Data-driven Modeling of Stratiform and Convective Rain Area

Sara Shamekh[a] Pedro Angulo-Umana [b] Paul A. O'Gorman [c]

[a] *Courant Institute, New York University, New York*

[b] *School of Earth and Environmental Sciences, Seoul National University, Seoul, South Korea*

[c] *Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge*

Corresponding author: Sara Shamekh, ss18284@nyu.edu

ABSTRACT: Understanding how the fractional area of different rain types, including shallow, deep convective, and stratiform, varies with large-scale environmental conditions is fundamental to advancing our knowledge of moist atmospheric processes. Yet this relationship remains poorly quantified, especially at instantaneous timescales where observational uncertainty and inherent variability pose challenges. In this study, we tackle this problem by integrating satellite-derived rain area data from TRMM with environmental variables from ERA5, using a two-step framework that combines conditional normalizing flows (CNFs) with symbolic equation discovery. The CNF model reconstructs the full conditional joint distribution of three rain area types, capturing their co-dependence and reproducing observed features such as the bimodal relationship between convective and stratiform area. From these distributions, we extract the conditional means and then apply equation discovery to derive compact analytical expressions that relate rain area fractions to large-scale thermodynamic variables. The resulting equations reveal interpretable and physically consistent predictors, such as lower-tropospheric stability, free-tropospheric moisture deficit, and boundary-layer buoyancy, that act as both triggering thresholds and predictors of rain area magnitude. These expressions capture well-known mechanisms in closed form, opening a path toward stochastic, physically grounded parameterizations. Our phase-space analysis clarifies how rain types evolve across sea surface temperatures and stability regimes, linking environmental variability to the spatial extent of convective systems—crucial for precipitation prediction and cloud–radiation interactions. Together, this framework provides a probabilistically robust, physically based approach for modeling rain area, with implications for both process understanding and parameterization.

1. Introduction

Accurately predicting precipitation remains a major challenge in atmospheric science, despite significant advances in modeling and computational power (Dai 2006; Wilcox and Donner 2007; Flato et al. 2014; Kooperman et al. 2018). A key reason is that precipitation, particularly from convective systems, is highly sensitive to subgrid-scale processes such as turbulence (Tompkins 2001b; Zuidema et al. 2017; Windmiller and Hohenegger 2019; Lochbihler et al. 2021; Shamekh and Gentine 2023; Orenstein et al. 2025), convective organization (Muller et al. 2022; Shamekh et al. 2023; Colin et al. 2019; Angulo-Umana and Kim 2023), and cloud microphysics (Singh and O'Gorman 2014; Charn et al. 2020; Morrison et al. 2020), all of which are difficult to explicitly resolve in global climate models (GCMs). These challenges are amplified at fast time scales (e.g., instantaneous or sub-hourly), where precipitation is more stochastic and sensitive to unresolved variability. Yet this time scale is particularly important for extremes and impacts (Lenderink et al. 2017). Conventional parameterizations primarily focus on vertical mass fluxes (Arakawa and Schubert 1974), such as those aimed at consuming Convective Available Potential Energy (CAPE), but often neglect the horizontal structure or spatial extent of precipitation.

However, recent observational evidence suggests that changes in total precipitation are more strongly tied to variations in rain area, encompassing both convective and stratiform regions, than to changes in rain intensity (defined as precipitation amount per unit area; Ahmed and Schumacher 2015; Zipser and Liu 2021). These findings underscore the need to better understand what controls the spatial extent of precipitation. While convective area is relatively well understood in radiative-convective equilibrium, (RCE; Manabe and Strickler 1964; Tompkins and Craig 1998), where radiative cooling constrains mass flux and thus convective area if the updraft velocity is unchanged (Agasthya et al. 2025; Robe and Emanuel 1996), the situation is more complex at the smaller spatial and temporal scales relevant to GCM grid cells and fast-time-scale variability. In these regimes, the atmosphere is far from equilibrium, and similar environmental conditions can give rise to very different convective activities due to unresolved processes and internal variability. This raises the question: can we derive predictive models for rain area at fast time scales where the systems are far from equilibrium?

To investigate this question, we use a generative machine learning model, specifically, a conditional normalizing flow (CNF) (Rezende and Mohamed 2015; Kobyzev et al. 2020; Winkler et al.

2019), to estimate the joint distribution of shallow convective, deep convective, and stratiform rain areas given the large-scale environmental variables. This approach captures not only the mean but also the variance, covariance, and higher-order structure of rain areas within $2° \times 2°$ domains, where similar large-scale states can produce diverse outcomes. Unlike bin-averaging techniques, which may obscure variability and introduce bias, the CNF enables conditional sampling, producing a range of plausible rain area configurations consistent with a given large-scale environment. Each sample can be viewed as a synthetic snapshot, akin to what a satellite might observe during a single overpass, providing a physically plausible realization of shallow, deep, and stratiform rain areas under specific atmospheric conditions.

To enable interpretability, we use a concise set of physically motivated predictors derived from ERA5 reanalysis (Hersbach et al. 2020), including precipitable water, equivalent potential temperature, CAPE, vertical velocity at 500 hPa, and integrated or averaged humidity and temperature in the boundary layer and free troposphere (Bretherton et al. 2004; Neelin et al. 2009; Muller et al. 2011). Precipitation data is taken from NASA's Tropical Rainfall Measuring Mission (TRMM) gridded rain-type classification product (2A23, Version 7; Houze Jr et al. 2015; Mission 2011), which distinguishes between deep convective, stratiform, and shallow rain (Awaka et al. 2009). Our dataset comprises ~300,000 samples from the ocean between 35°S and 35°N.

Finally, we apply equation-discovery methods to the output statistics, such as the mean and standard deviation of rain area types, to derive simple, closed-form expressions linking large-scale atmospheric variables to precipitation area. This approach bridges statistical learning and physical reasoning, offering a step toward more interpretable and physically grounded precipitation modeling.

The paper is organized as follows. Sections 2 and 3 introduce the data and methodology, respectively. Section 4.a presents the CNF-estimated joint distribution of stratiform and convective rain area conditioned on large-scale environmental variables. Section 4.b examines the structure and interpretation of the discovered equations that relate rain area to environmental drivers. Section 4.c analyzes how each term in the discovered equations varies across key environmental gradients, in particular atmospheric stability and sea surface temperature. Finally, Section 5 summarizes the findings and discusses implications for future parameterization development.

2. Data

a. Rain–Type Area Data

Our precipitation data are obtained from instantaneous retrievals by the Tropical Rainfall Measuring Mission (TRMM) satellite. The data are on a regular 0.05°×0.05° grid. In addition to near-surface precipitation rate (TRMM 2A25), we analyze the TRMM 2A23 radar product, which classifies observed tropical precipitation into shallow-convective, deep-convective, and stratiform regimes (Awaka et al. 1997). This classification is done for each rainy pixel. A rainy pixel is identified as belonging to the shallow regime if the storm's height is 1 km or more lower than the estimated 0°C-isotherm; to the deep-convective regime if the maximum radar echo is sufficiently strong, or if it stands out against the mean reflectivity of the surrounding region, and is not shallow; and as stratiform regime if either a so-called "bright-band" echo exists in the vertical radar reflectivity profile, indicating a layer of melting ice particles; or a rainy pixel's maximum radar echo along the bean direction is too weak to be labeled convective (Awaka et al. 2009).

Ten years of data (2000–2010) are extracted over four tropical–subtropical oceanic regions between 35°S–35°N: the Central Indian Ocean (CIO: 55°E–110°E, 35°S–10°N), the Maritime Continent warm pool (WMP: 105°E–178°E, 35°S–10°N), the tropical South America (SAM: 95°W–25°W, 35°S–20°N), and tropical Africa (AFC: 30°W–60°E, 35°S–35°N). In all cases, only ocean grid points within these domains are retained to simplify interpretation, so land areas are excluded. Observations are provided as snapshots along the TRMM satellite swath, which is segmented into $2^\circ \times 2^\circ$ boxes. A box is included if at least 90% of its area is covered by the swath, ensuring near-complete sampling. For each valid box we record both the surface rain amount and the fractional area covered by shallow convective, deep convective, and stratiform rain. Boxes are retained whenever the total precipitation is nonzero, meaning that at least one of the three rain types is present, though the others may be absent. Throughout the paper, rain area is denoted by the vector $\mathbf{A} = (A_{\text{shallow}}, A_{\text{deep}}, A_{\text{stratiform}})$, representing the fractional areas of shallow, deep, and stratiform precipitation, each defined relative to the total area of a $2^\circ \times 2^\circ$ grid box (see Fig. 1). Unless otherwise noted, all figures use the combined dataset from these four oceanic regions.

b. Large-Scale Environmental Fields

Predictor variables come from ERA5 and are coarse-grained to the same $2^\circ \times 2^\circ$ grid by block averaging; that is, by averaging all native ERA5 grid points that fall within each $2^\circ \times 2^\circ$ box. We illustrate the data collection method in Fig. 1. TRMM provides instantaneous snapshots, whereas ERA5 variables may represent either hourly means or instantaneous values, depending on the variable. We linearly interpolate the ERA5 fields to the exact TRMM over-pass times. At this horizontal scale, large-scale variables evolve slowly, so the interpolation error is negligible relative to the smoothing introduced by the common spatial averaging. Any residual timing offset is therefore unlikely to affect the inferred relationships between environment and rain area. Table 1 summarizes the variables used in this study and their data sources.

Simplified two-layer (or few-layer) representations of the atmosphere are commonly used in theoretical and idealized studies of convection to highlight key physical mechanisms (Betts and Miller 1986; Bretherton et al. 2004; Neelin et al. 2009; Muller et al. 2009; Ahmed and Neelin 2018; Abramian et al. 2025). In contrast, most machine learning–based parameterization efforts ingest the full vertical profiles, creating a large set of predictors across all atmospheric levels (Gentine et al. 2018; Yuval and O'Gorman 2020; Mooers et al. 2021; Hafner et al. 2024; Shamekh and Gentine 2023; Falga et al. 2025). While this comprehensive approach is feasible and often beneficial for neural network models, it poses major challenges for equation discovery, because most these methods are sensitive to input dimensionality; as the number of predictors increases, both search efficiency and interpretability deteriorate. To balance physical relevance with tractability, we adopt a two-layer structure that captures the dominant vertical organization of moist processes: (i) the planetary boundary layer (hereafter PBL; surface–900 hPa), and (ii) the free troposphere (hereafter FT; 850–400 hPa).

For each predictor, we compute a mean or vertically integrated value within each layer, reducing $\sim$20 vertical levels to two features per variable (tab. 1). This layer-based compression preserves the essential distinction between surface-coupled and free-tropospheric processes; an abstraction widely used in convection studies. Previous works have emphasized the importance of including information from one or both layers, though their relative roles remain debated. Our approach retains these key physical controls while keeping the input space small enough to enable efficient and interpretable equation discovery.

We choose predictors that theory and past observational studies (as cited above) identify as the primary thermodynamic and dynamic controls on moist convection:

- Integrated column water vapor (precipitable water, PW) and column-integrated relative humidity (CRH, defined as the ratio of PW to the saturation PW of the atmospheric column), each computed over the PBL and the FT. These quantify the ambient moisture availability and its vertical structure.
- Equivalent potential temperature (θ_e) and its saturated counterpart (θ_e^*), averaged over the PBL and the FT, equivalent to entropy or saturated entropy (and similar to moist static energy), used in combinations to measure moist convective instability and moisture.
- Surface temperature T_S and free-tropospheric average temperature (T_{FT}), which set the lower- and middle-level thermal state for convection.
- Vertical velocity at 500 hPa (ω_{500}), a dynamical proxy for large-scale ascent or subsidence.
- Convective available potential energy (CAPE), representing the parcel-integrated buoyant energy available for deep convection. CAPE is computed by lifting parcels of air departing at different model levels from below 350 hPa under pseudo-adiabatic, non-mixing assumptions, with simplified mixed-phase heating, and integrating their positive buoyancy relative to the environment.

We denote the resulting feature vector, which has 12 variables, by $\mathbf{X}_{\mathrm{LS}}$ (see Table 1).

3. Method

a. Motivation and workflow

Our goal is to express rain–area fractions as closed-form functions of the large-scale environmental state, $\mathbf{X}_{\mathrm{LS}}$. This task presents two key challenges. First, the relationship between $\mathbf{X}_{\mathrm{LS}}$ and rain area is inherently stochastic: even for fixed environmental conditions, instantaneous fractional rain area can vary widely due to unresolved subgrid processes and internal variability. This randomness is further amplified by observational sampling uncertainties as depending on the satellite overpass, the instrument may capture the entire convective system or just a portion of it, introducing additional spread in the observed rain area. A common workaround is to bin the

