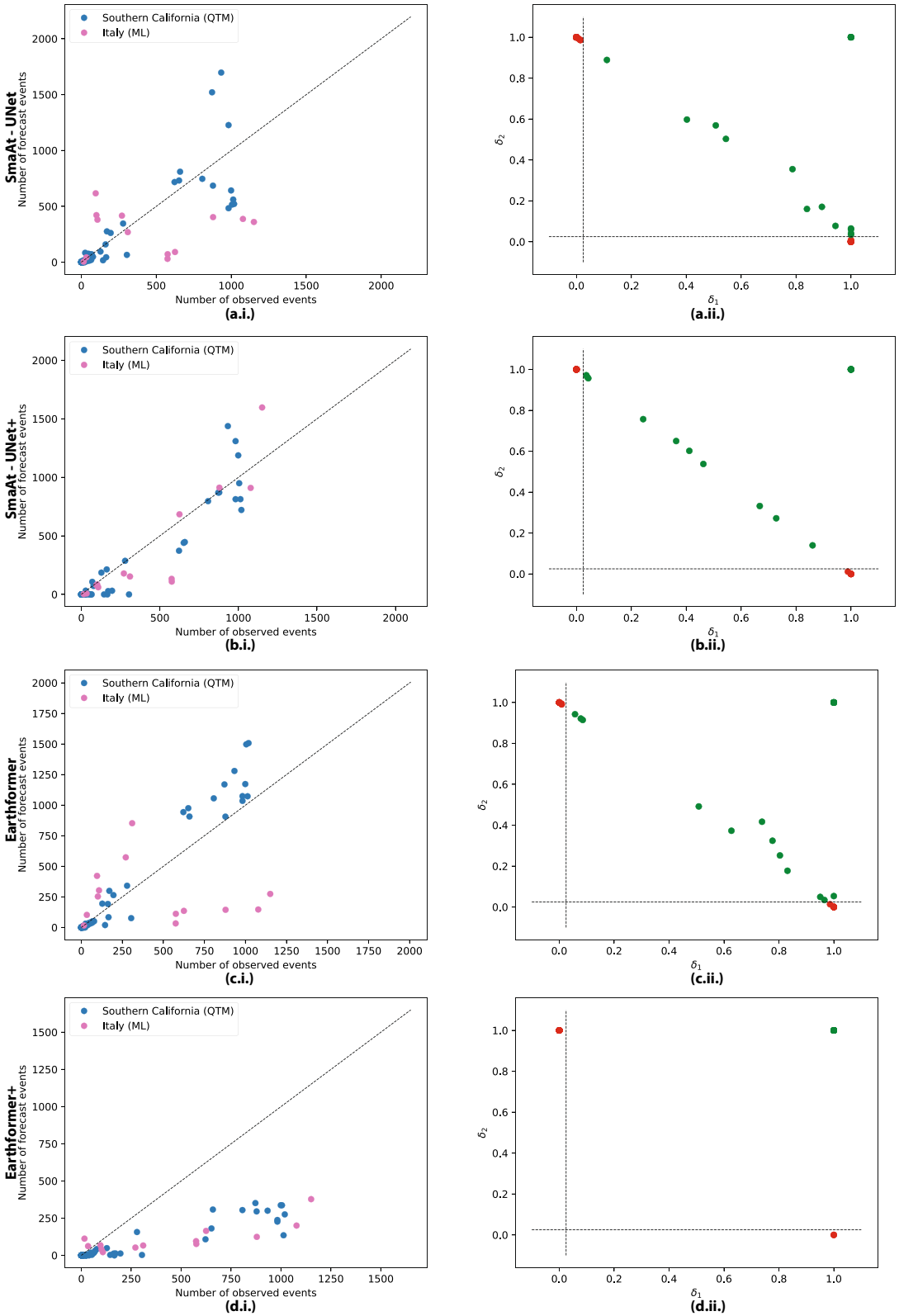


Fig. 8 N-test results on fine-tuning data test set: underforecasting and overforecasting. i) Observed versus forecast number of events. Points on or close to the $y = x$ line represent forecasts that are consistent with the observations. ii) N-test δ_1 and δ_2 values. Forecasts with $\delta_1 > 0.025$ and $\delta_2 > 0.025$ are consistent with the observations. BGS©UKRI 2025

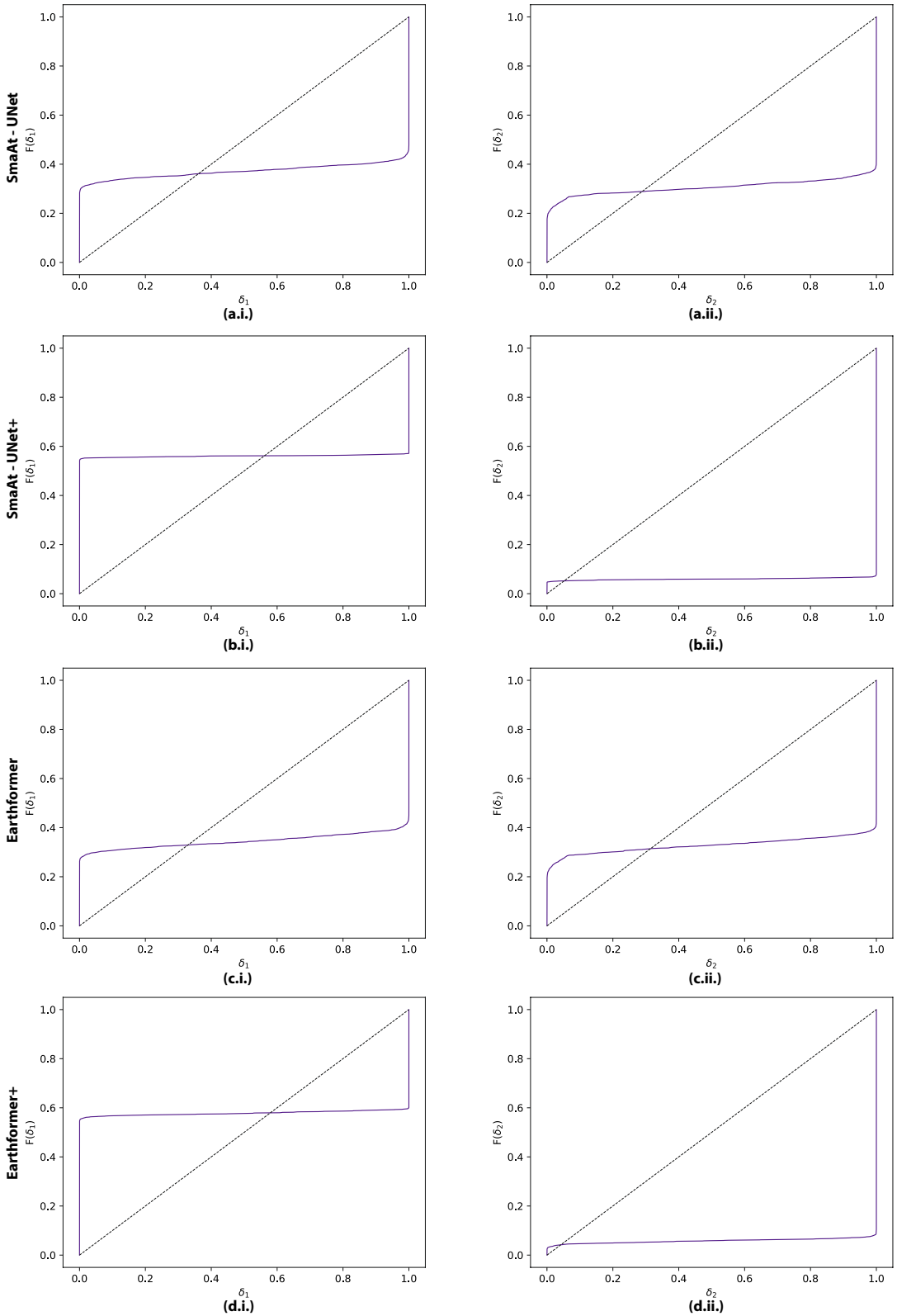


Fig. 9 Cumulative distributions of δ_1 and δ_2 values (bulk catalogue data test set). The uniform distribution, which corresponds to a perfectly calibrated model, is shown with a dashed line. In our case, most of the cumulative distributions of the quantile scores show underdispersion, which means that the observed data have less variation than the forecasts. Figures b.(**ii**) and d.(**ii**) indicate underprediction, as the cumulative distribution plot is almost constantly below the uniform distribution (Savran et al 2020). BGS©UKRI 2025

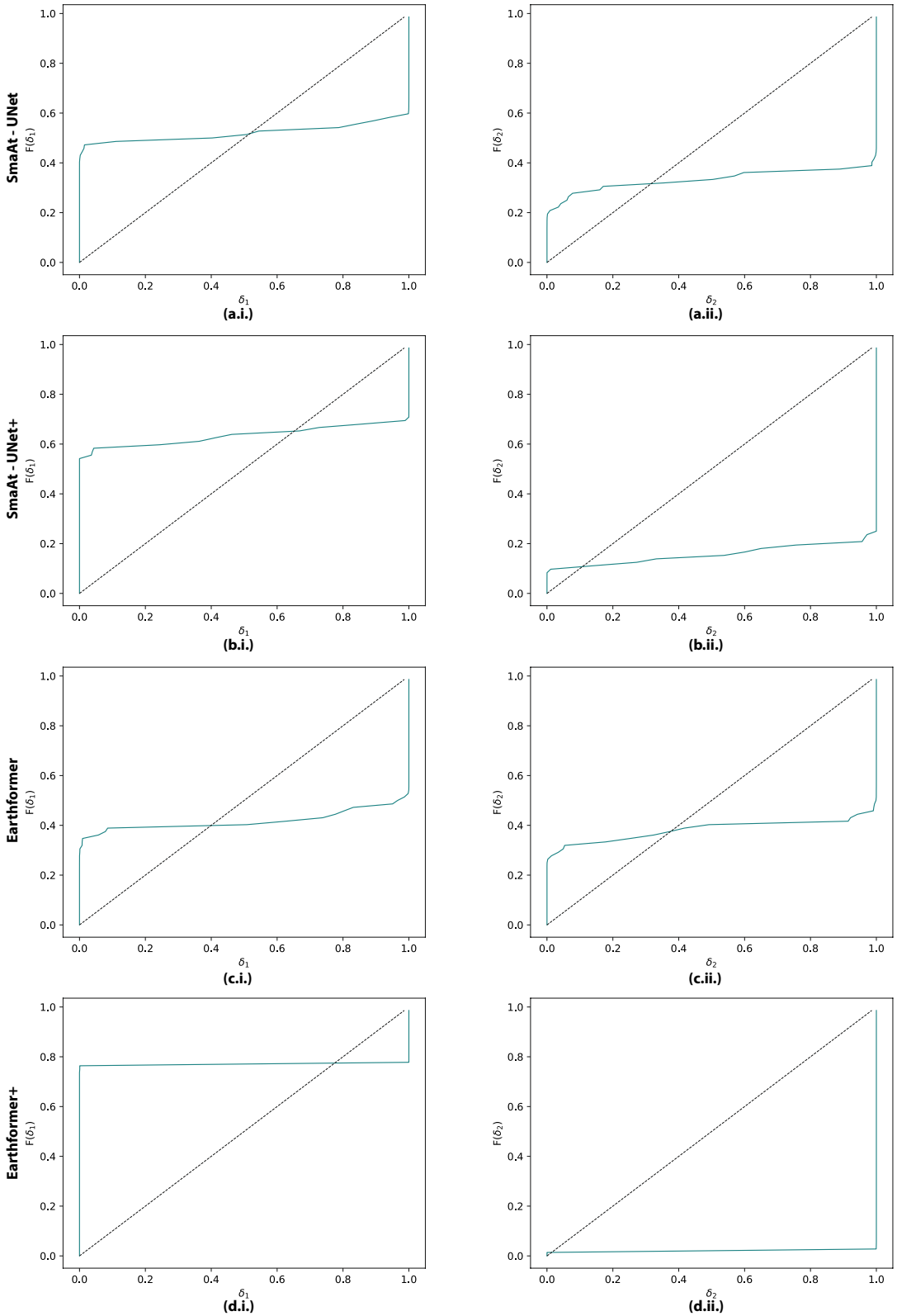


Fig. 10 Cumulative distributions of δ_1 and δ_2 values (fine-tuning data test set). The uniform distribution, which corresponds to a perfectly calibrated model, is shown with a dashed line. In our case, the cumulative distributions of the quantile scores show underdispersion, which means that the observed data have less variation than the forecasts. Figures b.ii) and d.ii) indicate underprediction, as the cumulative distribution plot is almost constantly below the uniform distribution (Savran et al 2020). BGS©UKRI 2025

