Computers & Geosciences 197 (2025) 105877

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

Analysis of earthquake detection using deep learning: Evaluating reliability and uncertainty in prediction methods

Sebastián Gamboa-Chacón [a,b,*], Esteban Meneses [a,b], Esteban J. Chaves [c]

[a] Costa Rica Institute of Technology (ITCR), Costa Rica
[b] National High Technology Center (CENAT), Costa Rica
[c] Volcanological and Seismological Observatory of Costa Rica (OVSICORI), National University (UNA), Costa Rica

ARTICLE INFO

Dataset link: https://github.com/SebasGamboa10/Reproducibility-and-Uncertainty-Assessment-in-EQTransformer.git

Keywords:
AI earthquake detection
Deep learning
EQTransformer
Reproducibility
Determinism

ABSTRACT

This study evaluates the performance and reliability of earthquake detection using the EQTransformer, a novel deep learning program that is widely used in seismological observatories and research for enhancing earthquake catalogs. We test the EQTransformer capabilities and uncertainties using seismic data from the Volcanological and Seismological Observatory of Costa Rica and compare two detection options: the simplified method (MseedPredictor) and the complex method (Predictor), the latter incorporating Monte Carlo Dropout, to assess their reproducibility and uncertainty in identifying seismic events. Our analysis focuses on 24 h-duration data that began on February 18, 2023, following a magnitude 5.5 mainshock. Notably, we observed that sequential experiments with identical data and parametrization yield different detections and a varying number of events as a function of time. The results demonstrate that the complex method, which leverages iterative dropout, consistently yields more reproducible and reliable detections than the simplified method, which shows greater variability and is more prone to false positives. This study highlights the critical importance of method selection in deep learning models for seismic event detection, emphasizing the need for rigorous evaluation of detection algorithms to ensure accurate and consistent earthquake catalogs and interpretations. Our findings provide valuable insights for the application of AI tools in seismology, particularly in enhancing the precision and reliability of seismic monitoring efforts.

1. Introduction

Technological advancements in conjunction with theoretical frameworks have revolutionized our understanding of the Earth interior and our ability to interact with it. In seismology, for instance, observatories all over the world, have exponentially increased the number of ultra-sensitive broadband instruments, fiber-optics, nodal arrays and the computational power for archiving and processing data with the aim of improving earthquake detection capabilities, specifically of smaller magnitude ($0 \leq M \leq 3.0$) events that occur along fault segments and may precede large and catastrophic ruptures (Spassiani and Sebastiani, 2016). The systematic increase in data since early and middle 2000s, when the digital era began for most seismological networks (Arrowsmith et al., 2022), have provided researchers with abundant information about the internal structure of the Earth, more complete earthquake catalogs and high quality recordings that allow to better understand fault mechanics and earthquake rupture dynamics.

However, this revolution comes at a cost. The total number of tebibytes of seismological data continues to increase in real time. As a result, traditional methods for earthquake detection and location, which are led by human intervention, are no longer sufficient. These methods struggle to fully capture the number of events generated during an earthquake sequence, especially the smaller magnitude earthquakes. These smaller events are generally obscured by ambient seismic noise, both having comparable frequency spectra and amplitude. Machine learning algorithms and artificial intelligence (AI) have significantly enhanced the ability of seismological observatories to detect and estimate earthquake hypocenter locations and magnitudes (Gürsoy et al., 2023). All these efforts have been potentiated by high-performance computing (HPC), enabling the scientific institutions to handle resource-intensive tasks, reducing execution times, thereby expediting scientific studies, interpretations and hazard assessments (Hassan et al., 2020).

All these advancements in science, specifically in seismology, have a direct impact on populations. Historically, hundreds of earthquakes around the world have caused significant loss of life and destruction, earning their classification as natural disasters. Costa Rica, known

* Correspondence to: National High Technology Center (CeNAT), 1.3 km north of the United States Embassy, San José, 10109, Costa Rica.
E-mail addresses: sgamboa@cenat.ac.cr, segamboachacon@gmail.com (S. Gamboa-Chacón).

https://doi.org/10.1016/j.cageo.2025.105877
Received 15 September 2024; Received in revised form 25 January 2025; Accepted 25 January 2025
Available online 2 February 2025

for its high seismic activity, has experienced numerous tragic events throughout its history. One of the most devastating was the 1910 earthquake in Cartago (UCR, 2015), which resulted in widespread destruction and loss of life. Similarly, other major earthquakes, such as the 1960 Valdivia earthquake in Chile and the 1964 Great Alaska earthquake, have left a lasting mark on global history, with an extensive list of such events highlighting the persistent threat of seismic activity.

An essential goal of science is to positively impact society, and seismology is no exception. With the rise of artificial intelligence, efforts to mitigate the impact of earthquakes on human lives have significantly increased. Around the world, studies have been conducted to improve early detection systems, enhance disaster response strategies, and ultimately reduce the risks associated with seismic events. One example of this kind of research (Jena et al., 2020), is focuses on using clustering analysis, convolutional neural networks (CNNs), and analytical hierarchy process (AHP) techniques to estimate earthquake risk and develop hazard maps for the Palu region. These advancements aim to safeguard lives by enabling more accurate predictions and timely interventions, reinforcing the critical role of science and technology in public safety.

Among the innovative algorithms that have been developed, EQTransformer (Mousavi et al., 2020) (hereafter referred to as EQT), a deep learning-based model, was designed to detect, phase-pick, and associate earthquakes from continuous seismic data. EQT leverages the power of deep learning to analyze seismic signals, offering an efficient and automated solution for earthquake detection. The EQT neural network has a multi-task structure with a deep encoder and three separate decoders.

The encoder employs 1D convolutions to extract spatial features from seismic waveforms and passes them to bidirectional Long Short-Term Memory (LSTM) layers, which capture temporal dependencies in both forward and backward directions. This is complemented by unidirectional LSTM layers, which ensure sequential processing for downstream tasks. Additionally, Network-in-Network (NiN) structures are utilized within the encoder to enhance the extraction of local, fine-grained features, while residual connections ensure stability during the training process by alleviating the vanishing gradient problem.

The self-attentive layers and transformers further refine the encoded representations, allowing the model to focus on the most relevant portions of the seismic signal. The decoders then process these high-level representations to produce confidence sequences for detecting earthquake events and identifying P and S phase arrivals. Each decoder specializes in a specific task: one detects earthquakes, while the others pinpoint P and S phase arrivals, ensuring a comprehensive multi-task approach.

This architectural design not only maximizes detection accuracy but also enables robust performance across a variety of seismic scenarios (Mousavi et al., 2020).

One of the novel features of EQT is its ability to provide uncertainties for the detection confidences, making the results more reliable. These uncertainties are approximated using a Gaussian distribution obtained through Monte Carlo Dropout. In 2016, this method was proposed (Gal and Ghahramani, 2016), which reinterprets dropout in deep neural networks as approximate Bayesian inference in deep Gaussian processes, enabling model uncertainty estimation without the computational expense of traditional Bayesian methods. This approach involves applying dropout during both training and inference, performing multiple forward passes to approximate the predictive distribution, and leveraging the variability in these predictions to gauge uncertainty. This method maintains computational efficiency and enhances test accuracy.

For earthquake detection and phase-picking, EQT provides two primary execution methods: a high-level method, referred to in the source code as Predictor (hereafter referred to as complex), which allows the configuration of multiple parameters for robust execution, and a low-level method, referred to in the source code as MseedPredictor (hereafter referred to as simplified), designed for basic execution with fewer adjustable options. Several studies (Jiang et al., 2021; Pita-Sllim et al., 2023) have shown promising results when using EQT, enhancing earthquake catalogs and providing a robust of seismotectonic characterization across different regions. Furthermore, several efforts (van der Laat et al., 2021; Castillo et al., 2024) that incorporate EQT methods have been developed aiming to generate automatic pipelines for daily seismological routines.

