JAMES | Journal of Advances in Modeling Earth Systems

REVIEW ARTICLE
10.1029/2023MS003715

Key Points:

- Inspired by the 2022 Model Hierarchies workshop, we present perspectives on how the field is evolving
- Firstly, we note the growing interest in the use of machine learning which presents another way to ascend or descend the model hierarchy
- Secondly, we discuss the role of the model hierarchy for actionable climate science and decision-making

Correspondence to:

L. A. Mansfield and A. Sheshadri,
lauraman@stanford.edu;
aditi_sheshadri@stanford.edu

Citation:

Mansfield, L. A., Gupta, A., Burnett, A. C., Green, B., Wilka, C., & Sheshadri, A. (2023). Updates on model hierarchies for understanding and simulating the climate system: A focus on data-informed methods and climate change impacts. *Journal of Advances in Modeling Earth Systems*, *15*, e2023MS003715. https://doi.org/10.1029/2023MS003715

Received 21 MAR 2023
Accepted 29 SEP 2023

Updates on Model Hierarchies for Understanding and Simulating the Climate System: A Focus on Data-Informed Methods and Climate Change Impacts

Laura A. Mansfield[1], Aman Gupta[1], Adam C. Burnett[1], Brian Green[1], Catherine Wilka[1], and Aditi Sheshadri[1]

[1]Department of Earth System Science, Stanford Doerr School of Sustainability, Stanford University, Stanford, CA, USA

Abstract The climate model hierarchy encompasses models of varying complexity along different axes, ranging from idealized models that elegantly describe isolated mechanisms to fully coupled Earth system models that aspire to provide useable climate projections. Based on the second Model Hierarchies Workshop, which took place in 2022, we present perspectives on how this field has evolved since the first Model Hierarchies Workshop in 2016. In this period, we have witnessed a dramatic increase in the use of (a) machine learning in climate modeling and (b) climate models to estimate risks and influence decision making under climate change. Here, we discuss the implications of these growing areas of research and how we expect them to become integrated into the model hierarchies framework.

Plain Language Summary The climate modeling community often describes the variety of models used to understand and simulate climate processes as a hierarchy in complexity. Simple idealized models exist at the bottom of the hierarchy and are useful for explaining underlying physics, while fully coupled Earth system models exist at the top of the hierarchy and aim to provide useable climate projections. We present perspectives on how the model hierarchy field is evolving, focusing on two noticeable changes in recent years. Firstly, models are increasingly using machine learning, and secondly, there has been a growing interest in the usability of climate models, for instance, for estimating risks associated with climate change. Here, we discuss the implications of these growing areas of interest and how we expect them to become integrated into the model hierarchies framework.

1. Introduction

Inspired by the perspectives that emerged from the second World Climate Research Programme (WCRP) Model Hierarchies Workshop, which took place at Stanford University in August-September 2022, we offer our thoughts on the future of climate model hierarchies, in particular to attempt to capture the potential future directions of evolution of the field.

The essence of the issue around the use of climate model hierarchies is that, while the Navier-Stokes equations have been known since the early nineteenth century, there is a glaring absence of first-principles theory linking the equations themselves to all their emergent complexities in various contexts. There is also the problem of processes across scales: motions on scales smaller than a model grid box and time step are not explicitly resolved, but significantly impact the resolved scales of motion and the processes they couple with. Furthermore, some Earth system processes such as sea-ice, land, and vegetation dynamics do not have closed-form mathematical representation, complicating their simulation in climate models and adding uncertainty in climate projections. Limitations in computing power also impose inherent trade-offs on which processes can be simulated and for how long. Simplified models, both analytical and numerical, thus have historically been of great use in advancing understanding of the fluid dynamics of the Earth's atmosphere and ocean, and have also aided the development of comprehensive global Earth system models.

The two key differences that made this second workshop stand out from the first, in our view, are (a) the dramatic growth of machine learning (ML), and data-informed methods more generally, as tools to aid modeling and understanding of the Earth system on weather, seasonal, and climate time scales, and (b) a fervent desire on the part of the modeling community to provide actionable science to downstream users of climate and weather models, which could include policy makers, private industry, and governments.

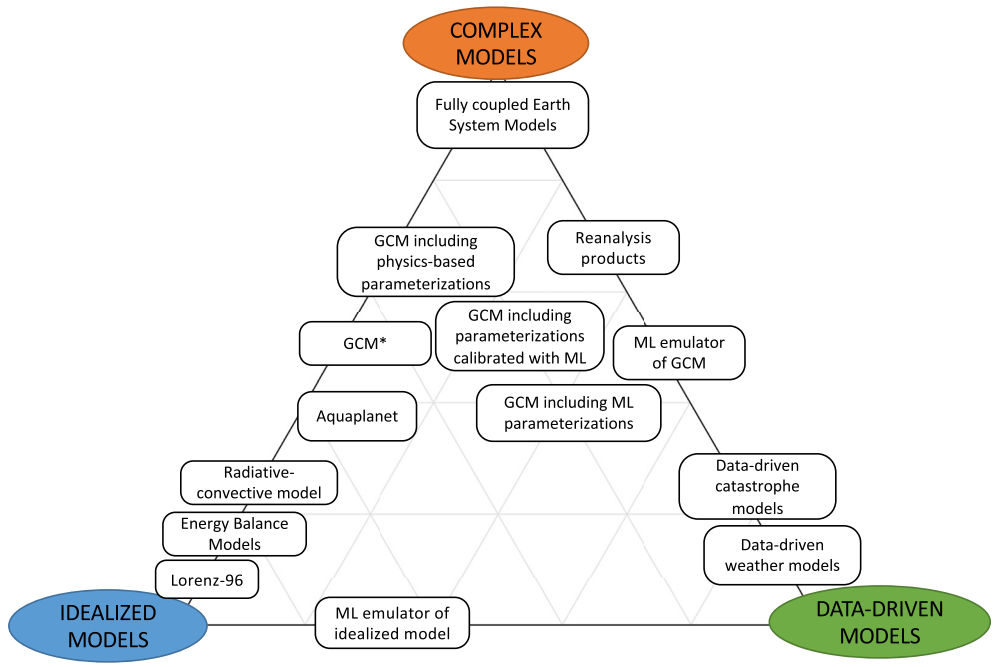

Figure 1. Diagram highlighting the range of models in the hierarchy, including idealized models, complex models, and purely data-driven models. Here "GCM" is used to mean atmosphere- or ocean-only general circulation models, in contrast to fully coupled Earth system models that include atmosphere, ocean, chemistry, land, vegetation, biosphere, sea-ice, and cryosphere components.

In Section 2, we introduce model hierarchies and define concepts discussed in this manuscript, such as complexity, generalizability, and interpretability. Given the rapid rise in popularity of ML and data-informed methods, in Section 3 we review their use in various aspects of modeling, including uncertainty quantification (UQ), subgrid-scale (SGS) modeling, and emulating various aspects of models, including the model itself. In Section 4, we discuss broadening the traditional scope of model hierarchy to incorporate the increasing desire for climate science that is useful in predicting and preparing for climate change impacts. In Section 5, we present conclusions and offer suggestions for the future of model hierarchies.

2. Model Hierarchies Today

With the advent of high performance computing, the general circulation model (or global climate model; GCM), has become the standard starting point for climate research. For many areas of research, however, a GCM is an insufficient and/or opaque tool, necessitating the use of models with different spatial scales, with different complexities, or that include different physical processes.

Following Held (2005), Bony et al. (2013), Jeevanjee et al. (2017), and Maher et al. (2019), among others, we consider climate model hierarchies as a set of ladders connecting our conceptual understanding of the physical principles of Earth's climate with comprehensive modeling and attempts at prediction of the Earth system in all its complexity. A "hierarchy" may also be defined as a spectrum of complexity along many different axes. Model resolution is one of the more commonly traveled axes, often traded against the size of the model domain or length of simulation. Others include the number of parameterized processes, the complexity of those parameterizations, and the coupling of multiple different models together, such as an oceanic model to an atmospheric model, or a climate model to an economic model. We can visualize different possible model configurations on different axes of the hierarchy, as in Figure 1 of Jeevanjee et al. (2017). The term "digital twin" is increasingly used to describe extremely high-resolution models at the "top" of the hierarchy, which produce data volumes on par with observations of the Earth (Bauer et al., 2021). There are many ways to arrange models into a hierarchy, but because the results of model experiments are interpreted in terms of the models' differences, careful choice of the hierarchy's axis is an important part of experimental design. In Figure 1, we offer a visualization of different models in terms

of their overall complexity (which can include the different axes described above) and the influence of data, which is becoming increasingly more relevant along with the prevalence of ML.

"Interpretability" and "generalizability" are two terms with specific meanings for a model hierarchy, and the goal of any hierarchy should be to produce results that are both interpretable and generalizable. We define these as follows:

Interpretability: the ability to translate a model into terms that are understandable to a human (Barredo Arrieta et al., 2020), that is, *an interpretable model ensures the behavior of a system follows known scientific laws*.

Generalizability: the degree to which a model maintains accuracy or skill when applied across different problems or regimes, that is, *a generalizable model should perform well under a range of scenarios*.

These definitions are not new in the context of model hierarchies, and it is typical to deem simpler models with fewer parameters ("more parsimonious") as "more interpretable" when compared against more complex models that include many competing mechanisms (Jeevanjee et al., 2017; Saravanan, 2021). In contrast, more idealized models within the hierarchy are more likely to be designed for a specific case, regime or scenario and may be less generalizable than complex models (e.g., a dry dynamical core). As we discuss in Section 3.3, when adopting artificial intelligence (AI) into the model hierarchy, these definitions of interpretability and generalizability remain relevant. In AI or ML models, results are "interpretable" if direct mechanistic connections can be made between model input variables and the output. Black-box models are not interpretable, but their results potentially can be made interpretable if the connections they find can be tested using a different model in the hierarchy. Generalizability tests the robustness of model output to the design of the experiment. Under what conditions do the model experiment's conclusions hold? Is the proposed mechanism or process always important for describing the behavior of the system? By systematically varying the simulated processes and the complexity of the system, model hierarchies are well suited to provide answers to these questions.

Useful prediction of the impacts of weather and climate on society require accurate representation of interactions across multiple scales and domains. For example, quantifying the potential impact of a hurricane requires knowing the likely track and intensity as well as the population density and built environment along its track. The rapidly growing area of "impacts research" explores the projected impact of climate change on economic and behavioral responses and feedbacks. As computing resources have increased, impacts research has grown more complex. Model hierarchies are often used both to study important processes in isolation and to evaluate causality by studying linkages between processes under simplified sets of assumptions. Just as model hierarchies can be useful in systematically varying a physical parameter space, so too can they be used to explore the possibility space of future human decisions.

