Pushing the Frontiers in Climate Modeling and Analysis with Machine Learning

Veronika Eyring[1,2*†], William D. Collins[3,4*†], Pierre Gentine[5], Elizabeth A. Barnes[6], Marcelo Barreiro[7], Tom Beucler[8], Marc Bocquet[9], Christopher S. Bretherton[10], Hannah M. Christensen[11], Katherine Dagon[12], David John Gagne[12], David Hall[13], Dorit Hammerling[14], Stephan Hoyer[15], Fernando Iglesias-Suarez[1], Ignacio Lopez-Gomez[15,16], Marie C. McGraw[17], Gerald A. Meehl[12], Maria J. Molina[18,12], Claire Monteleoni[19,20], Juliane Mueller[21], Michael S. Pritchard[22,13], David Rolnick[23,24], Jakob Runge[25,26], Philip Stier[11], Oliver Watt-Meyer[10], Katja Weigel[2,1], Rose Yu[27] and Laure Zanna[28]

[1] Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany.
[2] University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany.
[3] Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
[4] University of California, Berkeley, Berkeley, CA, USA.
[5] Department of Earth and Environmental Engineering, Columbia University, New York, NY, USA.
[6] Dept. of Atmospheric Science, Colorado State University, Fort Collins, CO, USA.
[7] Institute of Physics, School of Sciences, University of the Republic, Montevideo, Uruguay.
[8] University of Lausanne, Lausanne, Switzerland.
[9] CEREA, École des Ponts and EdF R&D, Île-de-France, France.
[10] Allen Institute for Artificial Intelligence (AI2), Seattle, WA, USA.
[11] Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, Oxford, UK.
[12] NSF National Center for Atmospheric Research, Boulder, CO, USA.
[13] NVIDIA Corporation, Santa Clara, CA, USA.
[14] Colorado School of Mines, Golden, CO, USA.
[15] Google Research, Mountain View, CA, USA.
[16] Climate Modeling Alliance (CliMA), California Institute of Technology, Pasadena, CA, USA.
[17] Cooperative Institute for Research in the Atmosphere (CIRA), Colorado State University, Fort Collins, CO, USA.
[18] University of Maryland, College Park, MD, USA.
[19] Department of Computer Science, University of Colorado Boulder, Boulder, CO, USA.
[20] INRIA Paris, Paris, France.
[21] Computational Science Center, National Renewable Energy Laboratory, Golden, CO, USA.
[22] University of California, Irvine, Irvine, CA, USA.
[23] McGill University, Montreal, Canada.
[24] Mila - Quebec AI Institute, Montreal, Canada.
[25] Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Datenwissenschaften, Jena, Germany.
[26] Technische Universität Berlin, Berlin, Germany.
[27] University of California, San Diego, La Jolla, CA, USA.
[28] Courant Institute, New York University, New York, NY, USA.

[*]Corresponding author(s). E-mail(s): veronika.eyring@dlr.de; wdcollins@lbl.gov;
[†]These authors contributed equally to this work.

Abstract.

Climate modeling and analysis are facing new demands to enhance projections and climate information. We argue that now is the time to push the frontiers of Machine Learning (ML) beyond state-of-the-art approaches, not only by developing ML-based Earth system models with greater fidelity, but also by providing new capabilities through emulators for extreme event projections with large ensembles, enhanced detection and attribution methods for extreme events, and advanced climate model analysis and benchmarking. Utilizing this potential requires addressing key ML challenges, in particular generalization, uncertainty quantification, explainable artificial intelligence, and causality. This interdisciplinary effort requires bringing together ML and climate scientists, while also leveraging the private sector, to accelerate progress towards actionable climate science.

Keywords: Climate modeling and analysis, machine learning, extreme events, benchmarks, generalization, uncertainty quantification, explainable artificial intelligence, causal inference

Introduction

The World Climate Research Programme's Coupled Model Intercomparison Project (CMIP [1]) brings together multi-model climate projections to understand past, present and future climate changes. These simulations are performed with global coupled Earth system models (ESMs) that simulate the physical climate as well as biogeochemical cycles under a wide range of forcings, yet large uncertainties for example in precipitation remain [2]. This limits the models' ability to accurately project global and regional climate changes, as well as climate variability, extremes and their impacts on ecosystems on decadal and multi-decadal timescales. In addition, the ever-increasing volumes of data makes the detection and understanding of patterns of variability and extreme events difficult. New machine learning (ML) methods promise great potential to address these challenges.

Machine Learning (ML) for Earth system science is rapidly expanding, with ML methods already being applied to a wide range of weather prediction applications [3, 4], a broad swath of additional climate change questions [5], and in diverse solution domains, including mitigation, adaptation, tools for individual and collective action, education, and finance [6].

For climate modeling and analysis, we argue that breakthroughs with ML can be achieved in multiple ways, in particular by (a) the development of hybrid ESMs where physical modeling is integrated with ML to maintain physical consistency and harvest ML versatility [7–9]; (b) ML-based emulation, where ML can provide fast and robust climate information including extreme event projections, allowing us to assess the envelope of recent weather possibilities; (c) ML-based detection and attribution of extreme events, where ML can advance understanding of the physical processes that underlie extreme occurrences; and (d) ML-enhanced climate model analysis and understanding of the Earth system, where ML can deliver powerful tools for analyzing high-dimensional datasets which are especially prevalent in Earth sciences, including the development of benchmarks [10, 11]. While ML has already made significant contributions to all these grand challenges, substantial advances in ML methods are required to fully exploit the potential of ML for climate modeling and analysis. These particularly include physical consistency of hybrid models that demonstrate the ability to realistically extrapolate to unseen climate regimes [12], uncertainty quantification [13], eXplainable Artificial Intelligence (XAI) to move away from ML as a black box [14], and causal inference methods that allow even more information to be extracted from Earth system data on how processes interact causally [15, 16].

In this Perspective, we focus on these key grand challenges in climate modeling and analysis that can be substantially improved with ML and discuss the fundamental advances in ML techniques that are required to advance across these grand challenges as schematically displayed in Figure 1 and summarized in Table 1. We also provide a perspective on remaining gaps, opportunities, and promising future directions. We argue that in order to achieve the full potential of ML for improved climate modeling and analysis, collaboration between academia and the private sector will be essential (see Box 1).

Figure 1. Schematic how ML can advance climate modeling and analysis. Each of these key sectors are discussed in this Perspective. While progress has been made, the full potential of ML for climate modeling and analysis remains to be reached as it is outlined in this Perspective.

Box 1. Collaboration between Academia and the Private Sector

Collaboration between academia and the private sector is crucial for advancing climate research and enhancing technology transfer. Collaborative projects may serve dual purposes - contributing to public service initiatives and commercial applications. For instance, academic research may lead to the development of ML-enhanced climate models that aid public policymakers in making informed decisions. Simultaneously, the private sector may leverage these models to create specialized climate services for industries such as energy, transport, or agriculture for commercial applications and solutions. Clear governance mechanisms respecting national and international laws need to be set in place that strike a balance between the public good derived from research outcomes and the proprietary interests of businesses while delineating responsibilities, addressing intellectual property concerns, defining licenses, data and code security, and privacy, as well as ensuring the ethical use of ML models.

Open data and source code initiatives play a pivotal role in fostering collaboration between academia and the private sector to accelerate progress in climate modeling and analysis with ML. This collaborative environment can be reinforced through joint publications on results, code or data descriptions, for example as in [11]. Academic institutions, often the creators of valuable climate datasets, contribute significantly by opening access to their data, fostering collaborative research. Simultaneously, the private sector's participation is facilitated by sharing proprietary datasets or tools, establishing a mutually beneficial exchange of information. Open source code, inherently distributed with licenses allowing users to freely view, use, modify, and distribute the source code, is a cornerstone of collaborative efforts. For the private sector, collaboration can become challenging in projects with copyleft (e.g., GNU General Public License (GPL)) or non-commercial (e.g., Creative Commons Attribution CC-BY-NC) licenses. Therefore, non-copyleft licenses, like Massachusetts Institute of Technology (MIT), Berkeley Software Distribution (BSD), or Apache Version 2.0, are preferred. These licenses, without restrictions for commercial use, offer flexibility for developers and organizations to choose how they use and distribute software, even incorporating it into proprietary projects. Open data and code initiatives not only facilitate seamless access, use, and contribution to tools and data but also foster transparency and innovation through shared code repositories, contributing to advancements beyond the state-of-the-art. The concept of ownership in the traditional sense is somewhat different in the context of open source software, as the collaborative nature of open source development allows multiple contributors to participate in shaping and enhancing the codebase. Still contributors might want to retain copyright to their specific contributions. The consortium developing the open source software should define the management, contributions, and utilization of the open source code within the consortium as well as Intellectual Property Rights (IPRs) for the specific contributions. Crucial considerations also include specifying the open source license, implementing Contributor License Agreements (CLAs) to define contribution terms, establishing governance structures for code decisions, and assigning responsibilities for code maintenance. Compatibility with consortium goals should be emphasized, ensuring alignment with the chosen open source license and integration into the collaborative project.

Collaborations that do not require further research or need to protect know-how are often performed without formal contracts. Otherwise, several governance models exist for collaborations between academia and the private sector. For example, a Non-Disclosure Agreement (NDA), which is a legal contract outlining the terms under which one party discloses confidential information to another, with the expectation that the recipient will not disclose the information to third parties, might be chosen during the phase of exploring possible collaboration opportunities while already exchanging ideas. For unfunded collaborations, a collaboration agreement can be established, defining a common research goal with all parties contributing research activities in roughly equal shares. While background information remains the property of each party, jointly developed foreground, i.e., intellectual property that is collaboratively created by two or more parties, can usually not be solely owned by the industry partner in this case. This model proves advantageous when academia and the private sector share common interests, enabling the long-term development of relationships without immediate financial commitments. It is particularly suitable for exploratory or precompetitive research, fostering a shared exploration of new ideas and solutions with risks and benefits distributed among the partners. A funded research project aligns with long-term innovation goals and provides the necessary financial support for sustained research efforts. It might also be required if the private sector wants to own jointly developed foreground, which is usually not possible in a collaboration agreement. Consultancy services are chosen when specific expertise and rapid solutions are needed. Consultants offer targeted recommendations for implementing solutions, making them valuable for efficient project management and execution. Which governance model is actually selected depends on the collaborative objectives, timeframes, and the depth of expertise required and might also depend on national or international laws of the participating parties (e.g., European Union State aid law).

Several companies are currently using weather forecasting as a first application in this research field that enables easier validation of foundational ML technology than ML for Earth system and climate modeling. For example, Microsoft has built a general-purpose foundation model for weather and climate based on Vision Transformers [17]. NVIDIA is developing a global forecasting model based on Spherical Fourier Neural Operators [18], generative Artificial Intelligence (AI) methods for downscaling and channel synthesis at km-scales [19], and collaborations with climate scientists on open benchmarks for hybrid AI-physics climate modeling [11]. NVIDIA has also open sourced its workflows for training large scale global AI weather simulators, together with US national lab scientists (https://github.com/NVIDIA/modulus-makani), in addition to tools for probabilistically assessing and intercomparing such systems' predictions for open community assessment [20], as well as collaborating in the open domain on applications beyond weather to climate simulation (https://github.com/ai2cm/ace). DeepMind and Google are developing ML models for global weather forecasting [1, 2], and Google also uses ML to make operational flood forecasts [21].