Nevertheless, little to none attention to EQT detection uncertainties and intricacies between the simplified and complex method have been investigated yet. Understanding the differences in performance and behavior between these two methods is essential for optimizing the use of EQT in various applications but also to generate realistic interpretations in seismological studies. This work aims to analyze, quantify and describe uncertainties in earthquake detection by EQTransformer. Reproducibility is a crucial aspect in scientific research, as accurate and consistent results are essential for researchers studying and analyzing critical characteristics of earthquakes and their uncertainties. Reproducibility is closely tied to deterministic outcomes, where consistent results are expected for identical experiments, identical data or algorithm runs. However, our observations clearly show variability in earthquake detection as a function of time when performing different executions of EQT while maintaining equal input variables, data and computer architectures. We aim to understand the factors contributing to this non-determinism and quantify its impact on the accuracy and reliability of EQT performance.

We analyzed the behavior of EQT focusing on the differences between the simplified and complex execution methods, particularly, the non-systematic earthquake detection effects introduced by the Monte Carlo Dropout. Given the complex nature of deep learning models, it is crucial to assess whether their execution is deterministic, that is, whether identical conditions yield consistent results in repeated runs. To achieve these objectives, we conducted a series of experiments comparing the outputs of EQT using both methods under varying computational setups. By systematically evaluating the results, we identified variations directly linked to the performance and nature of both algorithms. Not only does this analysis contribute to a deeper understanding of EQT's functionality and uncertainty, but also provides insights into the broader implications of using deep learning models for enhancing seismological catalogs.

2. Background

Costa Rica is part of the Central America volcanic front, where four tectonic plates (the Cocos plate, the Caribbean plate, the Panama microplate, and the Nazca plate) interact along the Middle America Trench (Protti et al., 1994; Montero et al., 1998). The local stress field, induced by this complex geodynamic system into the country, is translated into hundreds of very active tectonic faults with different length, geometry and seismic potential (Montero et al., 1998; Styron et al., 2020). The Volcanological and Seismological Observatory of Costa Rica (OVSICORI) at Universidad Nacional operates the largest and most modern geodynamic network in Central America and the Caribbean, composed by more than 200 instruments between broadband seismic stations, accelerometers, GNSS and multi-gas, for the permanent monitoring of the tectonic and volcanic activity in the country, generating alerts and official communications with governmental institutions and the general public.

In 2021, OVSICORI teamed up with the Costa Rica National High Technology Center (CeNAT) to develop a novel pipeline, known as the OKSP pipeline, for identifying and locating earthquakes from waveforms recorded by seismological stations across the country (van der Laat et al., 2021). Fig. 1 summarizes the multiple steps carried out by this pipeline, which incorporates the EQT algorithm as a core component. The OKSP pipeline begins with the collection of seismic

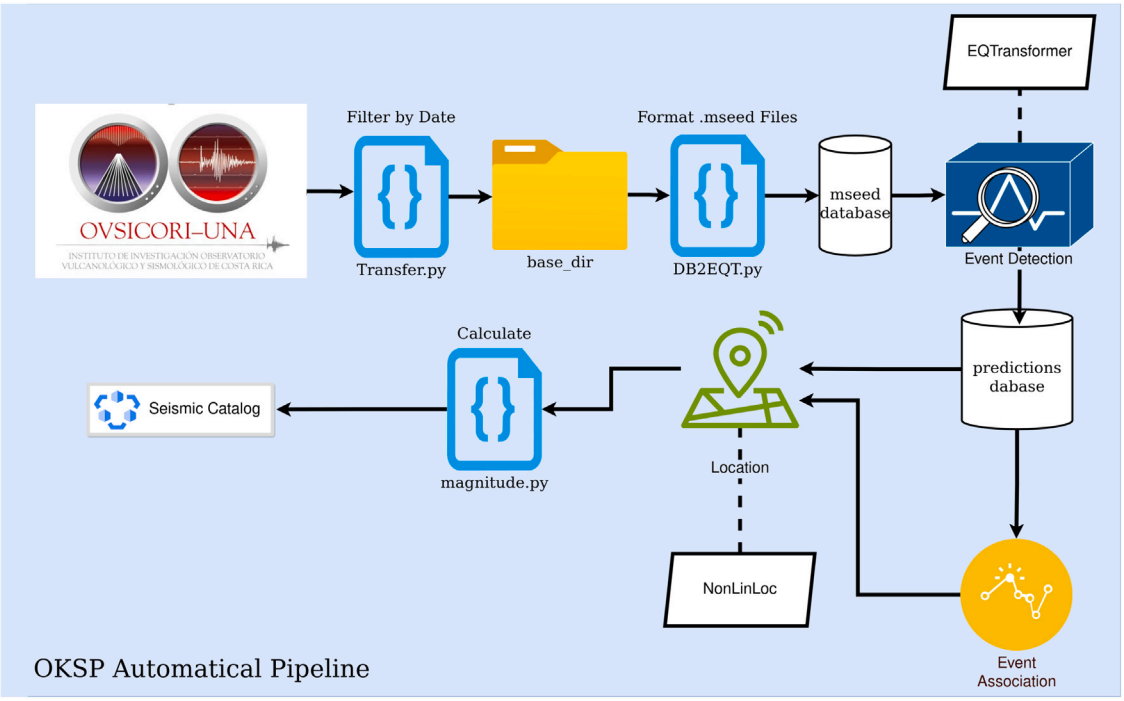

Fig. 1. OKSP pipeline. A schematic representation of the earthquake detection and phase identification process at the Costa Rica High Technology Center (CeNAT). This system utilizes three-component seismic data from OVSICORI-UNA to automatically generate a seismic catalog.

Table 1
Classification Metrics.

Metric	Result
Precision	0.8214
Recall	1.0000
F1-Score	0.9020

data, followed by the identification of P and S wave arrivals. Subsequently, the pipeline associates the detected phases, determines the event locations, calculates magnitudes, and ultimately provides an interactive map displaying the located events alongside a comprehensive seismological catalog.

While our study does not directly employ the OKSP pipeline, it emphasizes the importance of the EQT algorithm, as its performance is critical for accurate earthquake detection within the pipeline. In the study which involves the use of the OKSP pipeline, the detection capabilities of EQT were evaluated by analyzing the classification metrics Precision, Recall, and F1 Score. These metrics were derived from a large aftershock sequence recorded over five days at multiple stations in southern Costa Rica. The results, summarized in Table 1, were compared to traditional detection methods developed by OVSICORI (van der Laat et al., 2021). It is important to highlight that the OKSP pipeline employs the simplified EQT method.

Precision, with a value of 0.8214, indicates that 82.14% of the events detected by the EQT model were true positives, meaning actual earthquakes. This suggests that there is a 17.86% rate of false positives, where non-seismic events were incorrectly identified as earthquakes. Recall is perfect at 1.0000, signifying that the EQT model successfully detected all actual earthquake events that occurred during the period of study. The absence of false negatives is crucial for comprehensive seismic monitoring, ensuring no real events were missed.

The F1 Score, calculated as the harmonic mean of Precision and Recall, stands at 0.9020. This high F1 Score reflects a balanced performance of the EQT model, effectively combining both precision and completeness in earthquake detection. These metrics underscore the effectiveness of the EQT model in expanding the OVSICORI earthquake catalog. By setting the appropriate confidence threshold, it is possible to ensure high detection accuracy and completeness. The 85% confidence threshold was chosen as it strikes a balance between reducing false positives and maintaining a high signal-to-noise ratio, which is crucial for analyzing low-magnitude events (van der Laat et al., 2021).

There are algorithms similar to EQTransformer or based on it, such as EQCCT (M. Saad et al., 2023). This algorithm has demonstrated better results than EQT in terms of predictions and event detections. Using Japanese test data (M. Saad et al., 2023), EQCCT showed characteristics that ensure two consecutive runs on the same data yield consistent detections for the same events. However, EQT remains a central focus in research, as it is one of the most popular and widely available tools, serving as the foundation for the development of several new tools.

