Self-adaptive Extreme Penalized Loss for Imbalanced Time Series Prediction

Yiyang Wang[1,*], **Yuchen Han**[1], **Yuhan Guo**[2]
[1] College of Artificial Intelligence, Dalian Maritime University
[2]College of Transportation Engineering, Dalian Maritime University
{yywerica, han_yuchen, guoyuhan}@dlmu.edu.cn

Abstract

Forecasting time series in imbalanced data presents a significant research challenge that requires considerable attention. Although there are specialized techniques available to tackle imbalanced time series prediction, existing approaches tend to prioritize extreme predictions at the expense of compromising the forecasting accuracy of normal samples. We in this paper propose an extreme penalized loss function that relaxes the constraint on overestimating extreme events, thereby imposing great penalties on both normal and underestimating extreme events. In addition, we provide a self-adaptive way for setting the hyperparameters of the loss function. Then, both the proposed loss function and an attention module are integrated with LSTM networks in a decomposition-based framework. Extensive experiments conducted on real-world datasets demonstrate the superiority of our framework compared to other state-of-the-art approaches for both time series prediction and block maxima prediction tasks.

1 Introduction

Time series prediction plays an important role in various practical fields, including but not limited to meteorological forecasting, electricity demand prediction, and stock prices monitoring. In most cases, the collected data exhibit an imbalance, where unusual extreme events of extremely high or low values occur within the usual range of the time series. There is a growing emphasis on extreme events as they offer valuable insights for anticipating the most severe scenarios projected to occur within the forecast period. In this paper, our goal is to precisely predict extreme events in time series data without compromising the accuracy of predicting normal events.

Recent advancements for addressing time series prediction techniques have shown significant progress, shifting from statistical and machine learning techniques [Box *et al.*, 2015; Sapankevych and Sankar, 2009; Tyralis and Papacharalampous, 2017] to the utilization of deep learning methods, particularly with a focus on recurrent neural networks [Connor *et al.*, 1994; Zhou *et al.*, 2022; Siami-Namini *et al.*, 2019; Weerakody *et al.*, 2021]. However, most time series forecasting models fall short in accurately predicting extreme events at their actual severity primarily due to the utilization of simplistic loss functions such as mean square error during model training. Such loss functions prioritize the estimation of the conditional expectation of the target variable, which remains relatively unaffected even in cases where predictions for extreme event are inaccurate. Hence, they are likely to result in an underestimation of extreme event values located at the tail end of the distribution.

In recent years, this problem has received significant attention from not a few researchers. Plenty of efforts have been undertaken to precisely predict extreme events in time series, which can be classified into two categories based on the objects being predicted. One category involves forecasting the sequence of values within a future time window. The majority of research efforts in this field concentrate on the direct development of the loss function, with a particular emphasis on extreme events [Ding *et al.*, 2019; Polson and Sokolov, 2020; Wilson *et al.*, 2022a; Liu *et al.*, 2023]. Nevertheless, they usually improve the precision of extreme predictions at the expense of compromising the accuracy of normal events. There are also approaches that focus on designing classifiers to initially detect extremes and subsequently develop distinct predictors for both normal and extreme events [Hou *et al.*, 2021; Rajbhandari *et al.*, 2021; Li *et al.*, 2023]. However, the misjudgements in identifying extreme events may lead to significant errors. Besides, there is also research dedicated to forecasting the likelihood of extreme events [Laptev *et al.*, 2017; Zhu and Laptev, 2017], which has a distinct objective compared to forecasting the value itself. While the other category is concerned with predicting the maximum value within a prescribed time window, representing a variant of the time series prediction problem [Galib *et al.*, 2022; Galib *et al.*, 2023].

In this study, our main focus lies in the domain of time series prediction, with a specific objective to precisely forecast sequence of values within a predetermined forecasting period. In order to improve the predictive accuracy of commonly underestimated extreme events while preserving the prediction precision of normal events, we propose a novel approach involving a tripartite division in a proposed extreme penalized loss function instead of a bipartite one. Specifically, our main contributions can be summarized as follows.

1. A novel self-adaptive extreme penalized loss function is

*Corresponding author.

proposed to tackle the challenge of imbalanced time series prediction. By appropriately relaxing the constraint on overestimating extreme events, our proposed loss imposes significant penalties on both the normal and under-predicted extreme events, thereby enhancing the overall precision of predictions. Besides, the hyper-parameters of the proposed loss function are adaptively determined based on the provided data and task requirements.

2. A decomposition-based framework is carefully designed to enhance data predictability by reducing the disparity between extreme and normal samples through the elimination of long-term trends. Both the proposed loss and an attention module are integrated with long short-term memory (LSTM) networks within the framework to effectively capture the temporal characteristics of specific decomposition components.
3. Extensive experiments are carried out using real-world datasets, demonstrating the superior performance of our complete framework in accurately forecasting both normal and extreme events. Our results outperformed other state-of-the-art approaches for the tasks of both time series prediction and block maxima prediction.

2 Related Work

The methodologies for time series prediction typically derive from statistical techniques and machine learning approaches, involving methods such as the autoregressive integrated moving average model (ARIMA) [Box *et al.*, 2015], support vector machine (SVM) [Sapankevych and Sankar, 2009], extra-trees [Geurts *et al.*, 2006], random forest (RF) [Tyralis and Papacharalampous, 2017], among others. In recent years, deep neural networks have been widely considered in time series prediction owing to their remarkable ability for capturing intricate non-linear relationships. Among those deep learning networks, recurrent neural networks (RNNs) [Siami-Namini *et al.*, 2018], such as LSTMs [Karevan and Suykens, 2020], bidirectional LSTMs (BiLSTMs) [Siami-Namini *et al.*, 2019], gated recurrent networks [Weerakody *et al.*, 2021], transformer-based models [Wu *et al.*, 2021; Zhou *et al.*, 2021; Zhou *et al.*, 2022] demonstrate exceptional performance in forecasting time series data owing to their inherent ability to capture long-term dependencies. These approaches demonstrate high efficacy in predicting normal events in time series data. However, their performance is significantly compromised when it comes to forecasting extreme events due to the limited availability of extreme event data during training.