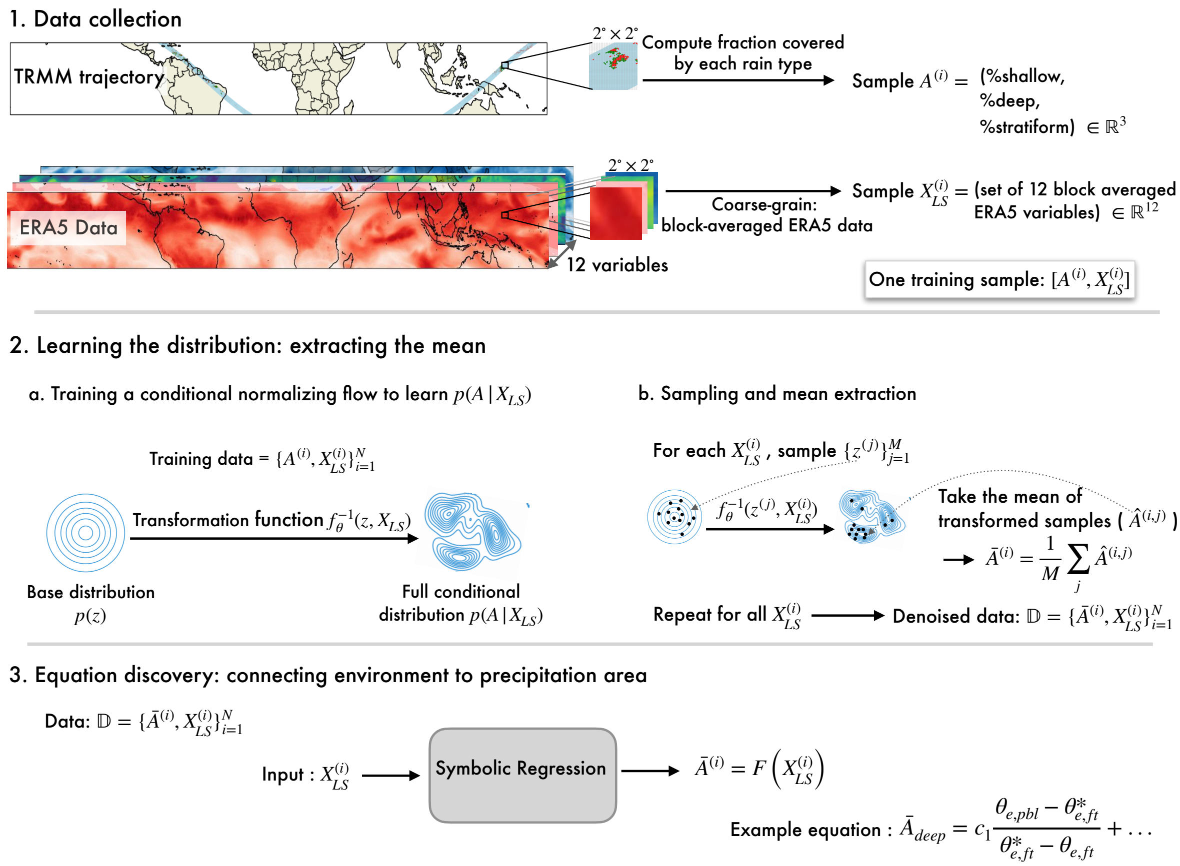

FIG. 1. **A schematic summarizing our methods going from observations to equations. (1) Data collection.** TRMM precipitation swaths are intersected with $2° \times 2°$ boxes. Within each box we compute the fractional area covered by shallow, deep, and stratiform rain, forming a target vector $A^{(i)} \in \mathbb{R}^3$. In parallel, ERA5 fields are coarse-grained to the same blocks and reduced to 12 large-scale variables, yielding a conditioning vector $X_{\rm LS}^{(i)} \in \mathbb{R}^{12}$. Each training example is a pair $\left(A^{(i)}, X_{\rm LS}^{(i)}\right)$. **(2) Learning the distribution and extracting the mean.** A conditional normalizing flow learns the full conditional distribution $p(A \mid X_{\rm LS})$ by transforming samples from a simple base density through a learned map $f_\theta^{-1}(z, X_{\rm LS})$. For each $X_{\rm LS}^{(i)}$, we draw M noise samples, obtain $\hat{A}^{(i,j)} \sim p(A \mid X_{\rm LS}^{(i)})$, and compute the conditional mean $\bar{A}^{(i)} = \frac{1}{M}\sum_{j=1}^{M} \hat{A}^{(i,j)}$. Repeating this for all i yields a noise-free dataset $\{(\bar{A}^{(i)}, X_{\rm LS}^{(i)})\}_{i=1}^{N}$. **(3) Equation discovery.** The extracted pairs are used in symbolic regression to identify interpretable equations of the form $\bar{A} = F(X_{\rm LS})$, linking large-scale environment to mean precipitation area. For clarity, the distributions in panel (2) are drawn in two dimensions; in practice, A is three-dimensional and the flow learns a distribution in $\mathbb{R}^3$.

data by one or two chosen predictors and fit the conditional mean within each bin. However, this strategy discards valuable cross-variable information, anchors the analysis to the chosen axes, and becomes unreliable in regions of sparse sampling. Alternatively, attempting to learn deterministic functional relationships directly from individual rain–area realizations would amount to fitting models to stochastic scatter rather than to reproducible conditional structure, resulting in unstable or poorly constrained mappings.

Second, the dimensionality of the predictor space poses a major limitation for equation discovery. While we reduce the vertical complexity of the data by averaging variables over two physically motivated layers, the input space still contains a dozen environmental variables. As the number of predictors increases, the equation discovery search space expands rapidly, making it harder to extract efficient and interpretable expressions.

To address these challenges, as illustrated in Fig. 1, we adopt a two-stage strategy. First, we train a conditional normalizing flow to learn the full probability distribution of the three rain type areas, conditioned on $\mathbf{X}_{\mathrm{LS}}$. Sampling from this model yields smooth, noise-resistant estimates of the distribution's moments, such as means, variances, and covariances, replacing highly scattered and stochastic rain-area data with stable conditional statistics that are robust to sampling noise. Second, we apply symbolic regression to the flow-derived moments. This separation allows the symbolic regression step to focus on extracting compact, interpretable relationships from the denoised data, rather than attempting to fit deterministic equations to inherently noisy realizations. Beyond its role in enabling equation discovery, the CNF-based conditional distributions are valuable in their own right, providing a flexible framework for sampling, uncertainty quantification, and diagnostic analysis of rain-area variability under fixed large-scale conditions.

To maintain tractability and improve interpretability, we avoid exposing all twelve predictors simultaneously. Instead, we conduct five parallel experiments, each retaining a shared thermal backbone consisting of surface temperature, and free-tropospheric temperature. To this core set, each experiment adds moisture- or buoyancy-related variables, represented consistently in both the boundary layer and the free troposphere:

$$\left\{\mathrm{PW}_{\mathrm{ft}}, \mathrm{PW}_{\mathrm{pbl}}\right\},\ \left\{\theta_{e,\mathrm{ft}}, \theta^*_{e,\mathrm{ft}}, \theta_{e,\mathrm{pbl}}, \theta^*_{e,\mathrm{pbl}}\right\},\ \left\{\mathrm{CRH}_{\mathrm{ft}}, \mathrm{CRH}_{\mathrm{pbl}}\right\},\ \left\{\Delta\theta_{e,\mathrm{ft}}, \Delta\theta_{e,\mathrm{pbl}}\right\},\ \left\{\theta_B, \theta_L\right\}.$$

We define θ_e deficits relative to saturation as

$$\Delta\theta_{e,\mathrm{ft}} = \theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}}, \qquad \Delta\theta_{e,\mathrm{pbl}} = \theta^*_{e,\mathrm{pbl}} - \theta_{e,\mathrm{pbl}}.$$

θ_B and θ_L are defined following Ahmed and Neelin (2018):

$$\theta_B = \frac{\theta_{e,\mathrm{pbl}} - \theta^*_{e,\mathrm{ft}}}{\theta^*_{e,\mathrm{ft}}}, \qquad \theta_L = \frac{\theta_{e,\mathrm{ft}} - \theta^*_{e,\mathrm{ft}}}{\theta^*_{e,\mathrm{ft}}}.$$

Note that Ahmed and Neelin (2018) average θ_e over the lower free troposphere (850–600 hPa), a thinner layer than the 850–400 hPa slab used here.

In complementary tests, we also examine the role of ω_{500}, as our dynamical predictor, to evaluate its contribution relative to the thermodynamic variables. These results are discussed in Appendix B.

Applying symbolic regression to these targeted subsets disentangles the influence of strongly correlated moisture variables and reveals which set delivers the simplest yet most accurate relationships. In combination with the generative step, this approach preserves stochastic information, suppresses sampling noise, and translates the reduced predictor space into concise, physically interpretable equations.

b. Generative machine learning for rain-area distribution appoximation

Rain-area exhibits substantial variability even under similar large-scale environmental conditions, motivating a probabilistic rather than deterministic prediction framework. We therefore aim to approximate the full distribution of rain-area, rather than a single value estimate, and model the conditional distribution $p(\mathbf{A} \mid \mathbf{X}_{\mathrm{LS}})$ using a conditional normalizing flow (CNF; Kobyzev et al. 2020; Rezende and Mohamed 2015; Winkler et al. 2019).

Let

$$\mathbf{A} = \left(A_{\mathrm{shallow}}, A_{\mathrm{deep}}, A_{\mathrm{stratiform}}\right) \in \mathbb{R}^3$$

denote the observed vector of rain areas associated with the three rain types, and let $\mathbf{X}_{\mathrm{LS}} \in \mathbb{R}^{12}$ denote the vector of large-scale environmental variables, defined in Section 2. The training dataset consists of N paired observations $\{(\mathbf{A}^{(i)}, \mathbf{X}^{(i)}_{\mathrm{LS}})\}_{i=1}^{N}$, where $\mathbf{A}^{(i)}$ denotes one observed instance of

Symbol	Name	Source	Layer	Units
A_{shallow}	Shallow precipitation area fraction	TRMM	Column (Retrieval)	–
A_{deep}	Deep convective precipitation area fraction	TRMM	Column (Retrieval)	–
$A_{\text{stratiform}}$	Stratiform precipitation area fraction	TRMM	Column (Retrieval)	–
PW_{pbl}	Precipitable water (total column water vapor)	ERA5	PBL integrated	kg m^{-2}
PW_{ft}	Precipitable water (total column water vapor)	ERA5	FT integrated	kg m^{-2}
CRH_{pbl}	Column relative humidity	ERA5	PBL integrated	–
CRH_{ft}	Column relative humidity	ERA5	PBL integrated	–
$\theta_{e,pbl}$	Equivalent potential temperature	ERA5	PBL mean	K
$\theta^*_{e,pbl}$	Saturated equiv. potential temperature	ERA5	PBL mean	K
$\theta_{e,\text{ft}}$	Equivalent potential temperature	ERA5	FT mean	K
$\theta^*_{e,\text{ft}}$	Saturated equiv. potential temperature	ERA5	FT mean	K
T_s	Near-surface air temperature (2 m)	ERA5	Surface	K
T_{ft}	Free-tropospheric temperature	ERA5	FT mean	K
ω_{500}	Vertical velocity at 500 hPa	ERA5	500 hPa level	Pa s^{-1}
SST	Sea surface temperature	ERA5	Surface (ocean)	K
CAPE	Convective available potential energy	ERA5	Column (diagnostic)	J kg^{-1}
$\Delta\theta_{e,ft}$	Free-tropospheric θ_e deficit (e.g., $\theta^*_{e,ft} - \theta_{e,ft}$)	ERA5	FT mean	K

TABLE 1. List of rain fraction variables obtained from TRMM retrievals, and large scale atmospheric variables used in the analysis and obtained from the ERA5 reanalysis. All ERA5 variables are interpolated to TRMM overpass time. PBL denotes planetary boundary-layer averages; FT denotes free-tropospheric averages above the diagnosed PBL (exact pressure bounds follow the Methods).

rain-area vector from TRMM and $\mathbf{X}_{\text{LS}}^{(i)}$ denotes the corresponding large-scale environmental state from ERA5.

Our objective is to approximate the true conditional distribution $p(\mathbf{A} \mid \mathbf{X}_{\text{LS}})$ with a parametric model $p_\phi(\mathbf{A} \mid \mathbf{X}_{\text{LS}})$. The true distribution is unknown and accessible only through a finite set of observed rain-area samples. Learning an accurate approximation is desirable for generating new rain-area realizations that are statistically consistent with the observed variability under a given large-scale state. Such realizations allow us to characterize conditional variability and to compute statistics beyond point estimates.

Constructing this approximation is challenging because the conditional distribution $p(\mathbf{A} \mid \mathbf{X}_{\text{LS}})$ is multi-dimensional and cannot be easily expressed in closed form. The role of the CNF is to provide a flexible parameterization of $p_\phi(\mathbf{A} \mid \mathbf{X}_{\text{LS}})$ that can be learned directly from data through likelihood-based training. This is achieved by defining the model distribution implicitly via an invertible transformation of a simple, well-characterized base distribution. In this work, the base distribution is a standard multivariate normal, $p_{\mathbf{Z}}(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, from which samples are transformed to the target distribution $p_\phi(\mathbf{A} \mid \mathbf{X}_{\text{LS}})$, as shown in Fig. 1. Once trained, this representation permits efficient generation of rain-area realizations consistent with the learned conditional distribution.

The mapping between the base distribution to the target distribution is defined through an invertible and differentiable function

$$f_\phi : (\mathbf{A}, \mathbf{X}_{\mathrm{LS}}) \mapsto \mathbf{z},$$

which transforms latent samples $\mathbf{z}$ into rain-area realizations conditioned on the large-scale state $\mathbf{X}_{\mathrm{LS}}$. The inverse mapping

$$\mathbf{A} = f_\phi^{-1}(\mathbf{z}, \mathbf{X}_{\mathrm{LS}})$$

allows computation of the exact likelihood for observed rain-area samples. The bijectivity of f_ϕ for fixed $\mathbf{X}_{\mathrm{LS}}$ ensures that the model admits an explicit conditional density and permits efficient generation of rain-area realizations for any prescribed large-scale state.

Model parameters ϕ are learned by maximum likelihood estimation, fitting $p_\phi(\mathbf{A} \mid \mathbf{X}_{\mathrm{LS}})$ to the observed data. Because the CNF defines p_ϕ implicitly through an invertible transformation of a known base distribution, the conditional log-density of an observed rain-area sample $\mathbf{A}$ given $\mathbf{X}_{\mathrm{LS}}$ can be evaluated exactly using the change-of-variables formula:

$$\log p_\phi(\mathbf{A} \mid \mathbf{X}_{\mathrm{LS}}) = \log p_{\mathbf{Z}}\big(f_\phi(\mathbf{A}, \mathbf{X}_{\mathrm{LS}})\big) + \log \left| \det \left(\frac{\partial f_\phi(\mathbf{A}, \mathbf{X}_{\mathrm{LS}})}{\partial \mathbf{A}} \right) \right|.$$

The training objective is to maximize the sum of this log-likelihood over all training samples.

The mapping f_ϕ is implemented using a conditional masked autoregressive flow (MAF, Papamakarios et al. 2017), in which the autoregressive structure operates over the components of $\mathbf{A}$. Conditioning on $\mathbf{X}_{\mathrm{LS}}$ is incorporated via a learned context vector that modulates each transformation layer. Architectural details are provided in SI.