Table 7 Training and fine-tuning CNNs and transformers using a temporal data split: Percentage of underpredictions (rejected δ_1 values) and overpredictions (rejected δ_2 values) (bulk catalogue data test set). BGS©UKRI 2025

		SmaAt-UNet	SmaAt-UNet+	Earthformer	Earthformer+
All	Rejected $\boldsymbol{\delta_1}$	31.87%	55.60%	29.65%	56.49%
	Rejected $\boldsymbol{\delta_2}$	23.04%	5.03%	25.46%	3.89%
Southern California	Rejected $\boldsymbol{\delta_1}$	20.00%	34.62%	22.31%	74.62%
	Rejected $\boldsymbol{\delta_2}$	56.92%	39.23%	50.00%	2.31%
Northern California	Rejected $\boldsymbol{\delta_1}$	21.90%	34.39%	26.67%	46.35%
	Rejected $\boldsymbol{\delta_2}$	31.75%	18.10%	21.27%	11.75%
New Zealand	Rejected $\boldsymbol{\delta_1}$	26.56%	49.59%	27.38%	46.66%
	Rejected $\boldsymbol{\delta_2}$	23.11%	0.23%	19.13%	2.93%
Italy	Rejected $\boldsymbol{\delta_1}$	36.68%	48.91%	36.68%	46.29%
	Rejected $\boldsymbol{\delta_2}$	16.59%	5.24%	14.41%	7.86%
Greece	Rejected $\boldsymbol{\delta_1}$	37.86%	57.01%	39.97%	58.22%
	Rejected $\boldsymbol{\delta_2}$	16.89%	6.18%	19.76%	4.22%
Japan	Rejected $\boldsymbol{\delta_1}$	41.82%	82.44%	24.40%	79.02%
	Rejected $\boldsymbol{\delta_2}$	20.54%	0.60%	44.64%	0.74%

Table 8 Training and fine-tuning CNNs and transformers using a temporal data split: Percentage of underpredictions (rejected δ_1 values) and overpredictions (rejected δ_2 values) (fine-tuning data test set). BGS©UKRI 2025

		SmaAt-UNet	SmaAt-UNet+	Earthformer	Earthformer+
All	Rejected $\boldsymbol{\delta_1}$	48.61%	55.56%	36.11%	77.78%
	Rejected $\boldsymbol{\delta_2}$	22.22%	11.11%	29.17%	2.78%
Southern California (QTM)	Rejected $\boldsymbol{\delta_1}$	45.76%	54.24%	33.90%	76.27%
	Rejected $\boldsymbol{\delta_2}$	20.34%	10.17%	25.42%	0.00%
Italy (ML)	Rejected $\boldsymbol{\delta_1}$	61.54%	61.54%	46.15%	84.62%
	Rejected $\boldsymbol{\delta_2}$	30.77%	15.38%	46.15%	15.38%

SmaAt-UNet on the Southern California and New Zealand catalogues separately. The Southern California catalogue is one of the most detailed catalogues that we have available, whereas the New Zealand catalogue contains the highest number of M4+ events and hence makes up the largest part of the training data. We investigate the use of a sequential data split, i.e. using older data for training and newer data for testing, as well as a random data split, where we randomly select the training and test examples.

4.4 Exploring how the use of different types of input maps influences performance

We also wish to explore how different types of input maps (rate, maximum magnitude and average depth) contribute to the overall performance. We therefore employ the same bulk catalogue dataset to train the SmaAt-UNet model first using only the rate maps and then using the rate and maximum magnitude maps as inputs. We compare the performance of the two trained models with each other as well as with the SmaAt-UNet model that uses all three types of maps. A random data splitting strategy is used in this case as well, hence look-ahead bias has been introduced here too. Nonetheless, this comparison still shows the difference in performance when different types of input maps are used.

5 Results and discussion

5.1 Behaviour of different deep learning architectures

As seen in Table 4, SmaAt-UNet is a much more compact model and the time needed for training it is significantly lower compared to the time needed to train the Earthformer. SmaAt-UNet is therefore more suitable for operational applications, as fast training times allow for further training at regular time intervals to continuously incorporate new data. In terms of performance, the SmaAt-UNet and the Earthfomer do similarly according to our evaluation metrics, leading to the conclusion that a larger number of parameters does not necessarily improve performance when the training dataset is relatively small, as

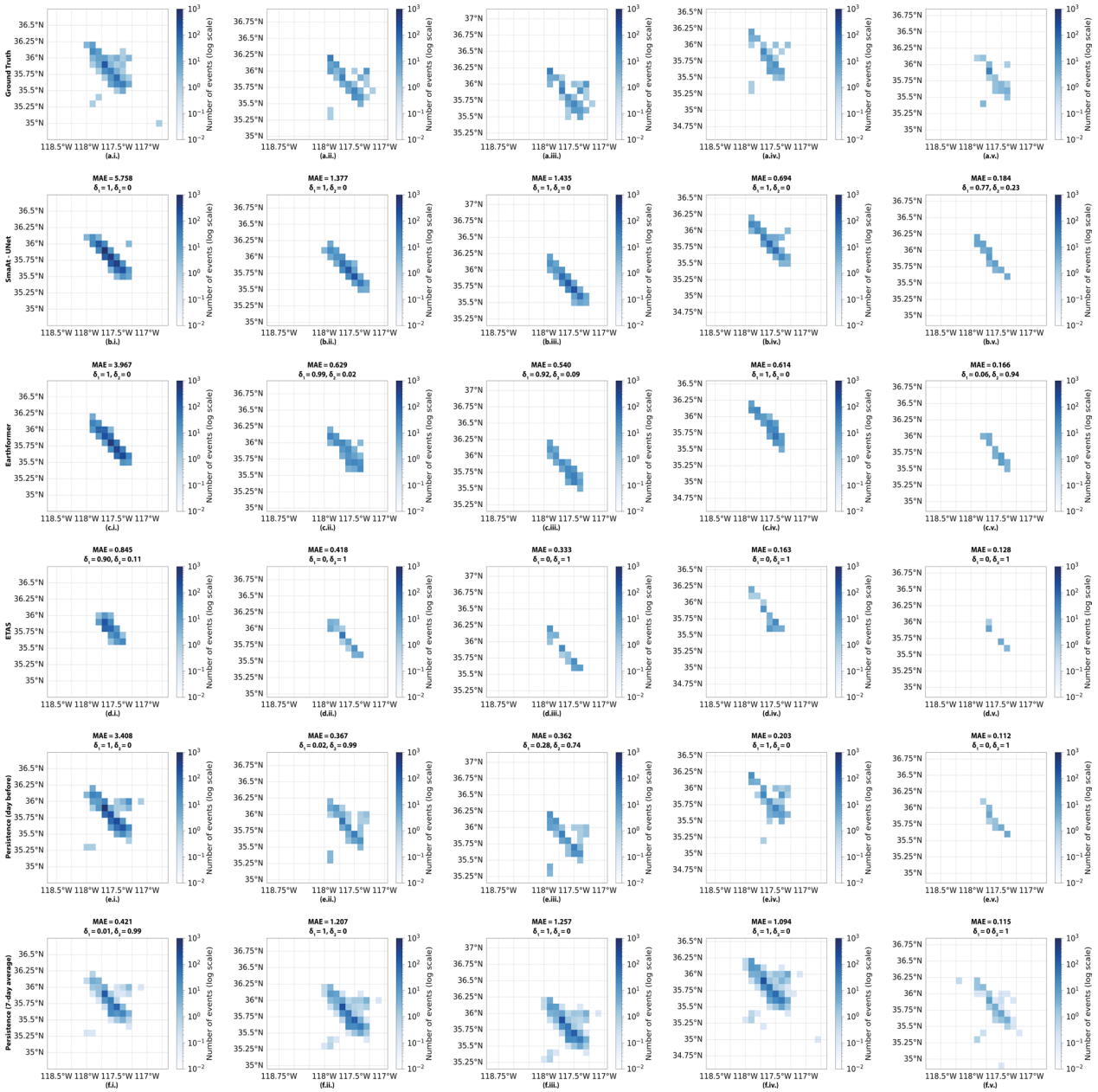


Fig. 11 Comparison between ML, ETAS and persistence forecasts following events that occurred within the 2019 Ridgecrest sequence in Southern California. The columns correspond to the following events: (i) 07/07/2019, M4.5, (ii) 10/07/2019, M4.2, (iii) 11/07/2019, M4.5, (iv) 12/07/2019, M4.9, (v) 26/07/2019, M4.7. BGS©UKRI 2025

in such cases the models are prone to overfitting (Lever et al 2016; Brigato and Iocchi 2021). The availability of earthquake catalogues is limited in duration, restricting the amount of data that can be used to train the models in this study. In such cases, the diversity and quality of the dataset play an important role in the success of the training process, as it is essential to expose the models to a dataset that is representative of as many different situations as possible in order to achieve generalisation. Furthermore, the Earthformer's inability to surpass the forecasting skill of the SmaAt-UNet model is indicative of the fact that transformer-based models need to be exposed to larger training datasets, as they lack the inductive bias that is inherent in convolutional neural networks (Dosovitskiy 2020; Gao et al 2022).

