Distilling Machine Learning's Added Value: Pareto Fronts in Atmospheric Applications

TOM BEUCLER[a,b] , ARTHUR GRUNDNER[c] , SARA SHAMEKH[d] , PETER UKKONEN[e] , MATTHEW CHANTRY[f] , RYAN LAGERQUIST[g,h]

[a] *Faculty of Geosciences and Environment, University of Lausanne, Lausanne, VD, Switzerland*
[b] *Expertise Center for Climate Extremes, University of Lausanne, Lausanne, VD, Switzerland*
[c] *Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany*
[d] *Courant Institute of Mathematical Sciences, New York University, New York, NY, USA*
[e] *Department of Physics, University of Oxford, Oxford, United Kingdom*
[f] *European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom*
[g] *Cooperative Institute for Research in the Atmosphere (CIRA), Colorado State University, Fort Collins, CO, USA*
[h] *National Oceanic and Atmospheric Administration (NOAA) Global Systems Laboratory (GSL), Boulder, CO, USA*

ABSTRACT: While the added value of machine learning for weather and climate applications is measurable, explaining it remains challenging, especially for large deep learning models. Inspired by climate model hierarchies, we propose that a full hierarchy of Pareto-optimal models, defined within an appropriately determined error-complexity plane, can guide model development and help understand the models' added value. We demonstrate the use of Pareto fronts in atmospheric physics through three sample applications, with hierarchies ranging from semi-empirical models with minimal tunable parameters (simplest) to deep learning algorithms (most complex). First, in cloud cover parameterization, we find that neural networks identify nonlinear relationships between cloud cover and its thermodynamic environment, and assimilate previously neglected features such as vertical gradients in relative humidity that improve the representation of low cloud cover. This added value is condensed into a ten-parameter equation that rivals the performance of deep learning models. Second, we establish a machine learning model hierarchy for emulating shortwave radiative transfer, distilling the importance of bidirectional vertical connectivity for accurately representing absorption and scattering, especially for multiple cloud layers. Third, we emphasize the importance of convective organization information when modeling the relationship between tropical precipitation and its surrounding environment. We discuss the added value of temporal memory when high-resolution spatial information is unavailable, with implications for precipitation parameterization. Therefore, by comparing data-driven models directly with existing schemes using Pareto optimality, we promote process understanding by hierarchically unveiling system complexity, with the hope of improving the trustworthiness of machine learning models in atmospheric applications.

SIGNIFICANCE STATEMENT: With machine learning permeating the geosciences, it becomes urgent to distinguish incremental progress from enduring knowledge discovery. We show that Pareto-optimal hierarchies transparently distill the added value of new algorithms using three atmospheric physics applications, providing a timely complement to post-hoc explainable artificial intelligence tools.

1. Introduction

Have recent advancements in machine learning (ML) led to discoveries in atmospheric science? The added value of ML for weather and climate applications is measurable, but it often remains challenging to understand. Taking advancements in data-driven, medium-range weather forecasting (Ben-Bouallegue et al. 2023) as an example, increasing reliance on complex architectures makes state-of-the-art models difficult to interpret.

Weyn et al. (2019, 2020) used convolutional neural networks (CNN) with approximately 200k and 700k learned parameters to produce global forecasts that outperformed climatology, persistence, and a low-resolution numerical weather prediction (NWP) model for lead times smaller than one week. Following the success of early approaches, Rasp et al. (2020) developed a benchmark dataset for data-driven weather forecasting, which facilitated the objective assessment of rapid developments (e.g., Clare et al. 2021; Scher and Messori 2021). Since then, for data-driven, medium-range forecasting, Rasp and Thuerey (2021) trained a $\approx$ 6.3M-parameter deep residual CNN, Keisler (2022) trained a $\approx$ 6.7M-parameter graph neural network (GNN), and Pathak et al. (2022) trained a $\approx$ 75M-parameter emulator combining transformers with Fourier neural operators. Recently, deep learning models started rivaling state-of-the-art, high-resolution, deterministic NWP models: Lam et al. (2022) via combined GNNs totaling $\approx$ 37M parameters, Bi et al. (2022) via a $\approx$ 256M-parameter Earth-like transformer, and Lang et al. (2024) via a $\approx$ 256M-parameter graph and transformer model. It is hard to pinpoint what makes these models so successful, even with modern explainable artificial intelligence (XAI; Buhrmester et al. 2021; Das and Rad 2020) tools, given that XAI requires certain assumptions to be satisfied (Mamalakis et al. 2022, 2023) and involves choosing which samples to investigate, which is challenging for large models and datasets.

Corresponding author: Tom Beucler, tom.beucler@unil.ch

The growing complexity of data-driven models for weather applications shares similarities with the development of general circulation models (GCM) that followed the first comprehensive assessment of global climate change due to carbon dioxide (Charney et al. 1979). Unlike data-driven weather prediction, where reducing forecast errors could warrant increased complexity, GCMs have traditionally been created to not simply project but also comprehend climate changes (Held 2005; Balaji et al. 2022). This implies that any additional complexity in an Earth system model should be well-justified, motivating *climate model hierarchies* that aim to connect our fundamental understanding with model prediction (e.g., Mansfield et al. 2023; Robertson and Ghil 2000; Bony et al. 2013; Jeevanjee et al. 2017; Maher et al. 2019; Balaji 2021).

Inspired by climate model hierarchies, we here show that modern optimization tools help systematically generate *data-driven model hierarchies* to model and understand climate processes for which we have reliable data. These hierarchies can: (1) guide the development of data-driven models that optimally balance simplicity and accuracy; and (2) unveil the role of each complexity unit, furthering process understanding by distilling the added value of ML for the atmospheric application of interest.

In this study, we showcase the advantages of considering a hierarchy of models with varying error and complexity, as opposed to focusing on a single fitted model (Fisher et al. 2019). After formulating Pareto-optimal model hierarchies (Sec. 2) and categorizing the added value of ML into four categories (Sec. 3), we apply our approach to three atmospheric processes relevant for weather and climate predictions (Sec. 4) to distill the added value of recently developed deep learning frameworks before concluding (Sec. 5).

2. Pareto-Optimal Model Hierarchies

In this section, we define Pareto optimality before discussing the definition of error and complexity.

a. Pareto Optimality

In multi-objective optimization, Pareto optimality represents a solutions set that cannot be improved upon in one criterion without worsening another criterion (e.g., Censor 1977; Miettinen 1999). The first step is to define a set of n real-valued model evaluation metrics $\mathcal{E} = \{\mathcal{E}_i\}_{i=1}^{n}$ that we wish to minimize (e.g., error, complexity). We call a model M_{opt} *Pareto-optimal* w.r.t. these metrics and w.r.t. a model family if there is no model in that family that strictly outperforms M_{opt} in one metric while maintaining at least the same performance in all other metrics (e.g., Lin et al. 2019). The *Pareto front* (PF) is the set of all Pareto-optimal models, which can be defined using logical statements:

$$\text{PF}_{\mathcal{E}} = \left\{ M_{\text{opt}} \mid \nexists M \text{ s.t. } \begin{cases} \forall i \, \mathcal{E}_i(M) \leq \mathcal{E}_i(M_{\text{opt}}) \\ \exists j \, \mathcal{E}_j(M) < \mathcal{E}_j(M_{\text{opt}}) \end{cases} \right\}. \tag{1}$$

Intuitively, when we select a model from the Pareto front, any attempt to switch to a different model would mean sacrificing the quality of at least one evaluation metric. Conversely, a model that can be replaced without compromising any evaluation metrics is described as *Pareto-dominated*. In practice, we often seek to balance evaluation metrics measuring error and complexity, which we discuss in the next subsections.

b. Error

We emphasize the importance of *holistic* evaluation, which employs several error metrics with different behaviors: Traditional regression or classification metrics, distributional distances, spectral metrics, probabilistic scoring rules, reliability diagrams (Haynes et al. 2023), causal evaluation (Nowack et al. 2020), etc. To facilitate the use of Pareto-optimal hierarchies, we recommend prioritizing proper scores, whose expectation is optimal if and only if the model represents the true target distribution (Bröcker 2009). For simplicity's sake, we will employ mean squared error (MSE) as our primary error metric for our study's applications. We make this choice because MSE is a proper score for deterministic predictions that can be efficiently optimized, while recognizing MSE's inherent limitations for non-normally distributed targets.

c. Complexity

To our knowledge, there are no universally accepted metrics for quantifying the complexity of data-driven models within the geosciences. In statistical learning, various complexity metrics, such as Rademacher complexity (e.g., Bartlett et al. 2005), rely on dataset characteristics, whereas others, like the VC dimension (Vapnik and Chervonenkis 2015), solely depend on algorithmic attributes. Here, we predominantly focus on two metrics that can be readily calculated from model attributes: the number of trainable parameters (N_{param}) and the number of features (N_{features}). We choose these metrics due to their simplicity and versatility across the broad spectrum of models considered in our study. As we will confirm empirically in Sec. 4, N_{param} and N_{features} can be used as (very) approximate proxies for generalizability, as models with fewer trainable parameters and features with more stable distributions tend to result in superior generalizability. We defer the exploration of additional complexity metrics, such as the number of floating point operations (FLOP), to future work.

Equipped with these definitions, we can now ask: Why, along the Pareto front in a well defined error-complexity

plane, does increasing complexity result in better performance? In the following section, we hence leverage Pareto optimality to distill machine learning's added value.

3. The Distillable Value of Machine Learning

Amid rapid progress in optimization, machine learning architectures, and data handling, it can be challenging to distinguish long-lasting progress in modeling from small improvements in model error. We postulate that progress is more likely to be replicable if we can *explain* how added model complexity improves performance in simple terms. Based on this postulate, we propose simple definitions to categorize a model's added value in the geosciences.

