Machine Learning for Clouds and Climate

Tom Beucler[1,2], Imme Ebert-Uphoff[3,4], Stephan Rasp[5], Michael Pritchard[1], Pierre Gentine[2]

[1]Department of Earth System Science, University of California, Irvine, CA, USA
[2]Department of Earth and Environmental Engineering, Columbia University, New York, NY, USA
[3]Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA.
[4]Electrical and Computer Engineering, Colorado State University, Fort Collins, CO, USA
[5]ClimateAi, Inc

Key Points:

- Machine learning (ML) helps model the interaction between clouds and climate using large datasets.
- We review physics-guided/explainable ML applied to cloud-related processes in the climate system.
- We also provide a guide to scientists who would like to get started with ML.

Corresponding author: Tom Beucler, tom.beucler@gmail.com

Abstract
Machine learning (ML) algorithms are powerful tools to build models of clouds and climate that are more faithful to the rapidly-increasing volumes of Earth system data than commonly-used semiempirical models. Here, we review ML tools, including interpretable and physics-guided ML, and outline how they can be applied to cloud-related processes in the climate system, including radiation, microphysics, convection, and cloud detection, classification, emulation, and uncertainty quantification. We additionally provide a short guide to get started with ML and survey the frontiers of ML for clouds and climate.

1 Introduction

Machine learning (ML) describes algorithms that learn to perform a task from data without being explicitly programmed for that task. This is in contrast to traditional algorithms, such as existing cloud parameterizations, that are explicitly programmed based on human expertise. Because ML can extract knowledge from large data volumes, it has revolutionized computer vision, natural language processing, and recommender systems. As we get an unprecedented amount of Earth system data from diverse observations (remote sensing, in situ measurements, citizen science) and models (high-resolution simulations, large ensembles of simulations), ML is quickly permeating geosciences (Karpatne et al. (2018); Bergen et al. (2019); Irrgang et al. (2021), Chap 1) while ML practitioners are increasingly interested in tackling climate change-related problems (Rolnick et al., 2019). Taking the example of numerical weather prediction, ML has already improved post-processing (Grönquist et al., 2020), statistical forecasting (McGovern et al., 2017), and nowcasting (Sønderby et al., 2020), along with promising attempts at purely data-driven weather forecasting (Rasp & Thuerey, 2020).

In contrast to numerical weather prediction, climate science exhibits challenges that have limited direct ML applications. First, many more of the variables used in climate models cannot be directly observed, such as cloud condensation rates and radiative effects, confining most ML attempts to the emulation of numerical models. Second, as understanding is often more important than accuracy for key challenges in clouds and climate science (Bony et al., 2015), climate scientists may avoid methods they deem uninterpretable. Third, it is notoriously hard to benchmark ML models of climate because perfectly labeled climate datasets are rare; two recent attempts had to rely on human labeling to benchmark shallow cloud classification (Rasp, Schulz, et al., 2020) and extreme weather events (Kashinath et al., 2021), while the most recent ML benchmark for data-driven weather forecasting exclusively relied on meteorological reanalysis data (Rasp, Dueben, et al., 2020). Finally, making long-term predictions in a changing climate is an extrapolation problem, meaning that ML algorithms solely trained on present-day climate data might fail to make predictions in the future (unobserved) climate (Schneider, Teixeira, et al., 2017; Beucler et al., 2020) with current methods.

In this final chapter, we argue that despite these challenges, ML algorithms are promising tools to consistently capture climate statistics from large datasets in their full complexity (Watson-Parris, 2020), which purely physical models struggle to do. After an overview of ML in Sec 2, we present promising applications of ML to clouds and climate in Sec 3, and give advice on how to get started with scientific ML in Sec 4 before concluding in Sec 5.

2 Machine Learning

To address the challenges specific to clouds and climate, we survey both physics-guided (Sec 2.3) and explainable ML (Sec 2.2) after introducing common ML tools (Sec 2.1).

Table 1. Acronyms used in this chapter.

Acronym	Description	Cross-reference
cGAN	conditional Generative Adversarial Network	3.4
CNN	Convolutional Neural Network	3.4, 3.7, 4.1
LRP	Layer-wise Relevance Propagation	2.2
ML	Machine Learning	All Sections
MRMS	Multi-Radar Multi-Sensor	2.2
NN	Neural Network	All Sections
RF	Random Forest	2.1, 2.3, 3.1, 3.3, 4.2
SR	Super-Resolution	3.4
XAI	eXplainable Artificial Intelligence	2.2, 3.5, 5

2.1 Overview of Machine Learning Tools

To identify the ML algorithm most appropriate for the task at hand, it is useful to introduce two hierarchies. First, we can classify algorithms based on the availability of external supervisory information from a human expert or other sources. In *supervised* learning, by far the most common ML approach, the algorithm learns to map an input (or features) to an output (or target) from a training dataset consisting of input-output pairs. The algorithm is then trained to make output predictions as close as possible to the training targets. This is usually done via the minimization of a cost or loss function (e.g. mean squared error, cross-entropy loss), which maps the algorithm's output to a single number that can be minimized. In contrast, *unsupervised* learning extracts features from data and without the need for labeled input-output pairs. Examples include dimensionality reduction (via e.g. principal component analysis, autoencoders, self-organizing maps, see Chap 8 of Géron (2019)), clustering algorithms (e.g. k-means, DBSCAN, Gaussian mixture models, see Chap 9 of Géron (2019)), and generative learning, which can be broadly defined as a probabilistic model for how a dataset is generated (e.g. variational autoencoders, generative adversarial networks, see Foster (2019)).

Second, we can classify algorithms based on the number of parameters that are fitted as part of the training process, as illustrated in Fig 1 (orange axis). It is always preferable to start with simple models because they are easier to interpret (green axis) and less computationally expensive to train. For *regression* tasks that seek to predict continuous outputs, linear regressions are arguably the simplest models, while logistic regressions are simple models for *classification* tasks that seek to predict categorical outputs. Decision trees, which use simple conditional statements (the "branches") to match each input's characteristics to corresponding outputs (the "leaves"), can be trained and averaged to form random forests (RFs) that exhibit lower variance in their predictions than single decision trees. Finally, neural networks (NNs), which have computational architectures loosely based on biological networks of neurons, are powerful non-linear regression and classification tools that are typically expensive to train if deep, i.e. composed of many neuron layers.

While in theory complex models such as NNs can fit any nonlinear mapping given enough data (Scarselli & Tsoi, 1998), this high representation power comes at a cost: Simple models such as linear regressions optimized using least squares have a unique set of optimal parameters, but the parameters of more complex models often need to be optimized stochastically with no guarantee of an optimal solution. As the number of re-

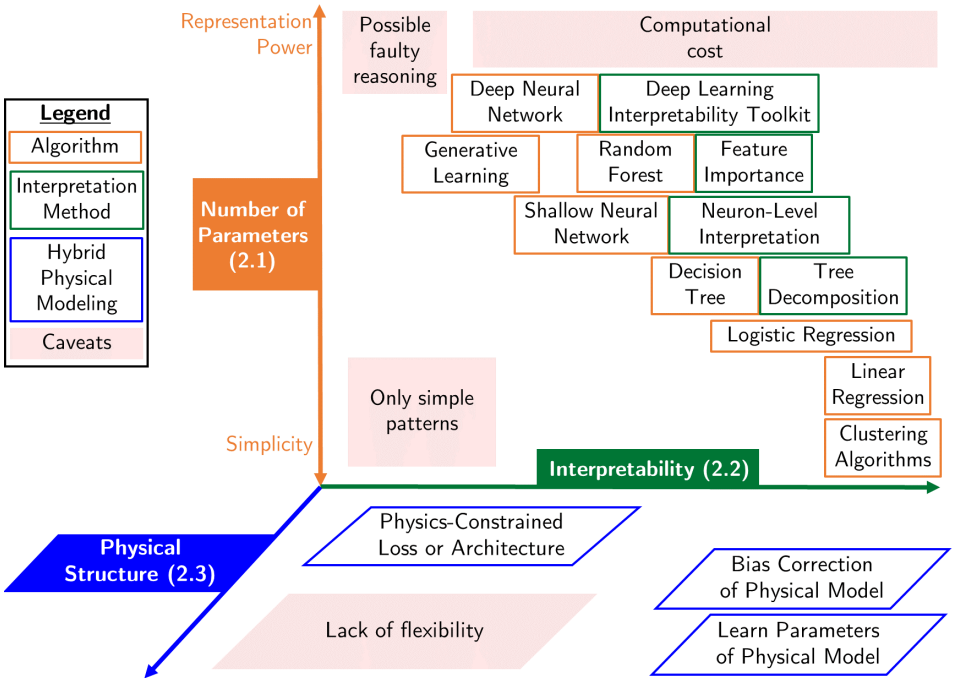

Figure 1. ML algorithms (orange boxes), corresponding interpretation methods (green boxes), and hybrid ML-physical modeling methods (blue boxes) organized in the three-dimensional space formed by the number of parameters (Sec 2.1), interpretability (Sec 2.2), and physical structure (Sec 2.3). Caveats of different configurations are indicated in light red boxes.

quired training samples increases with the number of model parameters, complex models tend to easily *overfit* the training dataset, i.e. they memorize the dataset "by heart", leading to poor performance when the model is applied to data samples it has not been trained on. To mitigate this issue, best ML practices require to split the available data into three datasets: A *training* dataset to optimize the model's parameters, a *validation* dataset to detect overfitting to the training dataset and to optimize the model's hyperparameters (e.g., number of layers of a NN), and a *test* dataset to evaluate the final model on data samples it has never seen before. In Earth science, splitting the data into independent training/validation/test datasets can be particularly challenging as samples tend to exhibit high spatiotemporal correlation (Karpatne et al., 2018), meaning that e.g., splitting the data by long time interval (such as years) may be preferable to a random split. Ying (2019) surveys common ML methods to avoid overfitting, including regularization of the model's parameters (e.g., L1, L2, Dropout) and early-stopping of the model's training.

If simple models fail to capture the data and a complex model is required, another caveat is that complex models such as NNs require large amounts of (usually labeled) data samples to train their many free parameters, often more than is available from e.g. observations. An elegant workaround, referred to as *transfer learning* is to build on a NN trained for a different but closely related task. Examples include leveraging NNs classifying the 50M images from the famous "ImageNet" dataset (Deng et al., 2009) to make predictions from satellite observations (Marmanis et al., 2015), and leveraging NNs trained on climate model data to make predictions from meteorological reanalysis data (Ham et al., 2019; Rasp & Thuerey, 2020). In these cases, building on existing NNs reduces not only the required sample size, but also development time and cost. That being said,

the complex resulting NN may still be hard to interpret, motivating methods to explain the strategies the NN uses to make accurate predictions.