Once training is accomplished, samples are generated from the learned conditional distribution $p_\phi(\mathbf{A} \mid \mathbf{X}_{\mathrm{LS}})$. To distinguish these model-generated realizations from the observed data, we denote them by $\widehat{\mathbf{A}}$. Specifically, samples are obtained by drawing $\mathbf{z}^{(j)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and applying the inverse transformation for a given large-scale state $\mathbf{X}_{\mathrm{LS}}^{(i)}$:

$$\widehat{\mathbf{A}}^{(i,j)} = f_{\phi}^{-1}\big(\mathbf{z}^{(j)}, \mathbf{X}_{\mathrm{LS}}^{(i)}\big) = \begin{bmatrix} \widehat{A}_{\text{shallow}}^{(i,j)} \\ \widehat{A}_{\text{deep}}^{(i,j)} \\ \widehat{A}_{\text{stratiform}}^{(i,j)} \end{bmatrix}. \tag{1}$$

Each sample $\widehat{\mathbf{A}}^{(i,j)}$ represents one realization drawn from p_{ϕ} given $\mathbf{X}_{\mathrm{LS}}^{(i)}$. Repeating this procedure M times yields Monte Carlo estimates of conditional mean implied by the model. For example,

$$\overline{A}_{\text{shallow}}^{(i)} = \frac{1}{M} \sum_{j=1}^{M} \widehat{A}_{\text{shallow}}^{(i,j)}, \tag{2}$$

with analogous expressions for deep and stratiform rain-area. Collecting $\bar{\mathbf{A}}$ yields a noise-reduced dataset,

$$\mathbb{D} = \left\{ \mathbf{X}_{\mathrm{LS}}^{(i)}, \overline{A}_{\text{shallow}}^{(i)}, \overline{A}_{\text{deep}}^{(i)}, \overline{A}_{\text{stratiform}}^{(i)} \right\}_{i=1}^{N},$$

which serves as input for downstream equation discovery (Section c). For notational simplicity, we drop the superscript (i); all rain-area quantities should be interpreted as conditional means estimated from the CNF model given $\mathbf{X}_{\mathrm{LS}}$.

c. Equation Discovery

We translate the estimated means (e.g. $\{\mathbf{X}_{\mathrm{LS}}, \overline{A}_{deep}\}$) into compact analytic relationships using symbolic regression. Specifically, we apply a python symbolic regression (PySR, Cranmer 2023), which employs genetic programming to search a user-defined library of functions for expressions that minimize error while penalizing unnecessary complexity. Symbolic regression has been successfully used to derive closed form equations for cloud fraction modeling (Grundner et al. 2024), ocean turbulent mixing (Zanna and Bolton 2020) and ocean eddy diffusivity (Sane et al. 2025).

For each rain-area (e.g., $\overline{A}_{\text{shallow}}, \overline{A}_{\text{deep}}$), we apply symbolic regression using PySR, fitting separate models for each target. We sweep over a range of expression depths and operator sets, and repeat the regression five times with different random seeds to ensure robustness. This procedure is performed across all experimental input sets defined in Section 2.a.b, allowing us to (i) identify which set of input variables is most predictive across all three rain-areas, and (ii) reduce sensitivity to stochastic variability in the optimization.

The output is a *Pareto front* of candidate formulas (Fig. S1; see also Beucler et al. 2025), each representing a trade-off between model complexity and accuracy. From this front, we select representative ”elbow” solutions that offer a favorable interpretability-performance balance. These are typically compact expressions with one or two additional terms beyond the simplest model. In addition to accuracy, we prioritize equations that include terms which appear consistently across runs with different random seeds. These selected equations can optionally be fine-tuned to improve their predictive performance following the approach of Grundner et al. (2024).

4. Results

a. Estimating the Joint Distribution of Rain Areas

We begin by evaluating how well the CNF reproduces the marginal and joint distributions of rain area for each rain type. Fig. 2 compares the observed and CNF predicted histograms of rain area for shallow convective (left), deep convective (center), and stratiform (right) types. For each environmental condition in the test set, we generated one sample (a triplet of rain type areas) from the CNF, mimicking the single-pass sampling characteristic of satellite observations. The model accurately reproduces the bulk of the distribution across all three rain types. For deep convection, the predicted distributions exhibit slightly heavier tails than observed, particularly for large-area events ($A_{deep} \gtrsim 15\%$) with frequencies below 10^2. This suggests the model slightly overestimates the likelihood of extreme spatial coverage, while still retaining high fidelity in reproducing overall statistical structure. For a complementary diagnostic, we present empirical CDFs of rain areas by rain type (Fig. S2), showing that the predicted distributions match the observed ones well across shallow, deep, and stratiform rain types, including at the distribution tails. Minor discrepancies are visible: the predicted CDF for deep convection slightly overestimates mid-range rain areas, while shallow rain is marginally shifted toward higher area fractions, both consistent with the model’s tendency to slightly overpredict area in less frequent regimes.

We now analyze the joint distribution of convective rain area, defined as the sum of the shallow and deep components, and stratiform rain area. Fig. 3 compares the observed (left) and predicted (right) two-dimensional histograms of convective area versus stratiform area, with color indicating

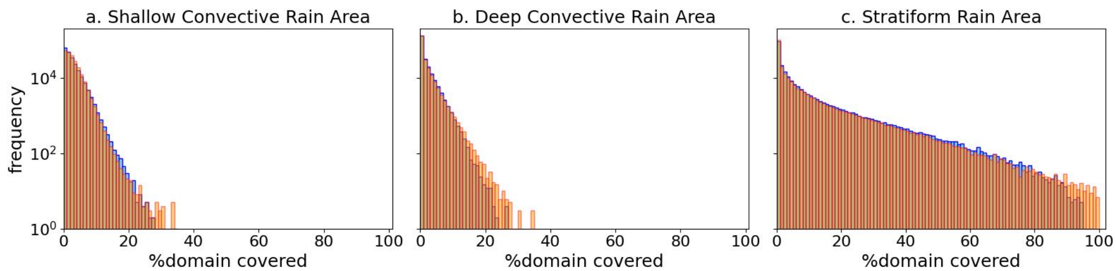

FIG. 2. Marginal distribution of a) shallow, b) deep convective, and c) stratiform rain areas. Blue indicates observed data, and orange represents the CNF generated distribution.

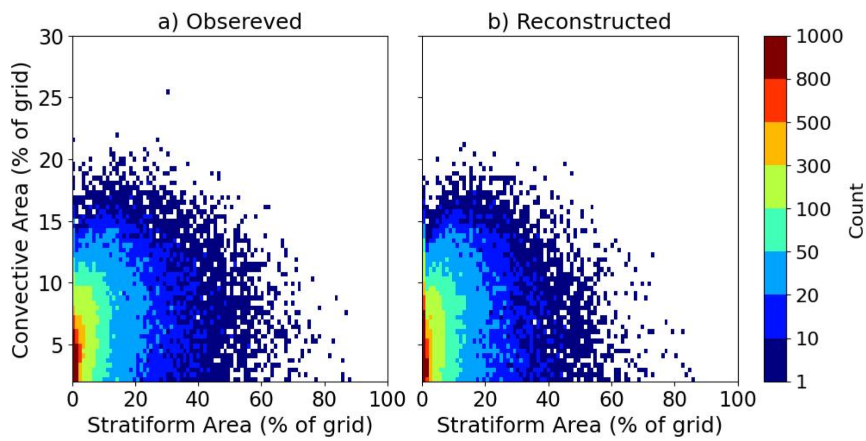

FIG. 3. Joint distribution of convective (shallow + deep) and stratiform rain area a) from observation and b) reconstructed by CNF. Color indicates number of occurrence in each bin.

number of occurrence for each bin in the joint plot. The model accurately reproduces the overall structure of the joint distribution, capturing the dense core centered around moderate stratiform area (0–20% of the grid) and convective area between 0–10%. It also captures the skewed triangular shape, where stratiform area reaches much larger values (80%).

To further examine structural relationships among three rain area components, we analyze the conditional means predicted by the CNF model. For each environmental condition, 300 samples are drawn and averaged to estimate the expected rain areas, $\overline{A}_{shallow}$, $\overline{A}_{deep}$, $\overline{A}_{strat}$ (see Sec. 3.b.2 and Eq. (2) for more detail). Fig. 4.a reveals two distinct bands in the joint distribution of stratiform and convective (the sum of shallow and deep) rain area. The upper band shows a strong positive correlation with a slope of $\beta = 0.14$, while the lower band exhibits a much weaker slope of $\beta = 0.03$. This bimodal structure suggests the presence of two regimes: one in which convective component occupy larger fraction of stratiform thus the covariance between the two is larger, and another in which stratiform rain extends more strongly for a small increase in convective area. When the analysis is restricted to the tropical belt between 20°S and 20°N, the lower branch largely disappears, and the data collapse onto a single dominant band with slope $\beta = 0.14$ (shown in Fig. S3). This transition suggests that the distinction between the two regimes is partly latitudinal in origin, with the stronger convective–stratiform coupling more prominent in the deep tropics.

A similar bimodal pattern has been reported in observational data by Schumacher and Funk (2023), who identified two distinct groupings of stratiform and deep convective rain area using GPM DPR data over $1^\circ \times 1^\circ$ grid cells. Their "Regime 1," associated with a higher convective fraction, was found primarily in the tropics, while "Regime 2," with lower convective contributions, dominated the extratropics. Although their grid resolution differs from ours and they consider only deep convection (excluding shallow convection), the emergence of similar regimes strengthens the interpretation that distinct dynamical processes underlie these patterns and our CNF can reproduce this pattern. Elsaesser et al. (2022) report that sustaining or growing stratiform anvils requires a system convective area fraction of approximately 15%, consistent with our finding of $\beta \approx 0.14$ for the upper band.

Fig. 4.b displays the relationship between deep and shallow convective rain area. Unlike the convective–stratiform case, their association is weaker and more diffuse, with most values clustered below 5% area. The dominant concentration (a bright orange core centered around 2–3% for both rain types) is largely driven by the tropical belt between 20°S and 20°N (see Fig. S3.b). Outside this region, the joint distribution becomes broader and less organized. Overall, compared to Fig, 4.a, shallow and deep convective areas exhibit a much weaker correlation.

This weak correlation can be partly attributed to differences in how deep and shallow convection are defined in the TRMM precipitation radar classification. Deep convection is identified based on strong radar echos, typically associated with intense updrafts and often spatially close to regions of stratiform rain. Shallow convection, by contrast, consists of isolated, low-topped cells with minimal vertical development and frequently occurs independently of deep systems. These two modes reflect distinct dynamical regimes that arise from differences in large-scale environmental conditions, even if those differences are subtle. Investigation of the underlying mechanisms is

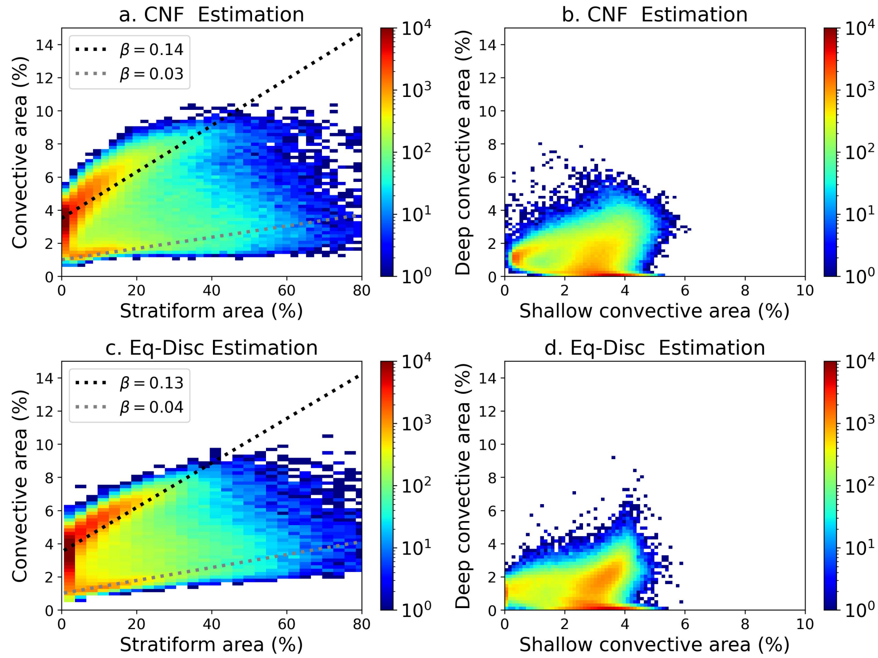

FIG. 4. Panels show two-dimensional histograms of rain areas predicted by the CNF model (means) and by the equation-discovery (Eq-Disc) model. The left panels (a, c) depict total convective area (deep + shallow) versus stratiform area, while the right panels (b, d) show deep versus shallow convective area. Color indicates frequency on a logarithmic scale. In panels (a) and (c), the black and gray dotted lines represent linear regressions between total convective and stratiform areas, with the estimated slopes shown in the legend.

beyond our scope, but the CNF's ability to reproduce these regimes without explicit training demonstrates the fidelity of its learned distribution.

b. Discovered Relationships

We now examine how large-scale environmental conditions relate to the mean of the conditional joint distribution of rain areas using equation discovery. To ensure interpretability and physical consistency, we constrain the equation discovery process to use parallel variables for the boundary layer and free troposphere. In our primary experiment, we include θ_e and θ_e^* for both layers $\left\{\theta_{e,\mathrm{ft}}, \theta^*_{e,\mathrm{ft}}, \theta_{e,\mathrm{pbl}}, \theta^*_{e,\mathrm{pbl}}\right\}$ along with T_S, and T_{ft}. We additionally test the inclusion of ω_{500} as a large-scale dynamical predictor. However, ω_{500} is not a purely prognostic environmental variable; it is, in part, a response to the occurrence of convection. Using ω_{500} to predict the spatial extent of rain therefore risks a causality issue: convection can drive large-scale ascent or be strongly coupled to it in quasi-equilibrium, yet the equation discovery would use that ascent to estimate the area of convection itself. Consistent with this caution, Retsch et al. (2022) used explainable AI applied to radar–reanalysis data over Darwin and found that mid-level vertical velocity is the single most influential large-scale predictor of total convective area, while noting the bidirectional coupling between convection and ω. In addition, Miller et al. (2025) show that mid-level vertical velocity emerges as the most important predictor for the deep-convection trigger function in a probabilistic ML framework. To avoid conflating association with causation while still assessing the role of ω_{500}, the main text reports equation discovery results excluding ω_{500}; for completeness, the corresponding results including ω_{500} are provided in Appendix B.

We emphasize that we do not assume a strictly one-way causal relationship between the large-scale environment and convection. It is well established that convection and its environment interact continuously and modify one another (Holloway and Neelin 2009), with feedbacks operating across a range of spatial and temporal scales. In this study, however, we adopt a pragmatic working assumption that, at the coarse spatial scale of $2^\circ \times 2^\circ$ and over the short temporal window considered here (approximately one hour), the large-scale environmental fields are only marginally modified by the ongoing convection. Under this assumption, the environmental variables can be interpreted as providing a useful description of the background state in which convection occurs, even though they may already reflect some degree of convective influence. The relationships identified in this work should therefore be interpreted as conditional and statistical, rather than strictly causal, and as representative of co-variability at the chosen spatial and temporal scales.

