Exascale computing and data handling: challenges and opportunities for weather and climate prediction

Govett, M., Bah, B., Bauer, P., Berod, D., Bouchet, V., Corti, S., Davis, C., Duan, Y., Graham, T., Honda, Y., Hines, A., Jean, M., Ishida, J., Lawrence, B. ORCID: https://orcid.org/0000-0001-9262-7860, Li, J., Luterbacher, J., Muroi, C., Rowe, K., Schultz, M., Visbeck, M. and Williams, K. (2024) Exascale computing and data handling: challenges and opportunities for weather and climate prediction. Bulletin of the American Meteorological Society, 105 (12). E2385-E2404. ISSN 0003-0007 doi: 10.1175/BAMS-D-23-0220.1 Available at https://centaur.reading.ac.uk/116611/

BAMS Article

Exascale Computing and Data Handling: Challenges and Opportunities for Weather and Climate Prediction

Mark Govett,[a] Bubacar Bah,[b] Peter Bauer,[c] Dominique Berod,[d] Veronique Bouchet,[e] Susanna Corti,[f] Chris Davis,[g] Yihong Duan,[h] Tim Graham,[i] Yuki Honda,[j] Adrian Hines,[k] Michel Jean,[l] Junishi Ishida,[m] Bryan Lawrence,[n] Jian Li,[h] Juerg Luterbacher,[o] Chiasi Muroi,[m] Kris Rowe,[p] Martin Schultz,[q] Martin Visbeck,[r] and Keith Williams[s]

KEYWORDS:
Atmosphere;
Ocean;
General circulation models;
Numerical weather prediction/ forecasting;
Optimization;
Climate

ABSTRACT: The emergence of exascale computing and artificial intelligence offer tremendous potential to significantly advance Earth system prediction capabilities. However, enormous challenges must be overcome to adapt models and prediction systems to use these new technologies effectively. A 2022 WMO report on exascale computing recommends "*urgency in dedicating efforts and attention to disruptions associated with evolving computing technologies that will be increasingly difficult to overcome, threatening continued advancements in weather and climate prediction capabilities*." Further, the explosive growth in data from observations, model and ensemble output, and postprocessing threatens to overwhelm the ability to deliver timely, accurate, and precise information needed for decision-making. Artificial intelligence (AI) offers untapped opportunities to alter how models are developed, observations are processed, and predictions are analyzed and extracted for decision-making. Given the extraordinarily high cost of computing, growing complexity of prediction systems, and increasingly unmanageable amount of data being produced and consumed, these challenges are rapidly becoming too large for any single institution or country to handle. This paper describes key technical and budgetary challenges, identifies gaps and ways to address them, and makes a number of recommendations.

SIGNIFICANCE STATEMENT: Earth system modeling and prediction stands at a crossroad. Exascale computing and artificial intelligence (AI) offer powerful new capabilities to advance Earth system predictions. However, models, assimilation, and data processing systems are increasingly unable to exploit these new technologies due to scientific, software, and computational limitations. Significant changes to the models including algorithms, software, and parallelism are needed to run models efficiently on diverse exascale systems. While AI offers significant potential, it is unclear the degree it can be developed and integrated into existing prediction systems. We recommend models be redesigned, linking science, software, and computing in codesign efforts to fully exploit exascale and AI. Special efforts are needed to recruit, train, and retain a highly skilled, interdisciplinary workforce. Given the high cost, shared computing and data facilities may become necessary.

DOI: 10.1175/BAMS-D-23-0220.1
Corresponding author: Mark Govett, markgovett@gmail.com
Manuscript received 23 August 2023, in final form 30 July 2024, accepted 1 August 2024

AFFILIATIONS: [a] NOAA/Global Systems Laboratory, Boulder, Colorado; [b] African Institute for Mathematical Sciences, Cape Town, South Africa; [c] European Centre for Medium Range Forecasts, Reading, United Kingdom; [d] Earth System Monitoring Division, WMO, Geneva, Switzerland; [e] Meteorological Service of Canada, Dorval, Quebec, Canada; [f] Institute of Atmospheric Science and Climate – CNR, Bologna, Italy; [g] National Center for Atmospheric Research, Boulder, Colorado; [h] Chinese Academy of Meteorological Sciences, Beijing, China; [i] Met Office, Exeter, United Kingdom; [j] Earth System Prediction Division, WMO, Geneva, Switzerland; [k] Center for Environmental Data Analysis, Science and Technology Council, Didcot, United Kingdom; [l] Infrastructure Commission, WMO, Geneva, Switzerland; [m] Japan Meteorological Agency, Tokyo, Japan; [n] Weather and Climate Computing, University of Reading, Reading, United Kingdom; [o] Science and Innovation Department, WMO, Geneva, Switzerland; [p] Argonne National Laboratory, Lemont, Illinois; [q] Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany; [r] GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany; [s] Atmosphere Physics and Parameterizations, Met Office, Exeter, United Kingdom