Machine Learning for Climate Modeling and Analysis

ML has great potential to substantially enhance our understanding of the Earth system and to reduce uncertainties in climate projections. In this section, we discuss key approaches in which climate modeling and analysis could be substantially enhanced with ML, in particular hybrid Earth system modeling, emulation of climate model simulations, extreme event detection and attribution, and climate model analysis and benchmarking (see also Table 1).

Hybrid Earth System Modeling

Approaches in which ML methods are combined and integrated into classical climate models, so called hybrid models (see Figure 2), have been proposed to be able to address many of the long-standing systematic biases and challenges faced by classical climate models [7, 8, 22]. Hybrid ESMs can be an integral part of initiatives like CMIP and can enhance classical models at all scales as proposed by [9].

Figure 2. Schematic diagram for integrating ML with process-modeling. (a) Clouds, convection, and gravity waves where subgrid-scale atmospheric processes are learned from short high-resolution simulations with ML (from [7]). Similar approaches exist for the ocean. (b) Modeling biological regulation processes (opening of the stomatal “valves” controlling water vapor flux from the leaves) with a recurrent neural network further coupled to a diffusion model (from [8]).

ML-based hybrid modeling and subgrid-scale parameterizations have been developed for different Earth system components, with first promising results for the atmosphere, ocean, and land already being achieved. Here, we provide some examples.

For the **atmosphere**, the largest sources of uncertainties in climate projections stem from the representation of clouds, aerosols, and their interaction, with significant structural biases remaining for example for the simulation of precipitation [23]. Advances in computing now allow for global storm-resolving model simulations of months to a few years [24], but not century-long projections, while low-level clouds and aerosols will continue to depend on parameterizations for their representation [9]. In this context, ML-based parameterizations have been developed to represent subgrid-scale physics as simulated by higher resolution model simulations [25, 26], including stochastic parametrizations [27]. Hybrid modeling has also shown remarkable success in correcting structural errors stemming from unresolved atmospheric processes in the bias-correction setting, producing stable, accurate multi-year simulations across a range of climates [28]. Several challenges of these approaches were identified early on, such as poor out-of-climate generalization [29], instabilities caused by interactions with the resolved dynamics of the parent model, disparities between offline skill (ML parameterization performance on the test set) and online skill (i.e., hybrid model performance) [30], and the violation of conservation laws [29]. Solutions to several of these problems have since been proposed, including architecture-based constraints to ensure conservation laws [31], incorporating symmetry to improve generalization [32], coupled online learning to prevent instabilities and biases [33], input restrictions to improve stability [28], causally-informed deep learning to respect the underlying physical processes [16], data-driven equation discovery [34, 35], and the use of transfer learning and climate-invariant inputs to improve generalization [12]. Results from these efforts are extremely promising. For example, [16] showed that a coarse-scale hybrid model aquaplanet simulation could accurately represent the Intertropical Convergence Zone and latitudinal patterns of precipitation and net radiation as represented by the high-resolution simulation (Figure 3).

Figure 3. Potential for reducing systematic errors in hybrid Earth system models. Zonal average climatologies of precipitation (left) and net radiative fluxes at the top of the atmosphere (right). Note that the causally-informed neural network simulation (red line) clearly captures both, zonal-mean precipitation (within the 95 confidence interval), and its variability (red dashed line) compared to the high-resolution SPCAM (Super parameterized Community Atmospheric model) simulation (black dashed line). Adapted from [16].

For the **ocean**, large uncertainties remain due to mesoscale eddies and other turbulent processes that are not fully resolved in most climate models [36]. Mesoscale eddies are turbulent features that play a key role in tracer transport, ocean heat uptake and thermosteric sea level changes [6]. To correctly capture the effect of ocean turbulence forcing on the large-scale and reduce associated uncertainties in climate projections, hybrid modeling approaches have been introduced [37, 38]. A similar approach to that in the atmosphere is taken, where data-driven ML parameterizations are learned from high-resolution climate model simulations, to augment existing coarse-resolution simulations. In this context, momentum-conserving convolutional neural networks (CNNs) and equation discovery have been studied to capture the effects of ocean mesoscale onto the large-scale. CNNs are known to capture complex structures [37, 39, 40] while equation discovery facilitates interpretable models. The generalization ability is best for symbolic expressions generated by the equation discovery model or sparse regression [37, 41]. These models perform better than state-of-the-art physics-based negative viscosity energetically constrained methods. These results encourage further development of hybrid ML ocean models in the long term. In the short term, these approaches will allow us to distill simple algebraic forms from the data through equation discovery, rendering more manageable models, and allowing us to capture the true physics, improve our understanding, and formalize previously purely empirical equations [37, 41].

For **land**, uncertainties in the terrestrial carbon cycle, such as projections in the land carbon sink, remain a major challenge [42]. These uncertainties can in part be tackled by automated and systematic reduction of uncertainties in land model structure and parameters [43]. Compared to the atmosphere or ocean, there is no equivalent to high-resolution, high-fidelity simulations for the land component, such that the main ways to improve models are through process representation and the use of observations. Land processes are further complicated by the fact that extremes are critical to land carbon and water cycles, dominating interannual variability and also the long-term carbon sink [44]. Hybrid modeling for the land provides a unique opportunity to combine ML with physical constraints or laws to better simulate and project terrestrial processes [8, 45]. The power of hybrid modeling lies in its ability to utilize existing and new observational data, coupled with physical understanding constraining land processes across a range of time scales. Fast processes, such as photosynthesis, can be constrained by data and are a good target for ML-based parameterizations, while slow processes, such as carbon allocation, do not have frequent observations and thus need to rely on physical knowledge as they cannot be derived from data alone. The advantage of hybrid modeling is its capacity to extrapolate and generalize beyond the scope of the observational data. This approach was recently developed for estimating ecosystem evapotranspiration [46], where a hybrid model showed a greater ability to generalize during extreme events compared to a pure ML model. Other successful cases of hybrid modeling for the land have combined traditional hydrologic modeling with ML to increase skill in predicting flood risk [47] and groundwater flow [48]. An ML component was also integrated within a physical model to learn total water storage with a neural network [49]. While these studies show early success in employing hybrid modeling for the land, there are several important considerations for future work. First, capturing extreme events on land (e.g., wildfires, floods, droughts) in the context of a changing climate is a high priority [50]. Second, data availability, sparsity, and observational uncertainties remain ongoing issues for land modeling. Variations across land datasets, unequal geographic distributions, and spatial and climatic biases in observations are key challenges for the use of data at scale, potentially biasing the retrievals [51].

Hybrid modeling as described above also introduces new challenges such as stability after coupling [30], differences between offline and online behavior [30, 33], and generalizability. The latter describes the question whether the models will be able to accurately project warming and extremes when they were trained against the current climate, rather than future climates. There may be unknown physical processes arising and the distribution of the data is likely changing with climate change. Thus, it is necessary to understand when models diverge and fail and take corrective actions. More comprehensive detection, analyses and metrics regarding their out-of-climate generalization and performance beyond time-averaged errors (e.g., on extremes) are needed. Ideally, the community will increasingly draw on the advances made in interpretable and explainable ML and other ML challenges to further advance hybrid models as we further discuss below.

Emulation of Climate Model Simulations

For climate modeling, many challenges remain including the relationship of model error and resolution [52, 53] and limits on near-term predictability due to internal variability of the climate system [54]. The emulation of weather and climate models with ML has demonstrated potential to accelerate resolution of these challenges and has therefore become a rapidly evolving field [3, 4, 10, 55]. Those algorithms aim to emulate a physically-based weather or climate model at a small fraction of its cost. In significant part, this speed-up arises by eliminating the mathematical condition that higher spatial resolution requires shorter time steps governing classical models that solve the full equations of motion. Some important applications are the use of those emulators to generate massive weather forecast and climate projection ensembles in order to better capture internal variability. Since the number of emulated simulations is several orders of magnitude larger than in the initial weather or climate forecast models, this is opening unique perspectives in the assessment of extreme events or very rare events (1 or 99 percentiles of the distribution), which often cannot be captured by the tens of ensemble members in the weather forecast or climate models. There is hope that much larger ensembles generated with emulators could capture such very rare events. There are caveats to the use of those emulator-based ensembles, especially related to checking whether they correctly capture the distribution generated by the emulated chaotic physical model. Emulators can also be used to answer scientific questions that would require running many climate model simulations and would therefore be computationally infeasible. Applications include the characterization of extreme event evolution or sampling [10, 56] and the emulations of regional-scale events [57, 58]. Again, in this context, care needs to be taken to systematically check that the emulator respects both the physical response and statistics of the host physical model. Advancing beyond emulation, climate models and observations have been optimally merged using a technique called transfer learning to better predict El Niño [59] or to better project climate change [60]. Transfer learning can improve the accuracy of climate predictions and projections spanning the past to the future by reducing systematic errors and increasing correlation to key observables in the recent climate record.

Extreme Event Detection and Attribution

Low-likelihood high impact (LLHI) extremes are a class of phenomena where the high but unknown risks of significant and negative societal and environmental effects are mismatched with inconsistent evidence and limited consensus regarding how LLHIs will evolve under global warming [61]. Two of the major obstacles to reducing the uncertainty in how LLHIs will change in warmer climates are the need to objectively yet rapidly search through petabytes of climate model projections while simultaneously harmonizing across highly diverse methods for detecting these extremes [62]. ML exhibits considerable promise to address these challenges. Deep learning approaches have enabled training algorithms to find and track extremes in climate model output at exascale speeds [63], and ML methods have been successfully deployed to study a wide variety of severe weather [64]. In addition, projections of LLHI evolution accompanied by quantifiable and objective measures of uncertainty can be generated using threshold-free Bayesian detection methods calibrated with Markov Chain Monte-Carlo [65]. Extreme phenomena have been identified using human-expert-labeled datasets of tropical cyclones, atmospheric rivers, and weather fronts in climate model output combined with deep [66] and convolutional neural networks [67]. Topological data analysis combined with support vector machines provide a threshold-free method for identifying atmospheric rivers in climate projections produced under a wide range of horizontal resolutions and climate scenarios [68]. Persistent phenomena such as hurricanes can readily and accurately be tracked using convolutional long-short term memory methods [69]. ML can also provide insights into the physical drivers of extreme phenomena and how these drivers will change in future projections [70]. In addition, certain applications of deep learning methods have shown the capability of generalizing from present-day to future climatic conditions, provided an extensive hyperparameter grid search is performed to find appropriate model hyperparameters [71]. Successful demonstrations that physical mechanisms can be learned from data rather than prescribed include analyses of the extreme precipitation circulation patterns and strongly rotating thunderstorms [71]. ML algorithms have also been used to emulate classical downscaling methods to enhance the horizontal spatial resolution of climate model simulations [72]. ML methods are exhibiting substantial potential to considerably accelerate projections of extremes in warmer climates. Recent applications include prediction of heat waves [73] and droughts [74]. These approaches advance addressing several long-standing challenges involving LLHIs, including the difficulty of sampling LLHIs from observations and climate model simulations of insufficient duration, and biases in projecting LLHIs involving physical processes that are under-resolved or highly parameterized in ESMs.

