# Enabling Smart Dynamical Downscaling of Extreme Precipitation Events with Machine Learning

**Xiaoming Shi**[1,2]

[1]Division of Environment and Sustainability, Hong Kong University of Science and Technology
[2]Department of Civil and Environmental Engineering, Hong Kong University of Science and Technology

**Key Points:**

- Dynamical downscaling at ∼1 km resolution produces reliable estimations of extreme rainfall but is computationally expensive.
- Machine learning (ML) makes smart dynamical downscaling (SDD) possible, where ML models filter out irrelevant large-scale patterns.
- We demonstrate that SDD can be enabled by deep neural networks, which do not necessarily have to involve sophisticated structures.

---

Corresponding author: Xiaoming Shi, `shixm@ust.hk`

## Abstract

The projection of extreme convective precipitation by global climate models (GCM) exhibits significant uncertainty due to coarse resolutions. Direct dynamical downscaling (DDD) of regional climate at kilometer-scale resolutions provides valuable insight into extreme precipitation changes, but its computational expense is formidable. Here we document the effectiveness of machine learning to enable smart dynamical downscaling (SDD), which selects a small subset of GCM data to conduct downscaling. Trained with data for three subtropical/tropical regions, convolutional neural networks (CNNs) retained 92% to 98% of extreme precipitation events (rain intensity higher than the 99th percentile) while filtering out 88% to 95% of circulation data. When applied to reanalysis data sets differing from training data, the CNNs' skill in retaining extremes decreases modestly in subtropical regions but sharply in the deep tropics. Nonetheless, one of the CNNs can still retain 62% of all extreme events in the deep tropical region in the worst case.

## Plain Language Summary

Climate scientists use supercomputers to simulate the climate and predict how it may change under global warming. Extreme precipitation, which can disrupt society by causing disasters like floods and landslides, is of great interest in climate studies. However, replicating severe rainstorms on a supercomputer, especially the storms in tropical and subtropical areas, is not easy. This is because those rainstorms often contain fine-scale details that cannot be represented confidently without extensive computational resources. If we use computationally affordable computer models to simulate those rainstorms, we obtain results with substantial uncertainties. If we use computationally expensive ones, we cannot simulate many scenarios and cannot be confident about the results. The power of machine learning in pattern recognition is here used to help modelers use their computational resources more efficiently. Instead of simulating all kinds of weather events, including unimportant ones, at high resolutions, we use machine learning algorithms to search coarse resolution climate data for those large-scale weather patterns that are more likely to cause severe rainstorms. Then modelers can make more efficient use of supercomputing resources by simulating severe weather events only and advance our understanding of them.

## 1 Introduction

Extreme precipitation events often disrupt society by causing disasters such as floods and landslides. Thus, predicting the response of precipitation extremes to global warming is crucial for our adaptation to climate change. Climate models agree well with each other on the potential response of extreme extratropical precipitation to global warming, but their results for subtropical and tropical extremes diverge (O'Gorman & Schneider, 2009). Predicting such changes is not straightforward, because the performance of numerical simulation of extreme precipitation is sensitive to model resolution (Li et al., 2018; Van Der Wiel et al., 2016), and grid spacings of current-generation climate models are still at coarse $\sim 1^\circ$ resolutions. Previous studies have demonstrated that to accurately predict future changes in extreme precipitation events, especially those associated with severe convection, it is necessary to resolve local storm dynamics with kilometer-scale grid spacings (Kendon et al., 2014, 2017). Such a high resolution is necessary not only because of the small spatial scale of convective cells, but also because the essential roles played by the interaction between convection and large-scale dynamics, air-sea coupling, and topographic forcing in determining the intensity of extreme events (Nie et al., 2016; Kendon et al., 2017; Rainaud et al., 2017).

Modelers have been attempting to refine global climate models' resolution, but the current highest resolution is only $\sim 25\,\mathrm{km}$ (Haarsma et al., 2016). A direct dynamical downscaling (DDD) approach has been adopted in the regional climate simulations at convection-permitting resolutions. Valuable findings have been obtained due to improved representation of fine-scale processes, but DDD at the convection-permitting resolution has a very high demand on computational resources (Prein et al., 2015).

Is there a way to avoid the expensive computational cost of long-term DDD but still allow a convection-permitting resolution? This question is the core problem we want to address in this study. When our concern is not the mean climate but instead a special kind of weather (e.g., extreme precipitation), we can save a tremendous amount of computational resources if we do not have to perform the DDD for every day of an extended period. In this study, we harness machine learning's power to fulfill the goal of selecting a small subset of GCM data for the dynamic downscaling of extreme precipitation events. We call this strategy smart dynamical downscaling (SDD).

Machine learning has been increasingly used in geoscience in recent years. In the atmospheric science community, it has applied to real-time nowcasting (Han et al., 2017; McGovern et al., 2017), physical parameterization (Brenowitz & Bretherton, 2019; Gagne et al., 2020), and weather forecasting (Weyn et al., 2019; Chattopadhyay et al., 2020). Previous authors have documented machine learning's potential to identify synoptic-scale patterns associated with extreme rainfall in the extratropics (Agel et al., 2018; Conticello et al., 2018; Knighton et al., 2019). The current study differs from previous ones in that we intentionally chose subtropical and tropical regions for potential applications on convective rainfall, which might be more challenging to capture based on large-scale circulation. Also, because the purpose of this study is to evaluate the potential of SDD, we used machine learning for the classification problem of circulation patterns, instead of attempting to predict the exact precipitation amount like other statistical downscaling studies (e.g., Sachindra et al., 2018).