Since deep learning networks are likely to underestimate extreme events as a result of the significant data imbalance, specialized techniques are required for extreme event prediction. Inspired by the principle of extreme value theory (EVT), the authors in [Ding *et al.*, 2019] have introduced a novel loss function, namely extreme value loss (EVL), with the aim of enhancing prediction accuracy for extreme events. Then, the authors in [Wilson *et al.*, 2022a] further extend the EVL for modeling geospatio-temporal extreme events. Moreover, the hyperparameters of their loss are not user-defined but rather acquired through data-driven learning. Then, the authors additionally utilize EVT and propose a novel loss function tailored to effectively capture zero-inflated spatiotemporal data [Wilson *et al.*, 2022b]. However, these approaches primarily emphasize the fitting of extreme value distributions and often neglect the comprehensive analysis of global data.

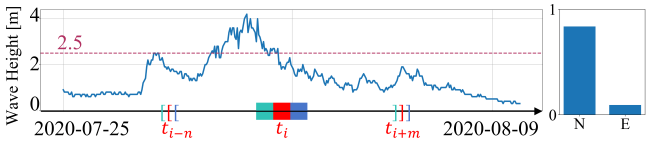

Figure 1: Visualization of imbalanced time series data. The red dotted line represents the warning threshold of wave height, adhering to the regulations of the World Meteorological Organization (WMO). Then, with the warning line, the histogram includes the ratio of normal events (denoted as "N") to extreme events (denoted as "E").

There exist alternative methodologies that aim to improve the prediction of extreme events without depending on EVT. The authors of [Liu *et al.*, 2023] propose a decomposition-based framework and propose an externally triggered loss to improve the accuracy of extreme event prediction. However, in the case of most time series, the external factors influencing the extreme events often pose challenges in effectively guiding the triggered loss. The authors in [Li *et al.*, 2023] propose a novel approach that concurrently learns extreme and normal predictors, while also introducing a selective back propagation mechanism for decision-making among them. However, there is a significant likelihood of misclassification in close proximity to the extreme threshold, leading to substantial inaccuracies in predictions.

There is also ongoing research focused on forecasting the likelihood of extreme events [Laptev *et al.*, 2017; Zhu and Laptev, 2017; Wambura *et al.*, 2020] or predicting the maximum value of extremes within a future time window [Galib *et al.*, 2022; Galib *et al.*, 2023]. However, their objective differs significantly from forecasting the sequence of value itself.

3 Methodology

3.1 Problem Statement

Consider a time series dataset that is divided into a sequence of temporally adjacent time windows, denoted as $\{w_i\}_{i=1}^{N}$, with each spaced one step apart. For each window, it contains a historical period defined by the time interval $[t_{i-n}, t_{i-1}]$, as well as a forecast period indicated by $[t_i, t_{i+m}]$, where n is the length of observed history series, and m is the prediction step length. Then, we let $\mathbf{y}_{t_i}^{-} := \{y_t\}_{t=t_{i-n}}^{t_{i-1}}$ be the series associated with the target-specific variable, and $\mathbf{z}_{t_i}^{-} := \{\mathbf{z}_t\}_{t=t_{i-n}}^{t_{i-1}}$ being the influencing factors. Therefore, the prediction problem can be described as

$$(\mathbf{y}_{t_i}^{-}, \mathbf{z}_{t_i}^{-}) \rightarrow \mathbf{p}_{t_i}^{+} := \{p_t\}_{t=t_i}^{t_{i+m}}. \tag{1}$$

The entire time series dataset is partitioned into training and test sets in accordance with chronological order, in a roughly $7:3$ ratio. At the training stage, effective predictors are learnt by minimizing the loss between the prediction $\mathbf{p}_{t_i}^{+}$ and the ground truth (GT) of the forecast, i.e., $\mathbf{y}_{t_i}^{+} := \{y_t\}_{t=t_i}^{t_{i+m}}$.

The mean square error (MSE), i.e.,

$$\text{MSE} := \sum_i \sum_{t=t_i}^{t_{i+m}} (x_t)^2, \tag{2}$$

where $x_t := p_t - y_t$, is commonly utilized as the loss function during training, calculating the squared discrepancy between the predicted and GT values. However, when dealing with imbalanced time series data, it is crucial to acknowledge that the majority of observations are within the normal range while a minority exhibit extreme values, as presented in Figure 1. In this context, employing MSE as the loss function for training predictors tends to lead to overfitting of the normal observations and underfitting of the extreme values. The predictors trained using MSE primarily focus on estimating the conditional expectation of the target variable, rather than its extreme values [Bishop and Nasrabadi, 2006]. Consequently, extreme values may be ignored as outliers, as these predictors prioritize enhancing performance on normal observations.

However, in the context of imbalanced time series data, there is a heightened emphasis on extreme events as they offer valuable insights for anticipating the most severe scenarios expected to occur within the forecast period. Using MSE as a training metric may underestimate extreme events, thereby potentially conveying misleading instructions to the public. Therefore, it is required to develop a novel loss function that compels predictors to prioritize extreme events with greater emphasis while maintaining the accuracy for normal events.

3.2 Self-adaptive Extreme Penalized Loss

Consider that in imbalanced time series prediction, not a few tasks solely focus on one-sided extremes (e.g. lying at the positive axis) such as the predictions of heavy waves, strong winds, severe air pollution, etc. For those instances, there is a widespread expectation that predictions related to extreme values should not be undervalued owing to their crucial role in precisely assessing potential harm. This naturally comes with the idea of increasing the penalty of extreme events [Ding *et al.*, 2019; Liu *et al.*, 2023].