Climate Model Analysis and Benchmarking

ML-based parametrizations that perform well in evaluations where they are not yet coupled online into the host ESM but rather trained, validated and tested offline on high-resolution model data, may exhibit surprising failure modes when coupled online within a climate model [30]. This all needs to be carefully tested. Tools such as the Earth System Model Evaluation Tool (ESMValTool [75]) facilitate the evaluation of ML-based online climate model simulations against Earth observations and other climate models. In addition, as ML for climate modeling efforts have matured, the community has recognized a growing need to develop metrics, datasets, and tools to benchmark ML performance in more rigorous and consistent ways [10, 11]. Another approach is data-centric AI, which focuses on how ML results can be improved by identifying ways to increase the quality and diversity of training data.

On the analysis side, climate networks reconstructed from statistical correlations of time series at grid points have been used together with measures from information theory to detect hidden structures in climate data [76]. ML has started to demonstrate its great potential to enhance climate model analysis through the application of causal inference, explainable Artificial Intelligence (XAI), non-linear multi-variate emergent constraints, and the development of more targeted observational products for model evaluation. Causal discovery algorithms learn causal dependencies beyond traditional correlation and regression methods [15]. Causal model evaluation compares causal dependencies as learned from observational data to the ones from climate models, thus enhancing process-oriented model evaluation [77, 78]. XAI can be applied to identify prototypical behavior linked to physics-based processes from images for Earth system science applications and with this provide a new approach for model evaluation [79]. ML methods have also been used to constrain uncertainties in multi-model projections based on process analysis and causal discovery [78] or the combination of emergent constraints on the global scale to reduce uncertainties on the regional scale [80], which is often more relevant for policymakers. In addition, ML-based approaches based on non-linear dimensionality reduction with variational autoencoders could help evaluating data intense high-resolution simulations [81].

Cross-cutting Challenges in ML Method Developments

Addressing key challenges in climate modeling and analysis with ML as discussed in the previous section does not only benefit from the application of current ML methods, but also requires addressing several challenges in ML method development that are shared by all these different applications. In this section, we focus on four ML challenges that have seen recent breakthroughs, but for which more work is needed in order to utilize full potential (see also Table 1). This particularly will require further progress in physical consistency and generalization, uncertainty quantification, explainable AI, and causal inference.

Physical Consistency and Generalization

Physical models are designed to be valid in a broad range of regimes, while ML models are usually trained to best fit a specific training set. Therefore, ML models can make inconsistent predictions when tested on out-of-distribution samples [12], such as warmer climates. There has been notable progress on making the quality of ML-based inference less sensitive to changes in the data, broadly referred to as robustness. Performance on outliers and extremes can be improved using custom losses that weigh extremes more without compromising mean predictions [82], or custom frameworks that normalize data using extreme value theory [83]. Physical consistency can be improved using custom losses that penalize physically-inconsistent predictions [84] or architectures that strictly enforce physical constraints [31, 34]. Overall, while improving robustness is application-dependent, we encourage conducting out-of-distribution tests over out-of-sample tests that are still independent and identically distributed with respect to the training data, addressing non-stationarity in the data if possible [12], and considering tests to ask whether the ML model can properly predict a causal intervention [15]. Making robustness tests a standard component of benchmark datasets for weather and climate would help establish the most generalizable ML frameworks on distinct cases, paving the way towards their routine use in climate science.

Uncertainty Quantification

Another challenge to be addressed in the ML space is uncertainty quantification of the predictive performance of ML models. Systematic uncertainties arise due to the choice of the ML model itself, and the variability of its predictions, e.g. due to the stochastic gradient descent methods used for training. Stochastic (aka statistical) uncertainty is also present due to noise in the data used for training, and the choice of predictive variables being an incomplete representation of the Earth system [54]. Therefore, even the best model of the Earth system cannot produce definitive predictions. However, stochastic and systemic uncertainty are not mutually exclusive and can be combined to address data sparsity and out-of-distribution generalization issues [85]. It is known that deep neural networks alone are not providing uncertainty estimates and tend to produce overconfident predictions. Therefore, uncertainty quantification is receiving growing interest in ML [86].

There are roughly two types of uncertainty quantification methods in deep learning. The first one focuses on robustness via employing parameterized distributions to describe stochastic uncertainty sampling over solutions to the loss minimization procedure during training or bootstrapping to approximate parent distributions. Perturbations are made to the inference procedure in initialization via deep ensemble [87], neural network weights via Monte Carlo dropout [86], and datasets via bootstrapping [88]. The other type is Bayesian such as variational autoencoders [89], which aims to model posterior beliefs of connection weights given the data. Bayesian methods are typically more robust in mean prediction, while confidence levels obtained from frequentist methods provide more extensive coverage over data variations [13].

Uncertainty quantification presents distinctive challenges for weather and climate projection. For weather forecasting, much progress has been made to ensemble forecasts, leading to increased forecast skills and more reliable probabilistic estimates. For climate projection, despite the effort in multi-model ensembles to quantify systematic uncertainty, the multi-scale nature of the system and its internal variability make it challenging to produce and validate reliable uncertainty estimates and risk assessments. Deep learning has also been used to create ensemble forecasts, including for medium-range weather systems [4], typically through Monte Carlo dropout [86] or deep ensembles [87]. Specifically, multiple deep learning models are trained by varying the dropout units or training data and then generate forecasts jointly. Recently, deep generative models have also been used for probabilistic forecasts [4, 90]. The accelerated inference enabled by deep learning emulators can in principle enable very large ensembles to quantify the uncertainty due to natural variability in weather forecasts, but also in climate projections [91].

eXplainable Artificial Intelligence

While most ML techniques have previously been viewed as a "black box," eXplainable Artificial Intelligence (XAI) methods have the potential to change how these tools are viewed and used in climate science by assisting scientists to determine whether the ML approach is obtaining the right answers for the right reasons [14]. XAI approaches are beginning to appear more frequently in ML climate studies, including for identifying sources of predictability within the climate system [92] and analyzing the physical impacts of climate change [71]. XAI methods can be used to ensure that neural network models are physically consistent with the true dynamics of the climate system [93]. Such model interpretation and visualization can help ML methods capture the physically salient aspects of a problem, operate within the limits of the training data, and help identify new scientific hypotheses [14]. For example, neural networks and their explainability tools can be harnessed to identify patterns of the forced signal within combined fields [94]. XAI can identify which oceanic patterns of sea surface temperature anomalies lead to the largest gains in predictability [95]. The applicability of XAI approaches originally trained for image classification are now being tested on climate prediction tasks. The sensitivity to the choice of XAI method and its specific parameters is still being resolved [96]. Furthermore, XAI methods are applied post-hoc to an otherwise "black box" model, and so, gaining insights from XAI into the decision-making process of the ML algorithm requires simplifications of the model itself [97]. As an alternative, scientists should therefore consider developing interpretable models which are built to incorporate the decision-making process explicitly into their structure in order to be completely understood by a human without the need for post-hoc methods [97].

Causal Inference

Standard ML methods, including deep learning, excel at learning highly non-linear statistical relationships from complex, large-scale datasets, and are being increasingly applied in Earth and environmental sciences [8]. However, research questions in climate science are often about causal relationships rather than purely statistical associations. Causal inference provides the theoretical foundations to utilize assumptions about the underlying system to answer causal questions from data [15]. Two main strands of causal inference are causal discovery, where the goal is to learn a qualitative causal graph from data, and causal effect estimation, where one assumes qualitative causal knowledge in the form of a graph and then quantifies the effect of hypothetical interventions, for instance, by utilizing causally informed ML models. Thus, causal inference complements ML well [35]. Causal methods have been employed in various contexts in climate science, see [15] for an in-depth overview.

Causal inference is currently used to tackle two major challenges in climate modeling and analysis. Firstly, causal models can inform subgrid-scale parameterizations in hybrid modeling to better respect the underlying physical processes in the ML model [16], which is crucial for modeling climate change. To this end, causal discovery [15] can be performed to estimate causal graphs from high-resolution models or observational data. This qualitative information can then help choosing which input variables to include in ML-based parametrizations, which is a formal way of feature selection. Second, causal inference can be used to evaluate and compare climate model output from projects such as CMIP [77, 78], with possible implications for reducing uncertainties of climate projections. Here the approach is to learn causal graphs separately from observational data as well as model output and then utilize graph comparison metrics to identify which physical models better simulate the causal relationships as learned from the observations. One may also directly assume a causal graph and compare the causal effect estimated.

Beyond the statistical challenges shared with pure ML methods, such as dealing with high-dimensional and spatially correlated data [35], the advantages and challenges of causal inference methods lie in the reliance on expert knowledge about the underlying system, from the presence of hidden confounders and the complexity of non-linear processes occurring across timescales, to the basic but often challenging problem of defining the causal variables of interest [15] or possible loss of causality when coarse-graining. More specifically, key challenges in causal inference, calling for advanced method development, are associated with the assumptions on which these methods often rest on: a) the data is generated from a causally stationary process when in practice many real-world processes are non-stationary; b) the data-generating causal model is acyclic, which may well not be true, especially, in the presence of feedback loops; and c) interdependencies are not coincidental but structural, and violations of this assumption may lead to incorrect conclusions [15, 35]. Tackling these challenges requires close collaboration between method developers and domain experts to define and incorporate assumptions into causal methods, as well as to develop benchmarks for evaluating methods on ground truth data [99]. If these challenges can be met, the primary advantages of causal methods lie in the intuitive interpretation of the causal graphs, their transparent way of stating assumptions, and their potential for better out-of-distribution performance, which increases trustworthiness in climate change projections.

Way Ahead

Innovative machine learning methods are rapidly providing new and transformative ways of modeling and projecting climate change and extracting information from massive data volumes. These are timely topics given the start of the IPCC's 7th Assessment cycle and the initiation of CMIP7. While the full potential of hybrid modeling will certainly not be reached in time for CMIP7 contributions, some proof-of-principle hybrid ESMs might well be ready to participate. This could include models where a subset of the physical or empirical parametrizations is replaced with ML-based parametrizations, for example for cloud cover and convection. The structure of CMIP is such that any climate model that can perform the DECK (Diagnostic, Evaluation and Characterization of Klima) and CMIP historical simulations can contribute to CMIP [1]. The upcoming CMIP7 ensemble can benefit from these developments to include some of these first ML-based hybrid ESMs, but also from the use of emerging ML techniques such as uncertainty quantification, XAI and causal inference to interpret simulations from these models in comparison to Earth observations. It will be important to benchmark the class of ML-based hybrid ESMs against classical climate models to assess potential improvements and to exploit ML-based non-linear multi-variate and transfer learning combined with other approaches to constrain uncertainties in climate projections with Earth observations.

ML shows great potential to improve ESMs by learning important subgrid-scale processes from high-resolution simulations and Earth observations, producing stable multi-year simulations with encouragingly small systematic errors. However, as we discussed in this Perspective, trust and generalizability of the ML models needs to be further improved by introducing climate invariant variables, physical constraints, or equation discovery and by further developing some of the main ML challenges including XAI, uncertainty quantification, and causality (see also Table 1). The increasing speed and fidelity of emulators will enable the creation of huge ensembles of hindcasts and forecasts. The unprecedented sampling of plausible but counterfactual climates could transform our understanding of the drivers and consequences of low-likelihood high-impact extremes. Stability in coupled-model simulations upon replacement of a numerical model component or parameterization with an ML-based parameterization, and improved coupled-model skill and projection capability, are benchmark activities that we foresee as being critically important as ML for climate continues to advance as a field.

