Feature scaling and attention convolutions for extreme precipitation prediction

Anonymous Author(s)

ABSTRACT

This work presents an in-depth exploration of the use of deep neural networks for the prediction of short-term severe precipitations. Notably, we are interested in the efficacy of attention-based network topologies for forecasting, specifically convolution vs. transformer approaches for forecasting, and the scaling methodologies for weather characteristics. As a consequence of our investigations, we have determined that feature scaling is crucial in the context of severe precipitation prediction. Attention-models with robust scalers outperform models with conventional scalers or models with raw inputs by a significant margin. An additional attention-augmented convolutional model, referred to as convolutional self-attention, surpasses the standard convolutional model by a wide margin.

CCS CONCEPTS

• **Applied computing** → **Environmental sciences**.

KEYWORDS

precipitation forecasting, neural networks, convolution, feature scaling

ACM Reference Format:

Anonymous Author(s). 2022. Feature scaling and attention convolutions for extreme precipitation prediction. In *Proceedings of KDD '22: 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Climate change is one of the most critical issues affecting humanity's future, and it will only grow in importance as extreme weather events increase in frequency and intensity. Among the most recent examples is the European floods that killed over 242 people and cost over 10 billion euros in damage in parts of Germany, France, Switzerland, Italy, and the United Kingdom. Total damages from Atlantic hurricanes in the United States are expected to total $67 billion in 2021. Hurricane Ida struck the Louisiana coast in August 2021, causing $64.5 billion in damage and killing 96 people. As climate change intensifies, these natural disasters will become more prevalent and destructive [14, 16].

To date, long-term weather and climate prediction models such as Extreme Value Analysis (EVA) and other numerical weather prediction (NWP) models are becoming increasingly inaccurate and unstable due to the unpredictability of extreme weather events as a direct result of climate change [16, 17]. It is in our great interest to provide an accurate and reliable system to predict and classify extreme weather events and estimate the damage caused.

Machine learning for climate forecasting [11, 12, 15] has risen in popularity as a consequence of recent developments in this field, such as deep learning [5, 7, 9, 13, 18], as well as the ability of some machine learning techniques to outperform numerical models in specific tasks [15]. A widely applied neural network architecture for this task is the Convolutional Neural Network (CNN) [9, 13, 18]. Convolutional layers in CNNs learn convolutional filters of size $K \cdot K$, with input and output dimensions D_{in} and D_{out}, respectively. The layer is parametrized by a 4D kernel tensor $\boldsymbol{W}$ of dimension $K \cdot K \cdot D_{in} \cdot D_{out}$ and a bias vector $\boldsymbol{b}$ of dimension D_{out}. As a result, CNNs are not spatially invariant to the input data and do not encode the position of each value in the input matrix.

Despite the success in applying CNNs in these studies, there are some challenges that should be addressed to further strengthen the applications of deep learning techniques in climate and environmental sciences. One major challenge is that, unlike static images in ImageNet, climate data is much more complex in nature, often spatio-temporal, highly nonlinear, noisy, non stationary, and correlated [6]. Consequently, forecasting algorithms must be developed appropriately in order to successfully leverage the potential of deep learning architectures.

With these limitations in mind, this study focuses on feature scaling methodologies and architectural designs that increase the accuracy and application of severe precipitation forecasting and so our contributions are two-fold. By applying Robust Scalers to a variety of precipitation variables, we have effectively normalised temporal observations and obtained considerably improved detection rates of extreme events. By supplementing the standard convolutional layer with a multi-head self-attention module, we have overcome a major drawback with convolutional neural networks, since the self-attention's field is always the whole input matrix. This considerably minimises the estimated errors in the final findings. Furthermore, we present a unique self-attention

KDD '22, August 14-18, 2022, Washington DC Convention Center

https://doi.org/XXXXXXX.XXXXXXX

enhanced convolutional layer that enables us to gain the advantages of the attention network on multi-dimensional feature maps. Compared to the Transformer attention mechanism, the suggested attention convolution achieves statistically comparable accuracy in identifying extreme events but with much reduced costs. We show its application to severe precipitation forecasting using real-world datasets.