3. Data-Driven Methods: The Emergence of Machine Learning

Data indirectly powers many models in the hierarchy already, either by providing initial or boundary conditions that resemble observations (e.g., sea surface temperatures or radiative equilibrium profiles, e.g., the relaxation temperature in Held-Suarez; Held & Suarez, 1994). Indeed, proper observations have been crucial in ensuring consistency with observed climate in state-of-the-art GCMs. However, the past decade has seen the emergence of a new data-driven paradigm, ML, in the existing climate model hierarchy. ML is generally defined as the development of algorithms that allow computers to learn without being explicitly programmed (Samuel, 1959). It is a subset of AI, which is the capability of computers to imitate human behavior. ML-driven modeling potentially provides a radically new approach to dynamical systems modeling, weather forecasting, and comprehensive climate modeling, to the extent of being classified as a new member in the model hierarchy. This section presents an evaluation of the recent developments, benefits, and limitations of ML-driven modeling and where it fits in the hierarchy. We present Figure 1 as a visual aid for describing a model in terms of both its reliance on ML-learned relationships and its complexity in the traditional model hierarchy sense.

3.1. An Introduction to Data-Driven Methods

The power of ML lies in its ability to learn complex, nonlinear relationships solely from data. An expanding set of satellite observations and high-resolution climate model integrations provide access to unprecedented volumes of climate data. This abundance of data makes the data-driven approach a great fit for developing faster and

possibly more accurate models, and even unlocks the potential to discover new science by learning previously unknown relationships. ML methods may not fit into the common concept of a model hierarchy based on process complexity. Still, as we discuss in this section, the diverse set of ML techniques at our disposal merits including them as a novel modeling approach, one that relies exclusively on data.

Climate modeling has already adopted numerous ideas from the field of AI and has, within a short period of time, witnessed a meteoric rise in ML-driven modeling (Reichstein et al., 2019). As discussed at the workshop, ML-assisted analyses have begun to pervade practically all aspects of the existing model hierarchy: from modeling fundamental partial differential equations (PDEs) and dynamical systems (Liu et al., 2022; Pathak et al., 2018a), to modeling and performing equation discovery for SGS processes (e.g., Brenowitz & Bretherton, 2019; Gentine et al., 2018; Rasp et al., 2018; Yuval & O'Gorman, 2020; Zanna & Bolton, 2020), to full-blown efforts to completely replace complex weather prediction models with a single ML model (Bi et al., 2022; Lam et al., 2022; Pathak et al., 2022). Moreover, rather than just being used to build new models, ML is also helping modelers improve existing models by aiding calibration and UQ, by providing emulators that approximate computationally expensive models, and by catalyzing the development of a new-class of data-driven parameterizations (e.g., Schneider et al., 2023).

The first application of ML in climate science dates back to the early 1960s (Glahn, 1964), not long after the first ML model was implemented (Samuel, 1959). For the next several decades, limited computing power discouraged widespread usability of ML (Balaji, 2021). As a result, ML did not gain traction in climate science until the late 1990s, when it was applied to perform a nonlinear variation of principal component analysis (Monahan, 2000), and to design a computationally superior radiation scheme for climate models (Chevallier et al., 1998; Krasnopolsky et al., 2005).

ML based models do not directly rely on the underlying physical principles, distinguishing themselves as new members in the climate model hierarchy.

- *Focus on Empiricism:* Orthogonal to all the existing models within the climate model hierarchy of Jeevanjee et al. (2017), which codify first principles from classical physics in varying degrees to model the large-scale flow, ML models learn the mechanics of the modeled process primarily through data and not from the underlying equations, that is, their reliance on data makes them more empirical than the other members in the hierarchy. In the context of numerical weather prediction (NWP), a traditional weather forecasting model uses an assimilated initial state along with the primitive equations and a suite of physical parameterizations to step that state forward in time and generate forecasts. A ML-based NWP, in contrast, would use the same assimilated state but instead input it to a trained Neural Network (NN) to generate new forecasts.
- *Novel approach to process-isolation:* Having models of varying complexity in the hierarchy facilitates process-isolation, allowing isolating the contribution of certain processes toward, for example, the global circulation or climate sensitivity. ML-based methods provide a fresh approach to process-isolation by stimulating the discovery of latent patterns and correlations.
- *Bidirectional interactions with the existing hierarchy:* The dependence of ML models on observations or existing model simulations for training and validation establishes a two-way relationship that connects them with all the existing members of the model hierarchy. As we will discuss, ML models not only benefit from training and validation data provided by the model hierarchy, but also can be trained on observations to generate state-of-the-art models for the hierarchy.

3.1.1. Machine Learning: Supervised Versus Unsupervised

The broad and rapidly advancing field of AI offers a multitude of models that use abstract "learning" algorithms (e.g., LeCun et al., 2015). These algorithms can be broadly partitioned into two sub-classes: supervised and unsupervised, with supervised learning algorithms being more widely adopted in many areas of research. The key factor distinguishing the two learning styles is the labeling of data. In supervised learning, the input to the ML model is mapped to a labeled output in order to develop a learning association, akin to inputting an image of an animal and labeling it as a cat, dog, or squirrel. After sufficient "training," the model eventually learns the defining features of each furry animal. Supervised learning algorithms, which include regression and classification, are thus suitable for making predictions on new inputs. In contrast, unsupervised learning algorithms mine a high-dimensional data stream to uncover systematic patterns, anomalies or clusters, akin to viewing a landscape photo and clustering areas into forest or savannah, or for dimensionality reduction. A subset of unsupervised algorithms, called generative algorithms, learn joint probability distributions which can be used to generate new

samples. This bird's-eye-view dichotomy, however, is starting to blur, with the emergence of a new class of "semi-supervised" learning (or weak supervision) algorithms that blend ideas from both supervised and unsupervised learning, for example, to train predictive models with unlabeled data. Table 1 provides a glimpse of the spectrum of ML algorithms that currently are being explored in climate modeling. For the remainder of Section 3, we focus on ML applications that are relevant to the model hierarchy. These tend to fall under the category of supervised learning for prediction, while unsupervised methods are most useful for data analysis (clustering, regime identification, dimension reduction).

3.1.2. The Basics of Supervised Learning

When described simply, supervised learning appears very similar to linear regression, where a function input is linearly mapped to the function output. The distinguishing feature of modern ML, however, is its ability to sift through large volumes of data and learn complex nonlinear relationships. Such learning cannot be achieved through simple linear transformations. "Deep learning" methods, such as deep NNs, comprise multiple learning layers through which the input passes, making them particularly tailored to learn complex mappings. The number of layers, which determines the complexity of the NN, can be instrumental in determining its learning skill. Within each layer, the layer-input is mapped to the layer-output using a nonlinear activation (mapping) function (LeCun et al., 2015; Schneider et al., 2017). A composite of multiple such nonlinear activations endows discriminatory powers to the NN. This nonlinearity is being leveraged to learn the multi-scale evolution of the climate system exclusively through data, bypassing the need to use the nonlinear equations of fluid flow.

3.2. AI in the Context of Model Hierarchies

The supervised learning approach outlined above has been central to several recent studies exploring ML methods in climate science. These studies predominantly explore how AI can (a) improve weather forecasting and climate prediction skill and speed, by developing enhanced forecast models, (b) develop novel data-driven physical parameterizations, or (c) quantify model uncertainty arising from parameterizations. Note the two distinct benefits that ML can provide: skill and speed. Points (a) and (b) are concerned with using ML to improve model skill, by training on high quality data sets such as observations or high resolution simulations. In contrast, point (c) relies on the speed-up that ML emulators provide for large ensemble generation and UQ.

3.2.1. Prediction (Supervised and Semi-Supervised Learning)

ML is increasingly being applied to attempt to improve weather forecasting and climate prediction. The central hypothesis being tested is whether data-driven models shine where equation-driven models struggle, by:

(a) learning the chaotic evolution of synoptic-scale weather systems and producing quick and skillful forecasts, even providing statistically robust subseasonal-to-seasonal (S2S) forecasts, and
(b) providing an improved representation of unresolved processes in order to provide more certain climate projections for the 21st century and beyond

3.2.1.1. ML Has Had Success in Weather Forecasting

Recent research suggests that ML has begun to break ground in terms of improving weather forecast accuracy. NN-based prediction models can now generate increasingly precise predictions of simple chaotic systems, such as those modeled by higher-order nonlinear PDEs (Pathak et al., 2018a). This numerical capability of NNs has been generalized by Liu et al. (2022) to develop faster deep learning numerical solvers that can replace traditional time-stepping numerical schemes, such as those heavily employed by climate models to solve the equations of motion. Rigorous tests on more advanced chaotic systems such as the Lorenz-96 model (L96) have reaffirmed the advantages of NN-based predictions (Chattopadhyay et al., 2020). Because different variables in an L96 model evolve on different timescales, this model can serve as an idealized proxy for the multi-scale evolution of the climate system. For this reason, such models have been frequently used in climate science to study chaos. In Chattopadhyay et al. (2020), NNs generated accurate predictions of the slowly evolving state variables in an L96 model, while struggling to capture the intermittent evolution of the fast variables. These results illustrate the latent power of ML, which could be leveraged for skillful weather forecasting and S2S predictability (Cohen et al., 2019).

