Characterizing Uncertainty in Deep Convection Triggering Using Explainable Machine Learning

GRETA A. MILLER,[a] PHILIP STIER,[a] AND HANNAH M. CHRISTENSEN[a]

[a] *Department of Physics, University of Oxford, Oxford, United Kingdom*

(Manuscript received 3 May 2024, in final form 2 December 2024, accepted 13 February 2025)

ABSTRACT: Realistically representing deep atmospheric convection is important for accurate numerical weather and climate simulations. However, parameterizing where and when deep convection occurs ("triggering") is a well-known source of model uncertainty. Most triggers parameterize convection deterministically, without considering the uncertainty in the convective state as a stochastic process. In this study, we develop a machine learning model, a random forest, that predicts the probability of deep convection, and then apply clustering of Shapley additive explanations (SHAP) values, an explainable machine learning method, to characterize the uncertainty of convective events. The model uses observed large-scale atmospheric variables from the Atmospheric Radiation Measurement constrained variational analysis dataset over the Southern Great Plains, United States. The analysis of feature importance shows which mechanisms driving convection are most important, with large-scale vertical velocity providing the highest predictive power for more certain, or easier to predict, convective events, followed by the dynamic generation rate of dilute convective available potential energy. Predictions of uncertain, or harder to predict, convective events instead rely more on other features such as precipitable water or low-level temperature. The model outperforms conventional convective triggers. This suggests that probabilistic machine learning models can be used as stochastic parameterizations to improve the occurrence of convection in weather and climate models in the future.

SIGNIFICANCE STATEMENT: Convective storms, which produce clouds and precipitation, are difficult to represent in models since they occur at scales smaller than a model grid box. The purpose of this study is to better understand why convection is sometimes easier or harder to predict with certainty. This is important because predicting where and when convection occurs in atmospheric models affects the energy, moisture, and momentum processes in these models, which is known to lead to errors in weather forecasts and climate projections. This work highlights the importance of representing uncertainty in processes like convection.

KEYWORDS: Deep convection; Convective parameterization; Machine learning

1. Introduction

Atmospheric convection plays a crucial role in Earth's climate system. By transporting mass and energy from the surface to the tropopause, convection impacts the global water, energy, and momentum budgets (Plant and Yano 2015). Convection also produces clouds that affect the radiation budget (Jones et al. 2024; Hartmann 2016). Severe weather events linked to atmospheric convection can result in substantial socioeconomic impacts such as loss of lives and property. It is therefore essential to represent convection accurately in atmospheric simulations.

To explicitly resolve individual convective cells and their intracloud motions, atmospheric models need to have a minimum spatial resolution on the order of 100 m (Bryan et al. 2003), but most present-day global atmospheric models range from a resolution of about 10–100 km. Additionally, convection can occur on the time scale of minutes, but most climate models have a time step on the order of 10 min (Cui et al. 2021; Christensen and Zanna 2022). Higher-resolution, kilometer-scale global models under development can resolve larger-scale convective storm structures (e.g., Hohenegger et al. 2023), but these simulations will not be practical in the near future for operational climate projections with multiple runs given computational limits.

Since convection temporal and spatial scales are much smaller than most present-day general circulation model (GCM) scales, convection cannot be explicitly resolved, so it is instead represented with a convection parameterization scheme. An important part of the convection parameterization scheme is the triggering of the convection parameterization, which determines where and when convection occurs in the model. Here, the "convective trigger" represents the activation of the convection parameterization in a model grid box from the Eulerian perspective. This study treats triggering similarly to most current convective trigger functions, which do not consider the history of convection and thus do not differentiate between different stages of the convection life cycle (Suhas and Zhang 2014). Generally, the basis of the trigger function is that convectively unstable air in the lower part of the atmosphere under certain atmospheric conditions can trigger convection in the model (Suhas and Zhang 2014). For example, in the NCAR Community Atmosphere Model, version 5 (CAM5),

Denotes content that is immediately available upon publication as open access.

Corresponding author: Greta A. Miller, greta.miller@physics.ox.ac.uk

DOI: 10.1175/JAS-D-24-0085.1

deep convection is triggered when the dilute convective available potential energy (dilute CAPE) exceeds a threshold of 70 J kg^{-1} (Cui et al. 2021).

Since the large-scale atmospheric processes that lead to the triggering of convection are not well understood, different weather and climate models use various convective trigger functions (Suhas and Zhang 2014). The primary differences between different convective trigger functions are how the source layer of convection is chosen within the troposphere and how the function realizes the atmospheric instability for convection development (Suhas and Zhang 2014). To initiate the release of the convective instability, some parameterization schemes use the gridscale vertical velocity (e.g., Kain and Fritsch 1990; Bechtold et al. 2001). Other models use boundary layer moisture convergence (Tiedtke 1989) or the amount of CAPE as in Zhang and McFarlane (1995). Xie and Zhang (2000) improved upon the CAPE trigger by adding a second condition, the dynamic generation rate of CAPE from large-scale advection (dCAPE). Xie et al. (2019) further modified dCAPE with an unrestricted parcel launch level (dCAPE ULL) to capture nocturnal elevated convection. Another modification of the CAPE trigger was the inclusion of entrainment in the calculation of CAPE (dilute CAPE) by Neale et al. (2008), which improved the simulation of El Niño–Southern Oscillation.

While there is a general understanding that local-scale convective instabilities cause convection, it is still difficult to map these local processes to coarser GCM scales in space and time. Despite various developments in the convective trigger function over the past several decades, the accurate prediction of the onset time, location, and evolution of convection continues to be a challenge in atmospheric models. Many convective triggers activate convection too often, resulting in models producing more drizzling rain than observations (Suhas and Zhang 2014). Errors in the convection trigger function have also been linked to inadequate simulations of the diurnal cycle of convection, the Madden–Julian oscillation, and the ITCZ seasonal migration and double peak biases (Suhas and Zhang 2014). Even with recent improvements, such as using the dynamic generation rate of CAPE (dCAPE) instead of CAPE for the convection trigger, there are still biases in the diurnal cycle and precipitation intensity compared to observations (Cui et al. 2021).

One promising approach for improving the representation of convection in atmospheric models is stochastic parameterizations. Since a given large-scale atmospheric state could correspond to multiple subgrid convective (or nonconvective) states, representing the occurrence of subgrid convection as a stochastic process is a natural choice. Furthermore, the stochastic parameterization of the convective trigger will become increasingly important as the spatial resolution of models increases and the scale separation between the large-scale and subgrid processes decreases; the assumption that each model grid box contains an ensemble of randomly distributed convective plumes is no longer valid (Berner et al. 2017). A few studies have specifically focused on the stochasticity of the convection trigger by adding a probabilistic component to a conventional trigger (Bright and Mullen 2002; Song et al. 2007; Rochetin et al. 2014). In a different study, the implementation of a stochastic deep convection parameterization scheme in NCAR CAM5 resulted in an improved precipitation frequency (Wang et al. 2016).

Another direction for improving the convective trigger function is the application of machine learning (ML) to learn nonlinear relationships directly from observations. Machine learning is particularly well suited to the convective trigger since the trigger function can easily learn a threshold for when to activate convection from observations. In contrast to conventional convective trigger functions which often use just one or two atmospheric state variables (e.g., vertical velocity, moisture), a ML trigger function can incorporate multiple different dynamic and thermodynamic variables to predict the occurrence of convection for a broader range of meteorological regimes. The inputs and models that perform best are useful in determining which physical assumptions are most accurate; using ML to develop a trigger can inform us about the physics of convection. In addition, since ML algorithms can predict a probability instead of a binary outcome, a ML trigger could be applied as a stochastic trigger parameterization in an operational model.

Both ML and stochastic parameterization methods can be used in combination, creating parameterizations that better represent the uncertainty learned from observations or high-resolution datasets (Christensen et al. 2024). For example, Nadiga et al. (2022) use a generative adversarial network, a probabilistic ML method, to generate vertical tendency profiles of temperature and humidity given the large-scale atmospheric state from reanalysis. Behrens et al. (2024) apply various stochastic ML approaches to parameterize subgrid convection and turbulence from a superparameterization embedded in a GCM.

Few studies have specifically focused on using ML for the convective trigger. Ukkonen and Mäkelä (2019) applied several ML methods to predict the occurrence of thunderstorms in Europe and Sri Lanka using ERA5 atmospheric reanalysis and lightning datasets. Zhang et al. (2021) applied gradient boosted trees to a variational analysis dataset over the Southern Great Plains, United States, and the Brazilian Amazon to predict the occurrence of convection. Chen et al. (2023) applied a convolutional neural network to ERA5 reanalysis and TRMM satellite precipitation to predict the occurrence of convection with location awareness. However, none of the studies consider the ML models within a probabilistic framework to characterize the uncertainty in predicting the occurrence of convection, which is particularly important as models approach the gray zone (Christensen and Zanna 2022).

This study aims to improve the understanding of sources of uncertainty in deep convective triggering to ultimately improve the representation of convection in weather and climate models with parameterized convection. A ML model is used to predict the probability of the occurrence of convection, and explainable ML methods are used to understand the uncertainty of convection triggering. The ML model, a random forest, is trained on observed large-scale atmospheric variables from the Atmospheric Radiation Measurement (ARM)-constrained variational analysis over the Southern Great Plains observational

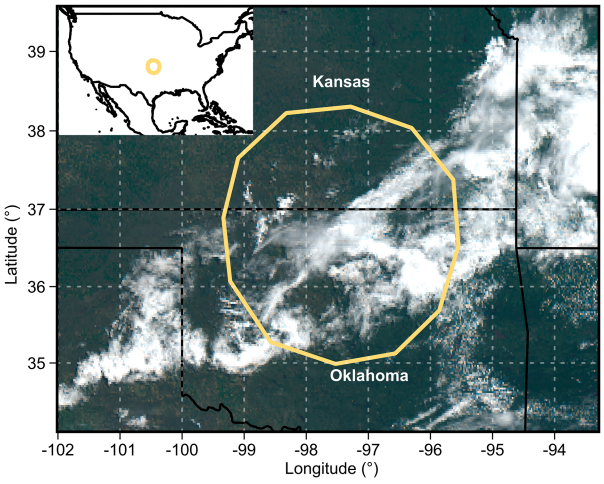

FIG. 1. Domain for the Southern Great Plains variational analysis, outlined by the yellow polygon, located in the central United States. [domain boundary adapted from Tang et al. (2019)]. The RGB image is from GOES-R Advanced Baseline Imager on 18 Aug 2017, using the 0.45–0.49-, 0.59–0.69-, and 0.846–0.885-μm wavelength bands (Schmit et al. 2017). Further information on the observational site can be found in Tang et al. (2019).

site. This approach builds off of Zhang et al. (2021) but within a probabilistic framework, which incorporates the aleatoric uncertainty in modeling convection occurrence. The model is evaluated against the observed time series of convection, the diurnal cycle, and the reliability of the probability of occurrence. To make the random forest explainable, the relative importance of each feature is used to assess which environmental factors are most closely linked to the occurrence of convection. The contribution of each of the environmental features to the predicted probabilities is then used to cluster the convective events and characterize certain and uncertain convective regimes.

We present the dataset and ML methods in section 2. The main results are presented in section 3, including the ML trigger skill and uncertainty analysis. In section 4, we discuss the results and implications for weather and climate modeling. Finally, section 5 concludes with a summary of the study.