2.2 Interpretable/Explainable Machine Learning

The literature often distinguishes *interpretable* ML, which generally refers to models that are designed to be a priori understandable, and *explainable* machine learning (known as eXplainable Artificial Intelligence or XAI in the literature), which refers to methods that try to explain a posteriori the prediction of a trained model, often for specific samples (Rudin, 2019). Both frameworks help design more transparent and hence trustworthy ML models. Two books (Molnar, 2019; Samek et al., 2019) provide in-depth discussion of both approaches. In this section we focus on XAI to help interpret trained ML models for climate science.

In Fig 1, methods to interpret ML algorithms are indicated with green boxes (simple methods such as linear regressions or decision trees are interpretable by construction): For example, tree decomposition follows the path of a decision tree to decompose a given prediction as the sum of each input's contribution. While methods that seek to interpret individual neurons of a neural network have been successful in computer vision (Olah et al., 2017, 2018; Carter et al., 2019; Yosinski et al., 2015; Bau et al., 2017), they are difficult to apply in climate science where objects have fuzzy boundaries (e.g. heat waves, atmospheric rivers). Exceptions involve cases where we can focus on only a few neurons, including McGovern et al. (2020) who ranked individual neurons in a NN trained for tornado prediction using the neurons' ability to discriminate output classes.

Therefore, XAI for climate science often focuses on understanding the prediction of overall ML models for specific samples, most notably via *attribution* methods. Attribution methods, most common in NNs for image classification, ask: Given a sample and its label, which pixels in the input image are most important to correctly predict the label? The result is a *heatmap* (or attribution map) indicating the most important areas in the input images. Attribution methods include saliency maps (Simonyan et al., 2013), SmoothGrad (Smilkov et al., 2017), Integrated Gradients (Sundararajan et al., 2017), Layer-wise relevance propagation (Montavon et al., 2017, 2018) (see Fig 2), DeepLIFT (Shrikumar et al., 2017), SHAP (Lundberg & Lee, 2017), GradCAM (Selvaraju et al., 2017), and occlusion methods (Fong & Vedaldi, 2019). Attribution methods have been successfully applied in atmospheric science (Gagne II et al., 2019; Brenowitz, Beucler, et al., 2020; Lagerquist et al., 2020; McGovern & Lagerquist, 2020; Barnes et al., 2020), and we refer the curious reader to recent reviews by McGovern et al. (2019) and Ebert-Uphoff and Hilburn (2020) on the use of XAI methods in meteorological applications.

Beyond attribution methods, other methods seeking to interpret ML models as a whole include *backwards optimization*, which calculates the input maximizing confidence in a given output (McGovern et al., 2019), and *ablation* studies (Raghu & Schmidt, 2020), which remove certain capabilities from the model's architecture and retrain the resulting model to test how important these capabilities are (see Sønderby et al. (2020) and Ebert-Uphoff and Hilburn (2020) for meteorological applications). Finally, when using XAI methods one should keep in mind (1) potential limitations of the attribution methods (Adebayo et al., 2018; Kindermans et al., 2017; Bansal et al., 2020), and (2) the importance of providing explanations that are tailored to the user's concerns and needs (Ras et al., 2018; Rutjes et al., 2019).

2.3 Physics-Guided Machine Learning

Even highly-interpretable ML models of physical processes such as clouds are often physically-inconsistent in two major ways, limiting their impact on climate science. First, they may violate physical laws we know should hold, such as mass and energy con-

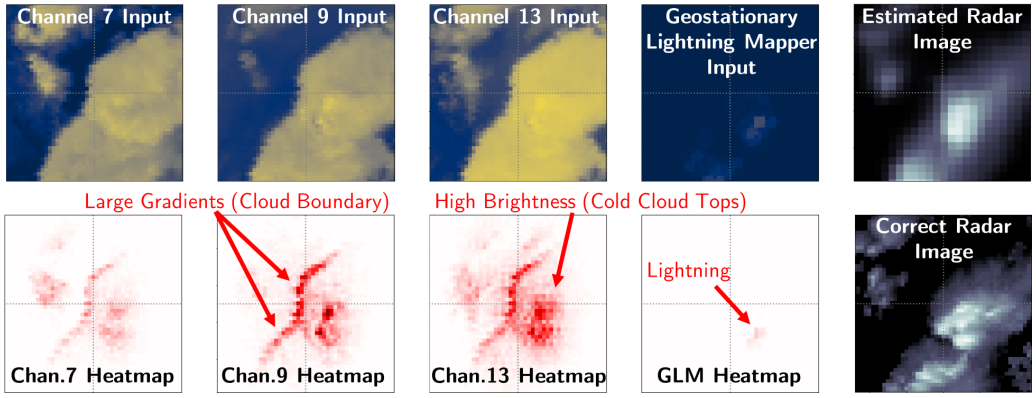

Figure 2. Use of an attribution method, namely Layer-Wise Relevance Propagation (LRP), to analyze a neural network model that seeks to translate imagery from the Geostationary Operational Environmental Satellite into Multi-Radar Multi-Sensor (MRMS) imagery. LRP results (bottom row) show *where* in the input channels the neural network is focusing when predicting the MRMS value for a single (central) pixel. Using LRP indicated for various samples three strategies for the neural network to trigger large MRMS values, namely i) cloud boundaries, ii) cold cloud tops, and iii) presence of lightning in the Geostationary Lightning Mapper channel. See Ebert-Uphoff and Hilburn (2020) for details.

servation. Second, they may fail to generalize to unseen conditions such as extreme weather events and shifts in the distribution of variables (e.g. geographical shift, climate change).

To address these issues, we can restrict the ML output space to physically-plausible solutions by adding physical structure to the ML framework. This is referred to as *physics-guided ML* (Willard et al., 2020; Reichstein et al., 2019) or hybrid ML-physical modeling (blue axis in Fig 1). We can group physics-guided ML approaches into three categories (three blue boxes in Fig 1), noting that the optimal approach depends on the task and data at hand.

First, we can integrate physical knowledge without changing the ML algorithm's architecture by adding a penalizing term to its loss function, akin to a Lagrange multiplier (Karpatne et al., 2017; Wu et al., 2020; Willard et al., 2020; de Bezenac et al., 2019). However, these soft constraints are often insufficient in climate science, where e.g. conservation laws need to hold exactly since climatic trends are driven by small energy imbalances in the Earth system. This motivates changing the model's architecture to enforce conservation laws to within machine precision, e.g., by using RFs that enforce conservation laws by construction, as long as these laws are a linear combination of the outputs (O'Gorman & Dwyer, 2018a), or by augmenting the NN's architecture with physical constraint layers so as to exactly enforce conservation laws (Beucler, Pritchard, et al., 2019). However, enforcing physical laws in this way is often not enough to properly generalize outside of the training set (O'Gorman & Dwyer, 2018b; Rasp et al., 2018; Reichstein et al., 2019; Beucler et al., 2020). To address this, the penalizing term in the loss function may enforce assumed (and often approximate) dynamics based on a physical model (Wu et al., 2020; Camps-Valls et al., 2018; Raissi et al., 2019; Pang et al., 2019; Raissi et al., 2017; Mao et al., 2020; D. Zhang et al., 2020; Sun et al., 2020; Gao et al., 2020; Wang et al., 2017). Even if imperfect, such constraints improve predictions in unseen conditions by reducing the range of possible outputs, unlike pure ML approaches that have to make predictions in much larger phase spaces (Karpatne et al., 2017; T. Yang et al., 2019).

If loss-based constraints do not provide enough physical structure, we can use ML to bias correct a physical prior or calibrate a physical model's free parameters. Examples include machine learning the parameters of Earth System Model parameterizations within the structure of the known governing equations (Schneider, Lan, et al., 2017), learning an effective diffusion coefficient to represent turbulent processes in the boundary layer rather than learning the full turbulent fluxes (Reichstein et al., 2019; Camps-Valls et al., 2018), or only learning certain coefficients of the Reynolds stress tensor to preserve Galilean invariance (Ling et al., 2016). A caveat of imposing too much physical structure is the resulting lack of flexibility to model the data; for instance, at standard climate model resolutions, even the best ML fit of an eddy-diffusion model (Siebesma et al., 2007) cannot properly capture the non-local transport associated with shallow convection in the boundary layer.

3 Application to Clouds and Climate

We now give concrete examples of how ML can be applied to clouds and climate. Sections are ordered by spatial scales whenever possible, from radiative transfer operating at the atomic level (Sec 3.1) to cloud classification at the planetary scale (Sec 3.7).

3.1 Radiative Transfer

Radiative transfer is defined as the energy transfer in the form of electromagnetic radiation. As solar radiation is the atmosphere's largest energy source, while thermal radiation to space is its largest energy sink, atmospheric models cannot forego calculating the heat transfer resulting from solar (shortwave) radiation and thermal (longwave) radiation. While we have excellent empirical knowledge of how molecules in the atmosphere absorb and emit radiation at each electromagnetic spectral line, it is computationally intractable to calculate radiative transfer line-by-line and atmospheric models rely on different levels of approximation, including (1) integrating radiative fluxes over spectral bands of predetermined width, (2) neglecting the three-dimensional nature of radiation by assuming strictly vertical fluxes (plane-parallel approximation); and (3) calculating radiation using a resolution in time and/or space that is coarser than the model's standard resolution.

Aside from recent work using RFs and NNs to reproduce the variability in surface solar irradiance resulting from the three-dimensional interaction between radiation and shallow cumuli (Gristey et al. (2020), addressing approximation 2), ML has mostly been used to improve the *temporal* resolution of radiative transfer (approximation 3). By replacing the original, computationally expensive radiative scheme with its ML-emulated counterpart (computationally inexpensive once trained), atmospheric models can be accelerated, with the possibility of calling the radiative scheme every time step to improve the quality of cloud-radiation interactions and numerical weather predictions. To keep using physical equations when they are available while still accelerating atmospheric models, it is also possible to only replace expensive computations within radiative schemes, such as gas optics (Ukkonen et al., 2020; Veerman et al., 2020). ML emulation of radiative transfer has been tested in meteorological (Chevallier et al., 2000; V. M. Krasnopolsky & Fox-Rabinovitz, 2006), global climate (Belochitski et al., 2011; V. M. Krasnopolsky et al., 2008; V. Krasnopolsky et al., 2010; Pal et al., 2019), and cloud-resolving (Roh & Song, 2020) models. Recent efforts using ML to emulate subgrid-scale thermodynamics (Rasp et al. (2018), see Sec 3.3) usually include subgrid-scale radiative cooling, making it an example where ML helps improve the *spatial* resolution of radiative transfer, e.g. by capturing the effect of subgrid clouds on grid-scale radiative cooling.

