Challenges and Benchmark Datasets for Machine Learning in the Atmospheric Sciences: Definition, Status, and Outlook

PETER D. DUEBEN,[a] MARTIN G. SCHULTZ,[b] MATTHEW CHANTRY,[a] DAVID JOHN GAGNE II,[c]
DAVID MATTHEW HALL,[d] AND AMY MCGOVERN[e]

[a] *European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom*
[b] *Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany*
[c] *National Center for Atmospheric Research, Boulder, Colorado*
[d] *NVIDIA Corporation, Santa Clara, California*
[e] *University of Oklahoma, School of Computer Science and School of Meteorology, Norman, Oklahoma*

(Manuscript received 10 November 2021, in final form 7 June 2022)

ABSTRACT: Benchmark datasets and benchmark problems have been a key aspect for the success of modern machine learning applications in many scientific domains. Consequently, an active discussion about benchmarks for applications of machine learning has also started in the atmospheric sciences. Such benchmarks allow for the comparison of machine learning tools and approaches in a quantitative way and enable a separation of concerns for domain and machine learning scientists. However, a clear definition of benchmark datasets for weather and climate applications is missing with the result that many domain scientists are confused. In this paper, we equip the domain of atmospheric sciences with a recipe for how to build proper benchmark datasets, a (nonexclusive) list of domain-specific challenges for machine learning is presented, and it is elaborated where and what benchmark datasets will be needed to tackle these challenges. We hope that the creation of benchmark datasets will help the machine learning efforts in atmospheric sciences to be more coherent, and, at the same time, target the efforts of machine learning scientists and experts of high-performance computing to the most imminent challenges in atmospheric sciences. We focus on benchmarks for atmospheric sciences (weather, climate, and air-quality applications). However, many aspects of this paper will also hold for other aspects of the Earth system sciences or are at least transferable.

SIGNIFICANCE STATEMENT: Machine learning is the study of computer algorithms that learn automatically from data. Atmospheric sciences have started to explore sophisticated machine learning techniques and the community is making rapid progress on the uptake of new methods for a large number of application areas. This paper provides a clear definition of so-called benchmark datasets for weather and climate applications that help to share data and machine learning solutions between research groups to reduce time spent in data processing, to generate synergies between groups, and to make tool developments more targeted and comparable. Furthermore, a list of benchmark datasets that will be needed to tackle important challenges for the use of machine learning in atmospheric sciences is provided.

KEYWORDS: Atmosphere; Neural networks; Numerical analysis/modeling; Uncertainty; Artificial intelligence; Bayesian methods; Data science; Decision trees; Deep learning; Machine learning; Neural networks; Other artificial intelligence/machine learning

1. Introduction

Many scientific domains have seen an unprecedented growth of machine learning tools and applications during the last decade. The toolbox of machine learning that allows learning complex nonlinear dynamics from data is also promising for many application areas within the Earth system sciences. This domain is data-rich—the European Centre for Medium-Range Weather Forecasts (ECMWF) has, for example, hundreds of petabytes of Earth-system-related data stored in its archive—but the Earth system is very large with many features and nonlinear, chaotic behavior resulting from a large variety of feedback processes on an extremely wide range of spatiotemporal scales. There has been a boom in machine learning for the Earth system sciences in recent years with many different applications of machine learning across all components of Earth system models—including atmosphere and atmospheric chemistry (Brenowitz and Bretherton 2018; Nowack et al. 2018), ocean (Sonnewald et al. 2021), land surface, sea ice and land ice (Andersson et al. 2021), severe weather (McGovern et al. 2019)—and across the entire workflow of weather and climate prediction models—observation processing (Aires et al. 2021), data assimilation (Brajard et al. 2020), forward models (Dueben and Bauer 2018), postprocessing, and dissemination (Gröenquist et al. 2021).

This recent boom has been fueled by the exponential growth of Earth system data over the last couple of decades (for both observation and model output data), the availability of powerful software tools that enable the creation of complex machine learning tools based on a couple hundred lines of Python code [such as TensorFlow (Abadi et al. 2015), Keras (Chollet et al. 2015), and PyTorch (Paszke et al. 2019)], and

Corresponding author: Peter D. Dueben, peter.dueben@ecmwf.int

DOI: 10.1175/AIES-D-21-0002.1

the recent development of computing hardware with processors optimized for deep learning and dense linear algebra at low numerical precision [e.g., Google's tensor processing units (TPUs) and NVIDIA's TensorCores], which allows one to train very complex machine learning tools with billions of degrees of freedom. While machine learning promises to help improve weather and climate predictions and to extract more information about the Earth system from the available data, machine learning also poses challenges for weather and climate prediction centers and in fact the entire scientific domain of Earth sciences (Düben et al. 2021). These include, for example, difficulties in the communication between domain scientists—who are used to numerical models based on process understanding and physical equations—and machine learning scientists—who work on an interdisciplinary challenge and focus on the data-science and statistical problems that are often defined via specific input and output data, and a specific loss function. Furthermore, the software and hardware tools that are used by domain scientists (often Fortran and CPU hardware) and machine learning scientists [mostly Python and graphics processing unit (GPU) hardware] are often different. However, note that, independent of machine learning, domain scientists are increasingly making use of Python for data analysis and visualization, and a number of modeling groups are porting their models to GPUs (Bauer et al. 2021).

While progress on the use and uptake of machine learning in Earth sciences has been breathtaking during the last 2–3 years, it has also become apparent that machine learning solutions that work well in other domains, such as image recognition, do not always provide the best possible solution for Earth sciences. Specifically, a physical domain that obeys physical laws of motion and interactions is not a cat video (Schultz et al. 2021; Karpatne et al. 2019). The benchmark image recognition problems, such as ImageNet, focus on the task of analyzing an image to determine if it contains an object from one of many predetermined classes. Images provide large datasets due to the billions of available images on the web, but each contains only three highly correlated color channels and roughly 10 million pixels. In contrast, atmospheric numerical simulation datasets for machine learning (ML) may contain from tens to thousands of simulations that have hundreds of related but distinct variables defined at roughly 100 million grid cells per variable and time step. This difference in dimensionality can necessitate different strategies for parallel computing and memory management. Image recognition machine learning tasks typically focus on analyzing the contents of a single image and "nowcasting" subsequent image frames, while Earth science machine learning tasks span the scope of analysis and prediction time scales from analyzing the present to predicting future states minutes to months in advance. While image recognition tasks typically assume data that are structured in space and translationally invariant, the atmosphere is often represented on special grids that respect the spherical shape of the globe (often unstructured or cubed sphere), with sparse observations, location-dependent influence of the Coriolis force, or with a vertical dimension that exhibits very different properties (and grid spacings) at the surface and at the top. Another issue in Earth sciences is the huge range of scales with many interactions occurring within and across scales. Such multiscale learning is still in its infancy but could be impactful across a broad set of physical and biological science problems (Alber et al. 2019).

