Deep Learning-based Block Maxima Distribution Predictor for Extreme Value Prediction

Kaneharu Nishino, Ken Ueno, Ryusei Shingaki
System AI Laboratory, Corporate Research & Development Center
Toshiba Corporation
Kawasaki, Japan
{kaneharu1.nishino, ken.ueno, ryusei1.shingaki}@toshiba.co.jp

ABSTRACT

Extremely large values often appear in various fields such as meteorology, hydrology and financial engineering, which may be caused by catastrophic events such as cyclones, flooding, and financial crisis. Although such extreme value prediction is crucial for risk management, it is difficult to forecast them by conventional machine learning because the number of extreme values is much smaller than the number of non-extreme values. The existing approaches forecast when and how large extremes occur, however, for practical use, some errors in forecast time may be ignored to take action such as evacuation against catastrophic events. In this study, instead of forecasting exact timing, we tackle an alleviated problem of predicting the maximum values over a certain length period, which are known to follow Generalized Extreme Value distribution (GEV). Therefore, we introduce GEV into deep learning and propose a method to forecast the maxima distribution by predicting of parameters of GEV. As a result of experiments, we report that the proposed method can correctly predict time series of the maxima for both of artificial data and real-world data.

KEYWORDS

Machine learning, Time series, Forecasting, Extreme Value Theory

1 Introduction

Time series prediction is one of the classical research topic. With the recent development of deep learning, recurrent neural networks such as GRU (Gated Recurrent Unit) [7] are attracting attention and much applied research has been conducted using these networks. These time series prediction methods are widely applied in the field of risk management related to extreme climate prediction [25, 10, 18, 21, 23], stock price alerts [8, 28], and network traffic anomaly detection [31].

Such prediction tasks often have a problem in that they involve extreme phenomena. For example, in extreme climate prediction, an extreme phenomenon corresponds to abnormal weather in which the values of temperature, wind speed or precipitation. are extremely high or low (called extreme values). Such extreme values are also observed in other fields as stock prices and network traffics. Since failures of the prediction can lead to catastrophic events in risk management, extreme prediction is an important task. There are many studies on extreme value prediction by typical time series models such as autoregressive integrated moving average (ARIMA) models [25, 31] and recent Deep Neural Network (DNN) models such as Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) [18, 21], which can handle non-stationary time series.

However, extreme value prediction is difficult for conventional machine learning methods [11]. Since extreme value data points occur only rarely, usual time series machine learning methods tend to learn mainly non-extreme data points. This is a sort of data imbalance problem, which previous studies shows most DNN suffer from [13, 20]. Thus, correct learning from extreme value is difficult and efforts to forecast extreme phenomena are prone to failure.

On the other hand, extreme values have been studied in statistics for a long time, and they are modelled with Extreme Value Theory (EVT). According to EVT, extreme values are defined as the maximum and minimum values over a certain number of data points, and the distribution of these values must be a Generalized Extreme Value (GEV) distribution. Although EVT is suitable for analysis of stationary data, EVT in non-stationary data is not sufficiently established [22, 9]. In recent years, approaches combining EVT and prediction techniques with deep learning have been studied [11, 8]. According to Ding, Yang, and He [11], the difficulty of extreme value prediction arises when the distribution of data does not follow a normal distribution. They deal with this problem by modelling the data distribution according to EVT. They handle the partial problem of predicting whether the future data points will exceed a threshold or not so that they introduce EVT to prediction model. Using the probability that future data points exceed the threshold as a weight, they add a correction to predicted values by GRU in order to predict future extremes more correctly.

In most of the research on time series data with extremes, the goal is to forecast both accurate timing and values of extreme phenomena. However, the accurate timing of extremes may be unnecessary for practical use. For example, in the case of disaster prevention planning for a reservoir, it may be sufficient to predict timing of extreme inflow with an accuracy of half-day to one-day for planning a deployment additional monitors to avoid flooding.

Since the deployment of monitors is planned on a daily basis, it is not needed to predict exact timing of extreme inflow events with hourly accuracy. In this case, prediction for exact value of maximum of inflow are required rather than its exact timing.

Therefore, this study aims to predict the maximum value over a period of a certain length, instead of predicting the exact timing of extremes. To tackle the maxima prediction problem, we propose a novel loss function to introduce GEV distribution to GRU so that GRU can learn based on the GEV distribution-based loss. The proposed model predicts the parameters of extreme distribution over a period of a certain length through GRU for non-stationary time series data. The results of experiments using artificial data indicate that the model correctly predicted the change in the maximum value distribution. Moreover, the results of experiments using real data of water reservoir inflow and energy consumption indicate that the model forecast the maximum inflow values more accurately than standard GRUs.

Our contributions concern two points. First, we set up a novel alleviated problem by assuming practical use of prediction. We relax the problem of traditional extreme value prediction with exact timing of extremes and reformulate the problem as prediction of extreme values over a period of a certain length. This reformulation enables to use the block maxima approach of EVT to predict future maxima, which is useful for predicting extreme value distribution in practical use. Second, we propose a method to predict the maxima by introducing GEV to GRU. We define a new loss function so that GRU can learn by GEV distribution-based loss function. The proposed method can predict the maxima more correctly than simple GRUs and may potentially be applicable to a wide variety of data.

2 Related Works

Extreme Values have been studied in statistics for a long time and have recently been handled as a problem in time series prediction with machine learning. In this section, we briefly introduce Extreme Value Theory and Deep Learning based approach as related works.