2. Methods

a. Data

The ML model in this study is developed using inputs from observed large-scale atmospheric variables from the ARM-constrained variational analysis dataset (Tang et al. 2019). This study focuses on a dataset from the Southern Great Plains (SGP) observational site, in Oklahoma, United States, shown in Fig. 1. The SGP site is a multidecade observatory that has surface meteorological observation stations, eddy flux towers, microwave radiometer stations, and solar and infrared observing stations (Tang et al. 2019). The observations are used in combination with geostationary satellite products, a gridded precipitation product, and Rapid Update Cycle (RUC, before 2012)/Rapid Refresh (RAP, after 2012) numerical weather reanalysis products (Tang et al. 2019). In the variational analysis method, the atmospheric state variables from the reanalysis are constrained by the observed surface and top of atmosphere fluxes (Tang et al. 2019). The variational analysis dataset is particularly useful for the application of a ML trigger because, unlike most reanalysis products, the dataset is dynamically and thermodynamically consistent with the observed fluxes including precipitation.

The variational analysis dataset averages the atmospheric state in a region across approximately 96°–99°W to 35°–38°N (Fig. 1). The vertical resolution of the dataset is 25 hPa, and the temporal resolution is 1 h. Inputs from the convectively active summer season (June–August) from 2004 to 2018 are used in the model.

To define the occurrence of deep convection, a precipitation threshold of 0.5 mm h^{-1} is used, following Suhas and Zhang (2014), Song and Zhang (2017), and Zhang et al. (2021). Using this precipitation threshold is not an exact indicator of convection since there is a lag between when convection initiates and when precipitation occurs. The threshold may also include some stratiform precipitation events. However, this reasonably large precipitation threshold is generally a good indicator of convection over the Southern Great Plains during the summertime (Suhas and Zhang 2014). Suhas and Zhang (2014) also tested a range of precipitation thresholds and found that the results of the trigger functions are not sensitive to the threshold value.

To predict the occurrence of convection, various dynamic and thermodynamic variables ("features") are used in the model (Table 1). Convective processes depend on the atmospheric state near the surface, such as surface temperature, humidity, and sensible and latent heat fluxes, as well as the vertical profiles of atmospheric variables. The features include dilute CAPE, column precipitable water, and large-scale vertical velocity, in addition to those used in Zhang et al. (2021). Since using all vertical levels of the dataset would likely require a more complex ML method to extract relevant convective signals, meaningful scalar features are derived from the vertical column instead. Some atmospheric features, such as temperature and vertical velocity, are averaged across the lower (700–800 hPa), middle (300–700 hPa), and upper (200–300 hPa) levels of the troposphere. Other features, such as convective inhibition (CIN) and dilute CAPE, are indicators of atmospheric stability and are calculated using vertical profiles of temperature and humidity.

b. Machine learning trigger

A random forest classifier is used to predict the probability of convection in this study, which is implemented using the Scikit-learn library (Pedregosa et al. 2011). The random forest algorithm is a ML method that averages an ensemble of decision trees (Breiman 2001). Each decision tree is trained on a bootstrapped subset of the dataset, and each decision split is limited to a random subset of the input features (Pedregosa et al. 2011). For a single tree, the predicted probability is the fraction of samples of the same class in a leaf. The probability

TABLE 1. Input features for the machine learning trigger. Notes: The lower, middle, and upper troposphere are defined by the pressure levels 700–800, 300–700, and 200–300 hPa, respectively. Table partially adapted from Zhang et al. (2021).

Predictors	Abbreviation	Feature group
Latent heat flux	Latent heat	Srf and buoyancy
Sensible heat flux	Sensible heat	Srf and buoyancy
Air temperature at the surface	T surface	Srf and buoyancy
Air relative humidity at the surface	RH surface	Srf and buoyancy
Lifting condensation level	LCL	Srf and buoyancy
Convective inhibition	CIN	Srf and buoyancy
Dynamic generation of dilute convective available potential energy	Dilute dCAPE	Vert. velocity
Vertical velocity in the lower troposphere	ω low	Vert. velocity
Vertical velocity in the midtroposphere	ω mid	Vert. velocity
Vertical velocity in the upper troposphere	ω high	Vert. velocity
Precipitable water in the column	PW	Moisture
Specific humidity in the lower troposphere	q low	Moisture
Specific humidity in the midtroposphere	q mid	Moisture
Specific humidity in the upper troposphere	q high	Moisture
Temperature in the lower troposphere	T low	Temperature
Temperature in the midtroposphere	T mid	Temperature
Temperature in the upper troposphere	T high	Temperature
Horizontal advective tendency of specific humidity in the lower troposphere	Adv. q low	Adv. q
Horizontal advective tendency of specific humidity in the midtroposphere	Adv. q mid	Adv. q
Horizontal advective tendency of specific humidity in the upper troposphere	Adv. q high	Adv. q
Horizontal advective tendency of dry static energy in the lower troposphere	Adv. s low	Adv. s
Horizontal advective tendency of dry static energy in the midtroposphere	Adv. s mid	Adv. s
Horizontal advective tendency of dry static energy in the upper troposphere	Adv. s high	Adv. s
Wind shear in the lower troposphere	Shear low	Shear
Wind shear in the midtroposphere	Shear mid	Shear
Wind shear in the upper troposphere	Shear high	Shear

of convection is calculated as the mean of the predicted probabilities of the trees in the forest (Pedregosa et al. 2011).

With random forests, the feature importance is shared, but diluted, among correlated features because of the random selection of features at each node (Breiman 2001). In contrast, other approaches, such as logistic regression or neural networks, can heavily favor one correlated feature over another or even give correlated features opposite signs (Flora et al. 2024). While multiple ML methods could be suitable, we used a random forest because of its simplicity to train and tune, and because tree-based methods in general can often outperform other methods like neural networks on tabular-style datasets (Lundberg et al. 2020).

The random forest is trained and evaluated using the environmental features from Table 1 to predict the occurrence of convection. The first 12 summers of the dataset (2004–15; 26 496 samples) are used for training, and the last 3 summers (2016–18; 6624 samples) are used for model evaluation. All features are standardized to have a mean of 0 and a standard deviation of 1 before model training. Further details of the random forest hyperparameter optimization are given in appendix A.

The atmospheric features represent the domain-averaged value at a given time, while the corresponding precipitation represents the total precipitation from 30 min before to 30 min after that time. The model uses the features to predict the occurrence of convection at the next time step (hour), i.e., 30–90 min after the feature variables are valid. This allows the model to learn to predict convection that will happen in the next hour, instead of diagnosing convection that has already happened in the current hour.

Both convective and nonconvective events at hourly temporal resolution are included as inputs to train the model. In the original dataset, the number of nonconvective events (majority class) far outnumbers the number of convective events (minority class), with a convective to nonconvective ratio of 0.095. To reduce the class imbalance and optimize training, the majority class was randomly undersampled by a factor of $1/r$. Since the dataset was resampled, the frequency of convective events will be different in the undersampled dataset compared to the original dataset (Miloshevich et al. 2023). Thus, the raw probabilities predicted from the undersampled dataset were adjusted to evaluate skill using the original dataset (Miloshevich et al. 2023). A more detailed explanation of the undersampling and probability scaling is given in appendix B.

The ML trigger is designed similarly to convection triggers in operational models; it identifies convection occurrence at every time step, including both the initiation and persistence of convection. Thus, the trigger must be designed to capture all occurrences of convection due to large-scale upward motion associated with processes such as synoptic-scale systems, existing convection, subgrid-scale dynamic instability, or the growth of the boundary layer (Xie et al. 2004).

c. *Evaluation metrics*

To evaluate the model skill, various scores are used to compare the probabilistic random forest to deterministic conventional triggers. For a deterministic comparison, precision (also

known as success ratio), recall (also known as probability of detection), and the F1 score are applied to the minority class (convective events). These evaluation metrics emphasize the performance on the minority class, which allows for a more nuanced evaluation given the imbalanced dataset. To compare the ML trigger to deterministic triggers, we assume a binary outcome of the random forest, where probabilities greater than 0.47 are convective events. The threshold of 0.47 ensures that the observed frequency of convection matches the predicted frequency. Precision P is the percentage of true positives (TPs) over the total of the predicted positive cases [TPs plus false positives (FPs)]:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{1}$$

Recall R is the percentage of TPs over the total of the true-positive cases [TPs plus false negatives (FNs)]:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{2}$$

The F1 score incorporates skill in both overprediction and underprediction and is equal to the harmonic mean of precision and recall:

$$\mathrm{F1} = \frac{2PR}{P + R}. \tag{3}$$

The F1 score optimally has a perfect value of one, and it is closer to one when both precision and recall are high.

While precision, recall, and F1 scores are deterministic scores, the triggers are additionally evaluated in a probabilistic framework with the Brier skill score (BSS):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}, \tag{4}$$

where BS is the Brier score for each trigger and $\mathrm{BS}_{\mathrm{ref}}$ is the reference Brier score. The Brier score is defined as

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2, \tag{5}$$

where p_i are the predicted probabilities of convection and o_i are the observed occurrences of convection, summed over the size of the dataset n. The reference Brier score $\mathrm{BS}_{\mathrm{ref}}$ uses the average frequency of observed convection during each hour of the day as the predicted probabilities. The BSS has a perfect value of one but can be negative if the predicted probabilities are worse than the reference. The BSSs for the deterministic triggers are also calculated with Eq. (4) but using the binary predicted convection occurrences instead of probabilities. In contrast to the deterministic scores, the BSS is not as sensitive to the minority class for an imbalanced dataset.

d. Conventional trigger functions

The random forest was also evaluated against four conventional CAPE-based trigger functions: CAPE, dilute CAPE, the dynamic generation rate of CAPE, and the dynamic generation rate of dilute CAPE. These trigger functions were tested on observations along with various other trigger functions by Suhas and Zhang (2014) and Song and Zhang (2017), who found that the dynamic generation rate of dilute dCAPE had the best performance. In this study, all four trigger functions use pseudoadiabatic parcels that ascend from the most unstable pressure level in the atmosphere below 600 hPa, known as the unrestricted launch level (Xie et al. 2019). Allowing the parcels to launch from above the boundary layer helps to capture elevated nocturnal convection (Xie et al. 2019).

CAPE is defined as the vertical integral of a parcel's buoyancy from the level of free convection to the level of neutral buoyancy:

$$\mathrm{CAPE} = -\int_{\ln p_{\mathrm{LFC}}}^{\ln p_{\mathrm{LNB}}} R_d(T_{v,\mathrm{parcel}} - T_{v,\mathrm{env}})d\ln p, \tag{6}$$

where R_d is the gas constant for dry air, p_{LNB} is the pressure at the level of neutral buoyancy, p_{LFC} is the pressure at the level of free convection, $T_{v,\mathrm{parcel}}$ is the virtual temperature of parcel, and $T_{v,\mathrm{env}}$ is the virtual temperature of environmental air. Originally developed for the Zhang–McFarlane scheme (Zhang and McFarlane 1995), CAPE-based triggers have been used in models such as the NCAR CAM model, WRF, and the U.S. Department of Energy E3SM Atmosphere Model. In models, when CAPE exceeds a threshold, typically 70 J kg^{-1}, the convection parameterization scheme is activated.

