

DOI 10.54596/2958-0048-2026-3-319-339

UDC 004.8:519.876

IRSTI 37.31.19

## A HYBRID MACHINE LEARNING APPROACH FOR REAL-TIME FORECASTING OF MINUTE-LEVEL SEISMIC SIGNAL AMPLITUDES

B. Sapy<sup>1</sup>, Zh.R. Umarova<sup>1\*</sup>, Zh.S. Kemelbekova<sup>2</sup>, A.T. Kalbayeva<sup>1</sup>, Yu. Jusoh<sup>3</sup>

<sup>1\*</sup>*Department of Information Systems and Modeling, M. Auezov South Kazakhstan University,  
Shymkent, Kazakhstan*

<sup>2</sup>*Department of Computer Science, M. Auezov South Kazakhstan University,  
Shymkent, Kazakhstan*

<sup>3</sup>*Faculty of Computer Science and Information Technology, University Putra Malaysia,  
Kuala Lumpur, Malaysia*

\*Corresponding author: [Zhanat-u@mail.ru](mailto:Zhanat-u@mail.ru)

### Abstract

Real-time interpretation of seismic waveform data is essential for monitoring signal dynamics and detecting abnormal activity. However, forecasting seismic amplitudes remains challenging because seismic signals are highly variable, noisy, and non-stationary. This study presents an automated, gap-aware real-time forecasting pipeline for minute-level seismic signal amplitudes built on IRIS FDSN waveform data, together with a rigorous, statistically grounded comparison of seven forecasting models: Naive persistence, ARIMA, Ridge regression, HistGradientBoosting, a standard Gated Recurrent Unit (GRU), an LSTM, and a physics-inspired regularized GRU (PI-GRU) that adds temporal smoothness and amplitude-growth constraints to the training loss. Models were evaluated on 81 independently sampled, non-overlapping forecasting windows drawn from a continuous 91-day observation period across four horizons (5, 15, 30, and 60 minutes), using paired non-parametric significance testing, a quantitative ablation study, and noise- and missing-data robustness testing, and the main findings were additionally verified under an alternative peak-priority signal aggregation. Classical models, particularly Ridge regression and HistGradientBoosting, achieved significantly lower forecasting error than every neural model tested (Friedman and Holm-corrected Wilcoxon tests,  $p < 0.0001$ ) and were robust to synthetic noise and missing data, while the proposed PI-GRU was, contrary to expectation, the weakest-performing model overall and did not produce smoother forecast trajectories than an unregularized baseline GRU. We report this as an honest negative result: as currently configured, the physics-inspired regularization examined here does not demonstrate an accuracy or stability advantage over classical machine learning baselines for this task. The principal contributions of this work are the automated real-time forecasting pipeline itself and a reusable, rigorous evaluation methodology, together with specific, testable hypotheses – concerning hyperparameter tuning, training budget, and calibration of the physics-inspired constraint – for why the neural models underperformed and how this might be addressed in future work.

**Keywords:** Seismic signal forecasting; physics-inspired neural networks; time series prediction; hybrid forecasting models; machine learning; real-time monitoring.

## СЕЙСМИКАЛЫҚ СИГНАЛ АМПЛИТУДАЛАРЫН МИНУТТЫҚ ДЕҢГЕЙДЕ НАҚТЫ УАҚЫТТА БОЛЖАУҒА АРНАЛҒАН ГИБРИДТІ МАШИНАЛЫҚ ОҚЫТУ ТӘСІЛІ

Сапы Б.<sup>1</sup>, Умарова Ж.Р.<sup>1\*</sup>, Кемельбекова Ж.С.<sup>2</sup>, Калбаева А.Т.<sup>1</sup>, Jusoh Yu.<sup>3</sup>

<sup>1\*</sup>Ақпараттық жүйелер және модельдеу кафедрасы, М. Әуезов атындағы

Оңтүстік Қазақстан университеті, Шымкент, Қазақстан

<sup>2</sup>Информатика кафедрасы, М. Әуезов атындағы Оңтүстік Қазақстан университеті,  
Шымкент, Қазақстан

<sup>3</sup>Компьютерлік ғылымдар және ақпараттық технологиялар факультеті,

Путра Малайзия университеті, Куала-Лумпур, Малайзия

\*Хат алмасуға жауапты автор: [Zhanat-u@mail.ru](mailto:Zhanat-u@mail.ru)

### Аңдатпа

Сейсмикалық толқындық деректерді нақты уақыт режимінде интерпретациялау сигнал динамикасын бақылау және аномальды белсенділікті анықтау үшін маңызды болып табылады. Алайда сейсмикалық амплитудаларды болжау әлі де күрделі міндет болып қала береді, себебі сейсмикалық сигналдар жоғары өзгергіштікпен, шу деңгейінің көптігімен және стационарлы емес сипатымен ерекшеленеді. Бұл зерттеуде IRIS FDSN толқындық деректерін пайдалана отырып, минуттық деңгейдегі сейсмикалық сигнал амплитудаларын нақты уақыт режимінде болжауға арналған автоматтандырылған, үзілістерге төзімді конвейер, сондай-ақ жеті болжау моделін – Naive (persistence), ARIMA, Ridge-регрессия, HistGradientBoosting, стандартты GRU, LSTM және оқыту функциясына уақыттық тегістік пен амплитуданың өсуін шектейтін қосымша қосатын физикалық қағидаларға негізделген регуляризацияланған GRU (PI-GRU) – қатаң, статистикалық тұрғыдан негізделген салыстыруы ұсынылады. Модельдер 91 күндік үздіксіз бақылау кезеңінен алынған, төрт горизонт (5, 15, 30 және 60 минут) бойынша 81 тәуелсіз, қиылыспайтын болжау терезесінде бағаланды, бағалауда жұптасқан параметрсіз статистикалық маңыздылық тесттері, сандық абляциялық талдау және шу мен деректердің жетіспеушілігіне төзімділік тесттері қолданылды, ал негізгі нәтижелер балама peak-priority агрегациясы бойынша қосымша тексерілді. Классикалық модельдер, әсіресе Ridge-регрессия мен HistGradientBoosting, барлық нейрондық модельдерге қарағанда статистикалық тұрғыдан едәуір төмен болжау қатесін көрсетті (Friedman және Holm түзетілген Wilcoxon тесттері,  $p < 0.0001$ ) және синтетикалық шу мен жетіспейтін деректерге төзімді болды, ал ұсынылған PI-GRU моделі, керісінше, барлық модельдер арасында ең нашар нәтиже көрсетті және тегістелмеген базалық GRU моделімен салыстырғанда тегісірек болжам траекторияларын қалыптастыра алмады. Біз мұны жасырмай, адал теріс нәтиже ретінде хабарлаймыз: қазіргі конфигурацияда осы зерттеуде қарастырылған физикалық қағидаларға негізделген регуляризация классикалық машиналық оқыту әдістерімен салыстырғанда дәлдік немесе тұрақтылық артықшылығын көрсете алмады. Бұл жұмыстың негізгі үлесі – автоматтандырылған нақты уақыттағы болжау конвейерінің өзі және қайта пайдалануға болатын, қатаң бағалау әдістемесі, сондай-ақ нейрондық модельдердің неліктен төмен нәтиже көрсеткені және болашақта мұны қалай жақсартуға болатыны туралы нақты, тексерілетін болжамдар болып табылады.

**Кілт сөздер:** Сейсмикалық сигналдарды болжау; физикалық принциптерге негізделген нейрондық желілер; уақыттық қатарларды болжау; гибриді болжау модельдері; машиналық оқыту; нақты уақыт режиміндегі мониторинг.

**ГИБРИДНЫЙ ПОДХОД МАШИННОГО ОБУЧЕНИЯ  
ДЛЯ ПРОГНОЗИРОВАНИЯ АМПЛИТУД СЕЙСМИЧЕСКИХ СИГНАЛОВ  
НА МИНУТНОМ УРОВНЕ В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ**

**Сапы Б.<sup>1</sup>, Умарова Ж.Р. <sup>1\*</sup>, Кемельбекова Ж.С.<sup>2</sup>, Калбаева А.Т. <sup>1</sup>, Jusoh Yu.<sup>3</sup>**

<sup>1\*</sup>*Кафедра Информационных систем и моделирования, Южно-Казахстанский  
университет имени М. Ауэзова, Шымкент, Казахстан*

<sup>2</sup>*Кафедра Информатики, Южно-Казахстанский университет имени М. Ауэзова,  
Шымкент, Казахстан*

<sup>3</sup>*Факультет компьютерных наук и информационных технологий,  
Университет Путра Малайзия, Куала-Лумпур, Малайзия*