We also explored alternative predictor sets, including precipitable water, CRH, and buoyancy-like terms following Ahmed and Neelin (2018). Among these, CRH combined with ω_{500} showed great predictive skill, though its performance was weaker for shallow convection and the resulting expressions were substantially more complex. For reasons of parsimony, interpretability, and comparable accuracy, we therefore focus on the θ_e-based equations as our primary results. Additional results and details from the alternative predictor sets are provided in the Supplementary Materials and in Appendix B. Section 4.b.1 to 4.b.3 summarizes the best-performing equations, chosen for their balance of simplicity, accuracy, and physical interpretability. In the remainder of this section, we compare the predicted joint distribution from discovered equations against those from the CNF.

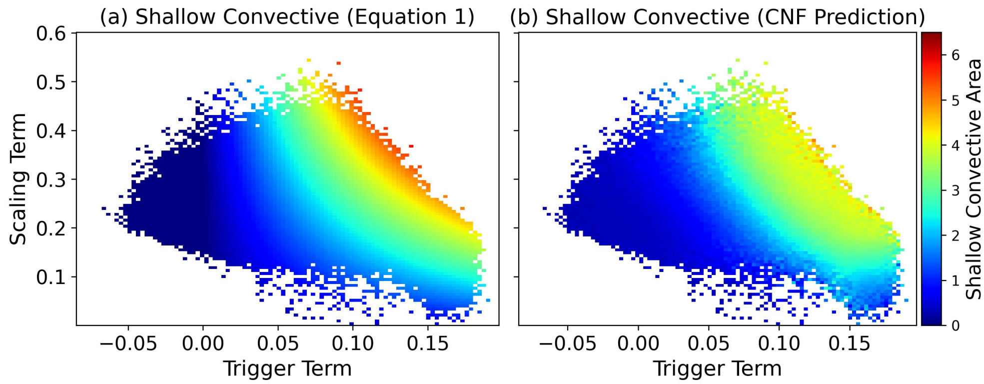

Fig. 5. Two-dimensional histograms illustrating the relationship between the trigger and scaling terms that control shallow convective rain area in Eq. (3). The horizontal axis shows the trigger term, and the vertical axis shows the scaling term. Color shading indicates the shallow convective rain area within each bin. The left panel shows areas computed directly from Eq. (3), while the right panel shows the mean areas from the CNF model.

Fig. 4.c and d show the joint distributions predicted by discovered equations. Panel c, which plots total convective versus stratiform area, captures the two distinct modes observed in the CNF-based distribution (Fig. 4.a). The slopes of the two regimes ($\beta = 0.13$ and $\beta = 0.04$) closely match the CNF predictions, indicating that the low-dimensional equations reproduce the distribution's structure. Compared to CNF, the equation's predictions show a tighter spread around the dominant modes, suggesting more constrained variability.

Panel d shows the joint distribution of deep and shallow convective areas. The main probability concentration is well reproduced, though the discovered equation slightly overestimates the core density and shows a narrower spread. The equations capture the dominant statistical structure of the learned rain-area distributions, supporting their use as interpretable reduced-form representations of precipitation behavior. We further assess skill using the coefficient of determination (R^2)[1] between the equations and CNF-derived means, obtaining values of 0.85, 0.89, and 0.88 for shallow, deep, and stratiform rain areas, respectively.

1) Shallow Convective Rain Area

The discovered equation for shallow area, Eq. (3), expresses the mean fractional area of shallow rain as the product of two dimensionless factors.

[1] R^2 measures the skill of prediction, with a value of 1 indicating perfect prediction and 0 indicating no skill beyond predicting the mean, and negative values indicating worse performance than simply predicting the mean.

$$\overline{A}_{\text{shallow}} = a_0\left(\frac{a_1\theta^*_{e,\text{ft}} - \theta^*_{e,\text{pbl}}}{\theta^*_{e,\text{ft}}}\right)\left(\frac{\theta_{e,\text{pbl}} - \theta^*_{e,\text{ft}}}{\theta^*_{e,\text{ft}}} + a_2\,\frac{\theta^*_{e,\text{ft}} - \theta_0}{\theta^*_{e,\text{ft}}}\right)^+ \tag{3}$$

where $(x)^+ \equiv \max(0,x)$, and $a_0 = 4.7\times10^2$, $a_1 = 1.05$, $a_2 = 1.4$, $\theta_0 = 302\text{K}$. The two factors are a scaling term and a trigger term, with the latter determining when shallow convection is active.

Within the trigger term, $\frac{\theta_{e,\text{pbl}}-\theta^*_{e,\text{ft}}}{\theta^*_{e,\text{ft}}}$ represents a dimensionless boundary-layer buoyancy surplus, measuring how much the the boundary-layer equivalent-potential temperature exceeds the saturation value aloft. A moist, well-mixed boundary layer capped by a cool free troposphere (small $\theta^*_{e,\text{ft}}$) makes this term large and positive, indicating sufficient buoyancy for up-drafts. The second term $\frac{\theta^*_{e,\text{ft}}-\theta_0}{\theta^*_{e,\text{ft}}}$, scaled by the fitted coefficient $a_2 \simeq 1.4$, depends only on the temperature of the free troposphere, given that θ_0 is a constant. As the free troposphere warms, this second term increases. With $a_2 > 1$, this enhanced warm-layer contribution, while keeping $\theta_{e,\text{pbl}}$ fixed, more than compensates for the simultaneous reduction in boundary-layer surplus. As a result, the sum is more likely to be positive, activating the trigger.

since $a_2 > 1$, a warmer free troposphere strengthens the second term faster than it erodes the boundary-layer-surplus term at fixed $\theta_{e,pbl}$, lowering the activation threshold and making triggering more likely. A colder free troposphere weakens the second term and raises the activation threshold. In practice, this formulation allows shallow rain area to expand when a moist boundary layer lies beneath either a cool inversion (given sufficient BL moisture) or a moderately warm free troposphere, while suppressing shallow convection when the boundary layer is dry.

The scaling term in Eq. 3, $\frac{a_1\,\theta^*_{e,\text{ft}}-\theta^*_{e,\text{pbl}}}{\theta^*_{e,\text{ft}}}$, depends only on the saturation equivalent potential temperature, and thus on temperature at fixed pressure. With the fitted value $a_1 > 1$, this ratio remains positive and increases as the free troposphere becomes warmer than the boundary layer, indicating greater lower-tropospheric stability. Normalizing by $\theta^*_{e,\text{ft}}$ renders the term dimensionless while preserving its interpretation as a stability weight. A more stable free troposphere suppresses deep updraft growth and favors a larger fractional area of shallow convection that detrains in the lower free troposphere.

Together, the scaling and the trigger terms capture the observed behavior of trade-wind cumuli, where extensive shallow convective area is favored when a strong inversion overlies a moist boundary layer that remains buoyant enough to penetrate the capping layer, while deep convection is suppressed. Previous observational studies reach overlapping but not identical conclusions: trade-wind studies emphasis boundary-layer moist entropy and lower-tropospheric stability (Bony et al. 2020; Schulz et al. 2021), whereas the RICO analyses identify lower–free-tropospheric humidity as the primary control (Nuijens et al. 2009). Eq. (3) contains no free-tropospheric humidity predictor, but it captures the boundary-layer moist entropy via $\theta_{e,pbl}$ and the lower-tropospheric stability via $\theta^*_{e,\mathrm{ft}} - \theta^*_{e,\mathrm{pbl}}$. In this sense, the explicit, thresholded form distills these controls into a single quantitative predictor of shallow-rain area.

Fig.5 maps each environmental sample into (trigger, scale)-space and colors the point by the resulting shallow-convective area. The area is near zero for negative trigger values, then increases once the trigger becomes positive, with larger scale values amplifying the growth. The analytic prediction (panel a) closely matches the CNF response (panel b), with only slight overestimation at the largest values, indicating that the compact equation captures both the activation threshold and magnitude learned by the CNF.

2) Deep Convective Rain Area:

Eq. (4) links large-scale thermodynamic conditions to deep convective rain-area, $\overline{A}_{\mathrm{deep}}$,

$$\overline{A}_{\mathrm{deep}} = b_0\left(\frac{\theta_{e,\mathrm{pbl}} - \theta^*_{e,\mathrm{ft}} + \theta_\epsilon}{\theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}} + \theta_\epsilon}\right)^+ + b_1\left(\frac{\theta^*_{e,\mathrm{pbl}} - \theta_1}{\theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}} + \theta_2} - b_2\right)^+ \tag{4}$$

with $b_0 = 2.1$, $b_1 = 0.35$, $b_2 = 7.9$, $\theta_\epsilon = 1\,\mathrm{K}$, $\theta_1 = 164\,\mathrm{K}$, $\theta_2 = 9\,\mathrm{K}$.

In this equation the first term models the boundary-layer buoyancy surplus versus free troposphere dryness. The numerator, $\theta_{e,\mathrm{pbl}} - \theta^*_{e,\mathrm{ft}} + 1$ (K), represents the thermodynamic contrast between the boundary layer and the saturated free troposphere, and thus the buoyancy of a rising parcel. The $\theta_\epsilon =$ 1 K offset arises naturally from the equation-discovery procedure as part of the optimal functional form, rather than being imposed a priori. A closely related buoyancy contrast, without this constant term, was previously proposed by Ahmed and Neelin (2018) in their model of precipitation. The denominator, $\theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}}$, scales the surplus by the free-tropospheric saturation deficit, suppressing convection when the mid-troposphere is too dry. Multiplication by the fitted gain $b_0\,(\sim 2.1)$ amplifies the response once this “buoyancy-over-dryness” ratio exceeds zero. Physically, the term turns on when the PBL can supply parcels that remain positively buoyant after penetrating the

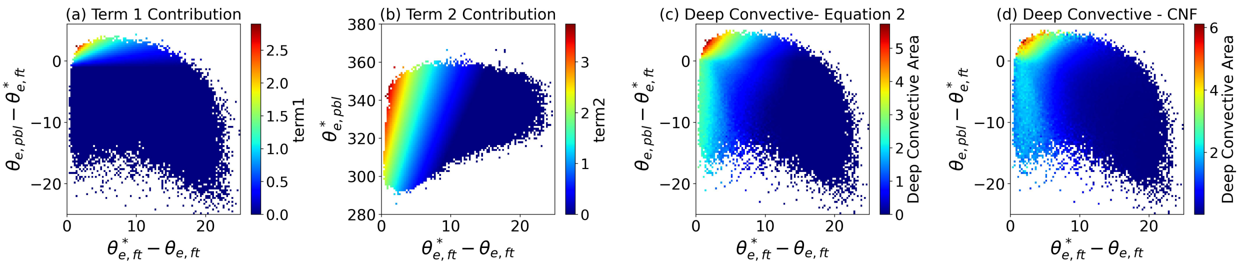

Fig. 6. Two–dimensional histograms illustrating the individual terms and overall performance of the deep-convection closure, Eq. (4). All panels use the free-tropospheric saturation deficit (K) on the x-axis. The y-axis shows the boundary-layer buoyancy surplus (K) in panels (a), (c) and (d), and the saturation value $\theta^*_{e,\mathrm{pbl}}$ in panel (b), consistent with the variables appearing in each term of the equation. Each bin ($\Delta x = \Delta y = 0.3\,\mathrm{K}$) aggregates all samples within it; color indicates the mean rain area per bin. (a) Contribution of the buoyancy-over-dryness term; (b) contribution of the humidity gate; (c) deep-convective area predicted by Eq. (4); and (d) by the CNF.

entraining free troposphere (Ahmed and Neelin 2018). As in the numerator, the 1 K offset in the denominator is determined by the equation-discovery algorithm.

The second term, $\left(\frac{\theta^*_{e,\mathrm{pbl}} - \theta_1}{\theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}} + \theta_2} - b_2\right)^+$ serves as a humidity gate. The numerator $\theta^*_{e,\mathrm{pbl}} - \theta_1$ increases with boundary-layer temperature, boosting the term when the sub-cloud layer is energetic. The denominator $\theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}} + \theta_2$ is the mid-tropospheric saturation deficit, padded by $\theta_2 \simeq 9\,\mathrm{K}$ to avoid singular behavior near saturation and also to scale down the numerator. Because the numerator depends on boundary-layer conditions while the denominator depends on free-tropospheric dryness, the ratio functions as an inverse-dryness index and becomes large when the PBL is warm and the free troposphere is relatively humid, but shrinks rapidly as the free troposphere dries. Since both the first and second terms depend on the FT saturation deficit, sufficiently dry mid-tropospheric conditions drive both terms to zero. Consequently, deep convective area can only expand when the mid-troposphere is sufficiently moist to limit entrainment-driven dilution, making explicit the moisture control predicted by entraining-plume theory.

In Eq. (4) the first term carries the buoyancy signal through the actual boundary-layer value $\theta_{e,\mathrm{pbl}}$, while the second term operates as a humidity boost term built around the saturation quantity $\theta^*_{e,\mathrm{pbl}}$. Although its numerator depends only on temperature, the second term remains sensitive to column moisture through its denominator, the free-tropospheric saturation deficit. In hot–dry environments, both numerator and denominator increase together, keeping the ratio small and the term effectively shut off. When the mid-troposphere is humid, the deficit shrinks, the ratio grows, and the second term activates, reinforcing the buoyancy surplus supplied by the first term. With the fitted coefficient $\theta_2 > 1$ the second term even relaxes slightly in moderately warmed yet still-humid mid-levels, consistent with the persistence of deep convection in "moist-mode" conditions, while still shutting down in hot, dry regimes. Hence, the compact formulation balances temperature and dryness without double-counting boundary-layer moisture, preventing spurious deep-convection activation and capturing the environments that favor it.

We visualize the terms in Eq. (4) to evaluate how each term maps the environmental phase space into deep-convective rain area. Fig. 6.a isolates the buoyancy–dryness contribution $b_0\left[(\theta_{e,\mathrm{pbl}} - \theta^*_{e,\mathrm{ft}} + \theta_\epsilon)/(\theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}} + \theta_\epsilon)\right]^+$. The signal is confined to the upper-left corner of phase space, where the boundary-layer surplus $y = \theta_{e,\mathrm{pbl}} - \theta^*_{e,\mathrm{ft}}$ is positive and the free-tropospheric deficit $x = \theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}}$ is small. Here the ratio inside the trigger operator is largest and the term approaches its fitted ceiling ($\sim 3\%$ area). Once the deficit exceeds ~ 8 K or the surplus turns negative, the term collapses, confirming that boundary-layer buoyancy alone cannot sustain deep convection when mid-level dryness is appreciable. Fig. 6.b shows the second term $b_1\left[(\theta^*_{e,\mathrm{pbl}} - \theta_1)/(\theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}} + \theta_2) - b_2\right]^+$. This term is amplified in warm boundary layers (high $\theta^*_{e,\mathrm{pbl}}$) and confined to small to moderate deficits; the 9 K buffer in the denominator damps the response beyond $x \approx 8$ K.