Xie and Zhang (2000) improved the CAPE trigger by developing dCAPE, which is the dynamic generation rate of CAPE from large-scale advection. dCAPE is calculated by

$$\mathrm{dCAPE} = \frac{\mathrm{CAPE}\left(T + \frac{\partial T}{\partial t}\delta t,\, q + \frac{\partial q}{\partial t}\delta t\right) - \mathrm{CAPE}(T, q)}{\delta t}, \tag{7}$$

where $\partial T/\partial t$ is the advective tendency of temperature due to large-scale advection, $\partial q/\partial t$ is the advective tendency of specific humidity due to large-scale advection, and δt is the time step (1 h). In operational models, when dCAPE exceeds a threshold, typically 65 J kg^{-1} h^{-1}, the convection parameterization scheme is activated.

The advective tendency of temperature is defined as the sum of the horizontal advective tendency (a), the vertical advective tendency (b), and the adiabatic expansion term (c):

$$\frac{\partial T}{\partial t} = \underbrace{-\mathbf{u}\cdot\nabla T}_{\text{(a)}} \underbrace{-\omega\frac{\partial T}{\partial p}}_{\text{(b)}} \underbrace{+\frac{\omega}{c_p\rho}}_{\text{(c)}}, \tag{8}$$

where $\mathbf{u}$ is the horizontal velocity vector, ω is the vertical velocity, c_p is the specific heat capacity of air, and ρ is the air density. The advective tendency of temperature is also equivalent to the advective tendency of dry static energy, which is sometimes used instead in the definition of dCAPE (e.g., Song and Zhang 2018).

The advective tendency of specific humidity is defined as the sum of the horizontal advective tendency and the vertical advective tendency:

$$\frac{\partial q}{\partial t} = -\mathbf{u} \cdot \nabla q - \omega \frac{\partial q}{\partial p}. \tag{9}$$

Neale et al. (2008) also improved the CAPE trigger by incorporating entrainment into the CAPE calculation (dilute CAPE). When a parcel ascends, its entropy S follows the equation:

$$\frac{\partial mS}{\partial z} = \frac{\partial m}{\partial z}\overline{S} = \varepsilon \overline{S}, \tag{10}$$

where m is the parcel mass and $\overline{S}$ is the environment entropy. The entrainment rate ε is set to 10^{-3} m^{-1}. As the parcel ascends, its entropy is updated with dilution, and then, the temperature and humidity at each level are obtained. In models, when dilute CAPE exceeds a threshold, typically 70 J kg^{-1}, the convection parameterization scheme is activated.

The fourth trigger function is dilute dCAPE, which combines the dynamic generation rate from advection [Eq. (7)] with the entrainment calculation [Eq. (10)]. The threshold used is the same for dCAPE, 65 J kg^{-1} h^{-1}.

To fairly compare the CAPE-based triggers to the ML trigger, values across the range of each CAPE-based trigger were tested as thresholds. For each trigger, the threshold with the highest F1 score on the same training set used for the random forest was chosen, and the optimized threshold was used for evaluation on the test set. The optimized thresholds for CAPE, dilute CAPE, dCAPE, and dilute dCAPE were 259 J kg^{-1}, 65 J kg^{-1}, 141 J kg^{-1} h^{-1}, and 66 J kg^{-1} h^{-1}, respectively. The optimized thresholds differ significantly from the operational thresholds (70 J kg^{-1}, 70 J kg^{-1}, 65 J kg^{-1} h^{-1}, and 65 J kg^{-1} h^{-1}, respectively), possibly because they are optimized on a single location and season instead of a full global model.

e. Explainable machine learning methods

To explain the random forest predictions, we used the feature importance methods, which estimate the impact on model performance, and feature relevance methods, which estimate the impact on the predicted probabilities (Flora et al. 2024). For feature relevance, we used Shapley additive explanations (SHAP) calculated using KernelSHAP from Lundberg and Lee (2017). A SHAP value represents the marginal contribution of a feature to a prediction while considering all permutations of features (Lundberg and Lee 2017). For an event, the SHAP values of each feature sum to the difference between the expected probability and the predicted probability. Individual SHAP values can be positive or negative, indicating that the random forest judges the given feature to lead to a more or less likely chance of convection, respectively. They are particularly useful for explainable ML because they provide both local (individual sample) and global (average across all features or samples) information about how the model works. The average magnitude of the SHAP value of each feature gives a global metric for feature relevance.

For feature importance, we used SAGE values, which estimate the global impact of features on model performance using the test set. SAGE values are calculated in a similar way to SHAP values, except that the impact is measured with model performance instead of the model output (Covert et al. 2020). We also compared the SAGE feature importance to impurity feature importance, which is based on the average decrease in impurity when a feature is used to split a leaf in the trees in the random forest (Breiman 2001). In addition, we compared single-pass permutation feature importance, which is proportional to the magnitude of the decrease in model skill when data corresponding to a single feature are randomly shuffled (Breiman 2001). The single-pass permutation feature importance was evaluated on the F1 score of the test set.

Since the random forest can dilute feature importance and relevance among correlated features, we also grouped the features into correlated groups, which are defined in Table 1. The grouped SAGE and SHAP rankings are calculated by summing the SAGE and SHAP values for the features in each group (Flora et al. 2024). This approach helps capture the collective effects of correlated features on the model (Flora et al. 2024).

f. Clustering into convective regimes

After the random forest is trained on the inputs, the SHAP values for each individual sample and feature are clustered using k-means clustering from the Scikit-learn ML library for Python (Pedregosa et al. 2011). Following a method similar to A. Douglas and P. Stier (2024, unpublished manuscript), clustering with SHAP values produces convective triggering regimes controlled by a similar combination of input features. Instead of clustering the original feature values, clustering the SHAP values allows for the interpretation to focus on the contribution of each feature to the predicted convection probabilities. To determine the optimal number of clusters, the elbow method was used (Kodinariya and Makwana 2013), in which the within-cluster variance is plotted against the number of clusters. The number of clusters used in this study, $n = 6$, was chosen where the decline in variance flattens with an increasing number of clusters. More details on the elbow method are shown in appendix C. The clusters from the SHAP values were then interpreted in terms of the model inputs and the uncertainty in the predicted probabilities. Since the SHAP values correspond to the unscaled predicted probabilities, the probabilities from the random forest were not scaled for the clustering analysis.

3. Results

a. Random forest convection trigger skill

The random forest trigger was trained to predict the probability of convection, and then, its skill was evaluated and compared to conventional CAPE-based triggers. The skill of the random forest is first evaluated on the test set using precision, recall, and F1 scores. On the full, unbalanced test dataset after the probability scaling, which is similar to the original observed dataset, the

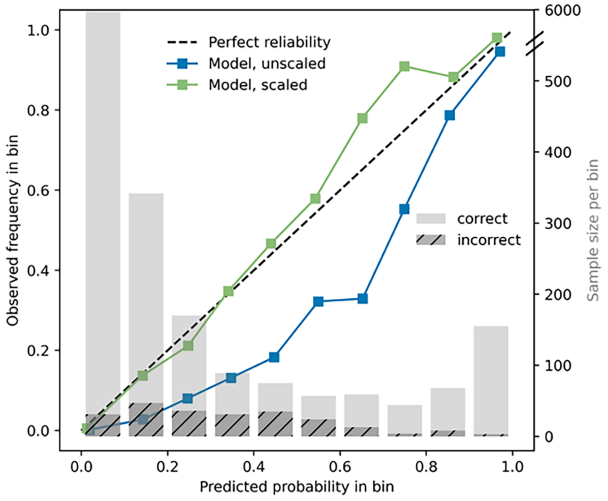

FIG. 2. Reliability curve for the random forest, before (blue line) and after (green line) the probability scaling. The reliability curve after the probability scaling is closer to a perfect reliability curve (black dashed line), indicating that the probability scaling makes the predictions more reliable. The histogram in gray shows the marginal totals, which are the sample sizes of the predicted probabilities in each bin. Most events are predicted as having near-zero probabilities (note the broken *y* axis for the histogram sample size).

precision, recall, and F1 scores were 0.77, 0.73, and 0.75, respectively. Since the precision is higher than the recall, the model tends to underpredict convective events; there were slightly fewer false positives than false negatives.

In contrast to the deterministic scores that assume a binary output from the random forest, the BSS evaluates the random forest probabilities directly. On the full, unbalanced test dataset after the probability scaling, the BSS is 0.47.

A reliability diagram can also be used to evaluate the predicted probabilities without transforming them into a binary output. Figure 2 shows how well the predicted probabilities from the random forest match the observed frequency of convection occurrence. The probabilities that were adjusted using the probability scaling are closer to a perfect reliability curve because the probability scaling systematically reduces the probability assigned to convective events. Figure 2 also shows the marginal totals, which represent the distribution of sample sizes of the predicted probabilities. Most events are predicted as having near-zero probabilities.

b. Diurnal cycle and comparison to conventional trigger functions

The random forest trigger was compared to the conventional triggers using both deterministic and probabilistic evaluation metrics. Figure 3 shows a comparison for precision, recall, F1, and BSS. The random forest outperforms all of the CAPE-based conventional triggers for precision, F1, and BSS. While CAPE and dilute CAPE have high recall because they tend to overpredict convection, the overpredictions also cause much lower precision, F1, and BSS. CAPE and dilute CAPE have negative BSS, indicating they perform worse than the baseline hourly diurnal frequency of observed convection. While the random forest trigger performs better than the conventional triggers for the deterministic metrics, it performs particularly well for the probabilistic metric, the BSS, since the random forest itself is probabilistic.

A promising ML trigger needs to simulate the diurnal cycle accurately, since the diurnal cycle has important impacts on the Earth system, such as different shortwave and net cloud radiative effects during the night and day. Figure 4 shows the diurnal cycle over the SGP for the random forest convection probability in comparison with the observed occurrence of convection and four CAPE-based triggers. The observed summertime diurnal cycle over the Southern Great Plains has two peaks in the frequency of convection throughout the day: a peak in the evening, which is connected to surface heating throughout the day, and a peak in the night/early morning,

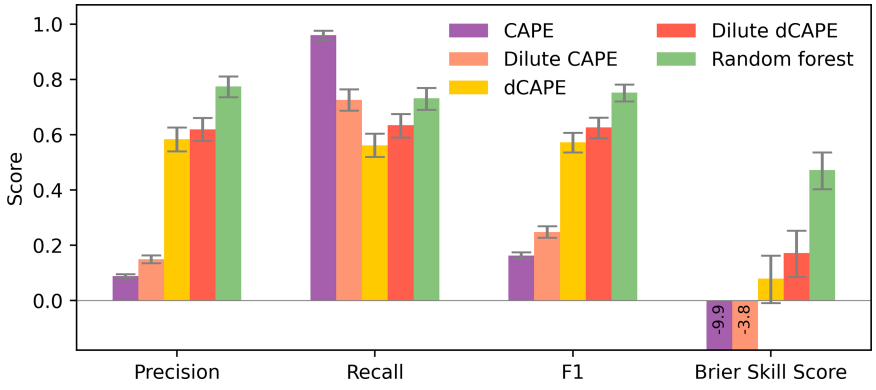

FIG. 3. Comparison of the random forest and conventional triggers for precision, recall, F1, and BSS. For all evaluation metrics, the values closer to 1 are better. The gray bars represent 95% confidence intervals from 1000 bootstrapped trials. The random forest outperforms all conventional triggers for precision, F1, and BSS. While CAPE and dilute CAPE tend to overpredict convection, leading to high recall, the overpredictions lead to poor precision and thus poor F1 and BSS overall.