Such proofs-of-concept have motivated NN-based weather forecasting initiatives like Scher (2018), Resnet (Rasp & Thuerey, 2021), Weyn et al. (2021), FourCastNet (Pathak et al., 2022), PanguWeather (Bi et al., 2022), and

Journal of Advances in Modeling Earth Systems
10.1029/2023MS003715

Table 1
Examples of Machine Learning (ML) Architectures and Their Uses in Climate Science

	ML architecture	Description/properties	Examples
SUPERVISED LEARNING	Regression (linear, non-linear, logistic, etc.)	Basic regression methods can be considered machine learning	Atmospheric chemistry parameterization (Nowack et al., 2018)
	Random forests, boosted forests	Ensemble of decision trees, useful for prediction	SGS parameterizations (Yuval & O'Gorman, 2020)
	Artificial Neural Networks (NNs, incl. recurrent NNs, convolutional NNs)	Layers of interconnected nodes and neurons that can learn highly non-linear relationships. Recurrent NNs used for time-series analysis, CNNs for image processing.	SGS parameterizations (e.g., Krasnopolsky et al., 2005) Weather forecasts (e.g., Pathak et al., 2022) Downscaling (e.g., Blanchard et al., 2022)
	Gaussian processes	Bayesian approach to learn predictive distributions	Uncertainty quantification (e.g., Carslaw et al., 2013)
UNSUPERVISED LEARNING	Generative models, for example, Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs)	Models that learn joint probability distributions. Useful for learning latent space (e.g., VAEs) and for generating new samples.	Dimensionality reduction (Mooers et al., 2022; Tibau et al., 2018) Stochastic SGS parameterizations (Perezhogin et al., 2023)
	K-means clustering	A clustering algorithm that identifies clusters/classes in multi-dimensional data	Identifying weather regimes (Michelangeli et al., 1995)
	Principal component analysis and its variants	A dimensionality reduction technique that can be used to identify patterns that capture the most variability	Jet variability (Thompson & Wallace, 2000)
SEMI-SUPERVISED LEARNING	Derivative-free optimization methods, for example, Ensemble Kalman Inversion	Train supervised model architectures (e.g., NNs) without direct observations of model output, or with missing/noisy observations	SGS parameterizations (Kovachki & Stuart, 2019; Lopez-Gomez et al., 2022)
	Transformers	NN that learns where to direct learning through attention mechanism (often referred to as a type of self-supervised)	Weather forecasts (e.g., Pathak et al., 2022)
	Image-style transfer	NNs trained with loss function that combines supervised and unsupervised loss functions	Identification of sparse/rare phenomena (e.g., Mahesh et al., 2023)

GraphCast (Lam et al., 2022). These initiatives aim to completely replace complex weather forecasting systems and their suite of physics parameterizations with standalone NNs that learn synoptic-scale variability directly from high-dimensional observations. Although they do not embed any physics (e.g., conservation laws) into the ML architecture, they can accurately learn the expected behavior of the system on short timescales. Once trained, the authors claim that NNs can operate up to 45,000× faster than typical NWP models, on fewer compute nodes (Pathak et al., 2022), offering a cost-effective alternative to produce reliable forecasts and larger ensemble generation. Currently, the focus has been on weather forecasting, with NN-based models achieving skillful forecasts for lead times consistent with atmospheric predictability (∼2 weeks), although they do not yet include data assimilation techniques. Next, we might expect advances that allow for longer-term prediction on S2S and climate timescales, where it becomes necessary to account for other components of the Earth system, such as land and ocean.

3.2.1.2. Climate Prediction Is More Complex Than Weather Forecasting

Predicting future climate is significantly more challenging than forecasting near-term weather. Unlike operational NWP models, climate models are not nudged to observations but instead self-consistently evolve through time. Therefore, multidecadal climate predictions require accurate model representation of a plethora of Earth system processes that induce climate variability over multiple timescales. Many of these processes, including shallow and deep convection, radiation, and gravity waves, are not resolved explicitly at typical climate model resolutions. Some processes, such as cloud microphysics, vegetation, and sea ice evolution, do not even have closed-form analytical formulations at any resolution. These processes are, therefore, parameterized. Almost all of these parameterizations are based on approximate single-column mass, momentum, or energy closures and, taken together, are the leading source of structural uncertainty in a model. These parameterizations also are often ill-tuned and form computational bottlenecks (Hourdin et al., 2017). ML models provide a fresh approach to developing the next generation of potentially faster and more physically consistent parameterizations, ones that benefit from a growing set of climate data. The underlying principle is fairly simple: if finding a closed-form representation for a physical process is intractable, one can instead learn its governing principles from data through physics-informed learning. This empowers the creation of a new class of empirical data-driven parameterizations for potentially all Earth system processes, one that can work in concert with the existing model hierarchy to provide better climate projections (Schneider et al., 2017).

This approach would require suitable quantities of training data, either from observations or high-resolution modeling. High-resolution models can provide training data for dynamics-driven SGS processes in the atmosphere and oceans, such as gravity waves, turbulence, convection, and cloud cover. These processes lie in the "gray zone," with scales O (1–100 km) which are under-resolved in typical climate models but are largely resolved in computationally intensive sub-kilometer scale models (e.g., atmospheric subgrid momentum fluxes: Yuval & O'Gorman, 2020; Yuval et al., 2021; Wang et al., 2022; ocean momentum forcing: Guillaumin & Zanna, 2021; Perezhogin et al., 2023; convection: Brenowitz & Bretherton, 2019; Gentine et al., 2018; clouds: Rasp et al., 2018; gravity waves: Sun et al., 2023) or large eddy simulations (e.g., eddy-diffusivity momentum flux: Lopez-Gomez et al., 2022; Shen et al., 2022). However, other coupled processes such as atmospheric chemistry, sea ice cover, and vegetation dynamics are not modeled explicitly at any resolution and may make use of observational data sets (e.g., ozone: Nowack et al., 2018; sea-ice: Andersson et al., 2021; vegetation: Chen et al., 2021). Alternatively, some studies simply use ML to emulate existing parameterizations at a lower computational cost (e.g., radiation: Chevallier et al., 1998; Krasnopolsky et al., 2005; aerosol microphysics: Harder et al., 2022; cloud microphysics: Andre Perkins et al., 2023). These studies have demonstrated that ML can successfully replace physics-based parameterizations to improve either model accuracy or speed. Some of the main challenges lie in maintaining stability once coupled online to a GCM (e.g., Brenowitz et al., 2020; Rasp, 2020), and in the interpretability and generalizability of model output, discussed further in Section 3.3. We suggest that ML methods may be most valuable when used in concert with parameterizations that remain as physics-based as possible, for example, in learning a specific parameter from data (e.g., Schneider et al., 2017, 2023), as discussed further in Section 3.2.2.

3.2.1.3. The Need for ML-MIPs

In the past 5 years, several modeling groups worldwide have adopted the data-driven approach to developing SGS parameterizations. This motivates the need for systematic comparison of emerging data-driven parameterizations. Similar to existing model intercomparison projects (MIPs), these comparisons could be instrumental in evaluating robustness and reproducibility of data-driven parameterizations. Such evaluations should focus on physical consistency, sensitivity, and generalizability of these parameterizations. Without sufficient care, the

NN-based approach to parameterizations could be susceptible to producing nonphysical outputs without accountability. Thus, in order to ensure maximum interpretability, the parameterizations from different modeling groups could be tested rigorously on key issues such as the sensitivity of the parameterization to the ML architecture and which training data are used, optimal objectives, validation, out-of-set performance, and UQ. These goals can be accomplished by introducing benchmark tests, analogous to those for existing model intercomparisons, in which the different parameterizations developed using different training data sets are re-trained on identical training data and optimized to minimize identical loss functions. Moreover, a key element of such tests could be to comprehensively evaluate process evolution and predictive skills on multiple timescales, in addition to obtaining a desired long-term climatology. Preliminary efforts in this direction are taking shape, with initiatives including WeatherBench (Rasp et al., 2020), WeatherBenchProbability (Garg et al., 2022), and ClimateBench (Watson-Parris et al., 2022), all of which demonstrate benchmark data sets designed for ML comparisons of standard meteorological variables and set the stage for more comprehensive, future ML-MIP projects.

3.2.2. Uses of ML for Uncertainty Quantification and Calibration

The 2022 Model Hierarchies Workshop also highlighted how ML increasingly is being used to improve model projections by aiding UQ. UQ is the estimation of uncertainties in model output. These uncertainties include lack of knowledge about future climate scenarios (scenario uncertainty), uncertainty in model parameters (parametric uncertainty), internal variability of the climate system, and discrepancies between a model and reality (structural uncertainty) (Hawkins & Sutton, 2009). UQ requires integrating a large number of model simulations to span a wide range of model inputs (parameters, scenarios, or initial conditions), which can be a challenge when working with comprehensive GCMs. This need arises because UQ studies usually assume model inputs to be defined by probability distributions, where many samples are required to obtain accurate probability distribution functions (e.g., using Bayesian methods), which are computationally expensive to generate for GCMs using traditional methods. ML methods are a way to emulate GCM output and make the cost of UQ computationally feasible. UQ is a well-established method for fairly small-scale problems with $O(10)$ inputs, which can make use of emulators such as regression splines, Gaussian processes, and polynomial chaos expansions (e.g., Bulthuis et al., 2019; Carslaw et al., 2013; Sraj et al., 2016). Recent advances in deep learning enable addressing higher-dimensional UQ problems with even $O(1{,}000)$ inputs (Lan et al., 2022).

Parametric UQ is closely related to calibration, where the chosen model parameters are carefully tuned to obtain model output consistent with some ground truth, such as a desired climate state or its variability. Bayesian methods are also popular for calibration, including Bayesian optimization (Kennedy & O'Hagan, 2001), ensemble Kalman inversion (e.g., Cleary et al., 2021), and history matching or iterative refocusing (Williamson et al., 2013). These methods involve assigning a prior probability distribution to parameters based on domain knowledge. If we do not have much knowledge of what the parameter values should be, this distribution can instead be chosen to be a wide uniform prior, known as a "vague" prior. The probability distribution is then constrained by data, often from observations, to derive a posterior probability distribution. Each of these methods requires either gradients (Bayesian optimization) or an ensemble of simulations (ensemble Kalman inversion, history matching), making ML-driven emulators such as Gaussian processes useful. Calibration should lead to improvements in model accuracy at any level within the model hierarchy (e.g., Couvreux et al., 2021; Hourdin et al., 2021).

Both calibration and UQ benefit the model hierarchy framework. They are useful for model development to better understand and improve GCM components (e.g., Carslaw et al., 2013; Couvreux et al., 2021; Dunbar et al., 2021; Guo et al., 2014; Williamson et al., 2017; Yang et al., 2012) and are relevant to impact studies concerned with risk and uncertainty (e.g., Clare et al., 2022; Edwards et al., 2021; Harrington et al., 2021).

3.3. Are Data-Driven Schemes Trustworthy?

Training exclusively on model or observational data can have its drawbacks. There is no clear way to decide whether the model has been trained sufficiently, whether it performs well on a previously unseen set of inputs (generalizability) or, most importantly, whether we can meaningfully interpret its output (interpretability). Trustworthy ML is a growing area of interest as users favor models that are fair, reliable, and robust, which becomes especially important for high-stakes decision making (e.g., McGovern et al., 2022). The precise definition of "trust" in a model depends on its application, but for integration into the model hierarchy, we aspire for a trustworthy ML model to be both generalizable and interpretable, as we do with any model. Simply put, we can trust

a model if it can both deduce how the Earth system responds to unseen forcing, and if we can understand how it made that deduction. The issue of trustworthiness is potentially more pronounced for ML models than for other models in the hierarchy which are typically derived using a robust set of scientific laws and assumptions, naturally making them more interpretable.