Other notable applications of ML to atmospheric radiative transfer are the retrieval of cloud properties from satellite images (Min et al., 2020) and statistical predictions of surface radiative fluxes. This includes solar forecasting, where ML algorithms have been

successful at post-processing physical forecasts, nowcasting, 6-hour forecasting (see review by Voyant et al. (2017), Gala et al. (2016)), and land-atmosphere modeling, where ML can help emulate surface radiative properties that are distorted by coarse climate models, e.g. sun-induced chlorophyll fluorescence spectra (Rivera et al., 2015).

3.2 Microphysics

Microphysics refers to small-scale (sub-μm to cm) processes that affect cloud and precipitation particles (Chap 3, 4, 12). Microphysical parameterization schemes, which model the effect of cloud and precipitation particle populations on weather and climate, currently face two major challenges (Morrison, van Lier-Walqui, Fridlind, et al., 2020): (1) how to represent this effect despite the impossibility of simulating all particles individually; and (2) uncertainties in microphysical process rates owing to critical gaps in cloud physics knowledge, especially for ice-phase processes. As microphysics links various components of Earth's atmospheric water and energy cycles, its overly simplistic representation, typically limited to one or two moments from the particle distribution of each hydrometeor species, remains a large source of uncertainty in numerical weather forecast and climate simulations (Zelinka et al., 2020).

Thankfully, simulations describing the hydrometeor size distribution more faithfully via a bin approach improve accuracy. Although such simulations are too computationally expensive to be run for long-term climate predictions, they can provide high-quality training data for ML emulation (which should typically be much faster), with potential progress towards addressing challenge (1). Gettelman et al. (2020) trained multiple NNs to emulate the formation of warm rain from a bin scheme, and the NN-powered climate simulation was able to match the bin scheme's accuracy at a significantly reduced computational cost. Seifert and Rasp (2020) trained a NN to emulate microphysical conversion rates in a two-moment scheme from a Monte Carlo super-droplet simulation, but while process rates were reproduced with greater accuracy, the resulting solutions to the collision-coalescence ordinary differential equations did not match the reference simulation as well as the heuristically-designed parameterization. Analysis hints at the ill-posed nature of two-moment schemes and highlights the importance of evaluating ML models in long-term simulations. To address structural uncertainty in existing parameterizations, Morrison, van Lier-Walqui, Kumjian, and Prat (2020) propose a Bayesian framework to flexibly relate microphysical process rates to moments of the hydrometeor size distribution via generalized power series (Loeb, 1991), while ML-based short-term forecasts tend to forego microphysical schemes by directly fitting parameters of the size distribution to observational data, e.g. for hail prediction (Gagne et al., 2017).

3.3 Convection

Atmospheric convection, defined as the vertical motion driven by air density differences, is notoriously hard to simulate because of its multi-scale nature, as it leads to the formation of stratocumulus decks ($\sim 10-100$km, Chap 6, 10) while interacting with planetary-scale dynamics ($\sim 10^4$km, Chap 9, 11). Since clouds are radiatively active at all scales while convection vertically transports heat and water from the surface to the atmosphere, misrepresenting convection and clouds in atmospheric models leads to large errors in the energy balance and hence remains the largest source of uncertainty in long-term climate predictions (Bony et al., 2015). In standard global climate models, convection and clouds are not explicitly represented and their effect on the climate is approximated by *subgrid closures*. Traditionally, designing subgrid closures involved a heuristic process and manual tuning to observations. Recent storm-resolving simulations can explicitly represent tropical deep convection (Stevens et al., 2019) but are too expensive for climate change simulations. Therefore, emulating the effect of convection and clouds in storm-resolving simulations using statistical algorithms, including ML, could provide a shortcut towards data-driven subgrid closure in global climate models.

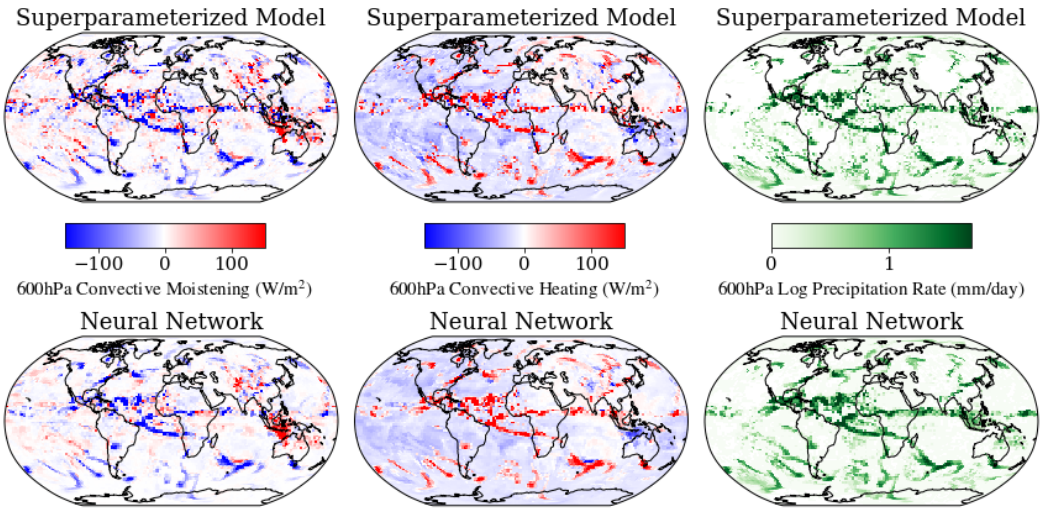

Figure 3. Lower-tropospheric subgrid moistening and heating tendencies (left) and subgrid precipitation (right) from a super-parametrized climate model with realistic boundary conditions (top) and as emulated by a deep NN (bottom). Adapted from Mooers, Pritchard, et al. (2020).

The first step in designing a data-driven subgrid closure is to create the training data; this can be done by coarse-graining a storm-resolving simulation to derive the correction term from coarse to high resolution (Yuval & O'Gorman, 2020; Brenowitz & Bretherton, 2019), by running a super-parameterized model that explicitly separates the storm and the coarse scales and directly provides that correction term (Gentine et al., 2018; Rasp et al., 2018), or by nudging a standard climate model to observations or meteorological reanalysis (Watt-Meyer et al., 2020; McGibbon & Bretherton, 2019). The second step is to choose the ML algorithm: While NNs are commonly used to emulate subgrid-scale parameterization because they usually yield the best fits (V. M. Krasnopolsky, 2013; V. M. Krasnopolsky et al., 2013; Gentine et al., 2018; Brenowitz & Bretherton, 2018), RF-based parameterizations have the advantage of making bounded predictions that respect linear physical constraints (O'Gorman & Dwyer, 2018b). This could explain why RFs tend to be more stable once coupled back to a climate model (Brenowitz, Henn, et al., 2020), although NN parameterizations can be designed to be unconditionally stable (Yuval et al., 2020; Brenowitz, Beucler, et al., 2020) and conserve mass/energy (Beucler, Rasp, et al., 2019). Recent success in emulating subgrid thermodynamics with Earth-like boundary conditions (Han et al. (2020), Mooers, Pritchard, et al. (2020), see Fig 3) are a promising step towards ML-powered long-term climate predictions, while simple proxys of subgrid parameterization such as the Lorenz 96 system (Lorenz, 1996) can help quickly test new ML algorithms (Crommelin & Vanden-Eijnden, 2008; Gagne et al., 2020) and frameworks (Rasp, 2020; Mouatadid et al., 2019).

3.4 Downscaling

Operational meteorological forecasts and climate predictions usually require variables at the local (e.g. 1km) scale, but global atmospheric models usually output variables at a coarser (e.g. 50-200km) scale, are not tuned for regional-scale predictions, and rely on physical variables that may not directly be societally relevant (e.g. momentum fluxes instead of wind bursts, heating fluxes instead of surface solar irradiance). Ground-based observations provide large datasets of such variables, which allows ML algorithms to be trained to predict societally-relevant variables at the local scale (Sharifi et al., 2019) from coarse-scale model output. We note that this is a particular case of statistical down-

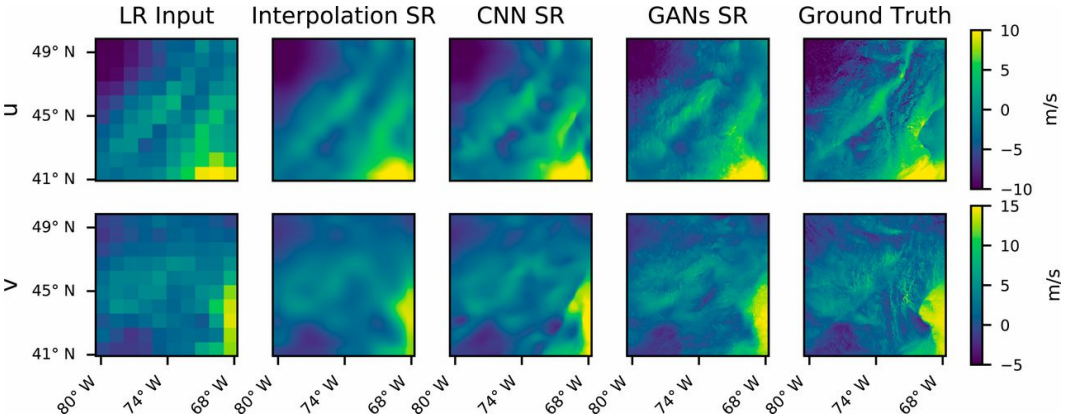

Figure 4. SR of East-West (U) and North-South (V) wind velocity from low-resolution (LR) Input using bilinear interpolation, a CNN, and GANs. Fig 3 from Stengel et al. (2020).

scaling (reviewed in Wilby et al. (1998)), which is a data-driven and computationally-inexpensive alternative to dynamical downscaling where a regional-scale physical model is run with boundary conditions derived from coarse-scale model output (Xue et al., 2014; Feser et al., 2011).