Fig 6.c, which combines the two analytic terms, reproduces the CNF pattern shown in Fig. 6.d: a fan-shaped lobe of high deep-convective area centered on $x < 5$ and $y > 0$ (K). Apart from a slight under-prediction for $x \approx 8$–10 K, where the CNF retains a faint tail, the activation boundary and the amplitude gradient align closely, giving an overall correspondence of $R^2 = 0.89$ between the analytic closure and the CNF predictions.

We additionally test how much predictive power each term of Eq. (4) carries on its own by re-fitting the closure using only one term at a time. Retaining the buoyancy–dryness term alone lowered the skill from $R^2 = 0.89$ to 0.65; keeping only the humidity gate still achieved $R^2 = 0.84$, but the tail of the distribution, high rain area at small saturation deficits and positive boundary-layer surplus, was poorly reproduced (not shown). Thus the second term explains most of the bulk variance, while the first term is essential for matching the high-intensity tail and restoring the full predictive skill of the two-term formulation.

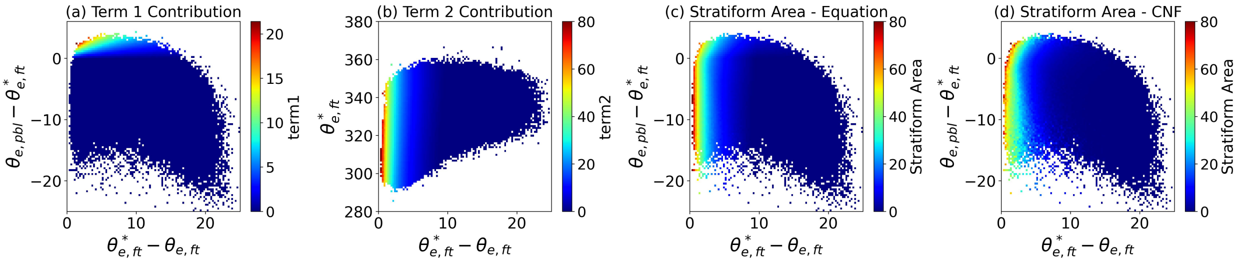

FIG. 7. Similar to Figure 6 but for stratiform rain area following Eq. (5)

3) STRATIFORM RAIN AREA:

The fitted expression for stratiform rain area is:

$$\overline{A}_{\mathrm{strat}} = c_0\left(\frac{\theta_{e,\mathrm{pbl}} - \theta^*_{e,\mathrm{ft}}}{\theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}} + \theta_\epsilon}\right)^+ + c_1\left(\frac{\theta^*_{e,\mathrm{ft}} + \theta_3}{\theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}} + \theta_\epsilon} - c_2\right)^+ \tag{5}$$

where $c_0 = 22,\ c_1 = 0.41,\ c_2 = 34,\ \theta_3 = 19.2$ K, $\theta_\epsilon = 1$ K.

The first term in Eq. (5), $c_0\left(\frac{\theta_{e,\mathrm{pbl}} - \theta^*_{e,\mathrm{ft}}}{\theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}} + \theta_\epsilon}\right)^+$, is structurally identical to the buoyancy–dryness term discovered for deep convection, except that the +1 K buffer is not present in the numerator and the fitted gain is much larger ($c_0 \approx 22$ versus $b_0 \approx 2$); this expression requires a strictly positive surplus to trigger. Physically, this term switches on only in environments where boundary-layer parcels would remain buoyant after penetrating a still–humid mid-level layer; conditions typically found in or near the active convective cores that give rise to stratiform shields (Houze 1997). The large coefficient c_0 compensates for the relatively small buoyancy surplus, scaling it up such that it matches the observed, often extensive, stratiform rain area. Under moist conditions, a modest buoyancy signal is thus amplified roughly tenfold relative to the deep-convection case, consistent with the observed fraction of stratiform area.

The second term operates purely as a mid-tropospheric humidity trigger term: $c_1\left(\frac{\theta^*_{e,\mathrm{ft}} + \theta_3}{\theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}} + \theta_\epsilon} - c_2\right)^+$. Both numerator and denominator refer to the same free-tropospheric layer, so the bracket is an inverse-dryness index. Because the denominator $\theta^*_{e,\mathrm{ft}} - \theta_{e,\mathrm{ft}}$ appears with only a 1K buffer, the ratio becomes large when the layer is nearly saturated. Yet even a modest increase in deficit rapidly drives the expression toward zero, as shown in Fig. 7.b where large values are confined to the left-most column of the phase space. The constant θ_3 raises the numerator so that moderately warm mid-level temperatures ($\theta^*_{e,\mathrm{ft}} \sim 300$–$320\,\mathrm{K}$) can still meet the threshold, while the large subtraction $c_2 = 34$ K enforces a strict moisture requirement.

Multiplication by $c_1 = 0.41$ keeps the contribution modest in absolute terms, yet, because the inverse-dryness bracket can reach $O(10^2)$ when the deficit is $\lesssim 1$ K, the term attains its maximum along the saturated edge and then dies off sharply with increasing dryness. Thus the term selectively boosts stratiform rain area in environments where the free troposphere is already moist, typical of the anvils and trailing stratiform clouds produced by mature convective systems, while leaving drier columns to be handled by the buoyancy–dryness addend.

Fig. 7 dissects the stratiform-rain area equation in Eq. (5). Panel (a) shows the buoyancy–dryness term, which contributes only when the boundary-layer surplus is positive and the free-tropospheric deficit is small; its contribution is thus confined to a thin wedge in the upper-left corner of phase space. Panel (b) isolates the humidity gate; because both numerator and denominator use free-tropospheric variables, its value depends almost exclusively on the saturation deficit x. The bright vertical stripe at $x \lesssim 2$K confirms that the gate activates only in nearly saturated mid-levels and is indifferent to the boundary-layer surplus plotted on the y-axis of the other panels. Adding the two contributions (panel c) reproduces the CNF prediction in panel (d): a broad plateau of 40–70 % stratiform area for very small deficits, tapering rapidly as dryness increases.

c. Sensitivity of Rain Area Terms to SST and Thermodynamic State

To diagnose the dependence of the analytic closures on large scale conditions, we project Eqs. (3) and (4) onto a two parameter phase space defined by SST and the saturation equivalent potential temperature contrast $\Delta\theta^*_e \equiv \theta^*_{e,\mathrm{pbl}} - \theta^*_{e,\mathrm{ft}}$, which measures convective inhibition between the boundary layer and the free troposphere. Equation (5) is analyzed separately using a different parameter pair. SST serves as a proxy for the background climate state and surface turbulent forcing of the boundary layer (Rieck et al. 2012; Zelinka et al. 2013), while $\Delta\theta^*_e$ is closely related to lower tropospheric stability (Klein and Hartmann 1993; Klein 1997; Wood and Bretherton 2006; Wood and Hartmann 2006), with more negative values indicating a more stable free troposphere relative to the boundary layer. The two axes are dynamically coupled through convective and wave activity. When convection is active, boundary-layer and free-tropospheric anomalies are tightly

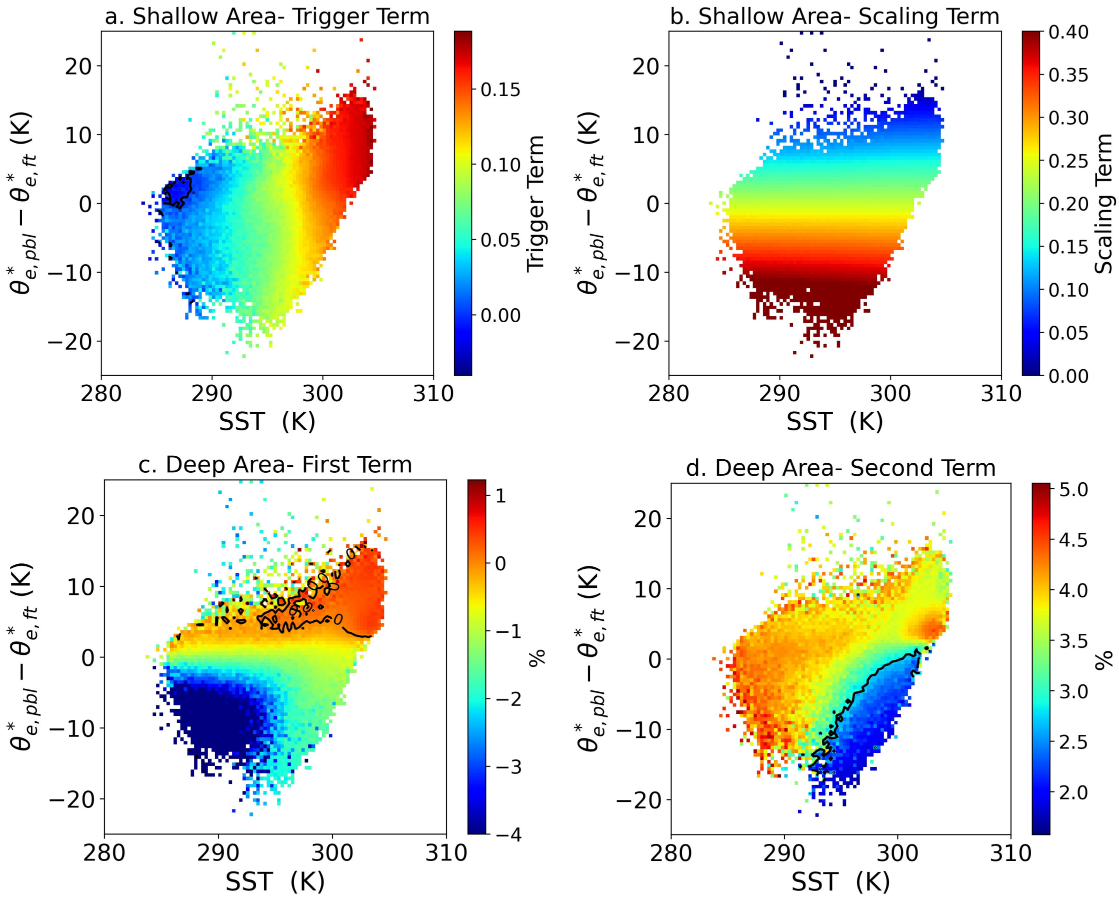

FIG. 8. Figure shows sensitivity of convective rain areas to surface temperature and atmospheric stability. All subplots share the same axes: x-axis is sea-surface temperature (SST, K); y-axis is $\Delta\theta^* = \theta^*_{e,pbl} - \theta^*_{e,ft}$ (K). Points are 2-D bins coloured by the mean value of the indicated analytic term; higher (warmer) colours denote larger positive values. A solid black contour marks the zero line of the corresponding trigger operator, i.e. where the term switches from inactive to active. (a) Shallow-convection trigger term (Eq. (3), second bracket). (b) Shallow-convection scaling term (always-positive prefactor in Eq. (3)). (c) Deep-convection buoyancy–dryness term (first term in Eq. (4)). (d) Deep-convection humidity gate (second term in Eq. (4))

linked, whereas under suppressed convection the free troposphere evolves more independently, allowing large positive $\Delta\theta^*_e$ to persist despite surface warming.

Fig.8.a displays the *trigger term* in Eq. (3). This term is positive over nearly the entire phase space; it is modulated primarily by SST and therefore exhibits a strong horizontal gradient, with higher values at warmer SSTs. Physically, warmer ocean temperatures raise $\theta_{e,\mathrm{pbl}}$ relative to the free–tropospheric reference $\theta^*_{e,\mathrm{ft}}$, so the triggering condition is easily met on the warm side of

the diagram. The scaling term (Fig. 8.b) increases monotonically with $\Delta\theta_e^*$, amplifying shallow-convective area under stronger stability. The largest shallow rain areas occur under warm SSTs, which activate the trigger, and a stable free troposphere, which enhances the scaling term. In the phase–space diagrams this corresponds to the lower–right part where warm SSTs coincides with large positive $\Delta\theta_e^*$ (see Fig. S6.a). Hence a warming ocean combined with any process that enhances lower–tropospheric stability (e.g. deep convection in nearby regions, subsiding subtropics) will shift the system toward this favorable corner, implying a systematic expansion of shallow-convective area under such large-scale environmental changes.

Figs. 8c–d show the two additive components of Eq. (4) in the ($\Delta\theta_e^*$, SST) phase space. Positive values are confined to warm SSTs and $\Delta\theta_e^* \gtrsim 0$, expanding with surface warming and shrinking as free-tropospheric stability increases. Activation forms a diagonal tongue in the upper–right quadrant, while regions with $\Delta\theta_e^* < 0$ are clipped to zero.

The second term of Eq. (4) is shown in Fig. 8d. Its contribution spans a broader diagonal swath, extending into moderately *negative* $\Delta\theta_e^*$. Thus, even with some convective inhibition ($\Delta\theta_e^* < 0$), a sufficiently warm and moist boundary layer beneath a moist free troposphere can still trigger deep convection through this pathway. The active region collapses in the lower–right corner of the phase space, where a warm (stable) and extremely dry free troposphere suppresses deep convection (Fig. S7). Because moisture deficits enter the denominator in both terms, increasing dryness enlarges the denominator and shuts down the positive-part contributions. This explains both the disappearance of deep convection in dry free-tropospheric conditions and its suppression precisely where the shallow-convection closure (Eq. (3)) would otherwise predict maximal rain area.