3.3.1. Generalizability

The key issue of generalizability concerns whether an ML scheme can generate reliable predictions for climate states it was not previously trained on. Studies that have applied ML to weather forecasting with some success (e.g., Arcomano et al., 2020; Bi et al., 2022; Pathak et al., 2022; Rasp & Thuerey, 2021) assume that a sufficiently large training set contains almost all possible states of the Earth system. In such cases, an ML algorithm can be carefully trained and validated to avoid overfitting. Overfitting occurs when an ML model provides skillful prediction on the data it was trained on, but performs poorly on new data. This issue becomes particularly important for ML-based climate models applied to global warming scenarios. Even large volumes of training data based on the present-day climate cannot contain all possible states of a rapidly warming climate. Thus, the relationships learned during training may generalize poorly to drastically different climatologies, for example, extreme precipitation in a 2K-warmer globe. In such cases, extrapolation by the ML model can be meaningless. For example, Rasp et al. (2018), found that a NN-based SGS cloud parameterization embedded in a GCM performs poorly when applied to temperatures exceeding those in the training data. Pitfalls like these call into question the trustworthiness of ML schemes as opposed to physics-based models, which tend to be more climate-agnostic.

Espinosa et al. (2022) present a ML model that appears to generalize well, in the sense that it responds to new scenarios similarly to the model on which it was trained. However, the degree of generalizability presumably is sensitive to the complexity of the problem and the NN architecture (e.g., Rasp et al., 2018). Other ML methods such as random forests (Table 1) exhibit excellent online stability but also fail to generalize well (e.g., Brenowitz et al., 2020; Yuval & O'Gorman, 2020).

There is currently no universal solution to generalizability, but one helpful approach could be to design the training data more strategically. For example, by simply expanding the spatial coverage of their input data, Sit and Demir (2019) significantly improved flood projections using a recurrent NN, even predicting out-of-set extreme flooding events. Similarly, when designing emulators for existing GCMs (e.g., for faster prediction or to emulate high-resolution processes by training on parameterizations), incorporating a wide range of climate model scenarios covering both present-day and projected future climates reduces the chance of erroneous extrapolation (e.g., O'Gorman & Dwyer, 2018; Rasp et al., 2018; Watson-Parris et al., 2022). If the desired wide range of training data does not exist, this approach requires first creating the necessary set of GCM simulations for training. This makes traditional GCMs central for constraining ML models, and also reiterates the third point about bidirectional interactions that we mention in Section 3.1. Of course, unlike Sit and Demir (2019), who relied on observations, using this approach to improve generalizability seems muddled by the fact that the future projections of GCMs themselves are marked with uncertainties; training ML models on such data could amplify these uncertainties. In this case, encouraging consistency with the physical laws of the Earth system might salvage efforts to create generalizable models (e.g., Reichstein et al., 2019).

Physics-informed ML (or knowledge/theory-guided ML) couples physical knowledge to the ML architecture and offers one approach to enhance ML generalizability and trust (e.g., Gentine et al., 2021; Irrgang et al., 2021; Karpatne et al., 2017; Kashinath et al., 2021; Raissi et al., 2019a). Here, physical conservation laws are incorporated into ML algorithms, either by constraining the loss function (soft constraints, also called regularization; e.g., Brenowitz et al., 2020; Harder et al., 2022) or by more strictly enforcing conserved properties (hard constraints; e.g., Beucler et al., 2021a; Chattopadhyay et al., 2021). Variations of physics-informed ML include designing "climate-invariant" algorithms by rescaling inputs and outputs to avoid extrapolation (Beucler et al., 2021b) or incorporating equations governing the dynamics to build hybrid ML algorithms (e.g., Pathak et al., 2018b; Raissi et al., 2019a).

Yet another option is quality control (also called novelty detection or UQ), where the ML algorithm includes a quality control check that determines whether there are likely to be large errors because, for instance, the data are out of sample. This technique is increasingly used in the SGS parameterization community to develop "compound parameterizations", which use a ML model for within-sample data and revert to the physics-based model for out-of-sample prediction (Krasnopolsky et al., 2008; Sanford et al., 2022; Song et al., 2021). Bayesian

methods such as Bayesian NNs are a natural choice of ML architecture for estimating model uncertainty alongside predictions (e.g., Garg et al., 2022; Luo et al., 2022; Ortiz et al., 2022).

From a model hierarchies perspective, physics-informed ML seems the most promising method because it enables understanding individual physical systems with known underlying principles. However, it may be challenging to use this approach for many other Earth system processes, such as land ecology and biosphere processes, which might be too complex to have even approximate closed-form representations.

3.3.2. Interpretability

Even when an ML model can successfully generalize, it can be difficult to interpret its output. This leads us to the second major hurdle in the adoption of ML in climate science: interpretability, defined in Section 2 as the ability to translate a model into understandable terms.

Model hierarchies are imperative to understanding the mechanisms driving climate variability (e.g., Nabizadeh et al., 2019; Roach et al., 2022; Shaw & Smith, 2022). ML methods have high relevance in climate science, yet their prohibitively limited interpretability prevents their full assimilation into operational GCMs (e.g., Chantry et al., 2021; Irrgang et al., 2021). As discussed in Section 4, interpretability is crucial when using ML models for climate change impacts, risk assessment, and policy making. Investing time and resources to mitigate the effects of climate change based on the predictions of a non-interpretable model would be difficult to justify. This need is motivating the development of new ML models whose outputs can be explained or interpreted.

In practice, the first approach to interpret an ML model is to manually verify the correlations between outputs are reasonable (e.g., Rasp & Thuerey, 2021). Following this initial step, there exists a wide variety of more sophisticated methods that can be used to look inside the "black box" of ML models and explain the information hotspots that establish the relationships learned (Mamalakis et al., 2022; McGovern et al., 2019). These methods are often labeled as "explainable AI" (XAI), which can be defined as:

Explainable AI: AI that provides details or reasons to make its functioning clear or easy to understand (Barredo Arrieta et al., 2020).

Examples of XAI methods include saliency maps (e.g., Brenowitz & Bretherton, 2019), backward optimization (e.g., Gagne et al., 2019), layerwise relevance propagation (e.g., Labe & Barnes, 2021), and Shapley values (e.g., Espinosa et al., 2022). These XAI methods typically involve creating visualization maps which are used to interpret the weights of the trained model by connecting them to established theory (see McGovern et al., 2019 for a more comprehensive list of methods used in meteorology). They highlight why a model predicts a given output, in terms of which input variables are more influential. For example, wet-bulb temperatures are the primary predictor of large hail based on an ML model, which makes sense because above a certain temperature threshold, freezing rain is impossible (McGovern et al., 2019). This analysis requires experts to manually determine whether these relationships make sense in the context of current theory. Specifically, one would check that (a) correlations between input and output variables are not spurious and have a physical basis and (b) if there is a lack of correlations between given input and output variables, then this is due to a lack of a physical relationship. Verification is usually performed on a subset of the test data and cannot be applied to all predictions. This leads us to the question: how many test cases can we assess before assuming the interpretability of the model holds everywhere? In other words, is the interpretability itself generalizable?

The fact that these methods are applied post hoc—that is, *after* building, training and testing the model—highlights an important limitation of XAI: the process requires retrospectively verifying that the relationships are consistent with the existing physical theory. This does not mean that an ML model is inherently interpretable or, following our definition in Section 2, understandable to a human. For instance, it is possible that XAI methods highlight the right correlations for the wrong reasons (e.g., a sample bias or limited training data) (Rudin, 2019). Rudin points out the subtle difference between XAI and interpretable AI, which can be defined as

Interpretable AI: building AI models that follow steps that can be translated into understandable terms (Rudin, 2019).

While XAI provides reasons, in terms of correlations, for why a model makes a particular prediction, interpretable AI aims to build models where the computations themselves can be interpreted, making them less black-box and more transparent. This approach could ensure that known causal relationships exist in model predictions.

Currently, only a few methods allow interpretability tools to be built into their ML architecture (Barnes et al., 2022; Chen et al., 2019; Sonnewald & Lguensat, 2021). Barnes et al. (2022) introduce an interpretable convolutional NN that classifies the phase of the Madden-Julian Oscillation by separately classifying sub-regions of the data and selecting the phase with the highest probability based on similarities across all sub-regions. Sonnewald and Lguensat (2021) use a classification algorithm to first select the most likely global dynamical ocean regime, from which circulation changes are predicted using supervised learning. These methods both involve careful design based on expert knowledge of the possible regimes within the climate system, much like a traditional physics-based model, and once built, can also harness the computational speed-ups offered by ML. Similarly, physics-informed or hybrid approaches as discussed above offer a way to incorporate interpretable steps into a model (e.g., by enforcing conservation laws, Beucler et al., 2021a; Chattopadhyay et al., 2021). Alternatively, techniques such as data-driven equation discovery can learn closed-form equations for SGS parameterizations, which are more easily interpreted (e.g., Mojgani et al., 2022; Raissi et al., 2019b; Zanna & Bolton, 2020).

All of the above interpretable ML methods are still relatively new within climate modeling and have been successful at revealing known relationships, but they have not yet discovered new scientific insights. With further research into interpretable ML for climate modeling, one could envision how these methods may be incorporated into the model hierarchy framework and possibly assist prediction in a transparent way.

3.4. Model Hierarchies Tomorrow: Incorporating Data-Driven Methods

Efforts like XAI have demonstrated the potential to uncover new statistical relationships between model inputs and outputs. In addition to improving existing climate models, and complementing physics-based models in the hierarchy through better parameterizations, calibration, and UQ, some claim that ML-based models may also provide a novel approach to modeling physical systems and unraveling new climate patterns, teleconnections, and mechanisms through a potentially more nuanced process isolation and learning (Mamalakis et al., 2022)—which is one of the central goals of having a model hierarchy. Incorporation of ML algorithms has already led to novel discoveries in medicine (Stokes et al., 2020), astronomy (Valizadegan et al., 2022), chemistry (Hueffel et al., 2021) and mathematics (Davies et al., 2021). Thus far, ML has not led to an entirely new breakthrough in climate science. However, given the difficulty involved in predicting the future of scientific enterprise, this possibility may still exist.

The relationship between data-driven and physics-based models must be two-way, as we rely on physical models to validate the explainability of ML models in the hierarchy (e.g., Mahesh et al., 2023). Thus, the degree of data-drivenness involved in climate modeling, predominantly via ML algorithms, inspires us to add another axis to the existing notion of model hierarchies, where the existing axes describe the trade-off between model complexity and parsimony/ideology (Figure 1). One end of the hierarchy comprises idealized models that elegantly describe a system using closed-form equations, such as Lorenz-96 or energy balance models. One can ascend the model hierarchy by increasing the representation of dynamical processes (resolved dynamics), physical processes (such as boundary layer or bulk processes that are typically unresolved in GCMs; Jeevanjee et al., 2017), or resolution (Maher et al., 2019). At the top of this hierarchy, we can expect to find the most complex physical models that aim to achieve realistic prediction of the Earth system.

