A Multi-step-ahead Markov Conditional Forward Model with Cube Perturbations for Extreme Weather Forecasting

Chia-Yuan Chang,* Cheng-Wei Lu,* Chuan-Ju Wang
Research Center for Information Technology Innovation, Academia Sinica, Taiwan
{cychang, cwlu, cjwang}@citi.sinica.edu.tw

Abstract

Predicting extreme weather events such as tropical and extratropical cyclones is of vital scientific and societal importance. Of late, machine learning methods have found their way to weather analysis and prediction, but mostly, these methods use machine learning merely as a complement to traditional numerical weather prediction models. Although some pure machine learning and data-driven approaches for weather prediction have been developed, they mainly formulate the problem similar to pattern recognition or follow the train of thought of traditional time-series models for extreme weather event forecasting; for the former, this usually yields only single-step ahead prediction, and for the latter, this lacks the flexibility to account for observed weather features as such methods concern only the patterns of the extreme weather occurrences. In this paper, we depart from the typical practice of pattern recognition and time-series approaches and focus on employing machine learning to estimate the probabilities of extreme weather occurrences in a multi-step-ahead (MSA) fashion given information on both weather features and the realized occurrences of extreme weather. Specifically, we propose a Markov conditional forward (MCF) model that adopts the Markov property between the occurrences of extreme weather for MSA extreme weather forecasting. Moreover, for better long-term prediction, we propose three novel cube perturbation methods to address error accumulation in our model. Experimental results on a real-world extreme weather dataset show the superiority of the proposed MCF model in terms of prediction accuracy for both short-term and long-term forecasting; moreover, the three cube perturbation methods successfully increase the fault tolerance and generalization ability of the MCF model, yielding significant improvements for long-term prediction.

1 Introduction

Extreme weather forecasting is vital for efficient resource management and active warning systems (Rolnick et al. 2019). The further ahead that predictions can be made, the more beneficial in terms of allowing time for adjustment and reduction of damage due to the event. An accurate prediction of extreme weather is therefore crucial for both damage control and optimizing the management of government resources. However, extreme weather occurrences are by nature rare; due to the heavily skewed distribution of such rare events, extreme weather prediction remains one of the most challenging tasks in climate science and meteorology.

Traditionally, numerical weather prediction and generic circulation models are key tools to forecast the evolution of atmospheric states over time (Scher 2018). Both types of models acquire their predictions by solving discretized physical equations of the physics of the atmosphere. In recent years, combinations of machine learning models and physics formulation have been proposed for more accurate predictions (Reichstein et al. 2019; Krasnopolsky and Fox-Rabinovitz 2006; Holden et al. 2015; Scher and Messori 2018). For example, (Holden et al. 2015) learn relations between orbital parameters and climate fields from a climate model, whereas (Scher and Messori 2018) seek to predict the uncertainty of weather forecasts. Although such techniques are valuable for climate science and meteorology, they are focused on either extracting certain information from models or on combining information from different models, for which machine learning is thus regarded as a complement to traditional physical models.

Of late, pure machine learning and data-driven methods have found their way to weather analysis and prediction. Starting from simple regression to complicated neural network models, machine learning techniques have been applied to various weather tasks, including predicting hurricane routes (Kim 2019) and estimating precipitation (Hwang et al. 2019). As weather data are usually presented as an image with multiple weather features, convolutional neural networks (CNNs) are usually used; for example, CNN has been applied to predict the bounding box of extreme weather (Racah et al. 2017) and to emulate the complete physics and dynamics of generic circulation models (Scher 2018).

On the other hand, multi-step-ahead (MSA) prediction, rather than single-step, is more common in applications. For decades, the time-series literature has formulated MSA prediction with various training strategies, including direct and recursive strategies (Al-Qahtani and Crone 2013; Bontempi, Le Borgne, and De Stefani 2017; Chang, Chiang, and Chang 2007; Cheng et al. 2006; Hussein, Chandra, and Sharma 2016; Koesdwiady, El Khatib, and Karray 2018; Sorjamaa et al. 2007; Taieb and Atiya 2015; Taieb et al. 2012). For the direct strategy, separate prediction models are trained to directly predict each h-step ahead, whereas for recursive

*These authors contributed equally to this work.

strategies, the model is trained to forecast one step ahead, and afterwards, h-step-ahead forecasts are made by iterating the predictions h times, leveraging previously predicted values as inputs as needed. However, such studies are made under the prediction scenario of traditional time-series models such as the autoregressive model; that is, they model only dependency among output variable y at different times. For example, for extreme weather prediction, such an approach models relations among occurrences of extreme weather at different time points. As such, these studies lack the flexibility to include richer information from observed weather features such as temperature, pressure, and humidity.

Interestingly, prediction for extreme weather shares similar characteristics with default prediction in finance, as 1) both extreme weather and default are by nature rare, 2) MSA prediction is vital for both prediction tasks, and finally 3) observed covariates (e.g, temperature for extreme weather prediction and firm value for default prediction) are useful for prediction. For example, state-of-the art models for default prediction factorize the problem of MSA prediction into independent conditional forward models (Duan, Sun, and Wang 2012; Duffie, Saita, and Wang 2007), in which a multivariate feature vector is leveraged for prediction. However, these models are statistical models and thus naturally constrained by their functional rigidity. Moreover, two major differences between the problems of weather and default predictions make the former much more complicated and challenging both in terms of problem formulation and model training. First, whereas features for default prediction form a simple multidimensional vector as input, weather covariates are usually represented as multi-channel images with strong locality dependency; thus CNN-based architectures (e.g., naive CNN or CNN-LSTM (Klein, Wolf, and Afek 2015; Qiu et al. 2017; Xingjian et al. 2015)) are common choices for such a prediction problem. Second, whereas default can only happen once in general, extreme weather can happen more than once at the same location, making the conditional forward models for default prediction inapplicable to extreme weather prediction.

