Improving Convection Trigger Functions in Deep Convective Parameterization Schemes Using Machine Learning

Tao Zhang[1], Wuyin Lin[1], Andrew M. Vogelmann[1], Minghua Zhang[2], Shaocheng Xie[3], Yi Qin[3], Jean-Christophe Golaz[3]

[1] Brookhaven National Laboratory, Upton, New York, USA

[2] School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York, USA

[3]Lawrence Livermore National Laboratory, Livermore, California, USA

Key Points:

- A machine learning convective trigger function greatly outperforms four CAPE-based triggers at two distinct convective regimes.
- Insights are derived from the machine learning trigger that could be used to improve existing traditional CAPE-based triggers.
- Results suggest that a unified machine learning trigger function could be developed for use in climate models.

Abstract

Deficiencies in convection trigger functions used in deep convection parameterizations in General Circulation Models (GCMs) have critical impacts on climate simulations. A novel convection trigger function is developed using the machine learning (ML) classification model XGBoost. The large-scale environmental information associated with convective events is obtained from the long-term constrained variational analysis forcing data from the Atmospheric Radiation Measurement (ARM) program at its Southern Great Plains (SGP) and Manaus (MAO) sites representing, respectively, continental mid-latitude and tropical convection. The ML trigger is separately trained and evaluated per site, and jointly trained and evaluated at both sites as a unified trigger. The performance of the ML trigger is compared with four convective trigger functions commonly used in GCMs: dilute convective available potential energy (CAPE), undilute CAPE,

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2020MS002365.

dilute dynamic CAPE (dCAPE), and undilute dCAPE. The ML trigger substantially outperforms the four CAPE-based triggers in terms of the F1 score metric, widely used to estimate the performance of ML methods. The site-specific ML trigger functions can achieve, respectively, 91% and 93% F1 scores at SGP and MAO. The unified trigger also has a 91% F1 score, with virtually no degradation from the site-specific training, suggesting the potential of a global ML trigger function. The ML trigger alleviates a GCM deficiency regarding the overprediction of convection occurrence, offering a promising improvement to the simulation of the diurnal cycle of precipitation. Furthermore, to overcome the black box issue of the ML methods, insights derived from the ML model are discussed, which may be leveraged to improve traditional CAPE-based triggers.

Plain Language Summary

Deficiencies in convection trigger function, a set of conditions used to determine whether the convection will be activated at a given time in General Circulation Models (GCMs), have critical impacts on model simulated climate. This work presents a novel convection trigger function using a machine learning (ML) model. Environmental information on convective events are obtained from long-term data from the Atmospheric Radiation Measurement (ARM) program at its Southern Great Plains (SGP) and Manaus (MAO) sites, which represent two distinct convective regimes. The ML trigger is separately trained and evaluated per site, and jointly trained and evaluated at both sites as a unified trigger. The ML trigger substantially outperforms the four CAPE-based triggers commonly used in GCMs. The unified trigger virtually has no degradation from the site-specific training, suggesting some promise to develop a global trigger function. The ML trigger alleviates a GCM deficiency regarding the overprediction of convection occurrence, offering a promising improvement to the simulation of the diurnal cycle of precipitation. Furthermore, to overcome the black box issue of the ML methods, insights derived from the ML model are discussed, which may be leveraged to improve traditional CAPE-based triggers.

1 Introduction

Convection is critical to precipitation, heat and moisture transport, cloud amount and distribution, as well as to the global energy budget (Arakawa and Schubert, 1974; Arakawa, 2004). Properly representing convection is critical to successful numerical weather prediction (NWP) and climate simulations by general circulation models (GCMs). However, because current coarse-resolution models cannot explicitly resolve convection, it must be represented with convective parameterizations. Deficiencies in these parameterizations can cause a multitude of problems in simulation results. For example, GCMs often rain too frequently and at reduced intensity (Dai and Trenberth, 2004; Trenberth et al. 2003) and these deficiencies are conspicuously manifested in simulating the diurnal cycle of precipitation (Lee et al., 2007; Covey et al., 2016). Over land, convection is often overly active, and the diurnal cycle cannot be well-represented during the summer season (Dai, 2006; Lee et al., 2007). Over ocean, the amplitude of the diurnal cycle is typically too weak, and the peak time does not match observations (Dai, 2006). These problems are known to be closely related to convection trigger function used in the convective parameterizations (e.g., Xie and Zhang, 2000; Xie et al., 2004a; Dai and Trenberth, 2004; Lee et al., 2008; Xie et al., 2019; Zheng et al., 2019; Wang et al., 2020).

The convective trigger in a convective parameterization determines whether the convection will be activated at a given time. Dating back to early developments of cumulus parameterization, Kuo (1974, 1965) proposed a large-scale moisture convergence-based trigger. Fritsch and Chappell (1980), Kain and Fritsch (1993), and Rogers and Fritsch (1996) developed trigger functions by considering perturbations of vertical velocity and temperature on the large-scale low-level convergence to inhibit convection when the low-level upward motion was weak. Nowadays, most GCMs assume that convection is triggered when there is positive convective available potential energy (CAPE), a measurement of atmospheric instability and the potential energy that can be released by convection. For instance, in the Zhang-McFarlane (ZM) deep convection scheme (Zhang and McFarlane, 1995), convection is triggered when CAPE is larger than a specified threshold value. When this threshold is reached, the accumulated instability is released over a prescribed time scale. However, this type of trigger function tends to activate convection too frequently because CAPE is almost always positive in the tropics and can also be generated easily during daytime in the warm season over midlatitude lands due to surface solar heating. There are

published works that attempt to alleviate this problem by introducing dynamic and thermodynamic constraints to prevent CAPE from being released too early. For example, Xie and Zhang (2000) found that the observed precipitation correlates well with a positive dynamic CAPE generation rate (dCAPE), generated by the large-scale advective tendencies of temperature and moisture. They proposed a dCAPE-based trigger that assumes convection is triggered when there is positive dCAPE in addition to the presence of positive CAPE. Xie et al. (2019) further improved the dCAPE trigger with an unrestricted air parcel Launch Level (ULL), which allowed parcels to launch above the boundary layer to capture nocturnal elevated convection, and led to a dramatic improvement in the phase of the diurnal cycle of precipitation. Neale and Jochum (2008) introduced entrainment dilution into the CAPE calculation, which was used in the atmosphere models of the Community Earth System Model Version 2 (CESM2, Danabasoglu et al. 2020) and the Energy Exascale Earth System Model (E3SM, Xie et al. 2018; Rasch et al. 2019).

However, these triggers also suffer from large uncertainties and are ad hoc because the mechanism of deep convection occurrence is not fully understood; so most convective parameterization schemes are simplified to empirical CAPE-based conditions with artificial thresholds. Suhas and Zhang (2014) and Song and Zhang (2017) evaluated several trigger functions against data from the Atmospheric Radiation Measurement (ARM) user facility, including those used in the Arakawa-Schubert (AS) scheme (Arakawa & Schubert, 1974), the Bechtold scheme (Bechtold et al., 2001), the Donner scheme (Donner,1993), the Kain-Fritsch (KF) scheme (Kain, 2004), the Tiedtke scheme (Tiedtke, 1989), and the four variants of CAPE-based triggers (Zhang and McFarlane, 1995; Xie and Zhang, 2000; Neale et al., 2008). Their results revealed that the skill of those trigger functions leaves much room for improvement.

Machine learning (ML) methods have demonstrated substantial skill for classification and regression and have gained recent attention in the geosciences (Kurth et al., 2018; Ham et al., 2019). ML-based physical parameterizations have been developed for GCMs to improve climate simulations, such as for parameterization of convection and atmospheric chemistry (Brenowitz and Bretherton, 2018, 2019; Gentine et. al., 2018; Rash et. al., 2018; Silva et. al., 2019; Han et. al., 2020). ML methods have also been applied to predict specific physical quantities or events, such as rainfall rates (Tao et al., 2016; Miao et al., 2019), tropical cyclone genesis (Zhang et. al., 2019),

and ENSO (Ham et. al., 2019). Using ML in these geoscience applications can achieve a more accurate prediction by, for example, learning from high-resolution simulations or multi-source observations. The success is gauged not only to their flexibility to represent complex multi-variable non-linear structures, but also to their low computational cost once the models or schemes are well trained.

In this study, we use a ML model to construct a novel convection trigger function trained on the long-term variationally constrained ARM forcing dataset (VARANAL) at its Southern Great Plains (SGP) site, in the central US, and the Manaus (MAO) site, in the Amazon basin. The ML model XGBoost is trained on predictive factors that are associated with convective processes, such as surface heat fluxes, CAPE, lifting condensation level (LCL), convective inhibition (CIN), vertical distribution of temperature, humidity, wind shear, and large-scale advective tendencies of water vapor and dry static energy. We will show that the ML trigger scores equally high for both site specific and cross-site trainings and offers a promising improvement to simulation of the diurnal cycle of precipitation.

The paper is organized as follows: Section 2 describes the training datasets and predictors. Section 3 describes commonly used CAPE-based convection trigger functions and the ML XGBoost trigger functions (hereafter the XGB triggers) and the performance metrics used for assessment. Section 4 presents the performance of ML methods including comparison with previous studies. Conclusion and discussions are given in Section 5.

2 Data

Data used for this study are from the ARM continuous forcing and evaluation data products at its SGP and MAO sites (Xie et al. 2004b, Tang et al. 2016). The data were developed using NWP analyses constrained by the observed surface and top-of-atmosphere measurements using a variational analysis approach (Zhang and Lin, 1997; Zhang et al. 2001). At SGP, the Rapid Update Cycle (RUC) analysis product from the National Oceanic and Atmospheric Administration (NOAA) provides the background fields (Xie et al. 2004b), while for MAO the European Center for Medium-Range Weather Forecasts (ECMWF) analyses are used (Tang et al. 2016). Eleven boreal summer seasons (June, July, August) from 1999 to 2009 are used for SGP, and two years

of data are used for MAO from 2014 to 2015 that cover the Green Ocean Amazon (GoAmazon2014/15) field campaign (Martin et al., 2016). The datasets have a time resolution of 1 h for SGP and 3 h for MAO. Both datasets have a vertical resolution of 25 hPa.

A ML trigger is sensitive to how it is trained. For a trigger to be functional in a large-scale model, it should be trained with data accounting for all convective conditions. A broad subset of these conditions is represented here using data for continental mid-latitude summer and tropical convective conditions. This study illustrates the efficacy of the ML-based trigger and the sensitivity to how the trigger is trained. The ML models are evaluated by separately training for the two sites, as well as a joint training that combines the data from both sites. The training dataset contains a number of large-scale predictors summarized in Table 1. Because moist convection depends not only on the atmospheric state near the surface (temperature, specific humidity, etc.) but also on their vertical distribution (Emanuel, 1994), the predictors include scalar variables—such as surface heat fluxes, surface temperature and relative humidity, CAPE, LCL, and CIN—as well as the vertical profiles of temperature, specific humidity, wind shear, and advective tendencies. These predictors are all involved in triggering the convection and the ensuing precipitation processes. For example, CAPE is often used as the criterion to determine the occurrence of deep convection (Zhang and McFarlane, 1995), and surface heat fluxes are the main sources of heat and moisture for local convection that can disturb the large-scale atmospheric column (Kuo, 1974) and enhance surface precipitation (Tao et al., 1991). For strong convection, horizontal moisture convergence is a dominant source of water vapor. CIN represents how much kinetic energy must be added to a parcel to lift it to the level of free convection. If CIN is sufficiently large, deep convection can be suppressed. The LCL is included since it is a critical property of a convective air parcel that determines cloud base height and the thermodynamic structure of the convective cloud plume. Finally, previous work from observational and numerical studies demonstrated that vertical wind shear affects the intensity and organization of convection (Chen et. al., 2015). To simplify interpretation of the results, the vertical profiles of temperature, humidity, advective tendencies and wind shear are represented by three layers of 700-800, 300-700, and 200-300 hPa corresponding to, respectively, the lower, middle, and upper troposphere (Gyakum and Cai, 1990; Chen et. al., 2015). We note that the performance of the trigger is similar to when the values from the fully profiles are used (not shown).