Figure 1 introduces a new axis defined by the inclusion of data as an additional way to ascend or descend the model hierarchy, with purely statistical or "inductive" models on the lower right-hand corner, compared to theory-based or "deductive" models that exist along the upper left axis (see, e.g., Chapter 17 of Saravanan, 2021). A user may wish to consider the influence of data when choosing models most suited for a particular problem. For example, if the goal is to cheaply obtain an ensemble of forecasts to assess the likelihood of a heatwave, a purely data-driven model might be best suited (e.g., Weyn et al., 2021) but may not be so easily interpreted. If, in contrast, the goal is to query possible mechanisms for why a heatwave might occur, one might employ multiple models of differing complexity. While traditionally these models would be physics-based, explainable and/or interpretable AI methods could complement (rather than replace) them, by highlighting the main drivers of a given prediction and narrowing down the possible mechanisms to be explored.

Data play a role in all models in some way. Even idealized models such as aquaplanets are partially driven by data that describe the boundary conditions. However, some rely much more heavily on data: for example, reanalysis models, which involve assimilating large observational data sets into dynamical models. In the "middle" of the

hierarchy, we can find hybrid models containing both data-driven and physics-driven components: for example, a GCM with ML-based parameterizations which may achieve faster prediction or improved skill. These models achieve performance at the cost of interpretability, especially if multiple ML-based parameterizations are coupled together. They may be suitable for developing scientific understanding, but the private sector and governments may consider these less interpretable and therefore less trustworthy. They are distinct from GCMs that include physics-based parameterizations that have been tuned using ML techniques to improve prediction skill while also maintaining interpretability.

4. Model Hierarchies for Climate Change Impacts and Useable Climate Science

4.1. A Desire for Usability

One recurring theme of the workshop was the expressed desire by climate scientists for their work to be useful outside the scientific community. Also voiced was a growing desire within the private sector to make use of climate model projections. These ideas were the main topic of several presentations at the workshop and were given as an underlying motivation for a few others.

This movement toward usability (e.g., Gettelman & Rood, 2016) presumably is driven by the increasing magnitude and urgency of the anthropogenic climate change crisis (IPCC, 2018, 2021; Reidmiller et al., 2018). And as governments and the private sector increasingly accept that some degree of climate change adaptation is necessary, opportunities are emerging for scientists to direct their research toward climate change impacts (PCAST, 2023).

When motivating research in basic climate science, researchers often implicitly or explicitly assume a "linear model" of science informing policy, in which impartial scientists provide information to legislators and regulatory agencies who act in the public's best interest (Beck, 2011; Jasanoff & Wynne, 1998; Lahsen, 2005; Pielke & Roger, 2007; Sobel, 2021). However, as argued by Sobel (2021), this linear model does not presently apply to climate science: because of political roadblocks, better climate science does not directly translate to better climate policy.

Climate science can influence policy through other pathways, such as forming the basis for litigation (Held v. State of Montana, 2020; Muffett & Feit, 2017), or swaying political agendas or public opinion (Drake & Henderson, 2022), as in some extreme-event attribution studies (Jézéquel et al., 2020; Sobel, 2021). Results of basic research in climate dynamics may not always be actionable for mitigating or adapting to the impacts of climate change, and reducing scientific uncertainty in climate change projections may not lead to an increase in use of this information by policymakers and other end-users (Lemos & Rood, 2010). As a result of this impasse, many climate scientists are asking how they can use their scientific research to contribute toward solving challenges associated with the impacts of climate change.

4.2. A Need for Actionable Climate Risk Predictions

Government officials and agencies, nonprofit and community-based organizations, and corporations across various sectors increasingly desire to incorporate climate change knowledge into their planning for the future. Along with individuals, these groups are collectively commonly referred to as decision makers, or anyone "who may need to make decisions about climate change" (p. 25, National Research Council, 2011). For example, state and local governments along the Gulf Coast of the United States face concerns about rising sea levels and potentially increased risk of damages from hurricane landfall, while other regions face increased risk of wildfires and drought (Reidmiller et al., 2018). However, current GCMs have coarser grid spacings than typical city scales, necessitating the use of regional or downscaled models to simulate local impacts. Directing climate science toward understanding risks of specific local or regional climate change impacts, and working with local or regional governments and entities, is a way to work toward actionable climate science.

Model credibility and trustworthiness vary depending on the intended application and target audience. State-of-the-art GCMs, such as those used in the Coupled Model Intercomparison Project (CMIP) version 5 (CMIP5; Taylor et al., 2012) or version 6 (CMIP6; Eyring et al., 2016), have greatly enhanced our scientific understanding of global climate but are not always useful to or interpretable by non-scientists. One hurdle to the usability and interpretability of these models is their complexity. Another obstacle to the usability of CMIP6 models is their coarse spatial scale, typically 50–100 km, which is insufficient to resolve some features that are

important for regional-scale planning and disaster preparedness. For example, while hydrological features such as the North American Great Lakes affect local weather and climate, most CMIP5 models do not represent their regional climate impact at all (Briley et al., 2017, 2021). Regional climate models traditionally have proven very useful for assessing current risk at finer spatial scales (Giorgi, 2019), but these models have limited capabilities for projection because climate change is a global phenomenon and local variability can depend upon changes and teleconnections far afield.

Insurance companies represent another large sector planning for and already dealing with the use of climate data. These for-profit financial institutions use their own "catastrophe" or "cat" models, statistical models developed in-house, to calculate the risk of insurance payouts due to natural disasters, many of which are increasingly influenced by climate change (Hereid, 2022). The potential for ML to improve cat models is an area of exploration (Gualdi et al., 2022; Swiss Re, 2021). As the forced climate change signal becomes larger relative to the internal variability in a given location, cat models will need to account for a changing climate state, rather than rely solely on historical data.

The push by companies to incorporate climate change information into their business models is driven by both financial and regulatory pressures (Hereid, 2022). A recent proposal by the United States Securities and Exchange Commission (SEC) would require companies to disclose their climate change-related financial risks at the ZIP-code level (SEC, 2022). Demands such as these have "leap-frogged the current capabilities of climate science and climate models by at least a decade" (Fiedler et al., 2021) and reflect a disconnect between the detailed local-scale information sought by the financial and insurance industries and the coarser-scale projections typical of the global climate modeling community. A move toward integrating understanding of local-scale and near-term risks with global projections, or otherwise bridging the gap between the information sought by these decision makers and the information climate science currently can provide, would be a step in the right direction for useable climate science.

4.3. An Ecosystem of Models for Climate Change Impacts

One possible example of a model hierarchy applied to climate risk, in which models of differing complexity can be applied to the same problem, is downscaling. Downscaling may be achieved using a finer-scale dynamical model embedded in a GCM (Giorgi & Gutowski Jr, 2015), a statistical model (Thrasher et al., 2022), or ML (Blanchard et al., 2022). In a sense, these models of differing levels of complexity and connections to first principles could be viewed as an expanded hierarchy.

Apart from the example of downscaling, it is difficult to define the application of model hierarchies, as the climate science community commonly understands them, to climate risk assessment. Whereas a model hierarchy employs an array of models to simulate the same phenomenon at different levels of complexity, with the goal of developing fundamental understanding and idealized physical theory, it is not immediately clear how this could extend to modeling the risks of economic and human loss due to extreme weather. Statistical relationships or heuristics, such as a relationship between the wind speed of a hurricane and the severity of the resulting infrastructural damage (Emanuel, 2011), are fundamentally different from the physical understanding gained from a model hierarchy, and may require a reframing of the hierarchy.

The related concept of a model "ecosystem," described in a report by the President's Council of Advisors on Science and Technology (PCAST, 2023), offers a helpful complement to the more traditional model hierarchy. A model ecosystem resembles a model hierarchy in that it encompasses a variety of models used in concert to achieve a scientific purpose. But unlike a hierarchy, in which an array of models simulates the same phenomenon at different levels of complexity, a model ecosystem incorporates models of different processes from different disciplines. Such an ecosystem of models might include physical weather and climate models that estimate probabilities of extreme weather events such as wildfires or flooding, models of the local-scale variations in intensity and impacts of those events, and models that project damage due to those local impacts (PCAST, 2023). Bringing together these models from different disciplines would help make the resulting predictions more transparent, easily verifiable, trustworthy, and perhaps interpretable for downstream users, and would facilitate progress toward better, more actionable predictions of extreme weather risk due to climate change (PCAST, 2023).

While "interpretability" has a specific definition for ML, as described in Section 2, it is also a more generally desired quality for decision makers, who need to be able to draw defensible conclusions from climate model

projections that are deemed trustworthy. Accordingly, it is worth considering how the paradigm of the model hierarchy or ecosystem can aid in this task, especially as ML and other new methods and data sets are deployed, existing models are improved, and new ones are developed.

Though not directly related to the concept of a model hierarchy or ecosystem, another possible step toward making climate data more accessible to end users is the involvement of human interpreters. For example, most weather forecast bureaus use models to provide guidance in concert with teams who assess the model projections and communicate results to the public. Analogous figures for climate models, termed "climate interpreters" (Gettelman & Rood, 2016) or "climate translators" (Fiedler et al., 2021), could help frame model projections with easy-to-follow narratives or "storylines" (Shepherd et al., 2018) that still capture the key science, making them more interpretable, trustworthy, and salient for decision makers.

In summary, there is a growing demand for useable climate science, especially in the form of studying climate change impacts and adaptation. GCMs are too complex, difficult to interpret, and spatially coarse to offer useable information about local-scale climate change impacts. Instead, insurance companies and others outside academia are developing their own models to assess local climate change impacts and risks. Whereas GCMs fail to be sufficiently interpretable, credible, and useable for many decision makers, risk models could be designed with end users in mind, and at different levels of complexity depending on their research purpose. This set of models could form an "ecosystem" (PCAST, 2023), with each model tailored to a different purpose and climate scientists helping to develop and interpret them. Some of the same concepts central to model hierarchies thus may play a new and growing role in making climate science actionable.

5. Conclusions

While theming a conference around the principle of model hierarchies may predispose the attendees to favoring this paradigm, many of the talks and discussions at the 2022 Model Hierarchies Workshop also highlighted the need to look beyond a linear hierarchy when discussing model organization and use. Within fluid dynamics, a hierarchy refers to the level of simplification of the governing equations and which processes and features (such as moisture and continents) are included or omitted. This abstraction is possible because the simplified forms of complex fluid dynamical processes have long been known to retain fundamental information about physical truths. The validity of this reductionism, however, is less certain for areas of study which inherently bring together multiple interacting processes, and for applications where model results are only useful when they speak to specific impacts. An "ecosystem" of models could be an alternative path forward. These ideas came up the most when discussing impacts modeling, but an expanded organizational paradigm may also help to tie the use of ML into the traditional modeling framework. Due to the common need for interpretability and generalizability among both scientific users and decision makers, there is a demand for the hierarchy to provide models built on established theory as well as on data.