Inspired by conditional forward models for default prediction, we address the aforementioned issues regarding functional rigidity, locality dependency, and multiple occurrences using a conditional forward model that adopts the Markov property between occurrences of extreme weather for MSA extreme weather prediction. We term this a *Markov conditional forward* (MCF) model. Specifically, the usage of the Markov property between occurrences of extreme weather results in two conditional forward models between two consecutive prediction horizons, using which we recursively compose the prediction for all future prediction horizons. Such a Markov approach better leverages the observations and future realized occurrences for model training. Moreover, in contrast to the statistical default models in (Duan, Sun, and Wang 2012; Duffie, Saita, and Wang 2007), which impose distribution constraints in their models, the proposed MCF model is free from functional rigidity and thus can be learned by any neural architecture. Here, due to the locality dependency of weather features, we mainly propose using a CNN architecture to learn the conditional forward models.

Although the future realized occurrences of extreme weather are available at training, such information is unavailable at inference and thus we must use the predicted probabilities of occurrences to compute the prediction for each future prediction horizon. Similar techniques are commonly used in multiple-step-ahead time series prediction, such as the moving average (Said and Dickey 1984) and exponential smoothing (Gardner Jr 1985); in the field of neural networks, sequence-to-sequence models also take into account former predictions in the decoding phase (Chiu et al. 2018). However, these methods, which use predicted values to compute the next prediction, all accumulate error (Sorjamaa et al. 2007; Taieb et al. 2012); that is, bad predictions harm the performance of future predictions. We address this by perturbing the training labels to increase the fault tolerance and generalization ability of the MCF model, for which, considering the locality and temporal dependency of weather features, three novel *cube perturbation* methods are proposed.

To evaluate the proposed MCF model along with the three cube perturbation methods, we conduct extensive experiments on the ExtremeWeather dataset (Racah et al. 2017) for predicting two types of extreme weather events: tropical and extratropical cyclones. Experimental results demonstrate that the proposed MCF model achieves significantly better prediction performance than the direct-strategy model. We further show that the proposed three cube perturbation methods effectively prevent error accumulation and thus yield significant improvements when compared to the model without perturbation, especially for long-term prediction. Moreover, we conduct sensitivity analyses on the hyperparameters of the proposed cube perturbation methods, providing general guidance for perturbation and hyperparameter selection. Finally, interesting case studies with graphical presentation are also provided and discussed. In sum, the proposed approach advances the state of the art in extreme weather prediction along three dimensions:

(i) From a mathematical point of view, we propose a MCF model to better address MSA extreme weather prediction.
(ii) We successfully learn the MCF model with the use of neural networks with CNN architectures for MSA extreme weather prediction, yielding prominent results for both short-term and long-term predictions.
(iii) We propose three novel cube perturbation methods to effectively address error accumulation.

Furthermore, in addition to these three contributions, this paper also provides vision and approaches for research and applications that require multi-step-ahead prediction and simultaneously leverage the observed covariates and future realized occurrences for model training. To our best knowledge, this is the first work to formally frame such a problem with a Markov conditional forward model. For instance, with slight modifications, the proposed framework should be suitable for problems such as flood forecasting or patient readmission rate prediction concerning different future periods.

2 Problem Definition

We formulate multiple-step-ahead (MSA) prediction on extreme weather as follows. Suppose there are weather data of a specific rectangular area of size $I \times J$ for a period

of time with discrete time indices $\mathcal{T} = \{1, \cdots, T\}$. Given $\mathbf{x}_t = (x_t, x_{t-1}, \cdots)$ denoting the weather features observed on and before time t, where each $x_\ell \in \mathbb{R}^{d \times I \times J}$ is a tensor composed of all d-dimensional multivariate weather features in the rectangular data area at time ℓ and $\mathbf{y}_t = (y_t, y_{t-1}, \cdots)$ denoting extreme weather occurrences before time t, where each $y_r \in \{0, 1\}^{I \times J}$ contains extreme weather labels within the data area at time r, our goal is to predict extreme weather occurrences of the next H future observations for each location (i, j) in the data area; i.e., $\{y_{t+1}^{(i,j)}, \cdots, y_{t+H}^{(i,j)}\}$ for $i = 1, \cdots, I$ and $j = 1, \cdots, J$, where $y_t^{(i,j)} \in \{0, 1\}$ is the occurrence of extreme weather at given position (i, j) at time t. With the definitions above, our goal is further defined from a probability perspective as

$$\hat{y}_{t+h}^{(i,j)} = P_h\left(y_{t+h}^{(i,j)} = 1 \,\middle|\, \mathbf{x}_t, \mathbf{y}_t\right) \in [0, 1], \tag{1}$$

for $h = 1, 2, \cdots, H$, $i = 1, \cdots, J$, and $j = 1, \cdots J$. In Eq. (1), $\hat{y}_{t+h}^{(i,j)}$ denotes the prediction of the occurrence of the extreme weather event at time $t + h$ for position (i, j) in the area. Note that hereafter $x_t^{(i,j)} \in \mathbb{R}^d$ denotes the d-dimensional weather features (e.g., precipitation and temperature) at t for a specific position (i, j) in the data area.