2 RELATED WORK

2.1 Convolutional precipitation forecasting

Convolutional networks have been shown to be particularly effective in computer vision applications. According to [9], such networks have been used to examine large-scale "extreme precipitation circulation patterns" (EPCP). A traditional CNN classifier was developed to distinguish between EPCP and non-EPCP days. The CNN model accurately identifies 91% of severe precipitation days in the US midwest as EPCPs, with an overall accuracy of 88% across both classes. According to the classification results, this study also looked at how often EPCPs happen and how much rain there is on days when EPCPs happen.

Other network designs, such as the Convolutional LSTM Network [18] and the Temporal Fusion Transformer (TFT) [7], have been used to tackle comparable precipitation forecasting issues in addition to traditional convolutional networks.

These efforts, particularly [9], influenced our research and experimentation. However, our work puts a greater emphasis on developing more effective and efficient network topologies to increase prediction accuracy and deep learning model performance.

2.2 Attention mechanisms

Attention accentuates the relevant features of the input data while omitting the less relevant ones, with the concept that the network should spend more computational resources on a smaller but critical portion of the data [10]. Self-Attention, also known as intra attention, connects distinct points of a single sequence to compute a representation of the same sequence [19]. The self-attention mechanism allows the inputs to interact with one another (self) and determine who they should focus on more (attention). These interactions and attention scores are aggregated in the outputs. Self-attention, as compared to other attention methods, minimizes the total computing complexity of each layer of calculations by lowering the number of consecutive operations required. It may also reduce the maximum path length between any two input and output locations in a network with multiple layer types.

Using self-attention together with convolutions is shared by recent work in Natural Language Processing [21] and Reinforcement Learning [20, 21]. The convolution technique has a critical weakness in that it only works on local neighbourhoods, which means it misses out on global data. Self-attention, on the other hand, is capable of capturing long-range interactions that have been mostly applied to sequential modelling problems. In particular, [8] further proves that a multi-head self-attention layer is "at least as expressive as any convolutional layer".

The Transformer attention mechanisms revolutionizes the implementation of attention by dispensing recurrence and convolutions, relying solely on a self-attention mechanism [19].

3 METHODS

This study aims at tackling the problem of predicting and classifying extreme precipitation days by combining state-of-the-art deep learning representations with novel feature scaling and architectural design techniques. To achieve this, we build a self-attention augmented convolutional neural network to predict severe precipitation days using daily SLP and 500-hPa GPH anomalies. Each day's input data is processed via various layers to produce an output categorization of extreme precipitation (Class 1) or non extreme precipitation (Class 0). Our model receives a three-dimensional matrix with dimensions of $15 \cdot 35 \cdot 2$ for each day (i.e., latitude $\cdot$ longitude $\cdot$ 2 input variables). Precipitation data is used to produce ground-truth labels for model training.

In the following sections, we introduce the datasets we used in the experiments and the corresponding data processing methods. Then we formally describe the proposed Robust Scaler for precipitation features and the Attention-Augumented Convolutions network architecture.

3.1 Weather datasets

We compute extreme precipitation days across the United States Midwest from 1981 to 2019 using PRISM [3] 4 km daily precipitation [4]. We calculate the mean precipitation across a rectangular region including the Upper Mississippi Watershed and the eastern section of the Missouri Watershed ($37N$ to $48N$, $104W$ to $86W$) for each day. We determine the 95th percentile ($p95$) of daily precipitation using this regional daily precipitation time-series, and extreme precipitation days are classified as those that surpass the $p95$ threshold. The algorithm for this is presented in 1. We examine excessive precipitation days throughout all seasons since extreme precipitation can occur throughout the year in the Midwest.

The NCEP/NCAR-R1 reanalysis dataset [2], which offers worldwide coverage at $2.5 \cdot 2.5$ horizontal resolution, is used to determine daily mean sea level pressure (SLP) and 500-hPa geopotential height (GPH) anomalies. We examine atmospheric variables over a broader geographical region that includes the entire United States and nearby waters ($20N$ to $55N$ and $140W$ to $55W$). To eliminate uniform thermal dilatation produced by tropospheric warming, we first deduct the area-weighted average 500-hPa GPH trend over the atmospheric domain, maintaining spatially non-uniform

Algorithm 1: Precipitation percentile calculation

Input: array x_i, percentile m
Sort array.
$k = m * (n - 1)$
$f = floor(k)$
$c = ceil(k)$
if $f == c$ **then**
  Return x_f.
end if
$d_0 = x_f * (c - k)$
$d_1 = x_c * (k - f)$
Return $d_0 + d_1$

changes in 500-hPa GPH that may effect severe precipitation.