The ML model must be trained with both convective and non-convective events to determine the unique features required for convection. Because the definition of convection occurrence is ambiguous (Suhas and Zhang, 2014; Song and Zhang, 2017), exactly determining the onset of deep convection is nontrivial. In this study, the occurrence of deep convection is determined when the precipitation rate is greater than or equal to 0.5 mm/hour, following Suhas and Zhang (2014) and Song and Zhang (2017). Note that artificial and empirical thresholds may introduce some uncertainties. The precipitation threshold cannot distinguish convective and stratiform precipitation, but the threshold is justifiable as large precipitation in summer over midlatitude continents and in the tropics are commonly associated with convective events. It is noted that the cloud top or cloud thickness could also define the occurrence of deep convection. However, they still have uncertainties, depending on the evolution stage of the convective system. Further research could possibly explore a combination of precipitation and cloud geometries to determine convection that might lead to a more robust indicator when data are not missing.

3 Method

In this study, the ML trigger function is determined by the XGBoost algorithm, which is a state-of-the-art classifier that is widely used by data scientists to achieve perfect performance on many ML challenges (Chen and Guestrin, 2016; Vanichrujee et al., 2018; Zhong et al., 2018; Zamani Joharestani et al. 2019; Zhang et al., 2019; Zheng and Wu, 2019). For comparison, four variants of the CAPE-based trigger functions are evaluated: undilute CAPE, dilute CAPE, undilute dCAPE and dilute dCAPE.

3.1 CAPE-based convection trigger functions

CAPE-based triggers are commonly used in deep convection schemes such as the ZM scheme, which is one of the common deep convection schemes used in several climate models including the NCAR CAM and the Department of Energy E3SM Atmosphere Model (EAM). CAPE is defined as the vertical integral of the local buoyancy of a parcel from the launch level to the equilibrium level,

$$\mathrm{CAPE} = \int_{p_t}^{p_b} R_d (T_{vp} - T_{ve}) d \ln p \qquad (1)$$

where T_{vp} and T_{ve} are, respectively, the virtual temperature of the parcel and its environment. In the ZM scheme, the launch level is the level with the largest moist static energy within the boundary layer. The equilibrium level is where the air parcel become buoyancy neutral with respect to the environment. R_d is the gas constant for dry air. p_t is the pressure of the equilibrium level and p_b is the pressure of the launch level. This definition of CAPE is hereafter called undilute CAPE. By construction, the undilute CAPE trigger scheme prevents the accumulation of instability in the atmosphere beyond the threshold value and results in generally triggering convection too frequently and too early during the day (e.g., Xie and Zhang, 2000).

Neale et al. (2008) introduced the dilution effect of entrained air into the CAPE calculation (hereafter called dilute CAPE). In this trigger function, the entropy S of an ascending parcel is governed by,

$$\frac{\partial mS}{\partial z} = \frac{\partial m}{\partial z}\bar{S} = \varepsilon\bar{S} \qquad (2)$$

Where *m* is the mass of a parcel, and ε is the environmental entrainment of the rising air parcel per unit height which is assumed to be a constant (e.g., 10^{-3} m^{-1}). $\bar{S}$ is the entropy of the environmental air. The temperature and specific humidity of a parcel are updated when the entropy S at each height is obtained. Through the changes in the temperature and specific humidity of the air parcel, the latent heat from condensation and freezing is also involved in the CAPE calculation (Zhang, 2009). Convection is initiated when the dilute CAPE value exceeds the threshold (e.g., 70 J/kg as used for CAM5 low-resolution configuration). This definition of trigger function is used in the EAM (Golaz et al., 2019) and the NCAR CAM version 6 (CAM6) (Gettelman et al., 2019).

Xie and Zhang (2000) introduces the dynamic CAPE generation rate (dCAPE) trigger, which is a function of the large-scale advective tendencies of temperature and moisture and is defined by:

$$\text{dCAPE} = \frac{CAPE[T + adv(T)\delta t, q + adv(q)\delta t] - CAPE[T, q]}{\delta t} \qquad (3)$$

where T and q are temperature and specific humidity and *adv(T)* and *adv(q)* are the corresponding advective tendencies, which include both horizontal and vertical advections. δ*t* is the time interval, which is, respectively, 1 hour for the SGP data and 3 hours for the MAO data. This definition can be applied to either undilute or dilute CAPE. Deep convection is assumed to be initiated when the large-scale advection has a positive contribution to CAPE. In other words, the dCAPE value should

be greater than zero. Suhas and Zhang (2014) and Song and Zhang (2017) set the dCAPE threshold at 65 J/kg/h. The dCAPE trigger was found to be one of the best performers among serval trigger functions—such as the dilute CAPE, Bechtold, Tidetke, and heated condensation framework (HCF)—when compared against the ARM data at SGP, MAO, as well as the data collected from the GARP Atlantic Tropical Experiment (GATE) and the Tropical Ocean Global Atmosphere Coupled Ocean-Atmosphere Response Experiment (TOGA-COARE) (Suhas and Zhang, 2014; Donner and Phillps, 2003; Song and Zhang, 2017; Wang et al., 2020).

3.2 XGBoost convection trigger functions

XGBoost, short for eXtreme Gradient Boosting, is a state-of-the-art, tree-based ML classification and regression method. It is designed to be highly efficient, flexible and portable, and has been successfully applied in many applications, achieving extraordinary performance (Chen and Guestrin, 2016).

The gradient boosting algorithm is the foundation of XGBoost, which is an ensemble approach combining many basic weak ML methods into a more generalized model, illustrated in Figure 1. Typically, a decision tree is used as the basic weak ML model, which mimics how humans think and make decisions based on a series of rules that are organized in a tree shape (Quinlan 1987). Rather than training all models in isolation of one another, XGBoost works in a stage-wise manner, iteratively adding a tree that aims to correct the errors of prior trees and then combining all trees in a weighted average to make the final prediction. The advantage of the boosting methods is that the new model being added focuses on correcting the mistakes remaining from the previous models. In contrast, other ensemble methods train the models in isolation and the results from all trees are combined by averaging or applying the "majority rules" (e.g., as in the Random Forests method); however, training the models in isolation might simply lead to each making the same mistakes.

The XGBoost trigger functions (XGB triggers) are trained and evaluated by the long-term ARM continuous forcing data. The dataset consists of the selected environmental predictors shown in Table 1 and the predictand, which is a binary variable indicating whether or not convection is triggered. The trigger functions for SGP and MAO can be trained separately or jointly.

3.3 Performance metrics

The output of the trigger functions is whether deep convection occurs or not; so a 2x2 contingency table serves to count the number of the four possible outcomes in a two-category classification. If a trigger function correctly determines a convective or a non-convective event, the samples from the testing set are labeled, respectively, as true positive (TP; correct convection prediction) or true negative (TN; correct non-convection prediction). Those incorrectly determined are labeled as false positive (FP; overprediction) or false negative (FN; underprediction).

Precision (P) and Recall (R) are the two performance criteria that are calculated based on the contingency table, defined as (Van Rijsbergen 1986; Olson and Delen 2008):

$$\text{Precision (P)} = \frac{TP}{TP + FP} \qquad (4)$$

$$\text{Recall (R)} = \frac{TP}{TP + FN} \qquad (5)$$

Precision is the percentage of true positives in the total predicted positive case. Many overpredictions will lead to a low value in Precision. Recall is the percentage of true positives in the total actual positive cases. Many underpredictions will lead to a low value of Recall. Ideally, both precision and recall are 1.0. A better trigger function will produce more correct determinations with less overprediction and underprediction. The F1 score is a more robust performance metric than either P or R, which is the harmonic mean of the precision and recall, defined as:

$$\text{F1} = \frac{2PR}{P+R} \qquad (6)$$

The F1 score has a maximum value of one and will achieve a good performance only when both Precision and Recall are high. It is a widely used measurement to estimate the performance of ML methods (Alioto et al., 2015; Huang et al., 2018; Luo et al., 2019; Zhang et al, 2019).

To understand how the ML model performs overall across different categories (denoted by a and b below), we use the Macro-manner to compute Precision, Recall and F1 score, which computes these criteria independently for each category and then takes the average (Yang, 1999):

$$\begin{cases} \text{P}_a = \frac{TP}{TP+FP}\ ,\ \text{P}_b = \frac{TN}{TN+FN} \\ \text{P}_{macro} = \frac{P_a+P_b}{2} \end{cases} \qquad (7)$$

$$\begin{cases} \mathrm{R}_a = \frac{TP}{TP+FN}\ ,\ \mathrm{R}_b = \frac{TN}{TN+FP} \\ \mathrm{R}_{macro} = \frac{R_a+R_b}{2} \end{cases} \quad (8)$$

$$\begin{cases} \mathrm{F1}_a = \frac{2P_aR_a}{P_a+R_a}\ ,\ \mathrm{F1}_b = \frac{2P_bR_b}{P_b+R_b} \\ \mathrm{F1}_{macro} = \frac{F1_a+F1_b}{2} \end{cases} \quad (9)$$

4 Results

4.1 Performance of the trigger functions

The performance of each of the tested trigger functions is evaluated according to the contingency table and its derived performance criteria, including Precision, Recall and the F1 score. Table 2 lists the number of the samples used at the SGP and MAO sites. 80% of the total samples, randomly drawn, are used for training, and the remaining 20% for testing. The 'Positive' and 'Negative' columns represent that convection is triggered and not triggered. The XGB trigger functions are trained by the 'Train' samples and is evaluated by the 'Test' samples. To make the testing of the traditional CAPE-based trigger functions comparable to the ML trigger, they are also evaluated by the 'Test' samples.

Table 3 shows the four probable outcomes of the contingency table at the SGP and MAO sites for the CAPE-based and the XGB triggers. In this comparison, the XGB triggers are separately trained for the two sites. The XGB triggers achieve the best performance among the five, with smaller FP and FN, and larger TP and TN at both SGP and MAO. In contrast, the undilute CAPE trigger yields very large FP values, indicating heavy overprediction of convective cases. Dilute CAPE also tends to overpredict convective events, though far less severe (smaller FP) compared to the undilute CAPE trigger. The direct consequence is that the deep convection parameterization with these triggers would be activated too frequently. The FP values with dilute and undilute dCAPE triggers are much lower. This is not unexpected and is confirmed by previous works (Xie and Zhang, 2000; Xie et. al. 2004a; Suhas and Zhang, 2014). For the XGB triggers, FN is slightly greater than FP, which indicates that the number of overpredictions is lower than the number of underpredictions. In other words, it has a slight tendency to classify a case as non-convective. The presence of this

tendency is because the number of non-convective cases is far larger than the convective cases in the training set, as ML models tend to concentrate on predicting the major categories accurately while ignoring the minor ones (Estabrooks et al., 2004; Ramyachitra et al., 2014).