However, merely augmenting the penalty weight assigned to extreme events will result in a considerably magnified estimation of normal events, which is not in line with our expectations. Therefore, we propose a tripartite division of the penalty case instead of a bipartite one, to guarantee the performance on both normal and extreme events. A reduced penalty is imposed in cases of slightly overestimating extreme events. This is considered acceptable, as even if the eventual conditions prove to be less severe than initially predicted, the slight overstatement does not result in significant human and property losses when compared to the consequences of underestimating these risks. Moreover, given that extreme events constitute a significantly smaller proportion compared to normal events, it is more advantageous to relax the penalties associated with over-estimating extreme events in order to uphold the overall validity of the prediction.

Based on the above points, we propose an extreme penalized loss (EPL) that considers x_t as the independent variable and should comply with the following conditions:

1. For the EPL function, it is necessary that the magnitude of the penalty it imposes on significant predicted errors surpasses that of a linear function.

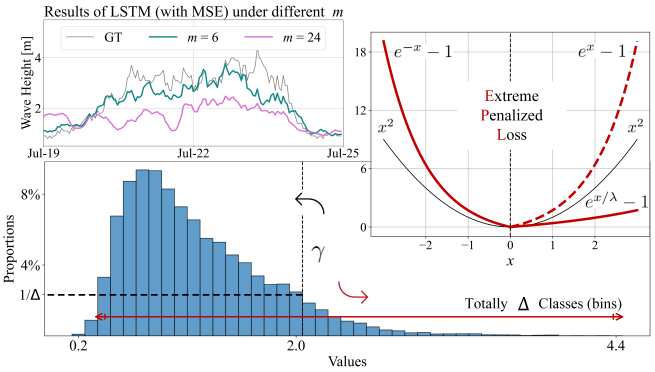

Figure 2: The upper left graph depicts the correlation between the underestimation degree of extreme events and m, highlighting the need to set $\lambda = m + 1$. We also present the proportion histogram of the training data of wave height, the self-adaptive setting of γ, as well as the graphical representation of function $f(x)$ from EPL.

2. Compared to normal events, the loss function should enhance the penalties for errors related to extreme events. Specifically, it is necessary to impose significant penalties on under-predicted errors of extreme events, while at the same time properly reducing the penalties for over-predictions, thereby strengthening the overall penalties for normal events and under-predicted extreme events.
3. Hyper-parameters of EPL function should possess self-adaptive capability, specifically in terms of determining the threshold for identifying extreme events and adjusting the penalty strength based on specific data characteristics, which will enhance the ability of EPL to generalize across various time series datasets while upholding a superior level of effectiveness.

Then, our EPL function is developed as

$$\text{EPL} := \sum_i \sum_{t=t_i}^{t_{i+m}} f(x_t), \tag{3}$$

$$f(x) = \begin{cases} x^2 & \tilde{x} < \gamma, \\ e^{-x} - 1 & \tilde{x} \geq \gamma,\ x < 0, \\ e^{x/\lambda} - 1 & \tilde{x} \geq \gamma,\ x \geq 0, \end{cases} \tag{4}$$

where $\tilde{x}$ denotes the indication from current step, e.g., when $x = x_{t_i}$, $\tilde{x} = y_{t_i}$. The parameter γ represents the threshold for detecting extreme events. It is noticed that in our EPL function, the loss function is defined as the traditional MSE loss ℓ_c for potentially normal events. The parameter λ indicates the penalty strength of over-predicted extreme events. More intuitively, the graph of $f(x)$ is displayed in Figure 2.

Both γ and λ are adaptively determined based on data characteristics and the predicted step size regarding to tasks. To determine γ, we first partition the entire training dataset into Δ classes by dividing the continuous data into bins with intervals equal to the maximum order of magnitudes minus one. For example, the intervals of wave height, wind speed and air quality are 0.1, 1 and 10, respectively. Intuitively, the bins that have a sample to population ratio lower than $1/\Delta$ will be considered as the intervals representing extreme events. However, bins with values close to zero should be excluded

from the pool of potential extreme events. Therefore, taking the time series of wind speed as an instance, the threshold γ for identifying the extreme event can be determined by selecting the interval $[\gamma, \gamma+1]$ with a sample to population ratio lower than $1/\Delta$, while ensuring that the sample to population ratio of interval $[\gamma-1, \gamma]$ is higher than $1/\Delta$. On the other hand, the value of λ clearly affects the penalty strength between the normal and extreme events, as shown in the upper left graph of Figure 2. Given the larger prediction steps, it becomes easier to under-predict the extreme values. Therefore, a higher value of λ should be employed to enhance the penalty on under-estimated extremes. Thus, we set $\lambda = m+1$ for all the prediction tasks and time series data.

Remark 1. *The proposed EPL loss in Equation 4 is specifically designed for time series predictions involving one-sided and positive extremes, making it unsuitable for tasks that care for two-sided extremes, such as the temperature prediction. However, the loss function on the negative side can be derived using similar methodologies. Then, by combining two-sided loss functions together, it is possible to extend our proposed EPL loss to handle situations with two-sided extremes.*

3.3 Prediction Framework

The pipeline of our prediction framework is illustrated in Figure 3, encompassing the modules of time-series decomposition, long-term trend prediction utilizing an attention-LSTM, and LSTM-based prediction for variations employing EPL.

Time-series Decomposition

Decomposition is a commonly employed technique for time-series data, enabling the conversion of complex signals into components that are more conducive to prediction. We in this framework employ the Empirical Wavelet Transform (EWT) [Gilles, 2013] to decompose the target time series $\{y_t\}$ into a long-term trend component $\{u_t\}$ and the residual variation $\{v_t\}$ satisfying $v_t = y_t - u_t$ for every time index t, as presented in Figure 3.