3. Methodology

We expanded on earlier work (van der Laat et al., 2021) by evaluating the uncertainties and consistency in earthquake detection carried out by EQT during two consecutive executions with the same parametrization and dataset. This task was performed for each detection method in EQT: the complex method (Predictor) and the simplified method (MseedPredictor). The seismic records from 5 stations operated by OVSICORI in the region surrounding the Poás Volcano in central Costa Rica were used. It is important to remember that the experimentation conducted for this study did not involve the use of the OKSP pipeline, but understanding the behavior of EQT is fundamental as this tool is an essential part of the pipeline. Since the objective was to reproduce the performance of both detection methods at each recording site, a total of 4 executions per seismic station were generated: 2 for the complex method and 2 for the simplified method.

For each station, we selected 24-h of data following the occurrence of the Magnitude 5.5 mainshock and part of the aftershock sequence that occurred on February 18, 2023, along the Norteastern flank of the Poás Volcano, near the town of Cinchona, Alajuela. This sequence is shown as green circles in Fig. 2, where the size of the circles represents earthquake magnitude and triangles correspond with the spatial distribution of broadband seismic stations around the study area. This

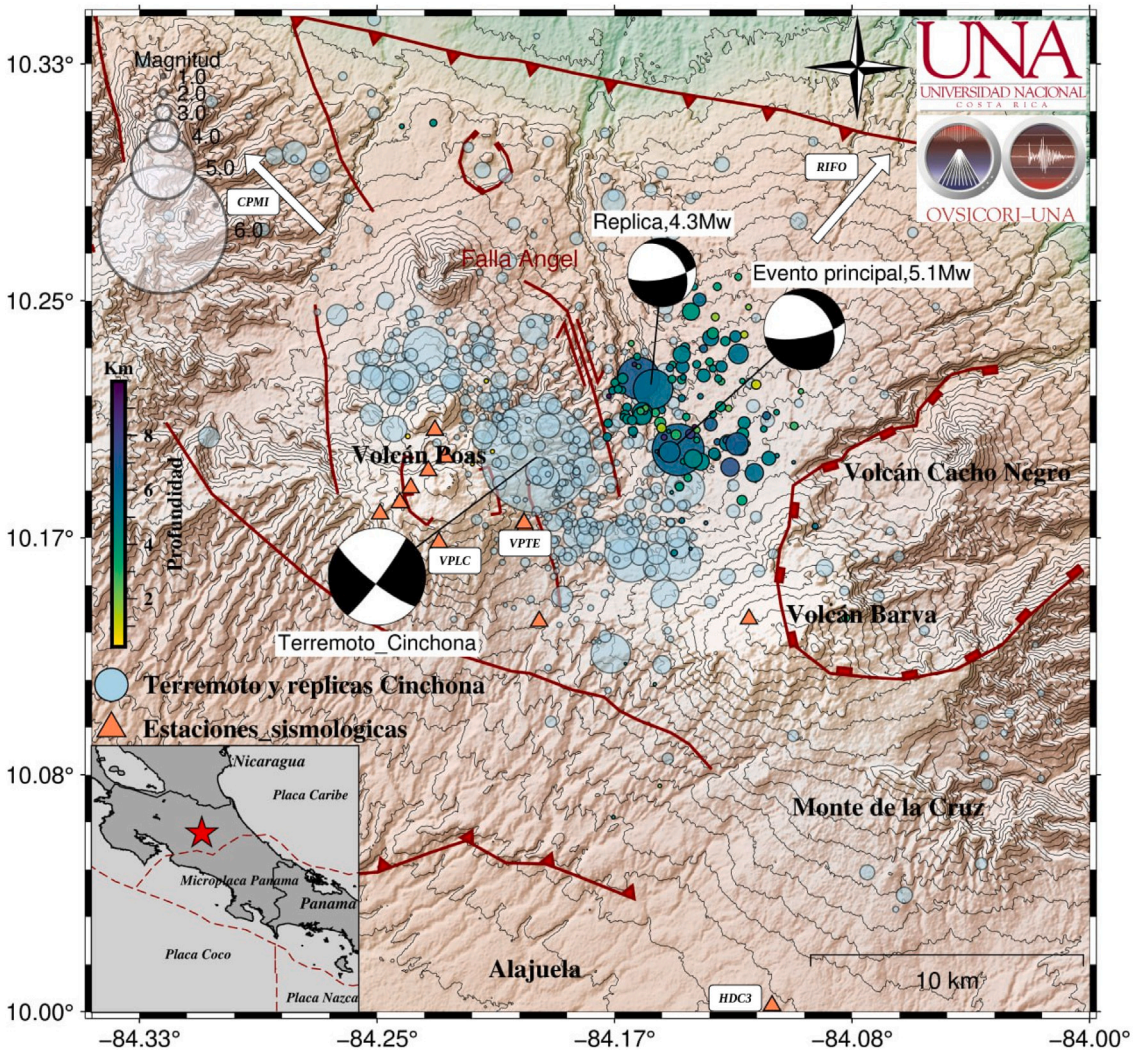

Fig. 2. Map of the study area. The map is showing the spatial distribution of the earthquake sequence generated by the February 18, 2023, M5.5 mainshock and its aftershocks (green circles) and its proximity to the January 8, 2009, M6.2 Cinchona earthquake sequence (light blue). In this figure, the size of the circles represent the magnitude of the earthquakes, while triangles correspond with the stations used for analyzing the seismic data in this study.

earthquake sequence is aligned parallel to the January 2009, M6.2 Cinchona earthquake sequence (shown as light blue circles), one of the most devastating events in the history of Costa Rica. During this event, multiple earthquake-triggered landslides caused the loss of 25 lives, left 17 people missing, and resulted in significant damage to public and private infrastructure (Instituto Costarricense de Electricidad and Universidad de Costa Rica, 2009), including hydroelectric dams of the Costa Rican Institute of Electricity (ICE), such as the Toro II and Cariblanco, which were partially affected. Therefore, characterizing the 2023 sequence is necessary for a better understanding of the seismotectonics and earthquake potential in the region.

3.1. Computer architectures and EQT detection functions

For analyzing the data, we initially considered four different computational architectures to explore the performance of our methods. These included three GPUs and one CPU, all detailed in Table 2. After evaluating the GPU performance specifications listed in the table, we decided to concentrate exclusively on the NVIDIA V100 GPU. This choice was driven by the V100's superior performance across several key metrics, including processing speed, memory capacity, memory bandwidth, and overall efficiency, making it the most suitable option for our analyses. By selecting the best-performing architecture, we aim to ensure that our results are both robust and consistent, minimizing any variability that could arise from using less capable hardware.

Recent advancements in earthquake detection models, such as the ECPickNet (Wang et al., 2023), emphasize the importance of selecting optimal computational architectures and parameter configurations to enhance performance, particularly in low signal-to-noise ratio environments. Their model integrates advanced architectures, including Convolution-Enhanced Transformers, to improve detection accuracy. While their work demonstrates significant improvements in processing efficiency and accuracy using advanced deep learning techniques, our study complements this by focusing on the practical implementation of EQTransformer under varying constraints. By leveraging the NVIDIA V100 GPU and exploring the impact of different execution methods, we provide a detailed analysis of the computational trade-offs and parameter settings that can influence the performance of EQT in diverse scenarios. This comparative perspective highlights the shared goal of optimizing earthquake detection workflows and underscores the broader applicability of our findings within the seismic detection community.

After selecting the optimal hardware, the focus shifted to two primary earthquake detection methods. The simplified execution method processes MiniSeed files from each station and runs a single pass without providing uncertainty estimates for the P and S phases or

Table 2
Specifications of the evaluated computational architectures.