From a ML perspective, downscaling is analogous to super-resolution (SR), which aims to obtain a high-resolution output from its low-resolution version (see review by W. Yang et al. (2019)). We note that SR is an ill-posed problem, as a single low-resolution image corresponds to an infinite number of high-resolution images from an unknown probability distribution. Convolutional neural networks (CNNs), which hierarchically extract data from image patches, have been successfully leveraged to super-resolve precipitation predictions (Vandal et al., 2017, 2019), satellite images (Pouliot et al., 2018), and idealized turbulent flows (Fukami et al., 2018). Despite outperforming bilinear interpolation baselines (Baño-Medina et al., 2020), CNNs typically underestimate extremes (Sachindra et al., 2018) as they tend to predict the average of all possible solutions to minimize the error at each pixel. While Sachindra et al. (2018) found that Relevance Vector Machine (a Bayesian approach to learning probabilistic sparse generalized linear models, Tipping (2000)) improved the SR of precipitation extremes, recent work has focused on generative modeling for SR (Singh et al., 2019), in particular conditional generative adversarial networks (cGANs, Mirza and Osindero (2014)). For instance, Stengel et al. (2020) used a sequence of two cGANs to super-resolve wind and solar power at the 2km-scale using 100km-scale climate model output, and showed that the resulting turbulent kinetic energy spectra and solar irradiance semi-variograms were more consistent with high-resolution climate model output than those generated by interpolation and simple CNNs (Fig 4). Finally, in an effort to represent the full distribution of possible super-resolved temperature and precipitation fields from coarse model output, Groenke et al. (2020) adapted recent work in normalizing flows for variational inference (Grover et al., 2020) to develop an unsupervised NN approach that generates the joint distribution of high and low-resolution climate maps.

3.5 Climate Analysis and Understanding

The successful ML attempts described above at modeling atmospheric processes already improve our understanding of climate dynamics. Taking the example of subgrid-scale thermodynamics parameterization (Rasp et al., 2018; Brenowitz & Bretherton, 2018; Yuval & O'Gorman, 2020), the ability of ML models that are local in time and space to capture the effect of clouds and fine-scale turbulence on large-scale thermodynamics demon-

strates that it is possible to approximately close large-scale moist thermodynamics equations without knowledge of small-scale stochasticity, convective-scale organization, and convective memory.

In addition, modern ML algorithms can be used to find relevant patterns and detect climate signals. Ebert-Uphoff and Deng (2012) used Bayesian networks, a type of graphical model, to detect causal relationships between low-frequency patterns in the atmosphere (see Runge et al. (2019) for a broad overview of causal discovery methods in climate science). Wills et al. (2020) show that signal-to-noise maximal pattern filtering extracts forced climate signals with up to 10 times fewer climate models, even on regional scales. Barnes, Hurrell, et al. (2019) and Barnes et al. (2020) show that XAI applied to a NN tasked with categorizing the year in a forced climate data record, based on inputs of global temperature maps, provides a simple new way to identify time-varying regional indicators of the external forcing, with the possibility of distinguishing e.g. greenhouse gas from aerosol forcing (Labe & Barnes, 2021). Toms et al. (2020) shows how XAI applied to NNs can also update views of El-Nino intrinsic predictability – compared to linear regression, NNs de-emphasize the roles of the Atlantic and Indian Ocean and boost the relevance of the northwest tropical Pacific. Applied to West coast temperature anomalies, the same approach reveals that NNs correctly identify near-coastal surface temperatures as the strongest source of predictability for 60-day forecasts. In contrast, linear regressions overemphasize El-Nino-related surface temperature signals in the tropical ocean that should only dominate at longer lead times. Beyond NNs, Barnes, Samarasinghe, et al. (2019) used probabilistic graph models to hypothesize a two-branch interaction between the Madden-Julian Oscillation and the North Atlantic Oscillation, while Di Capua et al. (2020) combined maximum covariance analysis with causal discovery networks to understand the causal influence of tropical convection on mid-latitude dynamics.

3.6 Uncertainty Quantification

Uncertainties can be divided into four main categories:

1. Observational: due to measurement and representation errors (Janjić et al., 2018);
2. Structural: due to incorrect model structure;
3. Parametric: due to incorrect model parameters;
4. Stochastic: due to internal climate variability or the chaotic nature of the flow (Deser et al., 2012; Lorenz, 1963).

These categories do not intend to be mutually exclusive but rather emphasize separate philosophical issues. For instance, consider the uncertain parameterization of cloud microphysics wherein multiple equivalently plausible equations to represent autoconversion rates are in use despite structural differences (structural uncertainty). Meanwhile, even within a given set of equations, the coefficients and parameters are left to be empirically constrained (parametric uncertainty). This requires confronting observational uncertainty (e.g. indirect radiometric measurement of precipitating cloud droplets from radar backscatter with associated sampling and inverse model uncertainty) as well as stochastic uncertainty (e.g. a perfect model can make counterfactual chaotic trajectories that will eventually diverge from nature necessitating multiple runs per comparison).

While observational, structural, and parametric errors should be reduced as much as possible, stochasticity should be reproduced as well as possible to increase the fidelity of the simulations (Berner et al., 2017). Many simple ML algorithms, such as individual feed-forward NNs, are primarily deterministic such that once trained their predictions do not characterize uncertainties by construction. Ensembles of such NNs, constructed by e.g. using different initial random weights or shuffling the training set, may capture *stochastic* uncertainty. But systematically characterizing *parameteric* uncertainties is at the heart of *data assimilation* (Evensen, 2009; Evensen et al., 1998; Eknes & Evensen,

1997), which adopts Bayesian approaches to infer the posterior distribution of either the state or parameters given some initial prior and observational uncertainty. However, despite characterizing the parameteric uncertainty (Dunbar et al., 2020), these Bayesian approaches do not always address structural uncertainty.

This motivates methods that inherently characterize uncertainties, such as Gaussian Processes (Camps-Valls et al., 2016) that e.g., were recently trained on perturbed climate simulation ensembles to better characterize aerosol forcing uncertainty from observed aerosol optical depth (Watson-Parris, Bellouin, et al., 2020) and cloud droplet number (McCoy et al., 2020). A strategy that has not yet seen many applications in Earth sciences is the use of generative models such as variational autoencoders (Pu et al., 2016) or generative adversarial networks (Radford et al., 2015; Z. Yang et al., 2019) to build intrinsically probabilistic models. These approaches aim at reproducing the underlying distribution as a function of the state and thus can be promising approaches for uncertainty quantification. Finally, dropout layers in NNs (Gal & Ghahramani, 2016) and Bayesian NNs (Khan & Coulibaly, 2006; Bate et al., 1998) extend determistic NNs to represent structural uncertainty, making them promising tools to quantify uncertainty.

3.7 Cloud Detection and Classification

The ability to detect and classify clouds from satellite or ground-based observations has a wide range of important applications, ranging from short-term forecasting of hazardous weather to gaining insights into the climate system by detecting the occurrence of different cloud types. Cloud detection and classification is a problem especially suited to modern ML as it mostly relies on image data for which CNNs are particularly well-suited. Some of the earliest work predates the surge of CNNs, namely J. Lee et al. (1990) and Tian et al. (1999) use fully connected NNs with engineered features to detect different cloud types (stratus, cirrus and cumulus). Wood and Hartmann (2006) use fully connected NNs to detect different types of cloud organization (closed/open cell shallow convection). Muhlbauer et al. (2014) create a climatology of these cloud classes.

Detecting different cloud types and patterns is of particular interest to the climate community since different cloud structures have widely different impacts on the Earth's energy balance (Bony et al. (2015), Chap 14). Mahajan and Fataniya (2019) provide an overview of methods to detect cloud properties from satellite imagery. Marais et al. (2020) and Rasp, Schulz, et al. (2020) try to detect larger cloud patterns from satellite imagery. Marais et al. (2020) use a CNN to classify satellite imagery (MODIS, VIIRS) into categories such as clear-air, closed-stratiform and high-altitude clouds. Rasp, Schulz, et al. (2020) focus on modes of organization in subtropical shallow cumulus clouds, based on cloud pattern classes (sugar, flower, fish and gravel) defined by Stevens et al. (2020). Watson-Parris, Sutherland, et al. (2020) used a CNN to detect pockets of open cells, trained on a small dataset of hand-labeled features, and used this algorithm to investigate their radiative impact.

A common challenge for these applications is the need for a large number of labeled data samples which are often hard to obtain. Solutions include (1) generating labels manually, (2) generating labels from other modalities that may only be available intermittently or in certain locations, (3) using transfer learning to reduce the number of samples needed (see Sec 2.1), and (4) using unsupervised learning. For example, both Marais et al. (2020) and Rasp, Schulz, et al. (2020) developed a labeling interface that allowed scientists to label many thousands of images. Yet even that amount of training labels is barely enough to train large modern CNNs. Thus both groups used CNNs pretrained on a huge number of natural images, fine-tuned on the cloud patterns, thus using both Strategies 1 and 3. An example of Strategy 2 is to use ground-based radar (only available in some locations) to generate convection labels for GOES satellite imagery (Y. Lee et al., 2021), or to use CloudSat data (only available intermittently) to generate cloud

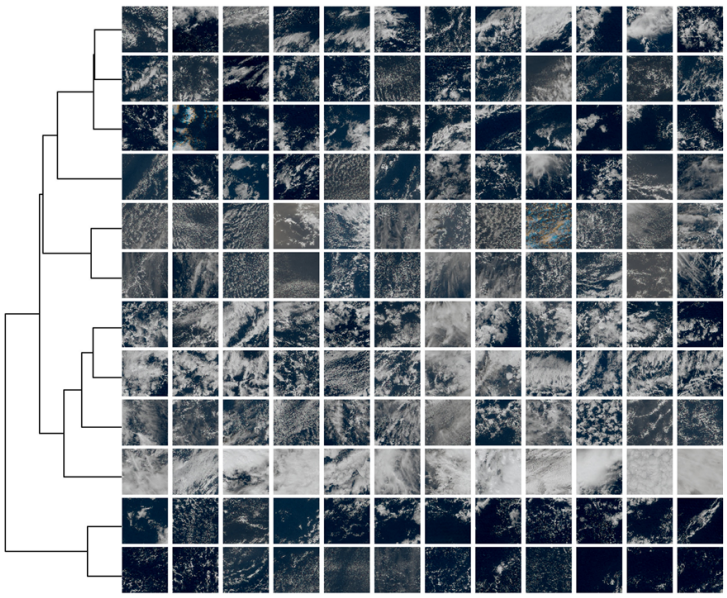

Figure 5. Cloud classification by applying hierarchical clustering to the latent space representation of satellite images in a NN trained with the Tile2Vec method. Each row shows samples from one distinct cloud category identified by the NN. Adapted from Fig 2 of Denby (2020).

type labels for Himawari-8 satellite imagery (C. Zhang et al., 2019). Similarly, (Zantedeschi et al., 2019) used a semi-supervised learning approach to leverage a small number of classified satellite images for creating a much larger labeled dataset to cloud types. Approaches for Strategy 4, unsupervised learning for cloud classification, range from very intuitive to quite abstract. Haynes et al. (2011) use k-means clustering on physical features to categorize satellite imagery into cloud types. Since the clustering is performed in a space with clear physical meaning the resulting cloud categories are easily interpretable. In contrast, recent methods propose to use deep NNs for this purpose. Denby (2020) uses an unsupervised neural network (Tile2Vec architecture by Jean et al. (2019)), then performs clustering in the abstract space of a NN layer. Visualization of samples shows some obvious similarity within each cluster (see Fig. 5); however, it is not yet clear what exactly the obtained cloud categories represent, or whether a network that is trained slightly differently would yield a similar categorization. Kurihana et al. (2019) also use an unsupervised NN approach for cloud classification, namely classifying satellite (MODIS) imagery using a convolutional auto-encoder. In summary, the motivation for NN-based approaches is that NNs might discover yet unknown patterns, but these patterns remain underexplored and the corresponding cloud categories may not be robust.