The long-term trend component is characterized by its inherent smoothness, guaranteeing a consistent magnitude of value without requiring the utilization of EPL. The variation component preserves almost all of the short-term fluctuations present in the original data. Thus, rather than implementing our proposed EPL for the original time series $\{y_t\}$, we opt to apply the EPL exclusively on the variation component $\{v_t\}$.

Decomposing the data and solely employing our proposed EPL to exclusively address the variation component bestows a distinct advantage. From the distributions of $\{y_t\}$ and $\{v_t\}$ depicted in Figure 3, it is observed that following the decomposition, there is a certain reduction in the disparity between extreme data samples and normal data samples, thereby enhancing the predictability of the data. Given the unique data characteristics of each component, we employ distinct strategies to effectively address them within the subsequent section of our prediction framework.

Learning Strategies

Since the long-term trend signal $\{u_t\}$ exhibits relatively smooth behavior devoid of significant fluctuations, we propose utilizing an LSTM network with an incorporated attention mechanism to accurately capture the long-term trend.

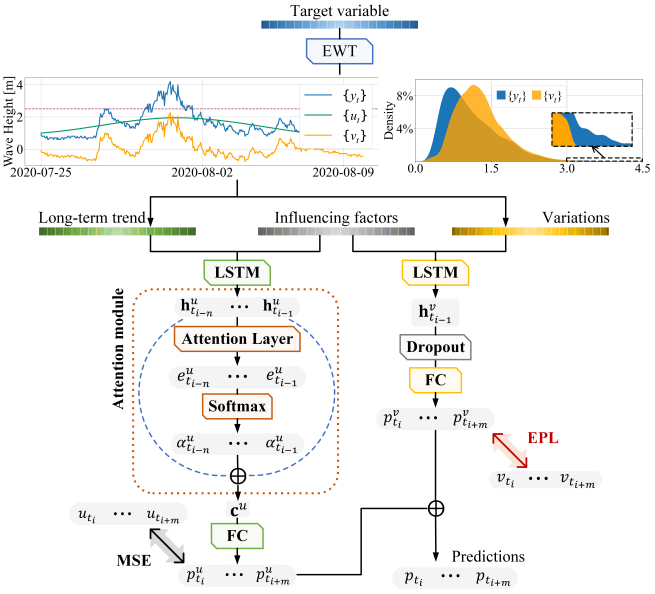

Figure 3: The flowchart of our framework is illustrated, along with the visualization of decomposition components using EWT, taking "wave height" as an example. Besides, the distribution comparisons between $\{y_t\}$ and $\{v_t\}$ are provided, wherein the distribution of $\{v_t\}$ is shifted by taking 0 as its minimum value to enhance clarity.

The flowchart of attention-based LSTM is depicted in Figure 3. With the hidden states of LSTM, denoted as $\{\mathbf{h}_t^u\}_{t=t_{i-n}}^{t_{i-1}}$, the attention module outputs $\mathbf{c}^u$ by calculating:

$$\mathbf{c}^u = \sum_{t=t_{i-n}}^{t_{i-1}} \alpha_t^u \mathbf{h}_t^u, \tag{5}$$

where the attention probability $\{\alpha_t^u\}_{t=t_{i-n}}^{t_{i-1}}$ is obtained by

$$\alpha_t^u = \mathrm{softmax}(e_t^u),\ e_t^u = \mathbf{W}_t \tanh(\langle \mathbf{h}_t^u, \mathbf{h}_{t_{i-1}}^u \rangle + \mathbf{b}_t), \tag{6}$$

for every $t = t_{i-n}, \ldots, t_{i-1}$. Afterwards, a fully-connected (FC) layer is employed to properly organize the learning outputs into the desired format of prediction. Given that the long-term trend is a signal with no obvious extremes within a given time window, we employ the MSE loss to train the attention-based LSTM network.

To predict the variation signal, we employ an LSTM network as its fundamental predictor. To alleviate the potential issue of model overfitting, we integrate a Dropout layer following the LSTM network. Then, an FC layer is utilized at last for generating the prediction variable. Given that the variation signal preserves the fluctuations of the un-decomposed time series signal, we utilize the proposed EPL loss to train the predictor associated with the variation signal. Moreover, during the test phase, the predictions of each individual component are added to generate the ultimate output.

Our primary focus is on innovating the loss function rather than constructing a new network for time-series prediction. The results in Section 4 will verify the superiority of our proposed EPL by achieving optimal performance even with the use of LSTM, compared to other state-of-the-art methods[1].

[1]Code will be available at https://github.com/Ldiper/EPL.

Wave Height										
Dataset	ARIMA	N-Beats	LSTM	BiLSTM	XFormer	EVL	SADI	NEC+	[Fu *et al.*, 2023]	Ours
BSG	0.36/0.686	0.39/0.598	0.39/0.565	0.36/0.579	0.38/0.584	0.39/0.563	0.38/0.611	0.37/0.638	0.35/0.699	0.27/0.814
DCN	0.30/0.647	0.28/0.590	0.29/0.623	0.28/0.646	0.32/0.538	0.24/0.699	0.35/0.499	0.26/0.524	0.27/0.648	0.22/0.776
NJI	0.32/0.557	0.31/0.585	0.33/0.541	0.32/0.565	0.31/0.625	0.28/0.661	0.39/0.341	0.26/0.713	0.34/0.532	0.18/0.771
XMD	0.20/0.639	0.23/0.398	0.27/0.359	0.27/0.368	0.26/0.408	0.21/0.559	0.20/0.525	0.17/0.655	0.24/0.418	0.16/0.699
Kaggle-WH	0.37/0.537	0.34/0.618	0.41/0.375	0.41/0.379	0.31/0.619	0.26/0.722	0.38/0.506	0.32/0.657	0.26/0.725	0.25/0.774
Wind Speed										
Dataset	ARIMA	N-Beats	LSTM	BiLSTM	XFormer	EVL	SADI	NEC+	[Zha *et al.*, 2022]	Ours
Kaggle-WS	2.74/0.706	4.63/0.077	3.29/0.713	3.26/0.718	3.34/0.703	3.45/0.683	2.57/0.736	3.46/0.683	2.09/0.789	1.91/0.803
Air Quality Index										
Dataset	ARIMA	N-Beats	LSTM	BiLSTM	XFormer	EVL	SADI	NEC+	[Chang *et al.*, 2020]	Ours
BJair	64.69/0.522	72.65/0.397	86.26/0.151	80.14/0.267	76.68/0.302	53.32/0.657	50.84/0.700	58.36/0.611	61.47/0.529	42.32/0.796
Italy	51.30/0.505	46.81/0.625	47.56/0.619	44.849/0.649	44.85/0.654	56.89/0.574	86.25/0.353	50.57/0.586	93.91/0.312	38.83/0.744