Figure 2 presents the performance (F1 score, Precision and Recall) of all trigger functions based on the results shown in Table 3. In terms of the F1 score, the XGB triggers perform the best with 91% and 93% accuracy at SGP and MAO, respectively. In comparison, the best performance for the CAPE-based trigger function, dilute dCAPE, is only 79% and 85% at the two sites. Since the pure CAPE (dilute or undilute) trigger functions activate the convection too frequently, they perform the worst. Such overprediction biases are largely reduced with both dCAPE triggers, leading to much improved Precision and F1 scores. The improvement of the XGB triggers relative to dilute dCAPE seems smaller at MAO compared to SGP. The time resolution could be a non-negligible factor for the ML performance. The time interval at MAO is 3-hour, while it is 1-hour at SGP. We test the SGP performance at a 3-hour interval and find the F1-score is reduced to 0.84, which is substantially lower than the F1-score of 0.91 with 1-hour interval data.

Although the ML trigger function attains the best performance, as alluded to above when sampling the data at longer time interval, the outcome may depend on the size of the training samples. This can be seen in Figure 3 that shows the training and testing performance as a function of training dataset size (also called the learning curve) for the XGBoost model. Here the F1 performance score of the XGB triggers are determined using the k-fold cross-validation method (Kohavi, 1995). This method is used to ensure that the performance evaluation has robust credibility, because the performance of ML algorithms usually also heavily depend on the choice of training set. The cross-validation method first divides the whole dataset into k subsets. Then each subset is in turn used for testing, with the remainder for training. After k iterations, the final performance reported by the k-fold cross-validation is the average of the F1 score of each iteration (here k is 5). Each point in the learning curve represents a model that is trained by a subset of the above training data, with its size given by the x-axis. The corresponding cross-validation score is always determined using the full testing set. As the figure shows, the testing performance increases as the size of the training dataset increases at SGP, suggesting that even more improvement is possible if a larger dataset were used to build the XGBoost trigger. In contrast, the trigger at MAO does not benefit as much

from more data. Note that the number of available cases at SGP is almost four-fold greater than at MAO, the difference in the learning curves implies a greater diversity of convective environments that must be mastered at SGP.

4.2 Time series and diurnal cycle

The performance of different convective triggers can be more clearly seen in their skills in predicting individual convective events. Figures 4 and 5 present the time series of convection occurrence predicted by all trigger functions at SGP and MAO, respectively. Here the XGB triggers are re-trained with the first three quarters of the entire VARANAL dataset for the purpose of displaying continuous time series and diurnal cycle for the last quarter. The training and testing data are therefore sequentially divided, while the previous training and testing data for Figure 2 and tabulated in Table 2 are randomly drawn. The last quarter of the dataset is used to evaluate the performance and presented in these two figures. The evaluation portion of the time series are the summer seasons from 2006/06/16 12:00 Coordinated Universal Time (UTC) to 2008/08/31 23:00 at SGP and from 2015/07/02 12:00 to 2015/12/31 21:00 at MAO. Precipitation is also shown for reference and is used to directly check whether deep convection is predicted when the precipitation rate is over 0.5mm/h.

For SGP, the convection events predicted by the XGB trigger is very close to the observations. The dilute dCAPE trigger also has a frequency of convection similar to that from observations and the XGB trigger, but it does not agree on timing. The other three CAPE-based trigger functions all activate convection too frequently, especially the undilute dCAPE and undilute CAPE. For MAO compared with observations, the XGB trigger and the dilute dCAPE show similarly good skill. The dilute CAPE activates deep convection surprisingly less frequently. Less frequent convective triggering however does not mean performing better. By inspecting the corresponding contingency table (not shown), it turns out that while total convective events (TP + FP) are underpredicted, false positive events (FP) are substantially overestimated. The resulting performance is still much poorer compared to using the dCAPE trigger, consistent with the findings in Song and Zhang (2017). The undilute CAPE and the undilute dCAPE give more overpredictions than their dilute

counterparts, especially the undilute CAPE trigger that dramatically overpredicts the convective events.

The diurnal cycle of precipitation is an important benchmark to measure the amount, frequency, intensity, and duration of precipitation. However, current climate and weather models have difficulties in capturing the diurnal variation of precipitation (Lee et al 2007, Covey et al. 2016). The role of the convective trigger function is particularly exemplified in Xie et al. (2019), which broadly improved the diurnal cycle of precipitation around the globe in E3SM using the dilute dCAPE trigger along with an unrestricted air parcel launch level (ULL) method. Figure 6 shows the diurnal frequency of convection predicted by XGBoost, dilute dCAPE, undilute dCAPE, and dilute CAPE. The undilute CAPE is not shown because it substantially overpredicts convective occurrences (i.e., Figures 4 and 5) and, if shown, the y-axis would be stretched, obscuring the differences to be discussed. At the SGP (Figure 6a), the dilute CAPE trigger (red) clearly would predict a very different diurnal occurrence frequency than observed, with a very strong and broad peak spanning early afternoon to early evening. The shape of diurnal occurrence predicted by the dilute dCAPE trigger (green) aligns reasonably well with observations, although its frequency is lower throughout most of the day. The performance of the undilute dCAPE (purple) is worse than dilute dCAPE (green), as it still activates convection too frequently throughout the day. This is also reflected in Figure 4. The XGB trigger mostly outperforms the dCAPE trigger in terms of diurnally varying counts of the convective occurrence. Similarly at MAO (Figure 6b), the better performance of the dCAPE and the XGB trigger in predicting diurnal convection is even more clear compared to the dilute CAPE trigger. Undilute dCAPE (purple) generally outperforms other triggers in the early morning (< 8 LST). Although it triggers convection more frequently than observations from the afternoon to nighttime, its positive bias from observations is similar in magnitude to the negative bias of the XGB trigger (orange) which both performs better than dilute dCAPE (green). We note that the XGB triggers underpredict the observed precipitation frequency at the two sites, which could be caused by the imbalance of the training dataset in which the number of non-convection cases is much larger than that of the convection cases. This is relevant because, as noted previously, ML models tend to concentrate on predicting the major category, which in this case is the non-convection events. We have tried restricting the training to use an equal number

of cases for the two categories by under-sampling the non-convective events, but the improvement was not significant because the total number of training dataset becomes smaller.

4.3 Interpretation of the XGB triggers

Though the XGB triggesr achieve better performance than the traditional CAPE-based schemes, it is a black-box model from which it is difficult to extract explicit knowledge due to its complicated structure consisting of hundreds of weak ML models, and the high dimensionality of the problem. For the sake of understanding how to improve convection parameterization schemes in use, simply replacing them with a machine-learning-based black-box model is less than satisfactory as explicit knowledge about the model is critically useful. The XGBoost method can provide the relative importance index of each predictor in its training process, which can help quantify the contribution of each predictor to the determination of convection. The decision tree is the foundation of the XGBoost method. In the tree structures, each non-leaf node selects one predictor and determines an optimal threshold, which partitions the dataset into two subsets. This process is crucial for building the tree. The principle is to make the subsets have a high purity (i.e., being composed of samples having the same category). Breiman (2001) proposed the mean decrease impurity importance (MDI) to measure the relative importance of each predictor by averaging the weighted impurity decrease with regard to this predictor over all nodes in one tree and then weighted averaging over all trees in the ensemble-based algorithm. Here the weight is defined as the fraction of subset under a node, and the impurity measure can be the variance, the Shannon entropy (Shannon, 1948), or the Gini index (Breiman, 2001).

Note that predictor indices sum to 1.0. Figure 7 displays the top 10 most important predictors when the XGB triggers are trained at SGP and MAO. Dilute dCAPE is the most important predictor for both SGP and MAO. Among the total of 21 predictors shown in Table 1, the relative importance index of the dilute dCAPE dominates at over 40%. The other important predictors standing out at SGP are the surface relative humidity and latent heat flux, while at MAO they are latent heat flux and low level temperature. Wind shear does not appear in the top 10 for either site.

Entrainment rate exerts a strong control on the degree of dilution and the magnitude of dCAPE, thereby affecting the occurrence and intensity of convection. Figure 8 shows the consistent

relationship between the relative importance index of the dilute dCAPE predictor from the XGB triggers and the XGB trigger performance as a function of entrainment rate. As can be seen, both are sensitive to the entrainment rate. The XGB triggers at SGP and MAO achieve the best performance when the entrainment rate is slightly lower than 1.0×10^{-3} m^{-1}, where the importance index of the dilute dCAPE is also the highest. This result is consistent with that of Song et al. (2017). (Please also see the comment in the response, may want to limit the left most point. If too close to 0, it is effectively undilute dCAPE

Accordingly, we can infer explicit knowledge about convection occurrence based on these key predictors using the decision tree that has easy interpretability. Figure 9 illustrates the decision process at SGP and MAO. Two rules can be derived from the decision tree. The first one is acquired by the root nodes, which classifies the dataset into two branches through thresholds of dilute dCAPE. The decision trees automatically suggest the optimal thresholds of 62 J/kg/h and 37 J/kg/h, respectively, at SGP and MAO. These thresholds are attained by maximizing the purity of each branch under the root node in terms of non-convective or convective cases. Cases are classified as non-convective when the dilute dCAPE values are less than or equal to the thresholds, while they are classified as convective when the values are greater than the thresholds. The performance under these thresholds, shown in Figure 10, is not worse than that using 65 J/kg/h as the threshold, which is a well-tuned value used by Suhuas and Zhang (2014; hereafter 'SZ threshold'), and is better than using 0 J/kg/h as the threshold as well (e.g., Xie et. al., 2019). Because a higher dilute dCAPE value implies a higher possibility of convection event, a higher threshold would lead to a better Precision score by limiting FP (Eq. 4) in terms of the convection cases. Although the Precision score based on the threshold from the decision tree at MAO is slightly lower than when using the SZ threshold, due to the smaller threshold value and fewer training data compared to what are used at SGP, the Recall score is better (Eq. 5) and helps achieve a good balance between Precision and Recall.

The other rule starts from the root node until reaching the leaf nodes. The rules in each path are connected by the 'AND' logic. Each leaf node represents a different combination of rules. We select the leaf nodes marked in the dashed boxes in Figure 9 to generate a reduced number of simple "Machine Learning Explicit Rules (MLER)" for classification of the whole dataset. These

leaf nodes were selected because they can correctly separate convective/non-convective events to the greatest extent and cover most of the dataset. In other words, these leaf nodes have the maximum purity. The rules at SGP are:

$$\text{NonConvective} \begin{cases} \text{dilute dCAPE} \leq 26\ \text{J/kg/h} \\ \text{RHsair} \leq 87\ \% \end{cases} \quad (10)$$

$$\text{Convective: dilute dCAPE} > 168\ \text{J/kg/h} \quad (11)$$

And the rules at MAO are:

$$\text{NonConvective} \begin{cases} \text{dilute dCAPE} \leq 23\ \text{J/kg/h} \\ \text{lhflx} \leq 114\ \text{W}/m^2 \end{cases} \quad (12)$$

$$\text{Convective: dilute dCAPE} > 66\ \text{J/kg/h} \quad (13)$$

The MLER trigger scheme not only contains the stricter dilute dCAPE threshold for non-convection and convection, but also incorporates thresholds of other important factors of XGB triggers, including surface air relative humidity and latent heat flux. Note that the rules in Eqs (10) - (13) are not closure conditions so they are not able to cover the entire dataset. As shown in the rightmost column of Figure 10, the MLER trigger covers 77% of SGP dataset and 62% of MAO dataset and achieves 89% and 99% F1 scores. It should be noted that the dilute dCAPE threshold rules cover 100% of the SGP and MAO dataset. If we evaluate the dCAPE threshold rules by the samples screened by the MLER rules, it would be an unfair comparison because the MLER rules have been appended to the dilute dCAPE rule. As seen in Figure 10, the SGP has a great Precision score (low over-prediction). This is because this MLER scheme imposes a stricter dilute dCAPE criterion on convection/non-convection events than both the root threshold of decision tree and the SZ threshold. Similarly, at MAO the criterion of dilute dCAPE in MLER to determine convection events is slightly larger than the SZ threshold, leading to a slight increment in terms of Precision score. On the other hand, due to the lower dilute dCAPE threshold and the additional latent heat flux constraint, the number of overpredictions is very low. According to Eq. 5, the Recall score has been improved significantly.