Experience from other scientific domains has demonstrated that the best way to find an optimal machine learning solution for a specific challenge and to enable cross-institutional collaborations and healthy competition is to define a benchmark dataset. Benchmark datasets should develop over time and "grow" in complexity (not necessarily in size) as machine learning solutions develop. The most prominent example for this evolution of benchmark data stems from image recognition, which started from a 70 000 sample dataset of 28 × 28 pixel images [Modified National Institute of Standards and Technology (MNIST); https://www.kaggle.com/avnishnish/mnist-original] and can now base developments on very large and complex benchmark datasets with millions of images (Russakovsky et al. 2015). Benchmark datasets are defined by the community. While many datasets can be proclaimed as "benchmarks" they are only successful if the machine learning community embraces them.

For atmospheric sciences, the first benchmark datasets are being developed and are starting to be used by several groups (see, e.g., Haupt et al. 2021) and there have been initial attempts to assemble lists of benchmark datasets (http://mldata.pangeo.io/). Currently, the most prominent benchmark in atmospheric sciences may be the WeatherBench benchmark datasets that aims to learn a global atmospheric model as a pure data science problem from global reanalysis data (Rasp et al. 2020). The WeatherBench dataset is very useful to test machine learning methods for their ability to represent atmospheric motion and scale interactions, to develop robust models that do not diverge when run for long forecast lead times, and to quantify prediction uncertainty. There is still only a small number of entries to the leaderboard (https://github.com/pangeo-data/WeatherBench; eight entries including five entries from the original paper), but there is more work that is not listed yet (Keisler 2022; Pathak et al. 2022), the respective paper (Rasp et al. 2020) has already been cited 37 times within the first 1.5 years after publishing. Nevertheless, many promising application areas for machine learning in atmospheric sciences will require benchmark datasets that are more targeted toward specific spatiotemporal scales and application needs. In air-quality research the concept of benchmark datasets is not well developed at present, but Betancourt et al. (2021) provides a rare example based on rich metadata aggregates from the global database of tropospheric ozone measurements (Schultz et al. 2017).

Benchmark datasets are needed to allow for a quantitative comparison between different machine learning tools when applied to a specific application of interest—as methods cannot be compared when trained using different datasets and tested using different diagnostics. Benchmark datasets also reduce the need for individual groups to develop their own datasets, which often takes more time than the work on the actual machine learning problem. This is especially important in light of the different expertise of domain scientists, who understand their data and know how to process terabytes of numerical model output, and ML experts, who have a superior knowledge of ML methods and their strengths and limitations.

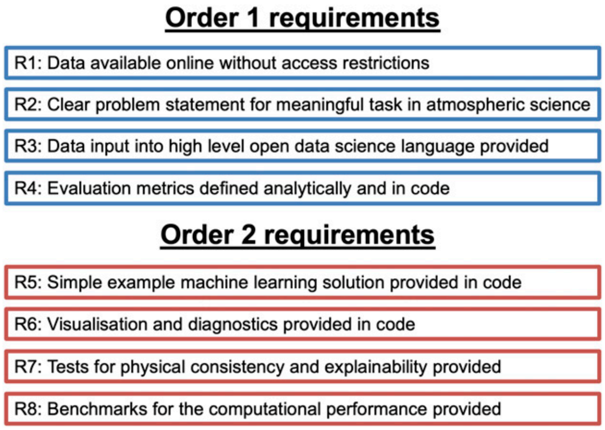
Besides saving a lot of work in constructing similar ML datasets from scratch over and over again, well-defined benchmark datasets can centralize and channelize data requests and thus reduce input/output (IO) and internet traffic. If, for example, 10 groups of scientists retrieve very similar datasets from a petabyte-scale model data archive, then many terabytes of data have to be loaded, processed, and reformatted 10 times to generate an ML dataset, which might have a size of several gigabytes or even less. If, on the other hand, each of these groups uses the same benchmark dataset, the data processing and copying will be limited to 10 times a few gigabytes, and such datasets can be made available online for easier access. Furthermore, the optimal machine learning solution for a specific problem in atmospheric sciences can often be used as a starting point for multiple applications. For example, a machine learning tool that performs well for WeatherBench may also be a promising candidate for use in bias correction (Bonavita and Laloyaux 2020) or for the improvement of global ensemble predictions (Gröenquist et al. 2021).

In general, there is a large interest by research groups specialized in machine learning to work on problems related to atmospheric sciences—mainly due to the numbers of data that are available and the societal impact of weather and climate predictions—the same is also true for large companies (Kurth et al. 2018; Weyn et al. 2021; Ravuri et al. 2021). Research groups in high performance computing, as well as hardware vendors, are also generally interested in improving the efficiency of weather and climate applications and the use of benchmarks in this domain as models of the atmosphere are one of the most important applications in high-performance computing at scale. We can therefore expect that these benchmarks will be picked up by several communities even without active promotion. However, the benchmarks need to be sufficiently documented and easy to use by nonspecialists.

This paper is not the first to discuss benchmark datasets in atmospheric sciences. However, this paper adds the following main elements to the existing literature: 1) provision of a clear definition of the ingredients that benchmark datasets should have (section 2) and 2) outline of scientific areas that will benefit from the definition of benchmark datasets in the future (section 3), including references to datasets that already exist. This paper also outlines (section 4) how these benchmark datasets could be collected and compared.

2. The definition of a benchmark dataset

In general, machine learning benchmark datasets fall into two categories: scientific and competition benchmark datasets. Scientific benchmark datasets represent a common platform for addressing a challenging domain problem shared by many research groups over a long timeframe. Members of the domain community apply new models to the benchmark to intercompare with existing approaches in a standardized fashion, reducing the confounding factors introduced by comparisons across bespoke datasets. In contrast, competition benchmark datasets generally focus on a more narrowly defined task that could be improved with focused short-term attention. Competition benchmarks often seek participation from a broader community of nondomain experts through platforms such as Kaggle. Some monetary reward is given to the teams that produce the best-scoring ML solution submitted within a limited timeframe.