We use zonal wind (u), meridional wind (v), and specific humidity (q) fields from the NCEP/NCAR-R1 reanalysis to evaluate differences in moisture flux on days with extreme precipitation and non-extreme precipitation patterns.

In addition to these datasets, the Global Historical Climatology Network daily (GHCNd) [1], a centralized database including daily climate summaries from satellites around the world, is adopted to evaluate our models' performance as an unseen real-world test dataset. The Global Historical Climate Network (GHCNd) is a collection of daily climate data from a range of sources that has been aggregated and subjected to a standard set of quality assurance criteria.

3.2 Feature scaling

Feature scaling is to center and scale samples independently of each feature, by computing the relevant statistics on the samples in the training set. In order to remove seasonal variability in the datasets, we apply Standard Scaler (1) and Robust Scaler (2) to the detrended daily GPH measurements and the SLP measurements. Afterwards, we compare the model performance under the following three different cases:

- **Raw inputs** can be defined as training models without applying any scalers. The purpose is to demonstrate the importance of feature scaling on the precipitation forecasting tasks.
- **Standard scaler** The daily normalized measurements are calculated by subtracting the grid-cell calendar-day mean from the daily measurements, and then performing division by the grid-cell calendar-day standard deviation. We henceforth apply a threshold to the calculated results to discover extreme precipitation days. As the most popular feature scaling method, the standard scaler method assumes a normal distribution within each feature and expunges the seasonal variability by removing the mean and scaling to unit variance. However, it may suffer from outliers or noises.
- **Robust scaler** The daily measurements are calculated by substracting the grid-cell calendar-day median from the daily measurements, and then scaling these data points according to the quantile range, Interquartile Range (IQR). IQR can be defined as the range between the $1st$ quartile ($25th$ quantile) and the $3rd$ quartile ($75th$ quantile). Unlike the standard scaler, features calculated from the robust scaler are robust to outliers, hence the name.

$$v = \frac{X - \mu}{\sigma} \tag{1}$$

where:

v is the computed sample (measurement) of each feature.
X is the value of the variable at the grid-cell.
μ is the grid-cell calendar-day mean.
σ is the grid-cell calendar-day standard deviation.

$$v = \frac{X - \tilde{\mu}}{Q_3(x) - Q_1(x)} \tag{2}$$

where:

v is the computed sample (measurement) of each feature.
X is the value of the variable at the grid-cell.
$\tilde{\mu}$ is the grid-cell calendar-day median.
$Q_3(x)$ is the $75th$ quantile of the grid-cell calendar-day samples.
$Q_1(x)$ is the $25th$ quantile of the grid-cell calendar-day samples.

The impacts of each scaler on model performance can be found in Section 4.

3.3 Attention augmented convolutions

We propose concatenating convolutional feature maps with a set of feature maps obtained by self-attention to enhance convolutional operators with this self-attention mechanism. Such an operation is defined below in Formula 5.

Given an input tensor of shape (H, W, D), which represents the height, width, and the feature depth of the anomalies respectively, we flatten it to a matrix $A(1, H \cdot W \cdot D)$ and perform multihead attention as proposed in the Transformer architecture [19]. The output of the self-attention mechanism for a single head h can be formulated as below (3):

$$Attn(X) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

where:

$Q = (X * W_q)$
$K = (X * W_k)$
$V = (X * W_v)$
W_q, W_k, W_v are the learned linear transformations that map the input X to queries (Q), keys (K), and values (V) respectively.

A single attention score calculation uses $O(A^2 \cdot h)$ space, with the multi-head attention calculation using $O(A^2 \cdot h^2)$ space. This is much more efficient than many other previous attention-based augmentations to the convolutional layer

due to omitting relative position embeddings. The multi-head attention calculation is performed on a single head at a time, and the output of the multi-head attention is concatenated to the output of the convolutional layer.

$$MHA(X) = Concat[Attn_1(X), Attn_2(X), ..., Attn_N h(X)] * W_m h \quad (4)$$

where:

$W_m h$ is the learned linear transformation that maps the concatenated attention scores to the output of the multi-head attention.