To sustain this rapidly evolving field, different communities need to work together. The full potential of ML for climate modeling and analysis with ML can only be met in an interdisciplinary approach, where the climate science community works closely with the ML community. Beyond this collaboration, this will demand new collaboration opportunities to be seriously approached between academia and the private sector (see Box 1). As the ML community becomes more aware of the potential of algorithms in society-relevant climate and Earth system research, large technology companies are increasingly interested in applying their capabilities to climate via interdisciplinary research with climate scientists, who are either employed directly or collaborate from academia. Private sector research may also be a valuable element in the development of more computationally efficient and scalable climate models as well as the developments of digital twins of the Earth which are defined as "an information system that exposes users to a digital replication of the state and temporal evolution of the Earth system constrained by available observations and the laws of physics" by [100]. As these applications venture into the realm of unseen climates, input from academic domain experts will become increasingly essential, opening new opportunities for joint efforts to push the frontiers of climate science.

The use of ML to better understand, model, and project the Earth system is a challenging and promising research field with accelerating progress in the past five years. Additional research efforts could have a high impact both to advance science and to address topics of critical importance and high relevance for human society. These topics include the need for much more reliable and localized predictions of near-term global environmental change and projections of the many options for mitigating this change under investigation. With enhanced ML-based climate modeling and analysis capabilities as discussed in this Perspective, we can look forward to substantial advancement of Earth system sciences to accelerate scientific understanding, modeling, as well as projecting climate change towards desperately-needed actionable climate science.

Acknowledgments

The authors acknowledge the Aspen Global Change Institute (AGCI) for hosting a workshop on "Exploring the Frontiers in Earth System Modeling with Machine Learning and Big Data" in June 2022 as part of its traditionally landmark summer interdisciplinary sessions (https://www.agci.org/event/22s3). The workshop was funded by NASA's Earth Science Division, NOAA's Climate Program Office, NSF's Directorate for Geosciences (GEO), and the travel support from the World Climate Research Programme (WCRP). VE, PG and FIS were funded by the European Research Council (ERC) Synergy Grant "Understanding and Modelling the Earth System with Machine Learning (USMILE)" under the EU Horizon 2020 research and innovation program (Grant agreement #855187). VE and KW were also supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the Gottfried Wilhelm Leibniz Prize awarded to Veronika Eyring (# EY 22/2-1). WDC was supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy (contract #DE-AC02-05CH11231) and used resources of the National Energy Research Scientific Computing Center (NERSC), also supported by the Office of Science of the U.S. Department of Energy (DOE), contract #DE-AC02-05CH11231. PG acknowledges additional funding from the National Science Foundation (NSF) Science and Technology Center (STC) Learning the Earth with Artificial Intelligence and Physics (LEAP), Award #2019625-STC and DOE Advanced Scientific Computing Research (ASCR) program. Additional support was received for EAB: NSF #AGS-1749261 and U.S. Department of Energy supported by the Regional and Global Model Analysis (RGMA) program; CSB: Allen Institute for AI; HMC: Natural Environment Research Council #NE/P018238/1; KD: U.S. DOE #DE-SC0022070 and NSF #IA 1947282 and NSF #1852977; DJG: NSF #1852977 and NSF #ICER-2019758; DaH: NVIDIA; DoH: U.S. DOE #DE-FE0032311; MCM: NSF #ICER-2019758; GAM: U.S. DOE #DE-SC0022070 and NSF #IA 1947282; MJM: U.S. DOE #DE-SC0022070 and NSF #IA 1947282 and University of Maryland Grand Challenges Grants Program #GC17-2957817; CM: NSF 2153040; JM: U.S. DOE #DE-AC36-08GO28308; MSP: NSF LEAP # 2019625-STC and DOE #DE-SC0023368 and DOE #DE-SC0022255; DR: Canada CIFAR AI Chairs program; JR: EU Horizon 2020 research and innovation program XAIDA #101003469 and ERC Starting Grant CausalEarth #948112; PS: ERC project RECAP under the European Union's Horizon 2020 research and innovation program #724602 and FORCeS project under the European Union's Horizon 2020 research program #821205; OWM: Allen Institute for AI; KW: EU Horizon 2020 European Eddy RIch ESMs (EERIE) #101081383; RY: U.S. DOE #DE-SC0022255 and NSF #IIS-2146343; LZ: Schmidt Futures #G-23-65185 and NSF #GG017158-01. We thank Jeremy Snyder (LBNL) for his work to create Figure 1.

Declarations

Conflict of interest/Competing interests. The authors declare no competing interests.

Availability of data and materials. No datasets were generated or analyses during the current study.

Code availability. No code was developed, utilized, or directly implicated during the current study.

Authors' contributions. VE and WDC jointly led the writing of the perspective, coordinated and oversaw the execution of the entire study. They conceived the concept of pushing the frontiers of climate modeling and analysis with ML for this study with support of PG and contributed to all sections of the perspective. All co-authors contributed to the concept, drafting and writing of the manuscript and to the discussions.

References

[1] Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J., Taylor, K.E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geoscientific Model Development **9**(5), 1937–1958 (2016). https://doi.org/10. 5194/gmd-9-1937-2016

[2] Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J.P., Engelbrecht, F., Fischer, E., Fyfe, J.C., Jones, C., Maycock, A., Mutemi, J., Ndiaye, O., Panickal, S., Zhou, T.: Future Global Climate: Scenario-

Based Projections and Near Term Information. In: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Pean, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M.I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J.B.R., Maycock, T.K., Waterfield, T., Yelekci, O., Yu, R., Zhou, B. (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, pp. 553–672. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA (2021). Chap. 4. https://www.ipcc.ch/report/ar6/wg1/

[3] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., Ravuri, S., Ewalds, T., Alet, F., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Stott, J., Vinyals, O., Mohamed, S., Battaglia, P.: GraphCast: Learning skillful medium-range global weather forecasting. arXiv (2022). https://doi.org/10.48550/ARXIV.2212.12794

[4] Price, I., Sanchez-Gonzalez, A., Alet, F., Ewalds, T., El-Kadi, A., Stott, J., Mohamed, S., Battaglia, P., Lam, R., Willson, M.: GenCast: Diffusion-based ensemble forecasting for medium-range weather. arXiv (2023). https://doi.org/10.48550/arXiv.2312.15796

[5] Monteleoni, C., Schmidt, G.A., Alexander, F., Niculescu-Mizil, A., Steinhaeuser, K., Tippett, M., Banerjee, A., Blumenthal, M.B., Ganguly, A.R., Smerdon, J.E., Tedesco, M.: Climate informatics. In: Yu, T., Chawla, N., Simoff, S. (eds.) Computational Intelligent Data Analysis for Sustainable Development. Data Mining and Knowledge Discovery Series, pp. 81–126. CRC Press, Taylor & Francis Group, London (2013)

[6] Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A.S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A.S., Maharaj, T., Sherwin, E.D., Mukkavilli, S.K., Kording, K.P., Gomes, C.P., Ng, A.Y., Hassabis, D., Platt, J.C., Creutzig, F., Chayes, J., Bengio, Y.: Tackling Climate Change with Machine Learning. ACM Computing Surveys **55**(2), 1–96 (2022). https://doi.org/10.1145/3485128

[7] Gentine, P., Eyring, V., Beucler, T.: Deep Learning for the Parametrization of Subgrid Processes in Climate Models, pp. 307–314. John Wiley & Sons, Ltd, Chichester, West Sussex, UK (2021). Chap. 21. https://doi.org/10.1002/9781119646181.ch21

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

[9] **Eyring, V., Gentine, P., Camps-Valls, G., Lawrence, D.M., Reichstein, M.: AI-empowered next-generation multiscale climate modeling for mitigation and adaptation. Nature Geoscience (2024, accepted).**

[10] Watson-Parris, D., Rao, Y., Olivi´e, D., Seland, Ø., Nowack, P., Camps-Valls, G., Stier, P., Bouabid, S., Dewey, M., Fons, E., et al.: ClimateBench v1. 0: A benchmark for data-driven climate projections. Journal of Advances in Modeling Earth Systems, 2021–002954 (2022). https://doi. org/10.1029/2021MS002954

[11] Yu, S., Hannah, W.M., Peng, L., Bhouri, M.A., Gupta, R., Lin, J., Lu¨tjens, B., Will, J.C., Beucler, T., Harrop, B.E., Hillman, B.R., Jenney, A.M., Ferretti, S.L., Liu, N., Anandkumar, A., Brenowitz, N.D., Eyring, V., Gentine, P., Mandt, S., Pathak, J., Vondrick, C., Yu, R., Zanna, L., Abernathey, R.P., Ahmed, F., Bader, D.C., Baldi, P., Barnes, E.A., Behrens, G., Bretherton, C.S., Busecke, J.J.M., Caldwell, P.M., Chuang, W., Han, Y., Huang, Y., Iglesias-Suarez, F., Jantre, S., Kashinath, K., Khairoutdinov, M., Kurth, T., Lutsko, N.J., Ma, P.-L., Mooers, G., David Neelin, J., Randall, D.A., Shamekh, S., Subramaniam, A., Taylor, M.A., Urban, N.M., Yuval, J., Zhang, G.J., Zheng, T., Pritchard, M.S.: ClimSim: An open large-scale dataset for training high-resolution physics emulators in hybrid multi-scale climate simulators (2023) [cs.LG]. https://doi.org/10.48550/arXiv.2306.08754

[12] Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., Ahmed, F., O'Gorman, P.A., Neelin, J.D., Lutsko, N.J., Gentine, P.: Climate-Invariant Machine Learning (2021). https://doi.org/10.48550/arxiv.2112.08440

[13] Wu, D., Gao, L., Chinazzi, M., Xiong, X., Vespignani, A., Ma, Y.-A., Yu, R.: Quantifying Uncertainty in Deep Spatiotemporal Forecasting. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD '21, pp. 1841–1851. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10. 1145/3447548.3467325

[14] McGovern, A., Lagerquist, R., Gagne, D.J., Jergensen, G.E., Elmore, K.L., Homeyer, C.R., Smith, T.:

Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. Bulletin of the American Meteorological Society **100**(11), 2175–2199 (2019). https://doi.org/10.1175/bams-d-18-0195.1

[15] **Runge, J., Gerhardus, A., Varando, G., Eyring, V., Camps-Valls, G.: Causal inference for time series. Nature Reviews Earth & Environment 10, 2553 (2023). https://doi.org/10.1038/s43017-023-00431-y**

[16] Iglesias-Suarez, F., Gentine, P., Solino-Fernandez, B., Beucler, T., Pritchard, M., Runge, J., Eyring, V.: Causally-informed deep learning to improve climate models and projections. Journal of Geophysical Research: Atmospheres 129(4), 2023–039202 (2024). https://doi.org/10.1029/2023JD039202

[17] Nguyen, T., Brandstetter, e.J., Kapoor, A., Gupta, J.K., Grover, A.: ClimaX: A foundation model for weather and climate. arXiv (2023). https://doi.org/10.48550/ARXIV.2301.10343

[18] Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., Kashinath, K., Anandkumar, A.: Spherical fourier neural operators: Learning stable dynamics on the sphere. arXiv preprint arXiv:2306.03838 (2023)