*\*Автор для корреспонденции: [Zhanat-u@mail.ru](mailto:Zhanat-u@mail.ru)*

**Аннотация**

Интерпретация сейсмических волновых данных в режиме реального времени имеет важное значение для мониторинга динамики сигналов и выявления аномальной активности. Однако прогнозирование сейсмических амплитуд остаётся сложной задачей, поскольку сейсмические сигналы характеризуются высокой изменчивостью, шумом и нестационарностью. В данном исследовании представлен автоматизированный, устойчивый к пропускам конвейер прогнозирования амплитуд сейсмических сигналов на минутном уровне в реальном времени на основе волновых данных IRIS FDSN, а также строгое, статистически обоснованное сравнение семи моделей прогнозирования: наивной модели (persistence), ARIMA, Ridge-регрессии, HistGradientBoosting, стандартной GRU, LSTM и регуляризованной GRU, основанной на физических принципах (PI-GRU), которая добавляет в функцию потерь ограничения временной гладкости и штраф за нереалистичный рост амплитуды. Модели были оценены на 81 независимом, неперекрывающемся окне прогнозирования, отобранном из непрерывного 91-дневного периода наблюдений, для четырёх горизонтов (5, 15, 30 и 60 минут); оценка включала парные непараметрические тесты статистической значимости, количественное абляционное исследование и тестирование устойчивости к шуму и пропускам данных, а основные результаты были дополнительно проверены при альтернативной peak-priority агрегации сигнала. Классические модели, в особенности Ridge-регрессия и HistGradientBoosting, показали статистически значимо более низкую ошибку прогнозирования по сравнению со всеми нейросетевыми моделями (тесты Фридмана и Уилкоксона с поправкой Холма,  $p < 0.0001$ ) и продемонстрировали устойчивость к синтетическому шуму и пропускам данных, тогда как предложенная модель PI-GRU, вопреки ожиданиям, показала наихудший результат среди всех моделей и не позволила получить более сглаженные траектории прогноза по сравнению с нерегуляризованной базовой GRU. Мы сообщаем об этом как о честном отрицательном результате, не скрывая его: в текущей конфигурации рассмотренная в этой работе физически обоснованная регуляризация не демонстрирует преимущества в точности или устойчивости по сравнению с классическими методами машинного обучения. Основным вкладом данной работы являются сам автоматизированный конвейер прогнозирования в реальном времени и воспроизводимая, строгая методология оценки, а также конкретные, проверяемые гипотезы относительно того, почему нейросетевые модели показали худший результат и как это можно улучшить в дальнейшей работе.

**Ключевые слова:** прогнозирование сейсмических сигналов; нейронные сети, основанные на физических принципах; прогнозирование временных рядов; гибридные модели прогнозирования; машинное обучение; мониторинг в реальном времени.

**Introduction**

Real-time seismic monitoring systems continuously collect waveform data to identify earthquakes and other ground motion events. Besides the post-event detection, it is convenient to estimate the transient behaviour of the seismic signal at the future time. Such short-term forecasting can aid anomaly detection, automated alerting systems, and monitoring of abnormal ground motion levels. However, predicting continuous seismic signals is complex, especially since the data is noisy and quite variable and often shows non-stationary behavior due to sudden

amplitude spikes. A high sampling rate, typically one sample per second or more, is typical for current seismic stations to measure waveforms. In operational systems, forecasting directly these types of raw waveform streams is not easy to perform due to the rapid signal variation of the waveforms together with the large data volume. In addition, latency of the network or interrupted transmission are likely to introduce temporary gaps in live data streams. Therefore, typical real-time forecasting applications usually employ small waveform representations that improve noise and computational difficulty [1].

The continuous seismic waveform is converted in this work to a minute level time series representation. This calculates the mean absolute amplitude of the filtered signal per minute. Because of this, the minute-mean representation minimizes the impact of the isolated spikes and transient artifacts that frequently dominate measuring peaks. While minute-level maximum amplitudes provide poor representation of the total seismic activity, the mean absolute amplitude gives us a much more stable and noise-resistant picture within a given timeframe.

Machine learning models are well suited to predicting such time series, as patterns can be learned directly from data [1,2,3]. Linear models like Ridge regression are computationally efficient and stable but they might not take into account nonlinear relationships. Tree-based ensemble methods like Histogram-based Gradient Boosting are capable of modeling nonlinear dependencies and often perform well with noisy signals. Gated Recurrent Unit (GRU) is an example of this type of recurrent neural network that is built to capture temporal dependencies in sequential data representation. Yet, purely data-driven neural networks may have erratic predictions or unrealistic spikes as the signal distribution changes. To solve this issue, we propose a physics-inspired regularization approach in this work. In place of developing an entire physical model of seismic wave propagation, physical constraints are introduced to the neural network loss function, following the general idea of embedding physically motivated penalties into deep learning models for seismic and geophysical forecasting [4,5]. Specifically, the model punishes excessive growth in the amplitude of the future signal amplitude and promotes gradual evolution with respect to the prediction over the expected time scale. This leads to a physics-inspired GRU (PI-GRU) model, to achieve both more stable and plausible prediction.

This work mainly contributes to develop a real-time forecasting pipeline for minutely scaled seismic amplitude. We use IRIS's FDSN service to automatically retrieve waveform data, convert the signal into a minute-mean amplitude series, train forecasting models in rolling historical windows, generate predictions at 5, 15, 30, and 60 minutes windows, and so on. After the prediction horizon elapses, true-future validation is done. Using the same operational pipeline, three forecasting models (Ridge regression, Histogram Gradient Boosting, and proposed PI-GRU model) are simulated and evaluated. Standard forecasting error metrics are employed to compare their performance, and predicted and observed amplitude trajectories are visualized in the real-time context [6,7].

One could point at recent works employing machine learning approaches for seismic signal analysis and short-term forecasting. For example, deep learning architectures have been employed for earthquake early-warning and forecasting, while more general physics-informed multi-step forecasting frameworks have also been proposed for related geophysical and environmental time series such as temperature [8,9]. Approaches like these can in principle model arbitrary nonlinear relationships in observed signals and have been shown capable of identifying some observable geophysical patterns.

One limitation of existing work is that these models are often not designed for forecasting continuous waveforms of seismic amplitude, rather they are aimed at event detection/classification tasks. Another major caveat is that most deep-learning models

experience stability issues when directly applied to noisy and non-stationary signals of this kind, resulting in erratic predictions manifested as spikes or unrealistically-oscillating model outputs [10]. Such instability would likely cause problems for real-time monitors that make use of this technology.

Yet another caveat of most previous work is that data-driven models typically do not leverage any domain knowledge to improve performance. Recently, physics-informed neural networks have been introduced in related structural and infrastructure monitoring contexts, such as railway bridge assessment [11], as well as in broader physics-informed learning frameworks [12]. However, most PDE-informed approaches make use of either full physical models or partial differential equations. Full models cannot be leveraged in operational real-time systems, creating a need for approaches that impose some physical constraints on data-driven models.

To address this limitation, here we introduce a hybrid short-term forecasting framework that combines classical machine learning regression models with a physics-inspired neural network. Rather than hard-coding full physical relationships into the network, we instead impose simple, but effective smoothness and amplitude growth constraints. Such an approach enables numerical stability while retaining the flexibility of a data-driven model [13,14].

### Materials and Methods

A source of the seismic data used in the study is an operational broadband seismic station through the IRIS (Incorporated Research Institutions for Seismology) data service. Real-time continuous waveform records are obtained using the FDSN (International Federation of Digital Seismographic Networks) information service interface. Sustained energy collection was performed by collecting vertical ground motion measurements on the selected station (QZ.CHM.20.HNZ). This station provides vertical ground motion with a sampling interval of approximately one second (sampling interval 1 second). This duration typically depends on sensor positioning. The ObsPy Python library for seismological data processing used for realizing waveform acquisition and preprocessing.

The waveform segment is initially detrended to remove linear offsets and then filtered with a band-pass filter (1-10 Hz). This frequency range was chosen to reduce background noise and very high frequency disturbances over many periods without disturbing the dominant features of local and regional seismic activity. Let  $x(t)$  be the filtered seismic signal.

IRIS gives waveform data at the second (or sub-second) level, however such precise forecasting at this resolution is unstable and is computationally expensive for real-time operation. Thus, the signal is aggregated to one-minute temporal resolution. In contrast to most seismic monitor systems that depend on peak value, this study utilizes the mean absolute amplitude in 1-minute and in minutes as being

$$A_m(k) = \frac{1}{N_k} \int_{i=1}^{N_k} |x(t_i)|, \quad (1)$$

where,  $A_m(k)$  is the average amplitude in min  $k$ ,  $x(t_i)$  are samples in minute filtered waveform and  $N_k$  is the number of available min samples.

This was a conscious decision. The minute-mean amplitude provides a much more stable and noise-proof measurement of ground compared to minute maxima, which are usually characterised by isolated spikes or transient artifacts. The cumulative series is therefore smoother in nature and useful for the short-run prognosis.

The measured energy-like quantity  $E_m(k)$  is also described in order to stay in relation to physical limits:

$$E_m(k) = A_m(k)^2. \quad (2)$$

This quantity is not the true physical energy of the seismic signal; rather, it is a relative measure of signal intensity used to prevent unrealistic growth in the predicted amplitudes within the physics-inspired loss function.