Finally, the multi-head attention scores are concatenated to the output of the convolutional layer:

$$AAConv(X) = Concat[Conv(X), MHA(X)] \quad (5)$$

where:

$Conv(X)$ is the output of the convolutional layer, and $MHA(X)$ is the output of the multi-head attention.

3.4 Network architectures

One of the proposed network architectures, the self-attention augmented convolutional neural network, is shown in Figure 1. Given a multi-channel input tensor formed by stacking the input maps of mean SLP and GPH anomalies, the network predicts if the precipitation of the area is above the $p95$ threshold. The network comprises two attention augmented convolutional layers, each with a filter of 16, a max-pooling and dropout layer.

A highway layer followed by a dense layer is utilized to make predictions. Furthermore, while the predictands of our existing model were found to be quite stable, the highway network was added to provide feature depth to the model, thus improving its predictive accuracy.

The Transformer architecture is applied in a similar fashion. Given the multi-channel inputs from the dataset, the model firstly flattens the matrix into a 1-dimensional array before passing it through a transformer. The transformer consists of a multi-head attention mechanism followed by a feed-forward network. The transformer is applied to the input tensor of shape (H, W, D) and outputs a tensor of shape (H, W, D). We further apply the highway network to the output of the transformer. The output of the network is then passed through a dense layer with a single scalar output.

The main benefit of the Transformer neural network, as opposed to the self-attention augmented convolutional network, is its ability to learn the attention mechanism from the input data. This is achieved by using the self-attention mechanism to learn the attention weights. As a result, the transformer is much deeper than the self-attention augmented convolutional network. This means that a more complex, non-linear function can be learned by training the transformer.

3.5 Training and validation

Each model was trained for 100 epochs with the batch size of 32. The Adam optimizer was used with a constantly decaying learning rate scheduler, with the minimum and maximum learning rates as 10e-4 and 10e-2, respectively. The categorical crossentropy error was implemented as the loss function.

3.6 Experiments

We first evaluate the impact of feature scaling mechanisms on precipitation forecasting performance: raw feature inputs (without scaling), standard scaled features, and robust scaled features. We have tested several other common feature scaling approaches, such as the Min-Max Scaler, the Max-Abs Scaler, and the Power Transformer Scaler. The specifics of such research (including outcomes, etc.) are not given in this study due to space constraints.

To quantify the performance of the proposed models, we implement the conventional CNN model and the Transformer as the baseline. The proposed Self-Attention CNN models are constructed and trained in two variants: without highway networks (SAConvNet) and with highway networks (SAConvNet + Highway). On the same datasets, two different feature scaling algorithms are used to compare the performance of these two new models to the conventional CNN and Transformer models.

To investigate the model's generalization and robustness, we train five models with identical hyperparameters in each trial. Each model is used to forecast excessive precipitation at five distinct thresholds, notably $p91$ - $p95$. These precipitation thresholds can be used to simulate real-world extreme precipitation scenarios.

Each experiment is repeated with 5 times. The arithmetic Mean of the performance metrics are calculated, together with the Standard Error of the Mean (SEM - Standard Error of the Mean, calculated by taking the standard deviation and dividing it by the square root of the sample size). This is to remove factors of randomness that might be introduced by the model training or evaluation process.

All possible prediction results can be divided into the following four cases:

- True Positive (TP): actual extreme weather events are classified as extreme.
- True Negative (TN): actual normal weather events are classified as normal.
- False Positive (FP): actual normal weather events are classified as extreme events (e.g., false alarms).
- False Negative (FN): actual extreme weather events are classified as normal.

Accordingly, the performance of the proposed models can be evaluated by using the following evaluation metrics:

- **Accuracy** measures the proportion of the correctly classified extreme weather events to the total weather data samples.