2.1 Extreme Value Theory

Extreme Value Theory (EVT) describes behaviors of maxima for the same number of data points in an approach called Block Maxima. Data periods are divided to non-overlapping sub-periods with same length and the maximum of each sub-period is defined as an extreme value. These extremes are known to follow the distribution called Generalized Extreme Value distribution (GEV) [17], which cumulative distribution function is written as:

$$GEV(y;\mu,\sigma,\xi) = exp\left(-\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right) \quad (1)$$

where random variable y is the maximum of the sub-period data sequence X, and μ,σ,ξ are its parameters. These parameters vary depending on the original distribution which generates original data sequence X. Its probability density function is written as:

$$gev(y;\mu,\sigma,\xi) = \begin{cases} \frac{1}{\sigma}(1+\xi(\frac{y-\mu}{\sigma}))^{-1-\frac{1}{\xi}} exp\left(-(1+\xi(\frac{y-\mu}{\sigma}))^{-\frac{1}{\xi}}\right) & (\xi \neq 0) \\ \frac{1}{\sigma} exp\left(-exp\left(-\frac{y-\mu}{\sigma}\right) - \frac{y-\mu}{\sigma}\right) & (\xi = 0) \end{cases} \quad (2)$$

The definition requires the following condition:

$$1+\xi\left(\frac{y-\mu}{\sigma}\right) > 0 \quad (3)$$

When y does not satisfy this condition, we assume that the density function $gev(y;\ \mu,\sigma,\xi)$ is 0.

EVT typically handles stationary data and assumed that the X is generated from stationary distribution. However, time series data of our interest are often non-stationary time series data. Since a general theory dealing with non-stationary data has yet to be formulated [22, 9], many studies proposed techniques to apply EVT for non-stationary data. They are roughly divided into two following approaches.

One approach to construct a non-stationary extreme value distribution is by introducing time-varying terms into GEV parameters μ,σ and use $\mu(t),\sigma(t)$ [12, 22, 26, 6, 14].

$$GEV(y(t);\mu(t),\sigma(t),\xi) = exp\left(-\left[1+\xi\left(\frac{y(t)-\mu(t)}{\sigma(t)}\right)\right]^{-\frac{1}{\xi}}\right) \quad (4)$$

In this model, $GEV(y(t);\mu(t),\sigma(t),\xi)$ varies continuously with time t. Though this approach seems to be a natural extension of GEV, no general formulation for the time-varying term $\mu(t)$ and $\sigma(t)$ is known. Since data analysis and design of time-varying terms are done manually one by one, it takes a lot of time to construct a prediction model.

The other method is to slice the non-stationary time series data into short subsequences and to apply EVT to each of them independently [29]. This approach assumes that the short subsequence can be approximated as stationary sequence and estimate s stationary GEV distribution for each subsequence. In this model, $GEV(y_k;\mu_k,\sigma_k,\xi_k)$ is estimated separately for each subsequence and $\mu_k,\sigma_k,$ and ξ_k vary discretely with subsequence k. This approach has a problem that if the subsequences are too short, the number of data points from which the maximum is sampled is reduced and it results in instability of estimation. However, if the variation of the original distribution is sufficiently slow compared to the length of the subsequence, this approach can obtain an approximation of the non-stationary extreme value distribution.

2.2 Deep Learning-based Extreme Value Prediction

On the other hand, with the recent development of deep learning [15, 3], deep learning based approach became to be studied to predict extreme values [18, 21].

Different from traditional methods such as ARMA [24] and ARIMA [4], Recurrent Neural Networks (RNNs) such as Long-Short Term Memory (LSTM) [16] and Gated Recurrent Unit

(GRU) [7] are capable of extracting highly abstract patterns as features from time series data. Recent studies (e.g., [18, 21]) have attempted to predict extremes by extracting features corresponding to phenomena (such as typhoon, cyclones, or traffic accidents) that generate extreme values using deep learning. However, these extremes rarely occur and they are difficult to learn because the number of samples to learn is very smaller than the number of non-extreme samples. This imbalance leads to a model that predicts only non-extremes without extremes.

Such data imbalance problem is a general problem in the field of machine learning and it has been tackled by many studies [13, 20]. In contrast to these approaches, the study of Ding et al. [11] focused on the property of extreme values and analyzed that why extremes cannot be learned is because the distribution of the data does not necessarily follow a normal distribution. They claimed that the distribution of data often has a heavy-tail, and that the deference between a normal distribution and a heavy-tail distribution results in different performance of RMSE loss, which is the loss function usually employed in various machine learning methods.

As partial problem of prediction, they tackled a prediction problem of whether a future point will exceed a certain threshold or not. According to the Peak over Threshold approach in EVT, the distribution of a random variable exceeding a threshold follows the Generalized Pareto Distribution (GPD), whatever distribution generates the random variable. They utilize this property to propose a GPD-based loss function EVL to learn the problem of discriminating whether a threshold is exceeded or not. They attempt to correctly predict where data will exceed the threshold and predict future values containing extremes using the probability of exceeding as a weighting factor to revise predicted values.

3 Proposed Approach

Most of the previous studies attempted to predict the value of a certain time point in the future. However, it is known that usual machine learning methods cannot learn extremes properly due to the rarity of them. Ding et al. [11] analyzes that it is because the data do not follow a normal distribution, and they handle the exceedance probability following GPD in order to deal with any distributions of data.