In near real-time situations minute-level data may be partially missing due to telemetry delays or intermittent outages at stations in live-time. In this regard, the minute time index is reconstructed as a continuous sequence. Linear interpolation fills in these gaps at different levels of the model. Forecasting is allowed only if a certain set of the recently received minutes are present in the time-space (gap-aware validation). This allows the forecasting models to be trained, tested, and evaluated even with a limited amount of recent data, while enabling the rolling-retraining pipeline to continue operating continuously as new data arrive [15].

The forecasting problems solved in this work are as follows. For a univariate time series over minute level mean absolute seismic amplitudes  $\{A_m(t)\}_{t=1}^T$ , the goal is to forecast the future sequence

$$\{A_m(t+1), A_m(t+2), A_m(t+3), \dots, A_m(t+H)\}, \quad (3)$$

where  $H$  defined as the prediction horizon.

All models in this study work on the same executed input series, generate input series, and make straight multi-step forecasts, not one-step recursively generated predictions. The design eliminates error in building up to an error and is more convenient for real-time applications. In general terms, the forecasting task is to learn mapping

$$f: \mathbb{R}^W \rightarrow \mathbb{R}^H, \quad (4)$$

such that

$$[A_m(t-W+1), \dots, A_m(t)] \xrightarrow{f} [\widehat{A}_m(t+1), \dots, \widehat{A}_m(t+H)] \quad (5)$$

where,  $W$  is the input window length (minutes).

As a first baseline, a multivariate Ridge regression model is used. Sliding windows of length  $W$  are extracted from the amplitude series and mapped directly to future values over the horizon  $H$ . Ridge regression model is based upon a linear relationship between recent observations and future values. For this window we construct the forecast as

$$\widehat{y}_t = W^T x_t, \quad (6)$$

where,  $W \in \mathbb{R}^{W \times H}$  is the weight matrix. The parameters are estimated by solving

$$\min_W \|Y - XW\|_2^2 + \alpha \|W\|_2^2, \quad (7)$$

where,  $X$  contains input windows,  $Y$  are the corresponding values of future minute-mean amplitudes, and  $\alpha$  is regularization strength.

The model, although simple, serves as stable reference and well performs in stationary seismic conditions.

A Histogram-based Gradient Boosting model is applied, in order to deal with nonlinear effects [1]. Whereas each input window is converted into a feature vector

$$\phi(x_t) = [\mu_k, \sigma_k, \max_k, \min_k], k \in \{5, 15, W\}' \quad (8)$$

where  $\mu_k$  and  $\sigma_k$  denote mean and standard deviation over a diversity of subwindows. In this section, each forecast step  $h \in \{1, \dots, H\}$  is modeled independently as

$$\widehat{A}_m(t+h) = \sum_{j=1}^J \eta_j T_j(\phi(x_t)), \quad (9)$$

where,  $T_j$  are shallow decision trees and  $\eta_j$  are learned weights. This model is flexible, faster to train, and robust to moderate noise.

To more explicitly model the temporal dependencies, we are using the GRU in the Gated Recurrent Unit. Based on the input sequence,  $\{z(t-W+1), \dots, z(t)\}$ , the GRU updates its hidden state as

$$h_t = \text{GRU}(z(t), h_{t-1}), \quad (10)$$

in which  $h_t$  encodes the recent seismic history. The prediction is made by a dense output layer:

$$\hat{z}_{t+1:t+H} = W_0 h_t + b_0. \quad (11)$$

This layer forecasts minute-mean amplitude features, but not maximum values (the peak values), which makes for a less erratic and more realistically physical network behaviour.

This can alleviate the high sensitivity to spikes and causes the amplitude series to be transformed and expressed as

$$z(t) = \log(1 + A_m(t)). \quad (12)$$

The training is carried out in this log domain and the predictions are mapped back with

$$\hat{A}_m(t) = \exp(\hat{z}(t)) - 1. \quad (13)$$

This translation stabilizes the optimization and enables improved behaviour over high variance intervals.

Physics-inspired regularization techniques have recently been used to incorporate domain knowledge into neural networks [2]. The GRU is trained using a composite loss function is written as

$$\mathcal{L} = \mathcal{L}_{data} + \lambda_s \mathcal{L}_{smoot} + \lambda_e \mathcal{L}_{phys}. \quad (14)$$

Data fidelity term (Huber Loss) is shown below

$$\mathcal{L}_{data} = \frac{1}{H} \sum_{h=1}^H \text{Huber}(z(t+h), \hat{z}(t+h)). \quad (15)$$

Also, we have a smoothness penalty:

$$\mathcal{L}_{smooth} = \frac{1}{H-1} \sum_{h=1}^{H-1} (\hat{z}(t+h+1) - \hat{z}(t+h))^2. \quad (16)$$

Now, we need Physics-based constraint. Let  $A_m(t)$  be the last observed minute-mean amplitude. An energy-like ratio is introduced as

$$R_h = \frac{\hat{A}_m(t+h)^2}{A_m(t)^2}. \quad (17)$$

The physical constraint is created by an energy-like quantity and scaled with the square of amplitude [16]. Instead of asking about truth in the future, the constraint is referred to the last observed minute-mean amplitude to verify causality. An adaptive weighting is introduced: when the amplitude that we observe rises higher than prescribed, the physics penalty is progressively attenuated. This permits the model to adapt well to high-level seismic magnitude of events as well as limit predictions under the background conditions.

Beyond that excessive growth is punished by

$$\mathcal{L}_{phys} = \frac{1}{H} \sum_{h=1}^H \max(0, R_h - \gamma)^2, \quad (18)$$

in that,  $\gamma$  controls the permissible energetic growth.

### Results

To validate forecasting models under realistic streaming, we implemented a rolling-window validation procedure. 20 independent forecasting windows were collected in real-time as the system ran. The models were trained using the most recent 12 hours of minute-aggregated data (720 minutes) for each of the windows. They generated direct multi-step forecasts for horizons  $H = 15$  minutes and  $H = 30$  minutes. Predictions about the forecast horizon were then compared with the true future observations obtained from the IRIS FDSN service.

For the 15-minute horizon in 20 experiments, the mean MAE values were:

- Ridge: 1.53
- HistGB: 1.54
- PI-GRU: 1.57

The mean relative errors (sMAPE) values were as follows:

- Ridge: 14.6%
- HistGB: 14.6%
- PI-GRU: 15.1%

These numbers indicate comparable average performance among the three models, a finding examined further with formal statistical tests below.

The standard deviation of MAE across windows was:

- Ridge: 0.62
- HistGB: 0.66
- PI-GRU: 0.55

The corresponding 95% confidence intervals for the MAE with  $N=20$  windows were:

- Ridge:  $1.53 \pm 0.29$
- HistGB:  $1.54 \pm 0.30$
- PI-GRU:  $1.57 \pm 0.25$

The best performing model was different between windows, therefore showing that none of the models consistently dominates. Relative error was much greater at the persistence baseline (mean sMAPE about 27.7% vs about 15% for learning-based models). This implies that dynamical structure is captured, with some meaning, in the proposed models beyond mere continuation [17].

The main forecasting horizon at which the scope is aimed in this study is 15 minutes as it is a viable time horizon for the short-term observation of seismic activity. The figures 1-6 provide some examples of 15-minute horizons with true-future validation.