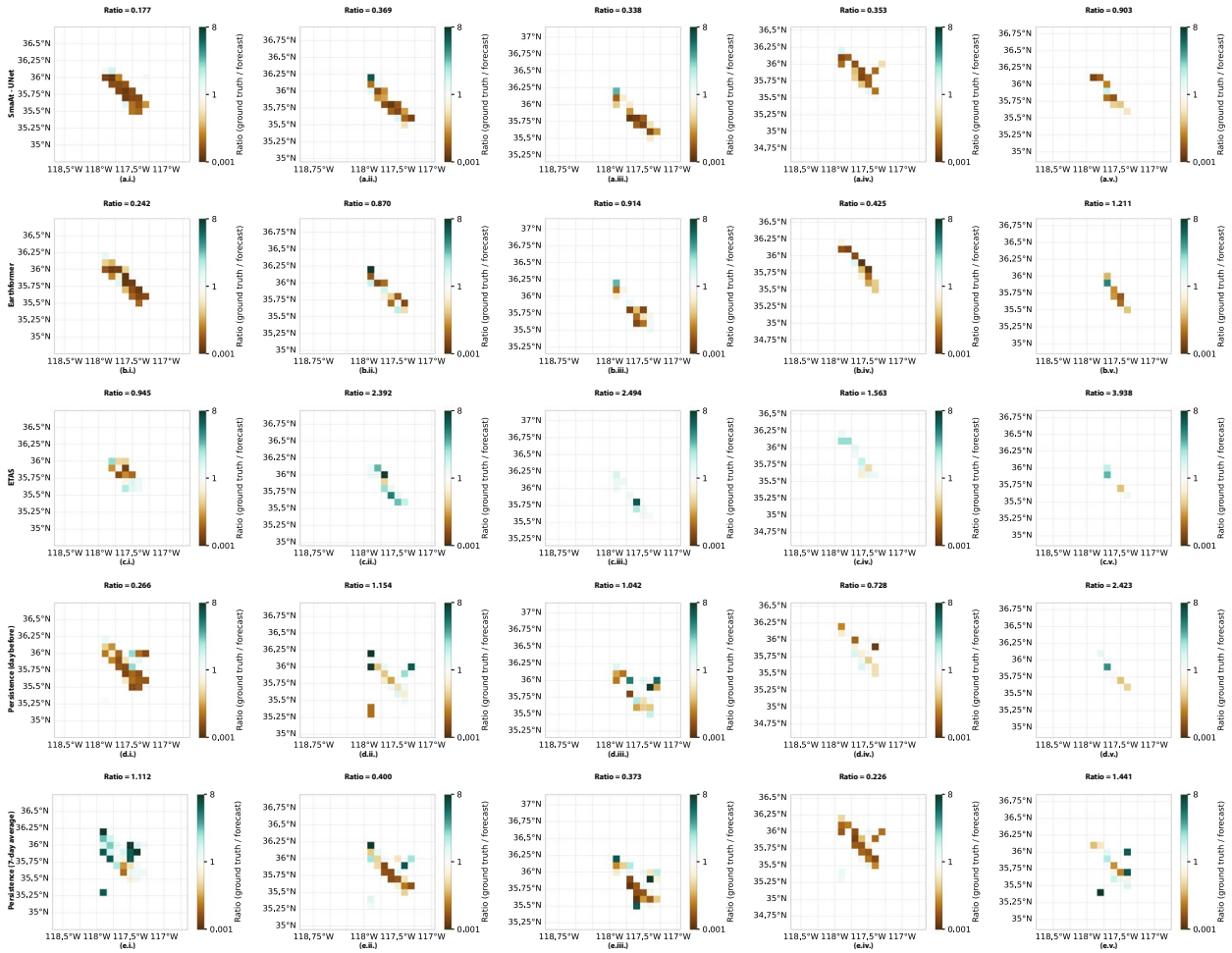


Fig. 12 Comparison between the ratios of ground truth to forecasts generated by ML, ETAS and persistence following events that occurred within the 2019 Ridgecrest sequence in Southern California. The columns correspond to the following events: (i) 07/07/2019, M4.5, (ii) 10/07/2019, M4.2, (iii) 11/07/2019, M4.5, (iv) 12/07/2019, M4.9, (v) 26/07/2019, M4.7. BGS©UKRI 2025

5.2 Qualitative and quantitative evaluation of forecasts in terms of rate and spatial distribution

Randomly selected examples of next-day forecasts produced for datapoints that belong to the bulk catalogue data test set can be seen in Fig. 3, where we qualitatively observe that the model outputs are generally consistent with the ground truth maps. Fig. 4 shows the model performance on both the bulk cataloguestandard and fine-tuning data test sets. The trained ML models seem to be able to produce forecasts that are relatively consistent with the observations in terms of number of events and spatial locations. This is evidenced by the values of the precision (0.710−0.718), the recall (0.679−0.721), the F1 score (0.698−0.715), the CSI (0.536−0.557) and the PRC AUC (0.701−0.718), which can be seen in Table 5. The FAR is low for both models (0.282-−0.290), which is also a positive result, as generating a large number of false alarms decreases the robustness of forecasting models. The MAE is equal to 0.330±1.206 for SmaAt-UNet and 0.244±0.568 for Earthformer, whereas the RMSE is 2.933±12.331 for SmaAt-UNet and 1.822±4.427 for Earthformer. SmaAt-UNet and Earthformer have similar performance, with SmaAt-UNet being slightly superior in terms of F1 score, CSI and PRC AUC and Earthformer having lower MAE and RMSE.. The fine-tuned models perform worse on the bulk catalogue data test set and have low recall and CSI scores, which indicates a low proportion of correctly classified actual positives and a large number of false negatives.

As can be seen in Table 6, the fine-tuned models perform better than the previous models when tested on data points from the fine-tuning dataset. This highlights the importance of fine-tuning ML models on data that are relevant to the application they will be used for. For example, if the goal is to build a model to forecast seismicity in a specific geographic region, it makes sense to

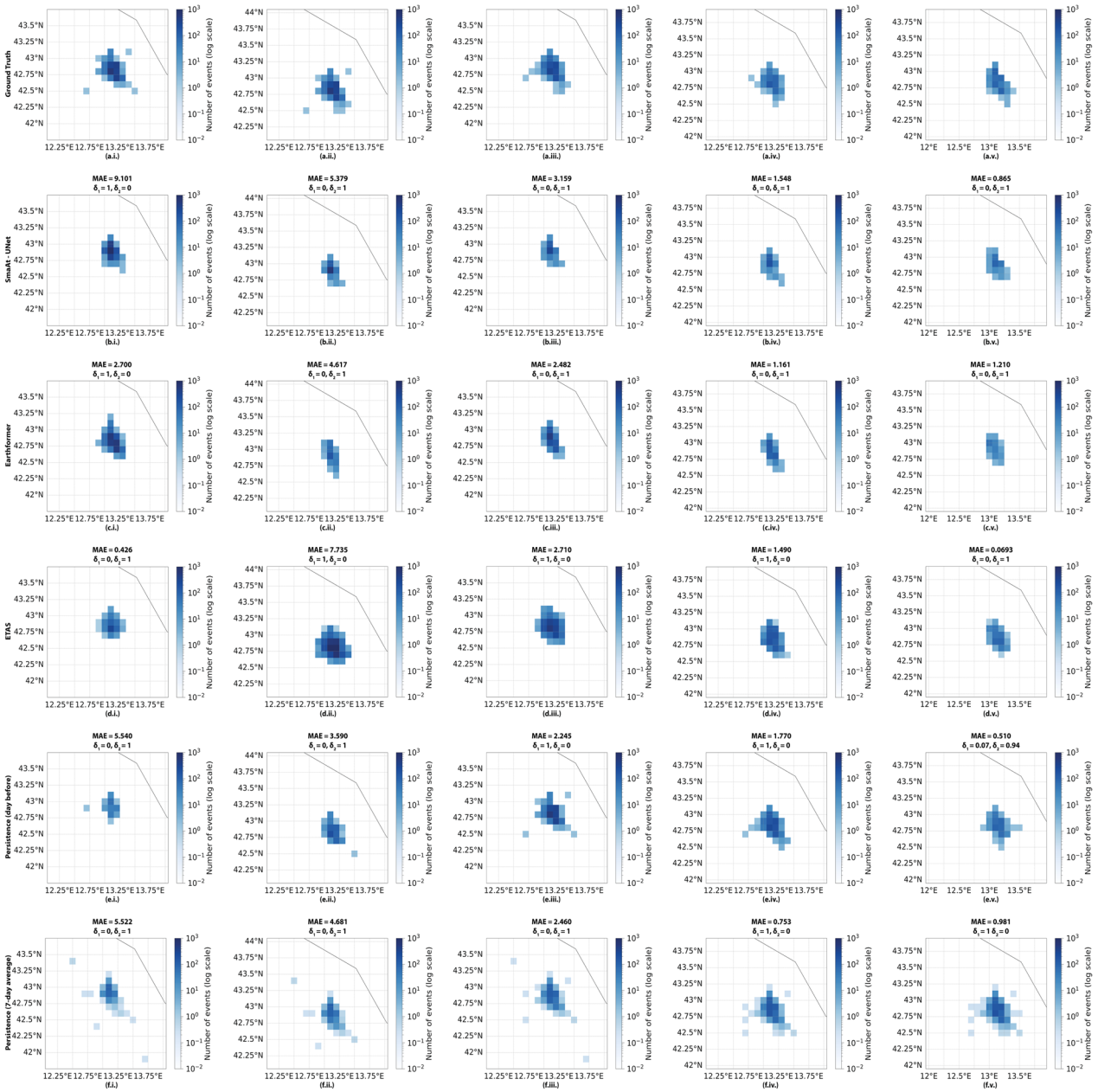


Fig. 13 Comparison between ML, ETAS and persistence forecasts following events that occurred within the 2016–2017 sequence in the Central Apennines. The columns correspond to the following events: (i) 30/10/2016, M4.0, (ii) 30/10/2016, M4.0, (iii) 31/10/2016, M4.0, (iv) 01/11/2016, M4.8, (v) 03/11/2016, M4.7. BGS©UKRI 2025

first train the model on a large bulk catalogue dataset that consists of data from different regions and then fine-tune on data specific to the region in which the model will be tested or used operationally. This indicates that the model is able to learn data properties that are inherent to specific catalogues, such as the level of completeness. SmaAt-UNet+ is the best performing model on the fine-tuning data test set but Earthformer+ also performs well, which again is evidenced by the evaluation metrics: precision (0.856–0.879), recall (0.672–0.716), F1 score (0.753–0.789), CSI (0.604–0.727), PRC AUC (0.768–0.801), FAR (0.121–0.144), MAE (0.373±0.504–0.543±0.724) and RMSE (3.532±4.691–4.762±5.850). The non fine-tuned models are less successful at forecasting the spatial distribution and number of events of the examples in the fine-tuning data test set. They achieve high precision

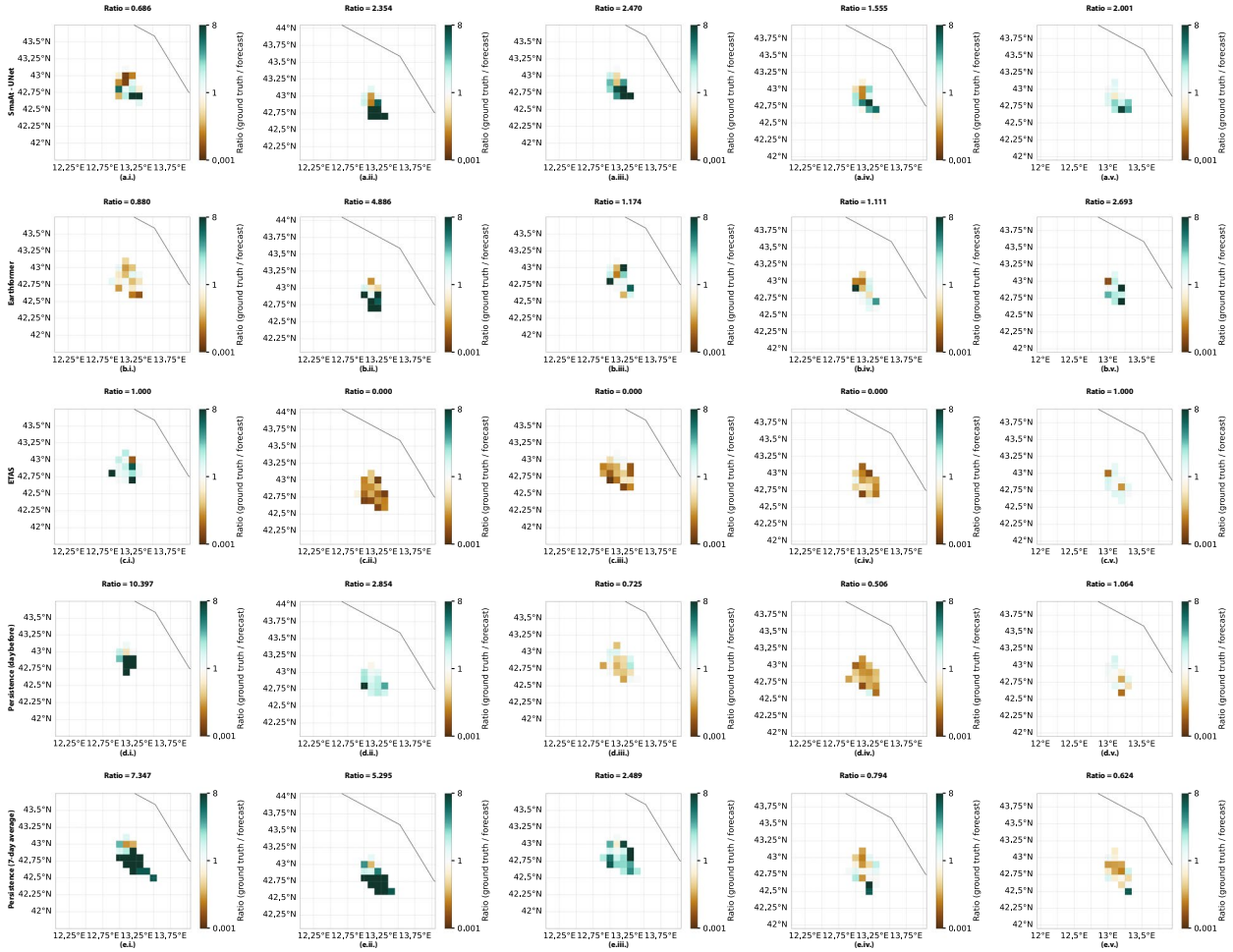


Fig. 14 Comparison between the ratios of ground truth to forecasts generated by ML, ETAS and persistence following events that occurred within the 2016–2017 sequence in the Central Apennines. The columns correspond to the following events: (i) 30/10/2016, M4.0, (ii) 30/10/2016, M4.0, (iii) 31/10/2016, M4.0, (iv) 01/11/2016, M4.8, (v) 03/11/2016, M4.7. BGS©UKRI 2025

values on this test set but the recall and CSI values are lower than those of the fine-tuned models, which means that when testing on the fine-tuning data test set, their use leads to more undetected actual positives compared to when the fine-tuned models are used.