Incidentally, is it always required to predict the extreme values with their exact time? Extreme value prediction is used for risk management in safety design or disaster prevention planning, such as prediction for flood [10], hurricanes or cyclones [18, 21], drought [23], stock price [8, 27], and anomaly network traffics [31]. In these cases, it is not always needed to forecast a time series with the same frequency as the sampling of data. For example, we consider the flood prediction. From the reservoir water level data sampled every minute, we would like to predict an event where the water level rises to an extreme level. However, for evacuation planning, it is not necessary to predict the exact time in minutes: prediction with an accuracy of half an hour or even an hour may be sufficient in this case. Moreover, for the planning of personnel deployment in the reservoir operation, it is sufficient to predict the timing with an accuracy of half a day or even a day.

Therefore, we assume that it is not necessary to predict the exact time of the extreme value for practical use. We can alleviate the conventional problem of extreme prediction needing both of exact timing and values into a new alleviated problem of prediction of maxima in a period of certain length. In this study, we aim to predict the maximum values in a certain length period in the future. Hereby we can introduce the Block Maxima approach of EVT to model any distribution of data with GEV distribution. Using a loss function based on GEV, we can solve the problem shown in [11] that the distribution of the data is not normal, and make it possible to learn extreme value data. In this study, we propose the method to predict the parameters of the GEV distribution of future periods using GRU.

3.1 Preliminaries

We formulate the alleviated problem setting here. We take multivariate time series data as input, and prediction of the maximum value of one dimension of the time series over a certain period in the future as output. Let the time series data of length T be

$$X_T = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_T] \quad (T \in \mathbb{N}) \tag{5}$$

and the length of the period to be predicted be $l \in \mathbb{N}$. We use the partial time series X_{kl} of this time series data up to time kl ($kl < T,\ k \in \mathbb{N}$) as input. The target y_{kl} of the prediction problem with input X_{kl} is the maximum of dth dimensional values over the period from kl +1 to $(k+1)l$, described as

$$y_{kl} = \max\left(x_{d,kl+1}, \cdots, x_{d,(k+1)l}\right) \tag{6}$$

where $x_{d,t}$ denotes dth dimensional value of $\boldsymbol{x}_t$.

3.2 Proposed Prediction Method

In order to predict maximum y_{kl} from input multivariate time series X_{kl}, we propose introducing GEV distribution into GRU, as shown in Figure 1. Using GRU and fully-connected layers, it obtains μ_{kl}, σ_{kl} and ξ_{kl} as parameters of the GEV distribution of the maximum y_{kl} from input time series X_{kl}.

In the proposed method, prediction is done as follows. First, input X_{kl} is fed to the GRU to obtain the hidden state $\boldsymbol{h}_{kl}$ at time kl.

$$\boldsymbol{h}_{kl} = GRU(X_{kl}) \tag{7}$$

Then the hidden state $\boldsymbol{h}_{kl}$ is input to a fully-connected two-layer neural network to calculate the distribution parameters μ_{kl}, σ_{kl} and ξ_{kl} as written by,

$$(\mu_{kl}, \sigma_{kl}, \xi_{kl})^T = W_2(W_1\boldsymbol{h}_{kl} + \boldsymbol{b}_1) + \boldsymbol{b}_2 \tag{8}$$

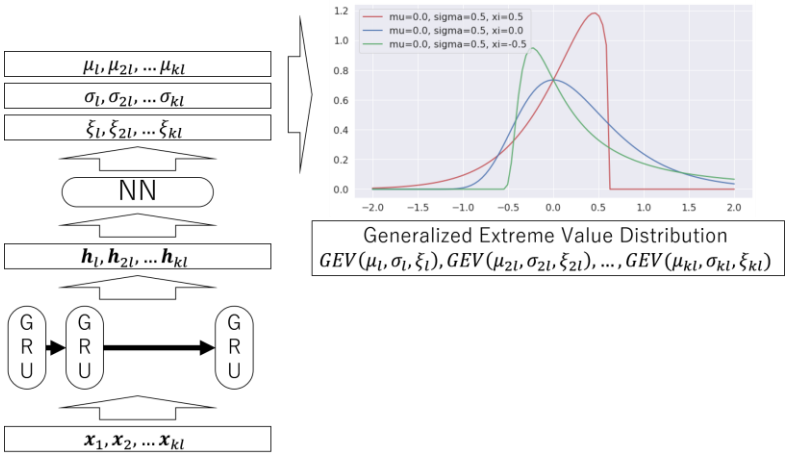


Figure 1 Illustration of proposed prediction process. Taking multivariate time series X_{kl} as input, it predicts the distribution parameters $\mu_{kl}, \sigma_{kl}, \xi_{kl}$ of the distribution of a future block maximum y_{kl}.

where $W_1, \boldsymbol{b}_1, W_2$ and $\boldsymbol{b}_2$ are parameters of the neural network. Hereby the predicted GEV distribution is obtained as $GEV(y_{kl}; \mu_{kl}, \sigma_{kl}, \xi_{kl})$.