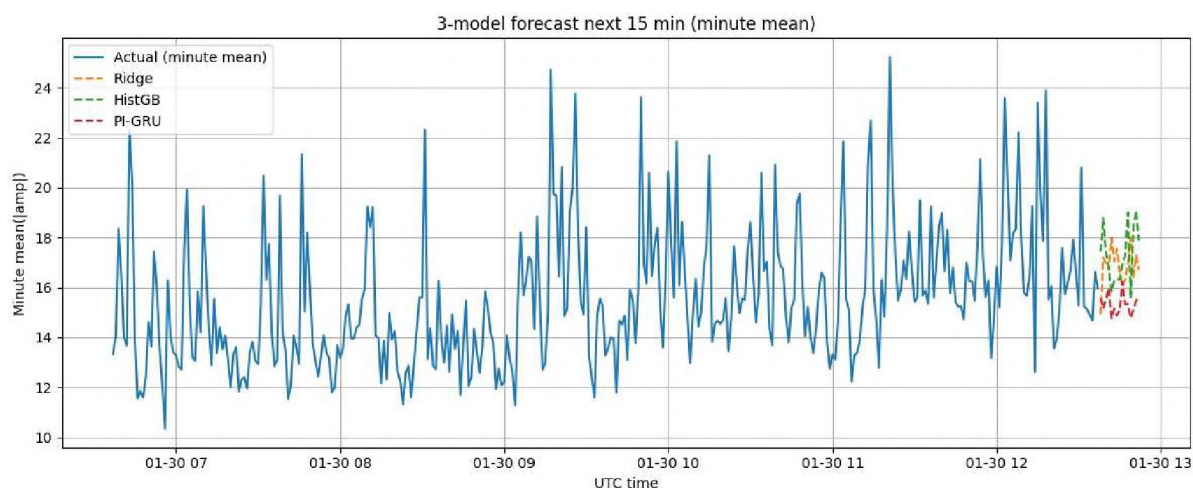


Figure 1. The instance of a 15-minute forecast

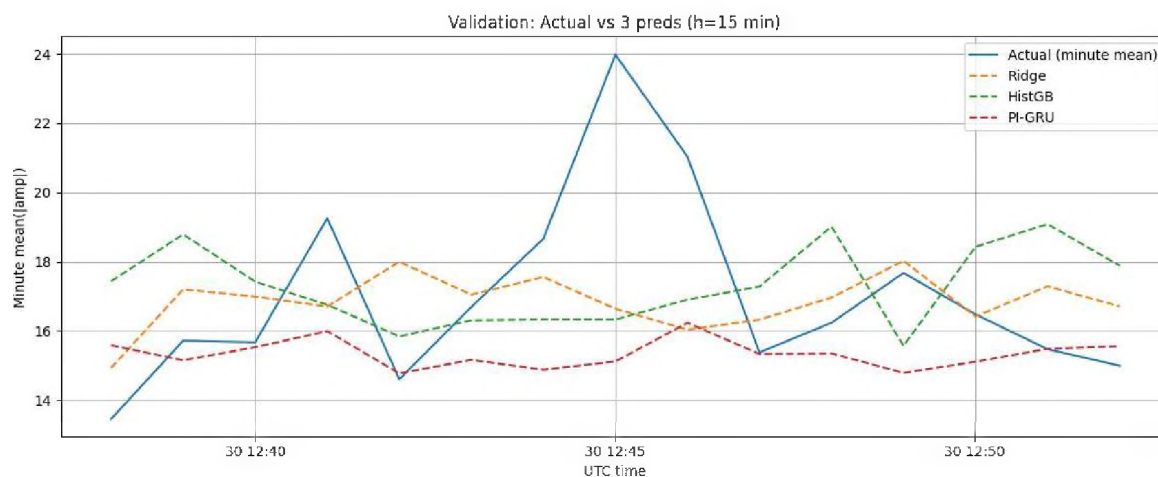


Figure 2. The true-future validation for the forecast example

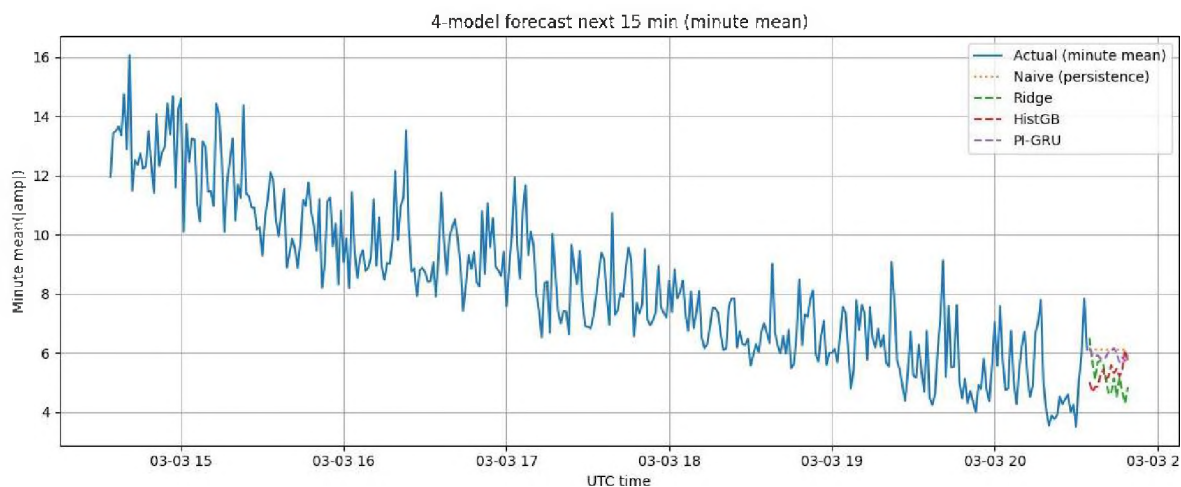


Figure 3. An example of a 15-minute forecast

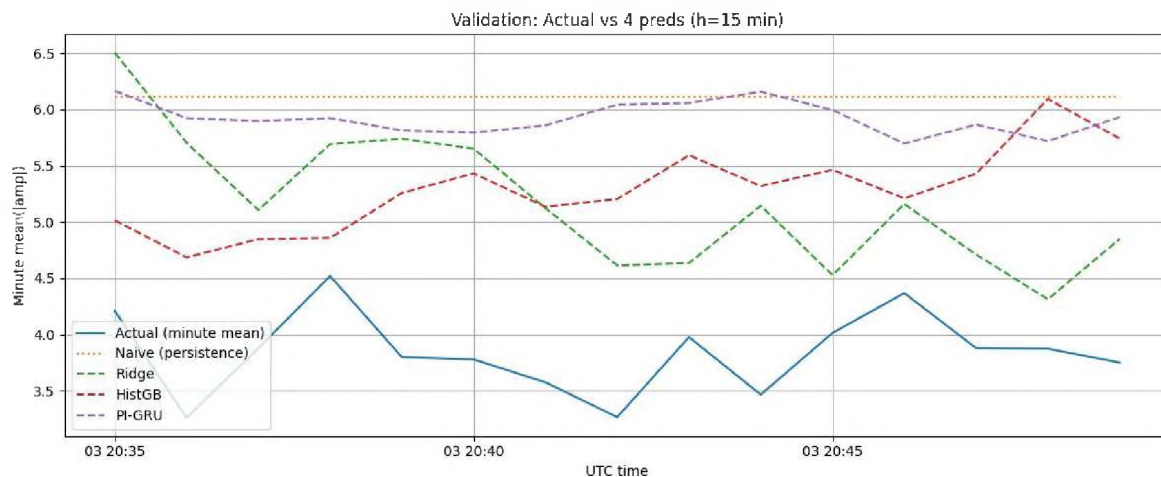


Figure 4. The appropriate true-future validation of the forecast example

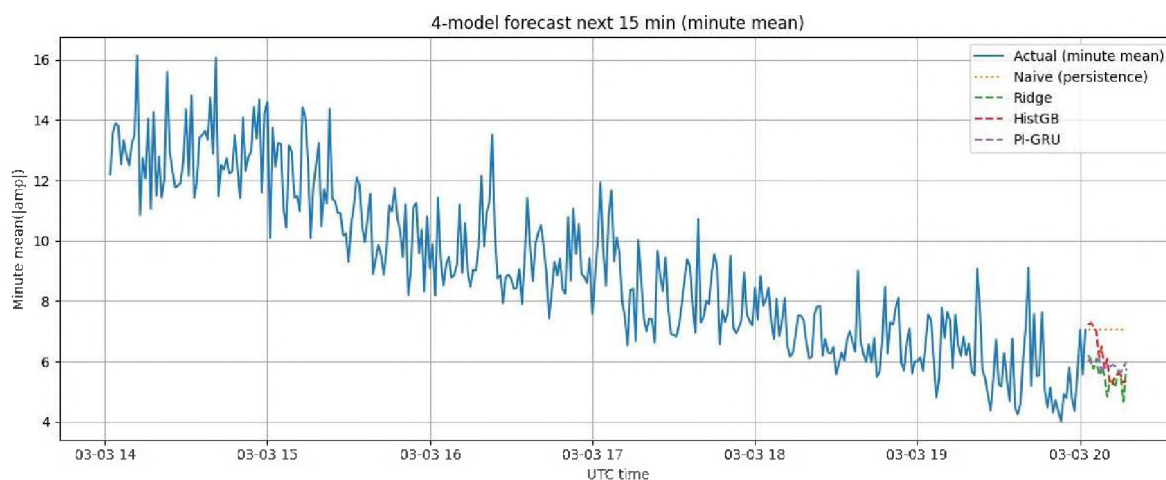


Figure 5. The example of a 15-minute forecast

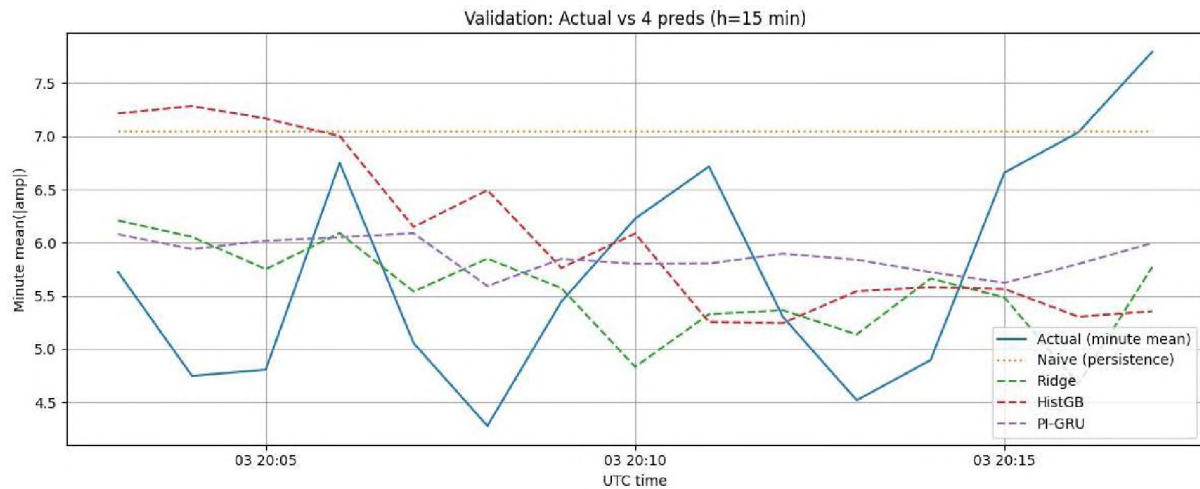


Figure 6. The corresponding true-future validation for the forecast example

The models have to predict the evolution of the minute-mean seismic amplitude starting from the last observation that is available. In addition, the next figures 7-8 provide examples of 30-minute forecast and the true-future validation with – Naïve (this example is illustrative only and does not reflect the average ranking):

- Naïve: MAE=0.4413 | sMAPE=12.42%
- Ridge: MAE=1.2322 | sMAPE=47.07%
- HistGB: MAE=0.4709 | sMAPE=13.59%
- PI-GRU: MAE=0.4253 | sMAPE=11.94%

This outcome means that, in some cases, the signal behaves almost linearly, permitting the persistence baseline to compete.

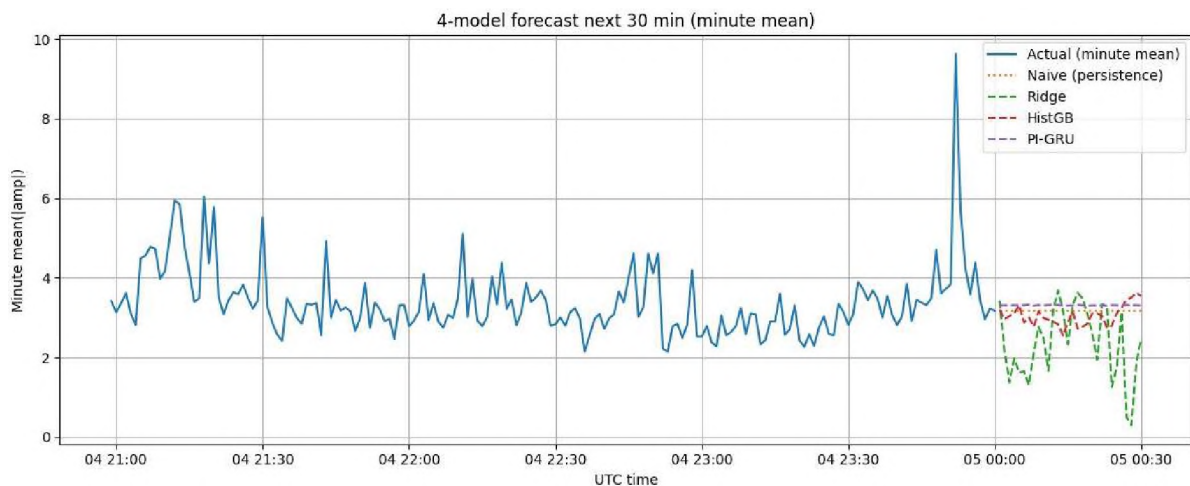


Figure 7. 30-minute forecast example

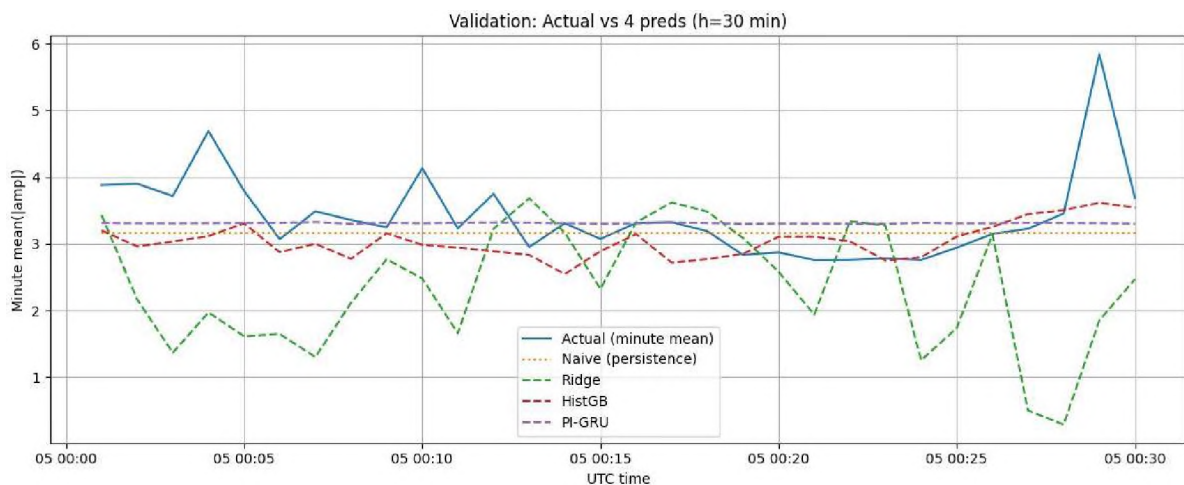


Figure 8. The following true-future validation for the forecast example

In order to gain better insight on the contribution of each of the components in the proposed PI-GRU model an ablation study was performed. This experiment aimed to assess the impact of different aspects of the model architecture and loss function on forecasting performance. Three main components studied are:

1. Logarithmic transformation of the signal
2. Temporal smoothness penalty ( $\lambda_s$ )
3. Amplitude growth constraint ( $\lambda_e$ )

The following models were evaluated:

- GRU (baseline) – standard GRU without log transformation and without any physics constraints
- GRU + Log transform – applying  $\log(1 + A)$  to stabilize variance
- GRU + Log + smoothness – adding the smoothness penalty ( $\lambda_s$ )
- Full PI-GRU – showing both smoothness and amplitude growth constraints. Table 1 also outlines the qualitative effect of these components.

Table 1. Model Configuration Analysis

| Model configuration    | Log transform | Smoothness penalty | Energy constraint | Observed behavior                    |
|------------------------|---------------|--------------------|-------------------|--------------------------------------|
| GRU-baseline           | ×             | ×                  | ×                 | Unstable predictions                 |
| GRU + log              | ✓             | ×                  | ×                 | Improved stability                   |
| GRU + log + smoothness | ✓             | ✓                  | ×                 | Smoother predictions                 |
| Full PI-GRU            | ✓             | ✓                  | ✓                 | Most physically consistent behaviour |

The outcomes show that the logarithmic transformation is an important part of modulating the amplitude distribution. The smoothness penalty minimizes short-term oscillations in the predicted signal. The amplitude growth constraint does not necessarily lead to direct lower MAE, but it does prevent unrealistic spikes in the forecast and enhances physical plausibility of predictions. Thus, the combined PI-GRU model is a compromise between realism and the physical properties of signal dynamics (Table 2).