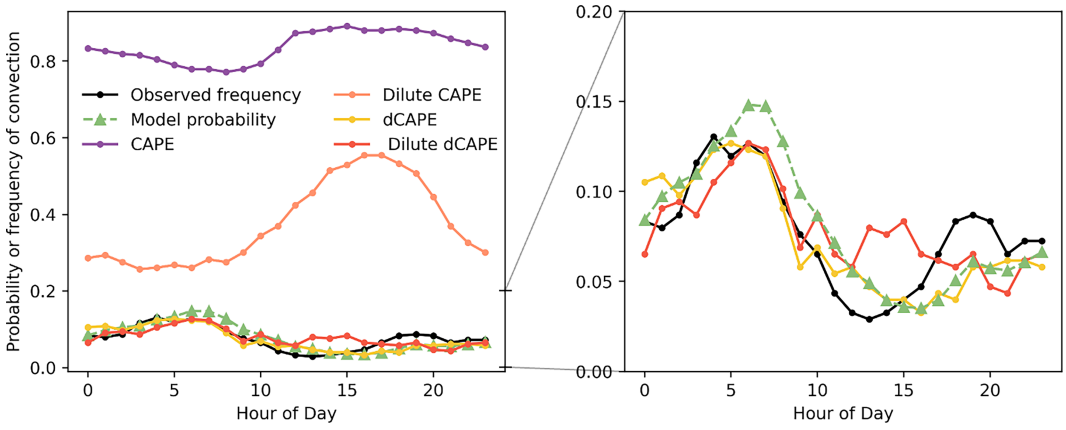

FIG. 4. Diurnal cycle of convection (local time) for the observations (black), random forest model probability (green), and four optimized conventional convection trigger frequencies for (left) all convection triggers and (right) a zoomed-in plot for the best-performing convective triggers: dCAPE, dilute dCAPE, and the random forest.

which is connected to mesoscale convective systems that propagate from the west (Zhao et al. 2017). Even with an optimized threshold, CAPE and dilute CAPE are poor predictors of convection since they overpredict many events. Of the conventional triggers, dCAPE and dilute dCAPE most closely match the observed diurnal cycle but have high hour-to-hour variability not seen in the observed cycle. Even though dilute dCAPE is found to be better than dCAPE in previous studies, dilute dCAPE does not capture the correct amplitude as well as dCAPE. The random forest slightly overestimates the observed amplitude but is also smoother than the dCAPE and dilute dCAPE triggers, which could be because it is more robust since it is able to consider multiple features. All of the better-performing triggers capture the nighttime convection but underpredict the evening peak in convection. In contrast to conventional trigger functions (Covey et al. 2016; Dai 2006), the diurnal cycle of the random forest trigger is not early in phase; instead, it is delayed by 1–2 h (time steps).

c. *Feature importance and relevance*

The application of a random forest to the ARM variational analysis can provide feature rankings to help gain insights into which environmental variables among the variables in the dataset are most closely linked to convection occurrence. Figure 5 shows the feature rankings using several feature importance and relevance measures. The different methods give different orderings for the features. Having different measures is therefore very useful, since comparing them indicates which conclusions are robust. Even though the details of the ordering can change, the broad picture shows consistency in the general rankings. Across all the methods, vertical velocity in the midtroposphere (ω mid) is the dominating environmental variable in predicting convection, followed by dilute dCAPE. The high predictive value of large-scale vertical velocity is consistent with known mechanisms of convection triggering, in which upward motion associated with convergence from large-scale dynamical processes helps convection overcome CIN and reach the level of free convection. It is also interesting that vertical velocity is found to have more predictive performance than dilute dCAPE, which was previously found to be the best-performing conventional trigger function (Suhas and Zhang 2014; Song and Zhang 2017).

Several other features including (in no particular order) vertical velocity in the upper and lower troposphere, column precipitable water, and surface relative humidity are also important features. The relative ranking of these features depends on the feature importance method. These features are generally consistent with previous studies, where surface properties, such as temperature and humidity, are important in determining the stability of the atmospheric column, and sufficient moisture throughout the atmospheric column is necessary to drive convection.

However, since the feature importance methods differ in ranking order because the methods used to measure importance are inherently different, caution should be used in taking any one method as the truth. This dataset provides a good example for why multiple methods should be applied to interpret feature importance results, whereas many previous studies only apply one method.

Since correlated features may have diluted feature relevance, the grouped SAGE and SHAP feature rankings are useful in understanding how correlated groups collectively impact the model and compare to other feature groups. In Figs. 5e and 5f, the vertical velocity features are ranked the highest for both SAGE and SHAP. The other important groups are the moisture features and the surface and buoyancy features, but their relative rankings differ for SHAP and SAGE. Overall, the grouped rankings reflect the individual rankings in that vertical velocity features dominate, and surface and moisture-related features follow, but the ranking of any individual method may not be robust.

While large-scale vertical velocity can cause convection, it can also be caused by convection. The large-scale vertical velocity in the variational analysis used in this study is more likely to be due to large-scale dynamics, given the large size of the domain (Fig. 1), which can contain many small-scale

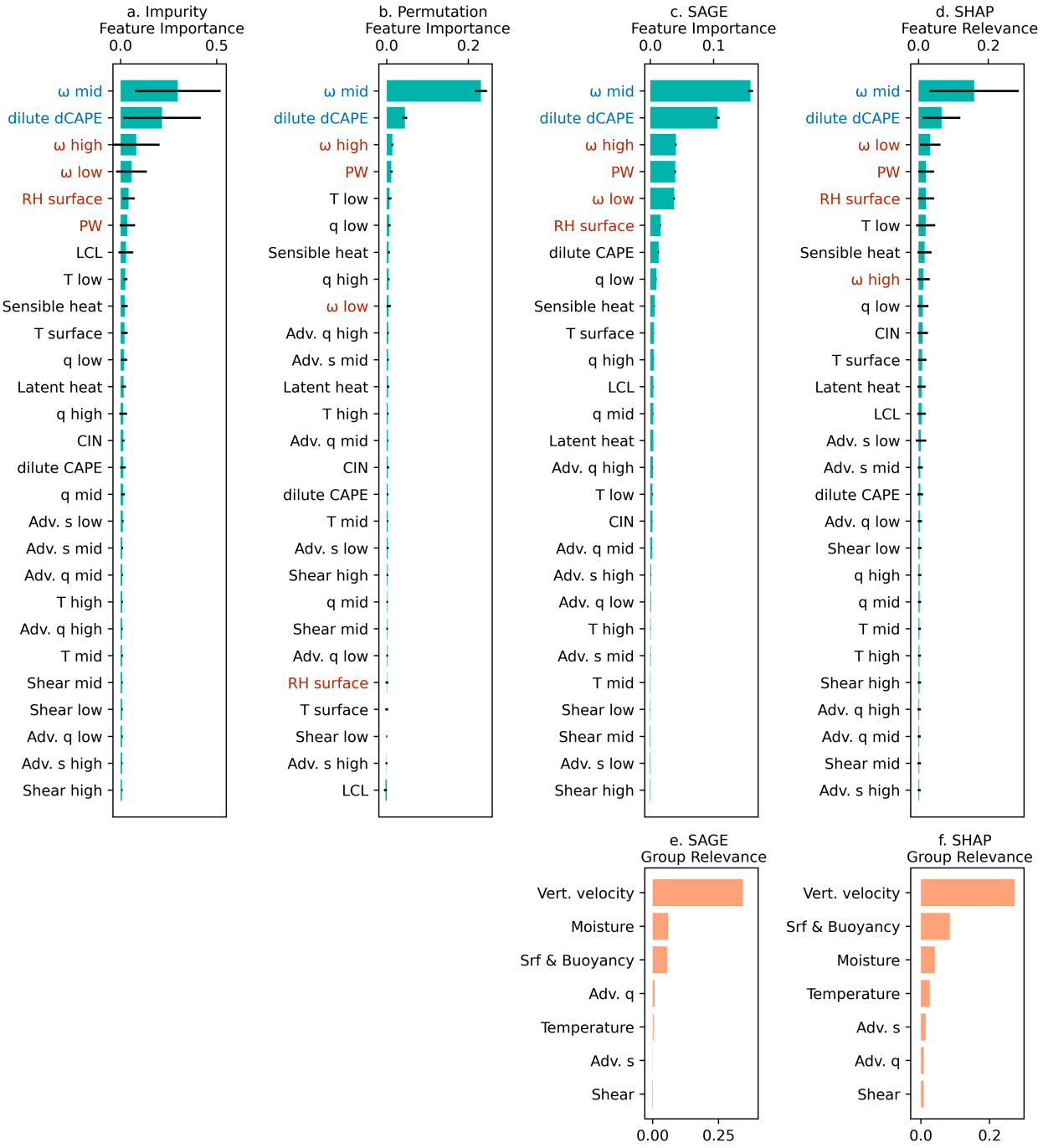

FIG. 5. Feature rankings show which inputs to the random forest are the most important predictors of convection for the dataset from the Southern Great Plains observational dataset. (a) Impurity feature importance, (b) single-pass permutation feature importance, (c) SAGE feature importance, and (d) SHAP feature relevance have similar, but not the same, feature rankings. The top two most important features for all metrics, in the blue text, are ω mid and dilute dCAPE. Features with higher importance but not the exact same ranking across all three metrics are in red text. The black bars represent the standard deviation of each feature importance value. The grouped feature rankings are shown for (e) SAGE and (f) SHAP values.

convective updrafts and downdrafts at one time. Conceptually, convection parameterizations do not typically output a vertical velocity tendency, under the assumption that the updrafts balance the downdrafts within a grid column. In addition, in atmospheric models, the convective trigger must be designed to initiate convection for both the start of convective events and persisting convection that may be triggered by previous convection. Thus, regardless of whether vertical

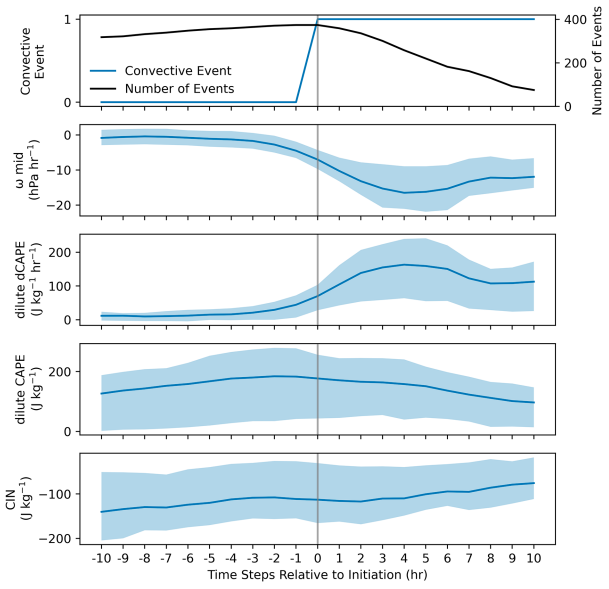

FIG. 6. Hourly averages (blue lines) and interquartile ranges (blue shaded areas) for various atmospheric variables before and during a convective event. The top panel shows the convective event, with no convection until time = 0 h, and persisting convection afterward. The black line in the top panel represents the number of samples averaged over each hour. Vertical velocity (ω mid) and dilute dCAPE show the changes in the few hours leading up to a convective event, but other variables such as dilute CAPE and CIN do not show a similar change.

velocity is due to large-scale dynamical processes, previous convection, or boundary layer growth, it tends to be a good indicator of convection. This is demonstrated in Fig. 6, which shows the average time series of the convective events over the SGP. Many hours before a convective event, vertical velocity tends to be close to zero. In the 1–2 h leading up to convection, upward vertical velocity begins to increase in magnitude and then remains high throughout the convection occurrence (note that negative ω indicates upward motion). Dilute dCAPE follows a remarkably similar path to vertical velocity leading up to convection, indicative of a close relationship between dilute dCAPE and vertical velocity, which is investigated further in the following section. Other atmospheric predictors, such as dilute CAPE and CIN, do not show a similar behavior.