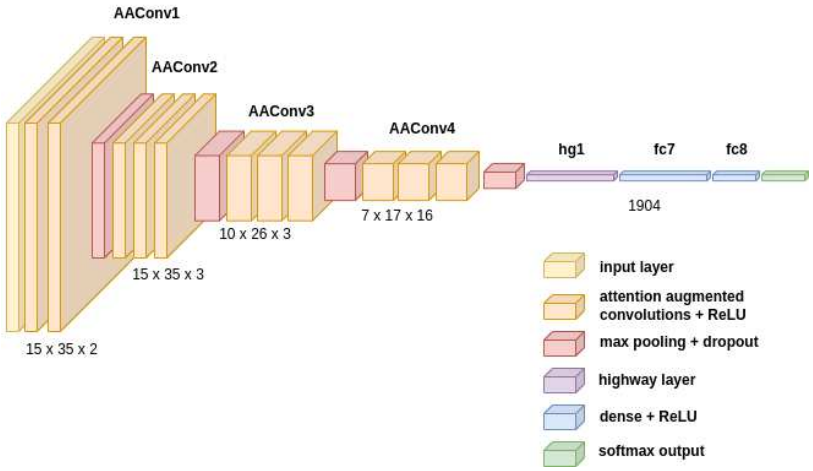

Figure 1: Proposed Network Architecture

- **Precision** measures the correctly classified extreme weather events, in the proportion of the total classified extreme weather events.
- **Recall or detection rate** measures the proportion of correctly classified extreme weather events by the models, in the actual extreme weather events, to measure the ability for detecting extreme weather events.
- **F1 score** measures the harmonic mean of the precision and recall.

It is expected that the models capture as many extreme events as possible, so as to effectively save lives and reduce damages. Therefore, the metrics above, accuracy, precision and recall, are considered as crucial ones when evaluating model performance.

4 RESULTS

4.1 Impact of feature scaling

This section shows the impact of feature scaling mechnisms on model performance. Due to length limitations, we only showcase the performance of the proposed self-attention augmented convolution model, with Standard Scaler and Robust Scaler. The case of other models is very similar with the proposed model. And these two scalers significantly outperform other scalers.

As is shown in the table 1, it is not surprising that, without feature scaling, the model predictions generate predominantly arbitrary results. This is caused by the seasonal variability of the weather data, which hence must be removed in feature engineering.

Meanwhile, both feature scalers achieve particularly promising results, with significant improvements on key performance metrics such as recall and accuracy. Although the detection rate under robust scaler (0.9265) is slightly lower than that under standard scaler, it is worth noting that the SEM of the model performance is much lower; specifically by three times. This demonstrates the major advantages of the robust scaler in handling outliers.

Figure 2 illustrates a visual representation of the feature scaling results. The GPH features are displayed on the left panel, while the SLP features are shown on the right panel.

4.2 Model performance

The model performance results of the experiments are as follows.

As is shown in the Table 2 & 3, the proposed Self-Attention CNN model with the highway networks achieves the best overall accuracy of 97% across both classifications (extreme precipitation vs. non-extreme precipitation), a 12% improvement as compared to the classical CNN model result by [9]. It is capable of accurately identifying more than 88% of severe precipitation days as extreme precipitation days (EPD), comparable to the classical CNN model. Extreme precipitation occurred on 64% of days classified as EPD patterns, 150% better than the classical CNN model result. And, only fewer than 12% of days, classified as non-EPD patterns, resulted in severe precipitation.

Disregarding the highway network, the proposed Self Attention CNN model achieves even better performance, nearly 94% recall rate, in successfully identifying severe precipitation days as EPD, with only as low as 6% of EPD classified as non-EPD. The overall accuracy of Self Attention CNN is still above 90%, which is still well above the classical CNN model results by 4%.

The proposed Self Attention CNN model also outperforms Transformer by at least 2% - 3% in terms of detection rate of extreme events. Such improvements are statistically significant, based on the calculated SEM. And the improvements are applicable to both performance of standard and robust scaler.

However, in terms of training costs, the proposed Self Attention CNN model is much more cost-efficient than the

Table 1: Feature Scaling on Model Performance (SAConvNet).