[19] Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Kashinath, K., Kautz, J., Pritchard, M.: Residual Diffusion Modeling for Km-scale Atmospheric Downscaling. arXiv (2023). https://doi.org/10.48550/arXiv.2309.15214

[20] Brenowitz, N.D., Cohen, Y., Pathak, J., Mahesh, A., Bonev, B., Kurth, T., Durran, D.R., Harrington, P., Pritchard, M.S.: A Practical Probabilistic Benchmark for AI Weather Models. arXiv (2023). https://doi.org/10.48550/arXiv.2401.15305

[21] Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., Reich, O., Gilon, O., Maor, R., Timnat, S., Shechter, T., Anisimov, V., Gigi, Y., Levin, Y., Moshe, Z., Ben-Haim, Z., Hassidim, A., Matias, Y.: Flood forecasting with machine learning models in an operational framework. Hydrology and Earth System Sciences **26**(15), 4013–4032 (2022). https://doi.org/10.5194/hess-26-4013-2022

[22] Eyring, V., Mishra, V., Griffith, G.P., Chen, L., Keenan, T., Turetsky, M.R., Brown, S., Jotzo, F., Moore, F.C., van der Linden, S.: Reflections and projections on a decade of climate science. Nature Climate Change **11**(4), 279–285 (2021). https://doi.org/10.1038/s41558-021-01020-x

[23] Eyring, V., Gillett, N.P., Rao, K.M.A., Barimalala, R., Parrillo, M.B., Bellouin, N., Cassou, C., Durack, P.J., Kosaka, Y., McGregor, S., Min, S., Morgenstern, O., Sun, Y.: Human Influence on the Climate System. In: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Pean, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M.I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J.B.R., Maycock, T.K., Waterfield, T., Yelekci, O., Yu, R., Zhou, B. (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, pp. 423–552. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA (2021). Chap. 3. https://www.ipcc.ch/report/ ar6/wg1/

[24] Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C.S., Chen, X., Du¨ben, P., Judt, F., Khairoutdinov, M., Klocke, D., Kodama, C., Kornblueh, L., Lin, S.-J., Neumann, P., Putman, W.M., R¨ober, N., Shibuya, R., Vanniere, B., Vidale, P.L., Wedi, N., Zhou, L.: DYAMOND: the DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains. Prog Earth Planet Sci **6**(61) (2019). https://doi.org/10.1186/s40645-019-0304-z

[25] **Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., Yacalis, G.: Could Machine Learning Break the Convection Parameterization Deadlock? Geophysical Research Letters 45(11), 5742–5751 (2018). https://doi. org/10.1029/2018gl078202**

[26] Grundner, A., Beucler, T., Gentine, P., Iglesias-Suarez, F., Giorgetta, M.A., Eyring, V.: Deep Learning Based Cloud Cover Parameterization for ICON. Journal of Advances in Modeling Earth Systems **14**(12), 2021– 002959 (2022). https://doi.org/10.1029/2021MS002959

[27] Behrens, G., Beucler, T., Iglesias-Suarez, F., Yu, S., Gentine, P., Pritchard, M., Schwabe, M., Eyring, V.: Improving atmospheric processes in earth system models with deep learning ensembles and stochastic parameterizations. Journal of Advances in Modeling Earth Systems (2024 (submitted)). arXiv preprint https://arxiv.org/abs/2402.03079

[28] Bretherton, C.S., Henn, B., Kwa, A., Brenowitz, N.D., Watt-Meyer, O., McGibbon, J., Perkins, W.A., Clark, S.K., Harris, L.: Correcting Coarse-Grid Weather and Climate Models by Machine Learning From Global Storm-Resolving Simulations. Journal of Advances in Modeling Earth Systems **14**(2) (2022). https://doi.org/10.1029/2021ms002794

[29] Rasp, S., Pritchard, M.S., Gentine, P.: Deep learning to represent sub-grid processes in climate models. Proceedings of the National Academy of Sciences **115**(39), 9684–9689 (2018). https://doi.org/10.1073/pnas.1810286115

[30] Brenowitz, N.D., Henn, B., McGibbon, J., Clark, S.K., Kwa, A., Perkins, W.A., Watt-Meyer, O., Bretherton, C.S.: Machine Learning Climate Model Dynamics: Offline versus Online Performance. In: NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning (2020). https://doi.org/10.48550/arxiv.2011.03081

[31] Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., Gentine, P.: Enforcing Analytic Constraints in Neural Networks Emulating Physical Systems. Physical Review Letters **126**(9), 98302 (2021). https://doi.org/10.1103/PhysRevLett.126.098302

[32] Wang, R., Walters, R., Yu, R.: Incorporating Symmetry into Deep Dynamics Models for Improved Generalization. arXiv (2020). https://doi.org/10.48550/ARXIV.2002.03061

[33] Rasp, S.: 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 (2020). https://doi.org/10.5194/gmd-13-2185-2020

[34] Grundner, A., Beucler, T., Gentine, P., Eyring, V.: Data-driven equation discovery of a cloud cover parameterization. Journal of Advances in Modeling Earth Systems 16(3), 2023–003763 (2024). https://doi.org/10.1029/2023MS003763

[35] Camps-Valls, G., Gerhardus, A., Ninad, U., Varando, G., Martius, G., Balaguer-Ballester, E., Vinuesa, R., Diaz, E., Zanna, L., Runge, J.: Discovering causal relations and equations from data. Physics Reports **1044**, 1–68 (2023). https://doi.org/10.1016/j.physrep.2023.10.005

[36] Couldrey, M.P., Gregory, J.M., Boeira Dias, F., Dobrohotoff, P., Domingues, C.M., Garuba, O., Griffies, S.M., Haak, H., Hu, A., Ishii, M., Jungclaus, J., K¨ohl, A., Marsland, S.J., Ojha, S., Saenko, O.A., Savita, A., Shao, A., Stammer, D., Suzuki, T., Todd, A., Zanna, L.: What causes the spread of model projections of ocean dynamic sea-level change in response to greenhouse gas forcing? Climate Dynamics **56**(1), 155–187 (2021). https://doi.org/10.1007/s00382-020-05471-4

[37] Zanna, L., Bolton, T.: Data-Driven Equation Discovery of Ocean Mesoscale Closures. Geophysical Research Letters **47**(17) (2020). https://doi.org/10.1029/2020GL088376

[38] Zhang, C., Perezhogin, P., Gultekin, C., Adcroft, A., Fernandez-Granda, C., Zanna, L.: Implementation and Evaluation of a Machine Learned Mesoscale Eddy Parameterization into a Numerical Ocean Circulation Model (2023). https://doi.org/10.48550/arXiv.2303.00962

[39] Guillaumin, A.P., Zanna, L.: Stochastic-Deep Learning Parameterization of Ocean Momentum Forcing. Journal of Advances in Modeling Earth Systems **13**(9) (2021). https://doi.org/10.1029/2021MS002534

[40] Bolton, T., Zanna, L.: Applications of Deep Learning to Ocean Data Inference and Subgrid Parameterization. Journal of Advances in Modeling Earth Systems **11**(1), 376–399 (2019). https://doi.org/10.1029/ 2018MS001472

[41] Ross, A., Li, Z., Perezhogin, P., Fernandez-Granda, C., Zanna, L.: Benchmarking of machine learning ocean subgrid parameterizations in an idealized model. Journal of Advances in Modeling Earth Systems (2023). https://doi.org/10.1029/2022MS003258

[42] Friedlingstein, P., O'Sullivan, M., Jones, M.W., Andrew, R.M., Bakker, D.C.E., Hauck, J., Landschu¨tzer, P., Le Quere, C., Luijkx, I.T., Peters, G.P., Peters, W., Pongratz, J., Schwingshackl, C., Sitch, S., Canadell, J.G., Ciais, P., Jackson, R.B., Alin, S.R., Anthoni, P., Barbero, L., Bates, N.R., Becker, M., Bellouin, N., Decharme, B., Bopp, L., Brasika, I.B.M., Cadule, P., Chamberlain, M.A., Chandra, N., Chau, T.-T.-T., Chevallier, F., Chini, L.P., Cronin, M., Dou, X., Enyo, K., Evans, W., Falk, S., Feely, R.A., Feng, L., Ford, D.J., Gasser, T., Ghattas, J., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gu¨rses, O., Harris, I., Hefner, M., Heinke, J., Houghton, R.A., Hurtt, G.C., Iida, Y., Ilyina, T., Jacobson, A.R., Jain, A., Jarn´ıkov´a, T.,

Jersild, A., Jiang, F., Jin, Z., Joos, F., Kato, E., Keeling, R.F., Kennedy, D., Klein Goldewijk, K., Knauer, J., Korsbakken, J.I., K¨ortzinger, A., Lan, X., Lefevre, N., Li, H., Liu, J., Liu, Z., Ma, L., Marland, G., Mayot, N., McGuire, P.C., McKinley, G.A., Meyer, G., Morgan, E.J., Munro, D.R., Nakaoka, S.-I., Niwa, Y., O'Brien, K.M., Olsen, A., Omar, A.M., Ono, T., Paulsen, M., Pierrot, D., Pocock, K., Poulter, B., Powis, C.M., Rehder, G., Resplandy, L., Robertson, E., R¨odenbeck, C., Rosan, T.M., Schwinger, J., Seferian, R., Smallman, T.L., Smith, S.M., Sospedra-Alfonso, R., Sun, Q., Sutton, A.J., Sweeney, C., Takao, S., Tans, P.P., Tian, H., Tilbrook, B., Tsujino, H., Tubiello, F., van der Werf, G.R., van Ooijen, E., Wanninkhof, R., Watanabe, M., Wimart-Rousseau, C., Yang, D., Yang, X., Yuan, W., Yue, X., Zaehle, S., Zeng, J., Zheng, B.: Global carbon budget 2023. Earth System Science Data **15**(12), 5301–5369 (2023). https://doi.org/10.5194/essd-15-5301-2023

[43] Dagon, K., Sanderson, B.M., Fisher, R.A., Lawrence, D.M.: A machine learning approach to emulation and biophysical parameter estimation with the Community Land Model, version 5. Advances in Statistical Climatology, Meteorology and Oceanography **6**(2), 223–244 (2020). https://doi.org/10.5194/ascmo-6-223-2020

[44] Humphrey, V., Berg, A., Ciais, P., Gentine, P., Jung, M., Reichstein, M., Seneviratne, S.I., Frankenberg, C.: Soil moisture–atmosphere feedback dominates land carbon uptake variability. Nature **592**(7852), 65–69 (2021). https://doi.org/10.1038/s41586-021-03325-5

[45] Shen, C., Appling, A.P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C.J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., Rahmani, F., Song, Y., Beck, H.E., Bindas, T., Dwivedi, D., Fang, K., H¨oge, M., Rackauckas, C., Mohanty, B., Roy, T., Xu, C., Lawson, K.: Differentiable modelling to unify machine learning and physical models for geosciences. Nature Reviews Earth & Environment **4**(8), 552–567 (2023). https://doi.org/10.1038/ s43017-023-00450-9

[46] Zhao, W.L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., Lin, C., Li, X., Qiu, G.Y.: Physics-Constrained Machine Learning of Evapotranspiration. Geophysical Research Letters **46**(24), 14496–14507 (2019). https://doi.org/10.1029/2019GL085291