Order 1 requirements

R1: Data available online without access restrictions

R2: Clear problem statement for meaningful task in atmospheric science

R3: Data input into high level open data science language provided

R4: Evaluation metrics defined analytically and in code

Order 2 requirements

R5: Simple example machine learning solution provided in code

R6: Visualisation and diagnostics provided in code

R7: Tests for physical consistency and explainability provided

R8: Benchmarks for the computational performance provided

FIG. 1. A summary of the requirements of a scientific benchmark. Order-1 requirements are considered to be absolutely essential in a benchmark definition, and order-2 requirements are recommended components for a complete definition.

In the atmospheric science domain such competitions have been previously organized by the American Meteorological Society Artificial Intelligence Committee and have included storm mode analysis (Lakshmanan et al. 2010), wind and solar energy prediction (McGovern et al. 2015), and quantitative precipitation estimation. More recent competitions have included the U.S. Bureau of Reclamation Sub-Seasonal Climate Forecasting Rodeo (U.S. Bureau of Reclamation 2019) and the World Meteorological Organization Seasonal-to-Subseasonal (S2S) Artificial Intelligence (AI) Challenge (World Meteorological Organization 2021). Competitions can help domain scientists identify promising new algorithms and talented students/early career professionals who could make further progress with more sustained support. Because of high financial stakes and short time frames, competition benchmark datasets usually centralize evaluation on a privately held portion of the data to penalize overfitting from repeated submissions to a public test metric. Scientific benchmarks contain all the data and rely on members of the community to self-report their performance, usually through peer-reviewed publications. We now define two orders of requirements to form a scientific benchmark datasets for the atmospheric science domain. While the order-1 requirements are absolutely essential to allow for quantitative comparisons, the order-2 requirements are still important to define a complete benchmark. The list is summarized in Fig. 1, and the points are described in detail in following lists. There are four order-1 requirements for scientific benchmark datasets:

R1) All data that are required to perform the benchmark testing must be available online without access restrictions. There must also be a descriptor of which data are used for testing/validation so results are standardized across different attempts. If possible, all data should be shared in a standardized file format [such as netCDF, gridded binary or general regularly distributed information in binary (GRIB), or hierarchical data format, version 5 (HDF5)] and follow common conventions for naming and units [such as the climate and forecast (CF) metadata conventions].

R2) There must be a clear problem statement defining the task to be solved through machine learning, and the task must be meaningful for atmospheric scientists. This problem statement must include an evaluation metric that is accepted by the domain expert community. If possible, the problem statements may also include tests on extrapolation of results, for example, to a changing climate, and information for the maximal error that may be acceptable in a specific application.

R3) Methods to read the data into a high-level open data science language platform, such as Python, R, or Julia with Jupyter notebooks, must be provided to make it as easy as possible to start working with the data. The chosen software environment should allow for the simple application of machine learning tools and be widely accepted by the ML community. If possible, the code should be available from a repository that allows external users to upload additional solutions.

R4) A well-defined and quantitative evaluation metric that is of use to the domain experts (e.g., skill score) must be defined and available in both the analytical formula and computer code. The availability as code will not only allow for a simple start when users are working with the benchmark dataset, it will also ensure that the quantitative metrics are calculated in exactly the same way by different users as there may be different choices possible that can have an impact on the final scores—for example, when sampling the probability density function to calculate the continuous ranked probability score (CRPS). If possible, a reference solution of conventional tools should be provided and described in sufficient detail.

Order-2 requirements for scientific benchmark datasets (on top of the order-1 list) are the following:

R5) A simple example machine learning solution for the machine learning problem must be available in code. This solution can be a very simple statistical or machine learning model, such as linear regression. However, it could also be a sophisticated machine learning solution that was published together with the benchmark dataset. The availability of an existing solution will facilitate the use of the dataset but will also enable the easy reproduction of published results to verify that the workflow is consistent. If the training for the benchmark solution is expensive, pretrained model weights should also be published with the dataset.

R6) Important visualizations and diagnostics that are used by the creators of the dataset must be made available in code. Again, this facilitates the use of the dataset and allows contributors to easily reproduce published results.

R7) If useful, additional tests for physical consistency and robustness, or tests to explore explainability and interpretability should be provided that are needed to convince domain scientists of the usefulness of the solution (such as a specific example of a weather event for a postprocessing tool of numerical weather predictions to check spatial and physical consistency of the solution).

R8) If useful, the computational performance of the machine learning solution that is published with the dataset but also of any new solution should be measured. This should include both training and inference speed, along with memory, storage requirements, and implementation instructions (e.g., for parallelization). Naturally, the performance numbers will heavily depend on the hardware and software that was used and will be difficult to compare directly. Relative comparisons against a baseline approach run on the same hardware may enable some level of performance comparisons across machines. However, the efficiency of machine learning solutions will likely become increasingly important as datasets and the complexity of machine learning solutions grow.

3. A suite of challenges and benchmark datasets to address them

In this section, we will outline areas for which benchmark datasets could help make progress in the atmospheric science domain. These include both application areas within atmospheric sciences but also technical developments for machine learning. We will present the different areas with needs for a benchmark dataset in order of importance—as judged by the authors—and have categorized the different benchmark datasets into three tiers. The list of benchmark challenges is summarized in Table 1. However, the individual entries are also explained in more detail in the text below. The challenges that are addressed by the first tier will be essential for the success of machine learning in atmospheric sciences in the coming years. Challenges of tier 2 are also important but are less general when compared with the first tier. Tier-3 challenges will also be important but only on a longer time scale.

Furthermore, we have categorized the benchmark datasets into three different groups for two additional categories (see Table 1). While some of the benchmark datasets could arguably be allocated to several groups, we have only allowed for a single allocation per category to keep the groups distinct. The first category distinguishes between benchmarks that are mainly handling 1) challenges due to the underlying physical, turbulent system, which is described by equations, and showing feedbacks and responses that can be explained physically; 2) challenges that relate to the uncertainties or availability of observations; and 3) challenges that concern developments or improvements of numerical models, for example, with regard to the computational speed, complexity, or technical developments. The second category distinguishes between benchmarks that handle 1) challenges that are very specific for the domain but will apply to a number of different machine learning applications within atmospheric science, 2) general challenges that are mainly concerning data handling, loading and cleaning, as well as data structures and dependencies, that

TABLE 1. This table provides an overview of the benchmark challenges that have been grouped by priority according to the authors.