We evaluated three machine learning models, a dual support vector machine (SVM) model, an 8-layer deep convolutional neural network (CNN), and a sophisticated 58-layer deep CNN, in classifying circulation patterns responsible for 6-hourly precipitation extremes. The performance of these machine learning models with increasing complexity is documented, and we found the deep CNN with a structure of intermediate-level complexity appears to suffice for SDD.

## 2 Data and Methods

### 2.1 Reanalysis and Satellite Data

We train machine learning models with reanalysis data of circulation and satellite data of 6-hourly precipitation. Our study focused on the areas surrounding three Asian cities, Hong Kong (HK), Manila (MN), and Singapore (SG), where extreme rainfall is often related to intense convection, to contrast applicability of the methodology developed here for subtropical and tropical climate.

The precipitation data we used are the final precipitation, Level 3 data of the Integrated Multi-satellitE Retrievals for Global Precipitation Measurement (GPM IMERG; Huffman et al., 2019). This data set have 0.1° spatial resolution and 30 min temporal resolution originally. We used the data set between the period of June 2000 to May 2019. Because the reanalysis data have a 6-hour temporal resolution, we average the original data in time to get the mean precipitation rate in 6-hour intervals. We also used area averaging of the precipitation data to coarse-grain the data onto a $0.5° \times 0.5°$ grid to ignore sporadic events that affect only a small area.

Multiple reanalysis data sets were used in the training and evaluation of machine learning models. For the SVMs' training, we use the NCEP/NCAR (National Centers for Environmental Prediction/National Center for Atmospheric Research) Global Reanalysis Products (Kalnay et al., 1996) to represent the state of the atmospheric circulation. This data set has $2.5° \times 2.5°$ horizontal resolution. We use data on eight pressure levels between 1000 hPa to 300 hPa. The variables we chose to depict the large-scale circulation include 7 three dimensional variables: geopotential height, relative humidity, temperature, $u$- and $v$-components of horizontal wind, vertical (pressure) velocity, and vorticity, in addition to 3 single-level variables — surface pressure, tropopause pressure, and precipitable water. The temporal resolution of the reanalysis data is 6 hours. The circulation variables were normalized with the mean and standard deviation at each level. The precipitation data from reanalysis were not used because they represent precipitation from large-scale circulation and significantly biased. Supplementary Figure S1 shows that precipitation data from the reanalysis data suggest inaccurate timing and intensities compared with GPM observation.

For the training of deep neural networks (R**a**Net and R**x**Net described below), we used the NCEP final (FNL) operational analysis data on 1°×1° grids (NCEP/NWS/NOAA, 2000). The NCEP/NCAR reanalysis data were not used for the training of CNNs, because its coarse resolution hampers the use of multiple convolutional layers. To reduce the computational cost in training the CNNs, we only used five variables (geopotential height, temperature, relative humidity, and $u$- and $v$-components of wind) on six pressure levels (300, 500, 700, 850, 925 and 1000 hPa).

Finally, to evaluate the sensitivity of the trained neural networks to potential model biases in climate simulations, we evaluated the performance of the FNL-trained neural networks with another two reanalysis data sets, ERA5 (Copernicus Climate Change Service, 2017) and JRA-55 (Japan Meteorological Agency, 2013). Those data were spline interpolated onto $1° \times 1°$ grid and used by the FNL-trained CNNs to make predictions.

### 2.2 Support Vector Machine

An SVM is a machine learning model for classification problems (Cristianini et al., 2000). At its core, an SVM finds a hyperplane in the feature space of data and separate points in the feature space into different groups. The hyperplane in feature space is defined as the set of points $\mathbf{x}$ satisfying

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{1}$$

The vector $\mathbf{w}$ and scalar $b$ for the best hyperplane are determined by an optimization procedure that maximizes the margin between two classes in the feature space. For a linearly separable problem, $\mathbf{w}$ and $b$ are entirely determined by those sample points that are closest to the best hyperplane. Those sample points are called support vectors. When data are not linearly separable, one can use a soft margin technique to allow a small number of misclassified instances.

Furthermore, in nonlinear classification problems, it is common to use a kernel function to replace dot product for operating the optimization algorithm in a transformed feature space implicitly. In our application, we used the Gaussian radial basis function,

$$G(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}\right) \tag{2}$$

where $\sqrt{2}\sigma$ is called kernel scale. Besides $\sigma$, the other hyperparameter for training an SVM is the box constraint which appears in the soft margin formula and determines the tolerance level of misclassification.

An SVM takes the NCEP/NCAR reanalysis data in the $15° \times 15°$ square region centered at one of the three cities as input. Each time slice is categorized as producing "significant rain" or "no significant rain" (with the 30th percentile of rain rate as the threshold), "light rain" or "heavy rain" (with the 60th, 70th, or 80th percentile as the threshold, see Section 3.1), based on next-6-hour precipitation in the $0.5° \times 0.5°$ cell centered at the same city. The SVMs were trained to classify the large-scale circulation patterns accordingly.