4.4 A unified machine learning trigger function

Training the XGB triggers at each grid box in a GCM will be extremely computationally intensive and, even if accomplished, the result would be undesirable for use in the model that requires basic

generalized parameterizations for simulation of variable climates. Therefore, building a unified scheme like the traditional parameterization scheme is more pragmatic for real use. Given data from the two sites in this study, one way to build a unified scheme for both sites is to verify whether the ML XGB model trained by one site also works for the other site. Another method is to train a unified scheme by joining dataset of both sites. Figure 11 shows the performance of the XGB triggers with these two approaches, along with the standalone XGB triggers for SGP and MAO for reference. The unified trigger built on the joint dataset has a 91% F1 score when evaluated by the combined testing dataset of SGP and MAO. Even when tested separately for each site, the unified trigger achieves, respectively, F1 scores of 91% and 92%. Thus, the unified trigger function performs as well as when the training is performed separately at each site. This result has an important implication that a suitably developed uniform ML trigger function, after accounting for all major convective regimes, may be feasible for all grids in a GCM.

To demonstrate the potential, we further apply the unified XGB triggers to the independent ARM Intensive Observation Period (IOP) data, including the SGP data in the summer of 1997 (SGP97) and the Tropical Warm Pool-International Cloud Experiment (TWP-ICE) in 2006 (TWP06), which are the testing data for the unified XGB triggers. Figure 12 compares this method and the traditional rule of dilute dCAPE. The unified trigger achieves, respectively, F1 scores of 80% and 86% at SGP97 and TWP06, which are far greater than the dilute dCAPE trigger with the threshold of 0 J/kg/hour. The results further prove that the joint XGB triggers are robust, and after being trained with more data that account for various cloud regimes, can potentially be applied globally.

We note that the performance is not very good when the trigger is trained at one site and tested at another site, though the performance is better when the trigger is trained at MAO then tested at SGP than the other way around. The relatively better performance with the MAO-trained trigger suggests that the trigger trained at MAO has a stronger generalization capability than that at SGP. The tropical atmosphere at MAO is likely ‘out of bounds’ relative to the conditions of the SGP training; therefore, the model trained at SGP cannot extrapolate to MAO. The rules in Eqs (10) – (13) can confirm this finding. The rules at SGP indicate that dilute dCAPE has a larger threshold than that of MAO when convection happens. Consequently, it would fail to predict many convection events at MAO when applying the dilute dCAPE trigger trained with the SGP dataset.

This illustrates the importance of using joining dataset from multiple sites that represents various convective regimes to train ML model for obtaining a unified trigger suitable for use in global climate models.

5 Discussion and Conclusion

In this study, we implemented a novel deep convection trigger function using the XGBoost method, which is a state-of-the-art ML classification model. We first develop the trigger function separately for the SGP and MAO sites based on the long-term VARANAL forcing data from the ARM program. The XGB triggers achieve 91% and 93% F1 scores at the SGP and MAO sites, respectively. Compared with the commonly used CAPE-based trigger functions, the new trigger function offers a substantial improvement. Among the latter, the best trigger function is dilute dCAPE that achieves 79% and 85% F1 scores at the SGP and MAO, respectively. Further investigation indicates that the CAPE-based trigger functions, especially the undilute CAPE and undilute dCAPE triggers, activate convection too frequently. The ML trigger functions alleviate the overprediction of convection occurrence and further demonstrate much better skill in capturing the diurnal cycle of convection.

To obtain explicit knowledge from the black-box ML trigger functions, a series of augmented rules are derived using a decision tree, which is built on the principal predictors identified by the XGB triggers. The rules drawn by the root node in the decision tree demonstrate better performance than traditional triggers using their default thresholds and are comparable to the triggers using well-tuned thresholds. The rules of the selected leaf nodes are stricter in determining convective or non-convective cases and improve the precision. However, the stricter roles are not closure conditions so some of the datasets are out of coverage. Therefore, in future work, supplemental rules would also be required to cover the rest of dataset.

This study demonstrates that a unified ML trigger function may potentially be developed for use in GCMs by jointly training on two sites that have distinctively different convective conditions, using data from SGP and MAO as an illustration. The jointly trained ML trigger function performs as well as those that were separately trained for the individual sites, while the trigger function

trained using data at one site is clearly not well suited for another site if data from the latter are not included in the training process. This result suggests that a functional ML convective trigger for GCMs need to be trained with data accounting for major representative convective regimes around the globe.

The insights obtained from examining the ML model also implies the formation of convection at different regions probably possesses distinct mechanisms or at least involves some different processes; hence it is necessary to develop a more comprehensive unified trigger scheme to better describe global convective process. The current study provides an exploration on this aspect by using the combined data from SGP and MAO sites. It would be interesting to include more data, such as oceanic VARANAL data from AMIE-Dynamo and TWP-ICE. Finally, it should be noted that training the ML model is sensitive to the dissimilar number of convection/non-convection events, where the training tends to key in on the conditions for the greater number of non-convective events. We have tried restricting the training to use an equal number of cases for the two categories by under-sampling the non-convective events, but the improvement was not significant because of the smaller size of the training dataset. In the future, it might be possible to achieve a better development when more data are available.

Data availability Statement

The long-term constrained variational analysis forcing data at SGP and MAO sites from the Atmospheric Radiation Measurement (ARM) program can be found at https://www.arm.gov/data. The values of the four variant CAPEs, dilute CAPE, dilute dCAPE, undilute CAPE, and undilute dCAPE, at SGP and MAO sites can be found at http://doi.org/10.5281/zenodo.4086008.

Author contributions

TZ and WL developed the ML XGBoost trigger function. All authors contributed to interpreting the results and writing the paper.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

This work was primarily supported by the Climate Model Development and Validation (CMDV) project and partially supported by the Energy Exascale Earth System Model (E3SM) project, funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research. Research activity at BNL was under Brookhaven National Laboratory contract DE-SC0012704 (T.Z., W.L., and A.M.V.). The work at LLNL was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., Heisler, L. E., Beck, T. A., Simpson, J. T., Tonon, L., et al.: A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing, Nature communications, 6, 10 001, 2015.

Arakawa, A.: The cumulus parameterization problem: Past, present, and future, Journal of Climate, 17, 2493–2525, 2004.

Arakawa, A. and Schubert, W. H.: Interaction of a cumulus cloud ensemble with the large-scale environment, Part I, Journal of the Atmospheric Sciences, 31, 674–701, 1974.

Bechtold, P., Bazile, E., Guichard, F., Mascart, P., and Richard, E.: A mass-flux convection scheme for regional and global models, Quarterly Journal of the Royal Meteorological Society, 127, 869–886, 2001.

Breiman, L.: Random Forests. Machine Learning 45, 5–32, 2001.

Brenowitz, N. D. and Bretherton, C. S.: Prognostic validation of a neural network unified physics parameterization, Geophysical Research Letters, 45, 6289–6298, 2018.

Brenowitz, N. D. and Bretherton, C. S.: Spatially Extended Tests of a Neural Network Parametrization Trained by Coarse-Graining, Journal of Advances in Modeling Earth Systems, 11, 2728–2744, 2019.

Chen, Q., Fan, J., Hagos, S., Gustafson Jr, W. I., and Berg, L. K.: Roles of wind shear at different vertical levels: Cloud system organization60and properties, Journal of Geophysical Research: Atmospheres, 120, 6551–6574, 2015.

Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794, ACM, 2016.

Covey, C., Gleckler, P. J., Doutriaux, C., Williams, D. N., Dai, A., Fasullo, J., Trenberth, K., and Berg, A.: Metrics for the diurnal cycle of precipitation: Toward routine benchmarks for climate models, Journal of Climate, 29, 4461–4471, 2016.

Dai, A.: Precipitation characteristics in eighteen coupled climate models, Journal of Climate, 19, 4605–4630, 2006.

Dai, A. and Trenberth, K. E.: The diurnal cycle and its depiction in the Community Climate System Model, Journal of Climate, 17, 930–951, 2004.

Dai, A., Trenberth, K. E., and Karl, T. R.: Effects of clouds, soil moisture, precipitation, and water vapor on diurnal temperature range, Journal of Climate, 12, 2451–2473, 1999.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D., DuVivier, A., Edwards, J., Emmons, L., Fasullo, J., Garcia, R., Gettelman, A., et al.: The community earth system model version 2 (CESM2), Journal of Advances in Modeling Earth Systems, 12, e2019MS001 916, 2020.

Donner, L. J.: A cumulus parameterization including mass fluxes, vertical momentum dynamics, and mesoscale effects, Journal of the Atmospheric Sciences, 50, 889–906, 1993.

Emanuel, K. A.: Atmospheric convection, Oxford University Press on Demand, 1994.

Estabrooks, A., Jo, T., and Japkowicz, N.: A multiple resampling method for learning from imbalanced data sets, Computational intelligence, 20, 18–36, 2004.

Fritsch, J. and Chappell, C.: Numerical prediction of convectively driven mesoscale pressure systems. Part I: Convective parameterization, Journal of the Atmospheric Sciences, 37, 1722–1733, 1980.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., and Yacalis, G.: Could machine learning break the convection parameterization deadlock?, Geophysical Research Letters, 45, 5742–5751, 2018.

Gettelman, A., Hannay, C., Bacmeister, J. T., Neale, R. B., Pendergrass, A. G., Danabasoglu, G., Lamarque, J.-F., Fasullo, J. T., Bailey, D. A., Lawrence, D. M., and Mills, M. J.: High Climate Sensitivity in the Community Earth System Model Version 2 (CESM2), Geophysical Research Letters, 46, 8329–8337, https://doi.org/10.1029/2019GL083978, 2019.

Golaz, J.-C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe, J. D., Abeshu, G., Anantharaj, V., Asay-Davis, X. S., Bader, D. C., et al.: The DOE E3SM coupled model version 1: Overview and evaluation at standard resolution, Journal of Advances in Modeling Earth Systems, 11, 2089–2129, 2019.

Gyakum, J.R., and Cai, M., 1990. An Observational Study of Strong Vertical Wind Shear over North America during the 1983/84 Cold Season, *J. Appl. Meteor. Clim.*, https://doi.org/10.1175/1520-0450(1990)029<0902:AOSOSV>2.0.CO;2

Ham, Y.-G., Kim, J.-H., and Luo, J.-J.: Deep learning for multi-year ENSO forecasts, Nature, 573, 568–572, 2019.

Han, Y., Zhang, G. J., Huang, X., and Wang, Y.: A moist physics parameterization based on deep learning, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 076, 2020.

Huang, B., Zhang, K., Lin, Y., Schölkopf, B., and Glymour, C.: Generalized score functions for causal discovery, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1551–1560, 2018.

Kain, J. S.: The Kain–Fritsch convective parameterization: an update, Journal of applied meteorology, 43, 170–181, 2004.