In geoscientific applications, we often work with features X that are functions of space, discretized in N_x spatial locations x, and of time, discretized in N_t timesteps t. We will consider a model M predicting a *target* vector Y from a multi-variate spatiotemporal field $X_{x,t}$ such that $Y = M[X_{x,t}]$. We define a model M as having added value w.r.t. a set of evaluation metrics $\{\mathcal{E}_i\}_{i=1}^n$ and a simple baseline model $\widehat{M}$ if the following is true: When applied to representative out-of-sample data, M is Pareto-optimal while the baseline model $\widehat{M}$ is not. Taking the data's spatiotemporal structure into account, we organize a model's added value into four mutually non-exclusive categories: functional representation, feature assimilation, spatial connectivity, and temporal connectivity (see Fig. 1).

a. Functional Representation

A fundamental aspect a model can improve is the functional representation of the observed relationship between a set of baseline features $\widehat{X}$—*fixed* features used to benchmark performance—and the target Y. We deem M to improve the functional representation of a baseline $\widehat{M}$ w.r.t. $\mathcal{E}$ if and only if:

$$\widehat{M}\left[\widehat{X}\right] \notin \mathrm{PF}_{\mathcal{E}} \;,\; M\left[\widehat{X}\right] \in \mathrm{PF}_{\mathcal{E}} \tag{2}$$

This improvement may stem from algorithms leading to better fits (e.g., gradient boosting instead of decision trees), improved optimization (e.g., the Adam optimizer and its variants instead of traditional stochastic gradient descent; Kingma and Ba 2014), improved parsimony (e.g., by decreasing the number of trainable parameters via hyperparameter tuning) or enforced constraints (e.g., positive concentrations and precipitation). Improvements in functional representation are readily visualized via partial dependence plots or their variants (marginal plots to avoid unlikely data instances, accumulated local effects to account for feature correlation, etc.; Molnar 2020). Note that Eq. 2 also captures improvements in probabilistic modeling by generalizing M to a stochastic mapping and including probabilistic scores (Gneiting and Raftery 2007; Haynes

et al. 2023) in $\mathcal{E}$. From an atmospheric science perspective, improving functional representation helps identify nonlinear regimes, identify model extremes, and describe how sensitive the prediction is to different features. However, it can be challenging to faithfully describe these sensitivities if key features are missing from the baseline set of features $\widehat{X}$.

b. Feature Assimilation

This motivates discussing the ability of a model to extract relevant information from a new feature set, which we refer to as *feature assimilation* and define as follows:

$$\widehat{M}\left[\widehat{X}\right] \notin \mathrm{PF}_{\mathcal{E}} \;,\; M[X] \in \mathrm{PF}_{\mathcal{E}} \tag{3}$$

where we emphasize the difference between the baseline $(\widehat{X})$ and new (X) sets of features. This improvement may stem from previously unconsidered features (e.g., variables whose quality has increased in recent datasets), the ability of models to handle features whose information could not be extracted in past attempts (e.g., the ability of deep learning to extract nonlinear relationships), or the discovery of a more compact feature set that facilitates learning or improves generalizability. The assimilation of new features may improve interpretability (e.g, if some features simplify the target's functional representation or if there are less features to consider), physical consistency (e.g., if the new feature set improves generalizability across regimes or consistency with physical laws), and predictability. While the features' spatial and temporal discretizations could technically be considered part of feature selection, we choose to discuss them separately in the following subsections, given the central role of space and time in geoscientific applications.

c. Spatial connectivity

To formalize a model M's improved ability to leverage the features' spatial information, we distinguish the spatial locations $\widehat{x}$ used to discretize the baseline $\widehat{M}$'s features from the spatial locations x used to discretize M's features:

$$\widehat{M}\left[\widehat{X}_{\widehat{x},\widehat{t}}\right] \notin \mathrm{PF}_{\mathcal{E}} \;,\; M\left[\widehat{X}_{x,\widehat{t}}\right] \in \mathrm{PF}_{\mathcal{E}} \tag{4}$$

This improvement may stem from the ability to:

1. handle features at different spatial resolutions (e.g., via improved pre-processing or handling of data). In atmospheric science, this can help consider multi-scale interaction and accommodate data from various Earth system models.

2. hierarchically process spatially adjacent data (e.g., via convolutional layers). In atmospheric science,
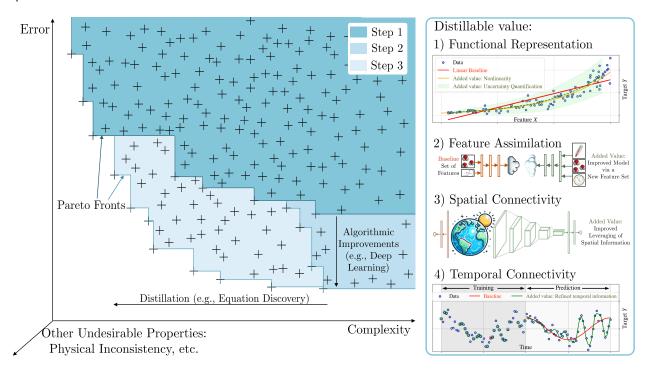
FIG. 1. Exploring Pareto fronts (sets of Pareto-optimal models) within a complexity-error plane highlights machine learning's added value. Crosses in step 1 denote existing models. Algorithms such as deep learning allow for the creation of efficient, low-error, albeit complex models (step 2). Knowledge distillation, through methods such as equation discovery, aims to explain error reduction, resulting in simpler, low-error models (step 3) and long-lasting scientific progress. For atmospheric applications, we argue that this added value can be classified into four categories: functional representation, feature assimilation, spatial connectivity, and temporal connectivity.

this acknowledges the high correlation between spatial neighbors due to, e.g., small-scale mixing.

3. capture long-range spatial dependencies (e.g., via self-attention mechanisms or a graph structure), such as teleconnections in the Earth system.

d. Temporal connectivity

To formalize a model M's improved ability to leverage the features' temporal information, we distinguish the timesteps $\widehat{t}$ used to discretize the baseline $\widehat{M}$'s features from the timesteps t used to discretize M's features:

$$\widehat{M}\left[\widehat{\boldsymbol{X}}_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{t}}}\right] \notin \mathrm{PF}_{\mathcal{E}} \, , \, M\left[\widehat{\boldsymbol{X}}_{\widehat{\boldsymbol{x}},\boldsymbol{t}}\right] \in \mathrm{PF}_{\mathcal{E}} \tag{5}$$

This improvement may stem from the ability to:

1. handle features at different temporal resolutions (e.g., via improved pre-processing or handling of data). This can help consider multiple timescales and accommodate data from various Earth system models.

2. process consecutive timesteps (e.g., via recurrent layers). This acknowledges the high temporal autocorrelation of Earth system data, which is a property of the underlying dynamical system.

3. capture long-term temporal dependencies (e.g., via gating or self-attention mechanisms) and cyclic patterns (e.g., via data adjustments and temporal Fourier transforms) such as the diurnal and seasonal cycles.

4. Atmospheric Physics Application Cases

Building upon the four types of added value outlined in the previous section, this section demonstrates that Pareto optimality guides model development and improves process understanding through three realistic, atmospheric modeling case studies. Each case includes machine learning prototypes with demonstrated performance from previous studies, along with Pareto-optimal models newly trained for this study.

a. Cloud Cover Parameterization for Climate Modeling

1) MOTIVATION

The incorrect representation of cloud processes in current Earth system models, with grid spacing of approximately 50-100 km (Arias et al. 2021), significantly contributes to structural uncertainty in long-term climate projections (Bony et al. 2015; Sherwood et al. 2014). Cloud cover parameterization, which maps environmental conditions at the grid scale to the fraction of the grid cell

occupied by clouds, directly affects radiative transfer and microphysical conversion rates, influencing the model's energy balance and water species concentrations.

Storm-resolving simulations with grid spacing below ≈ 5 km reduce uncertainty related to the interaction between storms and planetary-scale dynamics by explicitly simulating deep convection (Stevens et al. 2020). However, their large computational cost prohibits their routine use for ensemble projections (Schneider et al. 2017). Machine learning can learn the storm-scale behavior of clouds from short, storm-resolving simulations, potentially improving coarser Earth system models through data-driven parameterizations (Gentine et al. 2021).

In this section, we aim to understand the improvement gained from the higher-fidelity representation of storms and clouds. As illustrated in Fig. 2, we demonstrate that this knowledge can be symbolically distilled into an analytic equation that rivals the performance of deep learning.

2) Setup

We follow the setup described in Grundner et al. (2024), to which readers are referred for details. Fields are coarse-grained from storm-resolving ICON simulations (Giorgetta et al. 2018) conducted as part of the DYAMOND inter-comparison project (Stevens et al. 2019; Duras et al. 2021). The original simulations use a horizontal grid spacing of ≈ 2.5 km and 58 vertical layers below 21 km (the maximum altitude with clouds in the dataset). They are coarse-grained to a typical climate model resolution of ≈ 80 km horizontally and 27 vertical layers, converting the binarized, high-resolution condensate field (1 if the cloud condensate mixing ratio exceeds 10^{-6} kg/kg, 0 otherwise) into a fractional area cloud cover C (unitless).

To prevent strong correlations between the training and validation sets, the union of the "DYAMOND Summer" (Aug 10 to Sep 10, 2016) and "DYAMOND Winter" (Jan 30 to Feb 29, 2020) datasets was partitioned into six consecutive temporal segments. The second segment (approximately Aug 21 to Sep 1, 2016) and the fifth segment (approximately Feb 9–19, 2020) form the validation set. For all models, excluding traditional methods, the features are standardized to have a mean of zero and a standard deviation of one within the training set.

Once coarse-grained and pre-processed, we aim to map the environmental conditions $\boldsymbol{X}$ on vertical levels indexed by the background terrain-following height grid $\boldsymbol{z}$, to the cloud cover C on the same vertical levels. The variables $\boldsymbol{X}$ include the horizontal wind speed $\boldsymbol{U}$ [m/s], specific humidity $\boldsymbol{q_v}$ [kg/kg], liquid water mixing ratio $\boldsymbol{q_\ell}$ [kg/kg], ice mixing ratio $\boldsymbol{q_i}$ [kg/kg], temperature $\boldsymbol{T}$ [K], pressure $\boldsymbol{p}$ [Pa], and relative humidity $\mathbf{RH}$. Except for the "non-local NN" in Sec. 4.a.4, we simplify the mapping to a "vertically quasi-local" one where cloud cover at a given level depends only on the atmospheric variables $\boldsymbol{X}$ at the same level and

their first and second-order derivative with respect to $\boldsymbol{z}$. $\boldsymbol{X}$ also includes geometric height z [m] and surface variables: land fraction σ_f [%] and surface pressure p_s [Pa]. In summary, we approximate some mapping:

$$\underbrace{\left(\boldsymbol{X}, \frac{d\boldsymbol{X}}{d\boldsymbol{z}}, \frac{d^2\boldsymbol{X}}{d\boldsymbol{z}^2}, z, \sigma_f, p_s\right)}_{\in \mathbb{R}^{3\times 7+3}} \mapsto \underbrace{C}_{\in [0,1]} \qquad (6)$$

using a hierarchy of machine learning models. In the following subsections, we will show that Pareto-optimal hierarchies not only facilitate data-driven model development, allowing for the comparison of simple baselines with neural networks, but also promote process understanding when tracing the Pareto front.