[47] Yang, T., Sun, F., Gentine, P., Liu, W., Wang, H., Yin, J., Du, M., Liu, C.: Evaluation and machine learning improvement of global hydrological model-based flood simulations. Environmental Research Letters **14**(11), 114027 (2019). https://doi.org/10.1088/1748-9326/ab4d5e

[48] Wang, N., Zhang, D., Chang, H., Li, H.: Deep learning of subsurface flow via theory-guided neural network. Journal of Hydrology **584**, 124700 (2020). https://doi.org/10.1016/j.jhydrol.2020.124700

[49] Kraft, B., Jung, M., Körner, M., Koirala, S., Reichstein, M.: Towards hybrid modeling of the global hydrological cycle. Hydrology and Earth System Sciences **26**(6), 1579–1614 (2022). https://doi.org/10.5194/ hess-26-1579-2022

[50] Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., Shen, C.: Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. Journal of Hydrology **603**, 127043 (2021). https://doi.org/10.1016/j.jhydrol.2021.127043

[51] Nathaniel, J., Liu, J., Gentine, P.: MetaFlux: Meta-learning global carbon fluxes from sparse spatiotemporal observations. Scientific Data **10**(1) (2023). https://doi.org/10.1038/s41597-023-02349-y

[52] Peherstorfer, B., Willcox, K., Gunzburger, M.: Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization. SIAM Review **60**(3), 550–591 (2018). https://doi.org/10.1137/ 16M1082469

[53] Cutajar, K., Pullin, M., Damianou, A., Lawrence, N., Gonz´alez, J.: Deep gaussian processes for multi-fidelity modeling. arXiv preprint arXiv:1903.07320 (2019). https://doi.org/10.48550/arXiv.1903.07320

[54] Delaunay, A., Christensen, H.M.: Interpretable Deep Learning for Probabilistic MJO Prediction. Geophysical Research Letters **49**(16), 2022–098566 (2022). https://doi.org/10.1029/2022GL098566

[55] Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., Anandkumar, A.: FourCastNet: Accelerating Global High-Resolution Weather Forecasting using Adaptive Fourier Neural Operators. arXiv (2022). https://doi.org/10.48550/ ARXIV.2208.05419

[56] Beusch, L., Gudmundsson, L., Seneviratne, S.I.: Emulating Earth system model temperatures with MESMER: from global mean temperature trajectories to grid-point-level realizations on land. Earth System

Dynamics **11**(1), 139–159 (2020). https://doi.org/10.5194/esd-11-139-2020

[57] Doury, A., Somot, S., Gadat, S., Ribes, A., Corre, L.: Regional climate model emulator based on deep learning: Concept and first evaluation of a novel hybrid downscaling approach. Climate Dynamics **60**(5-6), 1751– 1779 (2023). https://doi.org/10.1007/s00382-022-06343-9

[58] Quilcaille, Y., Gudmundsson, L., Beusch, L., Hauser, M., Seneviratne, S.I.: Showcasing MESMER-X: Spatially Resolved Emulation of Annual Maximum Temperatures of Earth System Models. Geophysical Research Letters **49**(17), 2022–099012 (2022). https://doi.org/10.1029/ 2022GL099012

[59] Ham, Y.-G., Kim, J.-H., Luo, J.-J.: Deep learning for multi-year enso forecasts. Nature **573**(7775), 568–572 (2019). https://doi.org/10.1038/ s41586-019-1559-7

[60] Immorlano, F., Eyring, V., de Gouville, T.l.M., Accarino, G., Elia, D., Aloisio, G., Gentine, P.: Transferring climate change knowledge. arXiv (2023). https://doi.org/10.48550/arXiv.2309.14780

[61] IPCC: Summary for Policymakers. Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, pp. 3–32. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA (2021). [Masson Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Pean, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekci, R. Yu, and B. Zhou (eds.)]. https://doi.org/10.1017/9781009157896.001

[62] O'Brien, T.A., Wehner, M.F., Payne, A.E., Shields, C.A., Rutz, J.J., Leung, L.-R., Ralph, F.M., Collow, A., Gorodetskaya, I., Guan, B., Lora, J.M., McClenny, E., Nardi, K.M., Ramos, A.M., Tome, R., Sarangi, C., Shearer, E.J., Ullrich, P.A., Zarzycki, C., Loring, B., Huang, H., Inda-Diaz, H.A., Rhoades, A.M., Zhou, Y.: Increases in Future AR Count and Size: Overview of the ARTMIP Tier 2 CMIP5/6 Experiment. Journal of Geophysical Research: Atmospheres **127**(6) (2022). https://doi.org/10. 1029/2021jd036013

[63] Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E., Mahesh, A., Matheson, M., Deslippe, J., Fatica, M., Prabhat, Houston, M.: Exascale Deep Learning for Climate Analytics. arXiv (2018). https://doi.org/10.48550/ARXIV.1810.01993

[64] Salcedo-Sanz, S., P´erez-Aracil, J., Ascenso, G., Del Ser, J., Casillas-Perez, D., Kadow, C., Fister, D., Barriopedro, D., Garcıa-Herrera, R., Restelli, M., Giuliani, M., Castelletti, A.: Analysis, Characterization, Prediction and Attribution of Extreme Atmospheric Events with Machine Learning: a Review. arXiv (2022). https://doi.org/10.48550/ ARXIV.2207.07580

[65] O'Brien, T.A., Risser, M.D., Loring, B., Elbashandy, A.A., Krishnan, H., Johnson, J., Patricola, C.M., O'Brien, J.P., Mahesh, A., Ramirez, S.A., Rhoades, A.M., Charn, A., Dıaz, H.I., and, W.D.C.: Detection of atmospheric rivers with inline uncertainty quantification: TECA-BARD v1.0.1. Geoscientific Model Development **13**(12), 6131–6148 (2020). https://doi.org/10.5194/gmd-13-6131-2020

[66] Prabhat, Kashinath, K., Mudigonda, M., Kim, S., Kapp-Schwoerer, L., Graubner, A., Karaismailoglu, E., von Kleist, L., Kurth, T., Greiner, A., Mahesh, A., Yang, K., Lewis, C., Chen, J., Lou, A., Chandran, S., Toms, B., Chapman, W., Dagon, K., Shields, C.A., O'Brien, T., Wehner, M., Collins, W.: ClimateNet: an expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. Geoscientific Model Development **14**(1), 107–124 (2021). https://doi.org/10.5194/gmd-14-107-2021

[67] Liu, Y., Racah, E., Prabhat, Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., Wehner, M., Collins, W.: Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets (2016). https://doi.org/10.48550/arXiv.1605.01156

[68] Muszynski, G., Kashinath, K., Kurlin, V., Wehner, M.: Topological data analysis and machine learning for recognizing atmospheric river patterns in large climate datasets. Geoscientific Model Development **12**(2), 613– 628 (2019). https://doi.org/10.5194/gmd-12-613-2019

[69] Kim, S., Kim, H., Lee, J., Yoon, S., Kahou, S.E., Kashinath, K., Prabhat, M.: Deep-Hurricane-Tracker: Tracking and Forecasting Extreme Climate Events. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1761–1769 (2019). https://doi.org/10.1109/WACV.2019.00192

[70] Molina, M.J., O'Brien, T.A., Anderson, G., Ashfaq, M., Bennett, K.E., Collins, W.D., Dagon, K., Restrepo, J.M., Ullrich, P.A.: A Review of Recent and Emerging Machine Learning Applications for Climate

Variability and Weather Phenomena. Artificial Intelligence for the Earth Systems (2023). https://doi.org/10.1175/AIES-D-22-0086.1

[71] Molina, M.J., Gagne, D.J., Prein, A.F.: A benchmark to test generalization capabilities of deep learning methods to classify severe convective storms in a changing climate. Earth Space Sci. **8**(9) (2021). https://doi.org/10.1029/2020ea001490

[72] Vandal, T., Kodra, E., Ganguly, A.R.: Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. Theoretical and Applied Climatology **137**(1-2), 557–570 (2018). https://doi.org/10.1007/s00704-018-2613-3

[73] Miloshevich, G., Cozian, B., Abry, P., Borgnat, P., Bouchet, F.: Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data. Physical Review Fluids **8**(4) (2023). https://doi.org/10.1103/physrevfluids.8.040501

[74] Prodhan, F.A., Zhang, J., Pangali Sharma, T.P., Nanzad, L., Zhang, D., Seka, A.M., Ahmed, N., Hasan, S.S., Hoque, M.Z., Mohana, H.P.: Projection of future drought and its impact on simulated crop yield over South Asia using ensemble machine learning approach. Science of The Total Environment **807**, 151029 (2022). https://doi.org/10.1016/j. scitotenv.2021.151029

[75] Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Br¨otz, B., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., Davin, E.L., Davini, P., Debeire, K., de Mora, L., Deser, C., Docquier, D., Earnshaw, P., Ehbrecht, C., Gier, B.K., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hardiman, S., Hassler, B., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Koldunov, N., Lejeune, Q., Lembo, V., Lovato, T., Lucarini, V., Massonnet, F., Mu¨ller, B., Pandde, A., Perez-Zan´on, N., Phillips, A., Predoi, V., Russell, J., Sellar, A., Serva, F., Stacke, T., Swaminathan, R., Torralba, V., Vegas-Regidor, J., von Hardenberg, J., Weigel, K., Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP. Geoscientific Model Development **13**(7), 3383–3438 (2020). https://doi.org/10.5194/ gmd-13-3383-2020

[76] Dijkstra, H., Hernandez-Garcia, E., Masoller, C., Barreiro, M.: Networks in Climate. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA (2019). https://doi.org/10.1017/9781316275757

[77] Karmouche, S., Galytska, E., Runge, J., Meehl, G.A., Phillips, A.S., Weigel, K., Eyring, V.: Regime-oriented causal model evaluation of Atlantic–Pacific teleconnections in CMIP6. Earth System Dynamics **14**(2), 309–344 (2023). https://doi.org/10.5194/esd-14-309-2023

[78] Nowack, P., Runge, J., Eyring, V., Haigh, J.D.: Causal networks for climate model evaluation and constrained projections. Nature Communications **11**(1) (2020). https://doi.org/10.1038/s41467-020-15195-y

[79] Barnes, E.A., Barnes, R.J., Martin, Z.K., Rader, J.K.: This Looks Like That There: Interpretable Neural Networks for Image Tasks When Location Matters. Artificial Intelligence for the Earth Systems **1**(3) (2022). https://doi.org/10.1175/AIES-D-22-0001.1

[80] Schlund, M., Eyring, V., Camps-Valls, G., Friedlingstein, P., Gentine, P., Reichstein, M.: Constraining Uncertainty in Projected Gross Primary Production With Machine Learning. Journal of Geophysical Research: Biogeosciences **125**(11) (2020). https://doi.org/10.1029/2019jg005619

[81] Mooers, G., Pritchard, M., Beucler, T., Srivastava, P., Mangipudi, H., Peng, L., Gentine, P., Mandt, S.: Comparing storm resolving models and climates via unsupervised machine learning. Sci Rep (2023). https://doi. org/10.1038/s41598-023-49455-w. https://arxiv.org/abs/2208.11843

[82] Lopez-Gomez, I., McGovern, A., Agrawal, S., Hickey, J.: Global Extreme Heat Forecasting Using Neural Weather Models. Artificial Intelligence for the Earth Systems **2**(1), 220035 (2023). https://doi.org/10.1175/AIES-D-22-0035.1