1. Introduction

Continued development and use of Earth system models (ESMs) are at the core of our ability to address the complex challenges that society faces due to climate change. Society's demands for more accurate predictions and more comprehensive information for decision-making are needed to reduce impacts of extreme weather and to adapt to a rapidly changing climate. ESMs cover a wide range of time scales and requirements for computing based on time-to-solution constraints, length of simulations, and complexity of the processes and interactions within the modeling system.

Traditionally, the symbiotic relationship between models and high performance computing (HPC) relied on the ability to double the speed and capability of computers every few years at a fixed cost. However, the point was reached around 2005 where this paradigm no longer held. Processors are no longer benefiting from faster clock speeds, so increases in computing power have been achieved with more compute cores. This trend is expected to continue with future systems anticipated to have millions to hundreds of millions of compute cores. This change has resulted in new challenges including the enormous cost and energy requirements to drive systems of this magnitude and finding ways for ESMs to use such systems effectively and efficiently (Lawrence et al. 2018). Despite significant efforts made to optimize them, a 2017 report shows that most ESMs use less than 5% of the CPU processor's peak capabilities (Carman et al. 2017). Additional and substantial performance improvements targeting next-generation computing [CPU, graphical processing unit (GPU), and hybrid processors] are possible but will require rethinking how models are designed and developed (Bauer et al. 2021). Further, artificial intelligence (AI) offers untapped opportunities to alter how models are developed, observations are processed, and predictions are analyzed and extracted for decision-making (see the sidebar).

HPC architectures are becoming increasingly complex and diverse. As the Earth observation and prediction systems have become more sophisticated with increasing spatial resolution and scientific complexity, there is a corresponding increase in the volume and diversity of data that must be handled. This explosion of data is exacerbating the already severe challenges and barriers to data sharing, handling, input/output (I/O), and saving information. Therefore, the ESM community must reconsider current paradigms to both address the fundamental

The Role of AI for Prediction

The rapid advances in AI offer increasing potential to replace portions of prediction models, and data processing systems, or even build entirely new weather forecasting systems (Pathak et al. 2022; Bi et al. 2022; Lam et al. 2023; Price et al. 2023). Model results demonstrate similar predictive skill to traditional and numerical models, requiring a fraction of the computing resources to run them.

These results are encouraging, but limitations remain (Bonavita 2024). AI models are trained with data generated from physics-based prediction models. Until recently, they exclusively relied on reanalysis datasets, while latest efforts also aim to directly include observations and other sources. Accuracy of the AI models depend on sufficient coverage and completeness of training data used. Relying on historical data to train them, AI models face challenges being able to predict climate-induced, extreme weather events that occur rarely, if ever (Ebert-Uphoff and Hilburn 2023). The behavior of AI models for such events can be unpredictable. This is of critical importance since accurate prediction of extreme weather and climate change is where the biggest impacts to society lie.

Training AI models is an essential part of building a predictive capability. Even though AI models are rapidly improving, physics-based models will continue to be essential to provide an accurate evolution of three-dimensional fine-scale weather in time and space (Bauer et al. 2023). Therefore, continued investment and development of ESMs will be critical to providing improved weather and climate predictions.

and discontinuous changes in technology and ensure that continued advances meet societal needs for more accurate weather and climate predictions.

This paper is structured as follows: section 2 identifies the key changes in technologies, as well as opportunities, precipitated by exascale computing; section 3 provides a survey of major activities underway; section 4 identifies key technological challenges that must be overcome to have effective solutions. Section 5 discusses gaps and potential ways to address them. Section 6 summarizes and concludes with recommendations to the ESM community and stakeholders.

2. A disrupted horizon

The sustained increase in HPC capacity has been instrumental to advances in numerical weather, climate, and environmental prediction over the last few decades (Bauer et al. 2015). Heavens et al. (2013) describes the advancement of ESMs with more physical, biological, and chemical processes that provide a more accurate depiction of the climate system. Forecasting systems with increasing resolution and complexity, expansion of data assimilation and ensemble approaches, oceanographic, sea ice, and hydrological coupling are examples of how the growth and availability of HPC have improved weather and climate predictions but require significantly more computing.