The overall deep-convective area, then, forms a wedge that broadens with surface warming but contracts when free-tropospheric stability intensifies. Thus the net change in deep-convective area depends on the competition between SST increase (rightward shift) and any accompanying rise in free-tropospheric temperature (downward shift in $\Delta\theta_e^*$): warming alone broadens the active wedge, whereas concurrent stabilization of the FT offsets much of that expansion, leaving large swaths of ($\Delta\theta_e^* < 0$) space blue and convection-free (see Fig. S6.b). Although the axes of this phase space do not explicitly represent humidity, supplementary analysis (Fig. S7) shows that the convection-free regions correspond to environments with particularly dry free tropospheres. This confirms that the contraction of the active wedge arises not only from stability effects but also from

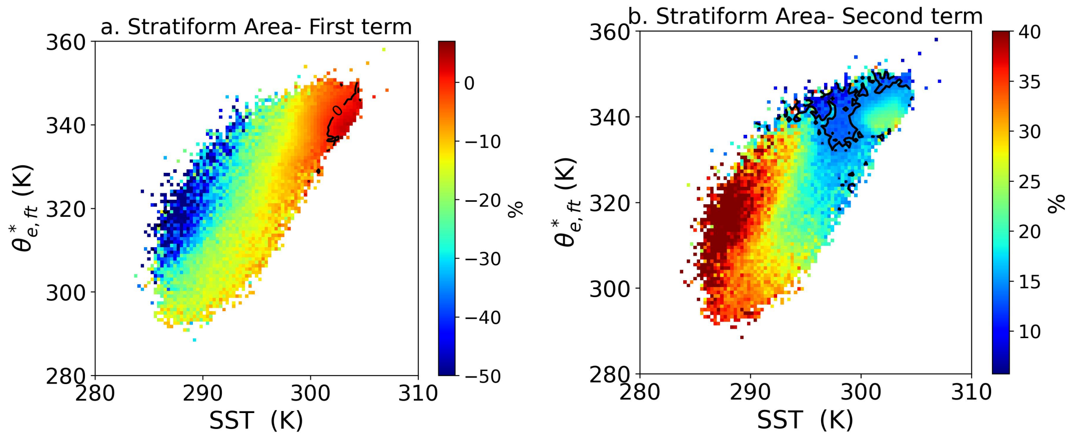

FIG. 9. Phase-space structure of the two components that enter the stratiform-area closure (Eq. (5)). (a) First (boundary-layer buoyancy–surplus) term; (b) Second (free-tropospheric deficit) term. Both panels are binned in SST (x-axis) versus free-tropospheric saturated equivalent potential temperature, $\theta^*_{e,ft}$ (y-axis). Color shading indicates the magnitude of each term as a percentage, and black contours trace the zero line where each trigger operator switches on.

the co-occurring dryness of the free troposphere. Previous studies (Tompkins 2001b; Derbyshire et al. 2004; Tompkins and Semie 2017; Stanfield et al. 2019) show that deep convection can be shut down over warm oceans when updrafts entrain dry free-tropospheric air; the resulting dilution reduces buoyancy. Conversely, organized convective systems can create a moist "halo" that shields subsequent updrafts from this dry environment—a form of "moisture memory" (Tompkins 2001a). Importantly, observational and modeling work on deep-precipitation onset further demonstrates that the transition to strong deep convection typically occurs only after the free troposphere moistens past a critical threshold (column-integrated or mid-tropospheric RH of $\sim 70-80\%$, Holloway and Neelin 2009; Schiro and Neelin 2019). Our phase-space plots captures this circumstance: the blue minimum at high SST and strongly negative $\Delta\theta^*_e$ pinpoints columns where free-tropospheric dryness, once entrained, stops convection from deepening, yet shallow convection is favored given that the boundary layer is moist.

Fig. 9 display the two contributors to the stratiform-area closure. In Fig. 9.a the color scale represents the first (buoyancy-surplus) term. Red shades cluster in the upper-right corner, where the free troposphere is warm and carries only a modest saturation deficit (see Fig. S7), while the PBL is still warmer/moister than $\theta^*_{e,\mathrm{ft}}$; under these circumstances the numerator is positive and the denominator small, yielding a large positive contribution. Moving toward cooler SST or toward scenes with a larger FT temperature drives the ratio negative, producing the blue, inactive region.

Fig. 9.b shows the second term, which depends only on free-tropospheric conditions. The numerator $\theta^*_{e,\text{ft}} + \theta_3$ scales directly with $\theta^*_{e,\text{ft}}$, but the denominator $\theta^*_{e,\text{ft}} - \theta_{e,\text{ft}} + \theta_\epsilon$ is more sensitive to changes in $\theta^*_{e,\text{ft}}$, so that the ratio is largest when $\theta^*_{e,\text{ft}}$ is low and the FT is close to saturation. Consequently this term is positive (yellow–red) along the lower-left flank of the cloud of points and decreases to near-zero (dark blue) as $\theta^*_{e,\text{ft}}$ rises. Combining the panels, warming the surface mainly amplifies the first term, whereas warming or drying aloft suppresses the second; the net stratiform area therefore hinges on the competition between boundary-layer heating and free-tropospheric moistening versus large-scale FT warming (see also Fig. S7).

Observational composites and idealized RCE/GCM experiments show that the fractional area of upper-tropospheric anvil cloud shrinks as SST rises and the upper troposphere stabilizes (Hartmann and Larson 2002; Zelinka et al. 2013; Bony et al. 2016; Saint-Lu et al. 2020; Stauffer and Wing 2022). Ito and Masunaga (2022) extend this picture with A-Train observations, demonstrating that anvil cloud fraction declines monotonically with increasing upper-tropospheric stability (their S_{UT}) while remaining largely insensitive to precipitation efficiency. Our study, in contrast, focuses on present-day variability (rather than forced climate change) and on precipitating stratiform rain area, a lower-level, optically thicker cloud regime, rather than the thin, non-precipitating anvil considered in the iris literature. Despite these differences in altitude, cloud type, and metric, the same moisture–stability control emerges. In our phase-space plot the buoyancy-surplus term is boosted by warm SST only when the free troposphere is moist, whereas the FT-deficit term collapses as $\theta^*_{e,\text{ft}}$ and the saturation deficit grow. The resulting blue tongue at warm SST (Fig.9 and Fig. S6.c) and high $\theta^*_{e,\text{ft}}$ echoes a similar shrinkage diagnosed by previous studies: a dry, highly stable lower free troposphere suppresses detrainment and horizontal spreading, whether the outcome is reduced high-cloud cover or diminished precipitating stratiform area.

5. Discussion and Conclusion

Precipitation forecasts hinge not only on intensity but on how much of a grid box is raining. Yet the fractional areas of shallow, deep-convective, and stratiform rain have seldom been modeled, even though they dominate sub-hourly precipitation variability and set the mass flux in convection schemes.

We close that gap with a two-step, data-driven framework. A conditional normalizing flow learns the full, stochastic joint distribution of the three rain-area fractions from large-scale thermodynamic and vertical velocity predictors, preserving their co-occurrence patterns. Symbolic regression applied to the CNF's conditional means distills this information into three concise "rain-area equations." Each takes the form of a buoyancy–humidity trigger term that identifies the environmental thresholds governing shallow, deep, and stratiform area.

For shallow convection, a buoyancy-surplus trigger and stability-based scaling reproduce its dependence on free-tropospheric stability and boundary-layer buoyancy. The deep convection expression reveals two key activation mechanisms: a buoyancy-over-dryness ratio that encapsulates parcel energetics relative to environmental dilution, and a humidity gate that suppresses rain area when the mid-troposphere is dry. The stratiform equation, by contrast, amplifies modest boundary-layer surpluses under near-saturated mid-levels and includes a sharp humidity threshold that isolates moist environments typical of mesoscale stratiform outflow. These results offer, for the first time, compact analytic closures for the co-occurrence of the three type rain areas of different convective regimes, grounded in physical intuition yet derived from high-dimensional generative models. This synthesis bridges observational, theoretical, and modeling perspectives and opens pathways toward improved understanding and representation of moist processes in weather and climate systems.

Our approach offers a promising path toward improving the representation of convection in large-scale models. In mass-flux parameterizations, the convective area plays a central role in scaling vertical transport of energy, mass, and moisture. Yet, this area is often prescribed using fixed fractions or simple empirical rules, limiting the ability of current schemes to respond dynamically to environmental variability. Our framework offers an alternative by predicting the area fractions of shallow, deep convective, and stratiform rain types as functions of the large-scale state. Importantly, it introduces stochasticity in a physically meaningful and data-driven way by sampling from the conditional distribution of rain area. Modeling rain area directly provides a mechanism for incorporating variability into convection schemes, which has been shown in previous studies to improve aspects of climate model performance (Berner et al. 2017). More generally, the need to model convective area and its variability is becoming increasingly important as climate models move toward higher resolutions and gray-zone grid spacings, where explicit convection is only partially resolved and accurate representation of convective area is critical.

Beyond its potential applications in mass-flux schemes, the framework enables a more detailed representation of the structure of convective systems. The separation into shallow, deep, and stratiform rain types reflects the diversity of vertical and horizontal convection structure observed in satellite data. This separation may also benefit other components of Earth system models, such as radiative transfer schemes, which rely on information about cloud area, type, and overlap to compute radiative fluxes. While additional work is needed to explore these applications in detail, our results suggest that environmental control of rain area can offer a more flexible and physically consistent interface between large-scale conditions and subgrid-scale processes.

The equations discovered here enable quantitative, process-level evaluation of precipitation structure through direct comparison with satellite products such as TRMM and GPM, which provide separate estimates of shallow, deep convective, and stratiform rain area. They offer a physically grounded benchmark for testing whether models, including high-resolution simulations, AI-based emulators, and statistical downscaling methods, reproduce the spatial organization of precipitation across convective regimes. This capability is particularly useful for generative machine learning approaches, which can produce large sample sets and estimate distributional moments. Although conventional climate models often represent convection as a single simplified process, our framework helps diagnose mismatches between environmental conditions and the expected convective footprint, providing a more interpretable reference for evaluating model performance across scales.

While this framework demonstrates that interpretable equations can be derived for rain-area statistics conditioned on large-scale environmental variables, important caveats remain. Many predictors co-vary with convection itself, complicating causal interpretation. The analysis also focuses on a limited set of variables and omits aspects of storm structure, such as convective organization, that may be important for extreme precipitation. Moreover, the present study models only conditional means, leaving the variance and covariance of rain areas as open questions. Addressing these limitations, along with incorporating organizational metrics, provides natural directions for future work.

Acknowledgments. This research received support through Schmidt Sciences, LLC, under the M^{2}LInES project. PAU was supported by the National Science Foundation (NSF grant AGS-2218829) and the National Oceanic and Atmospheric Administration MAPP program (NA21OAR4310343), as well as by the Korea Meteorological Administration Research and Development Program under Grant (RS-2025-02221669) and the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (RS-2024-00336160).

Data availability statement. This study used 10 years of ERA5 and TRMM precipitation radar data. All ERA5 datasets are openly accessible via the Copernicus Climate Data Store. The TRMM precipitation and rain type data are available from the University of Washington at `trmm.atmos.washington.edu`. All codes and software, including the machine-learning tools, can be found at this GitHub link.

APPENDIX A

Assessing the Predictive Value of Rain Area

As supplementary analysis, we assess the predictive value of rain area for rainfall rates using feedforward neural networks trained with identical sets of large-scale environmental predictors, with and without rain area included as an additional input. The networks consist of three hidden layers with 128, 128, and 64 units, ReLU activations, dropout regularization, and early stopping, and a softplus output layer to enforce non-negativity. The baseline configuration uses 12 large-scale environmental variables, while the augmented configuration adds rain area as a thirteenth predictor. This setup isolates the incremental information provided by rain area beyond that already contained in the large-scale state.

We carried out these experiments both regionally and globally, training models separately for each of the four oceanic regions as well as on the combined dataset. We further evaluated predictive skill across different rainfall categories: total versus convective precipitation, and the full distribution (hereafter referred to as full range) versus the extremes. We define extreme precipitation by first binning precipitable water (PW) into intervals of 0.5 mm. Within each PW bin, we compute the 95th percentile of the associated precipitation values and classify events exceeding this threshold as extremes. Comparing the predictive performance of the baseline and augmented networks provides a direct measure of the added value of rain area. We assess the impact of including rain-area

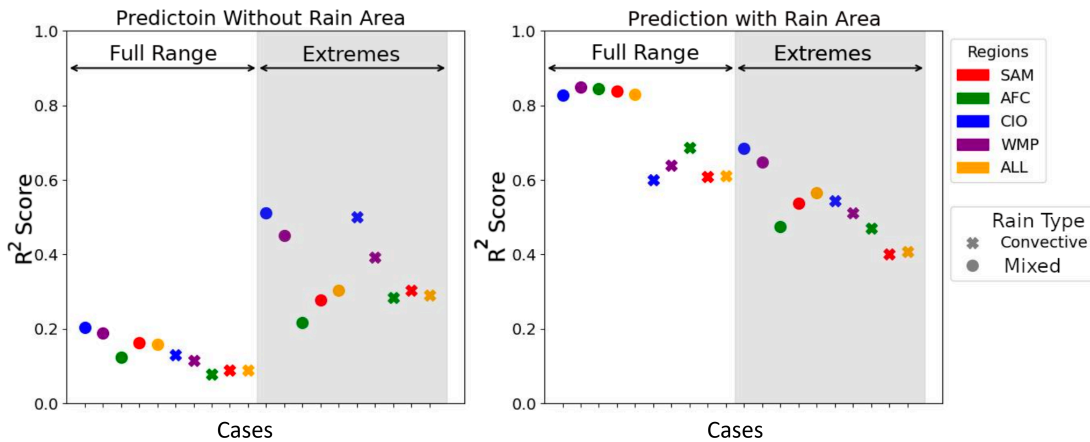

Fig. A1. Precipitation Predictability With and Without Rain-Area Information. Panels show the R^2 scores for predicting rain rate averaged over the grid box without (left) and with (right) rain-area information as input. Each point represents a different region (indicated by color), and marker shapes distinguish precipitation types: "x" for convective-only (including shallow and deep convective) and "o" for total precipitation (including shallow, deep convective, and stratiform). Shaded regions highlight predictions for extreme precipitation events; unshaded areas represent predictions over the full distribution of precipitation amounts.

information on the prediction of instantaneous precipitation by comparing two model setups: one without rain-area data (Fig A1, left) and one with it (right). Both setups are provided with the same set of 12 environmental predictors are described in section 2.b and listed in Tab. 1. The model variant on the right additionally receives rain-area fractions as input.

Across all cases, the inclusion of rain area as a predictor significantly improves precipitation predictability. Without this input, R^2 scores remain low across regions and rain types (convective-only and total), and for both the "Full Range" and "Extreme" subsets.

Including rain area markedly boosts performance, particularly for the "Full Range," with R^2 increasing by about 0.6 for total precipitation and 0.5 for convective-only cases. This confirms the strong predictive power of rain-area for rain rate prediction. For extreme events, the gain is more modest—around 0.25 in R^2 on average. In the WMP region, for example, the R^2 for extremes increases from 0.45 to 0.65, while the full-range R^2 rises more dramatically from approximately 0.2 to 0.85.