The qualitative (Fig. 3) and quantitative (Fig. 4 and Tables 5 and 6) evaluation based on metrics such as the CSI and the FAR show that both neural networks are able to produce forecasts that are consistent with the observations in terms of error metrics and spatial distribution of events provided they are trained on a dataset that is representative of the space and time they will be used in. The use of high-resolution catalogues (in SmaAt-UNet+ and Earthformer+) has not improved the performance on the bulk catalogue test set. This can be attributed to the fact that this dataset is mostly comprised of standard catalogues, which have larger magnitudes of completeness and thus contain smaller numbers of events. Introducing high-resolution data during fine-tuning might therefore have led to forecasting events that were not included in these initial catalogues. For similar reasons, SmaAt-UNet+ and Earthformer+ show improved performance compared to SmaAt-UNet and Earthformer when evaluating performance on the fine-tuning data test set.

5.3 Comparison of trained models with persistence

We compare the trained models against the persistence baseline, which assumes no change between consecutive time steps and uses either the map of the day before or the average map of the previous 7 days as the forecast. Looking at Tables 5 and 6, it is apparent that the trained models perform considerably better than persistence in terms of precision, recall, F1 score, CSI, PRC AUC, MAE and RMSE. However, the persistence model performs adequately considering that it is a model with zero parameters that can instantly produce a result without

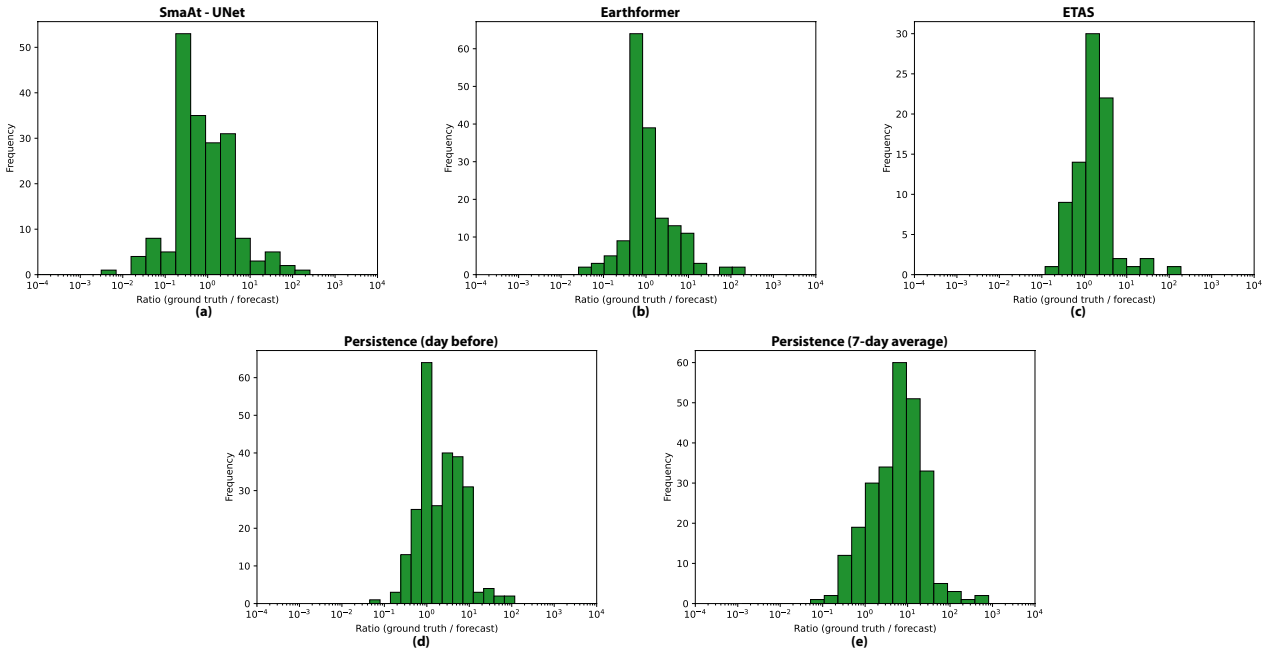


Fig. 15 Comparison between the ratios of ground truth to forecasts generated by ML, ETAS and persistence for all the examples in the Southern California and Italy test sets. BGS©UKRI 2025

the need for any training or computations. This is an expected behaviour due to the fact that seismicity shows clustering in time and space, making maps of consecutive days likely to be similar. In this sense, persistence is a better null hypothesis for forecasting than a Poisson assumption, hence its use as such in a variety of other applications (Hyndman and Athanasopoulos 2021; Mittermaier 2008; Knaff and Landsea 1997; Kumar et al 2024; Trebing et al 2021; Owens et al 2013; Stevenson et al 2022; Ghimire and Krajewski 2020; Bento et al 2022; Koprinska et al 2018; Pombo et al 2021; Voyant and Notton 2018; Chu et al 2017; Tziolis et al 2022).

5.4 Consistency between forecast and observed rates

The N-test indicates whether the number of observed events is consistent with the number of forecast events, with δ_1 values showing whether the models are underpredicting and δ_2 values showing whether the models are overpredicting. Tables 7 and 8 and Figures 5 and 6 show the percentage of underpredictions (rejected δ_1 values) and overpredictions (rejected δ_2 values). In most cases the percentage of rejected δ_1 and δ_2 values is smaller than the percentage of accepted values, which can be seen in Figs. 5, 6, 7 and 8. However, the models do not reproduce the distribution of events correctly, which is evidenced by the fact that the model is rejected more than 5% of the time for an α value of 0.05. This is also illustrated by the cumulative distributions of δ_1 and δ_2 values shown in Figs. 9 and 10, which are not uniformly distributed. The cumulative distributions show that we mostly have a case of underdispersion, which means that the variation of the observed data is less than that of the forecast data, whereas in a few cases the distribution of δ_2 values indicates underprediction (Savran et al 2020). Overall, the percentage of underpredictions is considerably greater than that of overpredictions, with the overpredictions usually not exceeding 10% of the test set. This can be seen in Figs. 5 and 6 and in column i) of Figs. 7 and 8, which show the number of observed events versus the number of forecast events. These figures also show that the introduction of high-resolution catalogues during fine-tuning increases the percentage of underpredictions and reduces the percentage of overpredictions. Looking at the daily number of observed events in weekly sequences that resulted in rejected δ_1 values and rejected δ_2 values (which can be seen in the supplementary material), we observe that many of the weekly sequences for which we under– or overpredict have occurred within longer earthquake sequences. Looking at Fig. 5ii), we observe that Earthformer exhibits the most balanced performance in terms of underpredictions and overpredictions. However, we observe that overall underprediction remains the most critical issue , which is an indication of the fact that the models learn the incompleteness of the training data. The latter highlights the need for further research in ML-based earthquake detection to address short-term

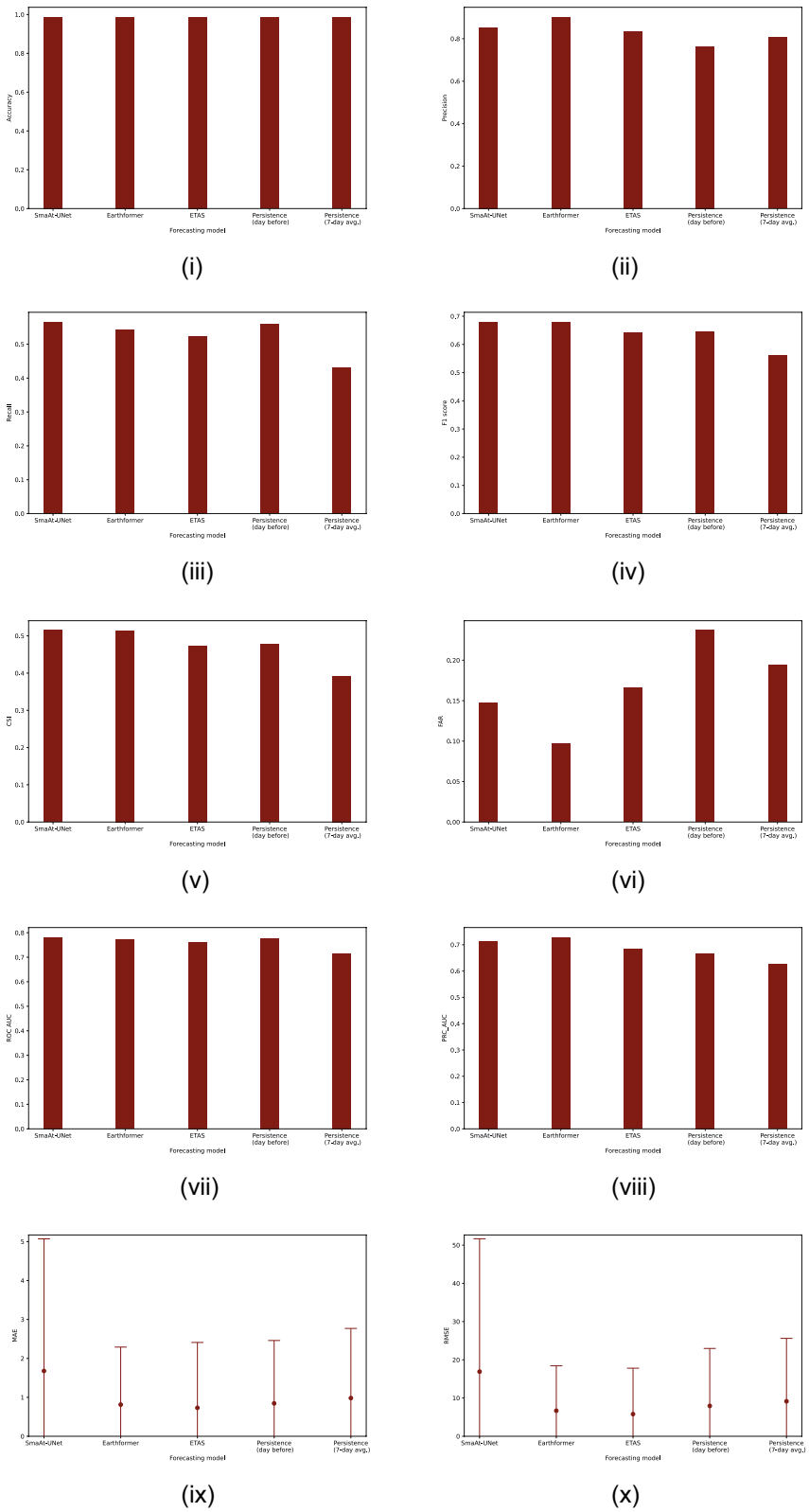


Fig. 16 Evaluation metrics on the Southern California and Italy test set. Comparison between SmaAt-UNet, Earthformer, ETAS, persistence (day before) and persistence (average of previous 7 days). BGS©UKRI 2025

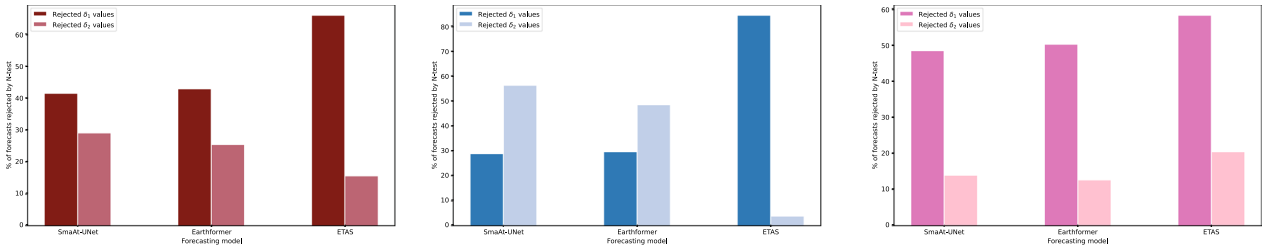


Fig. 17 Percentage of underpredictions (rejected δ_1 values) and overpredictions (rejected δ_2 values). Comparison between SmaAt-UNet, Earthformer and ETAS: (**a**) Southern California and Italy test set, (**b**) Southern California test set, (**c**) Italy test set. BGS©UKRI 2025

aftershock incompleteness in order to improve catalogue quality by detecting events that are currently missing.