Since dilute dCAPE appears to be the second-most important feature across all importance metrics and behaves as a precursor to convection in Fig. 6, we investigated the physical mechanisms linking dilute dCAPE and the occurrence of convection. Figure 7d shows the advective tendency temperature profile, which is used to calculate dilute dCAPE, following Eqs. (7)–(10), averaged for the convective and nonconvective events. In comparison with nonconvective events, the convective events have a negative total advective tendency of temperature in the middle and upper atmospheres, which decreases the environmental temperature profile, causing a larger temperature difference from the warmer air parcel and thus an increase in buoyancy.

In Fig. 7, the total advective tendency can be decomposed into the horizontal advection component [Eq. (8), term a], the vertical advection component [Eq. (8), term b], and a vertical adiabatic expansion term [Eq. (8), term c]. All three terms show differences between convective and nonconvective events, but the vertical adiabatic expansion term is the largest in magnitude and contributes the most to the total tendency (Fig. 7c, note the unequal horizontal scales). The vertical adiabatic expansion term is a product of a function of density, which is nearly identical for convection and nonconvection (Fig. 7f), and vertical velocity, which differs for the average convective and nonconvective cases (Fig. 7e). Therefore, the primary component in dilute dCAPE that makes it a good classifier of convection and nonconvection is vertical velocity. While dilute dCAPE is intended to capture more information than vertical velocity, such as heat and moisture convergence, the dataset in this study shows that vertical velocity tends to be the dominant link of dilute dCAPE to convection, and vertical velocity itself outperforms dilute dCAPE in the ML trigger. There are similar results for the advective tendency of humidity (not shown).

d. *Convective clusters and uncertainty*

To further explore model explainability, SHAP values were calculated for the random forest for each feature and event, before the SHAP values were clustered into six clusters using the *k*-means algorithm. Figure 8 (top row) shows the clustering results in the SHAP value space for several features. The scatters show distinctive clusters, which is expected since the *k*-means clustering was performed on the SHAP values. The clustering tends to be more defined for the most important features (e.g., ω mid, dilute dCAPE). Figure 8 also shows the same clusters but in the real feature space (bottom row, i.e., the actual atmospheric measurements used as predictors). Although the data were clustered using the SHAP values, the groupings are still somewhat apparent in the real feature space, which shows that the *k*-means clustering method still indirectly reflects the actual environmental state.

Figure 9 shows the predicted unscaled probabilities and observed occurrence of convection for each of the clusters. Even though the data were only clustered on the SHAP values, distinct convective and nonconvective groupings emerged, with an apparent split around a probability of 0.5. Based on the narrower distributions of predicted probabilities near zero and the large number of observed nonconvective events, the first and second clusters were characterized as certain nonconvective clusters. Based on the narrower distributions of probabilities near one and the large number of observed convective events, the fifth and sixth clusters were identified as certain convective clusters. In contrast, based on the large spread of probabilities between zero and the threshold, the third cluster was characterized as an uncertain nonconvective cluster. The fourth cluster was characterized as an uncertain convective cluster because of the large spread of probabilities between the threshold and one.

To further characterize each cluster, Fig. 10 shows the average feature values and SHAP values for each cluster. The

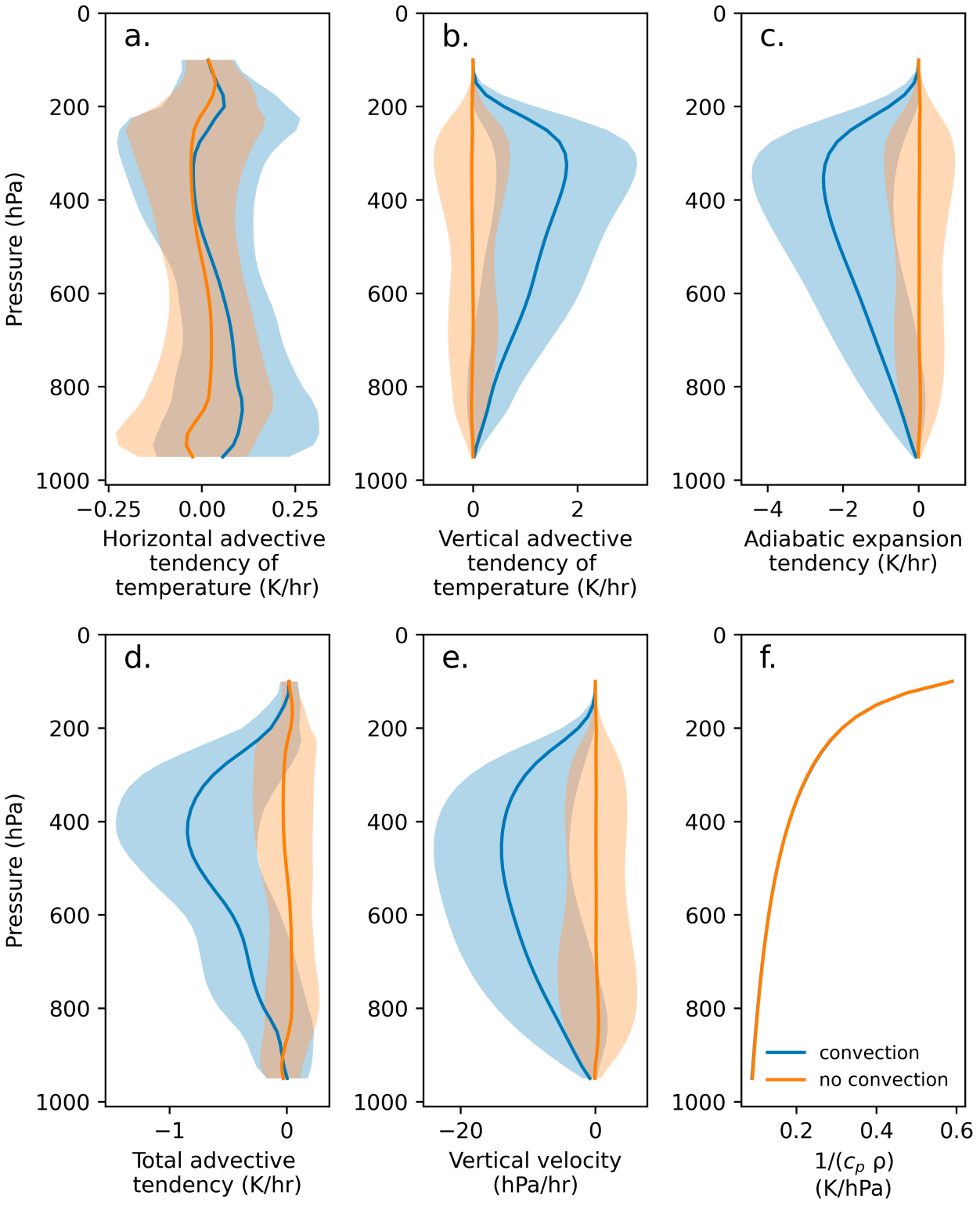

FIG. 7. Vertical profiles of the (a)–(c) terms that make up the (d) total advective tendency for temperature following Eq. (8), averaged for the convective (blue) and nonconvective (orange) events. The adiabatic expansion term (c) is the largest component of the total tendency, and the adiabatic term is equal to the product of (e) the vertical velocity and (f) a function of density, $1/(c_p\rho)$. The shaded areas represent one standard deviation from the mean. Note the unequal horizontal scales.

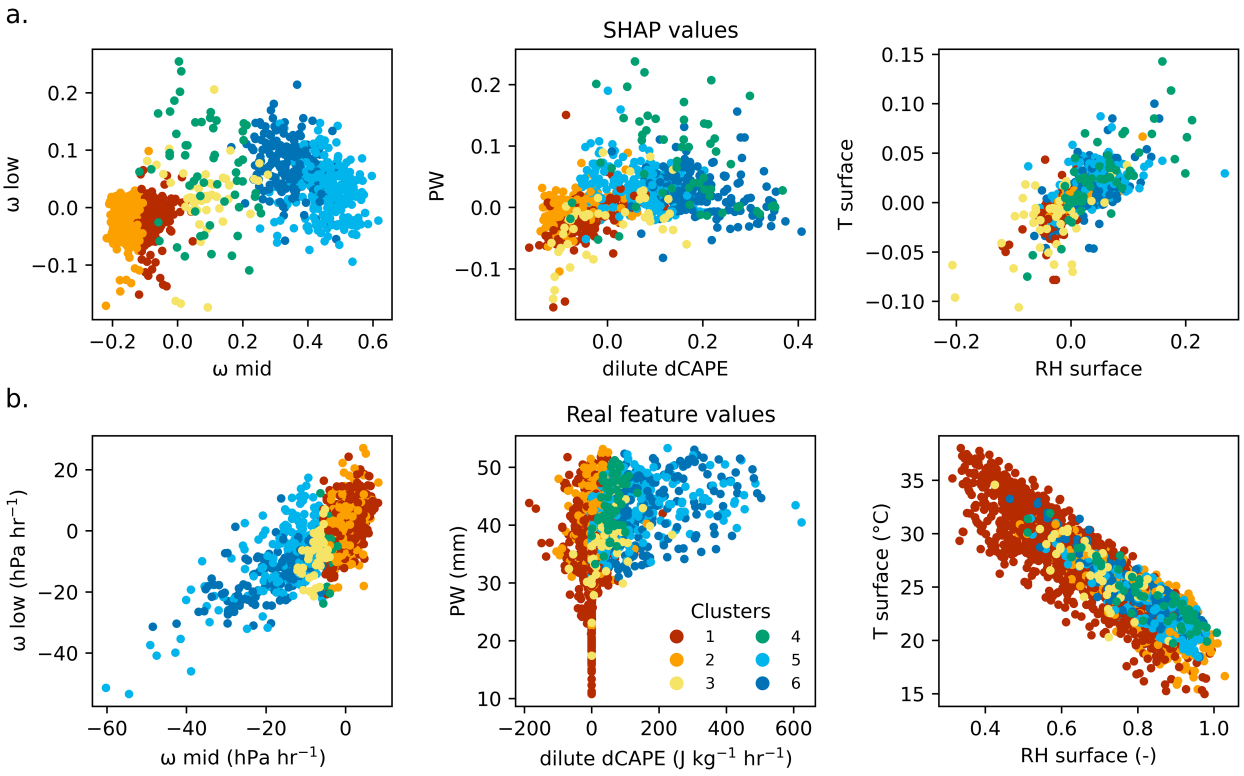

FIG. 8. (a) Scatters of the clusters in SHAP value space for several features. Since the *k*-means clustering was performed on the SHAP values, the clusters show distinct groupings in the SHAP value scatterplots. (b) Scatters of the clusters in real feature space for several features. Although the *k*-means clustering was performed on the SHAP values, the groupings of the clusters are still apparent in the real feature scatterplots.

certain convective events (clusters 5 and 6) have large upward (negative) vertical velocities that the ML model identifies as indicating a high probability of convection such that the SHAP value associated with vertical velocity is large and positive. For these certain convective events, the model considers features other than vertical velocity and dilute dCAPE very little. In contrast, the uncertain convective events (cluster 4) have a lower average upward vertical velocity and dilute dCAPE, and thus, the model uses other features such as precipitable water, low-level temperature, and surface relative humidity to predict the occurrence of convection. Without dynamical forcing, the model is less certain about predicting convection.