MATLAB R2019b was used to train SVMs. The SVMs were trained using Bayesian optimization to find out the best hyperparameters. Their performance was evaluated with 10-fold cross-validation, in which the input data set was partitioned into ten subsets. Each subset was sequentially used as the validation set, while the other nine were used for training. Performance metrics are based on ten-time averages.

### 2.3 Convolutional Neural Network

In its essence, a neural network transforms the signal from one layer of neurons to the next through a linear transformation and the use of a nonlinear activation function,

$$\mathbf{z}^{[k]} = \mathbf{W}^{[k]}\mathbf{a}^{[k-1]} + \mathbf{b}^{[k]}, \qquad \mathbf{a}^{[k]} = g^{[k]}(\mathbf{z}^{[k]}) \tag{3}$$

where $\mathbf{a}^{[k]}$ is the activation of Layer $k$, $\mathbf{W}^{[k]}$ is a weight matrix, and $\mathbf{b}^{[k]}$ is a bias vector. $g^{[k]}$ is a non-linear activation function. For Layer 0, the activation $\mathbf{a}^{[0]}$ is the vector of input data $\mathbf{x}$. A fully connected layer in a deep neural network connects every neuron in the previous layer to every neuron in the current layer. A convolutional layer, by contrast, has multiple filters, which are used to convolve a sub-block of the activation data from the previous layer and connect that subset of neurons to a neuron in the current layer.

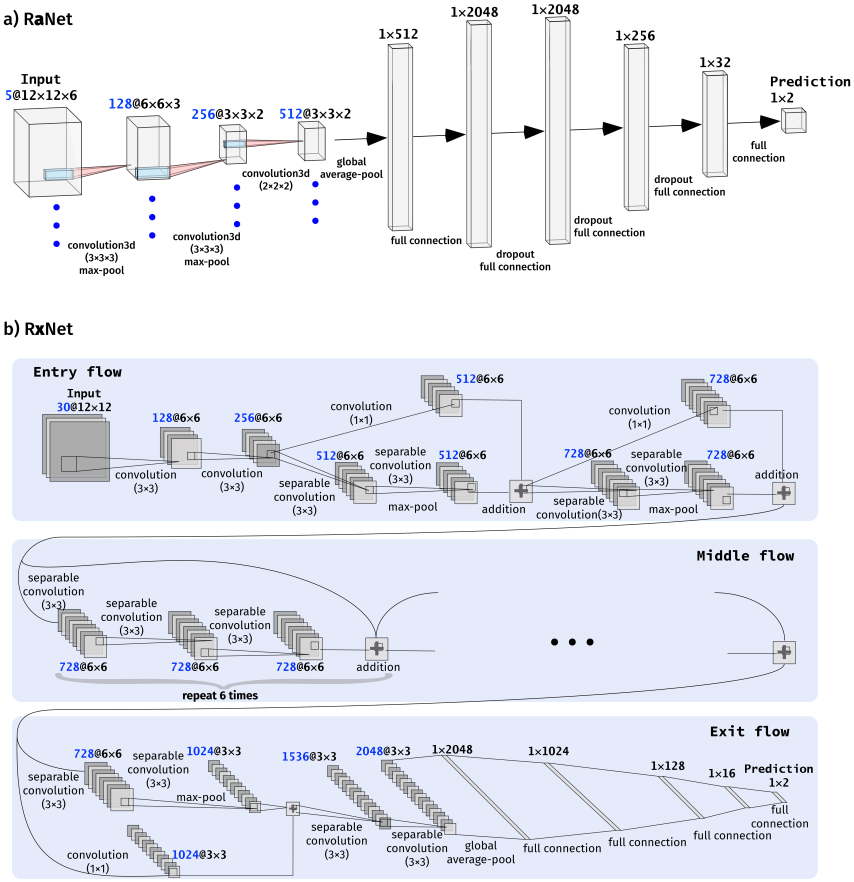

**Figure 1.** Structure of a) R**a**Net and b) R**x**Net. R**a**Net uses three-dimensional filters in the convolutional layers and leaky ReLU activation for all layers; the first two convolutional layers are followed by batch normalization layers which are not shown. R**x**Net uses two-dimensional filters in its regular convolution and channel-wise separable convolution operations, and used the ReLU activation function for all layers; all convolutional layers are followed by batch normalization layers which are not shown. Blue-font values before @ indicate the number of channels of each layer. The expression after @ indicates the size of activation arrays of a channel. The expression in brackets indicates the size of filters used by convolutional layers.

Two CNN structures are tested in this study (Fig. 1). They are motivated by the AlexNet (Krizhevsky et al., 2012) and Xception (Chollet, 2017) models, respectively, which showed excellent performance in computer vision competitions. This first CNN used in this study is named as R**a**Net (motivated by AlexNet, Fig. 1a). It has 3 convolution layers and 5 fully connected layers. Differing from the original AlexNet, R**a**Net uses three-dimensional filters in its convolutional layers; thus, its input layer has five channels (variables). By contrast, R**x**Net (motivated by Xception, Fig. 1b) treats the data on each pressure level as one individual variable; thus its input layer has 30 channels (5 variables × 6 levels). Such a design of **Rx**Net is used for closely following the original Xception model, which was applied to two-dimensional images. **Rx**Net is 58-layer deep and includes multiple residual connections.