Kain, J. S. and Fritsch, J. M.: Convective parameterization for mesoscale models: The Kain-Fritsch scheme, in: The representation of cumulus convection in numerical models, pp. 165–170, Springer, 1993.

Kuo, H.-L.: On formation and intensification of tropical cyclones through latent heat release by cumulus convection, Journal of the Atmospheric Sciences, 22, 40–63, 1965.

Kuo, H.-L.: Further studies of the parameterization of the influence of cumulus convection on large-scale flow, Journal of the Atmospheric Sciences, 31, 1232–1240, 1974.

Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E., Mahesh, A., Matheson, M., Deslippe, J., Fatica, M., et al.: Exascale deep learning for climate analytics, in: SC18: International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 649–660, IEEE, 2018.

Lee, M.-I., Schubert, S. D., Suarez, M. J., Held, I. M., Lau, N.-C., Ploshay, J. J., Kumar, A., Kim, H.-K., and Schemm, J.-K. E.: An analysis of the warm-season diurnal cycle over the continental United States and northern Mexico in general circulation models, Journal of Hydrometeorology, 8, 344–366, 2007.

Lee, M.-I., Schubert, S. D., Suarez, M. J., Schemm, J.-K. E., Pan, H.-L., Han, J., and Yoo, S.-H.: Role of convection triggers in the simulation of the diurnal cycle of precipitation over the United States Great Plains in a general circulation model, Journal of Geophysical Research: Atmospheres, 113, 2008.

Lin, J.-L., Kiladis, G. N., Mapes, B. E., Weickmann, K. M., Sperber, K. R., Lin, W., Wheeler, M. C., Schubert, S. D., Del Genio, A., Donner, L. J., et al.: Tropical intraseasonal variability in 14 IPCC AR4 climate models. Part I: Convective signals, Journal of climate, 19, 2665–2690, 2006.

Liu, Y., Guo, L., Wu, G., and Wang, Z.: Sensitivity of ITCZ configuration to cumulus convective parameterizations on an aqua planet, Climate dynamics, 34, 223–240, 2010.

Luo, R., Sedlazeck, F. J., Lam, T.-W., and Schatz, M. C.: A multi-task convolutional deep neural network for variant calling in single molecule sequencing, Nature communications, 10, 1–11, 2019.

Martin, S. T., Artaxo, P., Machado, L., Manzi, A. O., Souza, R., Schumacher, C., Wang, J., Biscaro, T., Brito, J., Calheiros, A., et al.: The Green Ocean Amazon experiment (GoAmazon2014/5) observes pollution affecting gases, aerosols, clouds, and rainfall over the rain forest, Bulletin of the American Meteorological Society, 98, 981–997, 2017.

Miao, Q., Pan, B., Wang, H., Hsu, K., and Sorooshian, S.: Improving monsoon precipitation prediction using combined convolutional and long short term memory neural network, Water, 11, 977, 2019.

Neale, R. B., Richter, J. H., and Jochum, M.: The impact of convection on ENSO: From a delayed oscillator to a series of events, Journal of climate, 21, 5904–5924, 2008.

Olson, D. L. and Delen, D.: Advanced data mining techniques, Springer Science & Business Media, 2008.

Quinlan, J. R.: Simplifying decision trees, International journal of man-machine studies, 27, 221–234, 1987.

Ramyachitra, D. and Manikandan, P.: Imbalanced dataset classification and solutions: a review, International Journal of Computing and Business Research (IJCBR), 5, 2014.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, Proceedings of the National Academy of Sciences, 115, 9684–9689, 2018.

Rogers, R. F. and Fritsch, J.: A general framework for convective trigger functions, Monthly weather review, 124, 2438–2452, 1996.

Silva, S. J., Heald, C. L., Ravela, S., Mammarella, I., and Munger, J. W.: A deep learning parameterization for ozone dry deposition velocities, Geophysical Research Letters, 46, 983–989, 2019.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379–423.

Song, F. and Zhang, G. J.: Improving trigger functions for convective parameterization schemes using GOAmazon observations, Journal of Climate, 30, 8711–8726, 2017.

Suhas, E. and Zhang, G. J.: Evaluation of trigger functions for convective parameterization schemes using observations, Journal of Climate, 27, 7647–7666, 2014.

Tang, S., Xie, S., Zhang, Y., Zhang, M., Schumacher, C., Upton, H., Jensen, M. P., Johnson, K. L., Wang, M., Ahlgrimm, M., et al.: Large- scale vertical velocity, diabatic heating and drying profiles associated with seasonal and diurnal variations of convective systems observed in the GoAmazon2014/5 experiment, Atmospheric Chemistry and Physics, 16, 14 249, 2016.

Tao, W.-K., Simpson, J., and Soong, S.-T.: Numerical simulation of a subtropical squall line over the Taiwan Strait, Monthly weather review, 119, 2699–2723, 1991.

Tao, Y., Gao, X., Hsu, K., Sorooshian, S., and Ihler, A.: A deep neural network modeling framework to reduce bias in satellite precipitation125products, Journal of Hydrometeorology, 17, 931–945, 2016.

Tiedtke, M.: A comprehensive mass flux scheme for cumulus parameterization in large-scale models, Monthly Weather Review, 117, 1779– 1800, 1989.

Trenberth, K. E., Dai, A., Rasmussen, R. M., and Parsons, D. B.: The changing character of precipitation, Bulletin of the American Meteorological Society, 84, 1205–1218, 2003.

Van Rijsbergen, C. J.: A non-classical logic for information retrieval, The Computer Journal, 29, 481–485, 1986.

Vanichrujee, U., Horanont, T., Pattara-atikom, W., Theeramunkong, T., and Shinozaki, T.: Taxi Demand Prediction using Ensemble Model Based on RNNs and XGBOOST, in: 2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES), pp. 1–6, IEEE, 2018.

Wang, Y.-C., Xie, S., Tang, S., & Lin, W.: Evaluation of an improved convective triggering function: Observational evidence and SCM tests. Journal of Geophysical Research: Atmospheres, 125, e2019JD031651. https://doi.org/10.1029/e2019JD031651. 2020.

Xie, S. and Zhang, M.: Impact of the convection triggering function on single-column model simulations, Journal of Geophysical Research: Atmospheres, 105, 14 983–14 996, 2000.

Xie, S., Zhang, M., Boyle, J. S., Cederwall, R. T., Potter, G. L., and Lin, W.: Impact of a revised convective triggering mechanism on Community Atmosphere Model, Version 2, simulations: Results from short-range weather forecasts, Journal of Geophysical Research: Atmospheres, 109, 2004a.

Xie, S., Cederwall, R. T., and Zhang, M.: Developing long-term single-column model/cloud system–resolving model forcing data using numerical weather prediction products constrained

by surface and top of the atmosphere observations, Journal of Geophysical Research: Atmospheres, 109, 2004b.

Xie, S., Wang, Y.-C., Lin, W., Ma, H.-Y., Tang, Q., Tang, S., Zheng, X., Golaz, J.-C., Zhang, G. J., and Zhang, M.: Improved diurnal cycle of precipitation in E3SM with a revised convective triggering function, Journal of Advances in Modeling Earth Systems, 11, 2290–2310, 2019.

Yang Y.: An evaluation of statistical approaches to text categorization, Information retrieval, 1(1-2): 69-90, 1999.

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., and Talebiesfandarani, S.: PM2. 5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data, Atmosphere, 10, 373, 2019.

Zhang, G. J. and McFarlane, N. A.: Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian Climate Centre general circulation model, Atmosphere-ocean, 33, 407–446, 1995.

Zhang, M. and Lin, J.: Constrained variational analysis of sounding data based on column-integrated budgets of mass, heat, moisture, and momentum: Approach and application to ARM measurements, Journal of the atmospheric sciences, 54, 1503–1524, 1997.

Zhang, M., Lin, J., Cederwall, R., Yio, J., and Xie, S.: Objective analysis of ARM IOP data: Method and sensitivity, Monthly Weather Review, 129, 295–311, 2001.

Zhang, T., Lin, W., Lin, Y., Zhang, M., Yu, H., Cao, K., and Xue, W.: Prediction of Tropical Cyclone Genesis from Mesoscale Convective Systems Using Machine Learning, Weather and Forecasting, 34, 1035–1049, 2019a.

Zhang, X., Li, T., Wang, J., Li, J., Chen, L., and Liu, C.: Identification of cancer-related long non-coding RNAs using XGBoost with high accuracy, Frontiers in genetics, 10, 735, 2019b.

Zheng, H. and Wu, Y.: A XGBoost Model with Weather Similarity Analysis and Feature Engineering for Short-Term Wind Power Forecasting, Applied Sciences, 9, 3019, 2019.

Zheng, X., Golaz, J.-C., Xie, S., Tang, Q., Lin, W., Zhang, M., Ma, H.-Y., and Roesler, E.: The Summertime Precipitation Bias in E3SM Atmosphere Model Version 1 over the Central United States, Journal of Geophysical Research: Atmospheres, 124, 8935–8952, 2019.

Zhong, J., Sun, Y., Peng, W., Xie, M., Yang, J., and Tang, X.: XGBFEMF: an XGBoost-based framework for essential protein prediction, IEEE transactions on nanobioscience, 17, 243–250, 2018.

Table 1. List of predictors for the machine learning trigger function. Layers used to represent the lower, middle, and upper troposphere are, respectively, 700-800, 300-700, and 200-300 hPa.

Predictors	Abbreviation
Latent heat flux	LHFLX
Sensible heat flux	SHFLX
Air temperature at the surface	Tsair
Air relative humidity at the surface	RHsair
Dilute dynamic CAPE generation rate	ddCAPE
Convective inhibition	CIN
Lifting condensation level	LCL
Temperature in the lower troposphere	T_low
Temperature in the middle troposphere	T_mid
Temperature in the upper troposphere	T_high
Specific humidity in the lower troposphere	q_low
Specific humidity in the middle troposphere	q_mid
Specific humidity in the upper troposphere	q_high
Horizontal advective tendency of water vapor at lower troposphere	q_adv_h_low
Horizontal advective tendency of water vapor in the middle troposphere	q_adv_h_mid
Horizontal advective tendency of water vapor in the upper troposphere	q_adv_h_high
Horizontal advective tendency of dry static energy in the lower troposphere	s_adv_h_low
Horizontal advective tendency of dry static energy in the middle troposphere	s_adv_h_mid
Horizontal advective tendency of dry static energy in the upper troposphere	s_adv_h_high
Wind shear in the lower troposphere	shear_low
Wind shear in the middle troposphere	shear_mid
Wind shear in the upper troposphere	shear_high

Table 2. The sample numbers for SGP and MAO

	SGP			MAO		
	All	Positive	Negative	All	Positive	Negative
All	22800	1991	20809	5840	935	4905
Train	18240	1569	16671	4672	738	3934
Test	4560	422	4138	1168	197	971

Table 3. Contingency table performance of convection trigger functions. For easy comparison, all values are normalized by the total number of the test cases. Perfect performance would be indicated by zeros in the FP and FN columns.