5.5 Comparison between ML-based and ETAS forecasts

Figures 11 and 12 show a comparison between forecasts that are generated by ML models and those generated by ETAS for the 2019 Ridgecrest sequence. In Fig. 11, we see the ground truth maps in the first row, the SmaAt-UNet and Earthformer forecasts in the second and third row, the ETAS forecasts in the fourth row and the persistence forecasts in the fifth and sixth row. We see forecasts starting from 7 July 2019 up to 26 July 2019 and observe that these look reasonable in terms of spatial distribution of events. SmaAt-UNet forecasts have the highest MAE values, whereas ETAS forecasts have the lowest MAE values. We also see that in cases of high seismicity, such as this one, both the ML and ETAS models are not able to accurately forecast the expected rates, as evidenced by the N-test values. Similarly, Figs. 13 and 14 show ML and ETAS forecasts following events that have occurred within the 2016–2017 Central Apennines sequence, from 30 October 2016 to 3 November 2016. Here, we observe that the ML and ETAS models all have similar MAE values and that the forecast number of events are mostly not consistent with the observations, as shown by the N-test values. Nevertheless, the forecasts visually look reasonable. Figure 15 shows the ratios of ground truth to forecast maps for all the examples in the Southern California and Italy test sets. We observe that both the ML and the ETAS forecasts are generally close to the ground truth values, as indicated by the peak around 1. The distributions of the ratios for forecasts produced by Earthformer and ETAS have a smaller spread than the distribution of the ratios in the case of SmaAt-UNet forecasts, which suggests that SmaAt-UNet produces forecasts with more variability compared to Earthformer and ETAS. The ground truth to forecast ratio distributions of the persistence models also have a peak around 1 and look relatively similar to that of SmaAt-UNet, indicating that persistence models are also able to generate forecasts that are relatively close to the ground truth maps.

Tables 9 and 10 and Figs. 16, 17 and 18, 19 show a comparison between the two machine learning models and an ETAS benchmark in terms of performance for the Southern California and Italy test sets. The performance of both approaches is similar in terms of spatial distribution of events, with all three models achieving decent performance. The two ML models are slightly superior to ETAS based on the F1 score, CSI, FAR and PRC AUC, but the difference is relatively small. In terms of the number of forecast events, SmaAt-UNet has a larger MAE and RMSE compared to Earthformer and ETAS, with ETAS having slightly lower error scores than Earthformer. As seen in Table 10, the ML models and ETAS tend to underpredict the number of events. However, ETAS has a considerably higher number of underpredictions and also a lower number of overpredictions compared to the ML models, as evidenced by the percentages of rejected δ_1 and δ_2 values. This can be attributed to the fact that the ETAS model was trained on an incomplete dataset. The cumulative distributions of the δ_1 and δ_2 values in Fig. 20 show that, similarly to the two ML models, we have a case of underdispersion in the ETAS forecasts. The main advantage of ML models is the fact that they can be trained within a few hours and then, once trained, they can instantly generate as many forecasts as needed. ETAS, on the other hand, needs an inversion procedure to estimate the parameters, which one could argue is comparable to ML training in terms of computational cost, and then needs to perform a large number of simulations in order to generate each forecast. These simulations are computationally expensive and can take from a few minutes to several hours to complete. ML models can therefore be particularly effective for generating real-time or near real-time forecasts, as well as for applications where the forecasting model needs to be regularly updated based on new data.

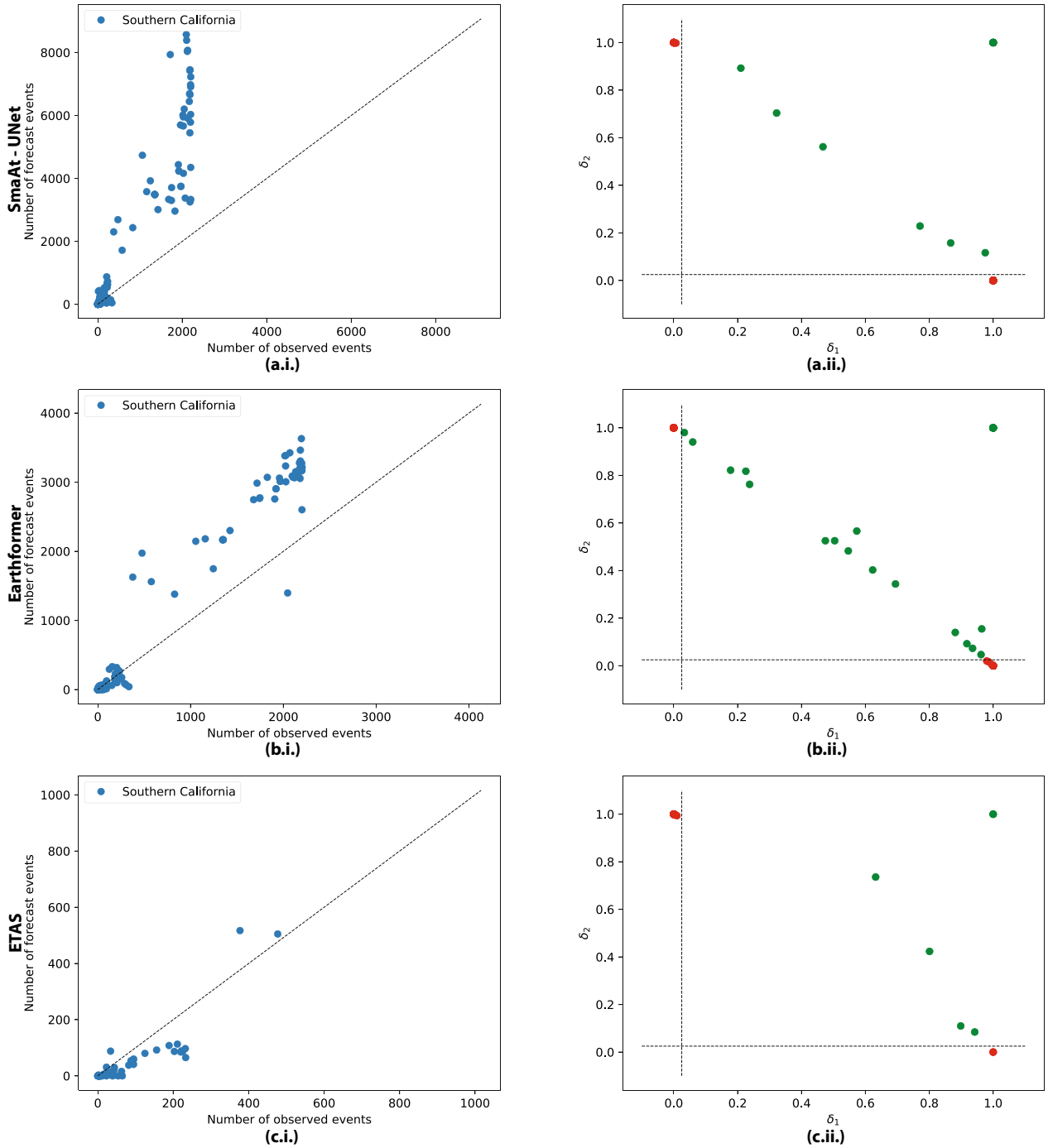


Fig. 18 Comparison between ML and ETAS forecasts. N-test results on the Southern California test set: underforecasting and overforecasting. (i) Observed versus forecast number of events. Points on or close to the $y = x$ line represent forecasts that are consistent with the observations. (ii) N-test δ_1 and δ_2 values. Forecasts with $\delta_1 > 0.025$ and $\delta_2 > 0.025$ are consistent with the observations. BGS©UKRI 2025

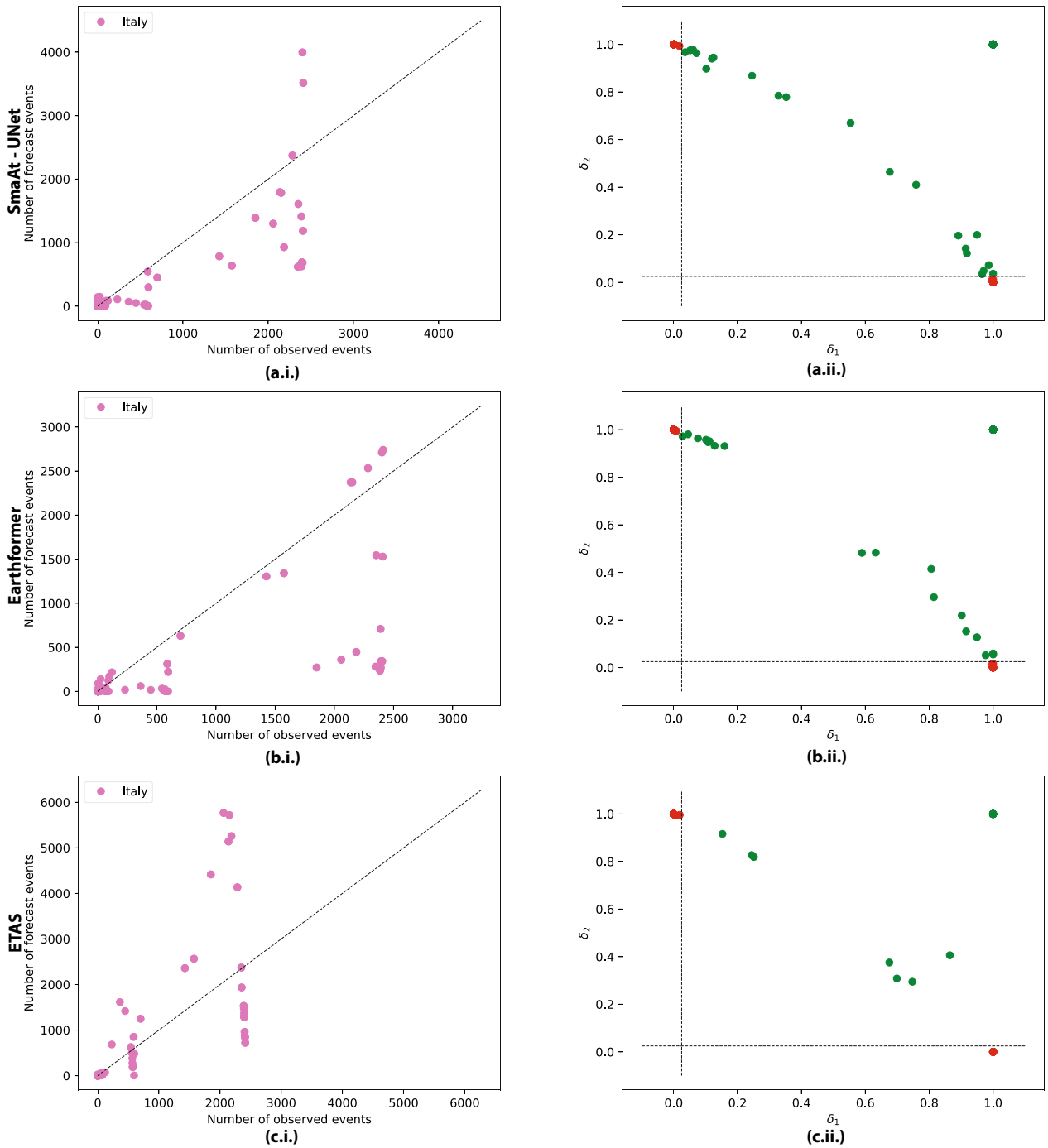


Fig. 19 Comparison between ML and ETAS forecasts. N-test results on the Italy test set: underforecasting and overforecasting. (i) Observed versus forecast number of events. Points on or close to the $y = x$ line represent forecasts that are consistent with the observations. (ii) N-test δ_1 and δ_2 values. Forecasts with $\delta_1 > 0.025$ and $\delta_2 > 0.025$ are consistent with the observations. BGS©UKRI 2025