When training the CNNs, we included the precipitation data for about 40 to 50 additional $0.5^\circ \times 0.5^\circ$ grid cells surrounding each of those three cities (and the accompanying circulation data) to obtain more samples, which helps prevent overfitting. The extent of the surrounding areas was determined by applying the trained SVMs to new nearby grid cells and evaluating the performance of the SVMs. Relatively high performance suggests the weather patterns governing precipitation at the new locations are similar to those at the original training location. Thus, it is appropriate to include the new grid cells' data to increase the total sample size. The exact extent of the selected HK, MN, and SG regions is shown in Supplementary Figure S2, with the selection threshold provided in the caption of Fig. S2.

6-hourly precipitation data of each $0.5^\circ \times 0.5^\circ$ cell within a selected region are used to categorize the corresponding time and location as producing "extreme rain" or "non-extreme rain" (with the 90th percentile of rain rate as the threshold). The input data for the neural networks are the FNL data spline-interpolated onto $12^\circ \times 12^\circ$ square regions, which are centered at each of the $0.5^\circ \times 0.5^\circ$ rain data cells and have $1^\circ \times 1^\circ$ resolution (Supplementary Figure S3). Input data for **Ra**Net are scaled perturbations. We define base-state profiles of geopotential height and temperature as their climatological means and the base-state profiles for $u$, $v$, and relative humidity as zero. The deviations of variables from base states are defined as perturbations and then scaled by their root-mean-square amplitudes. Because **Rx**Net treats the data on each pressure level as separate variables, input data of each channel for **Rx**Net are rescaled to be in the range of $-1$ to 1 using minimum and maximum values. When the FNL-trained CNNs are applied to ERA5 and JRA-55 data sets, leading-order model "biases" in these two data sets were removed by adjusting their mean and root-mean-square perturbation amplitude at each pressure level to be the same as FNL data.

For these two CNNs, 60% of the FNL data were used to train the models, and 20% used for validation, which helped decide if early stopping was needed during training. The other 20% data were held out as a test data set for evaluating trained models' performance. 70-15-15 partitioning of the train-validation-test data sets was also evaluated and did not cause a significant difference in results.

The two CNNs were trained to partition data into the categories of "extreme" and "non-extreme" rain, by iterating to minimize the weighted cross-entropy loss function,

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} w_j T_{ij} \log(Y_{ij})\,. \tag{4}$$

$T$ is training targets, $Y$ is predicted probability, $N$ is the number of instances, $K$ is the number of classes, and $w$ is the weighting factor. Instead of an unweighted loss function, this weighted loss function was used because the number of non-extreme events is much larger than that of extremes. The weighting factor $w$ is set to 0.95 for extreme events and 0.05 for non-extreme events. These weighting factors are determined by the approximate ratio of the number of events in the two categories. Therefore, predicting an extreme event wrong causes a much larger increase in the loss function than doing the same to a non-extreme event.

**Ra**Net and **Rx**Net were optimized using the Adam (adaptive moment estimation) optimizer in MATLAB through 30 epochs of iteration and with a learning rate of $1 \times 10^{-4}$. Training them with more iteration cycles can increase their accuracy and precision, but leads to deterioration in the recall, which suggests overfitting and is not favorable for retaining extreme events. Because of the high computational expense in training CNNs, we did not apply the Bayesian optimization here. Instead, the learning rate, CNN structures, and the number of training epochs were determined empirically through several rounds of experiments.

### 2.4 Performance Metrics

In the training of SVMs and CNNs, algorithms try to achieve the highest classification accuracy. However, because extreme events are only a small fraction of the data, accuracy of trained models is always intuitively high. Thus, in our discussion, we report the performance of trained models primarily with precision and recall. Precision quantifies the skill of a trained model in filtering out irrelevant circulation patterns, whereas recall quantifies how well the relevant patterns are retained. Specifically,

$$P_y^{\mathrm{M}} = \frac{|\{r > r_y\} \cap \{r' > r_y\}|}{|\{r' > r_y\}|}\,, \tag{5}$$

$$R_y^{\mathrm{M}} = \frac{|\{r > r_y\} \cap \{r' > r_y\}|}{|\{r > r_y\}|}\,, \tag{6}$$

where $P_y^{\mathrm{M}}$ and $R_y^{\mathrm{M}}$ are precision and recall of the model M when cases with precipitation rates greater than the $y$-th percentile, $r_y$, are labeled as positive. $r_y$ may *differ* from the actual threshold used in labelling data when training M. $\{r > r_y\}$ represent the set of instances for which real rain rate ($r$) exceeds $r_y$, and $\{r' > r_y\}$ is the set of instances for which the model M predicts rain rate ($r'$) exceeding $r_y$. $r'$ was not computed by the machine learning models explicitly, but rather the condition, $r' > r_y$, was judged by the model M.

## 3 Results

### 3.1 Dual SVM Model

We trained a pair of SVMs to select instances for extreme events. The first SVM (SVM1) tells whether the circulation at a time can produce “significant” rainfall or not, with the 30th percentile of rain rate as the threshold. The subset of circulation data, which SVM1 predicts to produce significant rain, is then adopted by the second SVM (SVM2), which uses a higher percentile (60th, 70th, or 80th) as its threshold for “extremes”. We found that this dual-SVM strategy can yield higher precision and recall than using a single SVM to directly predict “extremes”.