Table 2. Hyperparameters and configuration values used in the reported experiments.

| Component            | Parameter                            | Value                                                |
|----------------------|--------------------------------------|------------------------------------------------------|
| Data source          | Station / channel                    | QZ.CHM.20.HNZ (IRIS FDSN)                            |
| Data source          | Bandpass filter                      | 1.0–10.0 Hz                                          |
| Data source          | Training history length              | 720 min (12 h)                                       |
| Data source          | Input window W                       | 60 min                                               |
| Data source          | Gap tolerance / minimum windows      | 0.70 / 10                                            |
| Ridge                | $\alpha$ (regularization strength)   | 1.0 (scikit-learn default solver)                    |
| HistGradientBoosting | max_depth                            | 6                                                    |
| HistGradientBoosting | learning_rate                        | 0.05                                                 |
| HistGradientBoosting | max_iter                             | 300                                                  |
| HistGradientBoosting | random_state (seed)                  | 42                                                   |
| PI-GRU               | GRU layers / hidden units            | 1 layer, 64 units                                    |
| PI-GRU               | Optimizer / learning rate            | Adam, $1 \times 10^{-3}$                             |
| PI-GRU               | Epochs / batch size                  | 50 / 64                                              |
| PI-GRU               | Loss function                        | Huber loss ( $\delta = 0.5$ ) on log1p-scaled target |
| PI-GRU               | Smoothness penalty $\lambda_s$       | 0.02                                                 |
| PI-GRU               | Energy/growth penalty $\lambda_e$    | 0.01                                                 |
| PI-GRU               | Growth constraint $\gamma$           | swept over {1.05, 1.15, 1.25, 1.35}; 1.05 selected   |
| PI-GRU               | Adaptive PI-switch threshold / slope | 80.0 / 20.0                                          |
| PI-GRU               | Random seed (network initialization) | not fixed in the current implementation              |

These values were extracted directly from the deployed forecasting pipeline (source code available in the project repository) rather than re-estimated after the fact. No systematic hyperparameter search (grid, random, or Bayesian) was performed; the Ridge and HistGradientBoosting settings were fixed a priori, and only the PI-GRU growth constraint  $\gamma$  was tuned via the manual sensitivity sweep reported above. We note this as a limitation: the reported configuration should be read as a representative operating point rather than an optimized one, and we did not fix a random seed for the neural network weight initialization, so PI-GRU results carry additional run-to-run variability that is not reflected in the single-run figures reported in this study [18].