There was a clear sense from the workshop attendees that some type of model taxonomy is necessary, but whether we need a single unified paradigm or whether different applications may benefit from varying approaches is an ongoing topic of discussion. We do not pretend to have a definitive answer, but we hope that the community will continue to reflect on this topic, and on how the organization of model types and complexities can help shape the types of scientific questions we ask and their societal relevance.

Data Availability Statement

Data were not used, nor created for this research.

Acknowledgments
LAM, AG, BG, and AS acknowledge support from Schmidt Futures, a philanthropic initiative founded by Eric and Wendy Schmidt, as part of the Virtual Earth System Research Institute (VESRI). ACB, CW, and AS acknowledge support from the National Science Foundation through Grant OAC-2004492. BG and AS acknowledge support from Google LLC. We thank the World Climate Research Programme, the National Science Foundation, and Stanford University for their support towards the Model Hierarchies Workshop, and Niku Darafshi for assistance with organizing. We acknowledge V. Balaji, Alison Wing, Tapio Schneider, Kelly Hereid, Pier Luigi Vidale, R. Saravanan, and Pedram Hassanzadeh for several stimulating conversations before, during, and after the workshop. We thank all speakers at the workshop: Eviatar Bach, Michela Biasutti, Ashesh Chattopadhyay, Gang Chen, Po-Chun Chung, Mariana Clare, Pedram Hassanzadeh, Chengfei He, Kelly Hereid, Tien-Yiao Hsu, Nadir Jeevanjee, Wenwen Kong, Glenn Liu, Laura Mansfield, Brian Medeiros, Oliver Mehling, Gianluca Meneghello, Timothy Merlis, Ivan Mitevski, Andre Nogueira Souza, Lettie Roach, Richard Rood, R. Saravanan, Gavin Schmidt, Tapio Schneider, Ashwin Seshadri, Tiffany Shaw, Zhaoyi Shen, Isla Simpson, Adam Sobel, David Stainforth, Andrew Stewart, Joao Teixeira, Pier Luigi Vidale, Lei Wang, Yujie Wang, Duncan Watson-Parris, Allison Wing, Shaocheng Xie, Emily Zakem, Yi Zhang, Elisa Ziegler. We also thank all workshop participants for their engagement. The complete workshop schedule can be found in the Supplementary. We are grateful to the editor, reviewers Nadir Jeevanjee, Adam Sobel, and one anonymous reviewer for their insightful comments, which have considerably improved the manuscript.

References

Andersson, T. R., Hosking, J. S., Pérez-Ortiz, M., Paige, B., Elliott, A., Russell, C., et al. (2021). Seasonal Arctic sea ice forecasting with probabilistic deep learning. *Nature Communications*, *12*(1), 5124. https://doi.org/10.1038/s41467-021-25257-4

Andre Perkins, W., Brenowitz, N. D., Bretherton, C. S., & Nugent, J. M. (2023). Emulation of cloud microphysics in a climate model. *Authorea Preprints*. https://doi.org/10.22541/essoar.168614667.71811888/v1

Arcomano, T., Szunyogh, I., Pathak, J., Wikner, A., Hunt, B. R., & Ott, E. (2020). A machine learning-based global atmospheric forecast model. *Geophysical Research Letters*, *47*(9), e2020GL087776. https://doi.org/10.1029/2020GL087776

Balaji, V. (2021). Climbing down Charney's ladder: Machine learning and the post-Dennard era of computational climate science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *379*(2194), 20200085. https://doi.org/10.1098/rsta.2020.0085

Barnes, E. A., Barnes, R. J., Martin, Z. K., & Rader, J. K. (2022). This looks like that there: Interpretable neural networks for image tasks when location matters. *Artificial Intelligence for the Earth Systems*, *1*(3). https://doi.org/10.1175/AIES-D-22-0001.1

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bauer, P., Stevens, B., & Hazeleger, W. (2021). A digital twin of Earth for the green transition. *Nature Climate Change*, *11*(2), 80–83. https://doi.org/10.1038/s41558-021-00986-y

Beck, S. (2011). Moving beyond the linear model of expertise? IPCC and the test of adaptation. *Regional Environmental Change*, *11*(2), 297–306. https://doi.org/10.1007/s10113-010-0136-2

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P. (2021a). Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, *126*(9), 098302. https://doi.org/10.1103/PhysRevLett.126.098302

Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., et al. (2021b). Climate-invariant machine learning. https://doi.org/10.48550/arXiv.2112.08440

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2022). Pangu-weather: A 3D high-resolution model for fast and accurate global weather forecast. https://doi.org/10.48550/arXiv.2211.02556

Blanchard, A., Parashar, N., Dodov, B., Lessig, C., & Sapsis, T. (2022). A multi-scale deep learning framework for projecting weather extremes. arXiv preprint arXiv:2210.12137v1. https://doi.org/10.48550/arXiv.2210.12137

Bony, S., Stevens, B., Held, I. H., Mitchell, J. F., Dufresne, J.-L., Emanuel, K. A., et al. (2013). Carbon dioxide and climate: Perspectives on a scientific assessment. In G. R. Asrar & J. W. Hurrell (Eds.), *Climate science for serving society: Research, modeling and prediction priorities* (pp. 391–413). Springer Netherlands. https://doi.org/10.1007/978-94-007-6692-1_14

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357–4375. https://doi.org/10.1175/JAS-D-20-0082.1

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. https://doi.org/10.1029/2019MS001711

Briley, L. J., Ashley, W. S., Rood, R. B., & Krmenec, A. (2017). The role of meteorological processes in the description of uncertainty for climate change decision-making. *Theoretical and Applied Climatology*, *127*(3), 643–654. https://doi.org/10.1007/s00704-015-1652-2

Briley, L. J., Rood, R. B., & Notaro, M. (2021). Large lakes in climate models: A great Lakes case study on the usability of CMIP5. *Journal of Great Lakes Research*, *47*(2), 405–418. https://doi.org/10.1016/j.jglr.2021.01.010

Bulthuis, K., Arnst, M., Sun, S., & Pattyn, F. (2019). Uncertainty quantification of the multi-centennial response of the Antarctic ice sheet to climate change. *The Cryosphere*, *13*(4), 1349–1380. https://doi.org/10.5194/tc-13-1349-2019

Carslaw, K. S., Lee, L. A., Reddington, C. L., Pringle, K. J., Rap, A., Forster, P. M., et al. (2013). Large contribution of natural aerosols to uncertainty in indirect forcing. *Nature*, *503*(7474), 67–71. https://doi.org/10.1038/nature12674

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I., & Palmer, T. (2021). Machine learning emulation of gravity wave drag in numerical weather forecasting. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002477. https://doi.org/10.1029/2021MS002477

Chattopadhyay, A., Mustafa, M., Hassanzadeh, P., Bach, E., & Kashinath, K. (2021). Towards physically consistent data-driven weather forecasting: Integrating data assimilation with equivariance-preserving deep spatial transformers. https://doi.org/10.48550/arXiv.2103.09360

Chattopadhyay, A., Subel, A., & Hassanzadeh, P. (2020). Data-driven super-parameterization using deep learning: Experimentation with multiscale Lorenz 96 systems and transfer learning. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002084. https://doi.org/10.1029/2020MS002084

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems*, *32*. https://papers.nips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html

Chen, Z., Liu, H., Xu, C., Wu, X., Liang, B., Cao, J., & Chen, D. (2021). Modeling vegetation greenness and its climate sensitivity with deep-learning technology. *Ecology and Evolution*, *11*(12), 7335–7345. https://doi.org/10.1002/ece3.7564

Chevallier, F., Chéruy, F., Scott, N. A., & Chédin, A. (1998). A neural network approach for a fast and accurate computation of a longwave radiative budget. *Journal of Applied Meteorology and Climatology*, *37*(11), 1385–1397. https://doi.org/10.1175/1520-0450(1998)037<1385:ANNAFA>2.0.CO;2

Clare, M. C. A., Leijnse, T. W. B., McCall, R. T., Diermanse, F. L. M., Cotter, C. J., & Piggott, M. D. (2022). Multilevel multifidelity Monte Carlo methods for assessing uncertainty in coastal flooding. *Natural Hazards and Earth System Sciences*, *22*(8), 2491–2515. https://doi.org/10.5194/nhess-22-2491-2022

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., & Stuart, A. M. (2021). Calibrate, emulate, sample. *Journal of Computational Physics*, *424*, 109716. https://doi.org/10.1016/j.jcp.2020.109716

Cohen, J., Coumou, D., Hwang, J., Mackey, L., Orenstein, P., Totz, S., & Tziperman, E. (2019). S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *WIREs Climate Change*, *10*(2), e00567. https://doi.org/10.1002/wcc.567

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., et al. (2021). Process-based climate model development harnessing machine learning: I. A calibration tool for parameterization improvement. *Journal of Advances in Modeling Earth Systems*, *13*(3). https://doi.org/10.1029/2020MS002217

Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., et al. (2021). Advancing mathematics by guiding human intuition with AI. *Nature*, *600*(7887), 70–74. https://doi.org/10.1038/s41586-021-04086-x

Drake, H. F., & Henderson, G. (2022). A defense of useable climate mitigation science: How science can contribute to social movements. *Climatic Change*, *172*(1–2), 10. https://doi.org/10.1007/s10584-022-03347-6

Dunbar, O. R. A., Garbuno-Inigo, A., Schneider, T., & Stuart, A. M. (2021). Calibration and uncertainty quantification of convective parameters in an idealized GCM. *Journal of Advances in Modeling Earth Systems*, *13*(9). https://doi.org/10.1029/2020MS002454

Edwards, T. L., Nowicki, S., Marzeion, B., Hock, R., Goelzer, H., Seroussi, H., et al. (2021). Projected land ice contributions to twenty-first-century sea level rise. *Nature*, *593*(7857), 74–82. https://doi.org/10.1038/s41586-021-03302-y

Emanuel, K. (2011). Global warming effects on U.S. hurricane damage. *Weather, Climate, and Society*, *3*(4), 261–268. https://doi.org/10.1175/WCAS-D-11-00007.1

Espinosa, Z. I., Sheshadri, A., Cain, G. R., Gerber, E. P., & DallaSanta, K. J. (2022). Machine learning gravity wave parameterization generalizes to capture the QBO and response to increased CO_2. *Geophysical Research Letters*, *49*(8), e2022GL098174. https://doi.org/10.1029/2022GL098174

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Fiedler, T., Pitman, A. J., Mackenzie, K., Wood, N., Jakob, C., & Perkins-Kirkpatrick, S. E. (2021). Business risk and the emergence of climate analytics. *Nature Climate Change*, *11*(2), 87–94. https://doi.org/10.1038/s41558-020-00984-6

Gagne, D. J., II, 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. https://doi.org/10.1175/MWR-D-18-0316.1