		Category 1			Category 2		
No.	Benchmark challenge	Physical turbulent system	Observation	Modeling	Domain-specific challenges	General data challenges	Specific applications
Tier 1							
1	Weather and climate predictions based on machine learning			X			X
2	Trustworthy AI, explainable AI, and physical consistency	X			X		
3	Hybrid modeling and coupling			X	X		
4	Physical constraints	X			X		
5	Uncertainty quantification and representation			X	X		
6	Extreme value predictions		X		X		
7	Machine learning solutions in a changing climate	X				X	
Tier 2							
8	Model postprocessing and downscaling			X			X
9	Air-quality data interpolation		X				X
10	Varying types of data and fusion of diverse datasets		X		X		
11	Unstructured grids on the sphere			X		X	
12	Huge volume of data			X		X	
13	The emulation of model components			X			X
14	Detection of weather phenomena and pollution events		X				X
15	Multiscale interactions in space and time	X			X		
16	Data quality control		X			X	
17	Data quality across sites		X		X		
18	Nowcasting applications		X				X
19	Transfer learning	X			X		
20	Site-specific characteristics of observations		X		X		
Tier 3							
21	Missing data and irregular spacing of monitoring sites		X			X	
22	Autocorrelation and periodic patterns in time series		X			X	
23	Online and reinforcement learning	X			X		
24	Composite distributions of observations		X		X		
25	Human label consistency and bias		X			X	

should be addressed on a technical level, and 3) challenges that address specific applications within atmospheric sciences but for which solutions will often not be generalizable to other areas within the scientific domain—in contrast to the challenges in category 1.

The tier-1 benchmark challenges are as follows.

1) *Weather and climate predictions based on machine learning*: Learning the full prediction model from data is an interesting application that is easy to motivate. The WeatherBench benchmark dataset is already well placed for global predictions. Additional datasets could cover the advection of tracers (e.g., in the context of an atmospheric chemistry model) or could extend WeatherBench to higher resolutions, including more fields and fields with different statistical properties. Furthermore, diagnostic fields—such as CAPE—or quantities that are predicted but not prognostic variables—such as precipitation—could be learned. Specific example: This benchmark is basically already existing in form of WeatherBench (Rasp et al. 2020). However, WeatherBench could be extended with more data and more physical fields as the complexity of machine learning tools are growing over time [e.g., via the use of CMIP5 data for pretraining in a transfer learning approach Rasp and Thuerey (2021)]. Furthermore, as the machine learning solutions are approaching the scores of the operational predictions (Keisler 2022; Pathak et al. 2022), more sophisticated loss functions and diagnostics should be defined. These should also include probabilistic scores as machine learning provides very interesting opportunities to improve ensemble predictions.
2) *Trustworthy AI, explainable AI, physical consistency, and diagnostics to understand machine learning solutions* (Beucler et al. 2021; McGovern et al. 2019; Reichstein et al. 2019): While these topics are currently receiving significant attention for the application of machine learning in Earth system sciences and while the need for the use of these methods is intuitive in physical applications, their practical implementation is often difficult. This is due to physical conservation properties that are often only approximately realized in conventional models, incomplete knowledge about the level of errors in physical

consistency that can be accepted, physical feedbacks that are difficult to quantify, and a trade-off between the complexity and interpretability of machine learning solutions. There is therefore a need to develop benchmarks (plural) encouraging the construction of machine learning tools that follow physical reasoning and to build diagnostic tools to understand the functionality of complex machine learning tools and the consistency with physical feedback. Specific example: The generation of a benchmark dataset that is following a known physical equation of motion, including conservation properties (e.g., turbulent flow), and a machine learning task that is tested for the ability to fulfil the constraints.

3) *Hybrid modeling and coupling*: For many applications of machine learning the most promising approach combines machine learning techniques with conventional modeling. This allows one to incorporate physical consistency via the conventional model, and often the signal-to-noise ratio is improved when machine learning is used to predict a delta rather than the full signal (Watson 2019). However, there are still questions about how to best combine conventional models with machine learning. For example, how does one use machine learning tools within complex time-stepping schemes, either explicit or (semi)implicit? A benchmark problem may help identify the optimal approach for coupling machine learning and conventional methods. Specific example: Couple a conventional model of medium complexity (e.g., a simple dynamical core) with a machine learning tool to correct the model toward a truth dataset (e.g., from analysis or a simulation at much higher resolution).

4) *Physical constraints*: Many physical variables are bound by physical constraints, which lead to non-Gaussian frequency distributions. For example, precipitation or atmospheric constituent concentrations will never become negative and relative humidity can become saturated. Information about these physical constraints can improve the training of machine learning solutions. In particular, for predictions of extreme values that may be rare in the training datasets, it may be better to learn distributions rather than individual values. A benchmark dataset imposing one or more such constraints would help to compare various approaches for incorporating physical constraints in the model as well as assumptions about the distributions of physical fields. The non-Gaussian nature of certain variables raises questions about how to best preprocess data (e.g., log transform) and how to perform data normalization for deep learning applications. Specific example: A dataset from parameterization emulation with a clear description of the physical constraints that are present.

5) *Uncertainty quantification and representation*: There are many ways to use machine learning tools to improve uncertainty quantification in predictions. Examples include the use of Bayesian machine learning techniques, dropout techniques, the explicit prediction of model error, or the postprocessing of ensemble simulations (Gagne et al. 2020; Gröenquist et al. 2021). A benchmark dataset would help to compare these different approaches. Specific example: A dataset gathers information about uncertainty quantification from ensemble weather predictions and compares the estimate of the uncertainty quantified by a machine learning tool with this reference [e.g., as done in application A4 of the Machine Learning for Scalable Meteorology and Climate (MAELSTROM) datasets (https://www.maelstrom-eurohpc.eu/content/docs/uploads/doc6.pdf)].

6) *Extreme value predictions*: Forecasting problems are regression problems. Statistical forecasting methods tend to converge toward mean values and therefore perform poorly for extreme conditions, which are generally rare events in the atmosphere. For example, heavy precipitation events occur on less than 1% of days in temperate latitudes. While there are ML methods designed to improve the representation of extreme values in time series forecasts, these are difficult to compare for lack of a standardized benchmark dataset. Extreme weather events tend to have disproportionately larger impacts than their less-extreme counterparts but are also rarer and hard to identify deterministically. Standard machine learning methods are optimized to predict the most likely outcome for a given set of inputs, and generally perform poorly on rare events. However, machine learning architectures could be merged with parametric assumptions about the distributions of extremes to predict the probability of extreme events outside the space of the training data. A benchmark explicitly designed for this challenge could help drive a new wave of research in this direction. Specific example: Precipitation nowcasting (Ravuri et al. 2021) or downscaling that focuses on extreme precipitation events.