3.3 Learning Process

The learning process of the proposed method is performed by minimizing the negative log likelihood of the extreme value distribution $GEV(\mu_{kl}, \sigma_{kl}, \xi_{kl})$ using the observed y_{kl}. The parameters of the GRU $W_z, U_z, \boldsymbol{b}_z, W_r, U_r, \boldsymbol{b}_r, W_h, U_h, \boldsymbol{b}_h$ and the parameters of fully-connected layers $W_1, \boldsymbol{b}_1, W_2, \boldsymbol{b}_2$ are tuned through this process. Given the training data of length T and when $1 \leq k \leq n(T - l < nl < Tl, n \in \mathbb{N})$, the negative log likelihood loss function L_1 is written as:

$$L_1 = -\sum_{k=1}^{n} \ln GEV(y_{kl}; \mu_{kl}, \sigma_{kl}, \xi_{kl}) \tag{9}$$

However, $GEV(y; \mu, \sigma, \xi)$ takes 0 when $1 + \xi \frac{y-\mu}{\sigma} < 0$, and in this case L_1 diverges to infinity. Since the gradient of L_1 is zero in this case, it is no longer possible to learn parameters based on L_1 for data points where $1 + \xi_{kl} \frac{y_{kl}-\mu_{kl}}{\sigma_{kl}} < 0$. During the learning process, such cases often occur and lead to the failure of learning. Therefore we introduced another loss L_2 so that $1 + \xi_{kl} \frac{y_{kl}-\mu_{kl}}{\sigma_{kl}} > 0$ for all data points. The L_2 is defined as:

$$L_2 = \sum_{k=1}^{n} \max\left(-1 - \xi_{kl} \frac{y_{kl} - \mu_{kl}}{\sigma_{kl}}, 0\right) \tag{10}$$

Thus, the objective function L to be minimized in the proposed method is written as follows, with λ as a hyper-parameter.

$$L = L_1 + \lambda L_2 \tag{11}$$

The proposed method minimizes this loss function using Adam [19] to learn parameters.

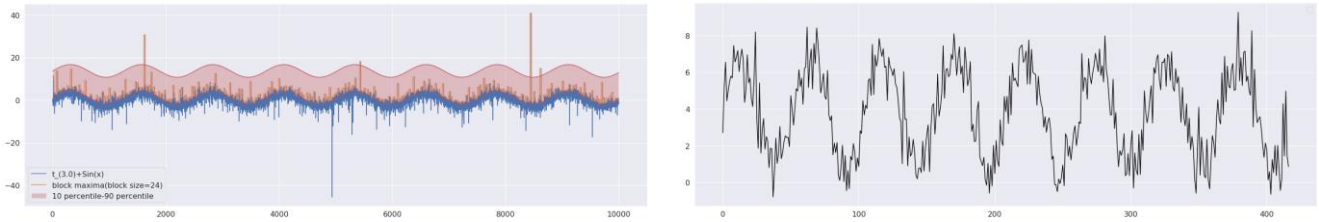
4. Experiments

We performed computational experiments to verify the performance of our proposed method. In this section, we describe the experiment setting and present the empirical results. The experiments are conducted on two artificial datasets and two real datasets. We compared our model with GRU and TPA-LSTM [27] as a baseline.

4.1 Datasets

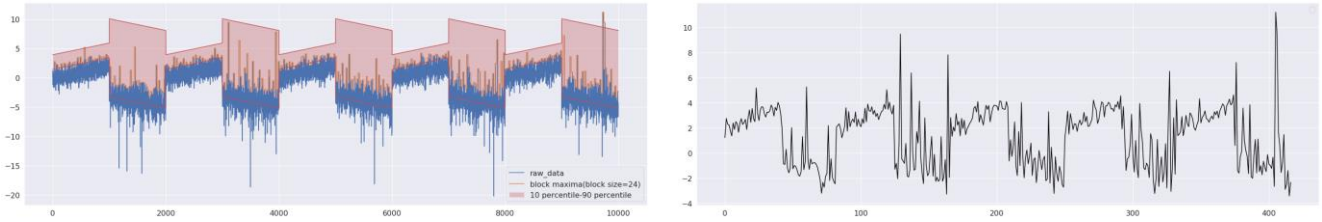
We use two artificial datasets and three real-world datasets.

- Artificial dataset 1: The original artificial dataset created by adding noise arising from a Student's t-distribution to a sinusoidal wave shown in Figure 2. Maxima were sampled with periods 24 points in length. The training data are for 0 - 6000, validation data are for 6000 - 8000, and test data are for 8000 - 10000.
- Artificial dataset 2: The original artificial dataset for an example of a case in which the distribution of data widely changes. We created the dataset by adding noise alternating between a normal distribution and a Student's t-distribution to base values that ascend and descend alternately. These data are shown in Figure 3.
- Unazuki Reservoir dataset: Water inflow data of Unazuki Reservoir and the precipitation at four points around Unazuki Reservoir [1]. We used this 5-dimensional data as input and the maxima of inflow over the following 24 hours as a target. The training data are for May 1, 2016 – April 30, 2017, validation data are for May 1, 2017 - July 31, 2017, and test data are for August 1, 2017 - December 31, 2017.
- Miyagase Reservoir dataset: Inflow data of Miyagase Reservoir [2] similar to Unazuki Reservoir dataset. The training data are for January 1, 2016 -December 31,



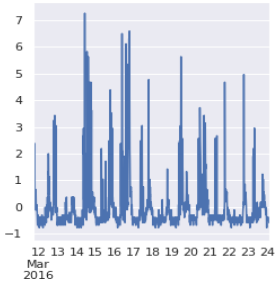
(a) Artificial Dataset 1. (b) Block Maxima of the artificial dataset 1.