Figure 2 shows the performance of the Dual SVM model trained with the data for the three cities, HK, MN, and SG. The precision of SVM1 for its training criteria, $P_{30}^{\mathrm{SVM1}}$, is around 0.7, and the recall of SVM1 for its training criteria, $R_{30}^{\mathrm{SVM1}}$, is between 0.48 and 0.59. These recall values are not very high. However, if we target to retrieve precipitation event with rain rate higher than the 90th and 99th percentiles, we can find that the corresponding recall, $R_{90}^{\mathrm{SVM1}}$ and $R_{99}^{\mathrm{SVM1}}$, is between 0.82 and 0.92 for HK and MN, and between 0.69 and 0.79 for SG. It should be noted that because we did not include rain rate lower than $0.05\,\mathrm{mm\,h^{-1}}$ in calculating the percentiles, SVM1 eliminates much more than 30% circulation data from all time slices. Precipitation rates in HK, MN, and SG exceed the corresponding 30th percentiles only in 14.5%, 28.9%, and 29.4%, respectively, of time slices of the 19 years (not shown).

Figure 2 also shows the performance of SVM2 for training criteria, and real extreme events defined by the 90th and 99th percentiles. For SG, we could not obtain a converged solution when the training criterion was set as the 80th percentile. Therefore, it is likely that those circulation patterns, responsible for the extreme events defined with the 80th percentile, are inseparable from others by an SVM.

The precision of SVM2 for the 90th and 99th percentiles (red and yellow bars in Fig. 2a-c) increases as the training criteria increase to become close to the evaluation criteria. However, those values are relatively low because SVM2 was trained with different criteria (e.g., the 70th percentile). The recall of SVM2 decreases as the training criteria increases. A higher training threshold means that we can filter out more “irrelevant” instances. However, it also increases our chance of losing actual extreme events due to misclassification. Based on Fig. 2, the SVM2 trained with the 70th percentile of rain rate appears to be the most balanced model for applications. If we target to retrieve extreme events defined by the 99th percentile in the selection, the SVM1 and the SVM2 trained

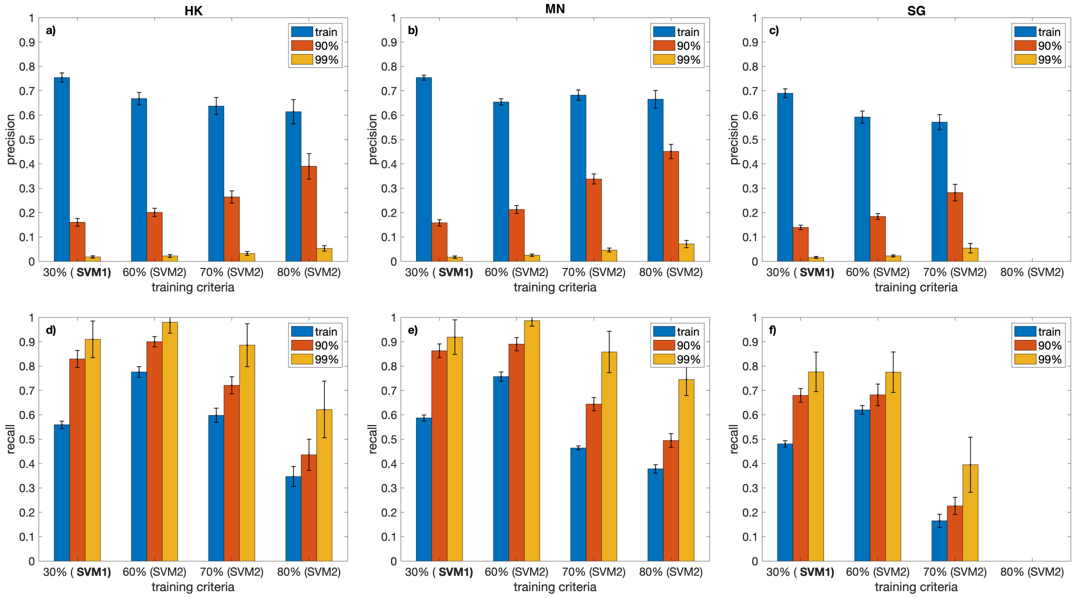

**Figure 2.** Precision (a–c) and recall (d–f) of the trained SVMs. a) and d) are the SVMs for HK, b) and e) for MN, c) and f) for SG. The SVMs were trained for the thresholds indicated below the horizontal axis, but their performance is evaluated against the training criteria and the 90th and 99th percentiles of rain rates.

with the 70th percentile can yield combined recall (product of the recall of SVM1 and SVM2) of $R_{99}^{\mathrm{SVM1}} R_{99}^{\mathrm{SVM2}} = 0.81,\ 0.79,$ and 0.31, for HK, MN, and SG, respectively.

The Dual SVM model's unsatisfactory performance for SG data suggests we cannot obtain a very reliable subset of data if we want to study extreme rainfall in the deep tropics with SVMs. Moreover, because we can only use the 70th percentile of rain rate in the training of SVM2, we still need to "waste" a substantial fraction of our computation to ensure the SVMs keep the most extreme events. Can we overcome these difficulties with deep neural networks?