Other important applications include detecting convection from satellite imagery (X. Zhang et al., 2019; Cintineo et al., 2020; Y. Lee et al., 2021), detecting convective storms from ground-based radar observations (Gooch & Chandrasekar, 2020), detecting cloud types from pictures taken from the surface (J. Zhang et al., 2018), and detecting fronts from numerical weather model output (Lagerquist et al., 2019, 2020). Finally, Liu et al. (2016), Racah et al. (2017) and Kashinath et al. (2021) describe an effort to build a database of expert-labeled tropical cyclones and atmospheric rivers in climate model output, using a custom labeling interface. This is especially relevant to estimate how the frequency of these extreme weather events will change under global warming.

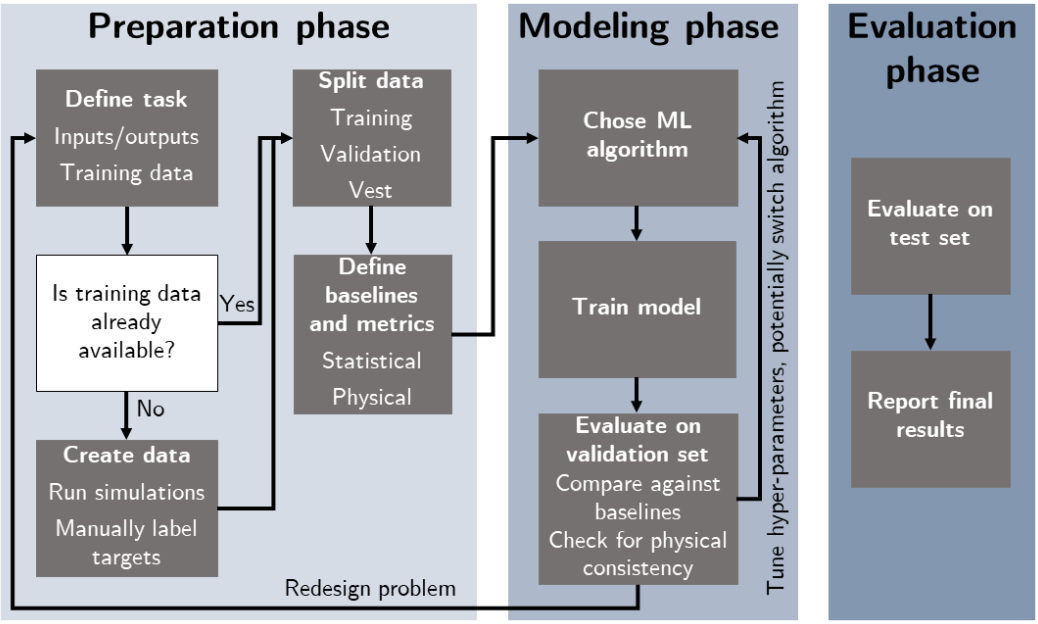

Figure 6. Workflow for a machine learning project in physical sciences.

4 Getting Started with Machine Learning

4.1 A Workflow for Scientific Machine Learning

The first step in a scientific ML workflow (Fig 6) is clearly defining a reproducible task to be solved by the ML algorithm and the data to be used for training. In some cases the data might already be available but in many other cases creating the training data is the first major challenge. For the use case of training a ML subgrid parameterization discussed in section 3.3, reference model simulations need to be run, often with different output requirements than traditional model simulations in terms of variables and spatiotemporal frequency. If the parameterization is to be trained on high-resolution simulations, an additional coarse-graining step has to be performed to provide the training dataset. For cloud detection and classification, as described in section 3.7, high-quality labels of the objects to be classified are often not readily available. Here, traditional algorithms can be used to create labels (Racah et al., 2017). However, as the motivation for using ML is to outperform traditional hand-crafted algorithms, crowd-sourcing has turned out to be a powerful tool to collect labels from domain experts (Kashinath et al., 2021; Rasp, Schulz, et al., 2020). The final task in creating a good dataset is to set aside representative and independent validation/test sets that can be used to monitor the performance of various ML algorithms on previously unseen data.

With the data available, the next step is training a ML model that is appropriate for the task. Before doing so, however, it is paramount to have solid baselines to compare to. These baselines can be traditional non-ML techniques or very simple linear ML algorithms such as linear or logistic regressions. Only with baselines and comparable evaluation is it possible to judge how well more complex algorithms are performing. This makes benchmark problems such as in WeatherBench (Rasp, Dueben, et al., 2020) crucial for the advancement of ML. With baselines set, the question turns to which ML algorithm to use. If possible, it is helpful to search for similar problems previously solved with ML (Reichstein et al., 2019). Generally, NNs and RFs are the most common modern ML techniques. For spatially-structured data, CNNs are often the algorithms of choice.

For time-series problems with memory, recurrent NNs are commonly used. Another important factor to consider when choosing an algorithm is the number of samples, as it will determine the complexity of the algorithm one can use as discussed in Sec 2.

Finally, it is important to visualize and analyze the models' predictions thoroughly and check for unexpected behavior that might not be visible from a simple validation score. Feature importance and other interpretability methods can help understand the inner workings of the algorithms and point towards potential shortcomings and ways to improve (Sec 2.2). If one detects physically unrealistic behavior, physics-guided ML methods to incorporate various degrees of physical structure (Sec 2.3) can help, especially when data are limited or when the algorithm will be applied in unfamiliar climate regimes.

Combined, all of these steps in a ML project can take several months with lots of trial and error. For this reason, it is paramount to make this workflow reproducible by, for example, using a version control system like Git and creating frequent checkpoints.

4.2 Machine Learning Software Ecosystem and Resources for Learning

The ML ecosystem changes significantly more rapidly than many other fields. For this reason, the software and learning resources listed below are also only a snapshot taken at the time of writing and might change in the future.

`Python` is by far the most popular programming language for modern machine learning and we will focus on Python for the remainder of this section. However, there are some alternatives: `R` still has a strong standing for statistical applications, e.g. post-processing, and is still being updated with the latest ML developments. `Matlab`, an old favorite, has also recently added support for deep learning. Finally, `Julia` is a newcomer with strong core support in the atmospheric science community and growing ML capabilities. In Python, `scikit-learn` is a well-established library that has implemented a huge number of supervised and unsupervised learning algorithms, such as linear regression, RFs, various clustering methods all the way up to simple NNs. For deep learning the two most popular choices in Python are `Tensorflow` with its easy-to-use wrapper `Keras`, and `Pytorch`. Most deep learning algorithms found in literature will have code in one of these two `Python` libraries available.

As for learning resources, there are a great number of books and courses available. For books, Géron (2019) is a great starting choice. So is Chollet (2017), which specifically focuses on deep learning. For courses, `deeplearning.ai` hosts several well-produced courses on `Coursera`, and we additionally recommend the deep learning courses of `fast.ai`.

5 Outlook

Exploration of ML for clouds and climate is still in its infancy with many underexplored research frontiers spanning observations, modeling and understanding.

For observations, the main challenge remains the lack of samples in the observational record: While the *volume* of observed data increases drastically from year to year, the number of observed events can still be very small for many applications (e.g., extreme events such as cyclones, atmospheric rivers, etc.) and often there are no labels. Traditional ML techniques to mitigate these issues, such as data augmentation (image translation, rotation, and mirroring to create more samples, Yu et al. (2017)), may only be valid for some two-dimensional data so long as the Earth's spherical geometry can be respected (Weyn et al., 2020). For that reason, it is promising to use transfer learning to fine-tune a ML model on the (potentially) sparse observational data of interest after training it on model data that can be generated at will (Sec 2.1), or meta-learning to efficiently adjust the ML algorithm's hyper-parameters with only a few samples (Finn et al., 2017; Rußwurm et al., 2020). Finally, the mathematical simplicity of (fitted) ML mod-

els compared to dynamical models makes them easier to integrate within data assimilation frameworks (Brajard et al., 2020), opening the door to ML-powered bias correction of operational atmospheric models (Bonavita & Laloyaux, 2020).

For modeling, ML emulations of individual processes such as clouds are flourishing in climate models and can replace known equations that are hard to discretize (Bar-Sinai et al., 2019) or fit process-resolving simulations when the equations are unknown as illustrated by the microphysics (Sec 3.2) and subgrid parameterization (Sec 3.3) cases. This could readily be extended to parameterize subgrid processes at progressively smaller scales as global climate model resolution improves. Despite recent progress, it should be noted that ML emulation for climate modeling faces challenges that are not fully addressed, such as stability when coupled back to the climate model and generalization to different climates (Sec 2.3). A related issue is the quantification of uncertainties with probabilistic or stochastic ML techniques (Sec 3.6). Attempts to date have also not explored the potential of multi-node GPU/TPU based-high performance computing for ML at scale due to software infrastructure challenges (Ben-Nun & Hoefler, 2019).