For the certain nonconvective events (clusters 1 and 2), the model uses downward (positive) or near-zero vertical velocity and low dilute dCAPE to predict a low probability of convection. Cluster 1 mostly contains daytime events, when the surface temperatures, latent heat fluxes, and sensible heat fluxes are usually higher. Cluster 2 mostly contains nighttime events, when the surface temperatures, latent heat fluxes, and sensible heat fluxes are usually lower. However, despite these cluster distinctions, the random forest primarily uses vertical velocity to predict the output for both. Uncertain nonconvective events (cluster 3) have a small average upward vertical velocity, so the model also relies on low-level temperature, surface relative humidity, and CIN to determine if the case is nonconvective.

These results are also reflected in Fig. 8, where, for example, in the bottom left panel, the certain convective cases (clusters 5 and 6) have a high upward vertical velocity. The uncertain events (clusters 3 and 4) have smaller, but still upward vertical velocities, but the uncertain convective events (cluster 4) tend to have higher precipitable water than the uncertain nonconvective events (cluster 3).

Figure 11 shows the SHAP dependence plots for vertical velocity, dilute dCAPE, precipitable water, and low-level temperature. Each curve on the SHAP dependence plot shows the partial predicted probability (or SHAP value) response as the feature value increases for each cluster. The certain convective events (clusters 5 and 6) show a distinct vertical velocity signal that is reflected in the higher SHAP values and thus higher probabilities of convection. The SHAP dependence plot for vertical velocity shows thresholding behavior, where events with vertical velocities exceeding approximately -5 hPa h^{-1} have a large positive contribution to the predicted probability. Even though the uncertain convective events (cluster 4) have similar vertical velocities as uncertain nonconvective events (cluster 3), they have distinct curves in precipitable water and low-level temperature, which differentiates them from nonconvective events.

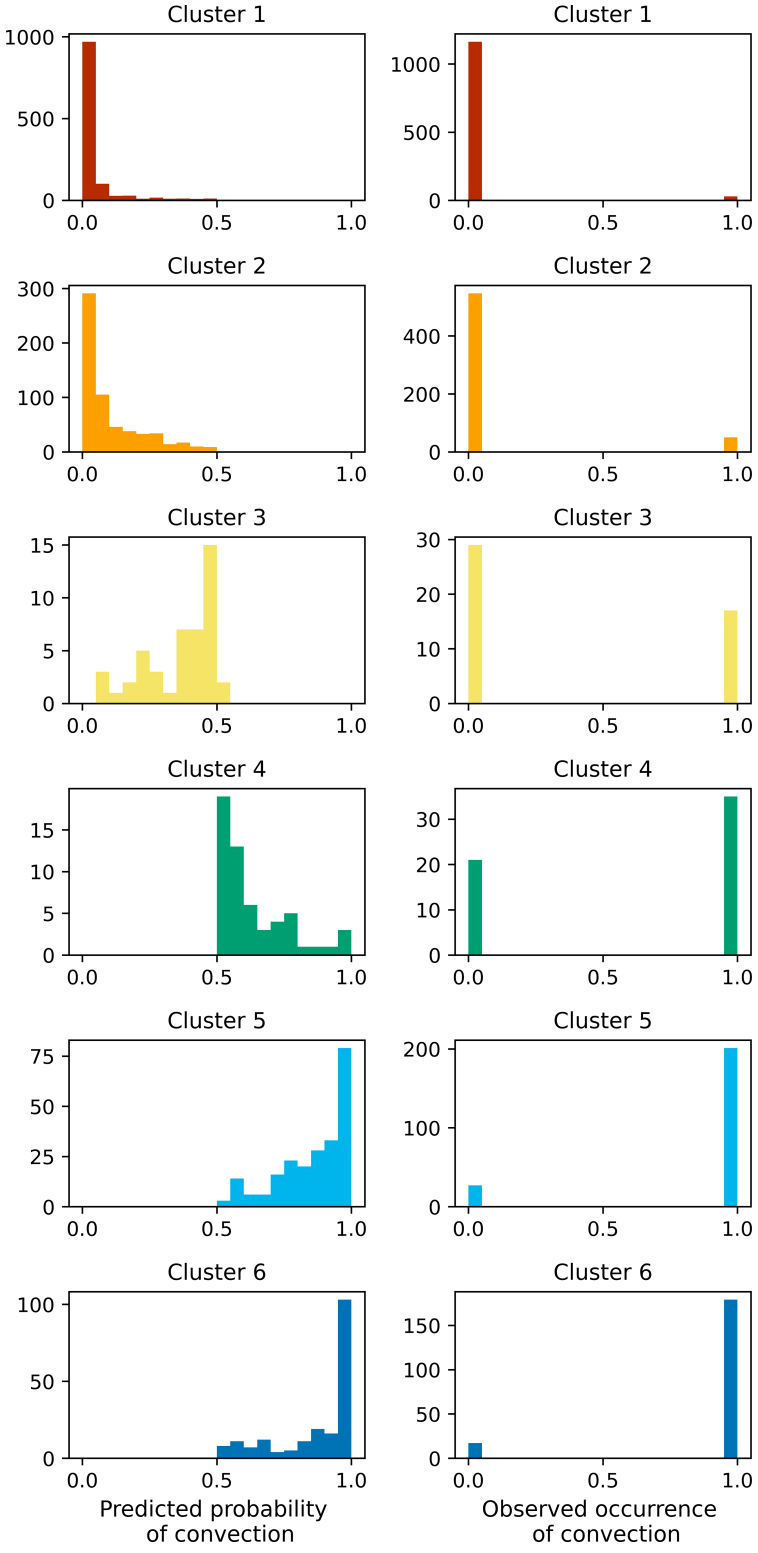

FIG. 9. Histograms of (left) the unscaled predicted probability of convection and (right) the observed occurrence of convection for each of the clusters. The clustering created distinct convective and nonconvective groups, with clusters mostly all above or below a probability of 0.5.

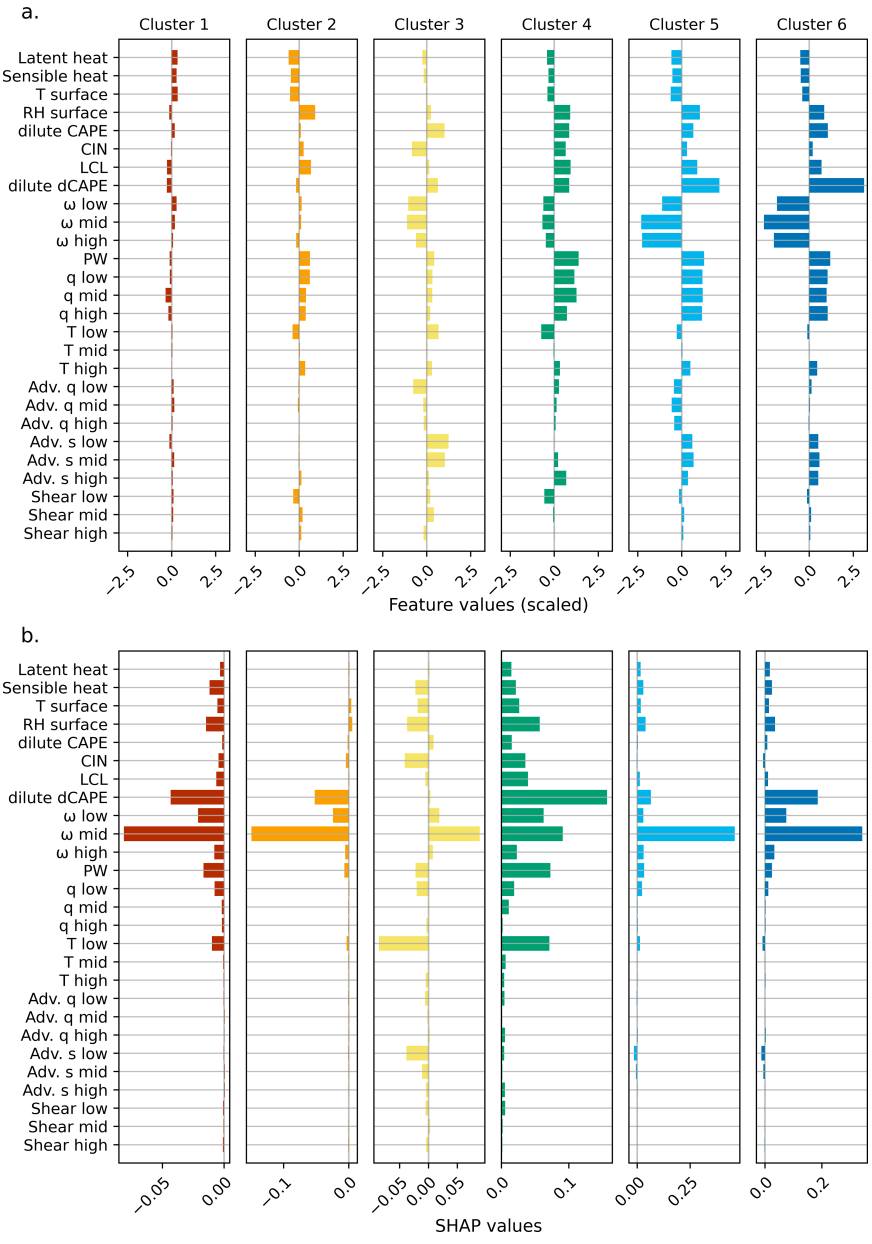

FIG. 10. Average characteristics for each cluster. (a) The feature values (i.e., the actual atmospheric measurements that have been scaled) and (b) the corresponding SHAP values (i.e., the marginal contribution of each feature to the predicted probability). The magnitude of the SHAP values for each cluster varies, as shown by the different *x*-axis scales in (b).

To sum up, using ML and SHAP value clustering, distinct convective groups can be produced to cluster atmospheric events into certain and uncertain, convective and nonconvective events. These results show that strong upward vertical velocity helps to predict convective events with greater certainty. If the vertical velocity is not strongly upward, then it is much less certain whether convection will occur; for these cases, other features such as precipitable water, low-level temperature, or surface relative humidity are essential in indicating whether events are likely convective or nonconvective.