5.6 Investigating the impact of individual catalogues and different data splitting strategies

In Tables 11 and 12, we see the evaluation metrics for the two ML models when those are trained and tested on randomly selected examples from the bulk catalogue dataset. These results are not directly comparable to those in Tables 5 and 7, as the test set used here is different. However, it is apparent that random splitting results in more similar data distributions between the training and test sets compared to a temporal split, leading

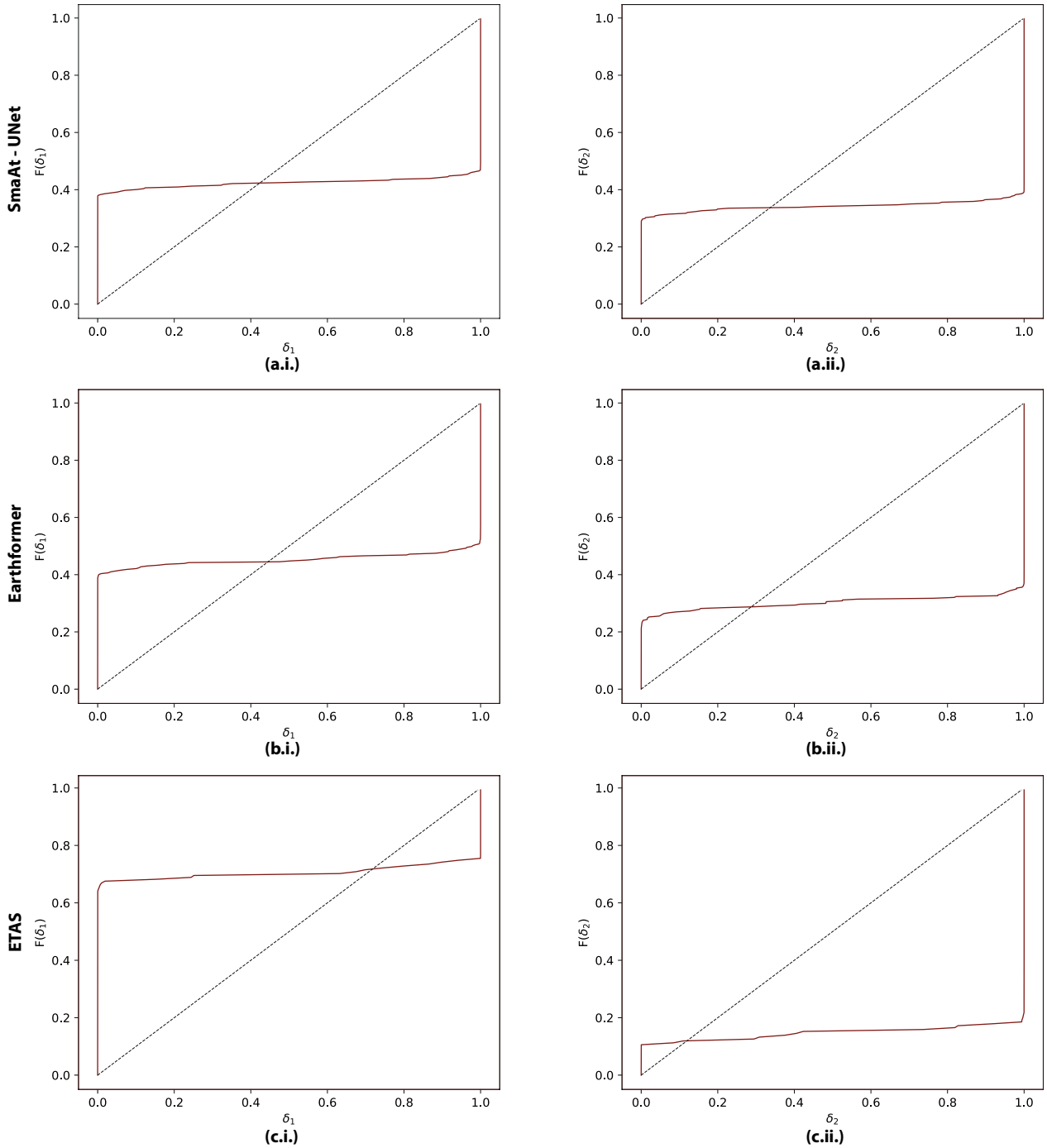


Fig. 20 Comparison between ML and ETAS forecasts. Cumulative distributions of δ_1 and δ_2 values (Southern California and Italy test set). The uniform distribution, which corresponds to a perfectly calibrated model, is shown with a dashed line. In our case, the cumulative distributions of the quantile scores show underdispersion, which means that the observed data have less variation than the forecasts (Savran et al 2020). BGS©UKRI 2025

to improved evaluation metric scores on the new randomly selected test set. This random splitting strategy introduces look-ahead bias (Peixeiro 2022), as it allows the model to see future data during training and then be tested on past data, which is not something that can be done in an operational scenario. Nevertheless, this highlights the importance of ensuring similarity between the data distributions used in training and inference settings

Table 9 Comparison of ML-based forecasts with ETAS forecasts: Classification and regression evaluation metrics (Southern California and Italy test set). BGS©UKRI 2025

Evaluation metric	SmaAt-UNet	Earthformer	ETAS	Persistence (day before)	Persistence (7-day avg.)
Accuracy ↑	0.987	0.988	0.988	0.985	0.984
Precision ↑	0.853	0.903	0.834	0.763	0.806
Recall ↑	0.566	0.544	0.524	0.559	0.433
F1 score ↑	0.680	0.679	0.644	0.646	0.563
CSI ↑	0.515	0.514	0.474	0.477	0.392
FAR ↓	0.147	0.097	0.166	0.237	0.194
ROC AUC ↑	0.782	0.772	0.761	0.778	0.715
PRC AUC ↑	0.714	0.729	0.684	0.667	0.626
MAE $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ ↓	1.678,3.393	0.813,1.478	0.731,1.679	0.845, 1.615	0.981,1.789
RMSE $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ ↓	16.903,34.751	6.687,11.744	5.800,12.007	7.946,15.027	9.151,16.457

Table 10 Comparison of ML-based forecasts with ETAS forecasts: Percentage of underpredictions (rejected δ_1 values) and overpredictions (rejected δ_2 values) (Southern California and Italy test set). BGS©UKRI 2025

		SmaAt-UNet	Earthformer	ETAS
All	Rejected $\boldsymbol{\delta_1}$	41.46%	42.86%	65.90%
	Rejected $\boldsymbol{\delta_2}$	29.13%	25.49%	15.61%
Southern California	Rejected $\boldsymbol{\delta_1}$	28.91%	29.69%	84.31%
	Rejected $\boldsymbol{\delta_2}$	56.25%	48.44%	3.92%
Italy	Rejected $\boldsymbol{\delta_1}$	48.47%	50.22%	58.20%
	Rejected $\boldsymbol{\delta_2}$	13.97%	12.66%	20.49%

for optimal ML model performance. Therefore, it is crucial to invest time and effort into creating training datasets that represent a wide range of scenarios that are likely to be encountered in an inference setting in order to train models that are capable of generalising well.

Table 13 shows the performance of SmaAt-UNet on the Southern California and New Zealand catalogues. Although the results are not directly comparable as different test sets have been used in each case, they offer valuable insights regarding the behaviour of ML forecasting models. We observe that the model that is trained on the Southern California catalogue is able to generate forecasts that are spatially consistent with the observations both in the sequential and the random data splitting scenario. When a random splitting strategy is used, the model is able to reproduce the number of events much better than when a sequential splitting strategy is used, as evidenced by the MAE and RMSE values as well as by the percentages of rejected δ_1 and δ_2 values. On the other hand, the model that is trained on the New Zealand catalogue is only able to produce forecasts that are consistent with the observations in terms of spatial distribution and number of events when a random data splitting strategy is used. This can be attributed to the fact that

Table 11 Training and fine-tuning CNNs and transformers using a random data split: Classification and regression evaluation metrics (standard catalogue data test set). For regression metrics (MAE and RMSE), the mean and the standard deviation of the errors are reported. BGS©UKRI 2025

Evaluation metric	SmaAt-UNet	SmaAt-UNet+	Earthformer	Earthformer+	Persistence (day before)	Persistence (7-day avg.)
Accuracy ↑	0.997	0.995	0.996	0.996	0.974	0.976
Precision ↑	0.953	0.785	0.931	0.844	0.522	0.373
Recall ↑	0.817	0.867	0.807	0.836	0.623	0.742
F1 score ↑	0.880	0.824	0.864	0.840	0.568	0.496
CSI ↑	0.786	0.701	0.761	0.725	0.397	0.330
FAR ↓	0.047	0.215	0.069	0.156	0.478	0.627
ROC AUC ↑	0.908	0.932	0.903	0.917	0.804	0.670
PRC AUC ↑	0.886	0.827	0.870	0.841	0.578	0.534
MAE $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ ↓	0.064,0.139	0.302,0.795	0.066,0.182	0.467, 1.400	0.229,0.543	0.256,0.594
RMSE $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ ↓	0.596,1.432	2.516,7.725	0.561,1.614	4.617,15.784	1.670,4.811	1.832,5.046

Table 12 Training and fine-tuning CNNs and transformers using a random data split: Percentage of underpredictions (rejected δ_1 values) and overpredictions (rejected δ_2 values) (bulk catalogue data test set). BGS©UKRI 2025

		SmaAt-UNet	SmaAt-UNet+	Earthformer	Earthformer+
All	Rejected $\boldsymbol{\delta_1}$	29.02%	14.31%	24.85%	11.96%
	Rejected $\boldsymbol{\delta_2}$	3.43%	32.06%	4.41%	33.19%
Southern California	Rejected $\boldsymbol{\delta_1}$	47.83%	23.91%	51.09%	17.39%
	Rejected $\boldsymbol{\delta_2}$	8.70%	53.26%	5.43%	55.43%
Northern California	Rejected $\boldsymbol{\delta_1}$	32.03%	17.75%	26.41%	12.55%
	Rejected $\boldsymbol{\delta_2}$	5.63%	32.47%	9.96%	34.63%
New Zealand	Rejected $\boldsymbol{\delta_1}$	29.83%	11.53%	23.50%	9.94%
	Rejected $\boldsymbol{\delta_2}$	2.49%	40.68%	2.71%	42.03%
Italy	Rejected $\boldsymbol{\delta_1}$	25.40%	23.81%	22.22%	17.46%
	Rejected $\boldsymbol{\delta_2}$	6.35%	12.70%	4.76%	15.87%
Greece	Rejected $\boldsymbol{\delta_1}$	19.28%	12.80%	16.55%	10.58%
	Rejected $\boldsymbol{\delta_2}$	2.05%	13.99%	3.41%	15.53%
Japan	Rejected $\boldsymbol{\delta_1}$	44.26%	20.22%	43.72%	20.77%
	Rejected $\boldsymbol{\delta_2}$	6.01%	43.72%	8.20%	39.89%