Finally, leveraging modern ML tools to improve our understanding of the climate from large datasets remains mostly untapped. While recent data-driven equation discovery tools (Long et al., 2019; S. Zhang & Lin, 2018) show promising preliminary results in physics (Brunton et al., 2016; Rudy et al., 2017) and oceanography (Zanna & Bolton, 2020) in simple settings, partial differential equation discovery tools have not yet been applied to cloud processes and climate modeling in Earth-like settings. XAI helps translate successes in emulation into improved climate understanding (Sec 2.2), but it does not extract causes of particular phenomena, motivating causal research to improve our theoretical understanding of the climate system (Runge et al., 2019). Causal inference (e.g. Granger causality) and causal discovery (e.g. causal effect networks) methods have been successfully applied to establish a feedback of ocean surface temperatures on the North-Atlantic Oscillation (Mosedale et al., 2006) and analyze Arctic drivers of midlatitude winter circulation (Kretschmer et al., 2016), but they remain under-explored in the analysis of convection and cloud-related climactic processes. A recent exception is Hirt et al. (2020) who used linear causal graph analysis to show that low model resolution decreased the frequency of convective initiation by reducing upward mass flux at gust fronts. Finally, as generative modeling can combine dimensionality reduction with prediction, further exploring generated latent spaces may reveal new sources of predictability in observational and simulation data (Mooers, Tuyls, et al., 2020).

Acknowledgments
TB and MP acknowledge funding from NSF grants OAC1835863 and AGS-1734164. MP acknowledges partial funding from the DOE "Enabling Aerosol-cloud interactions at GLobal convection-permitting scalES (EAGLES)" project (74358), funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Earth System Model Development (ESMD) program. Additionally, this material is based upon work supported by the National Science Foundation under Grants ICER-2019758 (IE) and OAC-1934668 (IE). PG thanks NSF grant AGS-1734164 as well as USMILE European Research Council grant. We thank Griffin Mooers for Fig 3 and Elizabeth Barnes, Alexei Belochitski, Derek Chang, Vladimir Krasnopolsky, Jeremy McGibbon, Sylvia Sullivan, Duncan Watson-Parris, and two anonymous reviewers for helpful advice that greatly improved the content and clarity of the manuscript.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Advances in neural information processing systems* (pp. 9505–9515).

Baño-Medina, J. L., García Manzanas, R., Gutiérrez Llorente, J. M., et al. (2020). Configuration and intercomparison of deep learning neural models for statistical downscaling.

Bansal, N., Agarwal, C., & Nguyen, A. (2020). Sam: The sensitivity of attribution methods to hyperparameters. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8673–8683).

Barnes, E. A., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2019). Viewing forced climate patterns through an ai lens. *Geophysical Research Letters*, *46*(22), 13389–13398.

Barnes, E. A., Samarasinghe, S. M., Ebert-Uphoff, I., & Furtado, J. C. (2019, August). Tropospheric and stratospheric causal pathways between the MJO and NAO. *J. Geophys. Res. D: Atmos.*, *124*(16), 9356–9371.

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020). Indicator patterns of forced change learned by an artificial neural network. *Journal of Advances in Modeling Earth Systems (JAMES)*. doi: 10.1029/2020MS002195

Bar-Sinai, Y., Hoyer, S., Hickey, J., & Brenner, M. P. (2019). Learning data-driven discretizations for partial differential equations. *Proceedings of the National Academy of Sciences*, *116*(31), 15344–15349.

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., & De Freitas, R. M. (1998). A bayesian neural network method for adverse drug reaction signal generation. *European journal of clinical pharmacology*, *54*(4), 315–321.

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6541–6549).

Belochitski, A., Binev, P., DeVore, R., Fox-Rabinovitz, M., Krasnopolsky, V., & Lamby, P. (2011). Tree approximation of the long wave radiation parameterization in the ncar cam global climate model. *Journal of Computational and Applied Mathematics*, *236*(4), 447–460.

Ben-Nun, T., & Hoefler, T. (2019, August). Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. *ACM Comput. Surv.*, *52*(4), 1–43.

Bergen, K. J., Johnson, P. A., Maarten, V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid earth geoscience. *Science*, *363*(6433).

Berner, J., Achatz, U., Batte, L., Bengtsson, L., Cámara, A. d. l., Christensen, H. M., ... others (2017). Stochastic parameterization: Toward a new view of weather and climate models. *Bulletin of the American Meteorological Society*, *98*(3), 565–588.

Beucler, T., Pritchard, M., Gentine, P., & Rasp, S. (2020). Towards physically-consistent, data-driven models of convection. *arXiv preprint arXiv:2002.08525*.

Beucler, T., Pritchard, M., Rasp, S., Gentine, P., Ott, J., & Baldi, P. (2019). Enforcing analytic constraints in neural-networks emulating physical systems. *arXiv preprint arXiv:1909.00912*.

Beucler, T., Rasp, S., Pritchard, M., & Gentine, P. (2019). Achieving conservation of energy in neural network emulators for climate modeling. *arXiv preprint arXiv:1906.06622*.

Bonavita, M., & Laloyaux, P. (2020). Machine learning for model error inference and correction. *Earth and Space Science Open Archive*, 36. doi: 10.1002/essoar.10503695.1

Bony, S., Stevens, B., Frierson, D. M., Jakob, C., Kageyama, M., Pincus, R., ... others (2015). Clouds, circulation and climate sensitivity. *Nature Geoscience*, *8*(4), 261–268.

Brajard, J., Carassi, A., Bocquet, M., & Bertino, L. (2020). Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the lorenz 96 model. *arXiv preprint arXiv:2001.01520*.

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *arXiv preprint arXiv:2003.06549*.

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298.

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2728–2744.

Brenowitz, N. D., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., ... Bretherton, C. S. (2020). Machine learning climate model dynamics: Offline versus online performance. *arXiv preprint arXiv:2011.03081*.

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, *113*(15), 3932–3937.

Camps-Valls, G., Martino, L., Svendsen, D. H., Campos-Taberner, M., Muñoz-Marí, J., Laparra, V., ... García-Haro, F. J. (2018). Physics-aware gaussian processes in remote sensing. *Applied Soft Computing*, *68*, 69–82.

Camps-Valls, G., Verrelst, J., Munoz-Mari, J., Laparra, V., Mateo-Jiménez, F., & Gómez-Dans, J. (2016). A survey on gaussian processes for earth-observation data analysis: A comprehensive investigation. *IEEE Geoscience and Remote Sensing Magazine*, *4*(2), 58–78.

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., & Olah, C. (2019). *Exploring neural networks with activation atlases.* Distill.

Chevallier, F., Morcrette, J.-J., Chéruy, F., & Scott, N. (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.

Chollet, F. (2017). *Deep learning with python.* Manning Publications Company.

Cintineo, J. L., Pavolonis, M. J., Sieglaff, J. M., Wimmers, A., Brunner, J., & Bellon, W. (2020). A deep-learning model for automated detection of intense mid-latitude convection using geostationary satellite images. *Weather and Forecasting*, 1–57.

Crommelin, D., & Vanden-Eijnden, E. (2008). Subgrid-scale parameterization with conditional markov chains. *Journal of the Atmospheric Sciences*, *65*(8), 2661–2675.

de Bezenac, E., Pajot, A., & Gallinari, P. (2019). Deep learning for physical processes: Incorporating prior scientific knowledge. *Journal of Statistical Mechanics: Theory and Experiment*, *2019*(12), 124009.

Denby, L. (2020). Discovering the importance of mesoscale cloud organization through unsupervised classification. *Geophysical Research Letters*, *47*(1), e2019GL085190.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: the role of internal variability. *Climate dynamics*, *38*(3-4), 527–546.

Di Capua, G., Runge, J., Donner, R. V., van den Hurk, B., Turner, A. G., Vellore, R., ... Coumou, D. (2020). Dominant patterns of interaction between the tropics and mid-latitudes in boreal summer: causal relationships and the role

of timescales. *Weather and Climate Dynamics*, *1*(2), 519–539.

Dunbar, O. R., Garbuno-Inigo, A., Schneider, T., & Stuart, A. M. (2020). Calibration and uncertainty quantification of convective parameters in an idealized gcm. *arXiv preprint arXiv:2012.13262*.

Ebert-Uphoff, I., & Deng, Y. (2012). Causal discovery for climate research using graphical models. *Journal of Climate*, *25*(17), 5648–5665.

Ebert-Uphoff, I., & Hilburn, K. A. (2020). Evaluation, tuning and interpretation of neural networks for working with images in meteorological applications. *Bulletin of the American Meteorological Society (BAMS)*.

Eknes, M., & Evensen, G. (1997). Parameter estimation solving a weak constraint variational formulation for an ekman model. *Journal of Geophysical Research: Oceans*, *102*(C6), 12479–12491.

Evensen, G. (2009). The ensemble kalman filter for combined state and parameter estimation. *IEEE Control Systems Magazine*, *29*(3), 83–104.

Evensen, G., Dee, D. P., & Schröter, J. (1998). Parameter estimation in dynamical models. In *Ocean modeling and parameterization* (pp. 373–398). Springer.

Feser, F., Rockel, B., von Storch, H., Winterfeldt, J., & Zahn, M. (2011). Regional climate models add value to global model data: a review and selected examples. *Bulletin of the American Meteorological Society*, *92*(9), 1181–1192.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*.

Fong, R., & Vedaldi, A. (2019). Explanations for attributing deep neural network predictions. In *Explainable ai: Interpreting, explaining and visualizing deep learning* (pp. 149–167). Springer.

Foster, D. (2019). *Generative deep learning: teaching machines to paint, write, compose, and play*. O'Reilly Media.

Fukami, K., Fukagata, K., & Taira, K. (2018). Super-resolution reconstruction of turbulent flows with machine learning. *arXiv preprint arXiv:1811.11328*.

Gagne, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine learning for stochastic parameterization: Generative adversarial networks in the lorenz'96 model. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001896.

Gagne, D. J., McGovern, A., Haupt, S. E., Sobash, R. A., Williams, J. K., & Xue, M. (2017). Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Weather and forecasting*, *32*(5), 1819–1840.

Gagne II, D. J., Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, *147*(8), 2827–2845.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

Gala, Y., Fernández, Á., Díaz, J., & Dorronsoro, J. R. (2016). Hybrid machine learning forecasting of solar radiation values. *Neurocomputing*, *176*, 48–59.

Gao, H., Sun, L., & Wang, J.-X. (2020). Phygeonet: Physics-informed geometry-adaptive convolutional neural networks for solving parametric pdes on irregular domain. *arXiv preprint arXiv:2004.13145*.

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

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Gettelman, A., Gagne, D. J., Chen, C.-C., Christensen, M., Lebo, Z., Morrison, H., & Gantos, G. (2020). Machine learning the warm rain process.