3) Existing Baselines and Polynomial Regression

We start with the Sundqvist baseline (red star in Fig. 2; Sundqvist et al. 1989), the standard cloud cover parameterization in ICON. The Sundqvist parameterization implemented in ICON represents cloud cover as a monotonically increasing function of relative humidity, provided it exceeds a critical threshold $\mathrm{RH_{crit}}$ (Roeckner et al. 1996):

$$\mathrm{RH_{crit}} \overset{\mathrm{def}}{=} \mathrm{RH_{top}} + \left(\mathrm{RH_{surf}} - \mathrm{RH_{top}}\right) \exp\left(1 - [p_s/p]^n\right), \qquad (7)$$

where our implementation includes 4 tunable parameters: $\{\mathrm{RH_{surf}}, \mathrm{RH_{top}}, \mathrm{RH_{sat}}, n\}$, with values listed in appendix A. When $\mathrm{RH} > \mathrm{RH_{crit}}$, cloud cover is given by the model:

$$M_{\mathrm{Sundqvist}} : (p, p_s, \mathrm{RH}) \mapsto C_{\mathrm{Sundqvist}}, \qquad (8)$$

whose output is:

$$C_{\mathrm{Sundqvist}} \overset{\mathrm{def}}{=} 1 - \sqrt{\frac{\min\{\mathrm{RH}, \mathrm{RH_{sat}}\} - \mathrm{RH_{sat}}}{\mathrm{RH_{crit}} - \mathrm{RH_{sat}}}}. \qquad (9)$$

To account for marine stratocumulus (low) clouds (Mauritsen et al. 2019), we use two different sets of 4 parameters for land and sea using a land fraction threshold of 0.5. The Sundqvist scheme is parsimonious with only 8 trainable parameters, but it performs poorly against high-resolution data, with MSE values as large as $626\%^2$ despite having been re-tuned to our training set.

Hypothesizing that the Sundqvist scheme's large error is due to its lack of features, we test the effect of adding features one-by-one. For that purpose, we apply forward sequential feature selection, which greedily adds features using a cross-validated score (here MSE), to a standard multiple linear regression model that includes polynomial combinations of all available features, up to a maximum degree of 3.

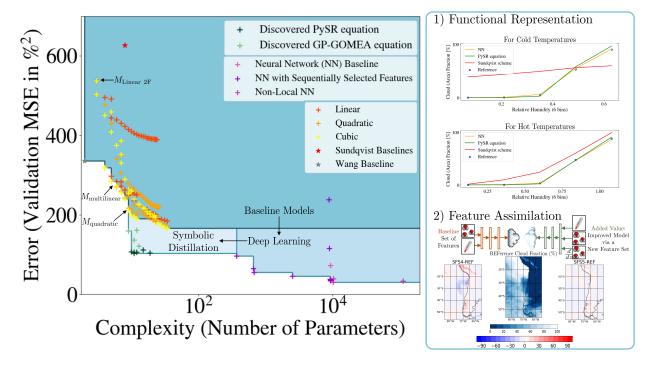We find that linearly including temperature as a feature alongside relative humidity is enough to outperform

FIG. 2. Pareto-optimal model hierarchies quantify the added value of machine learning for cloud cover parameterization. Machine learning better captures the relationship between cloud cover and its thermodynamic environment and assimilates features like vertical humidity gradients. (Left) We progressively improve traditional baselines via polynomial regression (red, orange, and yellow crosses), significantly decrease error using neural networks (pink and purple crosses), and finally distill the added value of these neural networks symbolically (green crosses). (Right) Both the neural network (orange line) and its distilled symbolic representation (green line) better represent the functional relationship between cloud cover and its environment, aligning more closely across temperatures with the reference storm-resolving simulation (blue dots) than the Sundqvist scheme (red line) used in the ICON Earth system model. "Cold" and "Hot" refer to the validation set's first and last temperature octiles. Additionally, machine learning models assimilate multiple features absent in existing baselines, including vertical humidity gradients. The smaller discrepancy between the 5-feature scheme ('SFS5') and the reference ('REF'), compared to the 4-feature scheme ('SFS4'), demonstrates improved representation of the time-averaged low cloud cover in regions such as the Southeast Pacific, thereby reducing biases in current cloud cover schemes that plague the global radiative budget.

the Sundqvist baseline, providing our simplest example of *feature assimilation*:

$$M_{\text{Sundqvist}} [\text{RH}] \notin \text{PF}_{\mathcal{E}} \ , \ M_{\text{Linear 2F}} [\text{RH}, T] \in \text{PF}_{\mathcal{E}}, \quad (10)$$

where $M_{\text{Linear 2F}}$ is a two-feature linear regression whose output $C_{\text{Linear 2F}}$ is:

$$\frac{C_{\text{Linear 2F}}}{\alpha_{2F}} = 1 + \frac{\text{RH}}{\text{RH}_{2F}} - \frac{T}{\Delta T_{2F}} \quad (11)$$

with parameters α_{2F}, RH_{2F}, and ΔT_{2F} listed in appendix A. This feature assimilation result corresponds to a well known result: accurate cloud cover parameterization requires temperature in addition to relative humidity.

Adding more features to linear (red crosses), quadratic (orange crosses), and cubic (yellow crosses) polynomials further reduces error. However, incorporating a simple physical constraint:

$$q_\ell + q_i = 0 \implies C = 0, \quad (12)$$

i.e., that the absence of condensates implies no clouds, further improves the results. Note that this constraint can also be derived from data using a binary decision tree. Accounting for this constraint, polynomial models achieve MSEs less than half of the Sundqvist scheme's (see Fig 2's lower error, higher complexity polynomial models). Focusing on the Pareto-optimal, multilinear and quadratic model outputs:

$$\frac{C_{\text{multilinear}}}{\alpha_1} = 1 + \frac{\text{RH}}{\text{RH}_1} - \frac{T}{\Delta T_1} + \frac{q_i}{q_1} - H_1 \frac{d\text{RH}}{dz}, \quad (13)$$

$$\frac{C_{\text{quadratic}}}{\alpha_2} = 1 + \frac{\text{RH}}{\text{RH}_{2,1}} - \frac{T}{\Delta T_{2,1}} + \frac{q_i}{q_{2,1}} + \frac{q_\ell}{q_{2,2}} - \frac{q_i q_\ell}{q_{2,3}^2}$$
$$+ H_2 \left(1 + \frac{\text{RH}}{\text{RH}_{2,2}} - \frac{T}{\Delta T_{2,2}} \right) \frac{d\text{RH}}{dz}, \quad (14)$$

with parameters listed in appendix A. Analyzing this first Pareto front unveils the role of each added feature in better representing cloud cover. First, it highlights the role of cloud condensates, both liquid and ice, in accurately de-

scribing cloud cover. Second, among all possible vertical gradients, Pareto-optimal models select the negative gradient of relative humidity to depict the role of inversions in increasing cloud cover, which is particularly important for low clouds, including marine stratocumulus clouds. However, upon closer inspection, these contributions are not always physically consistent: while the multilinear model correctly captures the increase of cloud cover with ice, it incorrectly decreases cloud cover when there is humidity above. Conversely, the quadratic model accurately captures the inversion role when temperature and humidity are fixed to their training set mean but wrongly assumes that inversions' effect on cloudiness depends on humidity and temperature, and the model's decrease in cloudiness with increasing mixed-phase cloud condensates suggests that its $q_i q_\ell$ term is not an interpretable sensitivity to liquid and ice but rather a bias correction term relying on these variables.

Combining these insights with the Pareto-optimality of the simplified "Wang" scheme (gray star in Fig. 2; Wang et al. 2023, based on the scheme of Xu and Randall 1996), which outputs:

$$C_{\text{Wang}} = \min\left\{1, \left[1 - e^{-\alpha_{\text{Wang}}(q_\ell + q_i)}\right] \times \text{RH}^{\beta_{\text{Wang}}}\right\}, \quad (15)$$

with $(\alpha_{\text{Wang}}, \beta_{\text{Wang}})$ in appendix A, suggests that the dependence of cloud cover on condensates is fundamentally non-linear. This justifies the use of deep learning to quickly explore which features can be combined to improve the non-linear, *functional representation* of cloud cover.

4) Deep Learning

We start with the baseline model from Grundner et al. (2022), which uses a 3-layer-deep, 64-neuron-wide multi-layer perceptron with batch normalization after the second hidden layer. Hyperparameters were optimized using the SHERPA Python library (Hertel et al. 2020). This NN estimates the target cloud cover with high fidelity (MSE=73%^{2}), but at the cost of increased complexity, as it has a total of 9345 trainable parameters. Note that the models from Grundner et al. (2022) were not designed to minimize the number of trainable parameters, so we do not overly focus on this complexity metric. The MSE can be further lowered to 33%^{2} with vertically non-local NNs, which map the entire atmospheric column of inputs to the entire column of outputs without inductive bias. However, this small error gain is deemed insufficient given the increased complexity cost.

Instead, we make the "vertically quasi-local" assumption (Eq. 6) and deploy a hierarchy of Pareto-optimal NNs, with features selected sequentially using cross-validated MSE. The five most informative features for NNs are:

$$\text{RH} \rightarrow q_i \rightarrow q_\ell \rightarrow T \rightarrow \frac{d\text{RH}}{dz}. \quad (16)$$

Unlike features selected by polynomial models, these can be non-linearly combined to yield high-quality predictions, as shown in the right panels of Fig. 2. First, NNs improve the functional representation of cloud cover by accurately modeling the sharp increase in cloud cover above a temperature-dependent relative humidity threshold – a highly nonlinear, bivariate behavior that simple schemes struggle to capture. Formally:

$$M_{\text{Sundqvist}}[\boldsymbol{X}] \notin \text{PF}_{\mathcal{E}} \,, \, M_{\text{NN}}[\boldsymbol{X}] \in \text{PF}_{\mathcal{E}}, \quad (17)$$

where M_{NN} is a Pareto-optimal NN at the bottom-right of the (complexity, error) plane. Second, by incorporating vertical relative humidity gradients, NNs can capture stratocumulus decks off the coasts of regions like California, Peru/Chile, and Namibia/Angola. This improvement is especially visible when comparing the error map of an NN using the first four features from Equation 16 to that of an NN additionally incorporating $d\text{RH}/dz$ (Fig 2, bottom-right panel).

While indicative of how accurately cloud cover can be parameterized, such improvements are often insufficient to be considered "discoveries" as they remain hard to explain, even with post-hoc explanation tools (Fig. 8 & 9 of Grundner et al. 2022). Therefore, improvements in functional representation and feature assimilation need to be further distilled into sparse models that scientists can readily interpret.

5) Symbolic Distillation and Equation Discovery

For this purpose, we use symbolic regression libraries, which optimize both the parameters and structure of an analytical model within a space of expressions. Symbolic regression yields expressions with transparent out-of-distribution behavior (asymptotics, periodicity, etc.) (La Cava et al. 2021), making them well-suited for high-stakes societal applications (Rudin 2019) and the empirical distillation of natural laws (Schmidt and Lipson 2009). To avoid overly restricting the analytical form of the distilled equation, genetic programming is used. Genetic programming evolves a population of mathematical expressions using methods such as selection, crossover, and mutation to improve a fitness function (Koza 1994). Given that genetic programming scales poorly with the number of features (Petersen et al. 2019), our NN feature selection results are used to restrict our features to those listed in Equation 16. Using NN results is appropriate since no assumption is made about the type of equation to be discovered.