Architecture	*Core clock speed*	*Main memory size*	*Memory clock speed*	*Memory bandwidth*	*Power consumption (TDP)*
NVIDIA TESLA V100 PCIe	1246 MHz	32 GB	1758 MHz	900.1 GB/s	250 Watt
NVIDIA TESLA K40 PCIe	745 MHz	12 GB	3004 MHz	288.4 GB/s	245 Watt
NVIDIA TESLA P6	1012 MHz	16 GB	6008 MHz	192.2 GB/s	90 Watt
CPU Intel Xeon Silver 4214R	2.40 GHz (24 cores)	128 GB	2933 MHz	107.3 GB/s	100 Watt

Table 3
Main configuration parameters for complex and simplified execution methods.

Parameter	Complex	Simplified
Estimate uncertainty	True	N/A
Number of Monte Carlo sampling	50	N/A
Detection threshold	0.85	0.85
P threshold	0.9	0.9
S threshold	0.7	0.7
Use multiprocessing	True	N/A
gpuid	0	0
gpu limit	None	None

earthquake detection predictions. This approach is suitable for larger datasets as it is more memory-efficient, bypassing the pre-processing step and working directly with the downloaded MiniSeed files.

In contrast, the complex execution method offers more detailed and customizable options. Although more demanding to implement, it allows for performance testing and the exploration of various parameter settings. This method requires pre-processed data and is better suited for smaller datasets, typically covering a period of a few days to a month. The pre-processing steps involve several crucial stages to prepare the seismic data for effective analysis. Raw seismic data from each station is filtered to remove noise and irrelevant frequencies, this includes applying band-pass filters to isolate the relevant frequency ranges for earthquake event detection. Furthermore, the complex method supports lower threshold values for detection and picking, leveraging EQTransformer's strong resistance to false positives.

Additionally, the complex method provides uncertainty estimates for P and S phases and earthquake detection probabilities through the use of Monte Carlo Dropout. This technique, (Gal and Ghahramani, 2016), enables the approximation of Bayesian inference over the network's weights. Dropout, a regularization method typically used during training to prevent overfitting, is applied at test time to impose a Bernoulli distribution over the network's weights. By performing multiple forward passes with dropout enabled, the method samples from the posterior distribution over models, which can be interpreted as generating a set of predictions. The variability among them reflects the model's uncertainty. The greater the dispersion between the predictions, the higher the uncertainty, providing a measure of confidence in the model's results.

Considering the existence of these two distinct methods, it becomes imperative to ensure uniform configuration for each execution. In Table 3 we provide a summary of the main configuration parameters used for both methods.

The parameters shown in Table 3 represent key configurations that can be adjusted in the EQT tool for each execution method. These settings enable users to tailor the tool's behavior to meet specific requirements and accommodate varying computational resources. By offering this flexibility, the EQT tool can be effectively adapted to diverse datasets and computational environments, thereby enhancing its usability and efficiency. Annex provides a detailed description of the primary parameters, refer to the repository at GitHub.

In this study, we utilized the original pre-trained model of EQT, which was developed using the Stanford Earthquake Dataset (STEAD). STEAD is a comprehensive, high-quality dataset specifically curated for seismic applications, comprising labeled seismic waveform data from diverse tectonic settings and geographic regions around the world. The pre-training process of EQT involved leveraging this dataset to fine-tune the model's ability to detect and classify seismic phases. By training on STEAD, EQT captures a wide range of seismic patterns, ensuring robust generalization across different seismic environments. Our decision to employ the original pre-trained model for both execution methods was motivated by the need to maintain consistency and standardization in our analyses. This approach eliminates the variability that could arise from retraining the model on local datasets, allowing us to focus on evaluating the performance and reproducibility of the detection methods under consistent conditions.

As the complex method incorporates Monte Carlo Dropout, we defined 50 runs. The runs refer to the times the model is executed with dropout enabled in inference mode to generate multiple predictions on the same input. EQT implement a dropout after every layer of the neural network and use it during both training and prediction (Mousavi et al., 2020). The confidence scores of each iteration are then averaged in order to get a final score. This number of runs in our case is determined by evaluating the percentage of matching events between experiments. This approach is analogous to the Elbow method in clustering analysis, where the optimal number of clusters is identified by finding the point where the reduction in the sum of squared errors (SSE) slows significantly (Humaira and Rasyidah, 2020). Similarly, in our case, we identified the point where increasing the number of runs leads to diminishing improvements in the percentage of matching events. The number of runs should be determined based on the specific analysis being conducted, as it can depend on factors such as the model employed, threshold settings, and the region. Our value is not definitive or generalized, and its suitability may vary for different datasets or experimental conditions. Fig. 3 shows this relationship, illustrating that with 50 runs, we achieved over 90% matching accuracy. Beyond this point, additional runs yielded progressively smaller gains, mirroring the behavior observed in the Elbow method when the SSE reduction begins to taper off.

For each station, we analyzed the number of events detected as a function of time for the two equal and consecutive experiments. This allowed us to track down possible errors or variations in earthquake detection per site. Furthermore, we compared the number of events per hour for each station across the two experiments. This comparison helped to identify any specific h during which differences occur, providing insights into the possible sources of discrepancy.

Finally, for each detection method, we compared the detection results from each experiment at each recording site, by applying a match filter algorithm to the detected origin time of the events, allowing a lag time of about ± 10 s and ensuring that all detections were performed on the same station channel (East, North or Vertical). This comprehensive analysis allows us to understand the functionality and better interpret the results from EQT.

4. Results and discussion

As previously introduced, we selected the Northeastern flank of the Poás Volcano, near the town of Cinchona, Alajuela, Costa Rica, to evaluate the performance of EQT for detecting earthquakes, as shown in Fig. 2.

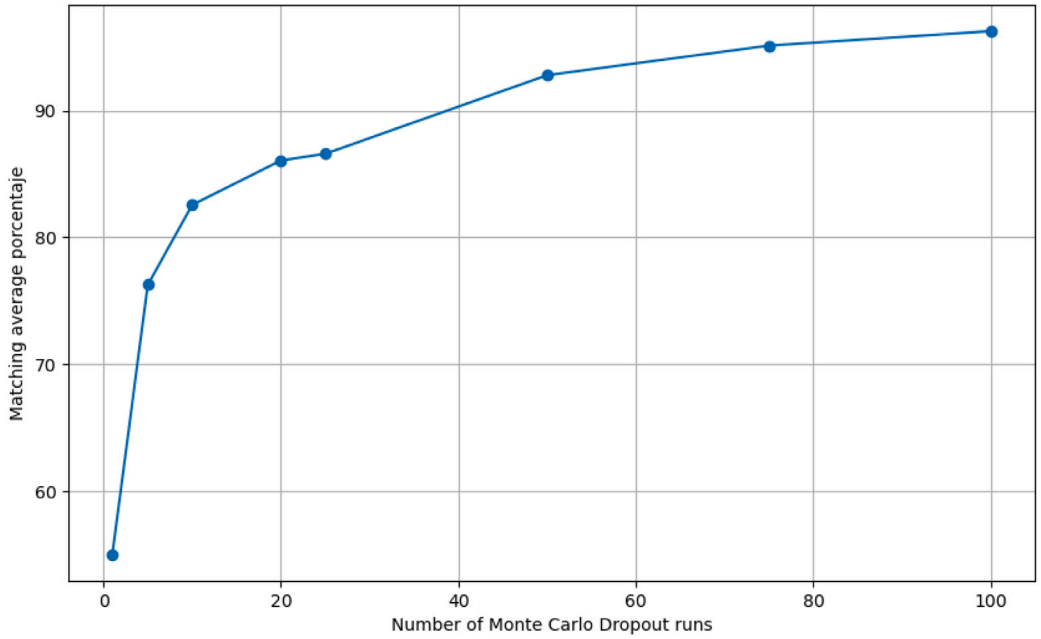

Fig. 3. Figure showing the matching percentage average between to experiments vs runs using Monte Carlo Dropout. Note that with 50 runs, we achieved a matching percentage higher than 90%. This indicates that, beyond this point, further runs yield progressively smaller improvements in matching accuracy.