Gooch, S. R., & Chandrasekar, V. (2020). Improving historical data discovery in weather radar image data sets using transfer learning. *IEEE Transactions on Geoscience and Remote Sensing*.

Gristey, J. J., Feingold, G., Glenn, I. B., Schmidt, K. S., & Chen, H. (2020). On the relationship between shallow cumulus cloud field properties and surface solar irradiance. *Geophysical Research Letters*, *47*(22), e2020GL090152.

Groenke, B., Madaus, L., & Monteleoni, C. (2020). Climalign: Unsupervised statistical downscaling of climate variables via normalizing flows. In *Proceedings of the 10th International Conference on Climate Informatics (CI 2020)*.

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2020). Deep learning for post-processing ensemble weather forecasts. *arXiv preprint arXiv:2005.08748*.

Grover, A., Chute, C., Shu, R., Cao, Z., & Ermon, S. (2020). Alignflow: Cycle consistent learning from multiple domains via normalizing flows. In *Aaai* (pp. 4028–4035).

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year enso forecasts. *Nature*, *573*(7775), 568–572.

Han, Y., Zhang, G. J., Huang, X., & Wang, Y. (2020). A moist physics parameterization based on deep learning. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002076.

Haynes, J. M., Jakob, C., Rossow, W. B., Tselioudis, G., & Brown, J. (2011). Major characteristics of southern ocean cloud regimes and their effects on the energy budget. *Journal of Climate*, *24*(19), 5061–5080.

Hirt, M., Craig, G. C., Schäfer, S. A., Savre, J., & Heinze, R. (2020). Cold-pool-driven convective initiation: using causal graph analysis to determine what convection-permitting models are missing. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 2205–2227.

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J. (2021). Will artificial intelligence supersede earth system and climate models? *arXiv preprint arXiv:2101.09126*.

Janjić, T., Bormann, N., Bocquet, M., Carton, J., Cohn, S., Dance, S., ... others (2018). On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *144*(713), 1257–1278.

Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., & Ermon, S. (2019). Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 3967–3974).

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, *31*(8), 1544–1554.

Karpatne, A., Watkins, W., Read, J., & Kumar, V. (2017). Physics-guided neural networks (pgnn): An application in lake temperature modeling. *arXiv preprint arXiv:1710.11431*.

Kashinath, K., Mudigonda, M., Kim, S., Kapp-Schwoerer, L., Graubner, A., Karaismailoglu, E., ... others (2021). Climatenet: an expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development*, *14*(1), 107–124.

Khan, M. S., & Coulibaly, P. (2006). Bayesian neural network for rainfall-runoff modeling. *Water Resources Research*, *42*(7).

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., ... Kim, B. (2017). The (un) reliability of saliency methods. *arXiv preprint arXiv:1711.00867*.

Krasnopolsky, V., Fox-Rabinovitz, M., Hou, Y., Lord, S., & Belochitski, A. (2010). Accurate and fast neural network emulations of model radiation for the ncep coupled climate forecast system: climate simulations and seasonal predictions. *Monthly Weather Review*, *138*(5), 1822–1842.

Krasnopolsky, V. M. (2013). The application of neural networks in the earth system sciences. neural network emulations for complex multidimensional mappings. *Atmospheric and Oceanic Science Library*, *46*.

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

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2008). Decadal climate simulations using accurate and fast neural network emulation of full, longwave and shortwave, radiation. *Monthly Weather Review*, *136*(10), 3683–3695.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2013). Using ensemble of neural networks to learn stochastic convection parameterizations for climate and numerical weather prediction models from data simulated by a cloud resolving model. *Advances in Artificial Neural Systems*, *2013*.

Kretschmer, M., Coumou, D., Donges, J. F., & Runge, J. (2016). Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. *Journal of Climate*, *29*(11), 4069–4081.

Kurihana, T., Foster, I., Willett, R., Jenkins, S., Koenig, K., Werman, R., ... Moyer, E. (2019). Cloud classification with unsupervised learning. In *9th international workshop on climate informatics (CI2019)*.

Labe, Z. M., & Barnes, E. A. (2021). Detecting climate signals using explainable ai with single-forcing large ensembles. *Earth and Space Science Open Archive ESSOAr*.

Lagerquist, R., McGovern, A., & Gagne II, D. J. (2019). Deep learning for spatially explicit prediction of synoptic-scale fronts. *Weather and Forecasting*, *34*(4), 1137–1160.

Lagerquist, R., McGovern, A., Homeyer, C. R., Gagne, D. J., & Smith, T. (2020). Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Monthly Weather Review*.

Lee, J., Weger, R. C., Sengupta, S. K., & Welch, R. M. (1990). A neural network approach to cloud classification. *IEEE Transactions on Geoscience and Remote Sensing*, *28*(5), 846–855.

Lee, Y., Kummerow, C. D., & Ebert-Uphoff, I. (2021). Applying machine learning methods to detect convection using goes-16 abi data. *Submitted to Atmospheric Measurement Techniques (in review)*.

Ling, J., Kurzawski, A., & Templeton, J. (2016). Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics*, *807*, 155–166.

Liu, Y., Racah, E., Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., ... others (2016). Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv preprint arXiv:1605.01156*.

Loeb, D. E. (1991). Series with general exponents. *Journal of mathematical analysis and applications*, *156*(1), 184–208.

Long, Z., Lu, Y., & Dong, B. (2019). Pde-net 2.0: Learning pdes from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics*, *399*, 108925.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, *20*(2), 130–141.

Lorenz, E. N. (1996). Predictability: A problem partly solved. In *Proc. seminar on predictability* (Vol. 1).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765–4774).

Mahajan, S., & Fataniya, B. (2019). Cloud detection methodologies: Variants and development—a review. *Complex & Intelligent Systems*, 1–11.

Mao, Z., Jagtap, A. D., & Karniadakis, G. E. (2020). Physics-informed neural networks for high-speed flows. *Computer Methods in Applied Mechanics and Engineering*, *360*, 112789.

Marais, W. J., Holz, R. E., Reid, J. S., & Willett, R. M. (2020). Leveraging spatial textures, through machine learning, to identify aerosols and distinct cloud types from multispectral observations. *Atmospheric Measurement Techniques*, *13*(10), 5459–5480.

Marmanis, D., Datcu, M., Esch, T., & Stilla, U. (2015). Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, *13*(1), 105–109.

McCoy, I. L., McCoy, D. T., Wood, R., Regayre, L., Watson-Parris, D., Grosvenor, D. P., ... others (2020). The hemispheric contrast in cloud microphysical properties constrains aerosol forcing. *Proceedings of the National Academy of Sciences*, *117*(32), 18998–19006.

McGibbon, J., & Bretherton, C. S. (2019). Single-column emulation of reanalysis of the northeast pacific marine boundary layer. *Geophysical Research Letters*, *46*(16), 10053–10060.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., ... Williams, J. K. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, *98*(10), 2073–2090.

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199.

McGovern, A., & Lagerquist, R. A. (2020). Using machine learning and model interpretation and visualization techniques to gain physical insights in atmospheric science. In *International conference on learning representations (ICLR 2020), AI for earth sciences workshop.*

McGovern, A., Lagerquist, R. A., & Gagne, D. J. I. (2020). Using machine learning and model interpretation and visualization techniques to gain physical insights in atmospheric science. In *International conference on learning representation (iclr2020).*

Min, M., Li, J., Wang, F., Liu, Z., & Menzel, W. P. (2020). Retrieval of cloud top properties from advanced geostationary satellite imager measurements based on machine learning algorithms. *Remote Sensing of Environment*, *239*, 111616.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Molnar, C. (2019). Interpretable machine learning. *Lulu. com*.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, *65*, 211–222.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

Mooers, G., Pritchard, M., Beucler, T., Ott, J., Yacalis, G., Baldi, P., & Gentine, P. (2020). Assessing the potential of deep learning for emulating cloud superparameterization in climate models with real-geography boundary conditions. *arXiv preprint arXiv:2010.12996*.

Mooers, G., Tuyls, J., Mandt, S., Pritchard, M., & Beucler, T. (2020). Generative modeling for atmospheric convection. *arXiv preprint arXiv:2007.01444*.

Morrison, H., van Lier-Walqui, M., Fridlind, A. M., Grabowski, W. W., Harrington, J. Y., Hoose, C., ... others (2020). Confronting the challenge of modeling cloud and precipitation microphysics. *Journal of Advances in Modeling Earth Systems*, e2019MS001689.

Morrison, H., van Lier-Walqui, M., Kumjian, M. R., & Prat, O. P. (2020). A bayesian approach for statistical–physical bulk parameterization of rain microphysics. part i: Scheme description. *Journal of the Atmospheric Sciences*, *77*(3), 1019–1041.

Mosedale, T. J., Stephenson, D. B., Collins, M., & Mills, T. C. (2006). Granger causality of coupled climate processes: Ocean feedback on the north atlantic oscillation. *Journal of climate*, *19*(7), 1182–1194.

Mouatadid, S., Gentine, P., Yu, W., & Easterbrook, S. (2019). Recovering the parameters underlying the lorenz-96 chaotic dynamics. *arXiv preprint arXiv:1906.06786*.

Muhlbauer, A., McCoy, I. L., & Wood, R. (2014, 7). Climatology of stratocumulus cloud morphologies: microphysical properties and radiative effects. *Atmospheric Chemistry and Physics*, *14*(13), 6695–6716. Retrieved from `https://www.atmos-chem-phys.net/14/6695/2014/` doi: 10.5194/acp-14-6695-2014

O'Gorman, P. A., & Dwyer, J. G. (2018a). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563.

O'Gorman, P. A., & Dwyer, J. G. (2018b). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. *Distill*, *3*(3), e10.

Pal, A., Mahajan, S., & Norman, M. R. (2019). Using deep neural networks as cost-effective surrogate models for super-parameterized e3sm radiative transfer. *Geophysical Research Letters*, *46*(11), 6069–6079.

Pang, G., Lu, L., & Karniadakis, G. E. (2019). fpinns: Fractional physics-informed neural networks. *SIAM Journal on Scientific Computing*, *41*(4), A2603–A2626.

Pouliot, D., Latifovic, R., Pasher, J., & Duffe, J. (2018). Landsat super-resolution enhancement using convolution neural networks and sentinel-2 for training. *Remote Sensing*, *10*(3), 394.

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., & Carin, L. (2016). Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems* (pp. 2352–2360).

Racah, E., Beckham, C., Maharaj, T., Kahou, S. E., Prabhat, M., & Pal, C. (2017). Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. In *Advances in neural information processing systems* (pp. 3402–3413).