Table 13 Investigating the impact of individual catalogues and different data splitting strategies: Evaluation of SmaAt-UNet performance when trained and tested on the Southern California (SC) and New Zealand (NZ) catalogues separately. The use of a sequential and a random data splitting strategy is explored. BGS©UKRI 2025

Evaluation metric	SmaAt-UNet SC sequential	SmaAt-UNet SC random	SmaAt-UNet NZ sequential	SmaAt-UNet NZ random
Accuracy ↑	0.991	0.997	0.992	0.997
Precision ↑	0.858	0.962	0.873	0.978
Recall ↑	0.820	0.891	0.401	0.923
F1 score ↑	0.838	0.925	0.549	0.949
CSI ↑	0.722	0.860	0.379	0.904
FAR ↓	0.142	0.038	0.127	0.022
ROC AUC ↑	0.908	0.945	0.700	0.961
PRC AUC ↑	0.841	0.927	0.641	0.951
MAE $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ ↓	1.160, 1.406	0.152, 0.244	0.146, 0.232	0.049, 0.085
RMSE $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ ↓	11.626, 14.438	1.574, 2.882	1.062, 1.694	0.407, 0.743
Rejected $\boldsymbol{\delta_1}$ ↓	52.31%	37.40%	40.83%	15.82%
Rejected $\boldsymbol{\delta_2}$ ↓	20.77%	13.82%	2.14%	3.39%

the Southern California catalogue covers a considerably smaller area compared to the New Zealand catalogue and is able to better capture the tectonic landscape of the region. This means that the distribution of older data is consistent with the distribution of newer data, making it possible to train a ML model that behaves adequately in a pseudo-prospective scenario. However, the model that is trained on the New Zealand catalogue behaves better in terms of number of forecast events, which is likely due to the fact that the training dataset built from the New Zealand catalogue is considerably larger in size and hence enables the model to learn not just the spatial distribution of events, but also how earthquake sequences behave in terms of daily rate. Overall, this experiment leads to the conclusion that the use of different datasets leads to different levels of predictability.

5.7 Exploring how the use of different types of input maps influences performance

Table 14 shows that use of maximum magnitude and average depth maps in addition to rate maps in the model inputs improves the overall forecasting performance in terms of error metrics and spatial distribution of events. This could be an indication that ML models are able to

Table 14 Exploring how the use of different types of input maps influences performance: Evaluation of SmaAt-UNet performance when trained and tested on the bulk catalogue dataset using rate, rate+magnitude and rate+magnitude+depth maps. BGS©UKRI 2025

Evaluation metric	SmaAt-UNet (rate)	SmaAt-UNet (rate+mag)	SmaAt-UNet (rate+mag+depth)
Accuracy ↑	0.996	0.997	0.997
Precision ↑	0.927	0.946	0.956
Recall ↑	0.787	0.819	0.840
F1 score ↑	0.851	0.878	0.894
CSI ↑	0.740	0.782	0.808
FAR ↓	0.073	0.054	0.044
ROC AUC ↑	0.893	0.909	0.920
PRC AUC ↑	0.858	0.884	0.899
MAE $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ ↓	0.084, 0.219	0.070, 0.159	0.063, 0.137
RMSE $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ ↓	0.277, 0.711	0.235, 0.537	0.214, 0.474
Rejected $\boldsymbol{\delta_1}$ ↓	22.06%	24.92%	23.67%
Rejected $\boldsymbol{\delta_2}$ ↓	7.26%	5.48%	4.67%

reproduce empirical laws of earthquake sequences. For example, the fact that the introduction of magnitude information improves performance could imply that ML models learn about magnitude-related productivity and hence forecast more events for larger-magnitude mainshocks than for mainshocks of smaller magnitudes. This is a known property of earthquake sequences that is accurately captured by ETAS models as well.

6 Conclusions

This study introduced a data-driven ML-based short-term spatiotemporal seismicity rate forecasting approach based on earthquake catalogues from different earthquake-prone regions. Our findings show that both tested deep learning models, the SmaAt-UNet and the Earthformer, perform similarly. On the one hand, a more balanced performance is observed when employing the Earthformer architecture; however, on the other hand, SmaAt-UNet needs considerably less training time as it is a smaller model with fewer parameters, which is an important advantage for operational applications. Overall, both models demonstrate potential as their performance is superior to that of the persistence model, a commonly used baseline that assumes no change between consecutive time steps. The ML models achieve similar performance to that of an ETAS benchmark on the Southern California and Italy test sets and are able to generate forecasts at significantly reduced processing times compared to ETAS. Once trained, ML models can instantly generate forecasts, whereas ETAS requires significant computational power to perform the large number of simulations that make up each forecast.

Our qualitative and quantitative analysis of the generated forecasts shows that the spatial distribution of forecast events is consistent with that of observed events, as ML models generally forecast more events in locations close to the mainshock and fewer events further away. We also observe that the use of maximum magnitude maps as additional inputs to the models enhances performance, which is indicative of the fact that the ML models learn about magnitude-related productivity. These are both known properties of earthquake sequences that are efficiently captured by disciplinary state-of-the-art approaches, such as the ETAS model.

The introduction of earthquake sequences built from high-resolution catalogues to the training process has had a considerable impact on the results, leading to the conclusion that different datasets exhibit a different level of predictability. This could be due to differences in data quality, which often influences ML performance, especially in cases where the training dataset is fairly limited in size. We also notice that training ML models on catalogues from different geographic regions leads to considerable differences in the results, which could be either due to varying data quality or be an indication of inherent differences in the tectonic landscapes of different geographic regions.

This study shows that the use of ML models for the development of data-driven short-term seismicity forecasting approaches shows some promise, as their pattern recognition ability can be used to uncover relationships hidden within the wealth of information in earthquake catalogues. However, in order for ML models to be efficient and achieve generalisation, they need to be provided with a diverse and adequately sized high-quality training set. Standard catalogues, which are available for longer terms, have relatively large magnitudes of completeness and are thus missing information that would be valuable for training ML models. The wealth of information that is present in high-resolution catalogues could potentially prove useful for building better ML forecasting models. However, their limited availability currently prevents training ML models exclusively on high-resolution catalogues. To quantify the impact of such datasets on forecasting performance, we need access to long-duration and high-accuracy ML catalogues. This will potentially enable the development of forecasting models that are capable of understanding triggering patterns of spontaneous events linked to cascading sequences (Ellsworth and Bulut 2018). The advantages for operational environments are important: the real- or near real-time development of ML catalogues, coupled with

computationally economical ML forecast models (such as SmaAt-UNet), will lead to improved understanding and tracking of the evolution of seismic crises.

Abbreviations

Adam:	Adaptive moment estimation
AUC:	Area under curve
CBAM:	Convolutional block attention module
CIG:	Cumulative information gain
CNN:	Convolutional neural network
ConvLSTM:	Convolutional long short-term memory
CSEP:	Collaboratory for the Study of Earthquake Predictability
CSI:	Critical success index
DSC:	Depthwise separable convolution
EM:	Expectation maximisation
ETAS:	Epidemic-type aftershock sequence
FAR:	False alarm ratio
FN:	False negative
FP:	False positive
FPR:	False positive rate
IGPE:	Information gain per earthquake
MAE:	Mean absolute error
MaxPool:	Maximum pooling
ML:	Machine learning
MSE:	Mean squared error
N-test:	Number test
PRC:	Precision-recall curve
QTM:	Quake template matching
ReLU:	Rectified linear unit
RMSE:	Root mean squared error
RNN:	Recurrent neural network
ROC:	Receiver operating characteristic
SmaAt-UNet:	Small attention UNet
STAI:	Short-term aftershock incompleteness
TN:	True negative
TP:	True positive
TPR:	True positive rate

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40623-025-02241-6.

Supplementary file 1.

Acknowledgements
The authors would like to thank the editor Bogdan Enescu for the support, as well as Leila Mizrahi and an anonymous reviewer, whose comments and suggestions considerably improved the quality of this manuscript. The authors would also like to thank Giorgos Papanastasiou and Chengjia Wang for the valuable discussions.

Author contributions
FD and MS conceptualised the study and designed the methodology. FD wrote the software, conducted the formal analysis, investigation and validation process and prepared the visualisations. MS, PP, BB, IM and AC supervised the process. FD and MS wrote the initial draft. PP, BB, IM and AC reviewed the manuscript.

Funding
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 955515 – SPIN ITN (www.spin-itn.eu).

Data availability
The earthquake catalogues used in this study are publicly available online at the websites referenced in the "Data" section of the paper: Southern California (https://doi.org/10.7909/C3WD3xH1), Northern California (https://doi.org/10.7932/NCEDC), New Zealand (https://doi.org/10.21420/0S8P-TZ38), Italy (https://doi.org/10.13127/ISIDE), Greece (https://www.gein.noa.gr/en/services-products/earthquake-catalogs/), Japan (https://www.hinet.bosai.go.jp/topics/JUICE/), Southern California QTM (https://scedc.caltech.edu/data/qtm-catalog.html), Italy ML (http://dx.doi.org/10.5281/zenodo.4662869). The implementation of the deep learning models, the training and test processes and the evaluation workflows rely on the use of open-source software. The basic components that were used are the Python programming language (van Rossum and Drake 2011), the PyTorch deep learning framework (Paszke et al 2019), the NumPy (Harris et al 2020), scikit-learn (Pedregosa et al 2011), Matplotlib (Hunter 2007) and seaborn (Waskom 2021) libraries and the pyCSEP toolkit (Savran et al 2022b). Code from the following GitHub repositories has been used: https://github.com/HansBambel/SmaAt-UNet, https://github.com/amazon-science/earth-forecasting-transformer, https://github.com/lmizrahi/etas and https://github.com/SCECcode/pycsep.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]The Lyell Centre, British Geological Survey, EH14 4AP Edinburgh, United Kingdom. [2]School of GeoSciences, University of Edinburgh, EH9 3FE Edinburgh, United Kingdom. [3]Dipartimento di Geoscienze, Università degli Studi di Padova, 35131 Padua, Italy.