Scaling Methods	Loss	Accuracy	Precision	Recall	AUC	F1_Score
Raw Feature	0.2200	0.9499	0.500	0.028	0.8559	0.9496
	±0.05	±0.0001	±0.0000	±0.0001	±0.0031	±0.0003
Standard Scaler	0.2357	0.9052	0.3384	**0.9369**	0.9678	0.9048
	±0.0127	±0.0074	±0.0139	±0.0098	±0.0021	±0.0074
Robust Scaler	0.2858	0.8930	0.3119	**0.9265**	0.9624	0.8927
	±0.0112	±0.0056	±0.0115	±0.0030	±0.0012	±0.0056

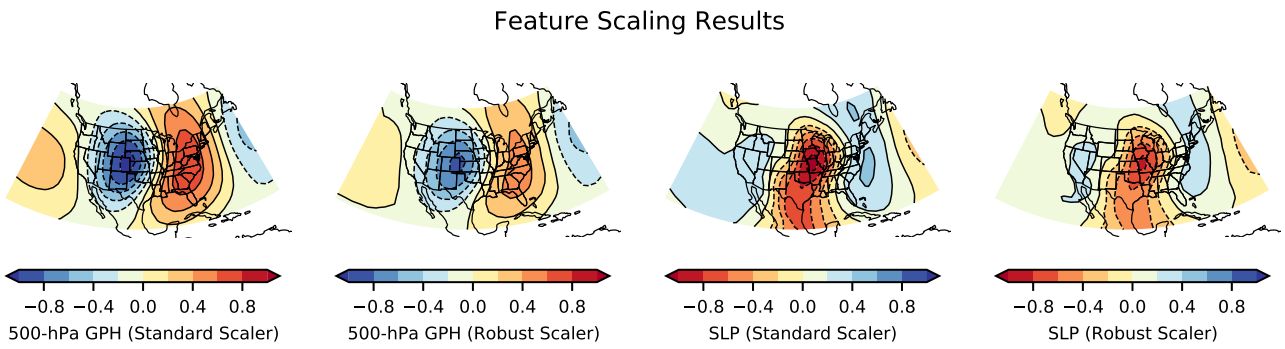

Figure 2: Precipitation Feature Scaling Results

Table 2: Precipitation Model Performance Under Standard Scaler.

Models	Training Time (s)	Loss	Accuracy	Precision	Recall	AUC	F1_Score
ConvNet	87	0.4177	0.8671	0.2613	0.9060	0.9533	0.8669
	±9	±0.0259	±0.0030	±0.0050	±0.0058	±0.0014	±0.0030
SAConvNet	274	0.2357	0.9052	0.3384	**0.9369**	0.9678	0.9048
	±17	±0.0127	±0074	±0139	±0.0098	±0.0021	±0.0074
SAConvNet + Highway	195	0.0876	**0.9688**	**0.6423**	0.8830	0.9763	0.9693
	±13	±0.0079	±0.0034	±0.0308	±0.0061	±0.0011	±0.0032
Transformer	3556	0.1079	0.9603	0.5667	0.9111	0.9759	0.9603
	±64	±0.0052	±0.0022	±0.0168	±0.0034	±0.0007	±0.0023
Transformer + Highway	5298	0.1298	0.9462	0.5023	0.9167	0.9748	0.9461
	±711	±0.0171	±0.0098	±0.0436	±0.0031	±0.002	±0.0099

Transformer, with a training runtime of 276 seconds, due to the simplicity of its architectural design. In contrast, the training runtime of the Transformer is 5855 seconds.

Table 3: Precipitation Model Performance Under Robust Scaler.

Models	Training Time (s)	Loss	Accuracy	Precision	Recall	AUC	F1_Score
ConvNet	84	0.4177	0.8671	0.2613	0.9060	0.9533	0.8669
	±7	±0.0259	±0.0030	±0.0050	±0.0058	±0.0014	±0.0030
SAConvNet	276	0.2858	0.8930	0.3119	**0.9265**	0.9624	0.8927
	±25	±0.0112	±0.0056	±0115	±0.0030	±0.0012	±0.0056
SAConvNet + Highway	177	0.1202	0.9616	0.5983	0.8598	0.9687	0.9616
	±17	±0.0175	±0.0059	±0.0441	±0.0296	±0.0027	±0.0059
Transformer	5855	0.1415	0.9514	0.5173	0.9153	0.9727	0.9513
	±394	±0.0109	±0.0049	±0.0309	±0.0046	±0.0005	±0.0050
Transformer + Highway	3717	0.1321	**0.9623**	**0.5908**	0.9032	0.9711	0.9622
	±201	±0.0104	±0.0045	±0.0368	±0.006	±0.0009	±0.0046

Table 2 & 3 show a comprehensive comparison of model performance under standard and robust scaler. Predominantly spearking, the robust scaler achieves very similar performance results to standard scaler, or slightly better, but with much lower SEM. This is applicable to all models including the CNN, the proposed Self Attention CNN and the Transformers.