3 Multi-step-ahead (MSA) Prediction Models

3.1 Direct Model

For MSA prediction, a direct and intuitive strategy is to treat each future prediction time point independently and build the model for each future time point $t + h$ solely based on the observations on and before time t to approximate the following probability:

$$\hat{y}_{t+h}^{(i,j)} = P_h^{\text{Direct}}\left(y_{t+h}^{(i,j)} = 1 \,\middle|\, \mathbf{x}_t\right). \tag{2}$$

For each $h \in \{1, \cdots, H\}$, we construct a (neural) model to approximate the probability in Eq. (2), resulting in H models in total for MSA prediction. We here formally define the model for the prediction of event occurrence at time $t + h$ as

$$\hat{y}_{t+h}^{(i,j)} = f_h(\tilde{\mathbf{x}}_t^{(i,j)}), \tag{3}$$

where $\tilde{\mathbf{x}}_t^{(i,j)}$ denotes the features extracted from $\mathbf{x}_t$ for location (i, j) to train the model. Note that as the occurrences of the extreme weather form binary labels (i.e., $y_{t+h}^{(i,j)} \in \{0, 1\}$), for each future time point $t + h$, the model in Eq. (3) simply treats the problem as a binary classification task.

3.2 Proposed Markov Conditional Forward (MCF) Model

For each future prediction time point $t + h$, as the information of the realized occurrences of the extreme weather (i.e., $\mathbf{y}_{t+h-1} = (y_{t+h-1}, y_{t+h-2}, \cdots)$) is generally available at the training stage, in this paper we propose a Markov conditional forward model that adopts the Markov property between occurrences of extreme weather for two consecutive future prediction time points to better leverage such useful information. Specifically, under the Markov assumption between y_{t+h} and y_{t+h-1}, for each future time point $t + h$ and

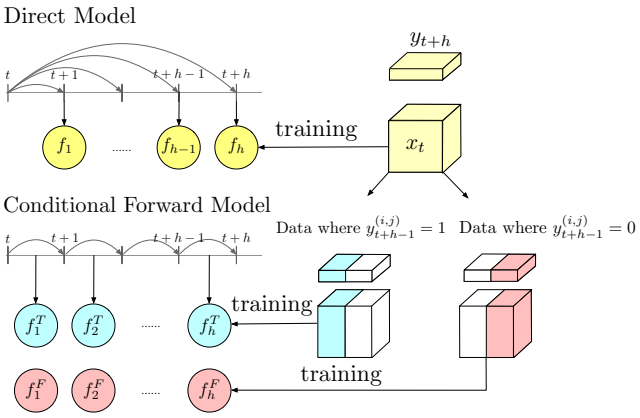

Figure 1: Direct and conditional forward models

a specific position in the data area (i, j), we formulate the inference for extreme weather prediction as

$$\begin{aligned}\hat{y}_{t+h}^{(i,j)} = & y_{t+h-1}^{(i,j)} \times p_{t+h-1,t+h}^{(i,j),T} \\ & + \left(1 - y_{t+h-1}^{(i,j)}\right) \times p_{t+h-1,t+h}^{(i,j),F},\end{aligned} \tag{4}$$

where

$$p_{t+h-1,t+h}^{(i,j),T} = P_{t+h-1,t+h}^{\text{Forward}}\left(y_{t+h}^{(i,j)} = 1 \,\middle|\, \mathbf{x}_t, y_{t+h-1}^{(i,j)} = 1\right), \tag{5}$$

$$p_{t+h-1,t+h}^{(i,j),F} = P_{t+h-1,t+h}^{\text{Forward}}\left(y_{t+h}^{(i,j)} = 1 \,\middle|\, \mathbf{x}_t, y_{t+h-1}^{(i,j)} = 0\right). \tag{6}$$

Above, Eq. (5) refers to the conditional probability of $y_{t+h}^{(i,j)} = 1$ given the observations on and before time t and the condition that the extreme weather occurs at previous time point $t+h-1$, whereas Eq. (6) refers to that given the observations on and before time t but with the condition that the extreme weather does not occur at previous time point $t + h - 1$.

Compared to the direct model, here we must train $2H$ models in total for MSA prediction as for each $h \in \{1, \cdots, H\}$, we construct two (neural) models to approximate the probabilities in Eqs. (5) and (6), respectively. Specifically, due to the availability of $y_{t+h-1}^{(i,j)}$ at the training stage, we use such information to separate the data into two groups—one with $y_{t+h-1}^{(i,j)} = 1$ and one with $y_{t+h-1}^{(i,j)} = 0$—to individually train models $f_h^T(\cdot)$ and $f_h^F(\cdot)$, respectively, as

$$p_{t+h-1,t+h}^{(i,j),T} = f_h^T(\tilde{\mathbf{x}}_t^{(i,j)}), \tag{7}$$

$$p_{t+h-1,t+h}^{(i,j),F} = f_h^F(\tilde{\mathbf{x}}_t^{(i,j)}), \tag{8}$$

for $h = 1, \cdots, H$, where $f_h^T(\cdot)$ and $f_h^T(\cdot)$ model the conditional probabilities in Eqs. (5) and (6) for forward period $(t + h - 1, t + h)$, respectively. Fig. 1 illustrates the concept of the direct and the proposed conditional forward models.