While there are distinct differences in the prioritization of the computational needs of the different applications,[1] the general issue of requiring massively enhanced computational and data handling capacities is similar everywhere. A current goal of the weather and climate communities is the development of global models capable of simulations having a 1–3-km resolution. Subkilometer or even large-eddy simulation (LES)-scale-limited-area models, nesting, and regional refinement are approaches to more accurately predict short-duration, high-impact weather events including heavy precipitation, fires, coastal inundation, and urban-scale events. Such models will facilitate the evolution from parameterization to direct simulation of sub-mesoscale processes including clouds, ocean eddies, surface hydrology, deep convection, and localized topographic forcing whose small-scale, transient dynamics are fundamental drivers of most weather and climate extremes (Satoh et al. 2019). This may well imply that future ESMs will

[1] In general, weather models must run in a specific period of time to be useful for short-term prediction. Time-to-solution requirements for climate prediction are less clearly defined based on many factors including model complexity, resolution, length of simulations, spinup time, and goals of the simulations.

always need to resolve at very fine scales to properly represent scale interactions that matter across forecast ranges from days to decades (Palmer and Stevens 2019). To make such advances energy-efficient and tractable, AI methods may be an effective replacement for some parameterizations and model components (Slater et al. 2023).

The increasing frequency and severity of extreme weather events around the world, combined with growing human population, necessitate improvements to the weather and climate prediction systems in order to provide timely and accurate information. Understanding changes and simulating their impacts—in terms of droughts, build environment, or food scarcity—to alleviate the growing costs of disasters in both lives and property could require more than a 100-fold increase in computational performance over the most powerful leadership-class HPC systems in use today (Schulthess et al. 2019). While development and deployment of such exascale computers are already underway, the need for green solutions to power and provide cooling for even larger HPC systems presents enormous challenges.

In general terms, an exascale supercomputer is defined as a system capable of achieving a sustained computational performance of 1 Exaflops (10^{18} floating point operations) per second, using a 64-bit floating-point arithmetic. However, exascale supercomputers are more than their computational capabilities. Figure 1 shows that in addition to computing hardware, fast memory, robust interprocessor communications, and large storage for I/O and analysis are also required.

Two aspects of computational capability are of particular importance to the weather and climate communities. The first is sustained computational performance. This requires a dedicated, high-speed network and distinguishes leadership-class HPC systems from distributed computing systems. The second aspect is the use of a 64-bit floating-point arithmetic—commonly referred to as double-precision. While 32-bit precision is used in most portions of the models, 64-bit precision is often required for physics and other areas. Recent investigations show that while ESMs can benefit from 16-bit precision, higher precision remains a requirement (Gan et al. 2013; Maynard and Walters 2019).

Until recently, most supercomputers worldwide achieved sufficient computational performance using traditional CPUs, and with few exceptions, these CPUs used the same x86 instruction set architecture. Portability across hardware from different vendors was enabled via a combination of shared-memory parallelization [Open Multiprocessing (OpenMP)], message passing interface (MPI), and standard-based programming languages—such as C, C++, and FORTRAN. Performance optimization was well understood for broad classes of science applications. This is no longer the case.

Exascale supercomputers being designed and deployed have increasingly diverse architectures: employing various combinations of many core CPUs, GPUs, and

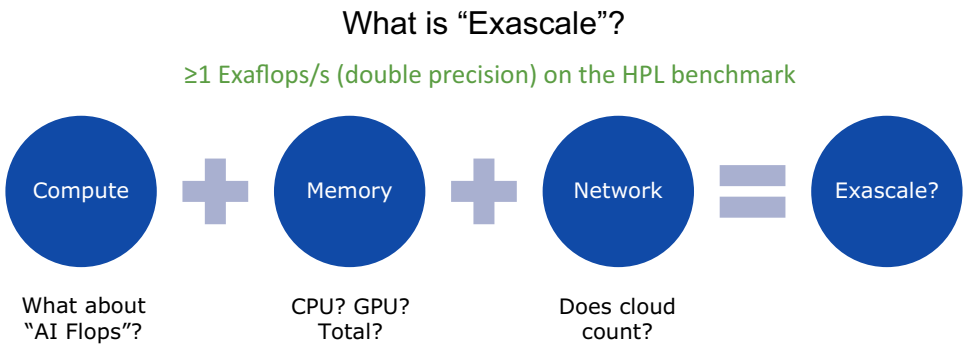

Fig. 1. Key elements of exascale include supercomputers with hundreds of thousands to millions of computational processors, hundreds of petabytes of high-speed memory, a robust system network capable of quickly moving information between processors, and large amounts of storage sufficient to support I/O and analysis requirements of the applications that run on them.