[83] Boulaguiem, Y., Zscheischler, J., Vignotto, E., van der Wiel, K., Engelke, S.: Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks. Environmental Data Science **1**, 5 (2022). https://doi.org/10.1017/eds.2022.4

[84] Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Kashinath, K., Mustafa, M., Tchelepi, H.A., Marcus, P.,

Prabhat, Anandkumar, A.: MeshfreeFlowNet: A Physics-Constrained Deep Continuous Space-Time Super-Resolution Framework. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. SC '20, pp. 1–15. IEEE Press, (2020). https://doi.org/10.1109/SC41405.2020.00013

[85] Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? arXiv (2017). https://doi.org/10.48550/ arXiv.1703.04977

[86] Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (2016). https://proceedings.mlr.press/v48/gal16.html

[87] Fort, S., Hu, H., Lakshminarayanan, B.: Deep ensembles: A loss landscape perspective. arXiv:1912.02757 (2019). https://doi.org/10.48550/ arXiv.1912.02757

[88] Osband, I., Blundell, C., Pritzel, A., Van Roy, B.: Deep Exploration via Bootstrapped DQN. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 29. Curran Associates, Inc., (2016). https://proceedings.neurips.cc/paper_files/paper/2016/file/8d8818c8e140c64c743113f563cf750f-Paper.pdf

[89] Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114 (2013). https://doi.org/10.48550/arXiv.1312. 6114

[90] Cachay, S.R., Zhao, B., Joren, H., Yu, R.: DYffusion: A Dynamics-informed Diffusion Model for Spatiotemporal Forecasting. arXiv (2023). https://doi.org/10.48550/arXiv.2306.01984

[91] Hall, D., Pathak, J., Kashinath, K., Pritchard, M., Messmer, P., Hariri, F., Brenowitz, N., Cohen, Y., Anandkumar, A., Posey, S.: NVIDIA's Earth-2: An Interactive Digital Twin of the Earth and its Subsystems. IUGG (2023)

[92] van Straaten, C., Whan, K., Coumou, D., van den Hurk, B., Schmeits, M.: Using Explainable Machine Learning Forecasts to Discover Subseasonal Drivers of High Summer Temperatures in Western and Central Europe. Monthly Weather Review **150**(5), 1115–1134 (2022). https://doi.org/10.1175/MWR-D-21-0201.1

[93] Mamalakis, A., Ebert-Uphoff, I., Barnes, E.A.: Explainable Artificial Intelligence in Meteorology and Climate Science: Model Fine-Tuning, Calibrating Trust and Learning New Science. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Mu¨ller, K., Samek, W. (eds.) xxAI - Beyond Explainable AI, pp. 315–339. Springer, Cham, Switzerland (2022). https://doi.org/10.1007/978-3-031-04083-2 16

[94] Rader, J.K., Barnes, E.A., Ebert-Uphoff, I., Anderson, C.: Detection of Forced Change Within Combined Climate Fields Using Explainable Neural Networks. Journal of Advances in Modeling Earth Systems **14**(7) (2022). https://doi.org/10.1029/2021ms002941

[95] Toms, B.A., Barnes, E.A., Hurrell, J.W.: Assessing Decadal Predictability in an Earth-System Model Using Explainable Neural Networks. Geophysical Research Letters **48**(12) (2021). https://doi.org/10.1029/2021gl093842

[96] Mamalakis, A., Barnes, E.A., Ebert-Uphoff, I.: Carefully Choose the Baseline: Lessons Learned from Applying XAI Attribution Methods for Regression Tasks in Geoscience. Artificial Intelligence for the Earth Systems **2**(1), 220058 (2023). https://doi.org/10.1175/AIES-D-22-0058.1

[97] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

[98] McGraw, M.C., Barnes, E.A.: Memory Matters: A Case for Granger Causality in Climate Variability Studies. Journal of Climate **31**(8), 3289– 3300 (2018). https://doi.org/10.1175/JCLI-D-17-0334.1

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

[100] National Academies of Sciences, Engineering, and Medicine,. Foundational Research Gaps and Future Directions for Digital Twins. Washington, DC: The National Academies Press (2024). https://doi.org/10.17226/26894

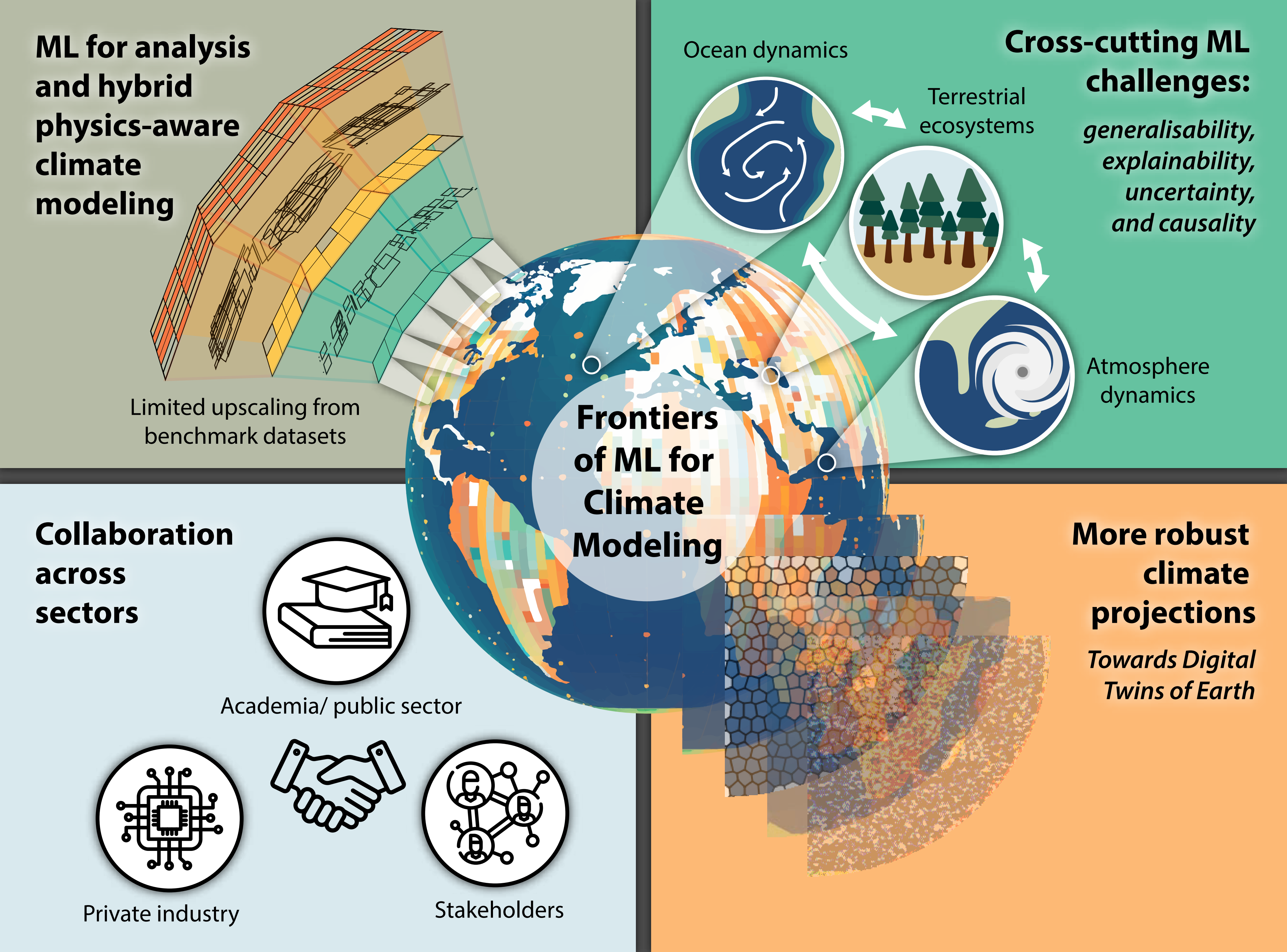
Table 1. Overview table summarizing challenges as well as potential ML-based solutions for (a) hybrid Earth system modeling, (b) emulation of climate model simulations, (c) extreme event detection and attribution, (d) climate model analysis and benchmarking, and (e) cross-cutting ML method developments, as discussed in this Perspective.

Challenges	**Potential of ML-based Solutions**
Hybrid Earth System Modeling	
Long-standing systematic errors and large uncertainties in climate projections in state-of-the-art climate and Earth system models	• Development of hybrid models, where physical modeling is integrated with ML to maintain physical consistency and harvest ML versatility [7–9]
Inhibitive computational expense of global storm-resolving model simulations, and dependence of coarser models on empirical parametrizations	• ML-based hybrid modeling and subgrid-scale parameterizations learning from higher resolution model simulations and Earth Observations [25, 26]
Poor out-of-climate generalization of hybrid models	• Incorporation of symmetries to improve generalization [32]; • Data-driven equation discovery [34,35]; • Transfer learning and climate-invariant inputs to improve generalization [12] • Symbolic expressions generated by the equation discovery model or sparse regression [37, 41] • Architecture-based physical constraints to ensure conservation laws [31]
Instabilities caused by interactions between ML parameterizations and resolved dynamics	• Causally-informed deep learning to respect the underlying physical processes [16]
Disparities between offline and online skill	• Coupled online learning to prevent instabilities and biases [33]
Violation of conservation laws	• Architecture-based physical constraints to ensure conservation laws [31] • Data-driven equation discovery with physical constraints [34] • Custom losses that penalize physically-inconsistent predictions [84]
Incorporation of the effects of mesoscale onto the large-scale	• Momentum-conserving CNNs [37,39,40]
Data availability, sparsity, and observational uncertainties/biases	• Meta-learning to learn new tasks from sparse data efficiently [51]
Constraints on processes across a range of time scales	• Combination of ML with physical constraints to simulate and project processes [8, 45]
Accurate simulation of extreme events	• More comprehensive analyses and metrics regarding the performance beyond time-averaged errors (e.g., on extremes) • Interpretable and explainable ML for understanding • Custom losses to weigh extremes more without compromising mean predictions [82] • Custom frameworks that normalize data using extreme value theory [83]
Emulation of Climate Model Simulations	
Uncertainty quantification	• Solutions to the trade-off between computational efficiency and prediction accuracy with multi-fidelity modeling, such as Gaussian or neural processes [52, 53] to combine simulation outputs and accelerate learning. • Use of ML-emulators to generate a massive ensemble of weather forecast and climate projection members to better capture internal weather and climate variability.
Separation of different sources of uncertainty	• Identification of physical conditions that affect prediction uncertainty based on a deep convolutional neural network forecast [54]