We analyzed 24-h time series data from five broadband stations within the study area using the two execution methods available in EQT, as detailed in Table 3. Fig. 4 presents the results for the seismic station VPTE, located at Poás Volcano, the closest station to the mainshock in this region. While this figure specifically showcases data from the VPTE station, related figures for the remaining stations can be accessed in the repository of this research at GitHub. This figure illustrates the cumulative number of detected events as a function of time, spanning from 00:00 on February 18 to 00:00 on February 19, 2023. Fig. 4a presents the results obtained using the complex method with Monte Carlo Dropout. It is important to highlight that the algorithm executes the specified number of runs, calculates the confidence value for each run, and then computes an average, which represents the final prediction confidence. EQTransformer does not display the individual results of each Monte Carlo Dropout run; instead, the shown prediction corresponds to the average across all runs. Fig. 4b shows the results using the simplified execution method.

We examined the data from all five stations similarly as shown in Fig. 4, executing the detection process twice to facilitate a comparative study. The comparison between each run or experiment, shown as pink and purple lines in Fig. 4, show clear evidence of non-determinism, regardless of the method used for earthquake detection. We noticed that for the complex method, which relies on the Monte Carlo Dropout for discriminating detections, the overall count of events presents less variance with respect to the simplified method. It is important to note that comparing these two methods can be challenging, as small differences in event detection can significantly affect the percentage difference between the two approaches.

For instance, for the same station, VPTE, the relative difference in earthquake count for the complex method resulted in ±2 events, whereas for the simplified method, the difference was approximately 15 times larger, ±30 events. For our analysis, both detection methods demonstrated that the second experiment often detected more events compared to the first. However, this pattern does not represent a consistent trend. In other cases or for different stations, the first experiment occasionally resulted in a higher number of detections than the second. This inconsistency suggests that the observed differences are not indicative of a systematic behavior.

As displayed in Fig. 4, the difference in the number of detections are scattered throughout the 24-h analysis period, inducing a time shift between the pink and purple curves for both detection methods. However, for the simplified method, a significant divergence begins around 6:00 am, where the differences between the two experiments increase noticeably.

Table 4
Events detected at multiple time-intervals.

Execution method		06:00	12:00	18:00	23:59
simplified	Exp1	156	395	565	661
	Exp2	167	427	586	691
complex	Exp1	11	46	73	81
	Exp2	13	47	75	83

We include zoomed-in plots in Fig. 4 in order to reinforce the observed variability obtained with both algorithms. For the complex method, for instance, the difference remains relatively constant within the zoomed-in time range, whereas for the simplified method, the difference increases within this area. As expected, the number of events increased following the main event at 08:24 UTC. It is important to note that, for the complex method, the difference in the number of events detected between executions remained relatively constant for the remainder of the day after the event, with only minor variations. In contrast, the simplified method exhibited a significant difference in the number of events detected between the two executions during the same period. Additionally, the difference in the number of events between executions is always higher for the simplified method.

To represent these changes, we considered specific time points: 6:00, 12:00, 18:00, and 23:59, as summarized in Table 4 for the VPTE station.

Table 4 reveals that the number of events decreases by approximately tenfold when using the complex method instead of the simplified method, despite consistent detection parameters and conditions. It is important to recall Table 3, where the threshold was kept constant for comparison purposes. However, lowering the threshold for the complex method could result in a higher number of detected events. The lower detection rates of the complex method, despite its robustness in uncertainty modeling, suggest that its threshold for event classification is more conservative compared to the simplified method. This aligns

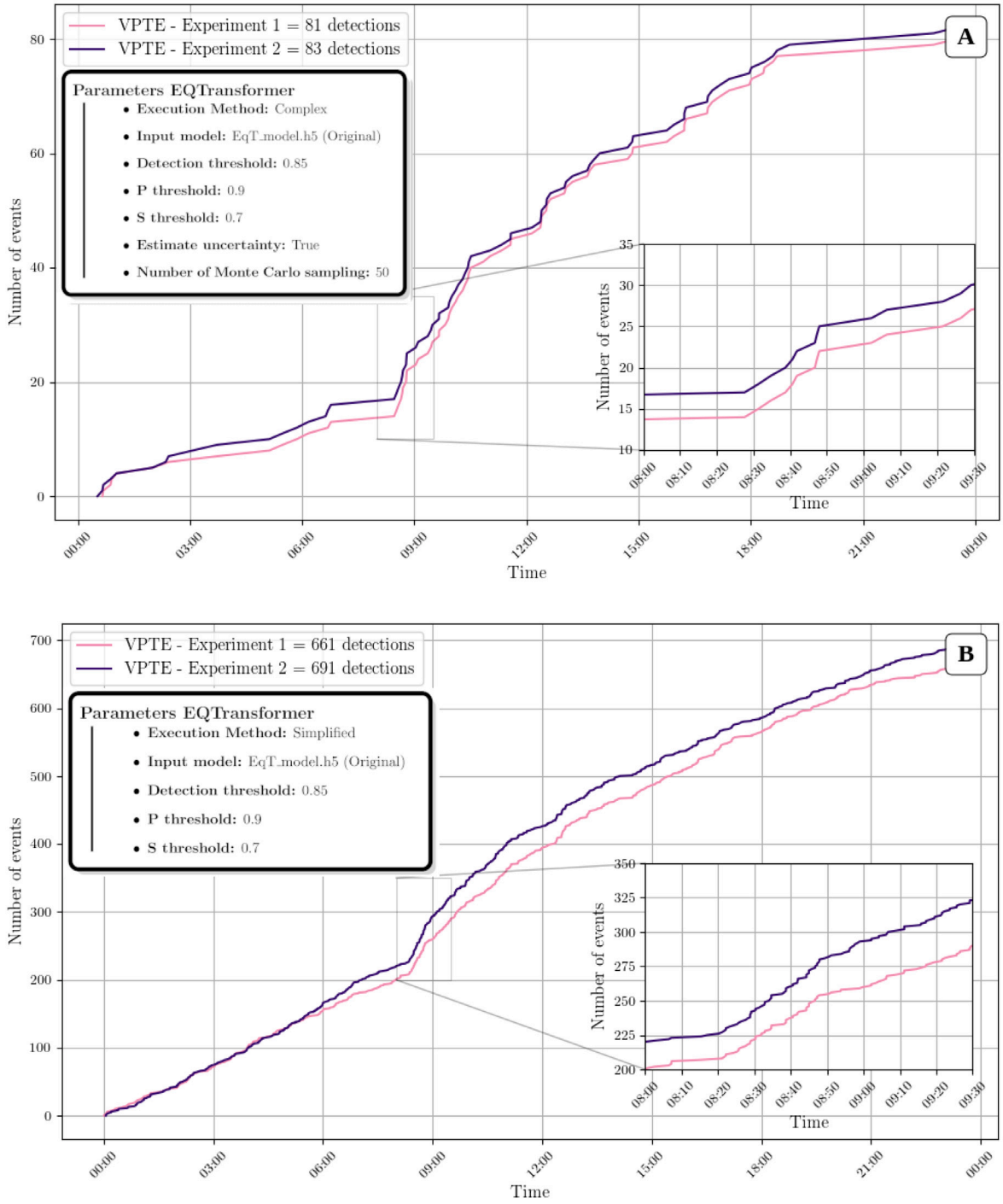

Fig. 4. Results from consecutive experiments performed using seismic station VPTE. For these experiments the parametrization and data were invariant. In panel (a), we show the cumulative number of events detected as function of time using the complex method of EQT, where purple and pink lines indicate the first and second run, respectively. Similarly, panel (b) highlight the results obtained for the same station, VPTE, but using the simplified function. The Purple and pink lines indicate the first and second experiment. Note the difference in the number of detections for both methods.

with the design intention of Monte Carlo Dropout, which prioritizes confidence in detections over volume. However, this conservative approach might result in the omission of low-energy events that could be critical in specific seismic studies, such as aftershock sequences or microseismic monitoring. Another viable approach to increasing the number of detections with the complex method is to use the median instead of the mean for predictions and its runs. The mean can be disproportionately influenced by outliers, potentially causing some detected events to fall below the threshold. For all five stations analyzed, we tested this approach by modifying the code to calculate detection confidence based on the median rather than the mean. This change resulted in approximately a 35% increase in the number of detections, with around 90% matching consistency between experiments after two identical executions. Notably, this percentage remained relatively consistent when using either the mean or the median. Despite these findings, we opted to continue using the mean to better understand the behavior of the original code. The results obtained using the median are available in the research repository at GitHub.