### 3.2 Convolutional Neural Networks

The performance of R**a**Net and R**x**Net is shown in Table 1. For the test set of FNL data, the precision of the two CNNs, $P_{90}^{\mathrm{RaNet}}$ and $P_{90}^{\mathrm{RxNet}}$, is between 0.23 and 0.33, which is not very impressive, but their recall, $R_{90}^{\mathrm{RaNet}}$ and $R_{90}^{\mathrm{RxNet}}$, is high, between 0.75 and 0.92. When evaluated for the 99th percentile, the recall of the CNNs, $R_{99}^{\mathrm{RaNet}}$ and $R_{99}^{\mathrm{RxNet}}$, reaches 0.93 to 0.98. Those high values contrast with the much lower recall values of the dual SVM models, especially for the SG region. Therefore, the deep neural networks R**a**Net and R**x**Net are indeed more powerful in recognizing large-scale patterns responsible for extreme events. The relatively low precision values partially result from the weighted cross-entropy loss, which ensures the high values of recall. We trained R**a**Net with unweighted cross-entropy loss. It exhibits a precision of 0.38–0.49, and recall (for the 90th percentile) drops to 0.58–0.67, leading to the misclassification of many extreme events.

Different climate models potentially have their intrinsic biases — can the CNNs trained with FNL data perform well when applied to climate simulation data? To evaluate the tolerance of R**a**Net and R**x**Net to potential GCM biases, we apply them to another two reanalysis datasets, ERA5 and JRA-55, to compute the performance metrics of the FNL-trained CNNs (while still using the GPM precipitation to label instances). Different reanalysis data sets are known to represent some parts of the general circulation differently (Kossin, 2015). Although we have adjusted the mean and ampli-

**Table 1.** Performance metrics of R**a**Net and R**x**Net. Three data sets, FNL, ERA5, and JRA-55, were used to evaluate the models. For FNL, only the test dataset (20% of all) was used to evaluate the performance of trained models, whereas, for ERA5 and JRA-55, entire data sets were used. The rows of "precision" and "recall" are computed for the training threshold, the 90th percentile values. The rows of "recall (99%)" is the recall when the trained models are evaluated for the 99th percentile values. "retention" refers to the fraction of data retained (as relevant to extreme events) by the trained models.

| | | HK Region | | MN Region | | SG Region | |
|---|---|---|---|---|---|---|---|
| | | R**a**Net | R**x**Net | R**a**Net | R**x**Net | R**a**Net | R**x**Net |
| accuracy | FNL | 0.936 | 0.961 | 0.897 | 0.933 | 0.900 | 0.920 |
| | ERA5 | 0.948 | 0.964 | 0.904 | 0.938 | 0.905 | 0.931 |
| | JRA-55 | 0.950 | 0.957 | 0.907 | 0.931 | 0.883 | 0.926 |
| precision | FNL | 0.238 | 0.331 | 0.229 | 0.307 | 0.230 | 0.274 |
| | ERA5 | 0.257 | 0.326 | 0.224 | 0.292 | 0.201 | 0.238 |
| | JRA-55 | 0.259 | 0.276 | 0.217 | 0.241 | 0.148 | 0.180 |
| recall | FNL | 0.921 | 0.858 | 0.832 | 0.748 | 0.770 | 0.749 |
| | ERA5 | 0.777 | 0.663 | 0.725 | 0.553 | 0.557 | 0.428 |
| | JRA-55 | 0.738 | 0.645 | 0.650 | 0.462 | 0.475 | 0.299 |
| recall (99%) | FNL | 0.985 | 0.983 | 0.955 | 0.935 | 0.927 | 0.936 |
| | ERA5 | 0.927 | 0.843 | 0.904 | 0.800 | 0.742 | 0.643 |
| | JRA-55 | 0.901 | 0.798 | 0.864 | 0.695 | 0.622 | 0.465 |
| retention | FNL | 0.082 | 0.055 | 0.126 | 0.084 | 0.120 | 0.098 |
| | ERA5 | 0.064 | 0.043 | 0.112 | 0.065 | 0.099 | 0.064 |
| | JRA-55 | 0.060 | 0.049 | 0.104 | 0.066 | 0.115 | 0.059 |

tude of ERA5 and JRA-55 data (Section 2.1) to correct leading order biases, significant changes in the performance of trained CNNs can still be found when applied to the ERA5 and JRA-55 data.

In Table 1, application of the FNL-trained CNNs to ERA5 data does not result in a large decrease in the accuracy and precision, but leads to a sharp drop in the recall, especially for the SG region. The recall corresponding to the training criterion (90th percentile) for the SG region is around 0.76 for the FNL test data set but drops to 0.56 and 0.43 for R**a**Net and R**x**Net, respectively, for the ERA5 data. When considering the 99th percentile, $R_{99}^{\mathrm{RaNet}}$ and $R_{99}^{\mathrm{RxNet}}$ are higher than 0.80 for the HK and MN regions with the ERA5 data, but are only 0.74 and 0.64, respectively, for the SG region.