The GP-GOMEA (light green crosses in Fig. 2; Virgolin et al. 2021) and PySR (dark green crosses in Fig. 2; Cranmer 2023) libraries were chosen for their ease of use and high relative performance compared to 12 other recent libraries (La Cava et al. 2021). They yielded over 500 closed-form equations for cloud cover, from which the 9 most physically consistent and lowest-error fits were

retained with their outputs always clipped to $[0, 1]$ (see Grundner et al. 2024 for details). By physically interpreting the learned parameters, the output of the Pareto-optimal PySR model may be written as:

$$C_{\text{PySR}} = \underbrace{\mathcal{I}_1\left(\text{RH}, T\right)}_{\text{Humidity/Temperature}} + \underbrace{\mathcal{I}_2\left(\frac{d\text{RH}}{dz}\right)}_{\text{Inversion}} + \underbrace{\mathcal{I}_3\left(q_\ell, q_i\right)}_{\text{Condensates}}, \quad (18)$$

where the first term $\mathcal{I}_1$ may be interpreted as a sparse, third-order Taylor expansion around the training-mean relative humidity ($\overline{\text{RH}} = 0.60$) and temperature ($\overline{T} = 257K$):

$$\begin{aligned} \mathcal{I}_1 &= \overline{C} + \left(\frac{\partial C}{\partial \text{RH}}\right)_{\overline{\text{RH}}, \overline{T}} \left(\text{RH} - \overline{\text{RH}}\right) - \left(\frac{\partial C}{\partial T}\right)_{\overline{\text{RH}}, \overline{T}} \left(T - \overline{T}\right) \\ &\quad + \frac{1}{2}\left(\frac{\partial^2 C}{\partial \text{RH}^2}\right)_{\overline{\text{RH}}, \overline{T}} \left(\text{RH} - \overline{\text{RH}}\right)^2 \\ &\quad + \frac{1}{2}\left(\frac{\partial C}{\partial \text{RH} \partial T^2}\right)_{\overline{\text{RH}}, \overline{T}} \left(T - \overline{T}\right)^2 \left(\text{RH} - \overline{\text{RH}}\right). \end{aligned}$$
$$(19)$$

Dominant Taylor series expansion terms are expected when discovering closed-form, subgrid-scale parameterizations (Jakhar et al. 2024), but as PySR is based on genetic programming, we find two more surprising terms:

$$\mathcal{I}_2 = H_{\text{PySR}}^3 \left[\frac{d\text{RH}}{dz} + \frac{3}{2}\left(\frac{d\text{RH}}{dz}\right)_{\max C}\right] \left(\frac{d\text{RH}}{dz}\right)^2, \quad (20)$$

where $H_{\text{PySR}} \approx 585m$ can be interpreted as a characteristic height for low-cloud humidity gradients, and $(d\text{RH}/dz)_{\max C}$ ($\approx -2/\text{km}$) is the value of the relative humidity gradient that maximizes cloud cover at the inversion level. The last term is:

$$\mathcal{I}_3 = -\frac{1}{\epsilon_{\text{PySR}}} \times \frac{1}{1 + 2\epsilon_{\text{PySR}}\left(\lambda_\ell q_\ell + \lambda_i q_i\right)}, \quad (21)$$

which is a monotonically increasing function of the condensates' concentrations, whose trainable parameters are provided in appendix A. Consistently, the Pareto-optimal GP-GOMEA equation has a term Q that sharply increases as liquid or ice concentrations exceed 0:

$$\frac{C_{\text{GOMEA}}}{\alpha_G} = 1 + \underbrace{\beta_G e^{\text{RH}/\text{RH}_G}}_{\text{Humidity}} + \underbrace{Q\left(q_\ell, q_i\right)}_{\text{Condensates}}, \quad (22)$$

$$Q = \gamma_G \ln\left[\epsilon_G + \frac{q_i}{q_{G,i}} + \delta_G \left(e^{\frac{q_\ell}{q_{G,\ell+}}} - 1\right)\right] - \frac{q_\ell}{q_{G,\ell-}}, \quad (23)$$

where the trainable parameters are listed in appendix A.

In addition to being easily transferable thanks to their low number of trainable parameters, the added value of these equations is transparent: the improved functional representation is explicit (see Eq. 19), and the assimilation of new features is interpretable (see Eq. 20). Finally,

scientific discovery may arise through the unexpected aspects of these equations that are robust across models, such as the difference between how cloud cover reacts to an increase in environmental liquid versus ice content. Indeed, at high resolution, cloud cover will become 1 as soon as condensates exceed a small threshold (here 10^{-6} kg/kg) independently of the water's phase. Then, how come cloud cover is more sensitive to ice than liquid in Eq. 21 ($\lambda_i \approx 3.8\lambda_\ell$) and Q increases much faster with ice than liquid (for $q_i + q_\ell \ll 1$) in Eq. 23?

These are in fact emerging properties of the subgrid distributions of liquid and ice (Grundner et al. 2024): As large values of cloud ice are rarely observed, larger spatial averages of cloud ice at coarse resolution means that many more high-resolution pixels contain low values of cloud ice compared to the liquid case, resulting in higher cloudiness for a given spatially-averaged condensate value. By assuming an exponential distribution for the subgrid liquid and ice content, we can even interpret λ_ℓ and λ_i as the rate of the respective exponential distributions. This allows us to hypothesize that the nonlinear relationship between condensates and cloud cover is not scale-invariant and requires separate treatments of liquid and ice, with implications for the interaction between microphysical processes and the radiative budget. While analyzing the feature importance of liquid and ice in neural networks could have suggested this difference, it would have been difficult to fully explain it and bridge spatial scales without distillation, confirming the importance of only treating deep learning as a first and not final step towards knowledge discovery. We now turn to a higher-dimensional problem in which spatial connectivity has to be considered: radiative transfer.

b. Shortwave Radiative Transfer Emulation to Accelerate Numerical Weather Prediction

1) Motivation

The energy transfer resulting from the interaction between electromagnetic radiation and the atmosphere, known as radiative transfer, is costly to simulate accurately. Line-by-line calculations of gaseous absorption at each electromagnetic wavelength (Clough et al. 1992) are too expensive for routine weather and climate models. Instead, models often use the correlated-k method (Mlawer et al. 1997), which groups absorption coefficients in a cumulative probability space to speed up radiative transfer calculations without significantly compromising accuracy. However, even the correlated-k method imposes a high computational burden (Veerman et al. 2021), forcing most simulations to reduce the temporal and spatial resolution of radiative transfer calculations, which can degrade prediction quality (Morcrette 2000; Hogan and Bozzo 2018).

This challenge has driven the development of ML emulators for radiative transfer in numerical weather prediction since the 1990s (Cheruy et al. 1996; Chevallier et al. 1998,

2000). ML architectures have become more sophisticated (e.g., Belochitski and Krasnopolsky 2021; Kim and Song 2022; Ukkonen 2022), but the primary goal remains to emulate the original radiation scheme as faithfully as possible. This allows the reduced inference cost of the ML model, once trained, to be leveraged for running the atmospheric model coupled with the emulator, enabling less expensive and more frequent radiative transfer calculations.

In this section, we examine how ML architectural designs impact the reliability of shortwave radiative transfer (covering solar radiation and wavelengths of 0.23-5.85 μm). As shown in Fig. 3, architectures that closely mimic the vertical bidirectionality of radiative transfer are Pareto-optimal, rivaling the performance of deep learning models with ten times more trainable parameters.

2) Setup

We follow the setup described in Lagerquist et al. (2022) and emulate the full behavior (gas optics and radiative transfer solver) of the shortwave Rapid Radiative Transfer Model (Mlawer et al. 1997) in the context of weather predictions made by version 16 of the Global Forecast System with 0.25° horizontal spacing. In addition to the realistic geography setup of Lagerquist et al. (2021), the ML models are trained with global data on the 127-level native pressure-sigma grid (henceforth referred to as η) with synthetic information about aerosols, trace gases, and hydrometeors' particle size distribution. We use data from most days between Sep 1 2018 and Dec 23 2020, with forecast lead times in $\{6, 12, 18, 24, 30, 36\}$ hours. In each forecast, 4000 grid points are randomly sampled on the global grid. The training set comprises 873,086 samples from 237 days, the validation set 479,806 samples from 126 days, and the test set 472,456 samples from 120 days. Features X can broadly be separated into 22 vertical profiles tied to the grid (X_η; listed in Appendix B) and four scalars with one value per atmospheric column: solar zenith angle ζ [°], surface albedo α_s, aerosol single-scattering albedo ω_0, and aerosol asymmetry parameter g. For convenience, the four scalars are converted into profiles by repeating their values at all 127 levels. We target the shortwave heating rate's vertical profile $(dT/dt)_{\text{SW}}$ [K day^{-1}], thus approximating a high-dimensional mapping:

$$\underbrace{(X_\eta, \zeta, \alpha_s, \omega_0, g, \sigma_e)}_{\in \mathbb{R}^{(22+4) \times 127}} \mapsto \underbrace{\left(\frac{dT}{dt}\right)_{\text{SW}}}_{\in \mathbb{R}^{1 \times 127}} \quad (24)$$

via an ML model hierarchy described in the next section.

3) Machine Learning Model Hierarchy

We deploy an ML model hierarchy using the same features X to isolate the added value of spatial connectivity

along the vertical coordinate η. For all models, cropping and zero-padding layers set the top-of-atmosphere (TOA; ~78 km above ground level) heating rate to zero, consistent with the RRTM model being emulated. Readers interested in the exact ML model architecture are referred to Lagerquist et al. (2022) and the accompanying code (https://zenodo.org/records/12557544).

We begin with a linear regression model (pink star), where the input layer is reshaped into a flattened vector and processed through linear layers without activation functions or pooling. This model lacks inherent spatial connectivity, treating all features (where one "feature" = one atmospheric variable at one vertical level) independently. Similarly, the multilayer perceptrons (MLP; orange crosses) flatten inputs, ignoring vertical connectivity, but process them with varying degrees of nonlinearity, controlled by two hyperparameters: depth (number of dense layers) and width (number of neurons per layer). We conduct a grid search with depth varying from $\{1, 2, 3, 4, 5, 6\}$ layers and width varying from $\{64, 128, 256, 512, 1024, 2048\}$ neurons, resulting in 36 MLPs. The activation function for every hidden layer (i.e., every layer except the output) is the leaky rectified linear unit (ReLU; Maas et al. 2013) with a negative slope $\alpha = 0.2$.

To incorporate vertical relationships between different levels, we train convolutional neural networks (CNN; red crosses) with 1D convolutions along the vertical axis. We conduct a grid search with depth varying from $\{1, 2, 3, 4, 5, 6\}$ convolutional layers and number of channels in the first convolutional layer varying from $\{2, 4, 8, 16, 32, 64\}$. After the first layer, we double the number of channels in each successive convolutional layer, up to a maximum of 64. The CNN's kernel size is restricted to 3 vertical levels, to enforce strong local connectivity.