	SGP				MAO			
	TP	FP	FN	TN	TP	FP	FN	TN
XGB	0.07	0.01	0.02	0.90	0.14	0.02	0.03	0.82
dilute_dcape	0.05	0.03	0.04	0.88	0.10	0.00	0.07	0.83
dilute_cape	0.05	0.17	0.04	0.74	0.04	0.09	0.13	0.74
undilute_dcape	0.07	0.13	0.03	0.78	0.14	0.05	0.03	0.78
undilute_cape	0.08	0.66	0.02	0.25	0.17	0.82	0.00	0.01

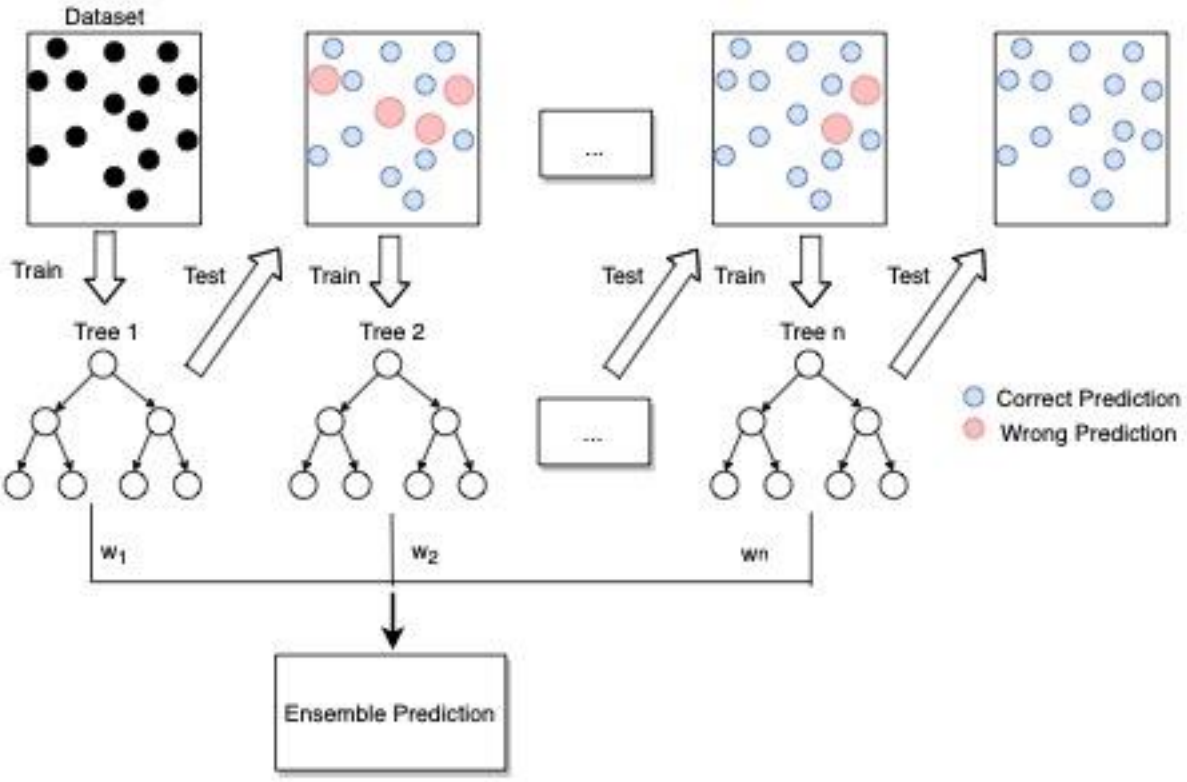

Figure 1. Flow diagram of gradient boosting machine learning method. The ensemble classifiers consist of a set of weak classifiers. The weights of the incorrectly predicted points are increased in the next classifier. The final decision is based on the weighted average of the individual predictions.

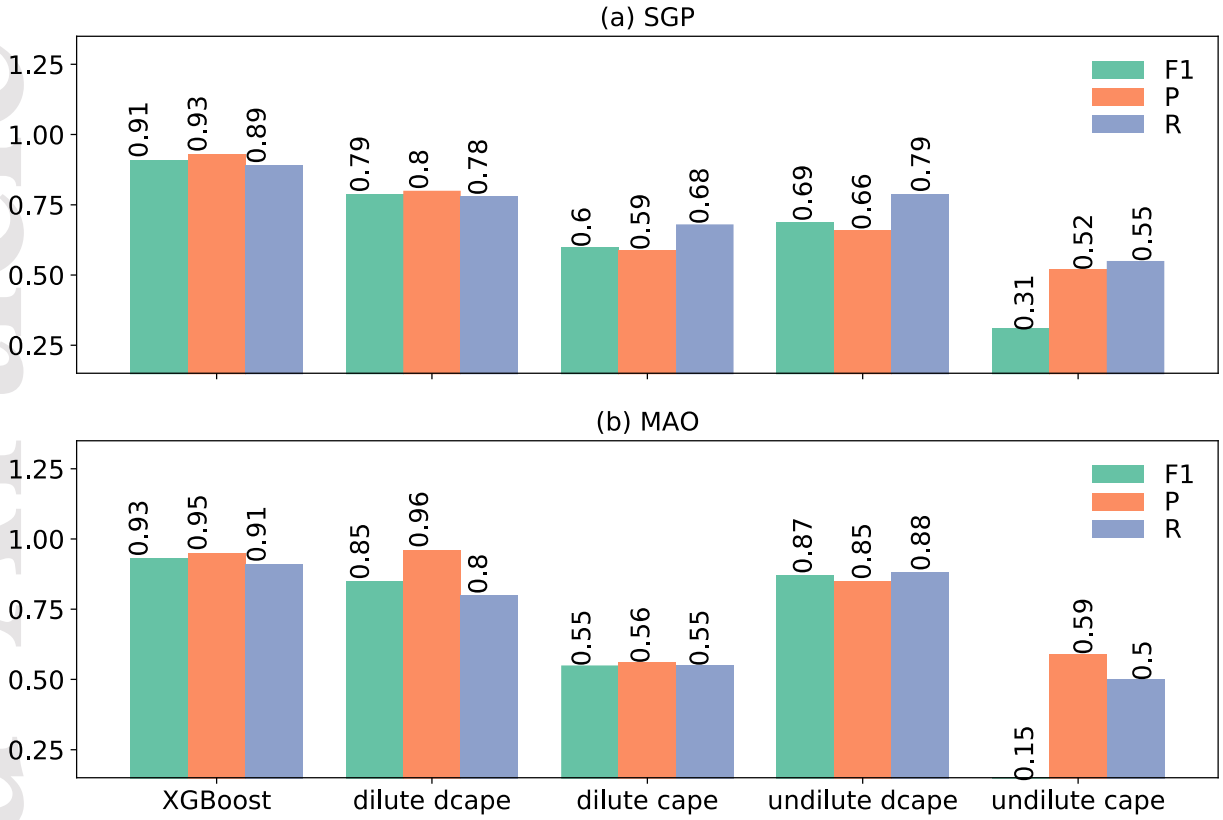

Figure 2. The performance of convection trigger functions at (a) SGP and (b) MAO. Results are shown for the XGB trigger, dilute CAPE, dilute dCAPE, undilute CAPE, and undilute dCAPE. Performance is quantified in terms of the F1 score, Precision (P), and Recall (R).

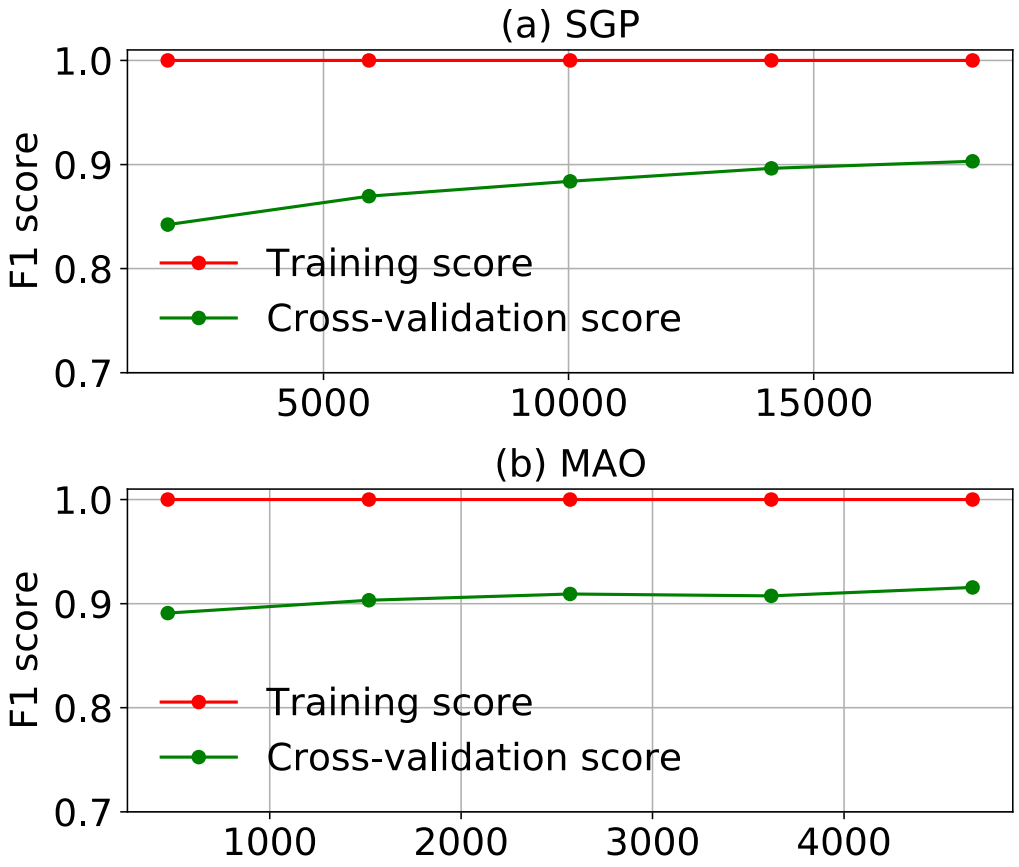

Figure 3. The learning curve of the XGB trigger at (a) SGP and (b) MAO. The learning curve is given as a function of the size of training samples for each site.

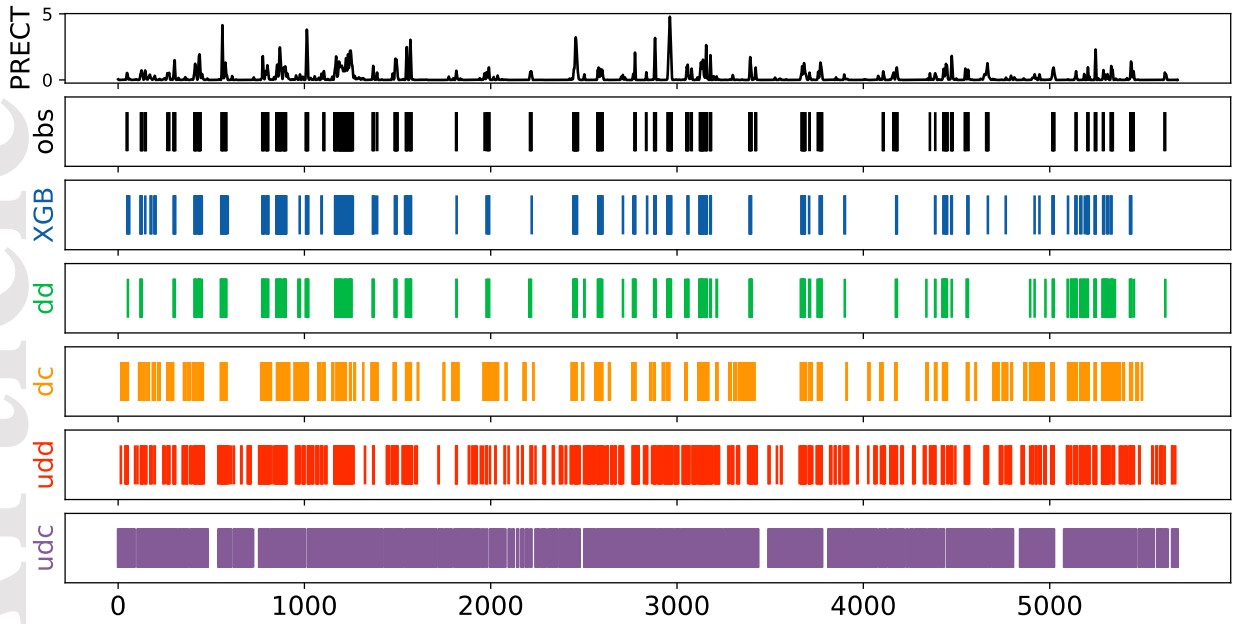

Figure 4. The time series of convection occurrence predicted by different trigger functions for the SGP. Results are shown for the XGB trigger, dilute dCAPE(dd), dilute CAPE (dc), undilute dCAPE (udd), and undilute CAPE (udc) at SGP. The occurrence of convection is indicated by a vertical line. The time series are for the summer seasons from 2006/06/16 12:00 to 2008/08/31 23:00. The XGB trigger is trained by the first three quarters of the time series of the whole dataset at SGP. The PRECT row is the time series of the observed precipitation. The obs row represents whether convection is triggered as determined by the prescribed observational precipitation criterion (0.5 mm/hr).