Figure 2 Artificial dataset 1. Student's t-distribution noise is added on a sine wave. (a) Raw data of artificial dataset 1. Blue line shows the value of the data set, its orange line shows the maximum values for each period of length 24, and the red shading shows intervals between 1st to 99th percentiles of the GEV distribution of maxima. (b) Series of maxima over periods of length 24 from artificial dataset 1.

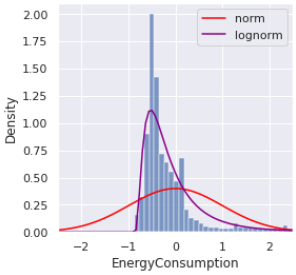


(a) Artificial Dataset 2 (b) Block Maxima of the artificial dataset 2.

Figure 3 Artificial dataset 2. Normal distribution noise is added on an ascending period and Student's t-distribution noise is added on descending period. (a) Raw data of artificial dataset 2. Blue: the value of the data set. Orange: the maximum values for each period. Red shading: intervals between 1st to 99th percentiles of maxima. (b) Series of Maxima taken from artificial dataset 2.



(a) Example of energy consumption data of Energy Dataset.



(b) Histogram of energy consumption data. Normal distribution, and Log-normal distribution are fitted and plotted as colored lines.

Figure 4 Energy Dataset.

2017, validation data are for January 1, 2018 - December 31, 2018, and test data are for July 1, 2019 - December 27, 2020.

- Energy dataset: Data of energy consumption in a low-energy building [5]. We used 27-dimensional data without energy use for lighting as input and the maxima of appliances' energy consumption over the following an hour as a target. The Energy consumption data and its histogram are shown as Figure 4, which shows that the energy consumption is not normally distributed. The training data are for January 11, 2016 - March 10, 2016, validation data are for March 11, 2016 - April 10, 2016, and test data are for April 11, 2016 - May 27, 2016.

4.2 Experimental Settings

In this experiment, the goal of prediction task is predicting maximum value in the future 24 hours. We used GRU for comparison of results of the artificial datasets. For the real-world datasets, the TPA-LSTM method was used in addition to GRU as a baseline.

Our method can directly predict the maximum distribution, but cannot predict maximum itself. Therefore, we picked a mode value of the predicted GEV distribution as a representative value of predicted maxima. On the other hand, GRU and TPA-LSTM are methods to predict output sequence from input sequence. There are two ways to predict future maxima using these sequence to sequence method: picking up the maxima from predicted sequence or predicting maxima directly. In picking up ways, GRU and TPA-LSTM predict future 24 hours' sequence, and then we pick up the maximum of the predicted sequence. We call results of this way as GRU_{PU} and $TPA\text{-}LSTM_{PU}$. In the other way, GRU and LSTM predict maxima of future sequence directly. The results of this directly way are shown as GRU_{Direct} and $TPA\text{-}LSTM_{Direct.}$ GRUs and our method have one hidden layer and the number of hidden units is selected from [10,25,50]. We employed ADAM for the optimization, which learning rate is selected from

[0.1, 0.01, 0.001]. These hyper-parameters are selected by grid-search using validation data.

4.3 Results

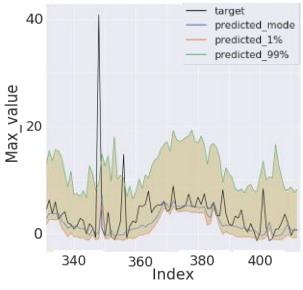
We evaluate the performance of the proposed method and benchmarks using MAE between the maxima of real data and predicted maxima. We report the results of datasets in Table 1.

We confirmed that proposed method can predict GEV parameters accurately using artificial datasets results. We visualized the predicted distribution of the future maxima and its parameters for artificial datasets. For artificial dataset 1, the prediction results by the proposed method are shown in Figure 5a. The predicted values of the parameters μ, σ, ξ for this dataset are shown in Figure 5b. The proposed method can estimate these parameters with RMSE 0.91 for μ, 0.013 for σ, and 0.016 for ξ. The result of artificial dataset 2 is shown in Figure 6a, b. For this dataset, RMSE of predicted parameters is 1.35 for μ, 0.072 for σ, and 0.034 for ξ. These results show that the proposed method can predict the GEV parameters along changes of the true parameters. Although the accuracy of the estimation of the distribution parameters is not yet sufficient and remains an issue for future work, we confirm that our method can detect the variation of future GEV distribution.

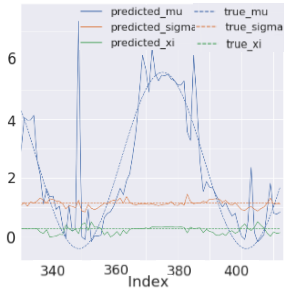
For real-world dataset, we visualize the result for the Energy dataset in Figure 7. Although predictions of GRU_{Direct} are high in the case of real value of less than 0, our proposed method can

Table 1 MAE of predicted values

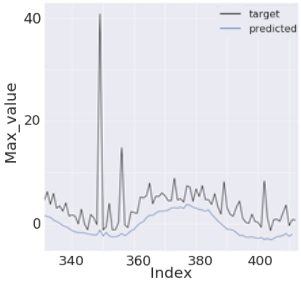
	GRU_{PU}	GRU_{Direct}	$TPA\text{-}LSTM_{PU}$	$TPA\text{-}LSTM_{Direct}$	Proposed method
Artificial 1	3.85	2.69	-	-	2.23
Artificial 2	2.75	1.51	-	-	1.17
Unazuki	0.702	0.497	0.568	0.661	0.444
Miyagase	0.653	0.385	0.431	0.368	0.379
Energy	0.972	0.879	0.964	0.792	0.761



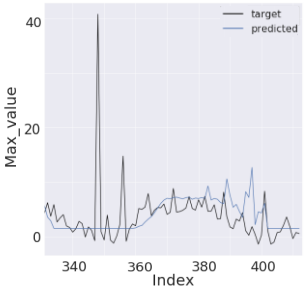
(a) Predicted distribution for the test data of artificial dataset 1.