At the inference stage, we adopt Eq. (4) to recursively obtain $\hat{y}_{t+h}^{(i,j)}$ by using $f_h^T(\cdot)$ and $f_h^F(\cdot)$. Note that at time t, for each position (i, j), as the only information we have is $\mathbf{x}_t$

and $y_t^{(i,j)}$ for predicting $y_{t+h}^{(i,j)}$ for $h = 1, \cdots, H$, we use the predicted value $\hat{y}_{t+h-1}^{(i,j)}$ to approximate $y_{t+h-1}^{(i,j)}$ in Eq. (4) for $h > 1$. To better illustrate how we produce the probabilities at inference, we here provide an example for $H = 2$. First, for $h = 1$, from Eq. (4), we have

$$\hat{y}_{t+1}^{(i,j)} = y_t^{(i,j)} \times f_1^T(\tilde{\mathbf{x}}_t^{(i,j)}) + \left(1 - y_t^{(i,j)}\right) \times f_1^F(\tilde{\mathbf{x}}_t^{(i,j)}), \quad (9)$$

with which we have the prediction for time $t + 2$ as

$$\hat{y}_{t+2}^{(i,j)} = \hat{y}_{t+1}^{(i,j)} \times f_2^T(\tilde{\mathbf{x}}_t^{(i,j)}) + \left(1 - \hat{y}_{t+1}^{(i,j)}\right) \times f_2^F(\tilde{\mathbf{x}}_t^{(i,j)}). \quad (10)$$

Note that the setting in Eq. (4) inherits the Markov assumption on the relation between y_{t+h} and y_{t+h-1}, which makes such a conditional forward approach tractable in practice. This is due to the fact that if we consider the p past realized occurrences (i.e., $y_{t+h-1}, y_{t+h-2}, \cdots, y_{t+h-p}$) as the conditions to build the conditional forward models, the number of models grows exponentially when p increases. For example, if a Markov chain of order 2 is considered, for each h, we must construct four models for the following four different conditions, $(y_{t+h-2} = 1, y_{t+h-1} = 1), (y_{t+h-2} = 1, y_{t+h-1} = 0), (y_{t+h-2} = 0, y_{t+h-1} = 1), (y_{t+h-2} = 0, y_{t+h-1} = 0)$. The number of models with p-Markovian models is thus $2^p H$, which for large values of p is computationally expensive.

4 Cube Perturbation

During training, the proposed MCF model in Eqs. (5) and (6) leverages the ground truth of the former time step (i.e., y_{t+h-1})[1] to learn the two conditional forward models in Eqs. (7) and (8) for period $(t + h - 1, t + h)$. However, during inference, with Eq. (4), we must use the estimated (predicted) $\hat{y}_{t+h-1}$ to compute the prediction for each future time point when $h > 1$ (see Eqs. (9) and (10)). This is widely done in multiple-step-ahead time series prediction, such as in the moving average (Said and Dickey 1984) and in exponential smoothing (Gardner Jr 1985); in the field of neural networks, sequence-to-sequence models also take into account former predictions in the decoding phase (Chiu et al. 2018). However, these methods, which use predicted values to compute the next prediction, also accumulate error; that is, a bad prediction degrades the performance of future predictions. In this work, we propose perturbing the training labels by adding noise to increase the fault tolerance and generalization ability of the MCF model.

A naive way to perturb the data—here termed *naive perturbation*—is to set a predefined perturbation probability and then randomly switch the label from 0 to 1 (or from 1 to 0) based on the probability. However, as extreme weather is usually a rare event, the number of occurrences is relatively small compared to that of extreme-weather-free data points. Therefore, to reflect such an unbalanced data distribution, we propose three cube perturbation methods based on the following two assumptions:

(i) Due to their sparsity, positive data are far more critical for model training than negative data.

[1]For simplicity, we omit the superscript (i, j) hereafter.

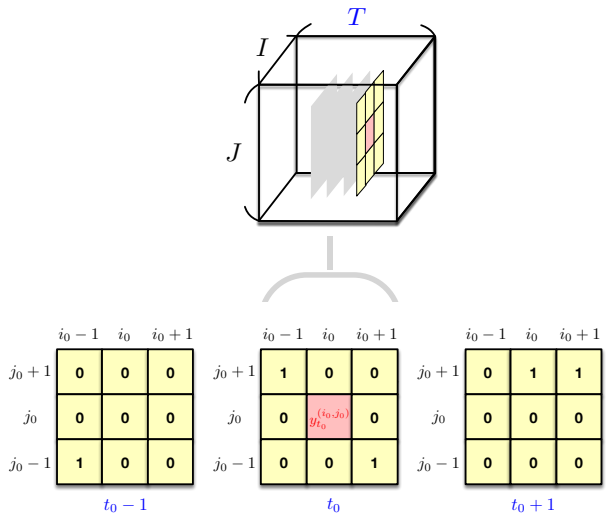

Figure 2: Cube perturbation candidates in SCP

(ii) Adjacent data points in terms of their physical locations on the given rectangle area or in terms of time can have a great impact on each other; i.e., there exist spatial and temporal correlations among labels in the obtained data.