The smaller improvement for extremes partly reflects their relatively higher predictability even without explicit rain–area information. More importantly, because extremes are defined as the upper percentiles of precipitation within each precipitable water bin, this conditioning preferentially selects situations in which the thermodynamic environment already strongly favors intense convection. Consistent with this, the rain-area distribution evolves from a highly skewed, gamma-like form across all events (Fig. 2) to a more Gaussian distribution with a larger mean for extreme events, indicating reduced relative variability once extremes are selected. In this regime, most snapshots are already characterized by large rain coverage, so variations in rain area provide limited additional information at the grid-cell scale. Nevertheless, including rain-area information still adds value by conveying aspects of system size and organization. Further improvements in predicting extremes may therefore require descriptors that explicitly capture storm structure and convective organization, beyond rain area alone (Shamekh et al. 2023; Angulo-Umana and Kim 2023; Colin et al. 2019). These results motivate the development of a model that directly represents the distribution of rain area as a function of environmental conditions. We also note that while rain area clearly emerges as a strong source of information for rain-rate prediction, part of this correspondence may reflect their common observational origin which includes the detailed convective state of the atmosphere, whereas the ERA5-based predictors provide a less direct estimate of the atmospheric state.

APPENDIX B

Role of Vertical Velocity

Vertical velocity at 500 hPa is included in our list of candidate input variables, but it is not used in the main analysis. The reason is its two-way relationship with convection: on the one hand, upward motion drives vertical transport of moisture, cloud formation, and ultimately precipitation; on the other hand, the presence of strong upward motion is itself a diagnostic indicator of ongoing convective activity. Relying on this variable therefore risks conflating cause and effect, and could lead the algorithm to exploit vertical velocity as a direct proxy for convection rather than learning the underlying physical controls from environmental state variables.

For completeness, however, we also examine how including vertical velocity in the input set affects the results. This sensitivity test provides context on the predictive power of this variable and clarifies to what extent it dominates, complements, or is redundant with other large-scale predictors of rain area.

In the equation discovery framework, we explicitly penalize the use of vertical velocity by assigning a higher complexity cost to this variable. In PySR, equation complexity is computed by summing the predefined complexity values of variables, constants, and operators. During Pareto optimization, this complexity is balanced against accuracy, and only equations that are not simultaneously more complex and less accurate than another are retained. This design choice ensures that the algorithm resorts to vertical velocity only if no other predictor provides equivalent information and if the added complexity remains within the allowable range. In practice, when θ_e is included as an input, the algorithm attains high predictive skill (R^2) without relying on vertical velocity, typically through simple, mostly linear relationships. Nevertheless, allowing ω_{500} yields a modest improvement in skill—for example, an increase in R^2 from about 0.90 to 0.92 for stratiform convection. When θ_e is available, this improvement extends across the full distribution of rain-area predictions.

For CRH, the situation is different. Including ω_{500} often allows the algorithm to identify substantially simpler expressions that still capture the key features of the rain-area distribution. Without access to vertical velocity, models based on CRH alone struggle to reproduce the distribution tails when compared with the conditional normalizing flow benchmark. In such cases, the discovered equations must become highly nonlinear and complex in order to approximate the extremes. By contrast, incorporating ω_{500} enables both a simpler functional form and an accurate representation of the tails, highlighting its unique role in complementing CRH as a predictor.

To illustrate the effect of including ω_{500} when CRH is used as a predictor, we present two sets of plots showing the stratiform and convective rain-area distributions with and without ω_{500}. In addition to the distributions, we provide the corresponding discovered equations. These side-by-side comparisons highlight the trade-off between simplicity and accuracy: without ω_{500}, the equations become more nonlinear and complex in order to capture the distribution tails, whereas including ω_{500} yields both simpler functional forms and yet accurate representations of the extremes.

Deep Convective Area with CRH:

Expression with ω_{500}:

$$\left(\mathrm{CRH}_{\mathrm{ft}}\left((T_{\mathrm{ft}}-T_0)(\mathrm{CRH}_{\mathrm{ft}}-\mathrm{CRH}_c)+c\right)-\omega_{500}\right)^{+} \tag{B1}$$

where $T_0 = 264.0(K)$, $CRH_c = 0.68, c = 1.8$

Expression without ω_{500}:

$$\left(\mathrm{CRH}_{\mathrm{pbl}}\left(\mathrm{CRH}_{\mathrm{ft}}^2\left(c_1\mathrm{CRH}_{\mathrm{pbl}}+(T_{\mathrm{ft}}-T_0)(\mathrm{CRH}_{\mathrm{ft}}^2-CRH_0^2)\right)-c_2\right)+c_3\right)^{+} \tag{B2}$$

where $c_1 = 4.556177,\ ,T_0 = 267.3(K),\ CRH_0 = 0.75,\ c_2 = 1.7,\ c_3 = 1.3$

Stratiform area with CRH:

Expression with ω_{500}:

$$(\omega_0-\omega_{500})\left(c\mathrm{CRH}_{\mathrm{ft}}+T_0-T_{\mathrm{ft}}\right)^{+} \tag{B3}$$

where $\omega_0 = 1.15,\ c = 62.7,\ T_0 = 233.6(K)$.

Expression without ω_{500}:

$$c_0\left[c_1\,\mathrm{CRH}_{\mathrm{ft}}^4\ -\ \mathrm{CRH}_{\mathrm{ft}}^3(T_0-T_{\mathrm{ft}})^{+}\ +\ \mathrm{CRH}_{\mathrm{ft}}\left(T_1-T_{\mathrm{ft}}+\frac{c_2}{c_3-\mathrm{CRH}_{\mathrm{ft}}}\right)^{+}\right] \tag{B4}$$

where $c_0 = 2.86,\ c_1 = 7.1,\ T_0 = 265.6(K),\ T_1 = 252.6(K),\ c_2 = 11.35, c_3 = 1.32.$

References

Abramian, S., C. Muller, C. Risi, T. Fiolleau, and R. Roca, 2025: How key features of early development shape deep convective systems. *npj Climate and Atmospheric Science*, **8 (1)**, 258.

Agasthya, L., C. Muller, and M. Cheve, 2025: Moist convective scaling: Insights from an idealised model. *Quarterly Journal of the Royal Meteorological Society*, **151 (766)**, e4902, https://doi.org/

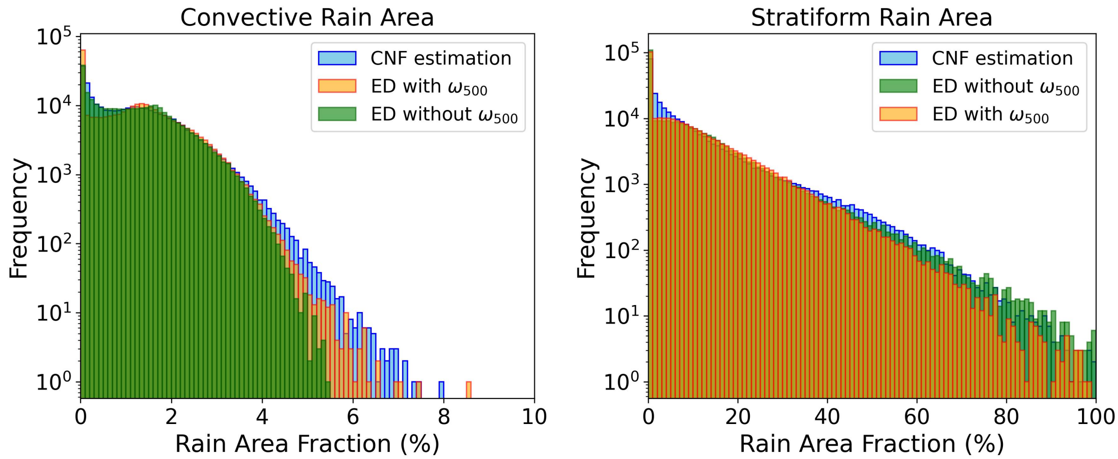

FIG. B1. Rain–area distributions predicted by equation discovery (ED) using CRH with and without ω_{500}, compared against the CNF benchmark (blue). Left: Convective area, corresponding to the CRH-based equations with ω_{500} (orange, Eq. (5)) and without ω_{500} (green, Eq. (5)). Right: Stratiform area, corresponding to the CRH-based equations with ω_{500} (orange, Eq. (5)) and without ω_{500} (green, Eq. (5)). Including ω_{500} yields substantially simpler equations that more accurately capture the tails of the distributions, while excluding it requires more complex expressions to approximate the same behavior.

https://doi.org/10.1002/qj.4902, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4902, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.4902.

Ahmed, F., and J. D. Neelin, 2018: Reverse engineering the tropical precipitation–buoyancy relationship. *Journal of the Atmospheric Sciences*, **75 (5)**, 1587 – 1608, https://doi.org/10.1175/JAS-D-17-0333.1, URL https://journals.ametsoc.org/view/journals/atsc/75/5/jas-d-17-0333.1.xml.

Ahmed, F., and C. Schumacher, 2015: Convective and stratiform components of the precipitation-moisture relationship. *Geophysical Research Letters*, **42 (23)**, 10,453–10,462, https://doi.org/https://doi.org/10.1002/2015GL066957, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GL066957, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015GL066957.

Angulo-Umana, P., and D. Kim, 2023: Mesoscale convective clustering enhances tropical precipitation. *Science Advances*, **9 (2)**, eabo5317, https://doi.org/10.1126/sciadv.abo5317, URL https://www.science.org/doi/abs/10.1126/sciadv.abo5317, https://www.science.org/doi/pdf/10.1126/sciadv.abo5317.

Arakawa, A., and W. H. Schubert, 1974: Interaction of a Cumulus Cloud Ensemble with the Large-Scale Environment, Part I. *Journal of Atmospheric Sciences*, **31 (3)**, 674–701, https://doi.org/10.1175/1520-0469(1974)031\textless{}0674:IOACCE\textgreater{}2.0.CO;2.

Awaka, J., T. Iguchi, H. Kumagai, and K. Okamoto, 1997: Rain type classification algorithm for trmm precipitation radar. *IGARSS'97. 1997 IEEE international geoscience and remote sensing symposium proceedings. remote sensing-a scientific vision for sustainable development*, IEEE, Vol. 4, 1633–1635.

Awaka, J., T. Iguchi, and K. Okamoto, 2009: Trmm pr standard algorithm 2a23 and its performance on bright band detection. *Japanese*, **87**, 31–52.

Berner, J., and Coauthors, 2017: Stochastic parameterization: Toward a new view of weather and climate models. *Bulletin of the American Meteorological Society*, **98 (3)**, 565 – 588, https://doi.org/https://doi.org/10.1175/BAMS-D-15-00268.1, URL https://journals.ametsoc.org/view/journals/bams/98/3/bams-d-15-00268.1.xml.

Betts, A. K., and M. J. Miller, 1986: A new convective adjustment scheme. part ii: Single column tests using gate wave, bomex, atex and arctic air-mass data sets. *Quarterly Journal of the Royal Meteorological Society*, **112 (473)**, 693–709, https://doi.org/https://doi.org/10.1002/qj.49711247308, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49711247308, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49711247308.

Beucler, T., A. Grundner, S. Shamekh, P. Ukkonen, M. Chantry, and R. Lagerquist, 2025: Distilling machine learning's added value: Pareto fronts in atmospheric applications. *Artificial Intelligence for the Earth Systems*, **4 (2)**, e240 078, https://doi.org/10.1175/AIES-D-24-0078.1, URL https://journals.ametsoc.org/view/journals/aies/4/2/AIES-D-24-0078.1.xml.

Bony, S., H. Schulz, J. Vial, and B. Stevens, 2020: Sugar, gravel, fish, and flowers: Dependence of mesoscale patterns of trade-wind clouds on environmental conditions. *Geophysical Research Letters*, **47 (7)**, e2019GL085 988, https://doi.org/https://doi.org/10.1029/2019GL085988, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.

1029/2019GL085988, e2019GL085988 10.1029/2019GL085988, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019GL085988.

Bony, S., B. Stevens, D. Coppin, T. Becker, K. A. Reed, A. Voigt, and B. Medeiros, 2016: Thermodynamic control of anvil cloud amount. *Proceedings of the National Academy of Sciences*, **113 (32)**, 8927–8932, https://doi.org/10.1073/pnas.1601472113, URL https://www.pnas.org/doi/abs/10.1073/pnas.1601472113, https://www.pnas.org/doi/pdf/10.1073/pnas.1601472113.

Bretherton, C. S., M. E. Peters, and L. E. Back, 2004: Relationships between water vapor path and precipitation over the tropical oceans. *Journal of climate*, **17 (7)**, 1517–1528.

Charn, A. B., W. D. Collins, H. Parishani, M. D. Risser, and T. A. O'Brien, 2020: Microphysical sensitivity of superparameterized precipitation extremes in the contiguous united states due to feedbacks on large-scale circulation. *Earth and Space Science*, **7 (7)**, e2019EA000 731, https://doi.org/https://doi.org/10.1029/2019EA000731, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019EA000731, e2019EA000731 10.1029/2019EA000731, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019EA000731.

Colin, M., S. Sherwood, O. Geoffroy, S. Bony, and D. Fuchs, 2019: Identifying the sources of convective memory in cloud-resolving simulations. *Journal of the Atmospheric Sciences*, **76 (3)**, 947–962.

Cranmer, M., 2023: Interpretable machine learning for science with pysr and symbolicregression.jl. URL https://arxiv.org/abs/2305.01582, 2305.01582.

Dai, A., 2006: Precipitation characteristics in eighteen coupled climate models. *Journal of climate*, **19 (18)**, 4605–4630.

Derbyshire, S. H., I. Beau, P. Bechtold, J.-Y. Grandpeix, J.-M. Piriou, J.-L. Redelsperger, and P. M. M. Soares, 2004: Sensitivity of moist convection to environmental humidity. *Quarterly Journal of the Royal Meteorological Society*, **130 (604)**, 3055–3079, https://doi.org/https://doi.org/10.1256/qj.03.130, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.03.130, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/qj.03.130.

Elsaesser, G. S., R. Roca, T. Fiolleau, A. D. Del Genio, and J. Wu, 2022: A simple model for tropical convective cloud shield area growth and decay rates informed by geostationary ir, gpm, and aqua/airs satellite data. *Journal of Geophysical Research: Atmospheres*, **127 (10)**, e2021JD035 599, https://doi.org/https://doi.org/10.1029/2021JD035599, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021JD035599, e2021JD035599 2021JD035599, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021JD035599.

Falga, R., S. Shamekh, and L. Zanna, 2025: Towards a unified data-driven boundary layer momentum flux parameterization for ocean and atmosphere. URL https://arxiv.org/abs/2511.01766, 2511.01766.

Flato, G., and Coauthors, 2014: Evaluation of climate models. *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, 741–866.

Gentine, P., M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis, 2018: Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, **45 (11)**, 5742–5751.