Figure 3 presents a comprehensive visual representation of the proposed Self Attention CNN model on the GHCNd database.

5 CONCLUSIONS AND FUTURE WORKS

This paper focuses on feature scaling techniques and network architectural designs that improve the accuracy and applicability of extreme precipitation forecasting. To achieve these goals, feature scaling mechanims are applied to different deep learning models to quantify their impacts on forecasting performance, and (self) attention mechanims are adopted in model designs to enhance convolution-based classification of extreme weather events. These efforts have led to significant improvements in detection rates of precipitation events and model efficiency. The applicability of these models and mechanisms has been demonstrated in experiments using real-world datasets.

In general, the proposed models are capable of identifying extreme weather days based mostly on climate variables. These variables are consistent with the pattern of high moisture flux. Thus, despite not having any external geographic information such as maps in the extreme precipitation time-series, the proposed models are still capable of uncovering physically meaningful features in the regions.

On the other hand, some extremes might still be missed by the algorithms. This is because extreme precipitation is often controlled by localized processes not reflected in the regional-mean daily precipitation. Additional input variables that reflect smaller-scale meteorological processes (e.g., vertical velocity, wind shear, and convective available potential energy) could further improve the models' accuracy for forecasting-related applications. The outputs of the models could likewise be adjusted to concentrate on a smaller region or to capture days with sub-regional variability.

Although the models show improved predictive accuracy across most tasks, the area on which we focused our extreme precipitation data may not completely reflect the model's predictive power. As addressed above, future work can include creating a better, possibly more scalable model to handle large amounts of data. Few-shot learning and meta-learning could also be used to significantly accelerate model convergence.

Additionally, by incorporating different types of information, such as geographical and climate data, the models could provide insight into different causes of extreme precipitation and extreme weather events.

Finally, because the models are trained on variables that are available globally, they could be readily retrained to analyse extreme precipitation and possibly even additional types of extreme weather in other regions, such as non-US regions. This approach could be useful in regions where the limitations of climate models in simulating precipitation processes lead to high uncertainty in future changes in extreme precipitation.

Although the initial case study is confined to small data sets, the results demonstrate that deep learning can provide critical insight into the physical processes underlying changes in climate extremes. Due to its generalizability to a range of extreme events and regions, it represents a promising tool for both scientific understanding and the planning and adaptation required to reduce vulnerability to current and future climate change.