Table 1: Quantitative comparisons (RMSE/R^2) of state-of-the-art approaches across different meteorological elements on evaluation datasets. The best scores are visually emphasized in red, whereas suboptimal outcomes are marked in blue. The reported scores for "wave height" and "air quality index" correspond to a prediction step of 24, whereas for "wind speed", the prediction step is 4.

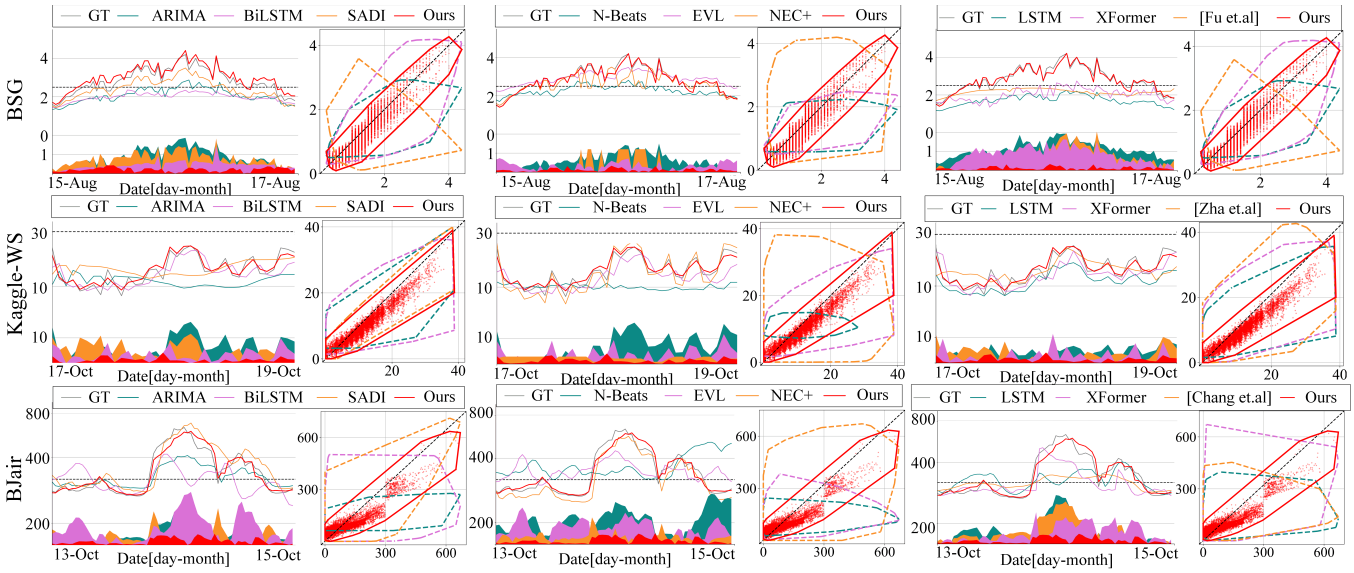

Figure 4: Quality comparisons with state-of-the-art methods. We not only present the sequence curves of prediction results but also show the absolute errors between these predictions and GTs. The scatter plots present the actual versus predicted values of our framework, along with the minimum convex hulls of all the compared approaches.

4 Experiments

In addition to the comparisons of time series predictions with state-of-the-art methods, we also provide the results of block maxima forecasting [Galib *et al.*, 2022]. Finally, we illustrate the impact of each component of the predictive framework.

4.1 Real-world Datasets

The experiments involve prediction tasks pertaining to 3 distinct variables, covering the comparisons on 8 datasets.

Wave Height

The wave height datasets are acquired from the National Marine Data Center[2] (NMDC) and the Kaggle competition[3]. The 4 datasets of NMDC are specifically associated with distinct coastal stations adjacent to China, namely "BSG", "DCN", "NJI" and "XMD". These datasets consist of hourly wave height data and corresponding met-ocean conditions spanning from January 2019 to December 2021. The dataset offered by the Kaggle competition originates from wave measuring buoys situated in close proximity to the Queensland coastline, which is named as "Kaggle-WH". This dataset involves hourly data collected from January 2017 to June 2019. According to WMO, the designated threshold for wave height warning is 2.5 meters.

Wind Speed

The dataset employed for wind speed prediction is acquired from the Kaggle competition[4], which includes hourly data collected from April 2006 to December 2016. Furthermore, it contains variables such as wind speed, temperature, humidity, and other relevant factors. The related weather station is located in an exposed terrestrial region within Szeged, Hun-

[2]https://mds.nmdis.org.cn

[3]https://www.kaggle.com/datasets/jolasa/waves-measuring-buoys-data-mooloolaba

[4]https://www.kaggle.com/datasets/budincsevity/szeged-weather

gary. Thus, the prescribed threshold for issuing a cautionary notice concerning land wind stands at 36 km/h, according to the regulations of WMO.