The U-net architecture (pink crosses; Ronneberger et al. 2015) builds on the CNN by incorporating skip connections that preserve high-resolution spatial information and an expansive path almost symmetric to the contracting path. Both of these model components improve the reconstruction of full-resolution data (here, a 127-level profile of heating rates). We conduct a grid search with depth (number of downsampling operations) varying from $\{3, 4, 5\}$ and number of channels in the first convolutional layer varying from $\{2, 4, 8, 16, 32, 64\}$, using the same channel-doubling rule as for CNNs. Finally, the U-net++ architecture (purple crosses; Zhou et al. 2019) enhances the U-net's skip pathways with nested dense convolutional blocks, potentially capturing more of the original spatial information and facilitating optimization. Our hyperparameters for U-net++ are the same as for the U-net.

4) Distilling Radiative Transfer's Bidirectionality

To better understand the added value of each architecture in representing shortwave absorption and scattering,
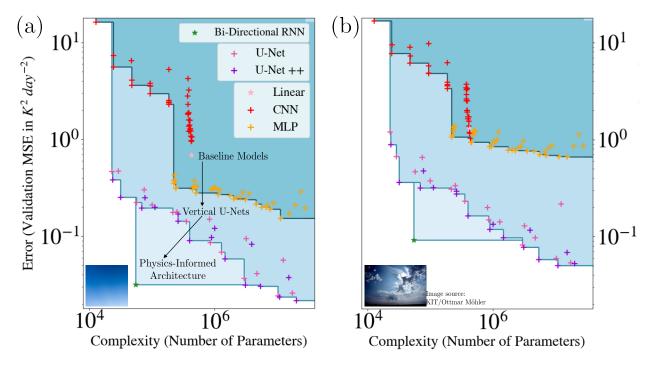
FIG. 3. Pareto-optimal model hierarchies guide the development of progressively tailored architectures for emulating shortwave radiative transfer. Panel (a) shows error vs. complexity on a logarithmic scale for the simple clear-sky cases dominated by absorption; panel (b) shows error vs. complexity for cases with multi-layer cloud, including both liquid and ice, where multiple scattering complicates radiative transfer. Convolutional neural networks (CNN; red crosses) with small kernels, multilayer perceptrons (MLP; orange crosses) that ignore the vertical dimension, and the simple linear baseline (light pink star) give credible results in the clear-sky case. However, they fail in the more complex case, which requires U-net architectures (dark pink and purple crosses) to fully capture non-local radiative transfer. The vertical invariance of the two-stream radiative transfer equations suggests a bidirectional recurrent neural network (RNN; green star) architecture, which rivals the skill of U-nets with a fraction of their trainable parameters.

we extract samples from our test set to form two distinct regimes. The simple, "clear-sky" regime (Fig. 3a) consists of profiles with no cloud (column-integrated liquid-water path = column-integrated ice-water path = 0 g m^{-2}), an oblique Sun angle (zenith angle > 60°), and little water vapor (column-maximum specific humidity < 2 g kg^{-1}). This restricts shortwave radiative transfer to gaseous absorption of solar radiation throughout the atmospheric column, well described by a simple exponential attenuation model such as Beer's law (Liou 2002). In contrast, the complex, "multi-cloud" regime (Fig. 3b) includes profiles with more than one cloud layer and at least one mixed-phase cloud layer. For this purpose, a "cloud layer" is defined as a set of consecutive vertical levels, such that every level has a total water content (liquid + ice) > 0 g m^{-3} and the layer has a height-integrated total water path (liquid + ice) ≥ 10 g m^{-2}. A mixed-phase cloud layer meets the above criteria plus two additional ones: both the height-integrated liquid water path and ice water path must be > 0 g m^{-2}. This regime is challenging to model as shortwave radiation is absorbed and scattered by both liquid and ice clouds, making the bidirectionality of radiative transfer essential to

capture. Out of the 472,456 test set profiles, 14,226 are in the clear-sky regime and 13,263 in the multi-cloud regime.

The first surprising result is the poor performance of CNNs: While Pareto-optimal for low complexity, with our simplest CNN having as few as 12,630 trainable parameters, none of the CNNs achieve an MSE below 0.95 K^2 day^{-2}, even in the clear-sky case. In contrast, MLPs systematically outperform our linear baseline model, showcasing the importance of nonlinearity. MLPs also outperform CNNs because they allow for non-local vertical connectivity. However, without information on which vertical levels are closest, MLPs connect every level together, resulting in high complexity: Increasing the number of trainable parameters by 100× does not even halve the MSE in the clear-sky case. This lack of inductive bias prevents generalization to more complex cases such as the multi-cloud regime, where the MSE climbs from ≈0.2-0.4 K^2 day^{-2} to ≈0.7-1.0 K^2 day^{-2}.

By ordering the features' vertical levels, both the U-net and U-net++ achieve lower errors with fewer trainable parameters. Most Pareto-optimal models are in the U-net++ family, suggesting that simple skip connections are insufficient; this highlights the non-locality of shortwave

radiative transfer, especially in the multi-cloud case. Fortunately, while shortwave radiation may instantaneously propagate information (e.g., the presence of mixed-phase clouds) throughout the entire atmospheric column, its transfer is approximately governed by bidirectional equations invariant in space. The two-stream equation holds exactly for the target Rapid Radiative Transfer Model data:

$$\frac{d}{d\eta}\begin{pmatrix}\mathcal{F}_+\\\mathcal{F}_-\end{pmatrix} = \mathcal{R}\left(\mathcal{F}_+, \mathcal{F}_-, X_\eta\right), \qquad (25)$$

where $\mathcal{F}_+$ and $\mathcal{F}_-$ are the upward and downward radiative fluxes (W m^{-2}) and the function $\mathcal{R}$ governs how these fluxes change with the vertical coordinate η, allowing us to update fluxes using neighboring fluxes only. As noted in Ukkonen and Chantry (2024), it is the vertical invariance of Eq. 25 that suggests we may treat radiative transfer with a network processing the vertical sequences of $\mathcal{F}_+$ and $\mathcal{F}_-$ in both directions with the same update rule at every level. Akin to standard atmospheric models, the target heating rate is then calculated by multiplying the vertical gradient of the net flux (downward minus upward) by the ratio of the gravity constant at the Earth's surface g to the specific heat capacity of dry air at constant pressure c_p:

$$\left(\frac{dT}{dt}\right)_{\text{SW}} = \frac{g}{c_p}\frac{d\eta}{dp}\frac{d\left(\mathcal{F}_- - \mathcal{F}_+\right)}{d\eta}. \qquad (26)$$

Following Ukkonen (2022), we divide the shortwave radiative fluxes ($\mathcal{F}_+, \mathcal{F}_-$) by the incoming TOA downward flux ($\mathcal{F}_-^{\text{TOA}}$), to implicitly inform the NN that the Sun is the original energy source. At each vertical level η, we then target the resulting ratios ($\mathcal{F}_+/\mathcal{F}_-^{\text{TOA}}$ and $\mathcal{F}_-/\mathcal{F}_-^{\text{TOA}}$) with a bidirectional RNN. The RNN has two 64-unit gated recurrent unit layers (one forward, one backward), which are concatenated before a dense layer for predicting the fluxes, resulting in only 54,786 trainable parameters. The RNN includes a multiplication layer with $\mathcal{F}_-^{\text{TOA}}$ and a custom layer to directly calculate shortwave heating rates, maintaining the same optimization objective as the other models. We find that the resulting RNN (green star) is clearly Pareto-optimal, competing with the U-net++ and achieving MSE below 0.1 K^2 day^{-2}, even in the complex, multi-cloud case. Our results highlight the advantages of physically constraining ML solutions using robust knowledge of the underlying system, extending the aquaplanet findings of Bertoli et al. (2023) to an Earth-like, operational setting, and paving the way towards fully hybrid physics-ML emulators of shortwave radiation (Schneiderman 2024). We now turn to a problem in which both spatial and temporal connectivities may be considered: precipitation parameterization.

c. Tropical Precipitation and Convective Organization

1) Motivation

Accurately representing precipitation processes in tropical regions enhances global Earth system models and forecasting tools, particularly for water management and flood risk in a changing climate (Seneviratne et al. 2021). Due to computational limitations, achieving horizontal resolutions below 25-50 km is challenging, which hinders the representation of subgrid processes causing precipitation (Stevens et al. 2020). These processes include tropical convection and its complex organization patterns (Bao et al. 2024), typically modeled using semi-empirical parameterizations that rely on a coarse representation of the thermodynamic environment and often exclude memory effects (Colin et al. 2019). These parameterizations generally approximate a mapping: $\langle \boldsymbol{X}_{t_0} \rangle \mapsto \langle P_{t_0} \rangle$, where at the time of interest t_0, both the spatially resolved environmental predictors $\boldsymbol{X}$ and the precipitation are averaged over the coarse grid cell:

$$\langle \boldsymbol{X}_{t_0} \rangle = \frac{1}{|\text{Grid Cell}|} \int_{\text{Grid Cell}} \boldsymbol{X}_{t_0, \boldsymbol{x}} d\boldsymbol{x}, \qquad (27)$$

where $\boldsymbol{x}$ represents a continuous, bidimensional horizontal coordinate. This spatial averaging removes convective organization information below the coarse grid's spatial scale. Using storm-resolving simulations as a reference, Shamekh et al. (2023) showed this blurs precipitation, adding irreducible uncertainty, especially for large values of precipitable water (PW). Given that high-resolution information is typically inaccessible in a parameterization context, we ask in this section: How much of this lost spatial granularity can we recover with temporal memory?