7) *Training and use of machine learning solutions in a changing climate*: Most machine learning techniques are good at interpolation but are unable to extrapolate. However, machine learning tools that are trained with data from today's climate but used in applications of climate change will naturally be confronted with weather situations that have not been observed during training. This problem could potentially be addressed through the generation of synthetic datasets (Meyer et al. 2021), including weather or climate model runs with various forcing scenarios (Molina et al. 2021), or by building machine learning tools that can extrapolate, for example, by incorporating physical constraints (Beucler et al. 2021; Yuval et al. 2021). A benchmark dataset would allow us to compare these approaches. Specific example: A WeatherBench-type dataset based on CMIP emission scenarios.

The tier-2 benchmark challenges are as follows.

8) *Model postprocessing and downscaling*: For this application, the challenges are closely related to applications in multiscale systems and machine learning applications for observations. However, the use of machine learning techniques for postprocessing is very prominent and may have a large impact on local predictions via the fusion of

forecast data with local information such as topography and local observations (Leinonen et al. 2021).

9) *Air-quality data interpolation*: The use of machine learning methods is very interesting for air-quality and pollution prediction as they allow the fusion of local observations with data from conventional pollution modeling, such as the Copernicus Atmosphere Monitoring Service (CAMS). This is important, as the density of observations is particularly sparse and inhomogeneous for pollution measurements, and local dynamics and pollution signatures depend heavily on local properties, such as topography, traffic, industrial emissions, and land use. The first challenges on pollution modeling with machine learning have been published (https://ai4eo.eu/; Betancourt et al. 2021), but additional benchmarks would be useful.

10) *Varying types of data and fusion of diverse datasets*: Machine learning is very useful for merging and fusing datasets. However, the machine learning often requires the use of data from various sources with uneven quality. For example, when using local temperature observations from different sources (e.g., countries), or when using high-quality data samples (such as measurements from national meteorological services) in combination with measurements from crowd-sourced or internet-of-things data. Data from high-quality sensors can also provide different types of data. For example, there is no consistent global radar specification. Algorithms developed for U.S.-based radar networks may not work in other countries whose radars operate at different frequencies. It is important to design a problem that is widely applicable and to note the limitations of a given technique. A reference dataset composed of disparate data sources would be useful to learn how to gain the best solution quality from diverse datasets. This could also be useful in the context of transfer learning. In particular, data assimilation frameworks may guide the fusion of training data as they already represent a framework for bringing various observational datasets together in a single application.

11) *Unstructured grids on the sphere*: Global weather and climate predictions describe the propagation and interaction of physical fields on the sphere. Although simple, regular longitude/latitude grids are suboptimal as they exhibit coordinate singularities with a dense cluster of grid points at each pole. Unstructured grids are generally more suitable. Naturally, the development of machine learning techniques that operate on unstructured grids would be very useful for global atmospheric science applications. However, most popular machine learning techniques, such as convolutional neural networks require structured grids. Therefore, the application of unstructured grids in deep learning solutions is a good candidate for a benchmark problem. One exception is the cubed-sphere grid as used in atmosphere models such as the Finite-Volume Cubed-Sphere dynamical core (FV^3), which has been used successfully for machine learning applications (Weyn et al. 2021).

12) *Huge volume of data*: For many applications of machine learning in Earth sciences, the data that are available for training will be limited. However, for a number of applications of machine learning in atmospheric sciences, there is almost no hard limit to the amount of training data available. For example, for applications that involve CMIP data or global ensemble predictions at high resolution, or for the emulation of physical parameterization schemes. Here, the capability to handle large datasets is often the limit. In particular, it is often difficult to fit data samples into memory if many high-resolution fields are required for global applications. A benchmark dataset that aims to optimize data handling may help to produce more efficient solutions for training and inference.

13) *The emulation of model components*: The emulation of parameterization schemes to speed up simulations has recently been studied by a number of groups, in particular for complex parameterization schemes such as superparameterization (Brenowitz and Bretherton 2018; Rasp et al. 2018; Chantry et al. 2021; Yuval et al. 2021). It may even be possible to learn improved parameterization schemes directly from observations. Here, it would be useful to define a benchmark dataset that would not only provide data for offline training but also allow for online testing within model simulations (see earlier discussion of online testing in section 3a). Training machine learning models of physical processes acting over vertical grid columns could be used as a blueprint for a number of applications in atmospheric sciences.

14) *Detection of weather phenomena and pollution events*: Machine learning tools are used successfully and routinely for feature detection in image recognition, but feature detection can also be performed when handling observations or modeling datasets (Lagerquist et al. 2020). Automatic feature detection can greatly reduce data volumes, for example, when storing the position and strength of a tropical cyclone instead of a three-dimensional field in climate model output, or when transmitting weather information from satellites. Feature detection may also enable the production of specific output fields within model simulations only when relevant thereby speeding up model simulations. Benchmark datasets would help to optimize the detection of features for physical fields.

15) *Multiscale interactions in space and time*: The atmosphere is a three-dimensional turbulent fluid and therefore exhibits interactions on all scales in both space and time. While encoder–decoder network architectures can be interpreted as multiscale solutions and are used successfully for atmospheric applications (e.g., Weyn et al. 2021; Ravuri et al. 2021), and while machine learning methods that use Fourier and spectral space show promising results for the modeling of turbulent fluids (Li et al. 2020), a clear evaluation of the ability of machine learning solutions to represent multiscale dynamics as present in the atmosphere still needs to be formulated, and this process could be supported by a benchmark problem.

16) *Data quality control*: Machine learning methods can map properties from one observational dataset to another,

even if the underlying measurements or physical fields do not allow for a direct physical comparison. Furthermore, machine learning tools can learn to detect anomalies. Machine learning is therefore well placed to detect and isolate erroneous observations. These capabilities can be applied within data assimilation frameworks, but also for the realization of a digital twin concept for the entire Earth system or for its components. Uncertainties for specific observations could be provided with the datasets and corrupted observations could be flagged for the user. Benchmark datasets for data fusion and anomaly detection would be useful for comparing different machine learning approaches.

17) *Data quality across sites*: Most meteorological variables can be measured by different instruments but unless the network of observations is specifically calibrated (e.g., Mesonets; McPherson et al. 2007; Brotzge et al. 2020), the data may vary in reliability. This can cause issues for machine learning, which assumes the data are correct and could be especially problematic if these data are being used as ground truth. It is important for benchmarks to acknowledge dataset limitations, and it could be useful to provide known error estimates and to enable the machine learning model to make use of the error estimates.