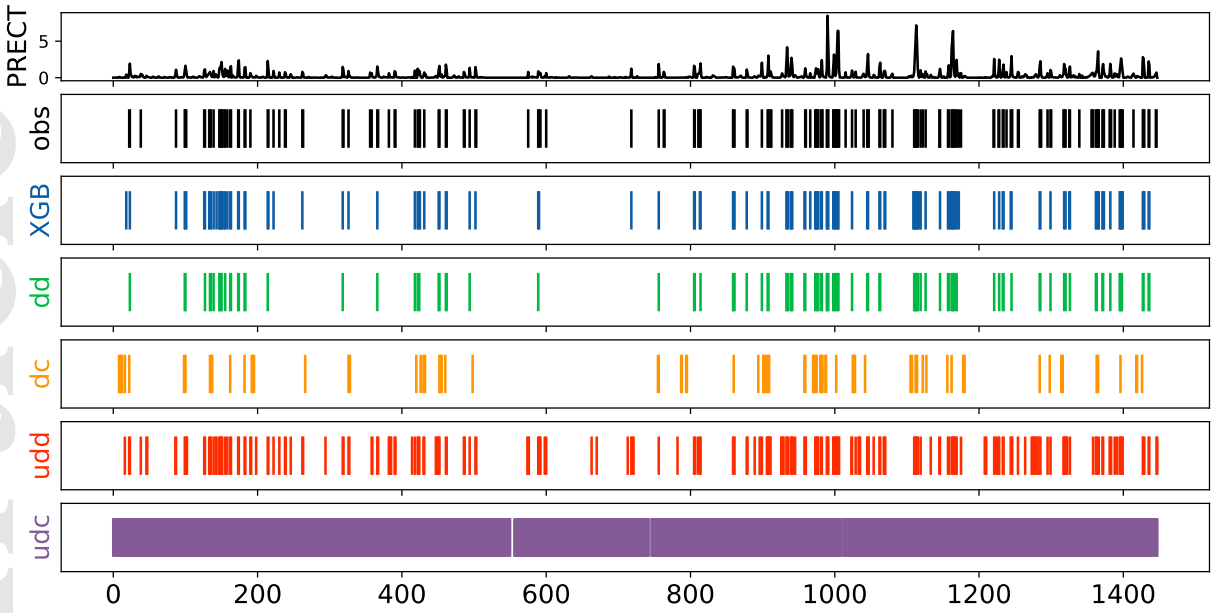

Figure 5. Same as the Fig 5, but for MAO. The time series is from 2015/07/02 12:00 to 2015/12/31 21:00.

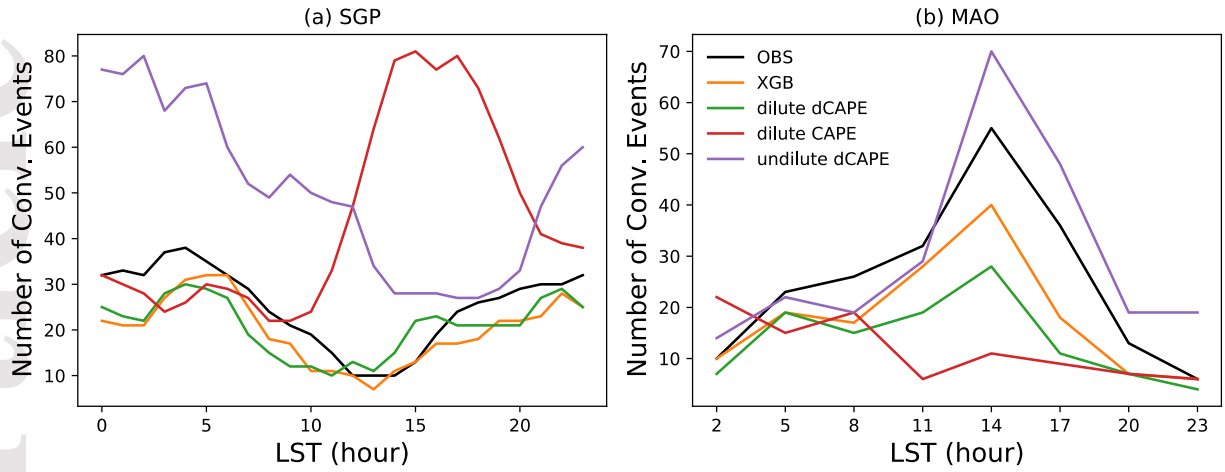

Figure 6. The diurnal cycle of convection occurrence predicted by different trigger functions at (a) SGP and (b) MAO. The training and testing datasets are same with Figure 4 and 5. Results are shown for XGB trigger, dilute dCAPE, dilute CAPE, and undilute dCAPE.

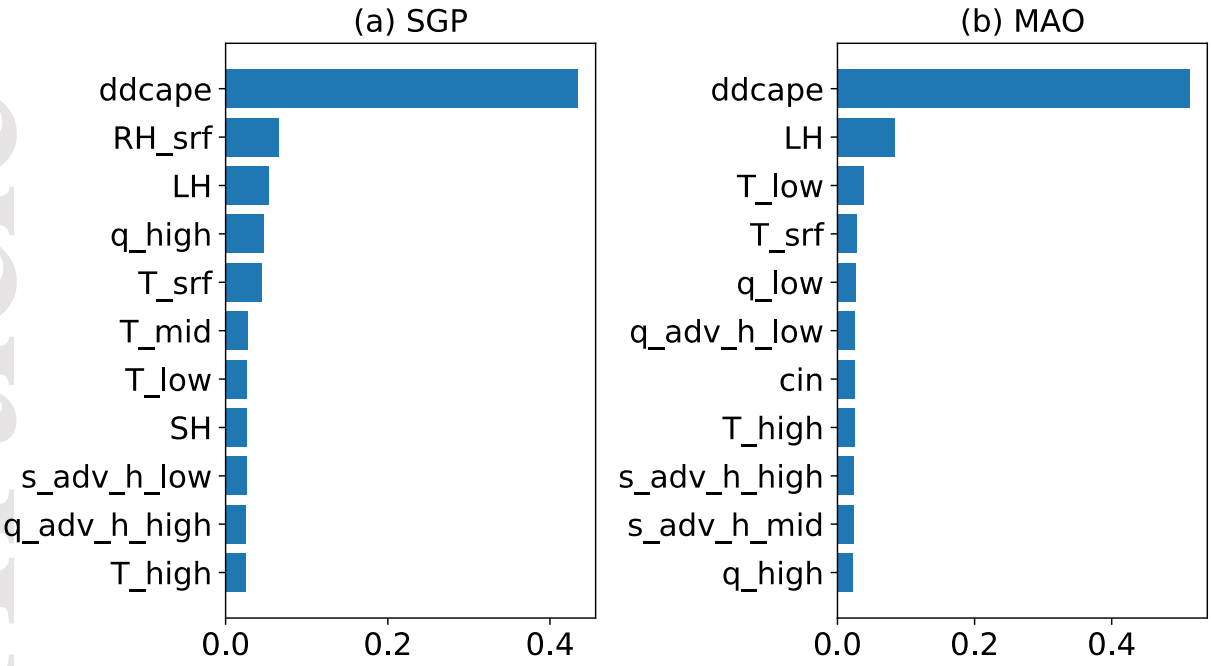

Figure 7. The relative importance indices (x axis) of the top ten individual predictors from the XGB trigger. Results are shown for (a) SGP and (b) MAO.

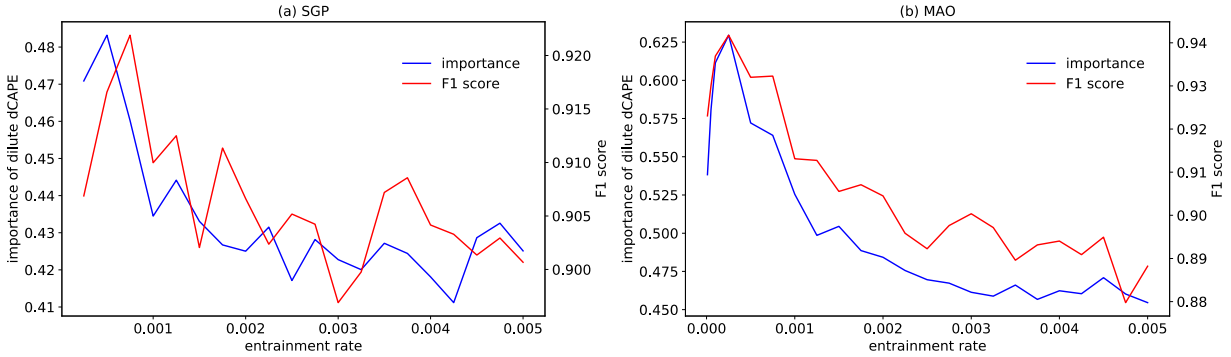

Figure 8. Relationship between the relative importance index (left y-axis) of dilute dCAPE predictor from the XGB triggers and the XGB trigger performance in terms of F1 score (right y-axis) as the function of entrainment rate (m^{-1}). Results are shown for (a) SGP and (b) MAO.