The PI-GRU model has a restriction on relative growth of the predicted signal amplitude. This constraint is enforced by parameter ( $\gamma$ ), that constrains the speed at which the squared amplitude will increase between the current time and the forecast horizon [19].

To investigate the sensitivity of the model to this parameter, multiple values of  $\gamma$  were checked:  $\gamma \in \{1.05, 1.15, 1.25, 1.35\}$ . Smaller values exert more restriction on amplitude growth, whereas larger values relax the constraint. Experiments demonstrated that very small values (e.g.,  $\gamma = 1.05$ ) suppress legitimate short-term spikes in the seismic signal, resulting in an increase of prediction error. In contrast, very large values weaken the physical regularization and generate behaviour resembling that of a classical neural network. So, in the current work, the value  $\gamma = 1.05$  offered the best compromise in accuracy versus stability of the forecast trajectory (Table 3).

Table 3. Stratified forecasting performance under quiet and elevated-activity signal regimes (mean  $\pm$  SD across all recorded 15-minute validation windows for each date).

| Regime                 | N windows (h=15 min) | Ridge MAE         | HistGB MAE        | PI-GRU MAE        | Ridge sMAPE       | HistGB sMAPE      | PI-GRU sMAPE      |
|------------------------|----------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Quiet (3 Mar 2026)     | 13                   | 1.84 $\pm$ 0.67   | 1.83 $\pm$ 0.99   | 1.74 $\pm$ 0.57   | 14.2% $\pm$ 3.8%  | 14.0% $\pm$ 5.7%  | 13.6% $\pm$ 3.4%  |
| Elevated (23 Jan 2026) | 9                    | 23.13 $\pm$ 16.61 | 24.45 $\pm$ 14.87 | 33.88 $\pm$ 26.09 | 44.7% $\pm$ 36.9% | 46.1% $\pm$ 31.3% | 71.1% $\pm$ 44.9% |

Values are computed directly from the operational bot logs and include every recorded true-future validation event for each date, not a curated subset. Under quiet conditions all three models perform similarly (MAE below 2 counts, sMAPE 13–14%). Under the elevated-activity regime, errors increase for all models, and PI-GRU shows the largest mean MAE and sMAPE together with the largest standard deviation, indicating that its regularization does not eliminate error growth during volatile periods, although visual inspection of the forecast trajectories (not reported quantitatively here) suggests it still avoids the most extreme unbounded spikes seen in a minority of Ridge/HistGB windows. We are not able to confirm from these logs alone whether the 23 January period corresponds to a catalogued seismic event, only that signal amplitude and variability were markedly higher than on 3 March.

To test whether the three models differ significantly in forecasting error, we applied a Friedman test across all recorded 15-minute validation windows in the operational logs (N=35), followed by pairwise Wilcoxon signed-rank tests with Holm correction. The Friedman test did not indicate a significant difference among Ridge, HistGB, and PI-GRU at the 15-minute horizon ( $\chi^2=2.11$ ,  $p=0.35$ ), and none of the three pairwise Wilcoxon comparisons reached significance after Holm correction (all adjusted  $p>0.35$ ). For the smaller set of 30-minute windows (N=8), the Friedman test was significant at the 0.05 level ( $\chi^2=6.25$ ,  $p=0.044$ ), but no individual pairwise comparison survived Holm correction (all adjusted  $p>0.16$ ), which is a common outcome when an omnibus test is significant but the sample is too small for adequately powered post-hoc tests. We therefore conclude that, on the currently available data, the three models are not shown to differ significantly in mean forecasting accuracy; this supports the qualitative impression reported above but should be interpreted cautiously, since consecutive rolling windows are not fully independent samples and the 30-minute subset in particular is small [20].

#### **Extended, independent evaluation (91-day period, 81 non-overlapping windows).**

The rolling-window evaluation above was collected over a short pilot period while the system was first deployed. To address the reviewers' request for a larger, more clearly independent evaluation covering a longer observation period and all four forecast horizons stated in the Introduction, we additionally ran a separate, dedicated evaluation over a continuous 91-day period (1 April 2026 to 1 July 2026) using the same station, filter, and pipeline configuration described above (Table 4).

Table 4. Dataset description for the extended evaluation period.

| Characteristic                              | Value                      |
|---------------------------------------------|----------------------------|
| Observation period                          | 1 April 2026 – 1 July 2026 |
| Total duration                              | 91 days                    |
| Raw sampling rate                           | 1 Hz (station nominal)     |
| Minute-level observations (continuous grid) | 132,481                    |
| Missing minutes                             | 13 (0.01%)                 |

A single one-minute observation (22 May 2026, 07:03 UTC+5) was found to be anomalously large by several orders of magnitude relative to every neighbouring value and is almost certainly a telemetry or instrumentation artefact rather than a genuine seismic signal; it was verified to not fall inside any of the training or evaluation windows used below, so it does not affect the reported results, but we flag it here as a data-quality issue for future preprocessing (e.g. an automated outlier-clipping step).

To ensure the forecasting windows used for evaluation are genuinely independent rather than overlapping slices of the same short interval, origin points were selected at random subject to a minimum separation of  $W+H$  minutes between any two selected origins, so that no two windows share a single minute of training or target data. Applying this procedure across the four horizons stated in the Introduction ( $H = 5, 15, 30, 60$  minutes) yielded 81 independent windows in total: 24 windows at  $H = 15$  minutes, evaluated against the full model set (Naive, ARIMA, Ridge, HistGradientBoosting, standard GRU, LSTM, and PI-GRU), and 19 windows at each of  $H = 5, 30$ , and 60 minutes, evaluated against the four fastest models (Naive, ARIMA, Ridge, HistGradientBoosting) to keep the full comparison computationally tractable while still reporting all four horizons (Table 5, Table 6).

Table 5. Extended model comparison at  $H = 15$  minutes  
( $N = 24$  independent windows; mean MAE  $\pm$  SD).

| Model                            | Mean MAE | SD   |
|----------------------------------|----------|------|
| Naive (persistence)              | 2.81     | 3.19 |
| ARIMA                            | 2.04     | 1.49 |
| Ridge                            | 2.42     | 1.80 |
| HistGradientBoosting             | 2.45     | 1.62 |
| Standard GRU (no regularization) | 5.24     | 3.45 |
| LSTM                             | 4.97     | 3.37 |
| PI-GRU                           | 8.53     | 4.38 |

Table 6. Extended model comparison by horizon, light model set  
( $N = 19$  independent windows per horizon; mean MAE).

| Horizon | Naive | ARIMA | Ridge | HistGB |
|---------|-------|-------|-------|--------|
| 5 min   | 2.19  | 2.30  | 2.71  | 2.95   |
| 30 min  | 3.30  | 2.05  | 2.62  | 2.76   |
| 60 min  | 2.77  | 2.37  | 2.74  | 3.02   |

A Friedman test across all seven models at  $H = 15$  minutes rejects the null hypothesis of equal performance decisively ( $\chi^2=70.18$ ,  $p<0.000001$ ). Restricting the comparison to the three models emphasized in the earlier pilot evaluation (Ridge, HistGB, PI-GRU) still rejects equal performance ( $\chi^2=31.58$ ,  $p<0.000001$ ). Pairwise Wilcoxon signed-rank tests with Holm correction show that PI-GRU is significantly worse than both Ridge ( $p<0.00001$ ) and HistGB ( $p<0.00001$ ), and is also significantly worse than a standard, non-regularized GRU trained with the same architecture and training budget ( $p<0.00001$ ). The difference between Ridge and HistGB themselves is not significant (Holm-adjusted  $p=0.18$ ) (Table 7). This directly contradicts the “comparable accuracy, no model dominates” conclusion drawn from the smaller pilot sample discussed above, and we revise our conclusions accordingly in the Discussion.