(b) Predicted parameters of μ (blue), σ (orange), ξ (green) and their true value (dashed line).

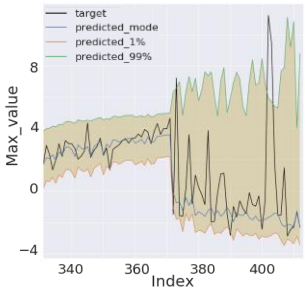


(c) Predicted values of GRU_{PU}.

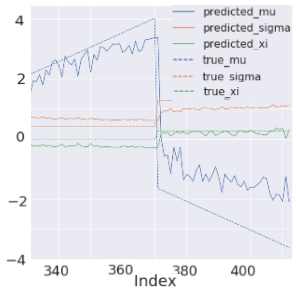


(d) Predicted values of GRU_{Direct}

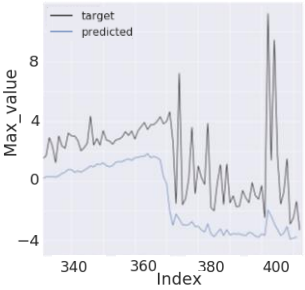
Figure 5 Results of the proposed method for the artificial dataset 1.



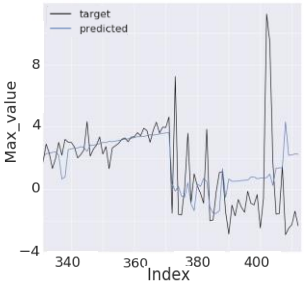
(a) Predicted distribution for the test data of artificial dataset 2.



(b) Predicted parameters of μ (blue), σ (orange), ξ (green) and their true value (dashed line).



(c) Predicted values of GRU_{PU}.



(d) Predicted values of GRU_{Direct}

Figure 6 Results of the proposed method for the artificial dataset 2.

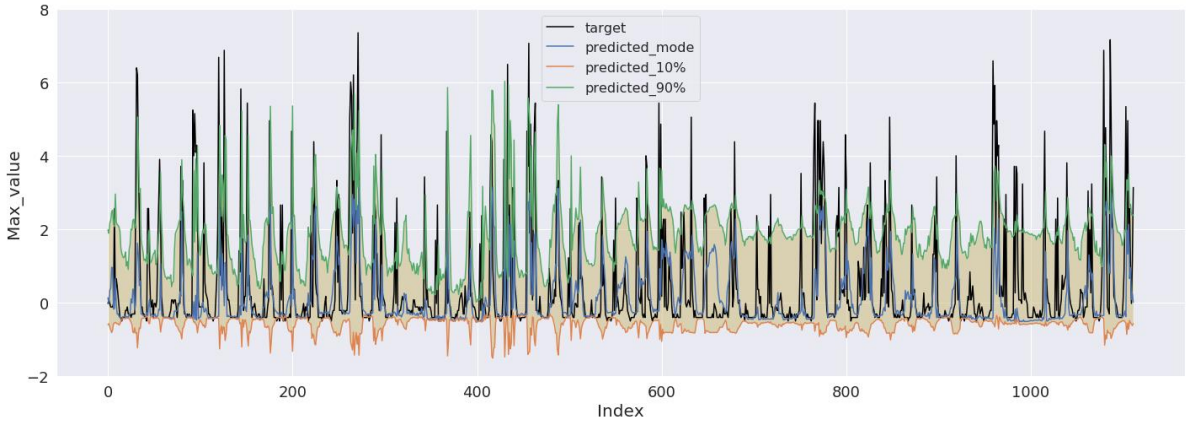
predict the mode correctly. It seems to have learned too much about high values. Our method can avoid this problem and correctly predict both high maxima and low maxima.

5 Discussion

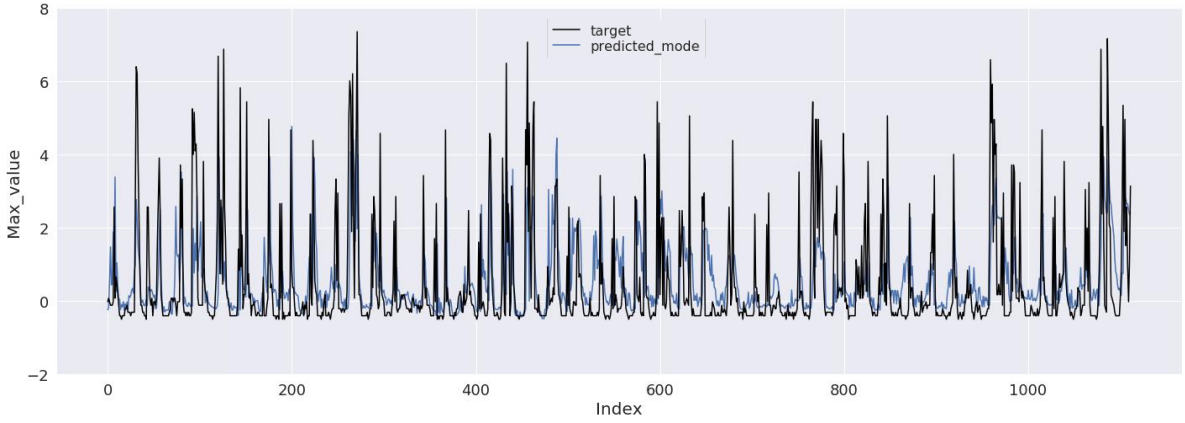
Results of GRU_{PU} and GRU_{Direct} indicate that picking-up approach cannot predict maxima. In the results for the artificial datasets, their predictions are close to the original signals before adding noise to create the artificial datasets rather than the maxima of noised signals. The direct approach can predict maxima more correctly because they are trained using only maxima as target data so that the imbalance of extremes and non-extremes are improved somewhat. This property is confirmed also in the results of real-world data. $TPA\text{-}LSTM_{Direct}$ also has this property.