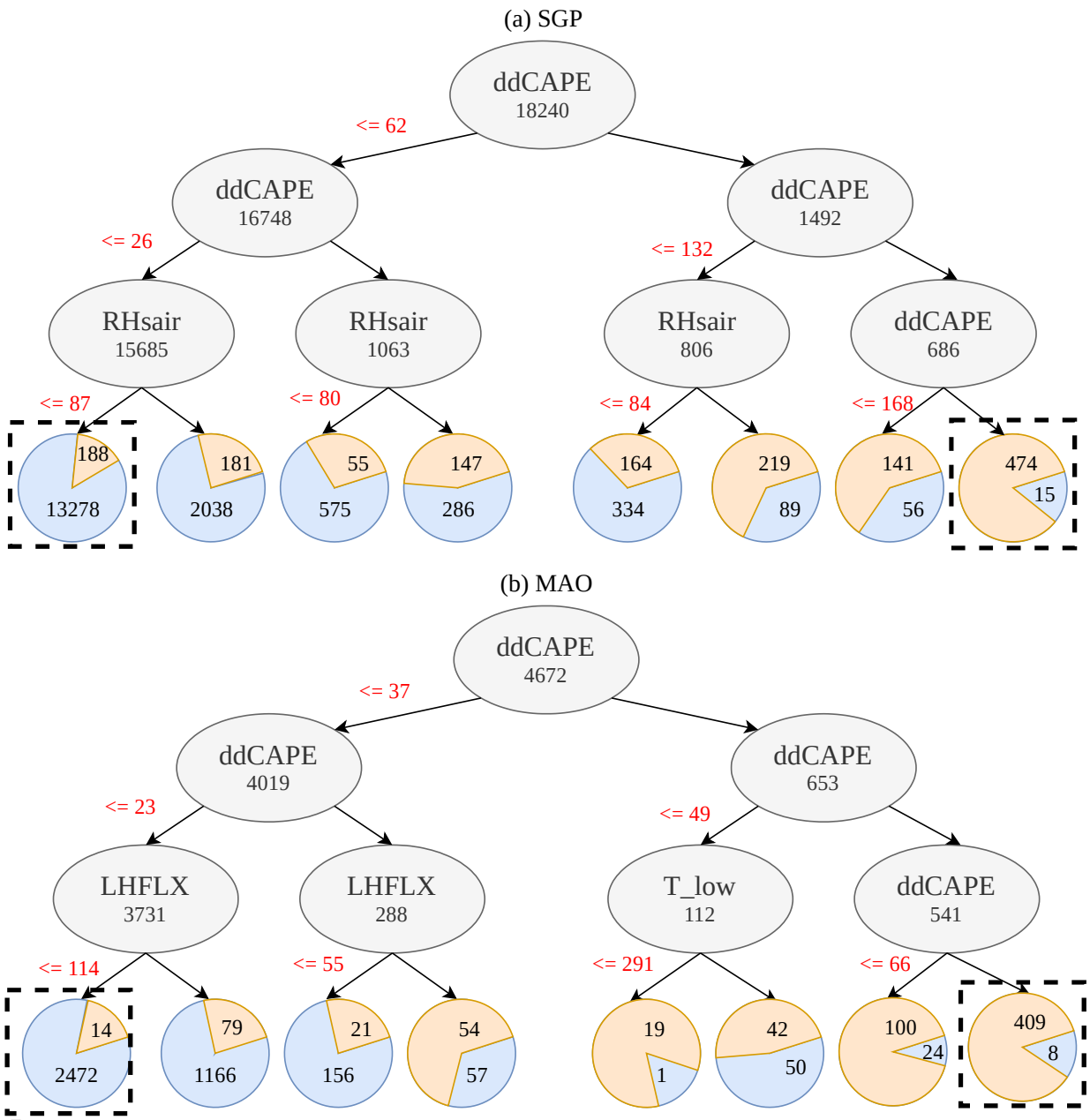

Figure 9. The decision tree built by the top three relative importance predictors identified by the XGB trigger for (a) SGP and (b) MAO. In the tree structures, each non-leaf node divides the dataset into two subsets by a predictor and its threshold. The numbers in the non-leaf nodes (gray ovals) represent the total number of events in the dataset of the current nodes. The blue color in the leaf nodes represent cases that are determined as being non-convective, while the brown color is convective.

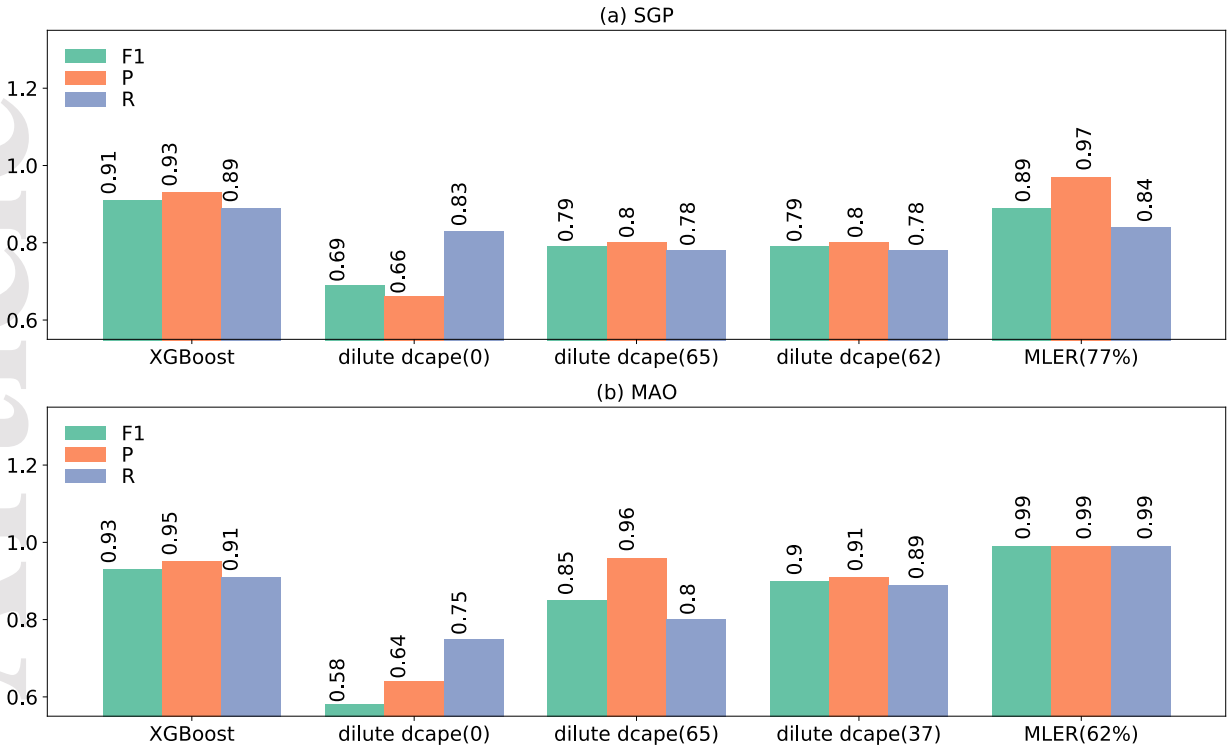

Figure 10. The performance of convection trigger functions for (a) SGP and (b) MAO. Results are given for the XGB trigger, dilute dCAPE for different thresholds (given in parentheses in J/kg/h), and following the 'MLER' rules given by Eqs. (10) – (13). The percentages given for the MLER results indicates the fraction of the cases that were in bounds for which a solution could be produced. Performance is quantified in terms of the F1 score, Precision (P), and Recall (R).

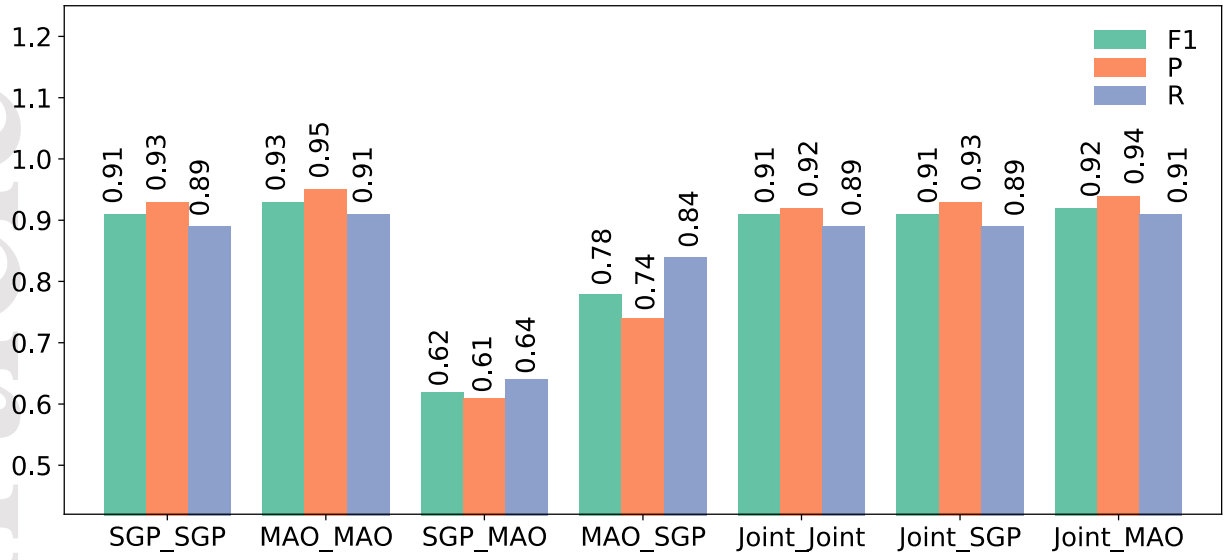

Figure 11. The performance of the XGB unified machine learning trigger function. The first two groups present the machine learning models trained and tested at the same site (as in Figure 2). The middle two groups present the machine learning models trained at one site and applied to the other site. The last three groups present the machine learning model jointly trained for the two sites and are tested jointly, as well as tested separately at each site.

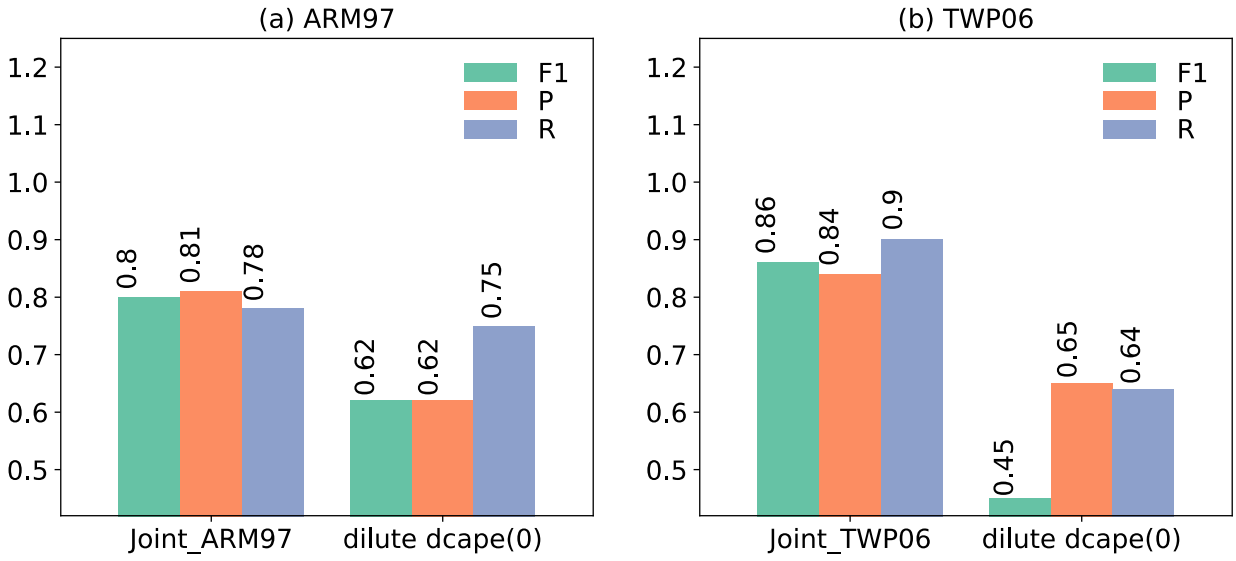

Figure 12. The performance of the XGB unified trigger function applied to SGP97 and TWP06 data.