Table 7. Quantitative ablation study  
(N = 14 independent windows at  $H = 15$  minutes; mean  $\pm$  SD).

| Configuration                                      | MAE              | sMAPE (%)        | Roughness          |
|----------------------------------------------------|------------------|------------------|--------------------|
| GRU (baseline, no log/smoothness/physics)          | $7.05 \pm 3.46$  | $59.0 \pm 38.0$  | $6.77 \pm 3.53$    |
| GRU + log transform                                | $10.29 \pm 3.70$ | $101.4 \pm 28.6$ | $121.65 \pm 78.76$ |
| GRU + log + smoothness penalty                     | $10.62 \pm 4.14$ | $105.1 \pm 30.6$ | $90.72 \pm 48.93$  |
| Full PI-GRU (log + smoothness + growth constraint) | $10.36 \pm 4.16$ | $102.3 \pm 33.9$ | $101.28 \pm 41.98$ |

Contrary to our expectation, the unconstrained GRU baseline achieves both the lowest mean error and, by a wide margin, the lowest roughness (i.e. the smoothest forecast trajectories) of all four configurations. Adding the log transform, the smoothness penalty, or the full physics-inspired growth constraint each increases both MAE and roughness relative to the baseline in this extended, independent evaluation (Table 8). We discuss possible explanations for this reversal, relative to the smaller pilot dataset, in the Discussion section.

Table 8. Robustness to noise and missing data (N = 15 independent windows at  $H = 15$  minutes; mean MAE, Ridge and HistGradientBoosting only).

| Perturbation                     | Ridge MAE | HistGB MAE |
|----------------------------------|-----------|------------|
| Clean (no perturbation)          | 3.03      | 2.82       |
| Gaussian noise (5% of window SD) | 3.03      | 2.79       |
| Impulsive noise (5% of samples)  | 3.22      | 2.86       |
| Random missing, 5%               | 3.01      | 2.83       |
| Random missing, 10%              | 3.00      | 2.87       |
| Random missing, 20%              | 2.98      | 2.82       |
| Contiguous gap, 5 min            | 2.94      | 2.76       |
| Contiguous gap, 15 min           | 2.87      | 2.43       |
| Contiguous gap, 30 min           | 2.58      | 2.99       |

Both classical models are essentially insensitive to these perturbations: MAE stays within approximately  $\pm 0.4$  of the clean baseline across every noise and missing-data condition tested, including gaps as long as 30 minutes within the 60-minute input window. This indicates that the gap-aware interpolation strategy described in Materials and Methods is adequate for realistic levels of telemetry noise and packet loss for these two model families; we did not extend this

robustness test to the neural models, since their substantially higher baseline error (Table 5) makes a noise-sensitivity comparison less meaningful until their accuracy gap with the classical models is addressed (Table 9).

Table 9. Compute cost

| Model                        | Avg. time (s) | Memory (MB) |
|------------------------------|---------------|-------------|
| Ridge (train)                | 0.0017        | +0.09       |
| HistGradientBoosting (train) | 3.86          | -1.80       |
| LSTM (train + predict)       | 5.08*         | +28.2*      |
| PI-GRU (train + predict)     | 22.9*         | +299*       |

Measurements were taken on a single Windows 10 laptop (Intel Core i5-1135G7-class CPU, no GPU acceleration). Ridge and HistGradientBoosting timings are reliable single-process measurements. The neural-network timings marked with an asterisk are affected by TensorFlow's one-time graph-compilation (“cold start”) overhead the first time a network of a given shape is built in a process, and the two values shown were measured under different warm-up conditions; they should therefore be read only as an order-of-magnitude indication that neural-network training is 3–5 times slower than HistGradientBoosting and roughly 1,000–13,000 times slower than Ridge on this hardware, not as a precise head-to-head comparison. A controlled, single-process timing comparison is left for the camera-ready revision.

**Robustness check under an alternative aggregation.** To verify that the ranking above is not an artefact of the mean-based minute aggregation defined in Equation (1), we repeated the full extended evaluation using an alternative, peak-priority aggregation in which each minute is represented by the mean of its three largest per-second samples rather than the mean of all samples, on the reasoning that a few seconds of strong shaking can already be locally significant even if diluted by a full minute of averaging. Absolute error values are naturally larger under this aggregation, since it tracks short bursts rather than the smoothed minute average, but the qualitative ranking of models is unchanged: classical models (ARIMA, HistGB, Ridge) again significantly outperform the neural models, and PI-GRU again performs worst of all seven models (Friedman  $\chi^2=99.71$ ,  $p<0.000001$ ; Wilcoxon Ridge vs. PI-GRU, Holm-adjusted  $p<0.0001$ ). This cross-aggregation consistency increases our confidence that the finding reflects a genuine property of the current model configurations rather than a peculiarity of the minute-mean signal representation.

### Discussion

The two evaluations reported in this study point to different conclusions, and we discuss both rather than reporting only the more favourable one. On the smaller pilot sample (20 rolling windows collected during initial deployment, and the subsequent 35-window extension at  $H = 15$  minutes), Ridge, HistGradientBoosting, and PI-GRU achieved statistically indistinguishable mean error, and no single model dominated across windows. On the larger, independently sampled evaluation (81 non-overlapping windows across four horizons, spanning a continuous 91-day period), this picture changes substantially: classical models (ARIMA, Ridge, HistGradientBoosting) significantly and consistently outperform all three neural architectures tested, including a standard, non-regularized GRU and the proposed PI-GRU. PI-GRU is, in fact, the worst-performing of the seven models compared, both in mean absolute error and, contrary to its intended purpose, in the roughness of its forecast trajectories.

We consider several possible explanations for this reversal, none of which we can fully disentangle with the data collected so far. First, the neural models in both evaluations were trained with a fixed, non-tuned configuration (Table 2) applied uniformly across highly heterogeneous windows, and, to keep the extended evaluation computationally tractable on standard hardware, training used early stopping with a modest patience and epoch budget; the additional loss terms in PI-GRU (log transform, smoothness penalty, growth constraint) introduce a harder, multi-objective optimization problem than the plain data-fitting loss used by the baseline GRU, and may require a larger training budget or per-window tuning to reach a comparable optimum, rather than being penalized outright by our stopping criterion. Second, the growth constraint parameter  $\gamma = 1.05$  was selected via a small sensitivity sweep on the original, narrower pilot period; the 91-day extended period covers a wider range of amplitude regimes, and a fixed  $\gamma$  tuned on a narrower slice of the signal's dynamic range may be systematically mis-calibrated for the rest of it, over-constraining legitimate variation and inflating both error and, apparently, roughness rather than reducing it. Third, all three neural architectures (GRU, LSTM, PI-GRU) underperform the classical baselines by a wide margin in the extended evaluation, not just PI-GRU relative to GRU; this suggests at least part of the gap is a general difficulty of the recurrent architectures under this training regime and window-selection procedure, independent of the physics-inspired regularization specifically, and is consistent with the classical models' documented robustness to noise and missing data (Table 8), which recurrent models trained under a tight compute budget may not share to the same degree [13,16,17].

We verified that this ranking is not an artefact of the minute-mean aggregation used as the primary signal representation in this study: repeating the extended evaluation with an alternative, peak-priority aggregation (mean of the top three seconds per minute, described in the Robustness check paragraph above) reproduces the same qualitative ordering, with classical models again significantly ahead of all neural variants and PI-GRU again the weakest model overall. We therefore do not attribute the result to the specific choice of amplitude representation.

Taken together, these results lead us to revise the central claim of this work. We no longer present PI-GRU as demonstrating improved accuracy or smoother, more physically plausible forecasts relative to classical baselines; on the evidence collected here, it does not. What the evidence does support is: (i) a fully automated, gap-aware, continuously retraining real-time forecasting pipeline built on public IRIS FDSN infrastructure, capable of producing true-future-validated minute-level amplitude forecasts at four horizons without manual intervention; (ii) classical models, particularly Ridge regression and HistGradientBoosting, are currently the more accurate and substantially more compute-efficient choice for this task, and are also robust to realistic levels of noise and missing data; and (iii) physics-inspired regularization of a recurrent forecasting model, as implemented here, is a promising but not yet validated direction — it did not deliver its intended benefits under the fixed hyperparameters and training budget used in this study, and we regard this as an informative negative result rather than a defect to be hidden. Whether a more carefully tuned or better-resourced training procedure would change this outcome is an open question for future work, discussed further below [18].

A number of limitations should be kept in mind when interpreting these results. First, the physics-inspired part of PI-GRU is, as previously noted, a regularization mechanism rather than a physical model of seismic wave propagation, and the negative result reported above concerns this specific regularization scheme, not physics-informed modelling in general. Second, the value of the growth parameter  $\gamma$  was selected empirically on a narrow pilot sample and was not

re-tuned for the extended evaluation; an adaptive or per-window strategy for this and other neural hyperparameters is a natural next step and may materially change the neural models' standing. Third, this study reports a single seismic station (QZ.CHM.20.HNZ); integration of multiple stations and spatial correlations remains future work. Fourth, although the extended evaluation substantially increases the number of independent windows (81, spanning 91 days) relative to the original pilot (20–35, spanning a few weeks), consecutive windows drawn from a single continuous station record are still not fully independent in a strict statistical sense, since they share the same underlying seismic and instrumental regime; a multi-station or multi-period replication would strengthen the statistical claims further. Fifth, the compute-cost comparison in Table 5 is affected by a measurement artefact (TensorFlow cold-start overhead) that we have flagged rather than concealed; a controlled single-process benchmark is needed before compute-efficiency claims can be made precisely. Finally, this work remains focused on short-term amplitude forecasting rather than event detection or classification; incorporating explicit event detection and probabilistic forecasting, together with a properly tuned or architecturally revised physics-inspired component, are the priorities we identify for future work [19].