With assumption (i), all of the three cube perturbation methods proposed in this paper perform label-switching on negative labels only and keep all positive labels unchanged. Moreover, the spatial and temporal correlations in assumption (ii) form the core idea of the three cube perturbation methods, for which two dimensions describe the physical locations of data points and one dimension describes the time factor. The three cube perturbation methods are introduced in detail in the following subsections. First, we define a set of indices of *neighbor* data points for data index (i_0, j_0, y_0) as

$$\begin{aligned}\mathcal{N}_{t_0}^{(i_0,j_0)} = &\{(i,j,t) \,|\, i \in \mathcal{I} \wedge j \in \mathcal{J} \wedge t \in \mathcal{T} \wedge \\ &|i - i_0| \le \kappa \wedge |j - j_0| \le \kappa \wedge |t - t_0| \le \kappa\} \setminus \\ &\{(i_0, j_0, t_0)\}, \end{aligned} \quad (11)$$

where $\mathcal{I} = \{1, \cdots, I\}$, $\mathcal{J} = \{1, \cdots, J\}$, and $\mathcal{T} = \{1, \cdots, T\}$; that is, the set $\mathcal{N}_{t_0}^{(i_0,j_0)}$ contains the indices of the neighbors of the data point at position (i_0, j_0) and time t_0. In Eq. (11), κ is related to the size of the cube in which the data points are considered as the neighbors of the center point; it is clear that the cube size is $(2\kappa + 1)^3$; for example, $\kappa = 1$ results in a $3 \times 3 \times 3$ cube perturbation.

4.1 Simple Cube Perturbation

We first propose simple cube perturbation (denoted as SCP hereafter) based on the aforementioned two assumptions. In SCP, a data point is a candidate for perturbation only if its label is negative and its neighboring data points contain positive data; hence, the label $y_{t_0}^{(i_0,j_0)}$ is a candidate for perturbation if

$$y_{t_0}^{(i_0,j_0)} \neq 1 \wedge \exists\, (i,j,t) \in \mathcal{N}_{t_0}^{(i_0,j_0)} \left(y_t^{(i,j)} = 1\right). \quad (12)$$

Fig. 2 shows an example for candidate selection, in which the red label $y_0^{(i_0,j_0)}$ is selected as a candidate as its original label is zero and it has more than one neighbors (i.e., five neighbors actually) having positive labels. With the perturbation candidates defined in Eq. (12), we set a universal rate $\xi^{\mathrm{SCP}} \in [0, 1]$, with which all perturbation candidates have the same probability to switch from the original label 0 to 1.

4.2 Neighborhood Probability Cube Perturbation

Instead of switching labels according to a universal rate ξ^{SCP}, we further propose a method termed neighborhood probability cube perturbation (abbreviated as NPCP hereafter), in which each position (i_0, j_0, t_0) is associated with its own probability for perturbation $\xi^{\mathrm{NPCP}}_{(i_0,j_0,t_0)}$, defined as

$$\xi^{\mathrm{NPCP}}_{(i_0,j_0,t_0)} = \frac{\left|\left\{(i,j,t) \;\middle|\; (i,j,t) \in \mathcal{N}^{(i_0,j_0)}_{t_0} \wedge y^{(i,j)}_t = 1\right\}\right|}{\left|N^{(i_0,j_0)}_{t_0}\right|}. \tag{13}$$

The intuition behind Eq. (13) is simple: for a given candidate location (i_0, j_0, t_0), the perturbation probability is proportional to how many of its neighbors have positive labels; that is, having more positive neighbors entails a higher perturbation probability.

4.3 Multinomial Cube Perturbation

A variant of the perturbation method is called multinomial cube perturbation (MCP). In probability theory, the multinomial distribution is a generalization of the binomial distribution, which for instance models the probability of counts for each side of a k-sided die rolled n times. Mathematically, we have k possible mutually exclusive outcomes, with corresponding probabilities $p_1, \cdots, p_k$, and n independent trials. In contrast to the previous two methods, MCP regards the choice of data points for perturbation as a multinomial distribution with $k = I \times J \times T$, for which the multinomial probability of the point at (i_0, j_0, t_0) is formulated as

$$\xi^{\mathrm{MCP}}_{(i_0,j_0,t_0)} = \frac{\left|\left\{(i,j,t) \;\middle|\; (i,j,t) \in \mathcal{N}^{(i_0,j_0)}_{t_0} \wedge y^{(i,j)}_t = 1\right\}\right|}{\sum\limits_{\forall \hat{i},\hat{j},\hat{t}} \left|\left\{(i,j,t) \;\middle|\; (i,j,t) \in \mathcal{N}^{(\hat{i},\hat{j})}_{\hat{t}} \wedge y^{(i,j)}_t = 1\right\}\right|}, \tag{14}$$

where $\hat{i} \in \mathcal{I}$, $\hat{j} \in \mathcal{J}$, and, $\hat{t} \in \mathcal{T}$. With the event probabilities defined in Eq. (14), we then roll the k-sided die n times to obtain n data points for perturbation, where n is associated with the hyperparameter for sampling rate $r^{\mathrm{MCP}} = n/k$.