It is important to note that this research focuses on comparing the executions of each method independently rather than making direct

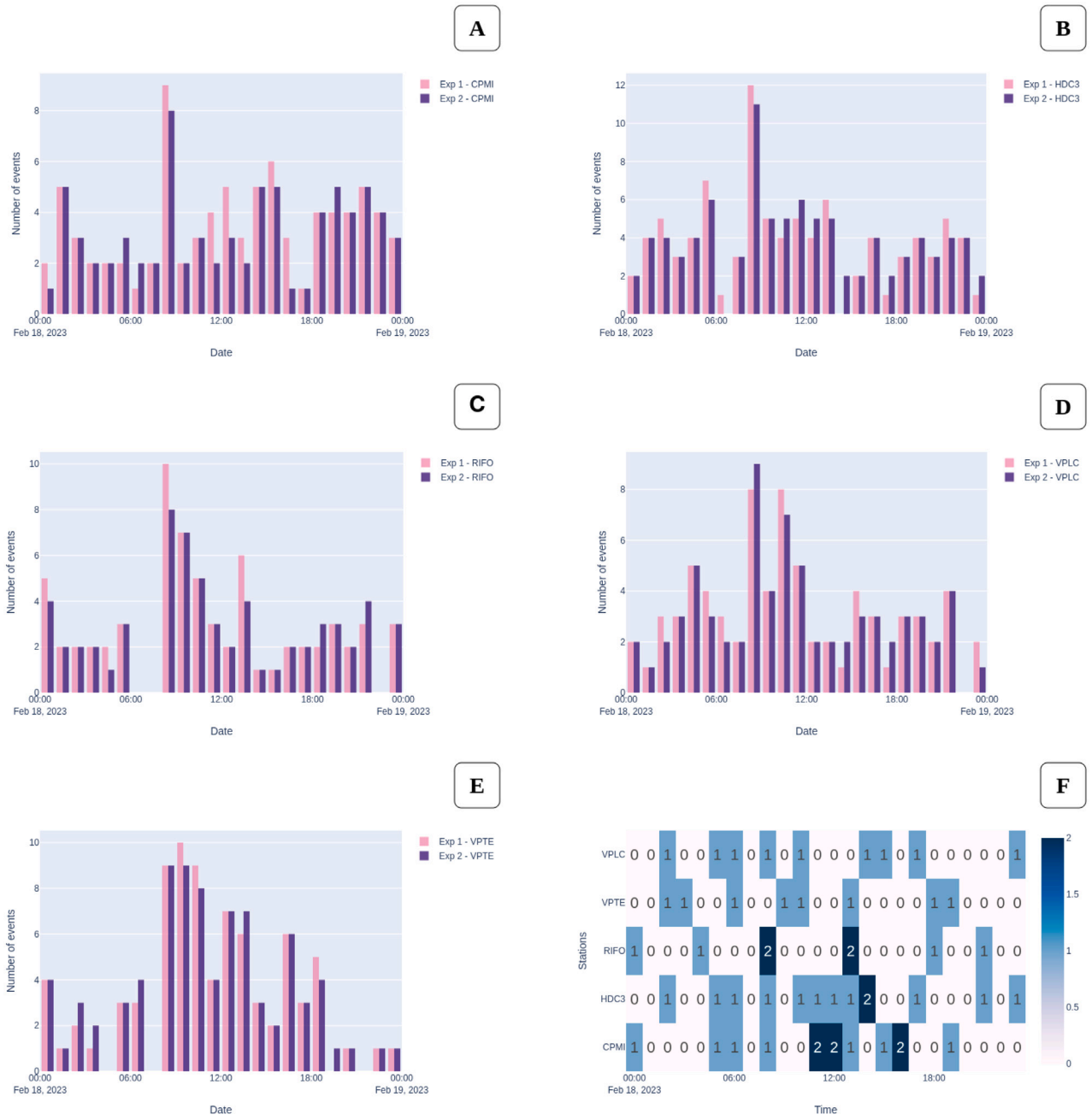

Fig. 5. A, B, C, D, E, Comparison plots of the number of events per hour for two exactly equal experiments using five stations. F, a heatmap of the difference of events per hour between the two experiments, **using complex execution method**.

comparisons between the two. For this reason, the thresholds were kept constant for both methods throughout the experiments. Establishing an equivalence of thresholds between these methods is inherently challenging, as their underlying algorithms differ significantly. Moreover, thresholds are not only model-dependent but can also vary based on the region being analyzed and other external factors influencing the detection process.

The simplified method, with its higher detection rates, offers a more comprehensive capture of events but at the cost of increased noise or false positives. This trade-off highlights the importance of selecting the appropriate method based on the specific objectives of the seismic study. For instance, a study focused on cataloging all possible events might favor the simplified method, while one aimed at precise characterization might opt for the complex approach.

Figs. 5 and 6 show a similar comparison between the number of seismic events detected per hour using both earthquake detection methods (Table 3) across the five seismic stations described above and shown in Fig. 2. The comparison is presented through subplots (A, B, C, D, E) for each station, and a general heatmap (F) that illustrates the difference in the number of events detected between two consecutive experiments using both execution methods. In the heat map, the color indicates the count difference in event detections between the 2 executions. This value is also indicated within each cell.

The differences in the number of detected events across the two experiments, indicate that, both of the detection methods introduces a certain level of variability, with the complex method being more reproducible or less variable than the simplified method by ~1 order of magnitude.