Grundner, A., T. Beucler, P. Gentine, and V. Eyring, 2024: Data-driven equation discovery of a cloud cover parameterization. *Journal of Advances in Modeling Earth Systems*, **16 (3)**, e2023MS003 763, https://doi.org/https://doi.org/10.1029/2023MS003763, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023MS003763, e2023MS003763 2023MS003763, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023MS003763.

Hafner, K., F. Iglesias-Suarez, S. Shamekh, P. Gentine, M. A. Giorgetta, R. Pincus, and V. Eyring, 2024: Interpretable machine learning-based radiation emulation for icon. *Authorea Preprints*.

Hartmann, D. L., and K. Larson, 2002: An important constraint on tropical cloud - climate feedback. *Geophysical Research Letters*, **29 (20)**, 12–1–12–4, https://doi.org/https://doi.org/10.1029/2002GL015835, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002GL015835, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2002GL015835.

Hersbach, H., and Coauthors, 2020: The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, **146 (730)**, 1999–2049.

Holloway, C. E., and J. D. Neelin, 2009: Moisture vertical structure, column water vapor, and tropical deep convection. *Journal of the atmospheric sciences*, **66 (6)**, 1665–1683.

Houze, R. A., 1997: Stratiform precipitation in regions of convection: A meteorological paradox? *Bulletin of the American Meteorological Society*, **78 (10)**, 2179 – 2196, https://doi.org/10.1175/1520-0477(1997)078⟨2179:SPIROC⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/bams/78/10/1520-0477_1997_078_2179_spiroc_2_0_co_2.xml.

Houze Jr, R. A., K. L. Rasmussen, M. D. Zuluaga, and S. R. Brodzik, 2015: The variable nature of convection in the tropics and subtropics: A legacy of 16 years of the tropical rainfall measuring mission satellite. *Reviews of Geophysics*, **53 (3)**, 994–1021.

Ito, M., and H. Masunaga, 2022: Process-level assessment of the iris effect over tropical oceans. *Geophysical Research Letters*, **49 (7)**, e2022GL097 997.

Klein, S. A., 1997: Synoptic variability of low-cloud properties and meteorological parameters in the subtropical trade wind boundary layer. *Journal of climate*, **10 (8)**, 2018–2039.

Klein, S. A., and D. L. Hartmann, 1993: The seasonal cycle of low stratiform clouds. *Journal of climate*, **6 (8)**, 1587–1606.

Kobyzev, I., S. J. Prince, and M. A. Brubaker, 2020: Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, **43 (11)**, 3964–3979.

Kooperman, G. J., M. S. Pritchard, T. A. O'Brien, and B. W. Timmermans, 2018: Rainfall from resolved rather than parameterized processes better represents the present-day and climate change response of moderate rates in the community atmosphere model. *Journal of advances in modeling earth systems*, **10 (4)**, 971–988.

Lenderink, G., R. Barbero, J. M. Loriaux, and H. J. Fowler, 2017: Super-clausius–clapeyron scaling of extreme hourly convective precipitation and its relation to large-scale atmospheric conditions. *Journal of Climate*, **30 (15)**, 6037 – 6052, https://doi.org/10.1175/JCLI-D-16-0808.1, URL https://journals.ametsoc.org/view/journals/clim/30/15/jcli-d-16-0808.1.xml.

Lochbihler, K., G. Lenderink, and A. P. Siebesma, 2021: Cold pool dynamics shape the response of extreme rainfall events to climate change. *Journal of Advances in Modeling Earth Systems*, **13 (2)**, e2020MS002 306, https://doi.org/https://doi.org/10.1029/2020MS002306, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002306, e2020MS002306 2020MS002306, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002306.

Manabe, S., and R. F. Strickler, 1964: Thermal equilibrium of the atmosphere with a convective adjustment. *Journal of Atmospheric Sciences*, **21 (4)**, 361–385.

Miller, G. A., P. Stier, and H. M. Christensen, 2025: Characterizing uncertainty in deep convection triggering using explainable machine learning. *Journal of the Atmospheric Sciences*, **82 (6)**, 1093 – 1111, https://doi.org/10.1175/JAS-D-24-0085.1, URL https://journals.ametsoc.org/view/journals/atsc/82/6/JAS-D-24-0085.1.xml.

Mission, T. R. M., 2011: Trmm precipitation radar rainfall rate and profile l2 1.5 hours v7, greenbelt, md, goddard space flight center earth sciences data and information services center (ges disc).

Mooers, G., M. Pritchard, T. Beucler, J. Ott, G. Yacalis, P. Baldi, and P. Gentine, 2021: Assessing the potential of deep learning for emulating cloud superparameterization in climate models with real-geography boundary conditions. *Journal of Advances in Modeling Earth Systems*, **13 (5)**, e2020MS002 385, https://doi.org/https://doi.org/10.1029/2020MS002385, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002385, e2020MS002385 2020MS002385, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002385.

Morrison, H., and Coauthors, 2020: Confronting the challenge of modeling cloud and precipitation microphysics. *Journal of Advances in Modeling Earth Systems*, **12 (8)**, e2019MS001 689, https://doi.org/https://doi.org/10.1029/2019MS001689, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001689, e2019MS001689 2019MS001689, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001689.

Muller, C., and Coauthors, 2022: Spontaneous aggregation of convective storms. *Annual Review of Fluid Mechanics*, **54 (1)**, 133–157, https://doi.org/10.1146/

annurev-fluid-022421-011319, URL https://doi.org/10.1146/annurev-fluid-022421-011319, https://doi.org/10.1146/annurev-fluid-022421-011319.

Muller, C. J., L. E. Back, P. A. O'Gorman, and K. A. Emanuel, 2009: A model for the relationship between tropical precipitation and column water vapor. *Geophysical Research Letters*, **36 (16)**, https://doi.org/https://doi.org/10.1029/2009GL039667, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009GL039667, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2009GL039667.

Muller, C. J., P. A. O'Gorman, and L. E. Back, 2011: Intensification of precipitation extremes with warming in a cloud-resolving model. *Journal of Climate*, **24 (11)**, 2784–2800.

Neelin, J. D., O. Peters, and K. Hales, 2009: The transition to strong convection. *Journal of the Atmospheric Sciences*, **66 (8)**, 2367–2384.

Nuijens, L., B. Stevens, and A. P. Siebesma, 2009: The environment of precipitating shallow cumulus convection. *Journal of the Atmospheric Sciences*, **66 (7)**, 1962 – 1979, https://doi.org/10.1175/2008JAS2841.1, URL https://journals.ametsoc.org/view/journals/atsc/66/7/2008jas2841.1.xml.

Orenstein, P., A. H. Sobel, S. J. Camargo, G. S. Elsaesser, and P. Garg, 2025: The global seasonal relationship between satellite-observed cold pools and rainfall. *Journal of Climate*, **38 (4)**, 989 – 1003, https://doi.org/10.1175/JCLI-D-23-0218.1, URL https://journals.ametsoc.org/view/journals/clim/38/4/JCLI-D-23-0218.1.xml.

Papamakarios, G., T. Pavlakou, and I. Murray, 2017: Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, **30**.

Retsch, M. H., C. Jakob, and M. S. Singh, 2022: Identifying relations between deep convection and the large-scale atmosphere using explainable artificial intelligence. *Journal of Geophysical Research: Atmospheres*, **127 (3)**, e2021JD035 388, https://doi.org/https://doi.org/10.1029/2021JD035388, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021JD035388, e2021JD035388 2021JD035388, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021JD035388.

Rezende, D., and S. Mohamed, 2015: Variational inference with normalizing flows. *International conference on machine learning*, PMLR, 1530–1538.

Rieck, M., L. Nuijens, and B. Stevens, 2012: Marine boundary layer cloud feedbacks in a constant relative humidity atmosphere. *Journal of the Atmospheric Sciences*, **69 (8)**, 2538 – 2550, https://doi.org/10.1175/JAS-D-11-0203.1, URL https://journals.ametsoc.org/view/journals/atsc/69/8/jas-d-11-0203.1.xml.

Robe, F. R., and K. A. Emanuel, 1996: Moist convective scaling: Some inferences from three-dimensional cloud ensemble simulations. *Journal of Atmospheric Sciences*, **53 (22)**, 3265–3275.

Saint-Lu, M., S. Bony, and J.-L. Dufresne, 2020: Observational evidence for a stability iris effect in the tropics. *Geophysical Research Letters*, **47 (14)**, e2020GL089 059.

Sane, A., B. Reichl, A. Adcroft, and L. Zanna, 2025: Machine learned equations for vertical mixing in the ocean surface boundary layer. https://doi.org/10.31219/osf.io/uab7v_v1.

Schiro, K. A., and J. D. Neelin, 2019: Deep convective organization, moisture vertical structure, and convective transition using deep-inflow mixing. *Journal of the Atmospheric Sciences*, **76 (4)**, 965–987.

Schulz, H., R. Eastman, and B. Stevens, 2021: Characterization and evolution of organized shallow convection in the downstream north atlantic trades. *Journal of Geophysical Research: Atmospheres*, **126 (17)**, e2021JD034 575, https://doi.org/https://doi.org/10.1029/2021JD034575, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021JD034575, e2021JD034575 2021JD034575, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021JD034575.

Schumacher, C., and A. Funk, 2023: Assessing convective-stratiform precipitation regimes in the tropics and extratropics with the gpm satellite radar. *Geophysical Research Letters*, **50 (14)**, e2023GL102 786, https://doi.org/https://doi.org/10.1029/2023GL102786, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023GL102786, e2023GL102786 2023GL102786, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023GL102786.

Shamekh, S., and P. Gentine, 2023: Learning atmospheric boundary layer turbulence. https://doi.org/10.22541/essoar.168748456.60017486/v1, URL http://dx.doi.org/10.22541/essoar.168748456.60017486/v1.

Shamekh, S., K. D. Lamb, Y. Huang, and P. Gentine, 2023: Implicit learning of convective organization explains precipitation stochasticity. *Proceedings of the National Academy of Sciences*, **120 (20)**, e2216158 120, https://doi.org/10.1073/pnas.2216158120, URL https://www.pnas.org/doi/abs/10.1073/pnas.2216158120, https://www.pnas.org/doi/pdf/10.1073/pnas.2216158120.

Singh, M. S., and P. A. O'Gorman, 2014: Influence of microphysics on the scaling of precipitation extremes with temperature. *Geophysical Research Letters*, **41 (16)**, 6037–6044, https://doi.org/https://doi.org/10.1002/2014GL061222, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014GL061222, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014GL061222.

Stanfield, R. E., H. Su, J. H. Jiang, S. R. Freitas, A. M. Molod, Z. J. Luo, L. Huang, and M. Luo, 2019: Convective entrainment rates estimated from aura co and cloudsat/calipso observations and comparison with geos-5. *Journal of Geophysical Research: Atmospheres*, **124 (17-18)**, 9796–9807.

Stauffer, C. L., and A. A. Wing, 2022: Properties, changes, and controls of deep-convecting clouds in radiative-convective equilibrium. *Journal of Advances in Modeling Earth Systems*, **14 (6)**, e2021MS002 917, https://doi.org/https://doi.org/10.1029/2021MS002917, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002917, e2021MS002917 2021MS002917, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021MS002917.

Tompkins, A. M., 2001a: On the relationship between tropical convection and sea surface temperature. *Journal of Climate*, **14 (5)**, 633 – 637, https://doi.org/10.1175/1520-0442(2001)014⟨0633:OTRBTC⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/clim/14/5/1520-0442_2001_014_0633_otrbtc_2.0.co_2.xml.

Tompkins, A. M., 2001b: Organization of tropical convection in low vertical wind shears: The role of water vapor. *Journal of the Atmospheric Sciences*, **58 (6)**, 529 – 545, https://doi.org/10.

1175/1520-0469(2001)058⟨0529:OOTCIL⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/atsc/58/6/1520-0469_2001_058_0529_ootcil_2.0.co_2.xml.

Tompkins, A. M., and G. C. Craig, 1998: Radiative–convective equilibrium in a three-dimensional cloud-ensemble model. *Quarterly Journal of the Royal Meteorological Society*, **124 (550)**, 2073–2097.

Tompkins, A. M., and A. G. Semie, 2017: Organization of tropical convection in low vertical wind shears: Role of updraft entrainment. *Journal of Advances in Modeling Earth Systems*, **9 (2)**, 1046–1068, https://doi.org/https://doi.org/10.1002/2016MS000802, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016MS000802, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2016MS000802.

Wilcox, E. M., and L. J. Donner, 2007: The frequency of extreme rain events in satellite rain-rate estimates and an atmospheric general circulation model. *Journal of Climate*, **20 (1)**, 53–69.

Windmiller, J., and C. Hohenegger, 2019: Convection on the edge. *Journal of Advances in Modeling Earth Systems*, **11 (12)**, 3959–3972.

Winkler, C., D. Worrall, E. Hoogeboom, and M. Welling, 2019: Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*.

Wood, R., and C. S. Bretherton, 2006: On the relationship between stratiform low cloud cover and lower-tropospheric stability. *Journal of Climate*, **19 (24)**, 6425 – 6432, https://doi.org/10.1175/JCLI3988.1, URL https://journals.ametsoc.org/view/journals/clim/19/24/jcli3988.1.xml.

Wood, R., and D. L. Hartmann, 2006: Spatial variability of liquid water path in marine low cloud: The importance of mesoscale cellular convection. *Journal of Climate*, **19 (9)**, 1748–1764.

Yuval, J., and P. A. O'Gorman, 2020: Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature communications*, **11 (1)**, 1–10.

Zanna, L., and T. Bolton, 2020: Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, **47 (17)**, e2020GL088 376, https://doi.org/https://doi.org/10.1029/2020GL088376, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.

1029/2020GL088376, e2020GL088376 10.1029/2020GL088376, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020GL088376.

Zelinka, M. D., S. A. Klein, K. E. Taylor, T. Andrews, M. J. Webb, J. M. Gregory, and P. M. Forster, 2013: Contributions of different cloud types to feedbacks and rapid adjustments in cmip5. *Journal of Climate*, **26 (14)**, 5007 – 5027, https://doi.org/10.1175/JCLI-D-12-00555.1, URL https://journals.ametsoc.org/view/journals/clim/26/14/jcli-d-12-00555.1.xml.

Zipser, E. J., and C. Liu, 2021: Extreme convection vs. extreme rainfall: A global view. *Current Climate Change Reports*, **7 (4)**, 121–130.

Zuidema, P., G. Torri, C. Muller, and A. Chandra, 2017: A survey of precipitation-induced atmospheric cold pools over oceans and their interactions with the larger-scale environment. *Surveys in Geophysics*, **38 (6)**, 1283–1305.