Garg, S., Rasp, S., & Thuerey, N. (2022). WeatherBench probability: A benchmark dataset for probabilistic medium-range weather forecasting along with deep learning baseline models. http://arxiv.org/abs/2205.00865

Gentine, P., Eyring, V., & Beucler, T. (2021). Deep learning for the parametrization of subgrid processes in climate models. In *Deep learning for the Earth sciences* (pp. 307–314). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119646181.ch21

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. https://doi.org/10.1029/2018GL078202

Gettelman, A., & Rood, R. B. (2016). In *Demystifying climate models* (Vol. 2). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48959-8

Giorgi, F. (2019). Thirty years of regional climate modeling: Where are we and where are we going next? *Journal of Geophysical Research: Atmospheres*, *124*(11), 5696–5723. https://doi.org/10.1029/2018JD030094

Giorgi, F., & Gutowski, W. J., Jr. (2015). Regional dynamical downscaling and the CORDEX initiative. *Annual Review of Environment and Resources*, *40*(1), 467–490. https://doi.org/10.1146/annurev-environ-102014-021217

Glahn, H. R. (1964). An application of adaptive logic to meteorological prediction. *Journal of Applied Meteorology and Climatology*, *3*(6), 718–725. https://doi.org/10.1175/1520-0450(1964)003<0718:AAOALT>2.0.CO;2

Gualdi, B., Binet-Stéphan, E., Bahabi, A., Marchal, R., & Moncoulon, D. (2022). Modelling fire risk exposure for France using machine learning. *Applied Sciences*, *12*(3), 1635. https://doi.org/10.3390/app12031635

Guillaumin, A. P., & Zanna, L. (2021). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002534. https://doi.org/10.1029/2021MS002534

Guo, Z., Wang, M., Qian, Y., Larson, V. E., Ghan, S., Ovchinnikov, M., et al. (2014). A sensitivity analysis of cloud properties to CLUBB parameters in the single-column community atmosphere model (SCAM5). *Journal of Advances in Modeling Earth Systems*, *6*(3), 829–858. https://doi.org/10.1002/2014MS000315

Harder, P., Watson-Parris, D., Stier, P., Strassel, D., Gauger, N., & Keuper, J. (2022). Physics-informed learning of aerosol microphysics. *Environmental Data Science*, *1*, E20. https://doi.org/10.1017/eds.2022.22

Harrington, L. J., Schleussner, C.-F., & Otto, F. E. L. (2021). Quantifying uncertainty in aggregated climate change risk assessments. *Nature Communications*, *12*(1), 7140. https://doi.org/10.1038/s41467-021-27491-2

Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108. https://doi.org/10.1175/2009bams2607.1

Held, I. M. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society*, *86*(11), 1609–1614. https://doi.org/10.1175/BAMS-86-11-1609

Held, I. M., & Suarez, M. J. (1994). A proposal for the intercomparison of the dynamical cores of atmospheric general circulation models. *Bulletin of the American Meteorological Society*, *75*(10), 1825–1830. https://doi.org/10.1175/1520-0477(1994)075<1825:APFTIO>2.0.CO;2

Held v. State of Montana. (2020). *Proposed final facts and conclusions of law, No. CDV-2020-307 (Montana First Judicial District, Lewis and Clark County) Hon*. Kathy Seeley, Judge Presiding. Retrieved from https://climatecasechart.com/case/11091/

Hereid, K. A. (2022). Hurricane risk management strategies for insurers in a changing climate. In J. M. Collins & J. M. Done (Eds.), *Hurricane risk in a changing climate* (pp. 1–23). Springer International Publishing. https://doi.org/10.1007/978-3-031-08568-0_1

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., et al. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*(3), 589–602. https://doi.org/10.1175/BAMS-D-15-00135.1

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., et al. (2021). Process-based climate model development harnessing machine learning: II. Model calibration from single column to global. *Journal of Advances in Modeling Earth Systems*, *13*(6). https://doi.org/10.1029/2020MS002225

Hueffel, J. A., Sperger, T., Funes-Ardoiz, I., Ward, J. S., Rissanen, K., & Schoenebeck, F. (2021). Accelerated Dinuclear Palladium catalyst identification through unsupervised machine learning. *Science*, *374*(6571), 1134–1140. https://doi.org/10.1126/science.abj0999

IPCC. (2018). Global Warming of 1.5°C: An Intergovernmental Panel on Climate Change (IPCC) Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways. In V. Masson-Delmotte, P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, et al., (Eds.), *The context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*.

IPCC. (2021). In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, et al., (Eds.), *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*.

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J. (2021). Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence*, *3*(8), 667–674. https://doi.org/10.1038/s42256-021-00374-3

Jasanoff, S., & Wynne, B. (1998). Science and decision making: Human choice and climate change. In S. Rayner & E. L. Malone (Eds.), *Human choice and climate change 1: The societal framework* (pp. 1–87). Batelle Press.

Jeevanjee, N., Hassanzadeh, P., Hill, S., & Sheshadri, A. (2017). A perspective on climate model hierarchies. *Journal of Advances in Modeling Earth Systems*, *9*(4), 1760–1771. https://doi.org/10.1002/2017MS001038

Jézéquel, A., Dépoues, V., Guillemot, H., Rajaud, A., Trolliet, M., Vrac, M., et al. (2020). Singular extreme events and their attribution to climate change: A climate service–centered analysis. *Weather, Climate, and Society*, *12*(1), 89–101. https://doi.org/10.1175/WCAS-D-19-0048.1

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., et al. (2021). Physics-informed machine learning: Case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *379*(2194), 20200093. https://doi.org/10.1098/rsta.2020.0093

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, *63*(3), 425–464. https://doi.org/10.1111/1467-9868.00294

Kovachki, N. B., & Stuart, A. M. (2019). Ensemble Kalman inversion: A derivative-free technique for machine learning tasks. *Inverse Problems*, *35*(9), 095005. https://doi.org/10.1088/1361-6420/ab1c3a

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Chalikov, D. V. (2005). New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Monthly Weather Review*, *133*(5), 1370–1383. https://doi.org/10.1175/MWR2923.1

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., Tolman, H. L., & Belochitski, A. A. (2008). Neural network approach for robust and fast calculation of physical processes in numerical environmental models: Compound parameterization with a quality control of larger errors. *Neural Networks*, *21*(2), 535–543. https://doi.org/10.1016/j.neunet.2007.12.019

Labe, Z. M., & Barnes, E. A. (2021). Detecting climate signals using explainable AI with single-forcing large ensembles. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2021MS002464. https://doi.org/10.1029/2021MS002464

Lahsen, M. (2005). Technocracy, democracy, and U.S. climate politics: The need for demarcations. *Science, Technology & Human Values*, *30*(1), 137–169. https://doi.org/10.1177/0162243904270710

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., et al. (2022). GraphCast: Learning skillful medium-range global weather forecasting. https://doi.org/10.48550/arXiv.2212.12794

Lan, S., Li, S., & Shahbaba, B. (2022). Scaling up Bayesian uncertainty quantification for inverse problems using deep neural networks. *SIAM/ASA Journal on Uncertainty Quantification*, *10*(4), 1684–1713. https://doi.org/10.1137/21M1439456

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lemos, M. C., & Rood, R. B. (2010). Climate projections and their impact on policy and practice. *WIREs Climate Change*, *1*(5), 670–682. https://doi.org/10.1002/wcc.71

Liu, X.-Y., Sun, H., Zhu, M., Lu, L., & Wang, J.-X. (2022). Predicting parametric spatiotemporal dynamics by multi-resolution PDE structure-preserved deep learning. https://doi.org/10.48550/arXiv.2205.03990

Lopez-Gomez, I., Christopoulos, C., Langeland Ervik, H. L., Dunbar, O. R., Cohen, Y., & Schneider, T. (2022). Training physics-based machine-learning parameterizations with gradient-free ensemble Kalman methods. *Journal of Advances in Modeling Earth Systems*, *14*(8), e2022MS003105. https://doi.org/10.1029/2022ms003105

Luo, X., Nadiga, B. T., Park, J. H., Ren, Y., Xu, W., & Yoo, S. (2022). A Bayesian deep learning approach to near-term climate prediction. *Journal of Advances in Modeling Earth Systems*, *14*(10). https://doi.org/10.1029/2022MS003058

Maher, P., Gerber, E. P., Medeiros, B., Merlis, T. M., Sherwood, S., Sheshadri, A., et al. (2019). Model hierarchies for understanding atmospheric circulation. *Reviews of Geophysics*, *57*(2), 250–280. https://doi.org/10.1029/2018RG000607

Mahesh, A., O'Brien, T., Loring, B., Elbashandy, A., Boos, W., & Collins, W. (2023). Identifying atmospheric rivers and their poleward latent heat transport with generalizable neural networks: ARCNNv1. EGUsphere. [preprint]. https://doi.org/10.5194/egusphere-2023-763

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. (2022). Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K. R. Müller, & W. Samek (Eds.), *Beyond explainable artificial intelligence, lecture notes in computer science* (Vol. 13200, pp. 315–339). Springer. https://doi.org/10.1007/978-3-031-04083-2_16

McGovern, A., Bostrom, A., Davis, P., Demuth, J. L., Ebert-Uphoff, I., He, R., et al. (2022). NSF AI institute for research on trustworthy AI in weather, climate, and coastal oceanography (AI2ES). *Bulletin of the American Meteorological Society*, *103*(7), E1658–E1668. https://doi.org/10.1175/BAMS-D-21-0020.1

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. https://doi.org/10.1175/BAMS-D-18-0195.1

Michelangeli, P.-A., Vautard, R., & Legras, B. (1995). Weather regimes: Recurrence and quasi stationarity. *Journal of the Atmospheric Sciences*, *52*(8), 1237–1256. https://doi.org/10.1175/1520-0469(1995)052<1237:WRRAQS>2.0.CO;2

Mojgani, R., Chattopadhyay, A., & Hassanzadeh, P. (2022). Discovery of interpretable structural model errors by combining Bayesian sparse regression and data assimilation: A chaotic Kuramoto–Sivashinsky test case. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *32*(6), 061105. https://doi.org/10.1063/5.0091282

Monahan, A. H. (2000). Nonlinear principal component analysis by neural networks: Theory and application to the Lorenz system. *Journal of Climate*, *13*(4), 821–835. https://doi.org/10.1175/1520-0442(2000)013<0821:NPCABN>2.0.CO;2

Mooers, G., Pritchard, M., Beucler, T., Srivastava, P., Mangipudi, H., Peng, L., et al. (2022). Comparing storm resolving models and climates via unsupervised machine learning. https://doi.org/10.48550/arXiv.2208.11843