18) *Nowcasting applications*: Machine learning has proven to be useful when performing short-term nowcasting, with predictions in the range of hours to days, when trained directly from observations. This approach can replace the combination of data assimilation and forecast models with a single tool. The time saved when generating predictions can be critical for short lead times and allows for more frequent forecasts. Furthermore, physical constraints such as conservation laws are less important over such short time periods as large errors do not accumulate. Several groups have produced nowcasting models but used different datasets (Sønderby et al. 2020; Ravuri et al. 2021). Benchmark datasets aimed at nowcasting a standard set of variables such as temperature and precipitation would help to identify the most promising algorithms.

19) *Transfer learning*: While there are initial examples of transfer learning for atmospheric sciences that have been published (e.g., Ham et al. 2019; Rasp and Thuerey 2021), benchmark datasets for transfer learning would help tackle a number of unanswered questions. For example, whether to perform extra training for a machine learning tool for data from a new model version if a modeling system is upgraded, or whether to train from data of all available model versions with an additional input variable that defines the model version. Another aspect in this regard is the generalization of geographic regions (e.g., Sha et al. 2020). For example: can a model trained on data over Europe be applied to the Indian subcontinent? Will physical constraints imposed on the models help or hinder transfer learning? Furthermore, it will be interesting to explore to what extent artificial data generators can be used for the preconditioning of deep learning networks.

20) *Site-specific characteristics of observations*: It can be difficult to generalize information from local observations to larger areas due not only to the change of variability and distributions of values when mapping from point measurements to grid values, but also due to local environment characteristics imprinted into observations. Often, factors that influence the local environment are unknown. It is therefore important to design a benchmark problem to gain insight into how tools trained from observations at a specific location can be retrained for other locations and how to generate datasets that can be generalized to large areas or even the entire globe.

The tier-3 benchmark challenges are as follows.

21) *Missing data and irregular spacing of monitoring sites*: Observations from surface measurements, ships, planes, and weather balloons are not distributed regularly over the planet. Even satellite observations do not typically span the entire globe and may show changing positions and biases over time. Machine learning shows promising results for gap filling between observations, but gap filling methods may need to be customized to the task at hand. They will, for example, differ between applications for which variability and covariance between grid points needs to be maintained, and applications for which uncertainty quantification for interpolated values is required. It would be useful to build benchmark datasets targeted at gap filling with these types of requirements.

22) *Autocorrelation and periodic patterns in time series*: The information in the Earth system is correlated in both space and time, and the Earth system shows variability on many different time scales from seconds to millennia. It is therefore important that limitations in the independence of data samples are acknowledged. However, how should correlated data (e.g., time series) be sampled for training? Should as many samples as possible be used for training to reduce overfitting, or should samples at a certain level of correlation be disregarded? Should the validation and test dataset be taken from either end of the time series, or should these datasets be taken from a number of time slices throughout the time series? A benchmark problem could help answer these questions.

23) *Online and reinforcement learning*: Both are still rarely used for relevant machine learning applications in atmospheric sciences. Online learning (Parisi et al. 2019) refers to machine learning models that are optimized on a sequential stream of data rather than randomly sampled minibatches. Reinforcement learning (Sutton and Barto 2018) optimizes the behavior of an agent through a reward function that is scored based on repeated interactions with a set environment. Both online and reinforcement learning are more challenging and computationally intensive to implement than the most widely used supervised learning. Online learning could be useful in settings where it is infeasible to store all the data passing through

a system. Reinforcement learning could optimize instrumentation design or data collection patterns and could be used to model decision-making procedures that are weather dependent. Benchmark problems would therefore help the atmospheric science community to understand how to formulate applications for these classes of machine learning methods in a meaningful way. At the very least, these benchmarks could serve as an educational exercise for domain scientists to get to know the intricacies of the methods.

24) *Composite distributions of observations*: Data distributions of observations can vary significantly by season, wind direction, and from a variety of other factors. As a result, datasets are often composed of multiple statistical distributions some of which may be significantly underrepresented and for which the mean distribution may differ significantly from the true physical distribution, for example, when a satellite cannot provide measurements if clouds are present. A benchmark aimed at learning composite distributions might allow for significant progress in the analysis of atmospheric observations and may also help to improve ML-based weather forecasting.

25) *Human label consistency and bias*: Many weather phenomena of interest are currently labeled by domain experts. However, manually labeled data comes with the potential for error and for disagreement among human experts. This is the case for data in other fields as well, but it can be particularly acute in meteorology (McGovern et al. 2021). Also, such human-provided labels can potentially contain bias (e.g., Allen et al. 2017; Anderson et al. 2007; Allen and Tippett 2015), which can then unintentionally be replicated in the machine learning model. If human data are provided for a benchmark dataset, a thorough analysis of potential biases must also be provided.

4. Summary

This paper provides a clear definition of benchmark datasets and a list of scientific areas that should be covered by benchmark datasets to facilitate scientific developments for the application of machine learning in atmospheric science in as targeted a fashion as possible. The paper includes references to datasets that already exist but, given the speed with which the scientific domain is evolving and the breadth of the application areas, this list cannot be complete. We aim to update the website (http://mldata.pangeo.io/) over time with new benchmark datasets made available for the topics listed within this paper.

Successful atmospheric science machine learning benchmarks have the potential to drive research in both the atmospheric science and machine learning communities for many years to come. To ensure a benchmarks' continued relevance, members of the community need to help support the maintenance of benchmark infrastructure. Benchmark maintenance can include preserving or expanding data access, auditing the datasets for systematic errors and biases, adjusting the choice of metrics as groups make progress or shift focus, creating tutorials and other documentation, hosting benchmark-focused sessions at conferences and workshops, and incorporating the data and models into other applications. Benchmark maintenance requires a convergent community effort spanning both the machine learning and domain sciences for them to have a continued impact.

A new discussion has developed in machine learning around so-called benchmark islands. The simple use of benchmarks and the possibility to evaluate results based on a few diagnostics can divert the focus from the actual scientific application and provoke chains of publications that do not improve the solution for the original application that motivated the benchmark. To counteract, the domain of machine learning in atmospheric sciences will need to constantly revisit the usefulness of the applications to either enhance scientific understanding or to improve our models for operational use (in terms of either efficiency or quality). This will require close cooperation between domain scientists and machine learning experts and some level of coordination, for example, through establishing regular benchmark review sessions at community relevant conferences.

Benchmark datasets can obviously only be made available if the data are not restricted to certain user groups. It is therefore important for machine learning developments that data are published as open source and that machine learning vanilla solutions and state-of-the-art reference results are made available as open source code. While this is often possible for model and reanalysis data, observational data are often only available with restricted access. This paper can be interpreted as a plea for the community of atmospheric sciences to continue to work toward open science wherever possible.