Compared to direct approach of baselines, Figure 5 and Figure 6 shows that our proposed method can predict maxima more correctly even if the data do not follow a normal distribution. The result of artificial dataset 2 indicates that although GRU_{Direct} can predict on the period of ascending values with normal distribution noise, it cannot predict correctly on the period of descending values with t-distribution noise. We consider that this is because the maxima distribution of the former period is closer to normal distribution than that of the latter period; since the GRU_{Direct} with RMSE loss can handle normal distribution data, it is better at maxima distribution from normal distribution noise. When the number of data points are small, the maxima distribution does not sufficient follow the GEV, so that GRU can predict maxima in this case. However, they cannot predict maxima and cannot capture decreasing trend on the latter period because GRU with RMSE cannot handle heavy tail distribution such as t-distribution. On the other hand, our proposed method can predict both of the former period and latter period. The proposed method maintains similar distribution parameters in the latter period, which implies that the loss function based on GEV distribution is less affected by the peak value generated by the t-distribution. This property is consistent with what was claimed in the previous study [11], and we believe that the proposed method has an advantage on extreme prediction of heavy tail distribution data.

The similar phenomenon was confirmed in the Energy, Unazuki Reservoir and Miyagase Reservoir datasets and also in artificial



(a) Predicted maxima distribution of the proposed method for the test data. Mode (blue), 10th percentile (orange), 90th percentile (green), interval between 10th and 90th percentiles (yellow shade) and real data (black).



(b) Predicted values by GRU_{Direct} (blue) and real data value (black) for the test data.

Figure 7 Results for Energy dataset.

dataset 1. For these datasets, prediction values of GRU and TPA-LSTM tend to have a small range of change and cannot keep up with actual data, which changes intensely at extremes. This is because they tend to be affected by large values of extremes, which has a too small number of samples to learn patterns of change. This weakness results in failure of prediction that predicted values are higher than actual non-extremes and lower than actual extremes. These results indicate that the EVT-based approach is valid for deep learning to improve the prediction for block maxima of real data.

However, it can be confirmed that a problem remains concerning proposed method with respect to the estimation accuracy of the distribution parameters. We can point out the possibility that the length of the period over which the maximum is taken is too short for maxima to follow GEV distribution. On the other hand, if a long period is assumed, there may be cases in which the distribution varies greatly within the period. In addition, increasing the number of data per period leads to a decrease in the number of maxima. This leads to a decrease in the amount of supervisor data that can be used for training, which causes overfitting. To avoid this, in future work we intend to consider improvement by suppressing the time variation of the hidden state of the GRU so that the predicted GEV changes more slowly, as in slow feature analysis (SFA) [30].

6 Conclusion

In this work, we set up the novel problem of predicting the maximum value of time series data in a certain period in the future. For practical use it can be unnecessary to predict the exact timing of extremes, and we tackle alleviated problem of predicting maxima on period of certain length.

We proposed a method to predict maxima by GEV distribution and GRU to solve this problem. Learning parameters using GEV distribution-based loss function often fail because GEV distribution is positive in a limited area of parameter value. Therefore, we introduce a new loss function for the objective function.

In an experiment using an artificial dataset, it was confirmed that the proposed method can detect the dynamics of maxima distribution, and that it is applicable to non-stationary time series extreme value prediction. In experiments using real data, it was confirmed that, in the case of using the modes of predicted distribution as predicted values, it is more accurate than GRU, a commonly used deep learning method. It may have potential for application to a wide variety of data.

On the other hand, the accuracy of parameter estimation is not yet sufficient. One of the causes is considered to be that the length of periods over which maxima are taken is too short for maxima to follow GEV distribution. In future work, we consider to employing the concept of SFA in GRU to improve estimation.