Received: 12 September 2024 Accepted: 18 June 2025
Published online: 25 November 2025

References

Armstrong JS (2001) Evaluating forecasting methods. Springer, Boston, pp 443–472. https://doi.org/10.1007/978-0-306-47630-3_20
Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473
Bento VA, Russo A, Dutra E et al (2022) Persistence versus dynamical seasonal forecasts of cereal crop yields. Sci Rep 12(1):7422. https://doi.org/10.1038/s41598-022-11228-2
Beroza GC, Segou M, Mostafa Mousavi S (2021) Machine learning and earthquake forecasting–next steps. Nat Commun 12(1):4761. https://doi.org/10.1038/s41467-021-24952-6
Brigato L, Iocchi L (2021) A Close Look at Deep Learning with Small Data. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp 2490–2497, https://doi.org/10.1109/ICPR48806.2021.9412492
Chollet F (2017) Xception: Deep Learning with Depthwise Separable Convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1800–1807, https://doi.org/10.1109/CVPR.2017.195
Chu Y, Pedro HT, Kaur A et al (2017) Net load forecasts for solar-integrated operational grid feeders. Solar Energy 158:236–246. https://doi.org/10.1016/j.solener.2017.09.052
Dascher-Cousineau K, Shchur O, Brodsky EE et al (2023) Using deep learning for flexible and scalable earthquake forecasting. Geophys Res Lett. https://doi.org/10.1029/2023GL103909
Dosovitskiy A (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
Ebert PA, Milne P (2022) Methodological and conceptual challenges in rare and severe event forecast verification. Natl Hazards Earth Syst Sci 22(2):539–557. https://doi.org/10.5194/nhess-22-539-2022
Ellsworth WL, Bulut F (2018) Nucleation of the 1999 izmit earthquake by a triggered cascade of foreshocks. Nat Geosci 11(7):531–535. https://doi.org/10.1038/s41561-018-0145-1
Gao Z, Shi X, Wang H, et al (2022) Earthformer: Exploring space-time transformers for earth system forecasting. In: NeurIPS 2022, https://www.amazon.science/publications/earthformer-exploring-space-time-transformers-for-earth-system-forecasting
Ghimire GR, Krajewski WF (2020) Exploring persistence in streamflow forecasting. JAWRA J Am Water Resour Assoc 56(3):542–550. https://doi.org/10.1111/1752-1688.12821

Science GNS (1970) New Zealand earthquake catalogue [Data set]. GNS Sci. https://doi.org/10.2142/0S8P-TZ38
Guo X, Yin Y, Dong C, et al (2008) On the Class Imbalance Problem. In: 2008 Fourth International Conference on Natural Computation, pp 192–201, https://doi.org/10.1109/ICNC.2008.871
Hardebeck JL, Llenos AL, Michael AJ et al (2024) Aftershock forecasting. Ann Rev Earth Planet Sci. https://doi.org/10.1146/annurev-earth-040522-102129
Harris CR, Millman KJ, van der Walt SJ et al (2020) Array programming with NumPy. Nature 585(7825):357–362. https://doi.org/10.1038/s41586-020-2649-2
Harte D (2017) Probability distribution of forecasts based on the ETAS model. Geophys J Int 210(1):90–104. https://doi.org/10.1093/gji/ggx146
Hewamalage H, Ackermann K, Bergmeir C (2023) Forecast evaluation for data scientists: common pitfalls and best practices. Data Min Knowl Discovery 37(2):788–832. https://doi.org/10.1007/s10618-022-00894-5
Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9(3):90–95. https://doi.org/10.1109/MCSE.2007.55
Hyndman R, Athanasopoulos G (2021) Forecasting: principles and practice (3rd edition). OTexts
ISIDe Working Group (2007) Italian seismological instrumental and parametric database (ISIDe) (Version 1). Istituto Nazionale di Geofisica e Vulcanologia (INGV). https://doi.org/10.1312/ISIDE
Japan Meteorological Agency (2024) The Seismological Bulletin of Japan. https://www.data.jma.go.jp/svd/eqev/data/bulletin/index_e.html
Jordan TH, Chen YT, Gasparini P et al (2011) Operational earthquake forecasting: state of knowledge and guidelines for utilization. Ann Geophys 54(4):315–391. https://doi.org/10.4401/ag-5350
Kagan YY, Jackson DD (1995) New seismic gap hypothesis: five years after. J Geophys Res: Solid Earth 100(B3):3943–3959. https://doi.org/10.1029/94JB03014
Kamath U, Graham K, Emara W (2022) Transformers for machine learning: a deep dive. Chapman and Hall/CRC, Boca Raton. https://doi.org/10.1201/9781003170082
Kamranzad F, Naylor M, Lindgren F et al (2025) Enhancing the ETAS model: incorporating rate-dependent incompleteness, constructing a representative dataset, and reducing bias in inversions. Geophys J Int. https://doi.org/10.1093/gji/ggaf156
Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
Knaff JA, Landsea CW (1997) An el niño-southern oscillation climatology and persistence (cliper) forecasting scheme. Weather and Forecast 12(3):633–652. https://doi.org/10.1175/1520-0434(1997)012
Koprinska I, Wu D, Wang Z (2018) Convolutional neural networks for energy time series forecasting. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp 1–8, https://doi.org/10.1109/IJCNN.2018.8489399
Kumar B, Haral H, Kalapureddy MCR et al (2024) Utilizing deep learning for near real-time rainfall forecasting based on radar data. Phys Chem Earth 135:103600. https://doi.org/10.1016/j.pce.2024.103600 (https://linkinghub.elsevier.com/retrieve/pii/S1474706524000585)
Lever J, Krzywinski M, Altman N (2016) Model selection and overfitting. Nat Methods 13(9):703–704. https://doi.org/10.1038/nmeth.3968
Mancini S, Segou M, Werner MJ et al (2022) On the use of high-resolution and deep-learning seismic catalogs for short-term earthquake forecasts: potential benefits and current limitations. J Geophys Res: Solid Earth. https://doi.org/10.1029/2022JB025202
Mittermaier MP (2008) The potential impact of using persistence as a reference forecast on perceived forecast skill. Weather Forecast 23(5):1022–1031. https://doi.org/10.1175/2008WAF2007037.1
Mizrahi L, Nandan S, Wiemer S (2021) The effect of declustering on the size distribution of mainshocks. Seismol Res Lett 92(4):2333–2342. https://doi.org/10.1785/0220200231
Mizrahi L, Nandan S, Wiemer S (2021) Embracing data incompleteness for better earthquake forecasting. J Geophys Res: Solid Earth. https://doi.org/10.1029/2021JB022379
Mizrahi L, Schmid N, Han M (2023). ETAS with fit visualization. https://doi.org/10.5281/zenodo.6583992
Mizrahi L, Dallo I, van der Elst NJ et al (2024) Developing, testing, and communicating earthquake forecasts: current practices and future directions. Rev Geophys. https://doi.org/10.1029/2023RG000823
Mousavi SM, Beroza GC (2022) Deep-learning seismology. Science 377(6607):eabm4470. https://doi.org/10.1126/science.abm4470
Mousavi SM, Beroza GC (2023) Machine learning in earthquake seismology. Ann Rev Earth Planet Sci 51:105–129. https://doi.org/10.1146/annurev-earth-071822-100323
NCEDC (2014) Northern California earthquake data center. Dataset, UC Berkeley Seismol Lab. https://doi.org/10.7932/NCEDC
NOAIG-CATALOGUE (2024) National Observatory of Athens Institute of Geodynamics Earthquake Catalogue. https://www.gein.noa.gr/en/services-products/earthquake-catalogs/
Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. J Am Stat Assoc 83(401):9–27
Omori F (1894) On the aftershocks of earthquakes. J College Sci 7:111–120
Owens MJ, Challen R, Methven J et al (2013) A 27 day persistence model of near-earth solar wind conditions: A long lead-time forecast and a benchmark for dynamical models. Space Weather 11(5):225–236. https://doi.org/10.1002/swe.20040
Paszke A, Gross S, Massa F et al (2019) PyTorch: an imperative style, high-performance deep learning library. Curran Associates Inc., Red Hook, p 12
Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
Peixeiro M (2022) Time Series Forecasting in Python. Manning
Pombo DV, Göçmen T, Das K, et al (2021) Multi-horizon data-driven wind power forecast: From nowcast to 2 days-ahead. In: 2021 International Conference on Smart Energy Systems and Technologies (SEST), pp 1–6, https://doi.org/10.1109/SEST50973.2021.9543173
Rainio O, Teuho J, Klén R (2024) Evaluation metrics and statistical tests for machine learning. Sci Rep 14(1):6086. https://doi.org/10.1038/s41598-024-56706-x
Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM et al (eds) Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. Springer International Publishing, Cham, pp 234–241
Ross ZE, Trugman DT, Hauksson E et al (2019) Searching for hidden earthquakes in Southern California. Science 364(6442):767–771. https://doi.org/10.1126/science.aaw6888
van Rossum G, Drake FL (2011) The Python Language Reference Manual. Network Theory Ltd
Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE 10(3):1–21. https://doi.org/10.1371/journal.pone.0118432
Savran WH, Werner MJ, Marzocchi W et al (2020) Pseudoprospective evaluation of UCERF3-ETAS forecasts during the 2019 ridgecrest sequence. Bull Seismol Soc Am 110(4):1799–1817. https://doi.org/10.1785/0120200026
Savran WH, Bayona JA, Iturrieta P et al (2022) pyCSEP: a python toolkit for earthquake forecast developers. Seismol Res Lett 93(5):2858–2870. https://doi.org/10.1785/0220220033
Savran WH, Werner MJ, Schorlemmer D et al (2022) pyCSEP: a python toolkit for earthquake forecast developers. J Open Sour Softw 7(69):3658. https://doi.org/10.2110/joss.03658
SCEDC (2013) Southern California earthquake center. Dataset, Caltech. https://doi.org/10.7909/C3WD3xH1
Schorlemmer D, Gerstenberger MC, Wiemer S et al (2007) Earthquake likelihood model testing. Seismol Res Lett 78(1):17–29. https://doi.org/10.1785/gssrl.78.1.17
Schorlemmer D, Werner MJ, Marzocchi W et al (2018) The collaboratory for the study of earthquake predictability: achievements and priorities. Seismol Res Lett 89(4):1305–1313. https://doi.org/10.1785/0220180053
Segou M (2020) The physics of earthquake forecasting. Seismol Res Lett 91(4):1936–1939. https://doi.org/10.1785/0220200127
Stevenson E, Rodriguez-Fernandez V, Minisci E et al (2022) A deep learning approach to solar radio flux forecasting. Acta Astronautica 193:595–606. https://doi.org/10.1016/j.actaastro.2021.08.004
Stockman S, Lawson DJ, Werner MJ (2023) Forecasting the 2016–2017 central apennines earthquake sequence with a neural point process. Earth's Future. https://doi.org/10.1029/2023EF003777
Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. MIT Press, Cambridge, MA, USA, NIPS'14, p 3104-3112

Tan YJ, Waldhauser F, Ellsworth WL et al (2021) Machine-learning-based high-resolution earthquake catalog reveals how complex fault structures were activated during the 2016–2017 central Italy sequence. The Seismic Record 1(1):11–19. https://doi.org/10.1785/0320210001
Trebing K, Stańczyk T, Mehrkanoon S (2021) SmaAt-UNet: precipitation nowcasting using a small attention-UNet architecture. Pattern Recogn Lett 145:178–186. https://doi.org/10.1016/j.patrec.2021.01.036
Tziolis G, Koumis A, Theocharides S, et al (2022) Advanced short-term net load forecasting for renewable-based microgrids. In: 2022 IEEE International Smart Cities Conference (ISC2), pp 1–6, https://doi.org/10.1109/ISC255366.2022.9922157
Utsu T (1961) A statistical study on the occurrence of aftershocks. Geophys Mag 30:521–605
Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS'17, p 6000-6010
Voyant C, Notton G (2018) Solar irradiation nowcasting by stochastic persistence: a new parsimonious, simple and efficient forecasting tool. Renew Sustain Energy Rev 92:343–352. https://doi.org/10.1016/j.rser.2018.04.116
Waskom ML (2021) seaborn: statistical data visualization. J Open Sour Softw 6(60):3021. https://doi.org/10.2110/joss.03021
Woo S, Park J, Lee JY et al (2018) CBAM: convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C et al (eds) Computer vision - ECCV 2018. Springer International Publishing, Cham, pp 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
Yano TE, Takeda T, Matsubara M et al (2017) Japan unified hIgh-resolution relocated catalog for earthquakes (JUICE): crustal seismicity beneath the Japanese Islands. Tectonophysics 702:19–28. https://doi.org/10.1016/j.tecto.2017.02.017
Yu M, Huang Q, Li Z (2024) Deep learning for spatiotemporal forecasting in Earth system science: a review. Int J Digit Earth 17(1):2391952
Zechar JD, Gerstenberger MC, Rhoades DA (2010) Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts. Bull Seismol Soc Am 100(3):1184–1195. https://doi.org/10.1785/0120090192
Zechar JD, Schorlemmer D, Liukis M et al (2010) The collaboratory for the study of earthquake predictability perspective on computational earthquake science. Concurr Comput: Pract Exp 22(12):1836–1847. https://doi.org/10.1002/cpe.1519
Zhang Z, Wang Y (2023) A spatiotemporal model for global earthquake prediction based on convolutional LSTM. IEEE Trans Geosci Remote Sens 61:1–12. https://doi.org/10.1109/TGRS.2023.3302316
Zlydenko O, Elidan G, Hassidim A et al (2023) A neural encoder for earthquake rate forecasting. Sci Rep 13(1):12350. https://doi.org/10.1038/s41598-023-38033-9

Publisher's Note