5 Experiments

5.1 Dataset and Settings

We conducted extensive experiments on global atmospheric states from 1979 to 1980 extracted from the ExtremeWeather dataset (Racah et al. 2017). The details of the dataset and the criteria of choosing the area for experiments are described in Appendices. In the experiments, we examine the proposed method on two extreme weather phenomena: tropical cyclones and extratropical cyclones.

To build the prediction models, we chose September 1979 and June 1979, respectively, as our training data. We then used the data of the same month in 1980 for testing; for validation, we chose the data from the month preceding the testing data to tune the hyperparameters of the perturbation methods; this corresponds to August 1980 and May 1980 for tropical and extratropical cyclones, respectively.

The prediction horizons were set to every 6 hours for a total of 48 hours. We took each pixel in an image as a training instance, labeling pixels within the 150×150 bounding box as 1 or 0 according to the extreme weather occurrence at that point. As the temporal resolution of data was 6 hours, we had $150 \times 150 \times 30 \times 4 = 2,700,000$ training instances in total; that is $I = 150$, $J = 150$, $T = 120$. In the experiments, we leveraged two machine learning algorithms—logistic regression (LR) and convolutional neural networks (CNN), the detailed settings of which are described in Appendices.

The prediction performance was evaluated with AUC, the area under the receiver operating characteristic (ROC) curve, which is a common evaluation metric for imbalanced data. For comparison purposes, except for the three cube perturbation methods, we also implemented naive perturbation, which randomly switches the label from 0 to 1 based on a universal predefined probability, ξ^{Naive}. Note that we selected the hyperparameters of each perturbation method with the best results on the validation data, after which we used these hyperparameters to compare their performance on the test data; in the experiments, we report the average performance over 20 repetitions.

5.2 Quantitative Results

Comparison of Direct and Markov Conditional Forward Models Table 1 first tabulates the performance of the MCF model and the direct model. As shown in Table 1, for both events (i,e., tropical and extratropical cyclones), MCF significantly outperforms the direct model in terms of AUC, especially for short-term prediction. Moreover, for extratropical cyclones, MCF even achieves a 13% performance improvement for the shortest prediction horizon (i.e., 6 hours). On average, the proposed method improves performance by over 4% compared to the direct model. These results clearly demonstrate the superiority of the MCF model, especially for short-term prediction.

Comparison of Perturbation Methods We here compare the performance of different perturbation methods, the goal of which is to enhance the fault tolerance and generalization ability of the MCF and thus generates more accurate predictions, long-term prediction in particular. Table 2 shows the experimental results of the four perturbation methods along with MCF; note that due to the superiority of the MCF model with CNN over other compared models shown in Table 1, we here only use the no-perturbation MCF model with CNN as our compared baseline. To reduce the complexity of hyperparameter search, we adopted a $3 \times 3 \times 3$ cube (i.e., $\kappa = 1$ in Eq. (11)) for the three cube perturbation methods; the effect of different cube sizes is discussed in Section 5.3. with this setting, we tuned hyperparameters ξ^{naive}, r^{mcp}, and ξ^{scp} of each perturbation method on the validation data; the best

Tropical cyclones									
Horizons (hours)	6	12	18	24	30	36	42	48	Average
LR (direct)	0.6682	0.6664	0.6641	0.6490	0.6472	0.6502	0.6576	0.6428	0.6557
LR (MCF)	0.7376	0.7213	0.7081	0.6747	0.6598	0.6634	0.6567	0.6428	0.6830
LR improvement (%)	$\mathbf{10.39}^{**}$	$\mathbf{8.25}^{**}$	$\mathbf{6.63}^{**}$	$\mathbf{3.96}^{**}$	$\mathbf{1.93}^{**}$	$\mathbf{2.04}^{**}$	$\mathbf{-0.14}^{*}$	0.00	4.16
CNN (direct)	0.6760	0.6760	0.6773	0.6629	0.6577	0.6601	0.6674	0.6518	0.6662
CNN (MCF)	0.7441	0.7309	0.7188	0.6864	0.6739	0.6756	0.6689	0.6506	0.6937
CNN improvement (%)	$\mathbf{10.07}^{**}$	$\mathbf{8.12}^{**}$	$\mathbf{6.13}^{**}$	$\mathbf{3.55}^{**}$	$\mathbf{2.46}^{**}$	$\mathbf{2.35}^{**}$	0.22	-0.18	4.13
Extratropical cyclones									
Horizons (hours)	6	12	18	24	30	36	42	48	Average
LR (direct)	0.8243	0.8113	0.7976	0.7831	0.7628	0.7429	0.7303	0.7242	0.7721
LR (MCF)	0.9361	0.8858	0.8426	0.8104	0.7821	0.7564	0.7376	0.7275	0.8098
LR improvement (%)	$\mathbf{13.56}^{**}$	$\mathbf{9.19}^{**}$	$\mathbf{5.64}^{**}$	$\mathbf{3.48}^{**}$	$\mathbf{2.53}^{**}$	$\mathbf{1.81}^{**}$	$\mathbf{1.00}^{**}$	$\mathbf{0.46}^{**}$	4.88
CNN (direct)	0.8244	0.8119	0.7979	0.7843	0.7734	0.7614	0.7489	0.7396	0.7802
CNN (MCF)	0.9360	0.8851	0.8421	0.8103	0.7852	0.7651	0.7532	0.7413	0.8148
CNN improvement (%)	$\mathbf{13.54}^{**}$	$\mathbf{9.02}^{**}$	$\mathbf{5.54}^{**}$	$\mathbf{3.32}^{**}$	$\mathbf{1.53}^{**}$	$\mathbf{0.49}^{**}$	$\mathbf{0.57}^{**}$	$\mathbf{0.23}^{**}$	4.43

"*" and "**" denote statistical significance at p-value ¡ 0.05 and ¡ 0.01, respectively, with a Student's t-test.