Air Quality Index

Two air quality datasets are acquired from the UCI repository, involving related data for two distinct locations: Beijing[5] and Italy[6]. The Beijing dataset encompasses hourly data spanning from January 2010 to December 2014, including the air quality index and a comprehensive range of meteorological variables. While, the dataset of Italy spans from March 2004 to April 2005 and comprises hourly air pollution index and other averaged responses obtained from a chemical multisensor equipped with five metal oxide sensors. On the other hand, the warning threshold for air quality index is set as 200, according to the regulation of WMO.

4.2 Experimental Setup

We compare the proposed method against various state-of-the-art approaches, including (1) ARIMA [Box *et al.*, 2015]; (2) N-Beats [Oreshkin *et al.*, 2020]; (3) LSTM [Hochreiter and Schmidhuber, 1997]; (4) BiLSTM [Schuster and Paliwal, 1997]; (5) XFormer [Vaswani *et al.*, 2017]; (6) EVL [Ding *et al.*, 2019]; (7) SADI [Liu *et al.*, 2023]; (8) NEC+ [Li *et al.*, 2023] on various time series prediction tasks. In particular, we also compare methods designed specifically for distinct tasks, including [Fu *et al.*, 2023] for wave height prediction, [Zha *et al.*, 2022] for wind speed prediction, and [Chang *et al.*, 2020] for air quality index prediction. Moreover, we also conduct comparisons on the task of block maxima forecasting, with a specifically designed approach named DeepExtrema [Galib *et al.*, 2022].

For all the experiments, a 3-layer LSTM with 128 nodes of layer widths is consistently utilized across different datasets. Besides, for model parameter settings, we set the learning rate to 0.002, the batch size to 256, the dropout rate to 0.2, the maximum epochs for training to 200, and we added an early stop mechanism to prevent model over-fitting. We use the following 3 metrics to evaluate the performance of each method on different datasets: (1) Root Mean Square Error (RMSE); (2) Coefficient of Determination (R^2); (3) F_1 Score.

4.3 Experimental Results

Time Series Prediction

We first validate our complete framework for time series prediction. The results, presented in Table 1, show the comparison results between our proposed method and other state-of-the-arts from quantitative analysis. We also offer qualitative comparisons in Figure 4, where two types of graphs are presented. One of them exhibits time sequence curves of specific temporal segments, showcasing not only the prediction outcomes generated by different approaches but also presenting the absolute errors between these predictions and GTs. The second graph displays scatter plots illustrating the actual versus predicted values of our method, along with the minimum convex hulls of the compared approaches[7].

[5]http://archive.ics.uci.edu/dataset/381

[6]http://archive.ics.uci.edu/dataset/360

[7]Due to space limits, more results are presented in the Appendix.

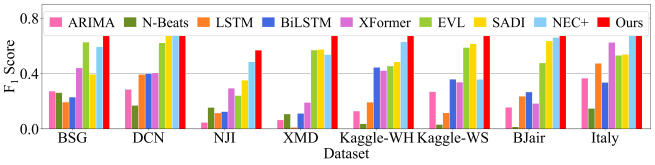

Figure 5: Performance comparison in terms of F_1 score for predicting extreme events only.

	EVL	SADI	NEC+	DeepExtrema	Ours
BSG	0.42/0.640	0.38/0.709	0.43/0.623	0.39/0.685	0.26/0.861
DCN	0.42/0.658	0.43/0.617	0.29/0.716	0.33/0.693	0.22/0.836
NJI	0.32/0.632	0.39/0.568	0.36/0.587	0.37/0.586	0.21/0.861
XMD	0.29/0.515	0.22/0.685	0.24/0.584	0.23/0.639	0.21/0.699
Kaggle-WH	0.45/0.377	0.37/0.613	0.33/0.667	0.27/0.777	0.24/0.821
Kaggle-WS	3.57/0.717	2.55/0.749	3.32/0.755	3.48/0.731	3.13/0.782
BJair	86.59/0.473	62.19/0.701	74.77/0.608	63.67/0.690	57.44/0.768
Italy	45.90/0.649	73.58/0.197	49.20/0.594	44.10/0.677	40.21/0.731

Table 2: Quantitative comparisons in terms of RMSE/R^2 on evaluation datasets for predicting extreme events only.

From the quantitative analysis we can see that our complete framework consistently acquires superior performance levels across both metrics for all datasets. Even when compared to recently proposed approaches such as SADI and NEC+, our complete framework has clear superiority, which can also be validated from Figure 4. It is worth noticing that the prediction results of SADI are relatively too smooth and underestimated, primarily attributable to its decomposition strategy. Although NEC+ outperforms SADI in predicting extremes, its classifier may cause misjudgement in identifying extreme events, thus leading to critical errors, as shown in Figure 4.

Extremes Prediction

Furthermore, we also evaluate the accuracy of each method in predicting extreme values (exceeding the predefined warning thresholds) for each meteorological element, as illustrated in Figure 5. It is observed that our complete framework outperforms all other approaches in terms of F_1 score. In addition, methods specifically designed for extreme events may be less effective than certain baselines, e.g., XFormer, thereby indicating their inherent instability.

Our approach is also feasible for effectively predicting the maximum value within a specified time window. We conduct further experiments to compare our framework with DeepExtrema [Galib *et al.*, 2022], a specialized approach designed for block maximum prediction. For the methods compared in Table 1, we evaluate the highest prediction values from time windows and compare them with the results achieved using DeepExtrema. From the comparisons in Table 2 we can see that our framework exhibits a superiority of over 20 percent in terms of R^2. Therefore, our complete framework achieves a commendable degree of precision in predicting the maximum value within a block of time window.