Muffett, C., & Feit, S. (2017). *Smoke and fumes: The legal and evidentiary basis for holding big oil accountable for the climate crisis*. The Center for International Environmental Law. https://www.ciel.org/wp-content/uploads/2019/01/Smoke-Fumes.pdf

Nabizadeh, E., Hassanzadeh, P., Yang, D., & Barnes, E. A. (2019). Size of the atmospheric blocking events: Scaling law and response to climate change. *Geophysical Research Letters*, *46*(22), 13488–13499. https://doi.org/10.1029/2019GL084863

National Research Council. (2011). *Informing an effective response to climate change* (p. 25). National Academies Press. https://doi.org/10.17226/12784

Nowack, P., Braesicke, P., Haigh, J., Abraham, N. L., Pyle, J., & Voulgarakis, A. (2018). Using machine learning to build temperature-based ozone parameterizations for climate sensitivity simulations. *Environmental Research Letters*, *13*(10), 104016. https://doi.org/10.1088/1748-9326/aae2be

O'Gorman, P. A., & Dwyer, J. G. (2018). 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. https://doi.org/10.1029/2018MS001351

Ortiz, P., Orescanin, M., Petkovic, V., Powell, S. W., & Marsh, B. (2022). Decomposing satellite-based classification uncertainties in large earth science datasets. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–11. https://doi.org/10.1109/TGRS.2022.3152516

Pathak, J., Hunt, B., Girvan, M., Lu, Z., & Ott, E. (2018a). Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical Review Letters*, *120*(2), 024102. https://doi.org/10.1103/PhysRevLett.120.024102

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., et al. (2022). FourCastNet: A global data-driven high-resolution weather model using adaptive fourier neural operators. https://doi.org/10.48550/arXiv.2202.11214

Pathak, J., Wikner, A., Fussell, R., Chandra, S., Hunt, B. R., Girvan, M., & Ott, E. (2018b). Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *28*(4), 041101. https://doi.org/10.1063/1.5028373

PCAST. (2023). *Extreme weather risk in a changing climate: Enhancing prediction and protecting communities*. Executive Office of the President of the United States of America. [President's Council of Advisors on Science and Technology] Retrieved from https://www.whitehouse.gov/wp-content/uploads/2023/04/PCAST_Extreme-Weather-Report_April2023.pdf

Perezhogin, P., Zanna, L., & Fernandez-Granda, C. (2023). Generative data-driven approaches for stochastic subgrid parameterizations in an idealized ocean model. arXiv preprint arXiv:2302.07984. https://doi.org/10.48550/arXiv.2302.07984

Pielke, J., & Roger, A. (2007). *The honest broker: Making sense of science in policy and politics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511818110

Raissi, M., Babaee, H., & Givi, P. (2019a). Deep learning of turbulent scalar mixing. *Physical Review Fluids*, *4*(12), 124501. https://doi.org/10.1103/PhysRevFluids.4.124501

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019b). 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. https://doi.org/10.1016/j.jcp.2018.10.045

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. https://doi.org/10.5194/gmd-13-2185-2020

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11). 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. https://doi.org/10.1073/pnas.1810286115

Rasp, S., & Thuerey, N. (2021). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, *13*(2), e2020MS002405. https://doi.org/10.1029/2020MS002405

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Reidmiller, D. R., Avery, C. W., Easterling, D. R., Kunkel, K. E., Lewis, K. L. M., Maycock, T. K., & Stewart, B. C. (2018). *Report-in-brief. Impacts, risks, and adaptation in the United States: The fourth national climate assessment, volume II*. U.S. Global Change Research Program. https://doi.org/10.7930/NCA4.2018.RiB

Roach, L. A., Eisenman, I., Wagner, T. J. W., Blanchard-Wrigglesworth, E., & Bitz, C. M. (2022). Asymmetry in the seasonal cycle of Antarctic sea ice driven by insolation. *Nature Geoscience*, *15*(4), 277–281. https://doi.org/10.1038/s41561-022-00913-6

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. https://doi.org/10.1038/s42256-019-0048-x

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Sanford, C., Kwa, A., Watt-Meyer, O., Clark, S., Brenowitz, N., McGibbon, J., & Bretherton, C. (2022). Improving the predictions of ML-corrected climate models with novelty detection. https://doi.org/10.48550/arXiv.2211.13354

Saravanan, R. (2021). *The climate Demon: Past, present, and future of climate prediction*. Cambridge University Press.

Scher, S. (2018). Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, *45*(22), 12616–12622. https://doi.org/10.1029/2018GL080704

Schneider, T., Behera, S., Boccaletti, G., Deser, C., Emanuel, K., Ferrari, R., et al. (2023). Harnessing AI and computing to advance climate modelling and prediction. *Nature Climate Change*, *13*(9), 887–889. https://doi.org/10.1038/s41558-023-01769-3

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), 12396–12417. https://doi.org/10.1002/2017GL076101

Shaw, T. A., & Smith, Z. (2022). The midlatitude response to polar sea ice loss: Idealized slab-ocean aquaplanet experiments with thermodynamic sea ice. *Journal of Climate*, *35*(8), 2633–2649. https://doi.org/10.1175/JCLI-D-21-0508.1

Shen, Z., Sridhar, A., Tan, Z., Jaruga, A., & Schneider, T. (2022). A library of large-eddy simulations forced by global climate models. *Journal of Advances in Modeling Earth Systems*, *14*(3). https://doi.org/10.1029/2021MS002631

Shepherd, T. G., Boyd, E., Calel, R. A., Chapman, S. C., Dessai, S., Dima-West, I. M., et al. (2018). Storylines: An alternative approach to representing uncertainty in physical aspects of climate change. *Climatic Change*, *151*(3), 555–571. https://doi.org/10.1007/s10584-018-2317-9

Sit, M., & Demir, I. (2019). Decentralized flood forecasting using deep neural networks. https://doi.org/10.48550/arXiv.1902.02308

Sobel, A. H. (2021). Useable climate science is adaptation science. *Climatic Change*, *166*(1), 8. https://doi.org/10.1007/s10584-021-03108-x

Song, H.-J., Roh, S., & Park, H. (2021). Compound parameterization to improve the accuracy of radiation emulator in a numerical weather prediction model. *Geophysical Research Letters*, *48*(20), e2021GL095043. https://doi.org/10.1029/2021GL095043

Sonnewald, M., & Lguensat, R. (2021). Revealing the impact of global heating on North Atlantic circulation using transparent machine learning. *Journal of Advances in Modeling Earth Systems*, *13*(8). https://doi.org/10.1029/2021MS002496

Sraj, I., Zedler, S. E., Knio, O. M., Jackson, C. S., & Hoteit, I. (2016). Polynomial chaos–based Bayesian inference of K-profile parameterization in a general circulation model of the tropical Pacific. *Monthly Weather Review*, *144*(12), 4621–4640. https://doi.org/10.1175/MWR-D-15-0394.1

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., et al. (2020). A deep learning approach to antibiotic discovery. *Cell*, *180*(4), 688–702. https://doi.org/10.1016/j.cell.2020.01.021

Sun, Y. Q., Hassanzadeh, P., Alexander, M. J., & Kruse, C. G. (2023). Quantifying 3D gravity wave drag in a library of tropical convection-permitting simulations for data-driven parameterizations. *Journal of Advances in Modeling Earth Systems*, *15*(5), e2022MS003585. https://doi.org/10.1029/2022MS003585

Swiss Re. (2021). Natural catastrophes in 2020: Secondary perils in the spotlight, but don't forget primary-peril risks. Retrieved from https://www.swissre.com/dam/jcr:ebd39a3b-dc55-4b34-9246-6dd8e5715c8b/sigma-1-2021-en.pdf

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Thompson, D. W. J., & Wallace, J. M. (2000). Annular modes in the extratropical circulation. Part I: Month-to-month variability. *Journal of Climate*, *13*(5), 1000–1016. https://doi.org/10.1175/1520-0442(2000)013<1000:AMITEC>2.0.CO;2

Thrasher, B., Wang, W., Michaelis, A., Melton, F., Lee, T., & Nemani, R. (2022). NASA global daily downscaled projections, CMIP6. *Scientific Data*, *9*(1), 262. https://doi.org/10.1038/s41597-022-01393-4

Tibau, X. A., Requena-Mesa, C., Reimers, C., Denzler, J., Eyring, V., Reichstein, M., & Runge, J. (2018). SupernoVAE: VAE based kernel PCA for analysis of spatio-temporal earth data. In *Proceedings of the 8th International Workshop on Climate Informatics: CI 2018* (pp. 73–76). University Corporation for Atmospheric Research. https://doi.org/10.5065/D6BZ64XQ

U.S. Securities and Exchange Commission. (2022). *The enhancement and standardization of climate-related disclosures for investors (SEC Proposed Rule Release No. 33-11042)*. SEC. Retrieved from https://www.sec.gov/rules/proposed/2022/33-11042.pdf

Valizadegan, H., Martinho, M. J., Wilkens, L. S., Jenkins, J. M., Smith, J. C., Caldwell, D. A., et al. (2022). ExoMiner: A highly accurate and explainable deep learning classifier that validates 301 new exoplanets. *The Astrophysical Journal*, *926*(2), 120. https://doi.org/10.3847/1538-4357/ac4399

Wang, P., Yuval, J., & O'Gorman, P. A. (2022). Non-local parameterization of atmospheric subgrid processes with neural networks. *Journal of Advances in Modeling Earth Systems*, *14*(10). https://doi.org/10.1029/2022MS002984

Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., et al. (2022). ClimateBench v1.0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, *14*(10). https://doi.org/10.1029/2021MS002954

Weyn, J. A., Durran, D. R., Caruana, R., & Cresswell-Clay, N. (2021). Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002502. https://doi.org/10.1029/2021MS002502

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., & Yamazaki, K. (2013). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dynamics*, *41*(7), 1703–1729. https://doi.org/10.1007/s00382-013-1896-4

Williamson, D. B., Blaker, A. T., & Sinha, B. (2017). Tuning without over-tuning: Parametric uncertainty quantification for the NEMO ocean model. *Geoscientific Model Development*, *10*(4), 1789–1816. https://doi.org/10.5194/gmd-10-1789-2017

Yang, B., Qian, Y., Lin, G., Leung, R., & Zhang, Y. (2012). Some issues in uncertainty quantification and parameter tuning: A case study of convective parameterization scheme in the WRF regional climate model. *Atmospheric Chemistry and Physics*, *12*(5), 2409–2427. https://doi.org/10.5194/acp-12-2409-2012

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), 3295. https://doi.org/10.1038/s41467-020-17142-3

Yuval, J., O'Gorman, P. A., & Hill, C. N. (2021). Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. *Geophysical Research Letters*, *48*(6), e2020GL091363. https://doi.org/10.1029/2020GL091363

Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, *47*(17), e2020GL088376. https://doi.org/10.1029/2020GL088376