Furthermore, as datasets become available at higher resolutions and as more data sources and in particular internet-of-things datasets are used, questions about the individual privacy need to be considered, for example, if mobile telephone data or information from social media are used in weather and climate science. This information needs to be anonymized, which may require degrading the quality of measurements, for example, with regard to spatial resolution.

Acknowledgments. We thank Umberto Modigliani and three anonymous reviewers for many comments that helped to improve the paper. Author Dueben gratefully acknowledge funding from the Royal Society for his University Research Fellowship as well as the ESiWACE project funded under Horizon 2020 823988. Dueben and author Chantry gratefully acknowledge funding from the MAELSTROM EuroHPC-JU project (JU) under 955513. The JU receives support from the European Union's Horizon Research and Innovation Programme and United Kingdom, Germany, Italy, Luxembourg, Switzerland, and Norway. Author Schultz acknowledges funding through ERC-2017-ADG 787576 (IntelliAQ) and the program Engineering Digital Futures of the Helmholtz Association's research field Information. Authors McGovern and Gagne acknowledges that this material is based upon work supported by the National Science Foundation under Grant ICER-2019758. Gagne acknowledges support by the National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the National Science Foundation (NSF) under Cooperative Agreement 1852977.

Data availability statement. There are no data accompanying this article.

REFERENCES

Abadi, M., and Coauthors, 2015: TensorFlow: Large-scale machine learning on heterogeneous systems. TensorFlow, https://www.tensorflow.org/.

Aires, F., P. Weston, P. de Rosnay, and D. Fairbairn, 2021: Statistical approaches to assimilate ASCAT soil moisture information—I. Methodologies and first assessment. *Quart. J. Roy. Meteor. Soc.*, **147**, 1823–1852, https://doi.org/10.1002/qj.3997.

Alber, M., and Coauthors, 2019: Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *npj Digital Med.*, **2**, 115, https://doi.org/10.1038/s41746-019-0193-y.

Allen, J. T., and M. K. Tippett, 2015: The characteristics of United States hail reports: 1955–2014. *Electron. J. Severe Storms Meteor.*, **10** (3), https://doi.org/10.55599/ejssm.v10i3.60.

——, ——, Y. Kaheil, A. H. Sobel, C. Lepore, S. Nong, and A. Muehlbauer, 2017: An extreme value model for U.S. hail size. *Mon. Wea. Rev.*, **145**, 4501–4519, https://doi.org/10.1175/MWR-D-17-0119.1.

Anderson, C. J., C. K. Wikle, Q. Zhou, and J. A. Royle, 2007: Population influences on tornado reports in the United States. *Wea. Forecasting*, **22**, 571–579, https://doi.org/10.1175/WAF997.1.

Andersson, T., and Coauthors, 2021: Seasonal Arctic sea ice forecasting with probabilistic deep learning. *Nat. Commun.*, **12**, 5124, https://doi.org/10.1038/s41467-021-25257-4.

Bauer, P., P. D. Dueben, T. Hoefler, T. Quintino, T. C. Schulthess, and N. P. Wedi, 2021: The digital revolution of Earth-system science. *Nat. Comput. Sci.*, **1**, 104–113, https://doi.org/10.1038/s43588-021-00023-0.

Betancourt, C., T. Stomberg, R. Roscher, M. G. Schultz, and S. Stadtler, 2021: AQ-Bench: A benchmark dataset for machine learning on global air quality metrics. *Earth Syst. Sci. Data*, **13**, 3013–3033, https://doi.org/10.5194/essd-13-3013-2021.

Beucler, T., M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, 2021: Enforcing analytic constraints in neural networks emulating physical systems. *Phys. Rev. Lett.*, **126**, 098302, https://doi.org/10.1103/PhysRevLett.126.098302.

Bonavita, M., and P. Laloyaux, 2020: Machine learning for model error inference and correction. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002232, https://doi.org/10.1029/2020MS002232.

Brajard, J., A. Carrassi, M. Bocquet, and L. Bertino, 2020: Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. *J. Comput. Sci.*, **44**, 101171, https://doi.org/10.1016/j.jocs.2020.101171.

Brenowitz, N. D., and C. S. Bretherton, 2018: Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.*, **45**, 6289–6298, https://doi.org/10.1029/2018GL078510.

Brotzge, J. A., and Coauthors, 2020: A technical overview of the New York State Mesonet standard network. *J. Atmos. Oceanic Technol.*, **37**, 1827–1845, https://doi.org/10.1175/JTECH-D-19-0220.1.

Chantry, M., S. Hatfield, P. Dueben, I. Polichtchouk, and T. Palmer, 2021: Machine learning emulation of gravity wave drag in numerical weather forecasting. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002477, https://doi.org/10.1029/2021MS002477.

Chollet, F., and Coauthors, 2015: Keras. https://keras.io.

Düben, P., and Coauthors, 2021: Machine learning at ECMWF: A roadmap for the next 10 years. ECMWF Tech. Memo. 878, 20 pp., https://www.ecmwf.int/node/19877.

Dueben, P. D., and P. Bauer, 2018: Challenges and design choices for global weather and climate models based on machine learning. *Geosci. Model Dev.*, **11**, 3999–4009, https://doi.org/10.5194/gmd-11-3999-2018.

Gagne, D. J., H. M. Christensen, A. C. Subramanian, and A. H. Monahan, 2020: Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz'96 model. *J. Adv. Model. Earth Syst.*, **12**, e2019MS001896, https://doi.org/10.1029/2019MS001896.

Gröenquist, P., C. Yao, T. Ben-Nun, N. Dryden, P. Dueben, S. Li, and T. Hoefler, 2021: Deep learning for post-processing ensemble weather forecasts. *Philos. Trans. Roy. Soc.*, **A379**, 20200092, https://doi.org/10.1098/rsta.2020.0092.

Ham, Y.-G., J.-H. Kim, and J.-J. Luo, 2019: Deep learning for multi-year ENSO forecasts. *Nature*, **573**, 568–572, https://doi.org/10.1038/s41586-019-1559-7.

Haupt, S. E., W. Chapman, S. V. Adams, C. Kirkwood, J. S. Hosking, N. H. Robinson, S. Lerch, and A. C. Subramanian, 2021: Towards implementing artificial intelligence post-processing in weather and climate: Proposed actions from the Oxford 2019 workshop. *Philos. Trans. Roy. Soc.*, **A379**, 20200091, https://doi.org/10.1098/rsta.2020.0091.

Karpatne, A., I. Ebert-Uphoff, S. Ravela, H. A. Babaie, and V. Kumar, 2019: Machine learning for the geosciences: Challenges and opportunities. *IEEE Trans. Knowl. Data Eng.*, **31**, 1544–1554, https://doi.org/10.1109/TKDE.2018.2861006.