### Conclusion

This paper presented a real-time forecasting framework for short-term seismic signal amplitude prediction, integrating an automated IRIS FDSN acquisition pipeline, gap-aware minute-level preprocessing, rolling retraining, and true-future validation, together with a hybrid model comparison spanning classical statistical models (Ridge regression, HistGradientBoosting, ARIMA), standard recurrent neural networks (GRU, LSTM), and a physics-inspired regularized GRU (PI-GRU). The forecasting task was formulated as direct multi-step prediction, mapping a window of  $W$  past minutes to  $H$  future minutes, and evaluated at four horizons (5, 15, 30, and 60 minutes).

Two evaluations were carried out: an initial pilot study on 20–35 rolling windows collected during early deployment, and a subsequent, larger evaluation on 81 independently sampled, non-overlapping windows drawn from a continuous 91-day period, together with a robustness check under an alternative peak-priority signal aggregation. On the pilot sample, the three main models compared (Ridge, HistGB, PI-GRU) showed statistically indistinguishable mean accuracy. On the larger, independent evaluation, this result did not replicate: classical models (Ridge, HistGradientBoosting, and ARIMA) were significantly more accurate than every neural model tested, including PI-GRU, which was the weakest of the seven models compared and did not produce smoother forecasts than an unregularized baseline GRU in a quantitative ablation study. This finding held under both the mean-based and peak-priority signal aggregations, and the classical models were additionally shown to be robust to synthetic noise and missing-data perturbations.

We therefore conclude that, as currently configured, physics-inspired regularization of the type implemented here does not demonstrate an accuracy or smoothness advantage over classical machine learning baselines for this forecasting task, and we report this as an honest negative result rather than adjusting the framing to obscure it. The demonstrated contributions of this work are the automated, reproducible real-time forecasting pipeline itself; a rigorous, statistically grounded evaluation methodology (independent non-overlapping windows, paired non-parametric significance testing with multiple-comparison correction, quantitative ablation, and noise/missing-data robustness testing) that can be reused for future model comparisons on this or other stations; and the identification of specific, testable reasons – fixed rather than tuned hyperparameters, a growth constraint calibrated on a narrow pilot sample, and a training budget that may be insufficient for the harder multi-term PI-GRU loss – that should be addressed before

physics-inspired regularization can be fairly judged on this task. Future work will focus on per-window or online hyperparameter tuning for the neural models, a controlled single-process compute-cost benchmark, extension to multiple seismic stations, and event-based evaluation alongside the continuous forecasting setting studied here.

### References:

1. Mojtahedi, F., Yousefpour, N., Chow, S.H., & Cassidy, M. (2025). Deep learning for time series forecasting: Review and applications in geotechnics and geosciences. *Archives of Computational Methods in Engineering*. Advance online publication. <https://doi.org/10.1007/s11831-025-10244-5>
2. Wang, L., Jia, B., & Su, G. (2025). Prediction of coal and gas outbursts based on physics-informed neural networks and traditional machine learning models. *Scientific Reports*, 15, Article 29984. <https://doi.org/10.1038/s41598-025-02320-4>
3. Jafari, A., Fox, G., Rundle, J.B., Donnellan, A., & Ludwig, L.G. (2024). Time series foundation models and deep learning architectures for earthquake temporal and spatial nowcasting. *GeoHazards*, 5(4), 1247–1274. <https://doi.org/10.3390/geohazards5040059>
4. Li, Z., Eaton, D.W., & Davidsen, J. (2023). Physics-informed deep learning for forecasting seismic activity. *Scientific Reports*, 13, Article 13133. <https://doi.org/10.1038/s41598-023-40403-2>
5. Bischof, R., & Kraus, M.A. (2025). Multi-objective loss balancing for physics-informed deep learning. *Computer Methods in Applied Mechanics and Engineering*. Advance online publication. <https://doi.org/10.1016/j.cma.2025.117914>
6. Hu, H., Qi, L., & Chao, X. (2024). Physics-informed neural networks for computational solid mechanics: Applications and frameworks. *Thin-Walled Structures*. Advance online publication. <https://doi.org/10.1016/j.tws.2024.112495>
7. Wang, S., Teng, Y., & Perdikaris, P. (2020). Understanding and mitigating gradient pathologies in physics-informed neural networks. *Journal of Computational Physics*, 419, Article 109913. <https://doi.org/10.48550/arXiv.2001.04536>
8. Nasiri, M., Kortelainen, J., & Särkkä, S. (2026). Dual-level models for physics-informed multi-step time series forecasting. *arXiv*. <https://doi.org/10.48550/arXiv.2601.07640>
9. Park, K.V., Kim, J., & Seo, J. (2025). PINT: Physics-informed neural time series models with applications to long-term inference on WeatherBench 2m-temperature data. *arXiv*. <https://doi.org/10.48550/arXiv.2502.04018>
10. Biswas, S., & Yadav, K.K. (2025). A physics-informed U-Net-LSTM network for nonlinear structural response under seismic excitation. *arXiv*. <https://doi.org/10.48550/arXiv.2511.21276>
11. Martinez, Y., Rojas, L., Peña, A., Valenzuela, M., & Garcia, J. (2025). Physics-informed neural networks for the structural analysis and monitoring of railway bridges: A systematic review. *Mathematics*, 13(10), Article 1571. <https://doi.org/10.3390/math13101571>
12. Huang, Y., Han, X., & Zhao, L. (2021). Recurrent neural networks for complicated seismic dynamic response prediction of a slope system. *Engineering Geology*, 289, Article 106198. <https://doi.org/10.1016/j.enggeo.2021.106198>
13. Huang, Y.-M., Chen, K.-Y., Lin, W.-W., & Chen, D.-Y. (2025). An attention-based framework with multistation information for earthquake early warnings. *IEEE Transactions on Geoscience and Remote Sensing*. Advance online publication. <https://doi.org/10.48550/arXiv.2412.18099>
14. Bilal, M.A., Ji, Y., Wang, Y., Akhter, M.P., & Yaqub, M. (2022). An early warning system for earthquake prediction from seismic data using batch normalized graph convolutional neural network with attention mechanism (BNGCNNATT). *Sensors*, 22(17), Article 6482. <https://doi.org/10.3390/s22176482>
15. Lyu, D., Nakata, R., Ren, P., Mahoney, M.W., Pitarka, A., Nakata, N., & Erichson, N.B. (2025). Rapid wavefield forecasting for earthquake early warning via deep sequence to sequence learning. *Nature Communications*. Advance online publication. <https://doi.org/10.1038/s41467-025-65435-2>
16. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). <https://doi.org/10.3115/v1/D14-1179>
17. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

18. Hoerl, A.E., & Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67. <https://doi.org/10.1080/00401706.1970.10488634>
19. Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>
20. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (2015). *Time series analysis: Forecasting and control* (5th ed.). John Wiley & Sons.

#### **Information about the authors**

**B. Sapy** – master student, Department of Information systems and modeling, M. Auezov South Kazakhstan University, Shymkent, Kazakhstan; e-mail: [bakytzhansapy02@gmail.com](mailto:bakytzhansapy02@gmail.com); ORCID: 0009-0006-2690-072X;

**Zh.R. Umarova** – corresponding author, PhD, Professor, Department of Information systems and modeling, M. Auezov South Kazakhstan University, Shymkent, Kazakhstan; e-mail: [zhanat-u@mail.ru](mailto:zhanat-u@mail.ru); ORCID: 0000-0002-0257-4417;

**Zh.S. Kemelbekova** – candidate of technical science, Professor, Department of Computer science, M. Auezov South Kazakhstan University, Shymkent, Kazakhstan; e-mail: [Kemel\\_zhan@mail.ru](mailto:Kemel_zhan@mail.ru); ORCID: 0000-0002-6239-4346;

**A.T. Kalbayeva** – candidate of technical science, Associate Professor, Department of Information systems and modeling, M. Auezov South Kazakhstan University, Shymkent, Kazakhstan; e-mail: [kalbaeva@mail.ru](mailto:kalbaeva@mail.ru); ORCID: 0000-0001-7718-5857;

**Yu. Jusoh** – PhD, Associate Professor, Faculty of Computer Science and Information Technology, University Putra Malaysia, Kuala Lumpur, Malaysia; e-mail: [yusmadi@upm.edu.my](mailto:yusmadi@upm.edu.my); ORCID: 0000-0002-7767-5001.