4. Discussion

Overall, the probabilistic random forest trigger outperforms conventional triggers and has similar performance to other nonprobabilistic ML models. However, while the ARM variational analysis training dataset offered the advantage that it was consistent with precipitation observations, it was limited in that it only represents a single location and is likely not generalizable to other regions. The simplistic precipitation threshold over the large region of approximately 3° may not have captured the convective events thoroughly; it likely

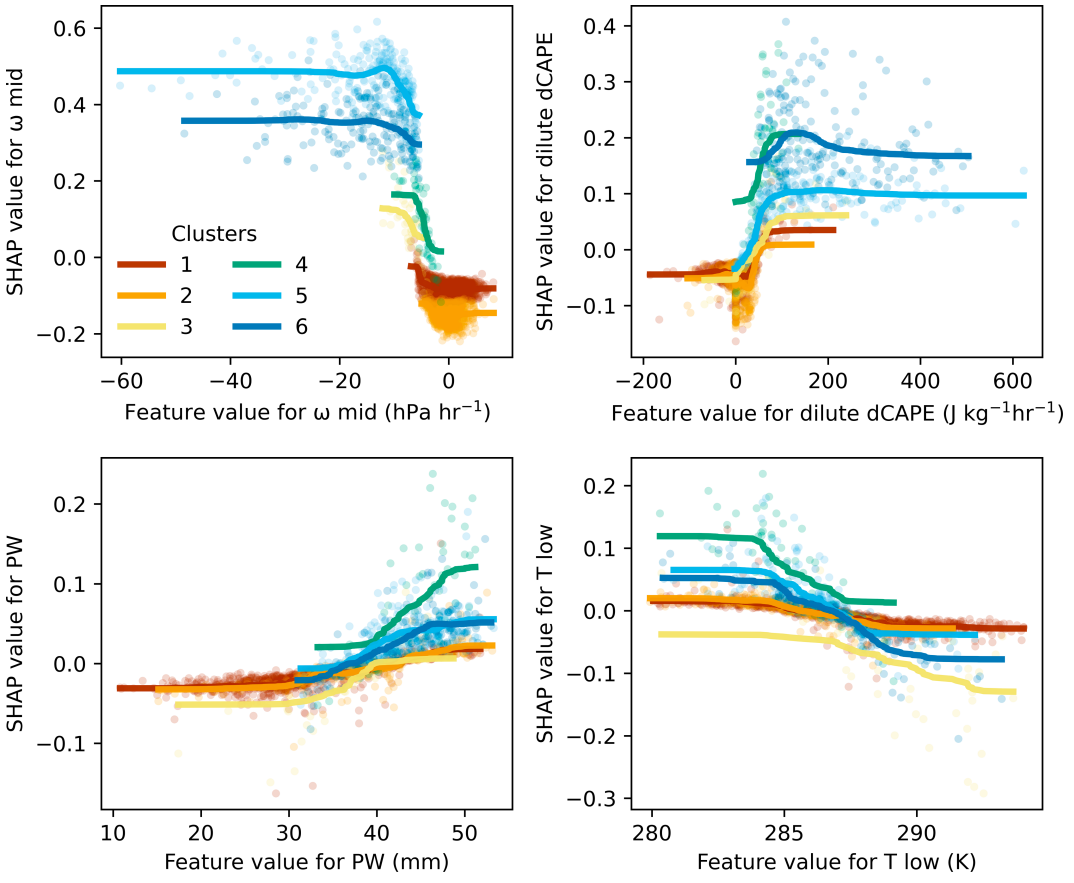

FIG. 11. SHAP dependence plots for vertical velocity (ω mid), dilute dCAPE, precipitable water (PW), and low-level temperature (T low) for each of the clusters. The SHAP dependence plots for ω mid and dilute dCAPE show thresholding behavior, in which events with high vertical velocity magnitudes and high dCAPE have a large positive marginal contribution to the predicted probability.

captured larger convective systems but limited the ability of the model to capture smaller convective storms. These limitations will be explored in future work, using higher-resolution, spatial datasets, and more refined definitions of convection occurrence, for the probabilistic trigger to be able to perform well globally.

The explainability methods in this paper are specific to the given dataset; the feature rankings do not inform us about the best overall features but rather the best features among the ones available. Additionally, correlations among features can dilute the feature importance and relevance values of correlated features, but the grouped feature rankings help to reduce misinterpreting the feature rankings due to correlations.

The probabilistic trigger in this study, with its well-represented diurnal cycle and robust reliability curve, could be a promising direction for improving weather and climate models. For example, the drizzle problem in GCMs may be alleviated with a probabilistic trigger that better captures the diurnal cycle. Probabilistic predictions of precipitation in weather forecasting may become more reliable with a more reliable convection trigger.

The ML approach used in this study allows us to consider a wide range of input features to select the most important ones. The feature importance results generally agree with previous studies, which show large-scale vertical velocity to be an important predictor of convection (Birch et al. 2014; Peters et al. 2013). Vertical velocity is indicative of low-level convergence, which can dynamically initiate convection. Convergence can also increase low-level moisture, which can help to initiate and sustain convection. However, vertical velocity can also be a result of existing convection, and this study does not establish a clear cause-and-effect relationship between vertical velocity and convection.

The clustering approach shows results that are physically consistent with previous studies but also goes well beyond as we find clusters that are not well represented by typical triggers (e.g., dCAPE) alone. The characteristic clusters found in this study could be linked to different types of convection. Future work will explore how certain and uncertain regimes correspond to different types of convection, such as synoptic-scale systems, which tend to be easier to predict an hour in advance or local instabilities that may be less predictable.

This study shows the potential for a probabilistic ML trigger trained on a global convection dataset to be implemented as a stochastic trigger in an operational model. To do this, convection would trigger in each time step if a random uniform number is

less than the predicted probability from the ML trigger for the given large-scale atmospheric state, unlike a deterministic thresholding approach. The random number field would have spatial and temporal coherence, which also provides an opportunity to represent memory in the convective life cycle. Implementing the trigger operationally would test the online performance, which is generally more challenging than achieving good offline performance.

5. Conclusions

In this study, we applied explainable ML methods with a probabilistic random forest to better understand how to predict the occurrence of deep atmospheric convection. The model was trained using atmospheric inputs from the variational analysis over the Southern Great Plains. Clustering of SHAP values derived from the random forest was used to interpret model uncertainty.

The model skill, as measured by precision, recall, F1 score, and Brier skill score, was better than conventional CAPE-based triggers. This may indicate that considering more atmospheric variables than a single CAPE-based threshold allows the trigger to better identify convective events. Vertical velocity in the midtroposphere was found to be the most important feature in the random forest, which suggests that the large-scale upward motion due to convergence plays an important role in triggering atmospheric instability. Dilute dCAPE, which was found to be the best convection trigger in previous studies, primarily relies on vertical velocity to classify convective events but is less important than vertical velocity in the random forest.

Clustering of the SHAP values produced groupings of certain and uncertain, nonconvective and convective events. Certain convective events were characterized by large upward vertical velocities, while certain nonconvective events were characterized by near-zero or downward vertical velocities. Uncertain events had smaller upward vertical velocities but could be differentiated into likely convective or likely nonconvective events using other features such as precipitable water, low-level temperature, or surface relative humidity.

This study showed that a probabilistic ML convection trigger may have the potential to be used in operational models to significantly improve predictions of convection. Since a given large-scale atmospheric state could correspond to multiple subgrid convective (or nonconvective) states, the probabilistic framework used in this study allows for a better understanding of uncertain atmospheric states. A probabilistic trigger like the one used in this study could be implemented as a stochastic parameterization in the future.

Acknowledgments. G. A. M. acknowledges the support from the U.K. National Environmental Research Council Award NE/S007474/1 and the University of Oxford Clarendon Fund. P. S. acknowledges the support from the FORCeS project under the European Union's Horizon 2020 research program with Grant Agreement 821205 and the CleanCloud project under the European Union's Horizon Europe research program and its UKRI underwrite. H. M. C. acknowledges the support from the U.K. Natural Environment Research Council Grant NE/P018238/1, a Leverhulme Trust Research Leadership Award, and the European Union EERIE project (Grant Agreement 101081383). The University of Oxford's contribution to EERIE is funded by U.K. Research and Innovation (UKRI) under the U.K. government's Horizon Europe funding guarantee (Grant 10049639). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Climate Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. We are also grateful to three anonymous reviewers for their constructive comments, which significantly improved this manuscript.

Data availability statement. The ARM variational analysis dataset used for this research is available at https://www.arm.gov/capabilities/science-data-products/vaps/varanal. The Python code for the machine learning training and evaluation in this study is available at https://github.com/gretamiller/convection_trigger_SGP.

APPENDIX A

Random Forest Hyperparameters

The random forest has several hyperparameters controlling the architecture setup that can be tuned to a given dataset, shown in Table A1. Bayesian optimization was used to tune the model hyperparameters [using the Scikit-optimize package (Pedregosa et al. 2011)]. We used four-fold cross-validation on the training set, splitting the 12-yr training dataset into evaluation sets of three consecutive years to avoid any issues with temporal correlations. Like the original random forest, the depth of the trees was not limited (Breiman 2001).

TABLE A1. Optimized random forest hyperparameters. Note: Hyperparameter descriptions from Scikit-learn ML library for Python (Pedregosa et al. 2011).

Hyperparameter	Description	Search range	Optimized value
Tree density	Number of individual trees that are averaged in the random forest	50–300	288
Max features	Maximum number of environmental inputs that each tree can use	4–12	9
Min samples split	Minimum number of samples required in a node to do another decision tree split	2–4	3
Min samples leaf	Minimum number of samples that are required to be at a node	1–4	3

APPENDIX B

Probability Scaling for Undersampled Dataset

To reduce the class imbalance, majority class undersampling was used, in which the majority class was randomly undersampled by a factor of $1/r$. Various undersampling ratios were tested, and undersampling by a factor of $r = 3$ (with a convective to nonconvective ratio of 0.285) was chosen to optimize performance metrics on both the undersampled and original training datasets, using the same k-fold cross-validation of the hyperparameter optimization. Since undersampling changes the frequency of positive events, the probabilities need to account for this change (Miloshevich et al. 2023).

Following the methods in Miloshevich et al. (2023), let $p_0(x)$ represent the probability of a nonconvective event given that $X = x$ in the original dataset. Then, the probability of a convective event is $p_1(x) = 1 - p_0(x)$. In the undersampled dataset, the probabilities of nonconvection $[p'_0(x)]$ and convection $[p'_1(x)]$ are given as

$$p'_0(x) = \frac{\frac{1}{r}p_0(x)}{\frac{1}{r}p_0(x) + p_1(x)}, \tag{B1}$$

and

$$p'_1(x) = \frac{p_1(x)}{\frac{1}{r}p_0(x) + p_1(x)}. \tag{B2}$$

To get an estimate of the true probabilities, the equations above can be inverted to give

$$p_0(x) = \frac{rp'_0(x)}{1 - (1 - r)p'_0(x)}, \tag{B3}$$

and

$$p_1(x) = \frac{p'_1(x)}{r + (1 - r)p'_1(x)}. \tag{B4}$$

The raw probabilities from the random forest predictions can then be tested on an original test set (not undersampled) after the probability rescaling [Eqs. (B3) and (B4)].

APPENDIX C

k-Means Clustering

After training the random forest, the Shapley additive explanations (SHAP) values were clustered using k-means clustering from the Scikit-learn ML library for Python (Pedregosa et al. 2011). The elbow method was used to determine the optimal number of clusters (Kodinariya and Makwana 2013). Figure C1 shows the within-cluster variance against the number of clusters. The number of clusters used in this study, $n = 6$, was chosen where the decline in variance flattens with an increasing number of clusters.