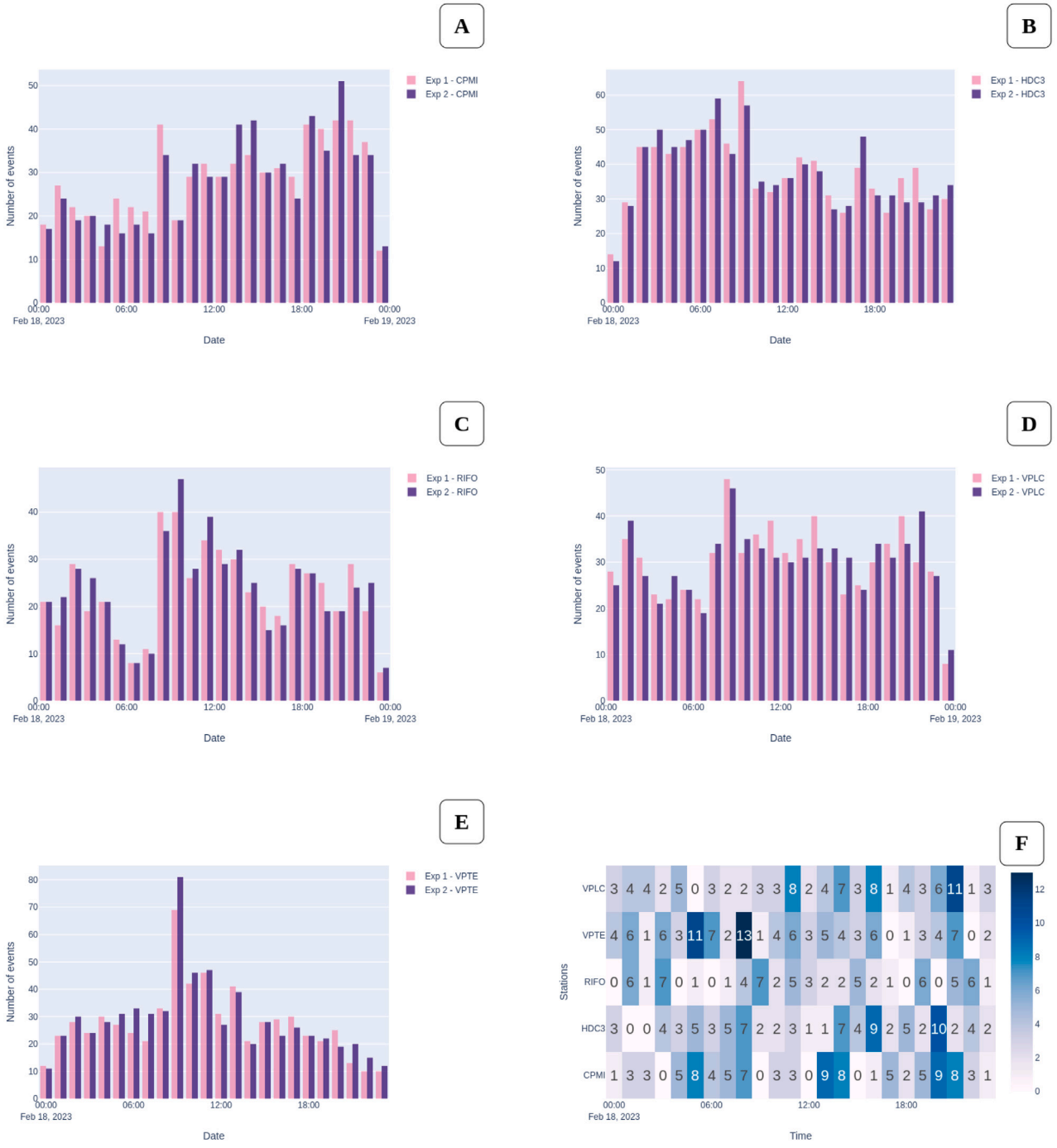

Fig. 6. A, B, C, D, E, Comparison plots of the number of events per hour for two exactly equal experiments using five stations. F, a heatmap of the difference of events per hour between the two experiments, **using simplified execution method**.

This non-determinism may result from the inherent stochastic nature of the detection methods or any potential issues in the computational process. For the case of the complex method, the random sampling process inherent to Monte Carlo Dropout results in different subsets of neurons being dropped out. This means that, even with identical input data and model parameters, the method may produce slightly different outputs in different runs, leading to variability in the number of detected seismic events. Since dropout is applied randomly in each forward pass, the predictions (and thus the detected events) can vary between runs. This stochastic nature is intended to simulate the model's behavior and also quantify uncertainties in event detections within the AI framework.

We developed a match filter technique to evaluate the consistency in event detection for all seismic stations with the aim of exploring time-dependent appearance of new detections, false positives and plausible temporal variations in the number of events detected. For this, we determine whether two events are identical across different executions by comparing the event start time (± 10 s), station, and detection channel (E, N or Z). We tested the match filter method for the two consecutive experiments either for the complex and the simplified method and computed the matching percentage between the experiments. Our findings are summarized in Fig. 7.

According to the filtering criteria described in the methodology section, the complex detection method shows that 85% to 95% of the

Fig. 7. Donut Plot representing the matching percentage between experiments for each station. (a) Using complex method. (b) Using simplified method.

events are identical in two different executions. In contrast, the simplified method exhibits significantly lower performance, with matched events ranging from 60% to 70%.

This significant difference in the matching percentage provides critical insights. For instance, for quick and straightforward detections, the simplified execution method is effective, offering a reasonable matching rate between experiments. However, it also raises a concern, given that about 30% to 40% are unmatched events and thus, results must be interpreted with caution. On the other hand, the complex execution method substantially increases the matching percentage, indicating a more robust approach. Although it detects fewer events, the majority of these events are reproducible across different executions, which is crucial for establishing the reliability of the tool when used by seismological research centers.

Analysis of methodological implications

The observed differences in detection counts and variability between the two methods underscore the trade-offs inherent in their respective approaches. The complex method, with Monte Carlo Dropout, introduces an element of controlled stochasticity that allows for the quantification of uncertainty in detections. However, this also means that reproducibility is inherently limited to a degree, as evidenced by the minor discrepancies in event counts between runs. Such stochastic behavior, while beneficial for uncertainty modeling, can pose challenges in applications where deterministic results are crucial.

In contrast, the simplified method, while more deterministic in design, displayed higher variability in event counts across runs. This increased variability could be linked to differences in algorithmic sensitivity or thresholds in signal detection. The substantial divergence observed in Fig. 4b, particularly around key times such as 6:00 AM, suggests that minor variations in signal processing parameters or noise conditions might disproportionately affect the outcome of this method.

Ensuring reproducibility in EQTransformer

An in-depth analysis of the EQT source code was conducted to thoroughly comprehend its architecture and address the issue of non-determinism identified in this study. EQT represents a sophisticated framework comprising numerous interdependent modules, libraries, and dependencies, all of which must operate in harmony to deliver the expected performance and reliability.

At the core of EQT's functionality lies TensorFlow, a widely recognized and versatile library that plays a pivotal role in the framework. TensorFlow facilitates GPU acceleration and multiprocessing, both critical for optimizing computational efficiency. Additionally, it underpins the neural network architecture, handling tasks such as layer configuration, training, inference, and data processing.

The inherent complexity of EQT, combined with its reliance on a diverse ecosystem of libraries, poses significant challenges to achieving determinism and reproducibility. Addressing this requires a deep understanding of how these components interact, particularly TensorFlow's handling of stochastic processes such as random seed initialization. Setting the random seed ensures that random processes are not entirely random; they become reproducible (Jain, 2023). By explicitly implementing tf.random.set_seed in both the complex and simplified execution methods, we have demonstrated that 100% reproducibility can be achieved, provided input conditions and parameters remain constant.

The effectiveness of this approach stems from its ability to control the random number generation processes that underpin many deep learning operations. Reproducibility is not merely a technical goal but a foundational requirement for scientific validation and the reliable dissemination of findings. By implementing this straightforward yet impactful solution, we mitigate the variability inherent in TensorFlow's processes, significantly enhancing the reliability and trustworthiness of EQT for researchers. Furthermore, this adjustment aligns with best practices in data science and machine learning, ensuring that EQT remains a robust tool for advancing seismological research.

5. Conclusions

The results obtained in this study reveal significant differences in seismic event detection when comparing the simplified execution method and complex execution method. The complex method consistently detects fewer events, approximately one-tenth compared to simplified. This outcome underscores the impact of the iterative Monte Carlo dropout used in the complex method, which appears to enhance model robustness by reducing false positives.

Moreover, there is a notable difference in the consistency of event detection between the two methods. The complex method exhibits minimal variability between repeated runs. In contrast, the simplified method shows considerable variability, with differences reaching up to one order of magnitude in some cases. This suggests that the complex method provides more reliable and reproducible results, which are crucial for accurate seismic analysis.

Regarding temporal patterns and major events, both methods tend to be stable on the number of events detected up to the mainshock. However, after this event, the simplified execution method shows a significant difference in the number of events detected throughout the remainder of the day, while the complex execution method maintains this difference on a much smaller scale. This indicates that although both methods are effective in identifying major events, the complex execution method sustains more consistent performance over extended periods.

Donut plots comparing the percentage of matched events between the two methods reveal that the complex method achieves a higher match rate (85% to 95%) compared to the simplified method (60% to 70%). This suggests that while the complex method detects fewer events, it does so with greater reliability and consistency in identifying the same events across repeated runs.