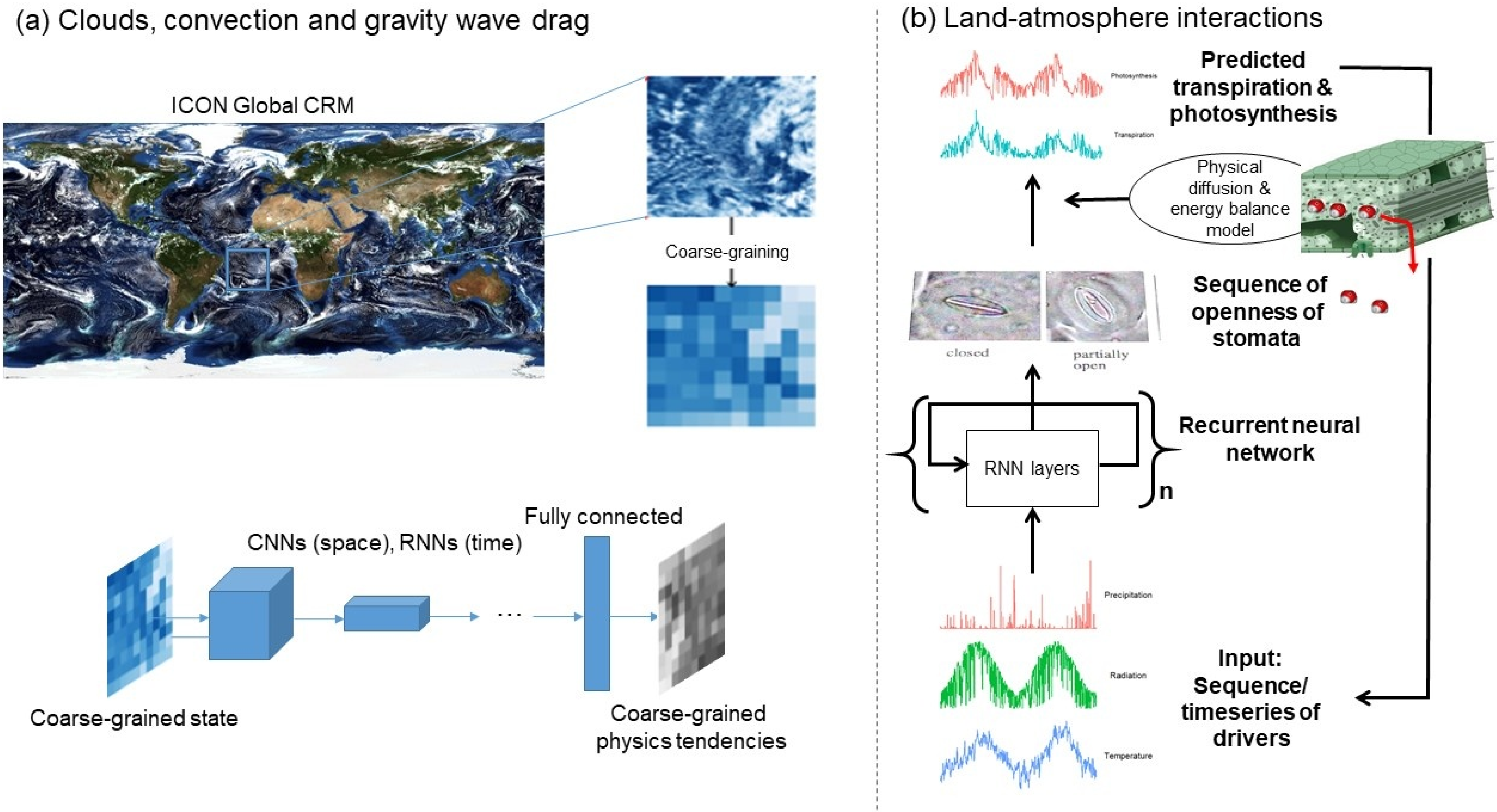
Sampling of very rare extreme or regional-scale events	● Larger ensembles generated with emulators [10, 56, 57, 58]
Improvement in projections and predictions	● Transfer learning [59, 60]
Extreme Event Detection and Attribution	
Objective and rapid searches through petabytes of climate model projections for detecting extremes	● Deep learning approaches for rapid detection [63]: ○ Human-labeled data sets combined with deep [66] and convolutional neural networks [67] ○ Convolutional long-short term memory methods [69] ● Quantifiable and objective measures with threshold-free methods: ○ Bayesian detection methods calibrated with Markov Chain Monte-Carlo [65] ○ Topological data analysis combined with support vector machines [68]
Harmonization of highly diverse methods of extreme event detection	● ML methods to study a wide variety of severe weather [64]
Generalization from present-day to future climatic conditions	● Derivation of insights from ML into the physical drivers of extreme phenomena and how these drivers will change in future projections [70] ● Extensive hyperparameter grid searches to find appropriate model hyperparameters can enable certain applications of deep learning methods to generalize from present-day to future climatic conditions [71].
Difficulty of sampling Low-likelihood high impact extremes from observations due to insufficient duration, or under-resolved or highly parameterized physical processes	● Emulation of classical downscaling methods with ML to enhance the horizontal spatial resolution of climate model simulations [72]. ● ML methods to considerably accelerate projections of extremes in warmer climates [73, 74]
Climate Model Analysis and Benchmarking	
Exhibition of surprising failure modes by ML models that perform well in offline test set evaluations when coupled within a climate model	● Evaluation of ML-based online climate model simulations against earth observations and other climate models, using tools like e.g. ESMValTool [75] ● Development of metrics, datasets, and tools to benchmark ML performance in more rigorous and consistent ways [10, 11] ● Data-centric AI to improve ML-results by identifying ways to increase the quality and diversity of training data.
Process-oriented model evaluation	● Causal model evaluation comparing causal dependencies as learned from observational data to the ones from climate models [77, 78]. ● XAI to identify prototypical behavior linked to physics-based processes from images for Earth system science applications [79].
Tighter constraints on uncertainties in multi-model projections	● Process analysis and causal discovery [78] ● Non-linear, multi-variable ML-based emergent constraints to reduce uncertainties for global and regional projections [80]
Availability and quality of Earth observations	● Use of ML methods to develop targeted observational products for model evaluation
Analysis and evaluation of data intense high-resolution simulations	● ML-based approaches based on non-linear dimensionality reduction with variational autoencoders [81] ● Climate networks reconstructed from statistical correlations of time series at grid points have been used together with measures from information theory to detect hidden structures in climate data [76].
Cross-cutting Challenges in ML Method Developments	
Physical consistency	● Custom losses that penalize physically-inconsistent predictions [84]

	● Architectures that strictly enforce physical constraints [31, 34]
Enhancement of robustness and generalization of ML predictions for out-of-distribution samples [12]	● Performance on outliers and can be improved using custom losses that weigh extremes more without compromising mean predictions [82] ● Custom frameworks that normalize data using extreme value theory [83]. ● Robustness tests addressing non-stationarity [12] and causal interventions [15]
Uncertainty Quantification	● Combination of aleatoric and epistemic uncertainty to address data sparsity and out-of-distribution generalization issues [85] ● Quantification of uncertainties through ○ Perturbations in the initialization via deep ensemble [87], neural network weights via Monte Carlo dropout [86], and datasets via bootstrapping [88]. ○ Bayesian methods e.g. for variational autoencoders [89]
Obtaining the right answers for the right reasons: eXplainable Artificial Intelligence (XAI)	● Identification and quantification of sources of predictability within the climate system [92, 95] ● Analysis of the physical impacts of climate change [71] ● Measures to ensure physical consistency with the true dynamics of the climate system [93]
Gaining insights from XAI into the decision-making process of the ML algorithm requires simplifications of the model itself	● Development of interpretable models which are built to incorporate the decision-making process explicitly into their structure to be understood without post-hoc methods [97]
Challenges for causal Inference: assumptions for methods may lead to incorrect conclusions [15, 35] e.g.: ● Assuming a causally stationary process when in practice many real-world processes are non-stationary ● Assumption of an acyclic causal model, which may not be true in the presence of feedback loops ● Structural rather than coincidental interdependencies	● Close collaboration between method developers and domain experts to define and incorporate assumptions into causal methods ● Development of benchmarks for evaluating methods on ground truth data [10, 11]

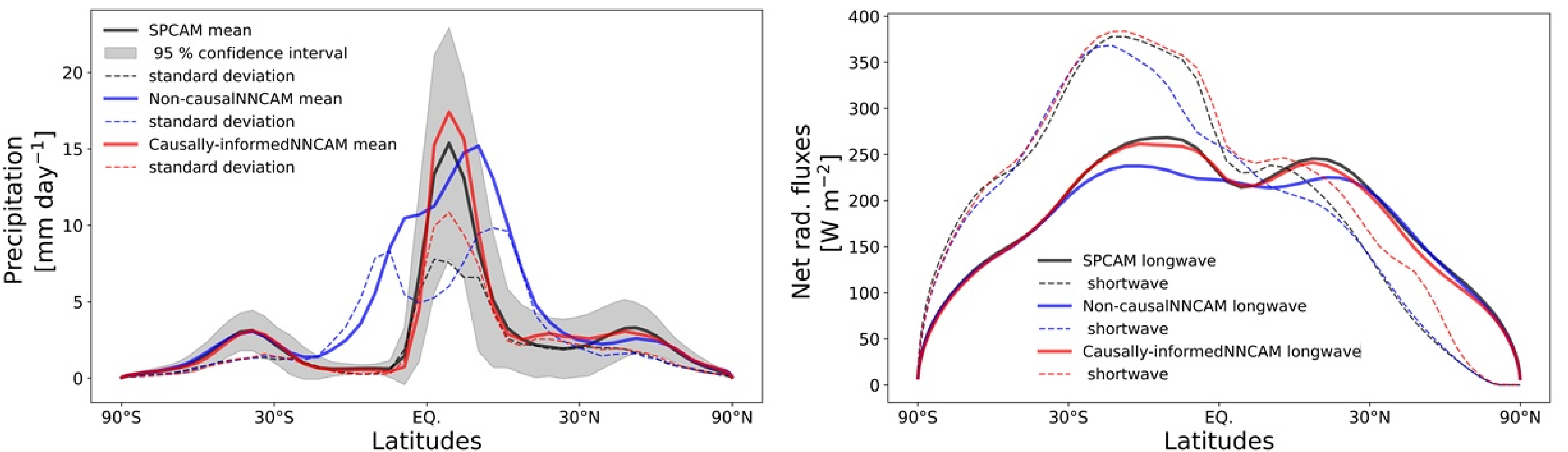
824



ML for analysis and hybrid physics-aware climate modeling
Limited upscaling from benchmark datasets
Ocean dynamics
Terrestrial ecosystems
Atmosphere dynamics
Cross-cutting ML challenges:
generalisability, explainability, uncertainty, and causality
Frontiers of ML for Climate Modeling
Collaboration across sectors
Academia/ public sector
Private industry
Stakeholders
More robust climate projections
Towards Digital Twins of Earth



(a) Clouds, convection and gravity wave drag
ICON Global CRM
Coarse-graining
CNNs (space), RNNs (time)
Fully connected
...
Coarse-grained state
Coarse-grained physics tendencies
(b) Land-atmosphere interactions
Photosynthesis
Transpiration
Predicted transpiration & photosynthesis
Physical diffusion & energy balance model
closed
partially open
Sequence of openness of stomata
RNN layers
n
Recurrent neural network
Precipitation
Radiation
Temperature
Input: Sequence/ timeseries of drivers


SPCAM mean
95 % confidence interval
standard deviation
Non-causalNNCAM mean
standard deviation
Causally-informedNNCAM mean
standard deviation
Precipitation [mm day^{-1}]
Latitudes
90°S
30°S
EQ.
30°N
90°N
SPCAM longwave
shortwave
Non-causalNNCAM longwave
shortwave
Causally-informedNNCAM longwave
shortwave
Net rad. fluxes [W m^{-2}]