2) Setup

To address this question, we build upon the framework of Shamekh et al. (2023), coarse-graining SAM-DYAMOND (Khairoutdinov et al. 2022) simulations from their native, high-resolution horizontal grid (4.34-km spacing at the equator) to a low-resolution grid (138.9-km spacing at the equator) representative of a coarse Earth system model. Over the tropical ocean (20°S-20°N), we map 6 environmental variables $\boldsymbol{X}$ – PW, 2-m specific humidity [kg kg^{-1}], 2-m temperature [K], surface sensible and latent heat fluxes [W/m^2], and sea surface temperature [K] – to 15-min precipitation rates $\langle P \rangle$ [mm h^{-1}]. We use three setups at a given low-resolution location x_{LR} including the high-resolution grid locations $x_{\textbf{HR}}$:

$$\text{Baseline:} \quad \underbrace{\langle \boldsymbol{X}_{t_0} \rangle_{x_{\text{LR}}}}_{\in \mathbb{R}^6} \mapsto \underbrace{\langle P_{t_0} \rangle_{x_{\text{LR}}}}_{\in \mathbb{R}_+}, \qquad (28)$$

Spatial Granularity: $\underbrace{\begin{pmatrix} \langle \boldsymbol{X}_{t_0} \rangle_{x_{\mathrm{LR}}} \\ \mathrm{PW}'_{t_0, \boldsymbol{x}_{\mathrm{HR}}} \end{pmatrix}}_{\in \mathbb{R}^{6+32\times32}} \mapsto \underbrace{\langle P_{t_0} \rangle_{x_{\mathrm{LR}}}}_{\in \mathbb{R}_+}, \quad (29)$

Temporal Memory: $\underbrace{\langle \boldsymbol{X}_{t_{\mathrm{Memory}}} \rangle_{x_{\mathrm{LR}}}}_{\in \mathbb{R}^{6 \times \mathrm{card}(t_{\mathrm{Memory}})}} \mapsto \underbrace{\langle P_{t_0} \rangle_{x_{\mathrm{LR}}}}_{\in \mathbb{R}_+}, \quad (30)$

where PW$'$ is PW's anomaly with respect to its grid cell-mean $\langle$PW$\rangle$ and memory is accounted for through up to two time steps in the past, in which case $t_{\mathrm{Memory}} = \{t_0, t_0 - 15\mathrm{min}, t_0 - 30\mathrm{min}\}$. Samples are extracted from 10 days of the simulation after spin-up, with 6 days for training, 2 for validation, and 2 for testing. To focus on precipitation intensity rather than triggering, samples with precipitation values below 0.05 mm/h are discarded, resulting in a total of $\approx 10^8$ samples.

3) Comparing Spatial and Temporal Connectivities

As illustrated in Fig. 4, we design three categories of NNs to learn the three mappings given by Eq. 28, 29, and 30. First, our baseline NNs (orange crosses) are simple MLPs with 5 layers of different widths, resulting in a number of trainable parameters ranging from $\approx 30 - 600 \times 10^3$ and never achieving MSEs below 0.4 mm^2 h^{-2}. The NNs using spatial granularity (purple crosses) additionally encode the high-resolution PW anomaly field through an encoder-decoder structure with a small bottleneck representing convective organization as a bidimensional latent variable, optionally regularized via data augmentation applied to PW$'$. MSEs are below 0.04 mm^2 h^{-2} for the deepest encoder-decoder architectures, which yield high reconstruction quality through more than 1M trainable parameters. The NNs leveraging temporal memory instead of spatial resolution (green crosses) use the last or last two previous timesteps to achieve competitive MSEs, all below 0.15 mm^2 h^{-2}. These "Memory-NNs" are built using two types of architectures: On the one hand, we flatten previous timesteps and feed them to MLPs conditioned on non-zero current precipitation, which ignores temporal connectivity. On the other hand, we use CNN-based temporal models that preserve temporal ordering through 1D convolutional layers, achieving MSEs around 0.1 mm^2 h^{-2} with ≈ 10 times fewer trainable parameters than the "High-Res Inputs-NNs" that use storm-scale information. Overall, our results showcase the added value of *spatial and temporal connectivities*, as $M_{\mathrm{MLP}} \left[\langle \boldsymbol{X}_{t_0} \rangle_{x_{\mathrm{LR}}} \right] \notin \mathrm{PF}_{\mathcal{E}}$ while:

$$M_{\mathrm{CNN}} \left[\langle \boldsymbol{X}_{t_{\mathrm{Memory}}} \rangle_{x_{\mathrm{LR}}} \right], \; M_{\mathrm{ED}} \left[\boldsymbol{X}_{t_0, \boldsymbol{x}_{\mathrm{HR}}} \right] \in \mathrm{PF}_{\mathcal{E}}^2. \quad (31)$$

4) Mitigating Low Spatial Resolution with Memory

While it may be unsurprising that leveraging spatio-temporal information decreases error, the competitive MSEs obtained with temporal memory but without high-resolution spatial information are promising for improving parameterizations. We ask: How can we explain that temporal memory helps recover a large portion of the spatial granularity information?

First, we can ask: Why would temporal memory of the coarse environment $\langle \boldsymbol{X}_{t_{\mathrm{Memory}}} \rangle_{x_{\mathrm{LR}}}$ be relevant on its own, given that at a given time t_0, the precipitation $\langle P_{t_0} \rangle$ can be exactly diagnosed from the high-resolution environment $\boldsymbol{X}_{t_0, \boldsymbol{x}_{\mathrm{HR}}}$? The informativeness of high spatial resolution is confirmed by the fact that our lowest-error model is M_{ED}. As written in Eq. 29, M_{ED} decomposes the high-resolution information into a low-resolution component and the inputs' spatial anomaly $\boldsymbol{X}'_{t_0, \boldsymbol{x}_{\mathrm{HR}}}$. Using this decomposition, we can write the following causal graph:

$$\begin{aligned} \boldsymbol{X}'_{t_0, \boldsymbol{x}_{\mathrm{HR}}} &\to P_{t_0} \leftarrow \langle \boldsymbol{X}_{t_0} \rangle_{x_{\mathrm{LR}}}, \\ \boldsymbol{X}'_{t_0, \boldsymbol{x}_{\mathrm{HR}}} &\to \boldsymbol{X}'_{t_0+\Delta t, \boldsymbol{x}_{\mathrm{HR}}} \leftarrow \langle \boldsymbol{X}_{t_0} \rangle_{x_{\mathrm{LR}}}, \\ \boldsymbol{X}'_{t_0, \boldsymbol{x}_{\mathrm{HR}}} &\to \langle \boldsymbol{X}_{t_0+\Delta t} \rangle_{x_{\mathrm{LR}}} \leftarrow \langle \boldsymbol{X}_{t_0} \rangle_{x_{\mathrm{LR}}}, \end{aligned} \quad (32)$$

where we posit that for a small enough time step Δt, the combination of the high-resolution anomaly and low-resolution information is enough to diagnose precipitation and step the entire system forward in time from t_0 to $t_0 + \Delta t$. Formally, this is a Markovian assumption and eliminates the need for temporal memory. Using Eq. 32 at times $t_0 - \Delta t$ and t_0, we see that in this framework, suppressing access to high-resolution information "unblocks" the following causal path:

$$\langle \boldsymbol{X}_{t_0 - \Delta t} \rangle_{x_{\mathrm{LR}}} \to \boldsymbol{X}'_{t_0, \boldsymbol{x}_{\mathrm{HR}}} \to P_{t_0}, \quad (33)$$

creating a conditional dependence of precipitation P_{t_0} on the past low-resolution environment $\langle \boldsymbol{X}_{t_0 - \Delta t} \rangle_{x_{\mathrm{LR}}}$ by removing their d-separation given current high-resolution anomalies $\boldsymbol{X}'_{t_0, \boldsymbol{x}_{\mathrm{HR}}}$ (Pearl et al. 2000). This explains the need for temporal memory without high-resolution information and provides a rationale for the Pareto optimality of the "simple-memory NN" M_{CNN} that considers the low-resolution inputs and their past timesteps.

Now, we can ask: How much information is lost in the process of replacing the high-resolution spatial anomaly with temporal memory of the low-resolution spatial means? The answer depends on the distance between low and high resolutions. There is no difference if these resolutions are equal, while we do not expect low-resolution memory to help at all if the low-resolution grid is coarse to the point that we cannot distinguish very different high-resolution situations. For the resolutions considered here, the bidimensionality of the encoder-decoder's bottleneck suggests that the high-resolution anomaly field can be
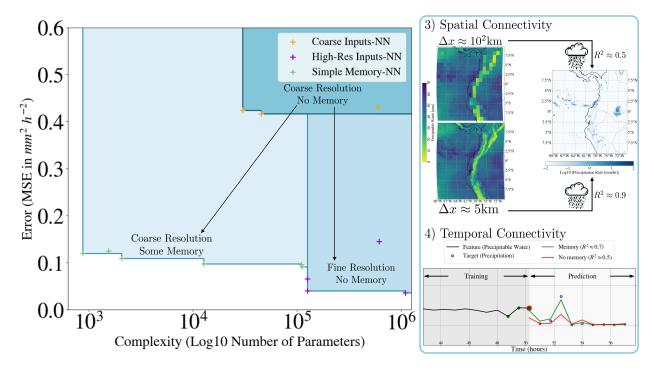
FIG. 4. Pareto-optimal model hierarchies underscore the importance of storm-resolving information in elucidating the relationship between precipitation and its surrounding environment, while also quantifying the recoverability of this information from the coarse environment's time series. (Left) Neural networks (NN) leveraging high-resolution spatial data (purple crosses) clearly outperform NNs that use only coarse inputs (orange crosses). However, this performance gap is largely mitigated when the coarse inputs' past time steps are included (green crosses). (Right) Processing the precipitable water (PW) field at a resolution of $\Delta x \approx 5$ km yields coefficients of determination $R^2 \approx 0.9$, clearly surpassing the $R^2 \approx 0.5$ attained by our best NN using PW fields at the coarse $\Delta x \approx 10^2$ km horizontal resolution. This performance gap is partially closed by incorporating two past time steps along with the current timestep, resulting in $R^2 \approx 0.7$. This suggests a partial equivalence of the environment's spatial and temporal connectivities in governing precipitation.

represented by a bidimensional variable for the purpose of modeling precipitation: **org**, which represents the aspects of convective organization that explain why precipitation can be different (stochastic) for the same low-resolution inputs (Moseley et al. 2016). This simplifies the question to how well we can represent **org** from the time series of low-resolution inputs. The encouraging results of our models leveraging temporal memory, along with the finding of Shamekh et al. (2023) that **org** is accurately predicted by an autoregressive model informed by the coarse-scale environment's past timesteps, suggest a mostly positive response. This reassuringly confirms that Earth system models with limited spatial resolutions can realistically represent coarse-scale precipitation as long as efforts to improve precipitation parameterization continue (Schneider et al. 2024). More broadly, this confirms that Pareto-optimal model hierarchies are useful in empirically establishing the partial equivalence between temporal memory and spatial granularity. This has practical applications for multi-scale dynamical systems that are not self-similar, where ergodicity cannot be used to deterministically infer detailed spatial information from coarse time series data.

5. Conclusion

In this study, we demonstrated that Pareto-optimal model hierarchies within a well defined (complexity, error) plane not only guide the development of data-driven models—from simple equations to sophisticated neural networks—but also promote process understanding. To distill knowledge from these hierarchies, we propose a multi-step approach: (1) use models along the Pareto front to hierarchically investigate the added value (functional representation, feature assimilation, spatial and temporal connectivities) that leads to incremental error decreases from the simplest to the most complex Pareto-optimal model; (2) generate hypotheses and propose models tailored to the system of interest based on this added value; and (3) once the models are sparse enough to be interpretable, reveal the system's underlying properties that previous theories or models may have overlooked. Beyond knowledge discovery, such hierarchies promote interpretability by explaining the added value of models that may initially seem overly complex, and sustainability by optimizing models' computational costs once their added value is justified.