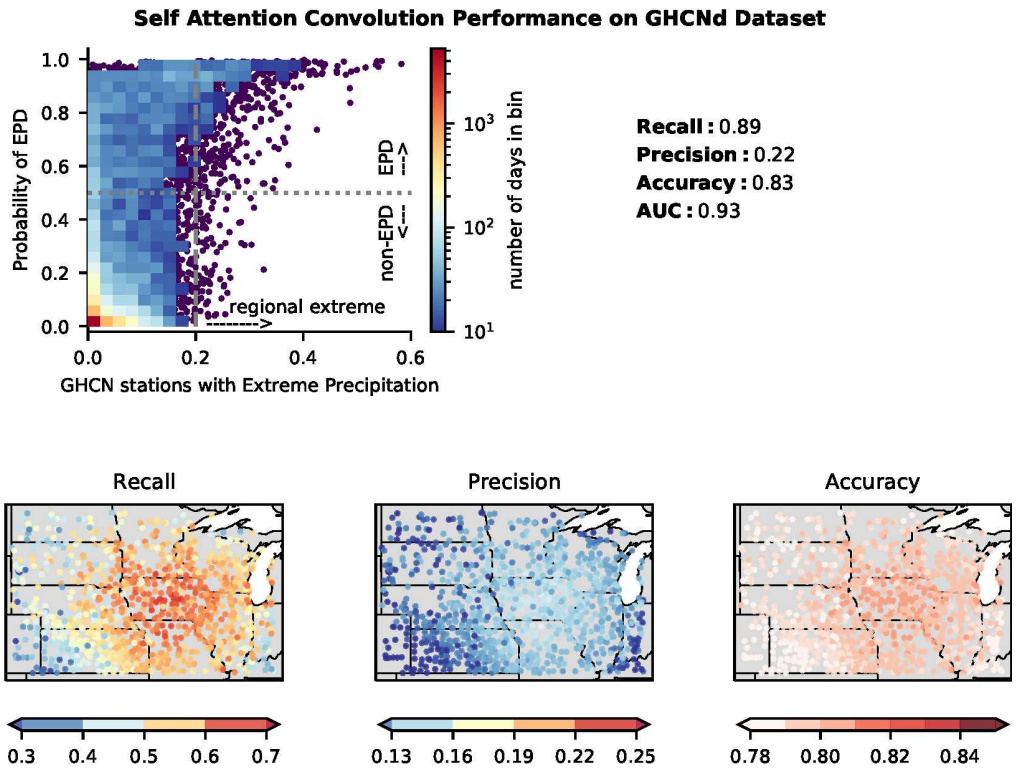

Figure 3: Self Attention Convolutions Performance

REFERENCES

[1] Global historical climatology network daily. URL https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily.

[2] Ncep/ncar reanalysis. URL https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html.

[3] Prism dataset, . URL https://prism.oregonstate.edu/.

[4] Prism daily precipitation, . URL https://ftp.prism.oregonstate.edu/daily/ppt/.

[5] Pedram Hassanzadeh Ashesh Chattopadhyay, Ebrahim Nabizadeh. Analog forecasting of extreme-causing weather patterns using deep learning, 2020.

[6] Pasha S. Chattopadhyay A., Hassanzadeh P. Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data, 2020.

[7] Daniel Salles Civitarese, Daniela Szwarcman, Bianca Zadrozny, and Campbell Watson. Extreme precipitation seasonal forecast using a transformer neural network, 2021.

[8] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers, 2020.

[9] Frances V. Davenport and Noah S. Diffenbaugh. Using machine learning to analyze physical causes of climate change: A case study of u.s. midwest extreme precipitation, 2021.

[10] Lindsay GW. Attention in psychology, neuroscience, and machine learning, 2021.

[11] Sijie He, Xinyan Li, Timothy DelSole, Pradeep Ravikumar, and Arindam Banerjee. Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. *CoRR*, abs/2006.07972, 2020. URL https://arxiv.org/abs/2006.07972.

[12] N Jones. How machine learning could help to improve climate forecasts. *Nature*, 548(379), 2017. doi: 10.1038/548379a.

[13] Yunjie Liu, Evan Racah, Prabhat, Joaquin Correa, Amir Khosrowshahi, David Lavers, Kenneth Kunkel, Michael F. Wehner, and William D. Collins. Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *CoRR*, abs/1605.01156, 2016. URL http://arxiv.org/abs/1605.01156.

[14] Stephen Ornes. Core concept: How does climate change influence extreme weather? impact attribution research seeks answers. *Proceedings of the National Academy of Sciences*, 115(33):8232–8235, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1811393115. URL https://www.pnas.org/content/115/33/8232.

[15] et. al. Peter B. Gibson, William E. Chapman. Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Commun Earth Environ 2*, 2021. URL https://doi.org/10.1038/s43247-021-00225-4.

[16] B. D. Santer, K. E. Taylor, T. M. L. Wigley, T. C. Johns, P. D. Jones, D. J. Karoly, J. F. B. Mitchell, A. H. Oort, J. E. Penner, and V. et al. Ramaswamy. A search for human influences on the thermal structure of the atmosphere, 2022.

[17] Sonia I. Seneviratne, Neville Nicholls, David Easterling, Clare M. Goodess, Shinjiro Kanae, James Kossin, Yali Luo, Jose Marengo, Kathleen McInnes, Mohammad Rahimi, and et al. *Changes in Climate Extremes and their Impacts on the Natural Physical Environment*, page 109–230. Cambridge University Press, 2012. doi: 10.1017/CBO9781139177245.006.

[18] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting, 2015.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[20] Baosong Yang, Longyue Wang, Derek Wong, Lidia S. Chao, and Zhaopeng Tu. Convolutional self-attention networks, 2019.

[21] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension, 2018.