Table 1: Performance (in terms of AUC) of direct and conditional forward models

hyperparameters are listed in parentheses after the name of each method in Table 2.

As shown in Table 2, the proposed three cube perturbation methods generally outperform naive perturbation. The results show that using naive perturbation usually worsens performance; for example, for the prediction of tropical cyclones at $h = 18$ hours, the AUC decreases around 4% with naive perturbation. In contrast, for both datasets, the three cube perturbation methods do not harm the original CNN-based MCF model. Moreover, for the prediction of tropical cyclones, some of the cube perturbation methods successfully and significantly enhance long-term prediction performance; e.g., there is a 4.76% improvement for $h = 36$ with the neighborhood probability cube perturbation method introduced in Section 4.2.

As for extratropical cyclones, from both Tables 1 and 2, we observe that the prediction for such an event is much easier than that for tropical cyclones, as clearly demonstrated by its high AUC values in both tables. In this case, although the improvement yielded by the proposed cube perturbation is relatively minor, it does not degrade performance and thus still yields results comparable to the model learned from data without perturbation.

5.3 Sensitivity Analysis

We further conducted a sensitivity analysis on the hyperparameters of different perturbation methods. Note that as the differences among the four perturbation methods are relatively insignificant on the extratropical cyclone data, we conducted the sensitivity analysis only on the tropical cyclone data; we present the results on the test data. The details of our sensitivity experiments are presented in appendices.

With the observations of sensitivity analysis, we summarize general guidelines for perturbation method and hyperparameter selection:

- Naive perturbation always yields the worst performance, especially when accompanied with an inappropriate perturbation rate.
- Multinomial cube perturbation is sensitive to neither sampling rate nor cube size, as event probabilities are endogenously generated from the data and thus successfully encode representative information; hence this can be regarded as a hyperparameter-free perturbation method.
- Increasing cube size does not improve performance for cube perturbation methods; thus, a $3 \times 3 \times 3$ cube is the best setting.
- Neighborhood probability cube perturbation yields the best performance, as shown in both Table 2 and appendices; most importantly, with a fixed cube size of $3 \times 3 \times 3$, it is hyperparameter-free.

6 Conclusion

This paper develops a Markov conditional forward (MCF) model to better address MSA extreme weather prediction, one of the most challenging prediction problems in climate science. Specifically, with the use of neural networks, we train two conditional forward models between two consecutive prediction horizons, by which we recursively compose the prediction for all future prediction horizons. To further enhance the fault tolerance and generalization ability of the MCF model, we present three unconventional cube perturbation methods. Experimental results on a real-world extreme weather dataset suggest that the proposed MCF model together with the cube perturbation methods yields significantly better performance than the direct model for both short-term and long-term predictions. Along with the contributions made in extreme weather prediction, this paper provides vision and approaches for research and applications that require MSA prediction and simultaneously uses the observed covariates and future realized occurrences for model training.

Tropical cyclones									
Horizons (hours)	6	12	18	24	30	36	42	48	Average
CNN (MCF)	0.7441	0.7309	0.7188	0.6864	0.6739	0.6756	0.6689	0.6506	0.6937
+Naive ($\xi^{\text{Naive}} = 5\%$)	0.7405	0.7144	0.6898	0.6614	0.6551	0.6555	0.6602	0.6438	0.6776
+Multinomial cube ($r^{\text{MCP}} = 10\%$)	0.7395	0.7266	0.7182	0.6863	0.6749	0.6791	0.6678	0.6481	0.6926
+Simple Cube ($\xi^{\text{SCP}} = 47.5\%$)	**0.7458**	**0.7317**	0.7266	0.6999	0.6908	0.7045	0.6820	0.6633	0.7056
+Neighborhood prob cube	0.7445	0.7308	**0.7272**	**0.7002**	**0.6922**	**0.7077**	**0.6846**	**0.6668**	**0.7068**
Improvement (%)	0.23	0.11	**1.16***	**2.01***	**2.71****	**4.76****	**2.35***	**2.48***	1.89
Extratropical cyclones									
Horizons (hours)	6	12	18	24	30	36	42	48	Average
CNN (MCF)	0.9360	0.8851	0.8421	0.8103	0.7852	0.7651	0.7532	0.7413	0.8148
+Naive ($\xi^{\text{Naive}} = 2.5\%$)	**0.9371**	0.8808	0.8288	0.7957	0.7716	0.7566	0.7470	0.7385	0.8070
+Multinomial cube ($r^{\text{MCP}} = 5\%$)	0.9365	0.8861	0.8425	0.8105	**0.7857**	0.7661	0.7545	0.7435	0.8157
+Simple Cube ($\xi^{\text{SCP}} = 47.5\%$)	0.9356	0.8855	0.8428	0.8101	0.7845	**0.7663**	**0.7558**	**0.7442**	0.8156
+Neighborhood prob cube	0.9368	**0.8873**	**0.8444**	**0.8112**	0.7849	0.7650	0.7538	0.7428	**0.8158**
Improvement (%)	0.12	0.24	0.27	0.12	0.06	0.16	0.35	0.39	0.12

"*" and "**" denote statistical significance at p-value ¡ 0.05 and ¡ 0.01, respectively, with a Student's t-test.
"Improvement (%)" denotes the percentage improvement of best perturbation method (in boldface) with respect to CNN conditional forward model without perturbation.