We have chosen three weather and climate applications in realistic geography settings to showcase the potential of

machine learning in bridging fundamental scientific discovery and increasing operational requirements. Each case demonstrates a different nature of discovery: data-driven equation discovery for cloud cover parameterization, physics-guided architectures, informed by the available Pareto-optimal solution, that reflect the bidirectionality and vertical invariance of shortwave radiative transfer, and spatial information hidden within time series of the coarse environment when diagnosing tropical precipitation. However, all three insights required retaining the full family of trained models and might have been overlooked had we focused on a single optimal model.

In all three cases, neural networks proved particularly advantageous as they can quickly explore large datasets and generate hypotheses about problems that are or appear high-dimensional. We nonetheless emphasize that within this framework, deep learning is an integral part but not the ultimate goal of the knowledge generation process. Simpler models rivaling the accuracy of deep neural networks, initially intractable, may emerge once the necessary functional representations, features, and spatio-temporal connectivities are distilled. From this perspective, a rekindled interest in multi-objective optimization and hierarchical thinking would open the door to extracting new, human-interpretable scientific knowledge from ever-expanding geoscientific data, while paving the way for the adoption of machine learning in operational applications by fostering informed trust in their predictions.

Data availability statement. The reduced data and code necessary to reproduce this manuscript's figures are available in the following GitHub repository: `https://github.com/tbeucler/2024_Pareto_Distillation`. The release for this manuscript is archived via Zenodo using the following DOI: `https://zenodo.org/records/13217736`. The cloud cover schemes and analysis code can be found at `https://github.com/EyringMLClimateGroup/`

`grundner23james_EquationDiscovery_CloudCover/tree/main` and are preserved at `https://doi.org/10.5281/zenodo.8220333`. The coarse-grained model output used to train and evaluate the data-driven models amounts to several TB and can be reconstructed with the scripts provided in the GitHub repository. This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bd1179. For radiative transfer, we use the ml4rt Python library (`https://github.com/thunderhoser/ml4rt`), with version 4.0.1 preserved at `https://doi.org/10.5281/zenodo.13160776`). The dataset for radiative transfer is stored at `https://doi.org/10.5281/zenodo.13159877`. The BRNN architecture used can be created with the file `peter_brnn_architecture.py`; all other architectures can be created with the scripts named `pareto2024_make_*_templates.py`. The precipitation example uses "Precip-org" (`https://github.com/Sshamekh/Precip-org`), a Python repository managed by Sara Shamekh. DYAMOND data management was provided by the DKRZ and supported through the projects ESiWACE and ESiWACE2. The full data is available on the DKRZ HPC system through the DYAMOND initiative (`https://www.esiwace.eu/services/dyamond-initiative`).

<div style="text-align:center">

APPENDIX A

Calibrated Parameters for Cloud Cover Parameterization

</div>

In this appendix, we provide the values of the calibrated parameters in this manuscript's equations.

For the Sundqvist scheme (Eq. 9): $\text{RH}_{\text{surf}} = 0.55$, $\text{RH}_{\text{top}} = 0.01$, $\text{RH}_{\text{sat}} = 0.9/0.95$ (over land/sea), and $n = 2.12$.

For the two-feature linear regression (Eq. 10): $\alpha_{2F} = 1.027$, $\text{RH}_{2F} = 0.827$, and $\Delta T_{2F} = 198.6$K.

For the multilinear model (Eq. 13): $\alpha_1 = 1.861$, $\text{RH}_1 = 1.898$, $\Delta T_1 = 265.9$K, $q_1 = 0.3775$g/kg, $H_1 = 104.3$m.

For the quadratic model (Eq. 14): $\alpha_2 = 2.070$, $\text{RH}_{2,1} = 2.021$, $\Delta T_{2,1} = 259.7$K, $q_{2,1} = 0.4741$g/kg, $q_{2,2} = 3.161$g/kg, $q_{2,3} = 0.1034$g/kg, $H_2 = 665.6$m, $\text{RH}_{2,2} = 2.428$, $\Delta T_{2,2} = 206.0$K.

For the Wang scheme: $\alpha_{\text{Wang}} = 0.9105$, $\beta_{\text{Wang}} = 914 \cdot 10^3$.

For the PySR equation, $\overline{C} = 0.4435$,

$$\left(\frac{\partial C}{\partial \text{RH}}\right)_{\overline{\text{RH}},\overline{T}} = 1.159, \left(\frac{\partial C}{\partial T}\right)_{\overline{\text{RH}},\overline{T}} = \frac{0.0145}{\text{K}}, \quad \text{(A1)}$$

$$\left(\frac{\partial^2 C}{\partial \mathrm{RH}^2}\right)_{\overline{\mathrm{RH}},\overline{T}} = 4.06, \left(\frac{\partial C}{\partial \mathrm{RH}\partial T}\right)_{\overline{\mathrm{RH}},\overline{T}} = \frac{1.32.10^{-3}}{\mathrm{K}^2},$$

$$(A2)$$

$H_{\mathrm{PySR}} = 585\mathrm{m}, \ (d\mathrm{RH}/dz)_{\mathrm{max}\ C} = -2/\mathrm{km}, \ \epsilon_{\mathrm{PySR}} = 1.06,$
$\lambda_\ell = 3.845.10^5, \lambda_i = 1.448.10^6.$

For the GP-GOMEA equation: $\alpha_G = 66, \ \beta_G = 1.36.10^{-4}, \mathrm{RH}_G = 11.5\%, \gamma_G = 0.194, q_{G,i} = 4.367\mathrm{mg/kg},$ $\delta_G = 0.774, q_{G,\ell+} = 88.05\mathrm{mg/kg}, q_{G,\ell-} = 5.61\mathrm{mg/kg},$ and $\epsilon_G \to 0$ (0 cloud cover assigned in the absence of condensates and model calibrated only with condensates present).

APPENDIX B

Features for Shortwave Radiative Transfer Emulation

The 22 features X_η with a vertical profile are temperature T [K], pressure p [Pa], specific humidity q [kg kg^{-1}], relative humidity **RH** [%], liquid water content **LWC** [kg m^{-3}], ice water content **IWC** [kg m^{-3}], downward and upward liquid water path **LWP$_\downarrow$**, **LWP$_\uparrow$** [kg m^{-2}], downward and upward ice water path **IWP$_\downarrow$**, **IWP$_\uparrow$** [kg m^{-2}], downward and upward water vapor path **WVP$_\downarrow$**, **WVP$_\uparrow$** [kg m^{-2}], O$_3$ mixing ratio [kg kg^{-1}], height above ground level z [m], height layer thickness Δz [m], pressure layer thickness Δp [Pa], liquid effective radius r_ℓ [m], ice effective radius r_i [m], N$_2$O concentration [ppmv], CH$_4$ concentration [ppmv], CO$_2$ concentration [ppmv], and aerosol extinction [m^{-1}].

References

Arias, P., and Coauthors, 2021: Climate change 2021: The physical science basis. *Contribution of working group14 I to the sixth assessment report of the Intergovernmental Panel on Climate Change*, 319–329.

Balaji, V., 2021: Climbing down charney's ladder: machine learning and the post-dennard era of computational climate science. *Philosophical Transactions of the Royal Society A*, **379 (2194)**, 20200 085.

Balaji, V., F. Couvreux, J. Deshayes, J. Gautrais, F. Hourdin, and C. Rio, 2022: Are general circulation models obsolete? *Proceedings of the National Academy of Sciences*, **119 (47)**, e2202075 119.

Bao, J., B. Stevens, L. Kluft, and C. Muller, 2024: Intensification of daily tropical precipitation extremes from more organized convection. *Science Advances*, **10 (8)**, eadj6801.

Bartlett, P. L., O. Bousquet, and S. Mendelson, 2005: Local rademacher complexities.

Belochitski, A., and V. Krasnopolsky, 2021: Robustness of neural network emulations of radiative transfer parameterizations in a state-of-the-art general circulation model. *Geoscientific Model Development*, **14 (12)**, 7425–7437.

Ben-Bouallegue, Z., and Coauthors, 2023: The rise of data-driven weather forecasting. *arXiv preprint arXiv:2307.10128*.

Bertoli, G., F. Ozdemir, S. Schemm, and F. Perez-Cruz, 2023: Revisiting machine learning approaches for short- and longwave radiation inference in weather and climate models, part i: Offline

performance. *ESS Open Archive*, https://doi.org/10.22541/essoar.169109567.78839949/v1.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2022: Panguweather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*.

Bony, S., and Coauthors, 2013: Carbon dioxide and climate: Perspectives on a scientific assessment. *Climate science for serving society: Research, modeling and prediction priorities*, 391–413.

Bony, S., and Coauthors, 2015: Clouds, circulation and climate sensitivity. *Nature Geoscience*, **8 (4)**, 261–268.

Bröcker, J., 2009: Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **135 (643)**, 1512–1519.

Buhrmester, V., D. Münch, and M. Arens, 2021: Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*, **3 (4)**, 966–989.

Censor, Y., 1977: Pareto optimality in multiobjective problems. *Applied Mathematics and Optimization*, **4 (1)**, 41–59.

Charney, J. G., and Coauthors, 1979: *Carbon dioxide and climate: a scientific assessment*. National Academy of Sciences, Washington, DC.

Cheruy, F., F. Chevallier, J.-J. Morcrette, N. A. Scott, and A. Chédin, 1996: Une méthode utilisant les techniques neuronales pour le calcul rapide de la distribution verticale du bilan radiatif thermique terrestre. *Comptes Rendus de l'Academie des Sciences Serie II*, **322**, 665–672.

Chevallier, F., F. Chéruy, N. Scott, and A. Chédin, 1998: A neural network approach for a fast and accurate computation of a longwave radiative budget. *Journal of Applied Meteorology and Climatology*, **37 (11)**, 1385–1397.

Chevallier, F., J.-J. Morcrette, F. Chéruy, and N. Scott, 2000: Use of a neural-network-based long-wave radiative-transfer scheme in the ecmwf atmospheric model. *Quarterly Journal of the Royal Meteorological Society*, **126 (563)**, 761–776.

Clare, M. C., O. Jamil, and C. J. Morcrette, 2021: Combining distribution-based neural networks to predict weather forecast probabilities. *Quarterly Journal of the Royal Meteorological Society*, **147 (741)**, 4337–4357.

Clough, S. A., M. J. Iacono, and J.-L. Moncet, 1992: Line-by-line calculations of atmospheric fluxes and cooling rates: Application to water vapor. *Journal of Geophysical Research: Atmospheres*, **97 (D14)**, 15 761–15 785.

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. *arXiv preprint arXiv:2305.01582*.

Das, A., and P. Rad, 2020: Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Duras, J., F. Ziemen, and D. Klocke, 2021: The dyamond winter data collection. *EGU general assembly conference abstracts*, EGU21–4687.

Fisher, A., C. Rudin, and F. Dominici, 2019: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, **20 (177)**, 1–81.