Our findings are critical for optimizing the use of EQTransformer and AI tools in seismological research. The complex execution method, with its enhanced consistency and reliability, is better suited for applications requiring high precision and reproducibility, making it more recommended for professional use, such as in seismological research institutions. Meanwhile, the simplified execution method, despite its higher event detection rate, may be more prone to variability and false positives. However, it offers the advantage of being easier to use and computationally lighter, making it suitable for non-professional tasks, such as training, academic purposes, and other less demanding applications.

Finally, incorporating TensorFlow's random seed initialization across both execution methods ensures reproducibility for all purposes, thereby transforming EQT into a more reliable and trustworthy tool for the scientific community. This enhancement directly addresses one of the most critical challenges in computational seismology and provides a solid foundation for future research.

Code availability

This research utilized multiple codes and tools, some of which were developed by us, in addition to the use of EQTransformer (Mousavi et al., 2020). As this research is an extension of the OKSP workflow developed in 2021 (van der Laat et al., 2021), we provide the necessary tools, code, and data to reproduce our results.

CRediT authorship contribution statement

Sebastián Gamboa-Chacón: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Esteban Meneses:** Writing – review & editing, Validation, Supervision, Project administration, Conceptualization. **Esteban J. Chaves:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partially supported by a machine allocation on Kabré supercomputer at the National High Technology Center, Costa Rica.

Hardware requirements

- **Operating System:** Linux 64-bit (cluster, server, or personal computer).
- **GPU Recommendation:** We recommend using an NVIDIA GPU to achieve faster results.

Programming language

- **Python:** All scripts and tools are developed in Python 3.

Software requirements

- **Conda Environment:** We recommend working within a Conda environment for consistency and ease of reproduction. To facilitate this, we provide a clone of our environment. Detailed instructions for setting up Conda can be found in the following tutorial: https://github.com/um-dang/conda_on_the_cluster.git
- **EQTransformer:** The EQTransformer tool can be accessed by cloning the following repository: https://github.com/smousavi05/EQTransformer.git
  Note: We strongly recommend using our provided Conda environment as it contains updated software libraries that we have actively used in this research.
- **Research Source Code:** The source code necessary for the detection stage, based on the OKSP pipeline (van der Laat et al., 2021), along with additional code and data required for reproducing the results, is available.
  Note: A README file is included in the repository, providing step-by-step instructions for use.
  The source code is available for download at the following link: https://github.com/SebasGamboa10/Reproducibility-and-Uncertainty-Assessment-in-EQTransformer.git

Data availability

This research utilized multiple codes and tools, some of which were developed by us, in addition to the use of EQTransformer (Mousavi et al., 2020). As this research is an extension of the OKSP workflow developed in 2021 (van der Laat et al., 2021), we provide the necessary tools, code, and data to reproduce our results. We have provided the necessary data for reproducibility, which can be found in a public repository at https://github.com/SebasGamboa10/Reproducibility-and-Uncertainty-Assessment-in-EQTransformer.git.

References

Arrowsmith, S.J., Trugman, D.T., MacCarthy, J., Bergen, K.J., Lumley, D., Magnani, M.B., 2022. Big data seismology. Rev. Geophys. 60 (2), e2021RG000769.

Castillo, E., Siervo, D., Prieto, G.A., 2024. Colombian seismic monitoring using advanced machine-learning algorithms. Seismol. Res. Lett. http://dx.doi.org/10.1785/0220240036.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. arXiv:1506.02142.

Gürsoy, G., Varol, A., Nasab, A., 2023. Importance of machine learning and deep learning algorithms in earthquake prediction: A review. In: 2023 11th International Symposium on Digital Forensics and Security. ISDFS, pp. 1–6. http://dx.doi.org/10.1109/ISDFS58141.2023.10131766.

Hassan, R., Hejrani, B., Zhang, F., Gorbatov, A., Medlin, A., 2020. High-Performance Seismological Tools (HiPerSeis). Geoscience Australia Canberra.

Humaira, H., Rasyidah, R., 2020. Determining the Appropiate Cluster Number Using Elbow Method for K-Means Algorithm. EAI, http://dx.doi.org/10.4108/eai.24-1-2018.2292388.

Instituto Costarricense de Electricidad and Universidad de Costa Rica, 2009. El terremoto de cinchona del 8 de enero de 2009 - informe completo. pp. 1–132, Red Sismológica Nacional (RSN: UCR-ICE).

Jain, A., 2023. Seeding success: A guide to setting seeds in data science. URL: https://medium.com/@abhishekjainindore24/seeding-success-a-guide-to-setting-seeds-in-data-science-5d24e00c3dc7.

Jena, R., Pradhan, B., Beydoun, G., Alamri, A.M., Ardiansyah, Nizamuddin, Sofyan, H., 2020. Earthquake hazard and risk assessment using machine learning approaches at Palu, Indonesia. Sci. Total Environ. 749, 141582. http://dx.doi.org/10.1016/j.scitotenv.2020.141582, URL: https://www.sciencedirect.com/science/article/pii/S0048969720351111.

Jiang, C., Fang, L., Fan, L., Li, B., 2021. Comparison of the earthquake detection abilities of PhaseNet and EQTransformer with the Yangbi and Maduo earthquakes. Earthq. Sci. 34 (5), 425–435. http://dx.doi.org/10.29382/eqs-2021-0038, URL: https://www.sciencedirect.com/science/article/pii/S1674451922000581.

van der Laat, L., Baldares, R.J.L., Chaves, E.J., Meneses, E., 2021. OKSP: A novel deep learning automatic event detection pipeline for seismic monitoringin costa rica. arXiv:2109.02723.

M. Saad, O., Chen, Y., Siervo, D., Zhang, F., Savvaidis, A., Huang, D., Igonin, N., Fomel, S., Chen, Y., 2023. EQCCT: A production-ready earthquake detection and phase-picking method using the compact convolutional transformer. IEEE Trans. Geosci. Remote Sens. PP, 1. http://dx.doi.org/10.1109/TGRS.2023.3319440.

Montero, W., Denyer, P., Barquero, R., Alvarado, G., Cowan, H., 1998. Map of Quaternary Faults and Folds of Costa Rica. Technical Report.

Mousavi, S.M., Ellsworth, W.L., Zhu, W., Chuang, L.Y., Beroza, G.C., 2020. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nat. Commun. 11 (1), 1–12.

Pita-Sllim, O., Chamberlain, C.J., Townend, J., Warren-Smith, E., 2023. Parametric testing of eqtransformer's performance against a high-quality, manually picked catalog for reliable and accurate seismic phase picking. Seism. Rec. 3 (4), 332–341. http://dx.doi.org/10.1785/0320230024.

Protti, M., Gundel, F., McNally, K., 1994. The geometry of the wadati-benioff zone under southern central america and its tectonic significance: results from a high-resolution local seismographic network. Phys. Earth Planet. Inter. 84 (1), 271–287. http://dx.doi.org/10.1016/0031-9201(94)90046-9, URL: https://www.sciencedirect.com/science/article/pii/0031920194900469.

Spassiani, I., Sebastiani, G., 2016. Exploring the relationship between the magnitudes of seismic events. J. Geophys. Research: Solid Earth 121 (2), 903–916.

Styron, R., García-Pelaez, J., Pagani, M., 2020. CCAF-DB: the Caribbean and Central American active fault database. Nat. Hazards Earth Syst. Sci. 20 (3), 831–857.

UCR, 2015. Terremoto de Cartago, 4 de mayo de 1910. — rsn.ucr.ac.cr. https://rsn.ucr.ac.cr/actividad-sismica/ultimos-sismos/26-sismologia/sismos-historicos/3486-otro.

Wang, S., Liu, F., Yin, X.-x., Chen, K., Cai, R., 2023. Employing convolution-enhanced attention mechanisms for earthquake detection and phase picking models. Front. Earth Sci. 11, http://dx.doi.org/10.3389/feart.2023.1283857.