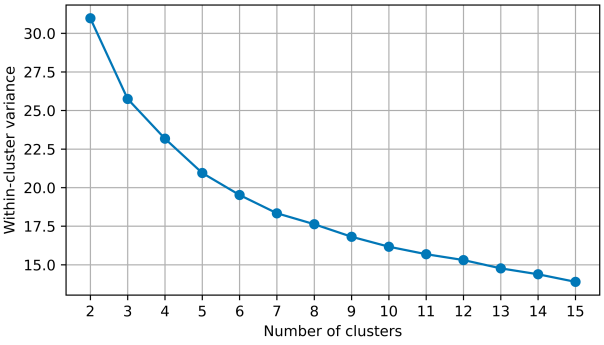

FIG. C1. Elbow curve plot for the k-means clustering of the SHAP values. Six clusters were chosen for the analysis since the within-cluster variance flattens as the number of clusters increases beyond six clusters.

REFERENCES

Bechtold, P., E. Bazile, F. Guichard, P. Mascart, and E. Richard, 2001: A mass-flux convection scheme for regional and global models. *Quart. J. Roy. Meteor. Soc.*, **127**, 869–886, https://doi.org/10.1002/qj.49712757309.

Behrens, G., T. Beucler, F. Iglesias-Suarez, S. Yu, P. Gentine, M. Pritchard, M. Schwabe, and V. Eyring, 2024: Improving atmospheric processes in earth system models with deep learning ensembles and stochastic parameterizations. *arXiv*, 2402.03079, https://doi.org/10.48550/arXiv.2402.03079.

Berner, J., and Coauthors, 2017: Stochastic parameterization toward a new view of weather and climate models. *Bull. Amer. Meteor. Soc.*, **98**, 565–588, https://doi.org/10.1175/BAMS-D-15-00268.1.

Birch, C. E., J. H. Marsham, D. J. Parker, and C. M. Taylor, 2014: The scale dependence and structure of convergence fields preceding the initiation of deep convection. *Geophys. Res. Lett.*, **41**, 4769–4776, https://doi.org/10.1002/2014GL060493.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Bright, D. R., and S. L. Mullen, 2002: Short-range ensemble forecasts of precipitation during the Southwest monsoon. *Wea. Forecasting*, **17**, 1080–1100, https://doi.org/10.1175/1520-0434(2002)017<1080:SREFOP>2.0.CO;2.

Bryan, G. H., J. C. Wyngaard, and J. M. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection. *Mon. Wea. Rev.*, **131**, 2394–2416, https://doi.org/10.1175/1520-0493(2003)131<2394:RRFTSO>2.0.CO;2.

Chen, M., H. Fu, T. Zhang, and L. Wang, 2023: ResU-Deep: Improving the trigger function of deep convection in tropical regions with deep learning. *J. Adv. Model. Earth Syst.*, **15**, e2022MS003521, https://doi.org/10.1029/2022MS003521.

Christensen, H., and L. Zanna, 2022: Parametrization in weather and climate models. *Oxford Research Encyclopedia of Climate Science*, Oxford University Press, 1–44, https://doi.org/10.1093/acrefore/9780190228620.013.826.

——, S. Kouhen, G. Miller, and R. Parthipan, 2024: Machine learning for stochastic parametrization. *Environ. Data Sci.*, **3**, e38, https://doi.org/10.1017/eds.2024.45.

Covert, I., S. Lundberg, and S.-I. Lee, 2020: Understanding global feature contributions with additive importance measures. arXiv, 2004.00668v2, https://doi.org/10.48550/arXiv.2004.00668.

Covey, C., P. J. Gleckler, C. Doutriaux, D. N. Williams, A. Dai, J. Fasullo, K. Trenberth, and A. Berg, 2016: Metrics for the diurnal cycle of precipitation: Toward routine benchmarks for climate models. *J. Climate*, **29**, 4461–4471, https://doi.org/10.1175/JCLI-D-15-0664.1.

Cui, Z., G. J. Zhang, Y. Wang, and S. Xie, 2021: Understanding the roles of convective trigger functions in the diurnal cycle of precipitation in the NCAR CAM5. *J. Climate*, **34**, 6473–6489, https://doi.org/10.1175/JCLI-D-20-0699.1.

Dai, A., 2006: Precipitation characteristics in eighteen coupled climate models. *J. Climate*, **19**, 4605–4630, https://doi.org/10.1175/JCLI3884.1.

Flora, M. L., C. K. Potvin, A. McGovern, and S. Handler, 2024: A machine learning explainability tutorial for atmospheric sciences. *Artif. Intell. Earth Syst.*, **3**, e230018, https://doi.org/10.1175/AIES-D-23-0018.1.

Hartmann, D. L., 2016: Tropical anvil clouds and climate sensitivity. *Proc. Natl. Acad. Sci. USA*, **113**, 8897–8899, https://doi.org/10.1073/pnas.1610455113.

Hohenegger, C., and Coauthors, 2023: ICON-Sapphire: Simulating the components of the Earth system and their interactions at kilometer and subkilometer scales. *Geosci. Model Dev.*, **16**, 779–811, https://doi.org/10.5194/gmd-16-779-2023.

Jones, W. K., M. Stengel, and P. Stier, 2024: A Lagrangian perspective on the lifecycle and cloud radiative effect of deep convective clouds over Africa. *Atmos. Chem. Phys.*, **24**, 5165–5180, https://doi.org/10.5194/acp-24-5165-2024.

Kain, J. S., and J. M. Fritsch, 1990: A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.*, **47**, 2784–2802, https://doi.org/10.1175/1520-0469(1990)047<2784:AODEPM>2.0.CO;2.

Kodinariya, T. M., and P. R. Makwana, 2013: Review on determining number of cluster in *K*-means clustering. *Int. J. Adv. Res. Comput. Sci. Manage. Stud.*, **1**, 90–95.

Lundberg, S. M., and S.-I. Lee, 2017: A unified approach to interpreting model predictions. *Proc. 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, CA, Curran Associates, Inc., 4768–4777, https://dl.acm.org/doi/10.5555/3295222.3295230.

——, and Coauthors, 2020: From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, **2**, 56–67, https://doi.org/10.1038/s42256-019-0138-9.

Miloshevich, G., B. Cozian, P. Abry, P. Borgnat, and F. Bouchet, 2023: Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data. *Phys. Rev. Fluids*, **8**, 040501, https://doi.org/10.1103/PhysRevFluids.8.040501.

Nadiga, B. T., X. Sun, and C. Nash, 2022: Stochastic parameterization of column physics using generative adversarial networks. *Environ. Data Sci.*, **1**, e22, https://doi.org/10.1017/eds.2022.32.

Neale, R. B., J. H. Richter, and M. Jochum, 2008: The impact of convection on ENSO: From a delayed oscillator to a series of events. *J. Climate*, **21**, 5904–5924, https://doi.org/10.1175/2008JCLI2244.1.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Peters, K., C. Jakob, L. Davies, B. Khouider, and A. J. Majda, 2013: Stochastic behavior of tropical convection in observations and a multicloud model. *J. Atmos. Sci.*, **70**, 3556–3575, https://doi.org/10.1175/JAS-D-13-031.1.

Plant, R. S., and J.-I. Yano, 2015: *Parameterization of Atmospheric Convection.* Series on the Science of Climate Change, Vol. 1, Imperial College Press, 1172 pp., https://doi.org/10.1142/p1005.

Rochetin, N., J.-Y. Grandpeix, C. Rio, and F. Couvreux, 2014: Deep convection triggering by boundary layer thermals. Part II: Stochastic triggering parameterization for the LMDZ GCM. *J. Atmos. Sci.*, **71**, 515–538, https://doi.org/10.1175/JAS-D-12-0337.1.

Schmit, T. J., P. Griffith, M. M. Gunshor, J. M. Daniels, S. J. Goodman, and W. J. Lebair, 2017: A closer look at the ABI on the GOES-R series. *Bull. Amer. Meteor. Soc.*, **98**, 681–698, https://doi.org/10.1175/BAMS-D-15-00230.1.

Song, F. F., and G. J. Zhang, 2017: Improving trigger functions for convective parameterization schemes using GOAmazon observations. *J. Climate*, **30**, 8711–8726, https://doi.org/10.1175/JCLI-D-17-0042.1.

——, and ——, 2018: Understanding and improving the scale dependence of trigger functions for convective parameterization using cloud-resolving model data. *J. Climate*, **31**, 7385–7399, https://doi.org/10.1175/JCLI-D-17-0660.1.

Song, Y., C. K. Wikle, C. J. Anderson, and S. A. Lack, 2007: Bayesian estimation of stochastic parameterizations in a numerical weather forecasting model. *Mon. Wea. Rev.*, **135**, 4045–4059, https://doi.org/10.1175/2007MWR1928.1.

Suhas, E., and G. J. Zhang, 2014: Evaluation of trigger functions for convective parameterization schemes using observations. *J. Climate*, **27**, 7647–7666, https://doi.org/10.1175/JCLI-D-13-00718.1.

Tang, S., C. Tao, S. Xie, and M. Zhang, 2019: Description of the ARM large-scale forcing data from the constrained Variational Analysis (VARANAL) version 2. US DOE Office of Science Tech. Rep. DOE/SC-ARM-TR-222, 33 pp., https://www.arm.gov/publications/tech_reports/doe-sc-arm-tr-222.pdf.

Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Mon. Wea. Rev.*, **117**, 1779–1800, https://doi.org/10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2.

Ukkonen, P., and A. Mäkelä, 2019: Evaluation of machine learning classifiers for predicting deep convection. *J. Adv. Model. Earth Syst.*, **11**, 1784–1802, https://doi.org/10.1029/2018MS001561.

Wang, Y., G. J. Zhang, and G. C. Craig, 2016: Stochastic convective parameterization improving the simulation of tropical precipitation variability in the NCAR CAM5. *Geophys. Res. Lett.*, **43**, 6612–6619, https://doi.org/10.1002/2016GL069818.

Xie, S., and M. Zhang, 2000: Impact of the convection triggering function on single-column model simulations. *J. Geophys. Res.*, **105**, 14 983–14 996, https://doi.org/10.1029/2000JD900170.

——, ——, J. S. Boyle, R. T. Cederwall, G. L. Potter, and W. Lin, 2004: Impact of a revised convective triggering mechanism on Community Atmosphere Model, Version 2, simulations: Results from short-range weather forecasts. *J. Geophys. Res.*, **109**, D14102, https://doi.org/10.1029/2004JD004692.

——, and Coauthors, 2019: Improved diurnal cycle of precipitation in E3SM with a revised convective triggering function. *J. Adv. Model. Earth Syst.*, **11**, 2290–2310, https://doi.org/10.1029/2019MS001702.

Zhang, G. J., and N. A. McFarlane, 1995: Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian climate centre general circulation model. *Atmos.-Ocean*, **33,** 407–446, https://doi.org/10.1080/07055900.1995.9649539.

Zhang, T., W. Lin, A. M. Vogelmann, M. Zhang, S. Xie, Y. Qin, and J.-C. Golaz, 2021: Improving convection trigger functions in deep convective parameterization schemes using machine learning. *J. Adv. Model. Earth Syst.*, **13**, e2020MS002365, https://doi.org/10.1029/2020MS002365.

Zhao, W., R. Marchand, and Q. Fu, 2017: The diurnal cycle of clouds and precipitation at the ARM SGP site: Cloud radar observations and simulations from the multiscale modeling framework. *J. Geophys. Res. Atmos.*, **122**, 7519–7536, https://doi.org/10.1002/2016JD026353.