Keisler, R., 2022: Forecasting global weather with graph neural networks. arXiv, 2202.07575v1, https://doi.org/10.48550/arXiv.2202.07575.

Kurth, T., and Coauthors, 2018: Exascale deep learning for climate analytics. *SC'18: Proc. Int. Conf. for High Performance Computing, Networking, Storage, and Analysis*, Dallas, TX, IEEE, 649–660, https://doi.org/10.1109/SC.2018.00054.

Lagerquist, R., J. T. Allen, and A. McGovern, 2020: Climatology and variability of warm and cold fronts over North America from 1979 to 2018. *J. Climate*, **33**, 6531–6554, https://doi.org/10.1175/JCLI-D-19-0680.1.

Lakshmanan, V., K. L. Elmore, and M. B. Richman, 2010: Reaching scientific consensus through a competition. *Bull. Amer. Meteor. Soc.*, **91**, 1423–1427, https://doi.org/10.1175/2010BAMS2870.1.

Leinonen, J., D. Nerini, and A. Berne, 2021: Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Trans. Geosci. Remote Sens.*, **59**, 7211–7223, https://doi.org/10.1109/TGRS.2020.3032790.

Li, Z., N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, 2020: Fourier neural operator for parametric partial differential equations. arXiv, 2010.08895v3, https://doi.org/10.48550/arXiv.2010.08895.

McGovern, A., D. J. Gagne, J. Basara, T. M. Hamill, and D. Margolin, 2015: Solar energy prediction: An international contest to initiate interdisciplinary research on compelling meteorological problems. *Bull. Amer. Meteor. Soc.*, **96**, 1388–1395, https://doi.org/10.1175/BAMS-D-14-00006.1.

——, R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

——, I. Ebert-Uphoff, D. J. Gagne II, and A. Bostrom, 2021: The need for ethical, responsible, and trustworthy artificial intelligence for environmental sciences. arXiv, 2112.08453, https://arxiv.org/abs/2112.08453.

McPherson, R. A., and Coauthors, 2007: Statewide monitoring of the mesoscale environment: A technical update on the Oklahoma Mesonet. *J. Atmos. Oceanic Technol.*, **24**, 301–321, https://doi.org/10.1175/JTECH1976.1.

Meyer, D., T. Nagler, and R. J. Hogan, 2021: Copula-based synthetic data augmentation for machine-learning emulators. *Geosci. Model Dev.*, **14**, 5205–5215, https://doi.org/10.5194/gmd-14-5205-2021.

Molina, M. J., D. J. Gagne, and A. F. Prein, 2021: A benchmark to test generalization capabilities of deep learning methods to classify severe convective storms in a changing climate. *Earth Space Sci.*, **8**, e2020EA001490, https://doi.org/10.1029/2020EA001490.

Nowack, P., P. Braesicke, J. Haigh, N. L. Abraham, J. Pyle, and A. Voulgarakis, 2018: Using machine learning to build temperature-based ozone parameterizations for climate sensitivity simulations. *Environ. Res. Lett.*, **13**, 104016, https://doi.org/10.1088/1748-9326/aae2be.

Parisi, G. I., R. Kemker, J. L. Part, C. Kanan, and S. Wermter, 2019: Continual lifelong learning with neural networks: A review. *Neural Networks*, **113**, 54–71, https://doi.org/10.1016/j.neunet.2019.01.012.

Paszke, A., and Coauthors, 2019: PyTorch: An imperative style, high-performance deep learning library. *33rd Conf. on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, BC, Canada, Neural Information Processing Systems, 8024–8035, http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pathak, J., and Coauthors, 2022: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv, 2202.11214v1, https://doi.org/10.48550/arXiv.2202.11214.

Rasp, S., and N. Thuerey, 2021: Data-driven medium-range weather prediction with a Resnet pretrained on climate simulations: A new model for WeatherBench. *J. Adv. Model. Earth Syst.*, **13**, e2020MS002405, https://doi.org/10.1029/2020MS002405.

——, M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci. USA*, **115**, 9684–9689, https://doi.org/10.1073/pnas.1810286115.

——, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, 2020: WeatherBench: A benchmark data set for data-driven weather forecasting. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002203, https://doi.org/10.1029/2020MS002203.

Ravuri, S., and Coauthors, 2021: Skilful precipitation nowcasting using deep generative models of radar. *Nature*, **597**, 672–677, https://doi.org/10.1038/s41586-021-03854-z.

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

Russakovsky, O., and Coauthors, 2015: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, **115**, 211–252, https://doi.org/10.1007/s11263-015-0816-y.

Schultz, M. G., and Coauthors, 2017: Tropospheric ozone assessment report: Database and metrics data of global surface ozone observations. *Elementa*, **5**, 58, https://doi.org/10.1525/elementa.244.

——, C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. Leufen, A. Mozaffari, and S. Stadtler, 2021: Can deep learning beat numerical weather prediction? *Philos. Trans. Roy. Soc.*, **A379**, 20200097, https://doi.org/10.1098/rsta.2020.0097.

Sha, Y., D. J. Gagne, G. West, and R. Stull, 2020: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: Daily precipitation. *J. Appl. Meteor. Climatol.*, **59**, 2075–2092, https://doi.org/10.1175/JAMC-D-20-0058.1.

Sønderby, C. K., and Coauthors, 2020: MetNet: A neural weather model for precipitation forecasting. arXiv, 2003.12140v2, https://doi.org/10.48550/arXiv.2003.12140.

Sonnewald, M., R. Lguensat, D. C. Jones, P. D. Dueben, J. Brajard, and V. Balaji, 2021: Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environ. Res. Lett.*, **16**, 073008, https://doi.org/10.1088/1748-9326/ac0eb0.

Sutton, R. S., and A. G. Barto, 2018: *Reinforcement Learning: An Introduction.* MIT Press, 552 pp.

U.S. Bureau of Reclamation, 2019: Forecast rodeo. https://www.usbr.gov/research/challenges/forecastrodeo.html.

Watson, P. A. G., 2019: Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction. *J. Adv. Model. Earth Syst.*, **11**, 1402–1417, https://doi.org/10.1029/2018MS001597.

Weyn, J. A., D. R. Durran, R. Caruana, and N. Cresswell-Clay, 2021: Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002502, https://doi.org/10.1029/2021MS002502.

World Meteorological Organization, 2021: Challenge to improve sub-seasonal to seasonal predictions using artificial intelligence. WMO, https://s2s-ai-challenge.github.io/.

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