4.4 Ablation Studies

We have demonstrated the superiority of our complete framework through a comparative analysis with other state-of-the-art approaches. However, it is essential to investigate the effectiveness of our proposed EPL function, explore the influence of the decomposition strategy, and assess the impact of

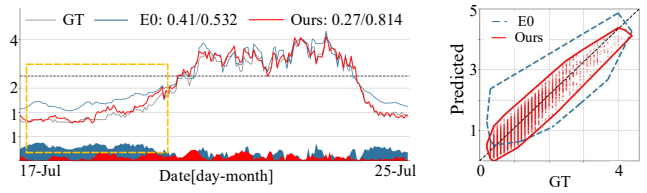

Figure 6: The necessity of setting a tripartite division of the penalty. Comparisons are performed on the "BSG" dataset under RMSE/R^2, using a 24-step prediction as an illustration example.

	E1	E2	E3	E4	Ours
Decomposition	×	×	√	√	√
Loss Function	MSE	EPL	MSE	EPL	EPL
Attention Module	√	√	√	×	√
Prediction Results					
RMSE/R^2	0.49/0.491	0.31/0.759	0.34/0.688	0.30/0.799	0.27/0.814

Table 3: Quantitative comparisons of ablation studies are performed on the "BSG" dataset, using a 24-step of wave height prediction as an illustration example.

incorporating the attention module. We conduct ablation experiments on the "BSG" dataset to facilitate in-depth analysis, as presented in Figure 6, Table 3 and Figure 7.

Effectiveness of EPL

Firstly, we conduct further detailed analysis on the necessity of designing the tripartite division of the penalty of EPL. We conduct a contrasting experiment, referred to as "E0", which employs a bipartite loss function, i.e., for $\tilde{x} < \gamma$, $f(x) = x^2$, while for other values, $f(x) = e^{|x|} - 1$. From the comparisons presented in Figure 6, it is evident that merely emphasizing penalties on extreme events leads to an overestimation of normal events, thereby diminishing the overall accuracy. Thus, it is crucial to develop the tripartite division of penalties as our EPL to enhance the overall performance.

To further validate the effectiveness of our proposed EPL, we establish E1, which employs an LSTM network with the same architecture as the one used for predicting $\{u_t\}$. Moreover, for E1, the traditional MSE is utilized as the loss function. Then, we additionally establish E2, which is configured identically to E1 except for the modification of the loss function. From the comparisons between E1 and E2, it is evident that our proposed EPL, when utilized as the loss function for predictions without decomposition, yields more than a 50 percent improvement compared to E1. Furthermore, we additionally conduct a control experiment wherein the configurations of our comprehensive framework remain unchanged, except for substituting EPL with MSE, denoted as E3. Using the decomposition strategy, our designed EPL demonstrates a remarkable improvement of over 18 percent improvement in R^2 and a significant reduction of more than 20 percent reduction in RMSE, thereby highlighting the effectiveness of our designed loss function.

Influence of Decomposition

Compared to E1, the E3 demonstrates a 40 percent improvement in R^2 due to the decomposition. With decomposition, the smooth long-term trend is divided, which dominates over the signal and facilitating more accurate predictions. Moreover, we have also demonstrated in Section 3.3 that through decomposition, the disparity between the extreme and normal samples is certainly reduced, thereby enhancing the predictability of the residual variation component. Moreover, it is evident that our framework achieves reduced prediction errors without employing the decomposition strategy, as evidenced by comparisons with E2. These findings collectively validate the necessity of employing a decomposition strategy for accurate prediction.

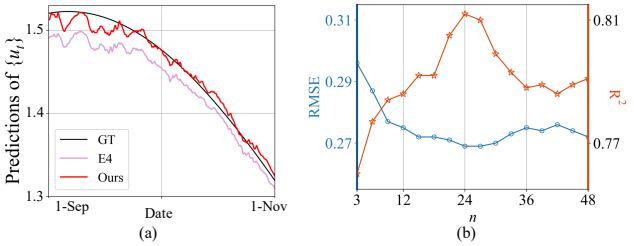

Figure 7: Ablation studies of (a) the visual impact of attention module, and (b) the sensitivity of parameter n.

Impact of Attention Module

Incorporating an attention mechanism into long-term trend prediction greatly enhances its ability to capture essential information and guides the attention of the network towards more stable parts of the data [Luo *et al.*, 2022]. The comparisons between E4 and our comprehensive framework demonstrate that the exclusion of the attention module leads to a slight decrease in prediction performance. We also offer visual comparisons in Figure 7(a), illustrating the differences between E4 and our proposed framework in terms of predicting the long-term trend component. Both of the comparisons validate the impact of incorporating an attention module.

Sensitivity of Parameter

In Figure 7(b), we additionally present a sensitivity analysis regarding the parameter n. It is observed that when predicting a 24-step of forecast, utilizing historical data of comparable length yields favorable performance.

5 Conclusion

In this research, we propose a self-adaptive loss function that employs a tripartite division of the penalty case rather than a bipartite one to guarantee the performance for both normal and extreme events. Then, a decomposition-based framework is developed that integrates the proposed EPL with an LSTM network as well as containing an attention-based LSTM, outperforming other state-of-the-art approaches for both time series prediction and block maxima prediction tasks.

Our main emphasis lies in the development of EPL that has been validated to be effective in certain tasks. However, in order to tackle the challenges of imbalanced time series prediction that involve data with unique temporal characteristics, it is recommended to integrate the proposed EPL with more advanced network architectures and decomposition techniques to improve the overall prediction performance.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities (No. 3132023523) and the Natural Science Foundation of Liaoning Province (No. 2023-MS-126).

References

[Bishop and Nasrabadi, 2006] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[Box *et al.*, 2015] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[Chang *et al.*, 2020] Yue-Shan Chang, Hsin-Ta Chiao, Satheesh Abimannan, Yo-Ping Huang, Yi-Ting Tsai, and Kuan-Ming Lin. An lstm-based aggregated model for air pollution forecasting. *Atmospheric Pollution Research*, 11(8):1451–1463, 2020.