REFERENCES

1. Database of dams. http://mudam.nilim.go.jp/chronology/summary/48, (Accessed on 03/17/2021)
2. Database of dams. http://mudam.nilim.go.jp/chronology/summary/42, (Accessed on 03/17/2021)
3. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: Advances in Neural Information Processing Systems 19 (NIPS' 06). pp. 153 – 160. MIT Press (2007), https://proceedings.neurips.cc/paper/2006/file/5da713a690c067105aeb2fae32403405-Paper.pdf
4. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time series analysis: forecasting and control. John Wiley & Sons (2015). https://doi.org/10.1111/jtsa.12194
5. Candanedo, L.M., Feldheim, V., Deramaix, D.: Data driven prediction models of energy use of appliances in a low-energy house. Energy and Buildings 140, 81–97 (April 2017). https://doi.org/10.1016/j.enbuild.2017.01.083
6. Cheng, L., AghaKouchak, A., Gilleland, E., Katz, R.W.: Non-stationary extreme value analysis in a changing climate. Climatic Change 127(2), 353–369 (Nov 2014). https://doi.org/10.1007/s10584-014-1254-5
7. Cho, K., van Merri¨enboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder–decoder approaches. In: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. pp. 103–111. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/W14-4012
8. Chu, V.W., Wong, R.K., Chi, C., Chen, F.: Extreme topic model for market ealert service. In: 2018 IEEE International Conference on Services Computing (SCC). pp. 145–152 (2018). https://doi.org/10.1109/SCC.2018.00026
9. Coles, S.: An introduction to statistical modeling of extreme values. Springer Series in Statistics, Springer-Verlag, London (2001) 10. Dawson, C., Abrahart, R., Shamseldin, A., Wilby, R.: Flood estimation at ungauged sites using artificial neural networks. Journal of Hydrology 319(1), 391–409 (2006). https://doi.org/10.1016/j.jhydrol.2005.07.032
11. Ding, D., Zhang, M., Pan, X., Yang, M., He, X.: Modeling extreme events in time series prediction. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1114–1122. KDD '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330896
12. El Adlouni, S., Ouarda, T.B.M.J., Zhang, X., Roy, R., Bobee, B.: Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. Water Resources Research 43(3), W03410 (2007). https://doi.org/10.1029/2005WR004545
13. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. 28(4), 594–611 (Apr 2006). https://doi.org/10.1109/TPAMI.2006.79
14. Gilleland, E., Katz, R.W.: extremes 2.0: An extreme value analysis package in r. Journal of Statistical Software, Articles 72(8), 1–39 (2016). https://doi.org/10.18637/jss.v072.i08
15. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation 18(7), 1527–1554 (Jul 2006). https://doi.org/10.1162/neco.2006.18.7.1527
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (Nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735
17. Jenkinson, A.F.: The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Quarterly Journal of the Royal Meteorological Society 81(348), 158–171 (1955). https://doi.org/10.1002/qj.49708134804
18. Kim, S., Kim, H., Lee, J., Yoon, S., Kahou, S.E., Kashinath, K., Prabhat, M.: Deep-hurricane-tracker: Tracking and forecasting extreme climate events. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1761–1769 (2019). https://doi.org/10.1109/WACV.2019.00192
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. pp. 1–15 (2015)
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (Oct 2017). https://doi.org/10.1109/ICCV.2017.324
21. Liu, Y., Racah, E., Prabhat, Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., Wehner, M.F., Collins, W.D.: Application of deep convolutional neural networks for detecting extreme weather in climate datasets. In: ABDA'16 – International Conference on Advances in Big Data Analytics. pp. 81–88 (2016), http://worldcomp-proceedings.com/proc/p2016/ABD6152.pdf
22. Mentaschi, L., Vousdoukas, M., Voukouvalas, E., Sartini, L., Feyen, L., Besio, G., Alfieri, L.: Non-stationary extreme value analysis: a simplified approach for earth science applications. Hydrology and Earth System Sciences Discussions pp. 1–38 (02 2016). https://doi.org/10.5194/hess-2016-65

23. Mishra, A.K., Desai, V.R.: Drought forecasting using stochastic models. Stochastic Environmental Research and Risk Assessment 19(5), 326–339 (Nov 2005). https://doi.org/10.1007/s00477-005-0238-4
24. Moran, P.A.: Hypothesis testing in time series analysis. Journal of the Royal Statistical Society: Series A (General) 114(4), 579–579 (1951). https://doi.org/10.2307/2981095
25. Nayak, A.K., Sharma, K.C., Bhakar, R., Mathur, J.: Arima based statistical approach to predict wind power ramps. In: 2015 IEEE Power Energy Society General Meeting. pp. 1–5 (2015). https://doi.org/10.1109/PESGM.2015.7286237
26. Sartini, L., Cassola, F., Besio, G.: Extreme waves seasonality analysis: An application in the mediterranean sea. Journal of Geophysical Research: Oceans 120(9), 6266–6288 (09 2015). https://doi.org/10.1002/2015JC011061
27 Shih, S. Y., Sun, F. K., & Lee, H. Y. (2019). Temporal pattern attention for multivariate time series forecasting. Machine Learning, 108(8), 1421-1441.
28. van den Berg, J., Candelon, B., Urbain, J.P.: A cautious note on the use of panel models to predict financial crises. Economics Letters 101(1), 80–83 (2008). https://doi.org/10.1016/j.econlet.2008.06.015
29. Vousdoukas, M.I., Voukouvalas, E., Annunziato, A., Giardino, A., Feyen, L.: Projections of extreme storm surge levels along europe. Climate Dynamics 47(9), 3171–3190 (Nov 2016). https://doi.org/10.1007/s00382-016-3019-5
30. Wiskott, L., Sejnowski, T.J.: Slow feature analysis: Unsupervised learning of invariances. Neural Comput. 14(4), 715–770 (Apr 2002). https://doi.org/10.1162/089976602317318938
31. Yaacob, A.H., Tan, I.K.T., Chien, S.F., Tan, H.K.: Arima based network anomaly detection. In: 2010 Second International Conference on Communication Software and Networks. pp. 205–209 (2010). https://doi.org/10.1109/ICCSN.2010.55