The JRA-55 data set appears to differ from the FNL data even more than the ERA5 data. Recall values of R**a**Net and R**x**Net, when applied to the JRA-55 data, become even lower than those for ERA5. For the HK region, the recall $R_{99}^{\mathrm{RaNet}}$ and $R_{99}^{\mathrm{RxNet}}$ are 0.90 and 0.80, respectively, with the JRA-55 data, which are still satisfactory. However for the SG region, $R_{99}^{\mathrm{RaNet}}$ and $R_{99}^{\mathrm{RxNet}}$ are only 0.62 and 0.47, respectively, with the JRA-55 data. These results suggest that if the CNNs are trained with one circulation data set and applied to the deep tropics in climate simulations, they may not capture all the circulation patterns in climate models that can generate extreme events when dynamically downscaled.

Overall, R**x**Net exhibits higher accuracy and precision than R**a**Net for all three regions. However, R**a**Net exhibits higher recall values and appears to be more resilient to potential model biases. For example, $R_{99}^{\mathrm{RaNet}}$ is consistently higher than $R_{99}^{\mathrm{RxNet}}$ by more than 0.10 in all three regions. However, the relatively higher recall comes with a price in computational cost, that is, less "irrelevant" data can be filtered out if the recall needs to be high. For example, when using ERA5 data for the MN region, R**a**Net retains twice as much data as R**x**Net.

## 4 Discussion and Summary

In this study, we demonstrated that the smart dynamical downscaling (SDD) of extreme rainfall is viable through deep neural networks, though the reliability of this method depends on climate regimes. For the subtropical regions (HK and MN), this methodology appears to be promising. The trained CNNs performed well, even when different reanalysis data sets were used to evaluate their performance. In the HK region, for example, 92% to 96% of circulation data can be filtered out as irrelevant patterns for extreme events. However, for the deep tropics (SG region), the CNNs' skill in retaining extremes significantly deteriorates when applied to different reanalysis data sets. For instance, R**x**Net has a recall of 0.94 for the 99th-percentile extreme events for FNL data but drops to 0.47 for JRA-55 data.

From simple SVMs to sophisticated CNNs, the model performance is always worse for the SG region than the other two regions. We speculate this is because the link between large-scale circulation and local precipitation in the deep tropics is just not as strong as those in subtropics. $k$-medoid clustering analysis (Supplementary Figure S4-S6) suggests that extreme precipitation events in the HK region are typically associated with warm-sector convection, frontal rainfall, and tropical cyclones (Fig. S4), of which the first type comprises the majority (Wu et al., 2020). Those weather patterns have distinct large-scale features. In contrast, extreme precipitation appears to be connected with squall lines and cold pools for the SG region (Porson et al., 2019), which exhibit significant variability at smaller grid scales (Fig. S6). It is probably not surprising that fitting small-scale features is more complicated than fitting large-scale ones.

Therefore, the SDD of extreme precipitation in the deep tropics appears to be challenging. One could use a threshold that is even lower than the 90th percentile to train CNNs to increase the recall for the 99th-percentile extreme events. However, such a strategy may not always be desirable because it increases the recall by sacrificing precision, thereby increasing the computational cost of downscaling simulations. It is also possible to include multiple reanalysis data when training CNNs to alleviate the problem of low tolerance to potential model biases. Lastly, using a model structure with an intermediate level of sophistication, like the R**a**Net here, may also be beneficial.

In subtropical regions, the potential of advanced deep neural networks, such as R**x**Net here, can be fully exploited to reduce computational expense while confidently retaining most of the circulation patterns causing extreme rainfall. In our study, the recall $R_{99}^{\mathrm{RxNet}} \geq 0.80$ for the HK region with all circulation data sets. The next step for our research is to apply deep neural networks to SDD of climate simulations and explore the response of extreme rainfall to global warming.

## Acknowledgments

The author thanks two anonymous reviewers for valuable comments and acknowledges the support of the Research Grants Council of Hong Kong SAR, China (Project No. AoE/E-603/18 and HKUST 26305720). FNL and JRA-55 data were obtained from the Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory (https://rda.ucar.edu/). The ERA5 data set was provided through Copernicus Climate Change Service Climate Data Store (https://cds.climate.copernicus.eu/) The GPM IMERG precipitation data were provided by the Goddard Earth Sciences Data and Information Services Center (GES DISC) (https://doi.org/10.5067/gpm/imerg/3b-hh/06).

## References

Agel, L., Barlow, M., Feldstein, S. B., & Gutowski, W. J. (2018). Identification of large-scale meteorological patterns associated with extreme precipitation in the us northeast. *Clim. Dynam.*, *50*(5-6), 1819–1839.

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *J. Adv. Model. Earth Syst.*, *11*(8), 2728–2744.

Chattopadhyay, A., Nabizadeh, E., & Hassanzadeh, P. (2020). Analog forecasting of extreme-causing weather patterns using deep learning. *J. Adv. Model. Earth Syst.*, *12*(2), e2019MS001958.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).

Conticello, F., Cioffi, F., Merz, B., & Lall, U. (2018). An event synchronization method to link heavy rainfall events and large-scale atmospheric circulation features. *Int. J. Climatol.*, *38*(3), 1421–1437.