[Connor *et al.*, 1994] Jerome T Connor, R Douglas Martin, and Les E Atlas. Recurrent neural networks and robust time series prediction. *IEEE Transactions on Neural Networks*, 5(2):240–254, 1994.

[Ding *et al.*, 2019] Daizong Ding, Mi Zhang, Xudong Pan, Min Yang, and Xiangnan He. Modeling extreme events in time series prediction. In *ACM SIGKDD*, 2019.

[Fu *et al.*, 2023] Yang Fu, Feixiang Ying, Lingling Huang, and Yang Liu. Multi-step-ahead significant wave height prediction using a hybrid model based on an innovative two-layer decomposition framework and lstm. *Renewable Energy*, 203:455–472, 2023.

[Galib *et al.*, 2022] Asadullah Hill Galib, Andrew McDonald, Tyler Wilson, Lifeng Luo, and Pang-Ning Tan. Deepextrema: A deep learning approach for forecasting block maxima in time series data. In *IJCAI*, 2022.

[Galib *et al.*, 2023] Abu Hasnat Galib, Aaron McDonald, Ping-Nam Tan, et al. Self-recover: Forecasting block maxima in time series from predictors with disparate temporal coverage using self-supervised learning. In *IJCAI*, 2023.

[Geurts *et al.*, 2006] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006.

[Gilles, 2013] Jérôme Gilles. Empirical wavelet transform. *IEEE Transactions on Signal Processing*, 61(16):3999–4010, 2013.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Hou *et al.*, 2021] Cheng Hou, Jiahui Wu, Bin Cao, and et al. A deep-learning prediction model for imbalanced time series data forecasting. *Big Data Mining and Analytics*, 4(4):266–278, 2021.

[Karevan and Suykens, 2020] Zahra Karevan and Johan A.K. Suykens. Transductive lstm for time-series prediction: An application to weather forecasting. *Neural Networks*, 125:1–9, 2020.

[Laptev *et al.*, 2017] Nikolay Laptev, Jason Yosinski, Li Erran Li, and Slawek Smyl. Time-series extreme event forecasting with neural networks at uber. In *ICML*, 2017.

[Li *et al.*, 2023] Yanhong Li, Jack Xu, and David C Anastasiu. An extreme-adaptive time series prediction model based on probability-enhanced lstm neural networks. In *AAAI*, 2023.

[Liu *et al.*, 2023] Hengbo Liu, Ziqing Ma, Linxiao Yang, Tian Zhou, Rui Xia, Yi Wang, Qingsong Wen, and Liang Sun. Sadi: A self-adaptive decomposed interpretable framework for electric load forecasting under extreme events. In *ICASSP*, 2023.

[Luo *et al.*, 2022] Qin-Rui Luo, Hang Xu, and Long-Hu Bai. Prediction of significant wave height in hurricane area of the atlantic ocean using the bi-lstm with attention model. *Ocean Engineering*, 266:112747, 2022.

[Oreshkin *et al.*, 2020] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *ICLR*, 2020.

[Polson and Sokolov, 2020] Nicholas G Polson and Vadim Sokolov. Deep learning for energy markets. *Applied Stochastic Models in Business and Industry*, 36(1):195–209, 2020.

[Rajbhandari *et al.*, 2021] Yaju Rajbhandari, Anup Marahatta, Bishal Ghimire, Ashish Shrestha, Anand Gachhadar, Anup Thapa, Kamal Chapagain, and Petr Korba. Impact study of temperature on the time series electricity demand of urban nepal for short-term load forecasting. *Applied System Innovation*, 4(3):43, 2021.

[Sapankevych and Sankar, 2009] Nicholas I. Sapankevych and Ravi Sankar. Time series prediction using support vector machines: A survey. *IEEE Computational Intelligence Magazine*, 4(2):24–38, 2009.

[Schuster and Paliwal, 1997] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[Siami-Namini *et al.*, 2018] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. A comparison of arima and lstm in forecasting time series. In *IEEE International Conference on Machine Learning and Applications*, 2018.

[Siami-Namini *et al.*, 2019] Soheil Siami-Namini, Nader Tavakoli, and Ahmad S Namin. The performance of lstm and bilstm in forecasting time series. In *Big Data*, 2019.

[Tyralis and Papacharalampous, 2017] Hristos Tyralis and Georgia Papacharalampous. Variable selection in time series forecasting using random forests. *Algorithms*, 10(4):114, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[Wambura *et al.*, 2020] Stephen Wambura, He Li, and Alemu Nigussie. Fast memory-efficient extreme events prediction in complex time series. In *ICRSA*, 2020.

[Weerakody *et al.*, 2021] Philip B. Weerakody, Kok Wai Wong, Guanjin Wang, and Wendell Ela. A review of irregular time series data handling with gated recurrent neural networks. *Neurocomputing*, 2021.

[Wilson *et al.*, 2022a] Tom Wilson, Ping-Nam Tan, and Liang Luo. Deepgpd: A deep learning approach for simulating geospatial temporal extreme events. In *AAAI*, 2022.

[Wilson *et al.*, 2022b] Tyler Wilson, Andrew McDonald, Asadullah Hill Galib, Pang-Ning Tan, and Lifeng Luo. Beyond point prediction: Capturing zero-inflated & heavy-tailed spatiotemporal data with deep extreme mixture models. In *ACM SIGKDD*, 2022.

[Wu *et al.*, 2021] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *NeurIPS*, 2021.

[Zha *et al.*, 2022] Wenting Zha, Jie Liu, Yalong Li, and Yingyu Liang. Ultra-short-term power forecast method for the wind farm based on feature selection and temporal convolution network. *ISA Transactions*, 129:405–414, 2022.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, 2021.

[Zhou *et al.*, 2022] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *ICML*, 2022.

[Zhu and Laptev, 2017] Lingxue Zhu and Nikolay Laptev. Deep and confident prediction for time series at uber. In *ICDM*, 2017.