Table 2: Performance of various perturbation methods

References

Al-Qahtani, F. H.; and Crone, S. F. 2013. Multivariate k-nearest neighbour regression for time series data—A novel algorithm for forecasting UK electricity demand. In *Proc. of the 2013 International Joint Conference on Neural Networks*, IJCNN'13, 1–8.

Bontempi, G.; Le Borgne, Y.-A.; and De Stefani, J. 2017. A dynamic factor machine learning method for multi-variate and multi-step-ahead forecasting. In *Proc. of the 2017 IEEE International Conference on Data Science and Advanced Analytics*, DSAA'17, 222–231.

Chang, F.-J.; Chiang, Y.-M.; and Chang, L.-C. 2007. Multi-step-ahead neural networks for flood forecasting. *Hydrological Sciences Journal* 52(1): 114–130.

Cheng, H.; Tan, P.-N.; Gao, J.; and Scripps, J. 2006. Multistep-ahead time series prediction. In *Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, PAKDD'06, 765–774.

Chiu, C.-C.; Sainath, T. N.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R. J.; Rao, K.; Gonina, E.; et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *Proc. of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP'18, 4774–4778.

Duan, J.-C.; Sun, J.; and Wang, T. 2012. Multiperiod corporate default prediction—A forward intensity approach. *Journal of Econometrics* 170(1): 191–209.

Duffie, E.; Saita, L.; and Wang, K. 2007. Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics* 83(3): 635–665.

Gardner Jr, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting* 4(1): 1–28.

Holden, P. B.; Edwards, N. R.; Garthwaite, P. H.; and Wilkinson, R. D. 2015. Emulation and interpretation of high-dimensional climate model outputs. *Journal of Applied Statistics* 42(9): 2038–2055.

Hussein, S.; Chandra, R.; and Sharma, A. 2016. Multi-step-ahead chaotic time series prediction using coevolutionary recurrent neural networks. In *Proc. of the 2016 IEEE Congress on Evolutionary Computation*, CEC'16, 3084–3091.

Hwang, J.; Orenstein, P.; Cohen, J.; Pfeiffer, K.; and Mackey, L. 2019. Improving subseasonal forecasting in the western US with machine learning. In *Proc. of the 2019 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, SIGKDD'19, 2325–2335.

Kim, S. 2019. Focus and Track: pixel-wise spatio-temporal hurricane tracking. Technical report, Lawrence Livermore National Lab. (LLNL), Livermore, CA (United States).

Klein, B.; Wolf, L.; and Afek, Y. 2015. A dynamic convolutional layer for short range weather prediction. In *Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR'15, 4840–4848.

Koesdwiady, A.; El Khatib, A.; and Karray, F. 2018. Methods to Improve Multi-Step Time Series Prediction. In *Proc. of the 2018 International Joint Conference on Neural Networks*, IJCNN'18, 1–8.

Krasnopolsky, V. M.; and Fox-Rabinovitz, M. S. 2006. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural Networks* 19(2): 122–134.

Qiu, M.; Zhao, P.; Zhang, K.; Huang, J.; Shi, X.; Wang, X.; and Chu, W. 2017. A short-term rainfall prediction model using multi-task convolutional neural networks. In *Proc. of the 2017 IEEE International Conference on Data Mining*, ICDM'17, 395–404.

Racah, E.; Beckham, C.; Maharaj, T.; Kahou, S. E.; Prabhat, M.; and Pal, C. 2017. ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. In *Proc. of the 2017 Advances in Neural Information Processing Systems*, NIPS'17, 3402–3413.

Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; et al. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature* 566(7743): 195–204.

Rolnick, D.; Donti, P. L.; Kaack, L. H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A. S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; et al. 2019. Tackling Climate Change with Machine Learning. *ArXiv Preprint ArXiv:1906.05433* .

Said, S. E.; and Dickey, D. A. 1984. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika* 71(3): 599–607.

Scher, S. 2018. Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters* 45(22): 12–616.

Scher, S.; and Messori, G. 2018. Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society* 144(717): 2830–2841.

Sorjamaa, A.; Hao, J.; Reyhani, N.; Ji, Y.; and Lendasse, A. 2007. Methodology for long-term prediction of time series. *Neurocomputing* 70(16-18): 2861–2869.

Taieb, S. B.; and Atiya, A. F. 2015. A bias and variance analysis for multistep-ahead time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems* 27(1): 62–76.

Taieb, S. B.; Bontempi, G.; Atiya, A. F.; and Sorjamaa, A. 2012. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Systems with Applications* 39(8): 7067–7083.

Xingjian, S.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-k.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Proc. of the 2015 Advances in Neural Information Processing Systems*, NIPS'15, 802–810.