Copernicus Climate Change Service. (2017). *ERA5: Fifth generation of ecmwf atmospheric reanalyses of the global climate.* Copernicus Climate Change Service Climate Data Store (CDS). Retrieved from `https://cds.climate.copernicus.eu`

Cristianini, N., Shawe-Taylor, J., et al. (2000). *An introduction to support vector machines and other kernel-based learning methods.* Cambridge University Press.

Gagne, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine learning for stochastic parameterization: Generative adversarial networks in the lorenz'96 model. *J. Adv. Model. Earth Syst.*, *12*(3), e2019MS001896.

Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., … others (2016). High resolution model intercomparison project (highresmip v1. 0) for cmip6. *Geosci. Model Dev.*, *9*(11), 4185–4208.

Han, L., Sun, J., Zhang, W., Xiu, Y., Feng, H., & Lin, Y. (2017). A machine learning nowcasting method based on real-time reanalysis data. *J. Geophys. Res.*, *122*(7), 4038–4051.

Huffman, G., Stocker, E., Bolvin, D., Nelkin, E., & Jackson, T. (2019). *GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree × 0.1 degree V06, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC).* doi: 10.5067/GPM/IMERG/3B-HH/05

Japan Meteorological Agency. (2013). *JRA-55: Japanese 55-year reanalysis, daily 3-hourly and 6-hourly data.* Boulder CO: Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. Retrieved from `https://doi.org/10.5065/D6HH6H41`

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., … others (1996). The ncep/ncar 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, *77*(3), 437–472.

Kendon, E. J., Ban, N., Roberts, N. M., Fowler, H. J., Roberts, M. J., Chan, S. C., … Wilkinson, J. M. (2017). Do convection-permitting regional climate models improve projections of future precipitation change? *Bull. Amer. Meteor. Soc.*, *98*(1), 79–93.

Kendon, E. J., Roberts, N. M., Fowler, H. J., Roberts, M. J., Chan, S. C., & Senior, C. A. (2014). Heavier summer downpours with climate change revealed by weather forecast resolution model. *Nature Clim. Change*, *4*(7), 570–576.

Knighton, J., Pleiss, G., Carter, E., Lyon, S., Walter, M. T., & Steinschneider, S. (2019). Potential predictability of regional precipitation and discharge extremes using synoptic-scale climate information via machine learning: An evaluation for the eastern continental united states. *J. Hydrometeor.*, *20*(5), 883–900.

Kossin, J. P. (2015). Validating atmospheric reanalysis data using tropical cyclones as thermometers. *Bull. Amer. Meteor. Soc.*, *96*(7), 1089–1096.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Li, J., Chen, H., Rong, X., Su, J., Xin, Y., Furtado, K., … Li, N. (2018). How well can a climate model simulate an extreme precipitation event: A case study using the transpose-amip experiment. *J. Climate*, *31*(16), 6543–6556.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., … Williams, J. K. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, *98*(10), 2073–2090.

NCEP/NWS/NOAA. (2000). *NCEP FNL operational model global tropospheric analyses, continuing from july 1999.* Boulder CO: Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. Retrieved from `https://doi.org/10.5065/D6M043C6`

Nie, J., Shaevitz, D. A., & Sobel, A. H. (2016). Forcings and feedbacks on convection in the 2010 pakistan flood: Modeling extreme precipitation with interactive large-scale ascent. *J. Adv. Model. Earth Syst.*, *8*(3), 1055-1072. doi: 10.1002/2016MS000663

O'Gorman, P. A., & Schneider, T. (2009). The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. *Proc. Natl. Acad. Sci. USA*, *106*(35), 14773–14777. Retrieved from `https://www.pnas.org/content/106/35/14773` doi: 10.1073/pnas.0907610106

Porson, A. N., Hagelin, S., Boyd, D. F., Roberts, N. M., North, R., Webster, S., & Lo, J. C.-F. (2019). Extreme rainfall sensitivity in convective-scale ensemble modelling over singapore. *Quart. J. Roy. Meteor. Soc.*, *145*(724), 3004–3022.

Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., … others (2015). A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Rev. Geophys.*, *53*(2), 323–361.

Rainaud, R., Brossier, C. L., Ducrocq, V., & Giordani, H. (2017). High-resolution air–sea coupling impact on two heavy precipitation events in the western mediterranean. *Quart. J. Roy. Meteor. Soc.*, *143*(707), 2448–2462.

Sachindra, D., Ahmed, K., Rashid, M. M., Shahid, S., & Perera, B. (2018). Statistical downscaling of precipitation using machine learning techniques. *Atmos. Res.*, *212*, 240–258.

Van Der Wiel, K., Kapnick, S. B., Vecchi, G. A., Cooke, W. F., Delworth, T. L., Jia, L., … Zeng, F. (2016). The resolution dependence of contiguous us precipitation extremes in response to co2 forcing. *J. Climate*, *29*(22), 7991–8012.

Weyn, J. A., Durran, D. R., & Caruana, R. (2019). Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *J. Adv. Model. Earth Syst.*, *11*(8), 2680–2693.

Wu, N., Ding, X., Wen, Z., Chen, G., Meng, Z., Lin, L., & Min, J. (2020). Contrasting frontal and warm-sector heavy rainfalls over south china during the early-summer rainy season. *Atmospheric Research*, *235*, 104693.