Gentine, P., V. Eyring, and T. Beucler, 2021: Deep learning for the parametrization of subgrid processes in climate models. *Deep learning for the Earth sciences: A comprehensive approach to remote sensing, climate science, and geosciences*, 307–314.

Giorgetta, M. A., and Coauthors, 2018: Icon-a, the atmosphere component of the icon earth system model: I. model description. *Journal of Advances in Modeling Earth Systems*, **10 (7)**, 1613–1637.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, **102 (477)**, 359–378.

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.

Grundner, A., T. Beucler, P. Gentine, F. Iglesias-Suarez, M. A. Giorgetta, and V. Eyring, 2022: Deep learning based cloud cover parameterization for icon. *Journal of Advances in Modeling Earth Systems*, **14 (12)**, e2021MS002 959.

Haynes, K., R. Lagerquist, M. McGraw, K. Musgrave, and I. Ebert-Uphoff, 2023: Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. *Artificial Intelligence for the Earth Systems*, **2 (2)**, 220 061.

Held, I. M., 2005: The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society*, **86 (11)**, 1609–1614.

Hertel, L., J. Collado, P. Sadowski, J. Ott, and P. Baldi, 2020: Sherpa: Robust hyperparameter optimization for machine learning. *SoftwareX*, **12**, 100 591.

Hogan, R. J., and A. Bozzo, 2018: A flexible and efficient radiation scheme for the ecmwf model. *Journal of Advances in Modeling Earth Systems*, **10 (8)**, 1990–2008.

Jakhar, K., Y. Guan, R. Mojgani, A. Chattopadhyay, and P. Hassanzadeh, 2024: Learning closed-form equations for subgrid-scale closures from high-fidelity data: Promises and challenges. *Journal of Advances in Modeling Earth Systems*, **16 (7)**, e2023MS003 874.

Jeevanjee, N., P. Hassanzadeh, S. Hill, and A. Sheshadri, 2017: A perspective on climate model hierarchies. *Journal of Advances in Modeling Earth Systems*, **9 (4)**, 1760–1771.

Keisler, R., 2022: Forecasting global weather with graph neural networks. *arXiv preprint arXiv:2202.07575*.

Khairoutdinov, M. F., P. N. Blossey, and C. S. Bretherton, 2022: Global system for atmospheric modeling: Model description and preliminary results. *Journal of Advances in Modeling Earth Systems*, **14 (6)**, e2021MS002 968.

Kim, P. S., and H.-J. Song, 2022: Usefulness of automatic hyperparameter optimization in developing radiation emulator in a numerical weather prediction model. *Atmosphere*, **13 (5)**, 721.

Kingma, D. P., and J. Ba, 2014: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koza, J. R., 1994: Genetic programming as a means for programming computers by natural selection. *Statistics and computing*, **4**, 87–112.

La Cava, W., B. Burlacu, M. Virgolin, M. Kommenda, P. Orzechowski, F. O. de França, Y. Jin, and J. H. Moore, 2021: Contemporary symbolic regression methods and their relative performance. *Advances in neural information processing systems*, **2021 (DB1)**, 1.

Lagerquist, R., D. Turner, I. Ebert-Uphoff, J. Stewart, and V. Hagerty, 2021: Using deep learning to emulate and accelerate a radiative transfer model. *Journal of Atmospheric and Oceanic Technology*, **38 (10)**, 1673–1696.

Lagerquist, R., D. D. Turner, I. Ebert-Uphoff, and J. Q. Stewart, 2022: Estimating full longwave and shortwave radiative transfer with neural networks of varying complexity. *Journal of Atmospheric and Oceanic Technology*, **40 (11)**, 1407–1432.

Lam, R., and Coauthors, 2022: Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*.

Lang, S., and Coauthors, 2024: Aifs-ecmwf's data-driven forecasting system. *arXiv preprint arXiv:2406.01465*.

Lin, X., H.-L. Zhen, Z. Li, Q.-F. Zhang, and S. Kwong, 2019: Pareto multi-task learning. *Advances in neural information processing systems*, **32**.

Liou, K.-N., 2002: *An introduction to atmospheric radiation*, Vol. 84. Elsevier.

Maas, A., A. Hannun, and A. Ng, 2013: Rectifier nonlinearities improve neural network acoustic models. *International Conference on Machine Learning*, Atlanta, Georgia, International Machine Learning Society, URL http://robotics.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf.

Maher, P., and Coauthors, 2019: Model hierarchies for understanding atmospheric circulation. *Reviews of Geophysics*, **57 (2)**, 250–280.

Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff, 2022: Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artificial Intelligence for the Earth Systems*, **1 (4)**, e220 012.

Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff, 2023: Carefully choose the baseline: Lessons learned from applying xai attribution methods for regression tasks in geoscience. *Artificial Intelligence for the Earth Systems*, **2 (1)**, e220 058.

Mansfield, L. A., A. Gupta, A. C. Burnett, B. Green, C. Wilka, and A. Sheshadri, 2023: Updates on model hierarchies for understanding and simulating the climate system: A focus on data-informed methods and climate change impacts. *Journal of Advances in Modeling Earth Systems*, **15 (10)**, e2023MS003 715.

Mauritsen, T., and Coauthors, 2019: Developments in the mpi-m earth system model version 1.2 (mpi-esm1. 2) and its response to increasing co2. *Journal of Advances in Modeling Earth Systems*, **11 (4)**, 998–1038.

Miettinen, K., 1999: *Nonlinear multiobjective optimization*, Vol. 12. Springer Science & Business Media.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: Rrtm, a validated correlated-k model for the longwave. *Journal of Geophysical Research: Atmospheres*, **102 (D14)**, 16 663–16 682.

Molnar, C., 2020: *Interpretable machine learning*. Lulu. com.

Morcrette, J.-J., 2000: On the effects of the temporal and spatial sampling of radiation fields on the ecmwf forecasts and analyses. *Monthly weather review*, **128 (3)**, 876–887.

Moseley, C., C. Hohenegger, P. Berg, and J. O. Haerter, 2016: Intensification of convective extremes driven by cloud–cloud interaction. *Nature Geoscience*, **9 (10)**, 748–752.

Nowack, P., J. Runge, V. Eyring, and J. D. Haigh, 2020: Causal networks for climate model evaluation and constrained projections. *Nature communications*, **11 (1)**, 1415.

Pathak, J., and Coauthors, 2022: Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.

Pearl, J., and Coauthors, 2000: Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, **19 (2)**, 3.

Petersen, B. K., M. Landajuela, T. N. Mundhenk, C. P. Santiago, S. K. Kim, and J. T. Kim, 2019: Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. *arXiv preprint arXiv:1912.04871*.

Rasp, S., P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, 2020: Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, **12 (11)**, e2020MS002 203.

Rasp, S., and N. Thuerey, 2021: Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, **13 (2)**, e2020MS002 405.

Robertson, A. W., and M. Ghil, 2000: Solving problems with gcms: General circulation models and their role in the climate modeling hierarchy.

Roeckner, E., and Coauthors, 1996: The atmospheric general circulation model echam-4: Model description and simulation of present-day climate. URL https://api.semanticscholar.org/CorpusID:14424015.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 234–241.

Rudin, C., 2019: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, **1 (5)**, 206–215.

Scher, S., and G. Messori, 2021: Ensemble methods for neural network-based weather forecasts. *Journal of Advances in Modeling Earth Systems*, **13 (2)**.

Schmidt, M., and H. Lipson, 2009: Distilling free-form natural laws from experimental data. *science*, **324 (5923)**, 81–85.

Schneider, T., L. R. Leung, and R. C. Wills, 2024: Opinion: Optimizing climate models with process knowledge, resolution, and artificial intelligence. *Atmospheric Chemistry and Physics*, **24 (12)**, 7041–7062.

Schneider, T., J. Teixeira, C. S. Bretherton, F. Brient, K. G. Pressel, C. Schär, and A. P. Siebesma, 2017: Climate goals and computing the future of clouds. *Nature Climate Change*, **7 (1)**, 3–5.

Schneiderman, H., 2024: Incorporation of physical equations into a neural network structure for predicting shortwave radiative heat transfer.

23rd Conference on Artificial Intelligence for Environmental Science, AMS Annual Meeting, Baltimore, MD, USA.

Seneviratne, S. I., and Coauthors, 2021: Weather and climate extreme events in a changing climate. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. R. Huang, K. Leitzell, E. Lonnoy, J. B. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, Eds., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1513–1766, https://doi.org/10.1017/9781009157896.013.

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.

Sherwood, S. C., S. Bony, and J.-L. Dufresne, 2014: Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, **505 (7481)**, 37–42.

Stevens, B., and Coauthors, 2019: Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, **6 (1)**, 1–17.

Stevens, B., and Coauthors, 2020: The added value of large-eddy and storm-resolving models for simulating clouds and precipitation. *Journal of the Meteorological Society of Japan. Ser. II*, **98 (2)**, 395–435.

Sundqvist, H., E. Berge, and J. E. Kristjánsson, 1989: Condensation and cloud parameterization studies with a mesoscale numerical weather prediction model.

Ukkonen, P., 2022: Exploring pathways to more accurate machine learning emulation of atmospheric radiative transfer. *Journal of Advances in Modeling Earth Systems*, **14 (4)**, e2021MS002 875.

Ukkonen, P., and M. Chantry, 2024: Representing sub-grid processes in weather and climate models via sequence learning. *Authorea Preprints*.

Vapnik, V. N., and A. Y. Chervonenkis, 2015: On the uniform convergence of relative frequencies of events to their probabilities. *Measures of complexity: festschrift for alexey chervonenkis*, Springer, 11–30.

Veerman, M. A., R. Pincus, R. Stoffer, C. M. Van Leeuwen, D. Podareanu, and C. C. Van Heerwaarden, 2021: Predicting atmospheric optical properties for radiative transfer computations using neural networks. *Philosophical Transactions of the Royal Society A*, **379 (2194)**, 20200 095.

Virgolin, M., T. Alderliesten, C. Witteveen, and P. A. Bosman, 2021: Improving model-based genetic programming for symbolic regression of small expressions. *Evolutionary computation*, **29 (2)**, 211–237.

Wang, Y., S. Yang, G. Chen, Q. Bao, and J. Li, 2023: Evaluating two diagnostic schemes of cloud-fraction parameterization using the cloudsat data. *Atmospheric Research*, **282**, 106 510.

Weyn, J. A., D. R. Durran, and R. Caruana, 2019: Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, **11 (8)**, 2680–2693.

Weyn, J. A., D. R. Durran, and R. Caruana, 2020: Improving data-driven global weather prediction using deep convolutional neural networks

on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, **12 (9)**, e2020MS002 109.

Xu, K.-M., and D. A. Randall, 1996: A semiempirical cloudiness parameterization for use in climate models. *Journal of the atmospheric sciences*, **53 (21)**, 3084–3102.

Zhou, Z., M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, 2019: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, **39 (6)**, 1856–1867.