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Raghu, M., & Schmidt, E. (2020). A survey of deep learning for scientific discovery. *arXiv preprint arXiv:2003.11755*.

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2017). Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*.

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707.

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and interpretable models in computer vision and machine learning* (pp. 19–36). Springer.

Rasp, S. (2020). Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: general algorithms and lorenz 96 case study (v1. 0). *Geoscientific Model Development*, *13*(5), 2185–2196.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). Weatherbench: A benchmark dataset for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, e2020MS002203. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002203` (e2020MS002203 2020MS002203) doi: https://doi.org/10.1029/2020MS002203

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689.

Rasp, S., Schulz, H., Bony, S., & Stevens, B. (2020). Combining crowd-sourcing and deep learning to explore the meso-scale organization of shallow convection. *Bulletin of the American Meteorological Society*.

Rasp, S., & Thuerey, N. (2020). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *arXiv preprint arXiv:2008.08626*.

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

Rivera, J. P., Verrelst, J., Gómez-Dans, J., Muñoz-Marí, J., Moreno, J., & Camps-Valls, G. (2015). An emulator toolbox to approximate radiative transfer models with statistical learning. *Remote Sensing*, *7*(7), 9347–9370.

Roh, S., & Song, H.-J. (2020). Evaluation of neural network emulations for radiation parameterization in cloud resolving model. *Geophysical Research Letters*, *47*(21), e2020GL089444.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., ... others (2019). Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*.

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.

Rudy, S. H., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2017). Data-driven discovery of partial differential equations. *Science Advances*, *3*(4), e1602614.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., ... others (2019). Inferring causation from time series in earth system sciences. *Nature communications*, *10*(1), 1–13.

Rußwurm, M., Wang, S., Korner, M., & Lobell, D. (2020). Meta-learning for few-shot land cover classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops* (pp. 200–201).

Rutjes, H., Willemsen, M., & IJsselsteijn, W. (2019). Considerations on explainable ai and users' mental models. In *Chi 2019 workshop: Where is the human? bridging the gap between ai and hci.*

Sachindra, D., Ahmed, K., Rashid, M. M., Shahid, S., & Perera, B. (2018). Statistical downscaling of precipitation using machine learning techniques. *Atmospheric research*, *212*, 240–258.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable ai: interpreting, explaining and visualizing deep learning* (Vol. 11700). Springer Nature.

Scarselli, F., & Tsoi, A. C. (1998). Universal approximation using feedforward neu-

ral networks: A survey of some existing methods, and some new results. *Neural networks*, *11*(1), 15–37.

Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, *44*(24), 12–396.

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

Seifert, A., & Rasp, S. (2020). Potential and limitations of machine learning for modeling warm-rain cloud microphysical processes. *Journal of Advances in Modeling Earth Systems*, *Accepted*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the ieee international conference on computer vision* (pp. 618–626).

Sharifi, E., Saghafian, B., & Steinacker, R. (2019). Downscaling satellite precipitation estimates with multiple linear regression, artificial neural networks, and spline interpolation techniques. *Journal of Geophysical Research: Atmospheres*, *124*(2), 789–805.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*.

Siebesma, A. P., Soares, P. M., & Teixeira, J. (2007). A combined eddy-diffusivity mass-flux approach for the convective boundary layer. *Journal of the atmospheric sciences*, *64*(4), 1230–1248.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Singh, A., White, B. L., & Albert, A. (2019). Downscaling numerical weather models with gans. In *Agu fall meeting 2019.*

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., ... Kalchbrenner, N. (2020). Metnet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*.

Stengel, K., Glaws, A., Hettinger, D., & King, R. N. (2020). Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences*, *117*(29), 16805–16815.

Stevens, B., Bony, S., Brogniez, H., Hentgen, L., Hohenegger, C., Kiemle, C., ... Zuidema, P. (2020, 1). Sugar, gravel, fish and flowers: Mesoscale cloud patterns in the trade winds. *Quarterly Journal of the Royal Meteorological Society*, *146*(726), 141–152. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3662` doi: 10.1002/qj.3662

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., ... others (2019). Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, *6*(1), 1–17.

Sun, L., Gao, H., Pan, S., & Wang, J.-X. (2020). Surrogate modeling for fluid flows based on physics-constrained deep learning without simulation data. *Computer Methods in Applied Mechanics and Engineering*, *361*, 112732.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*.

Tian, B., Shaikh, M. A., Azimi-Sadjadi, M. R., Haar, T. H. V., & Reinke, D. L. (1999). A study of cloud classification with neural networks using spectral and textural features. *IEEE transactions on neural networks*, *10*(1), 138–151.

Tipping, M. E. (2000). The relevance vector machine. In *Advances in neural infor-*

mation processing systems (pp. 652–658).

Toms, B. A., Kashinath, K., Yang, D., et al. (2020). Testing the reliability of interpretable neural networks in geoscience using the madden-julian oscillation. *Geoscientific Model Development Discussions*, 1–22.

Ukkonen, P., Pincus, R., Hogan, R. J., Pagh Nielsen, K., & Kaas, E. (2020). Accelerating radiation computations for dynamical models with targeted machine learning and code optimization. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002226.

Vandal, T., Kodra, E., & Ganguly, A. R. (2019). Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. *Theoretical and Applied Climatology*, *137*(1-2), 557–570.

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., & Ganguly, A. R. (2017). Deepsd: Generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1663–1672).

Veerman, M. A., Pincus, R., Stoffer, R., van Leeuwen, C., Podareanu, D., & van Heerwaarden, C. C. (2020). Predicting atmospheric optical properties for radiative transfer computations using neural networks. *arXiv preprint arXiv:2005.02265*.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, *105*, 569–582.

Wang, J.-X., Wu, J.-L., & Xiao, H. (2017). Physics-informed machine learning approach for reconstructing reynolds stress modeling discrepancies based on dns data. *Physical Review Fluids*, *2*(3), 034603.

Watson-Parris, D. (2020). Machine learning for weather and climate are worlds apart. *arXiv preprint arXiv:2008.10679*.

Watson-Parris, D., Bellouin, N., Deaconu, L., Schutgens, N. A., Yoshioka, M., Regayre, L. A., ... others (2020). Constraining uncertainty in aerosol direct forcing. *Geophysical Research Letters*, *47*(9), e2020GL087141.

Watson-Parris, D., Sutherland, S., Christensen, M., & Stier, P. (2020). A large-scale analysis of pockets of open cells and their radiative impact.

Watt-Meyer, O., Brenowitz, N., Bretherton, C. S., Clark, S., Henn, B. M., Kwa, A., ... Harris, L. (2020). Correcting weather models by learning nudging tendencies from hindcast simulations. In *Agu fall meeting 2020.*

Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *arXiv preprint arXiv:2003.11927*.

Wilby, R. L., Wigley, T., Conway, D., Jones, P., Hewitson, B., Main, J., & Wilks, D. (1998). Statistical downscaling of general circulation model output: A comparison of methods. *Water resources research*, *34*(11), 2995–3008.

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2020). Integrating physics-based modeling with machine learning: A survey. *arXiv preprint arXiv:2003.04919*.

Wills, R. C. J., Battisti, D. S., Armour, K. C., Schneider, T., & Deser, C. (2020, October). Pattern recognition methods to separate forced responses from internal variability in climate model ensembles and observations. *J. Clim.*, *33*(20), 8693–8719.

Wood, R., & Hartmann, D. L. (2006, 5). Spatial Variability of Liquid Water Path in Marine Low Cloud: The Importance of Mesoscale Cellular Convection. *Journal of Climate*, *19*(9), 1748–1764. Retrieved from `http://journals.ametsoc.org/doi/abs/10.1175/JCLI3702.1` doi: 10.1175/JCLI3702.1

Wu, J.-L., Kashinath, K., Albert, A., Chirila, D., Xiao, H., et al. (2020). Enforcing statistical constraints in generative adversarial networks for modeling chaotic dynamical systems. *Journal of Computational Physics*, *406*, 109209.

Xue, Y., Janjic, Z., Dudhia, J., Vasic, R., & De Sales, F. (2014). A review on regional dynamical downscaling in intraseasonal to seasonal simulation/prediction and major factors that affect downscaling ability. *Atmospheric research*, *147*, 68–85.

Yang, T., Sun, F., Gentine, P., Liu, W., Wang, H., Yin, J., ... Liu, C. (2019). Evaluation and machine learning improvement of global hydrological model-based flood simulations. *Environmental Research Letters*, *14*(11), 114027.

Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.-H., & Liao, Q. (2019). Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, *21*(12), 3106–3121.

Yang, Z., Wu, J.-L., & Xiao, H. (2019). Enforcing deterministic constraints on generative adversarial networks for emulating physical systems. *arXiv preprint arXiv:1911.06671*.

Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of physics: Conference series* (Vol. 1168, p. 022022).

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Yu, X., Wu, X., Luo, C., & Ren, P. (2017). Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. *GIScience & Remote Sensing*, *54*(5), 741–758.

Yuval, J., Hill, C. N., & O'Gorman, P. A. (2020). Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. *arXiv preprint arXiv:2010.09947*.

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

Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*.

Zantedeschi, V., Falasca, F., Douglas, A., Strange, R., Kusner, M. J., & Watson-Parris, D. (2019). Cumulo: A dataset for learning cloud classes. *arXiv preprint arXiv:1911.04227*.

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., ... Taylor, K. E. (2020). Causes of higher climate sensitivity in cmip6 models. *Geophysical Research Letters*, *47*(1), e2019GL085782.

Zhang, C., Zhuge, X., & Yu, F. (2019). Development of a high spatiotemporal resolution cloud-type classification approach using himawari-8 and cloudsat. *International Journal of Remote Sensing*, *40*(16), 6464–6481.

Zhang, D., Guo, L., & Karniadakis, G. E. (2020). Learning in model space: Solving time-dependent stochastic pdes using physics-informed neural networks. *SIAM Journal on Scientific Computing*, *42*(2), A639–A665.

Zhang, J., Liu, P., Zhang, F., & Song, Q. (2018). Cloudnet: Ground-based cloud classification with deep convolutional neural network. *Geophysical Research Letters*, *45*(16), 8665–8672.

Zhang, S., & Lin, G. (2018). Robust data-driven discovery of governing physical laws with error bars. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *474*(2217), 20180305.

Zhang, X., Wang, T., Chen, G., Tan, X., & Zhu, K. (2019). Convective clouds extraction from himawari–8 satellite images based on double-stream fully convolutional networks. *IEEE Geoscience and Remote Sensing Letters*, *17*(4), 553–557.