field-programmable gate arrays (FPGAs). Figure 2 illustrates two exascale systems installed in Japan and the United States with differing design approaches and hardware technologies. Developing software for such systems frequently involves the navigation of numerous, often vendor-specific, programming models and libraries. Model portability becomes a problem requiring significant expertise in many areas including—but not limited to—software engineering. Optimization of a prediction model for a specific compute architecture is an equally arduous task, where subtle design choices in GPU hardware from different vendors, for example, can lead to significant differences in performance.

Simultaneous with the shift in processor and system architectures, the volume and diversity of the output data produced with HPC systems continue to grow at rates that are equal to or faster than the computing cost. Reflecting this swell in the model output is the increase in both capacity and bandwidth of file systems associated with leadership-class computers which are costly to purchase and account for a significant and increasing percentage of HPC procurements. Further, data throughput is growing much slower than the computational performance, creating additional challenges in generating output from increasingly high-resolution models and ensembles. When models are run as part of complex workflows, I/O can create unanticipated bottlenecks in both model development (debugging, diagnostics, instrumentation) and downstream uses of the data.

Figure 3 illustrates some of the complexity in the workflows of today's operational weather prediction systems that is executed once or several times per day to produce the latest analysis (initial conditions) and forecasts. Observational data from various sources are received, preprocessed, and managed in object-based data stores to facilitate data selection, bias correction, and matching with model output. Analyses and forecasts are produced, and their output is postprocessed to generate products disseminated to a wide range of users and uses. Machine learning (in green) offers the potential to upgrade and accelerate processing across the workflow.

The massive expected growth in data will require new policies, technologies, and approaches to ingest, generate, distribute, analyze, compress, and store data. The assumption

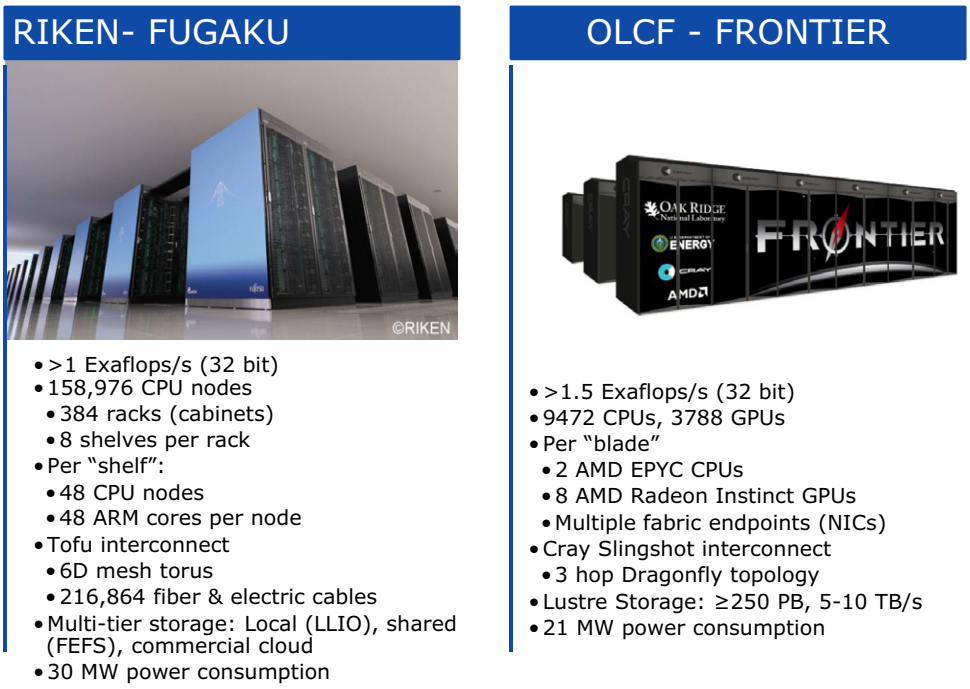

Fig. 2. Fugaku (RIKEN—2020) and Frontier (ORNL—2021) are two recently installed exascale supercomputers that illustrate the increasing hardware diversity on these systems including processors, interconnect, storage, and I/O. While Frontier is more power efficient due to the use of GPUs (21 megawatts vs 30 megawatts), power consumption on future systems is expected to continue increasing.

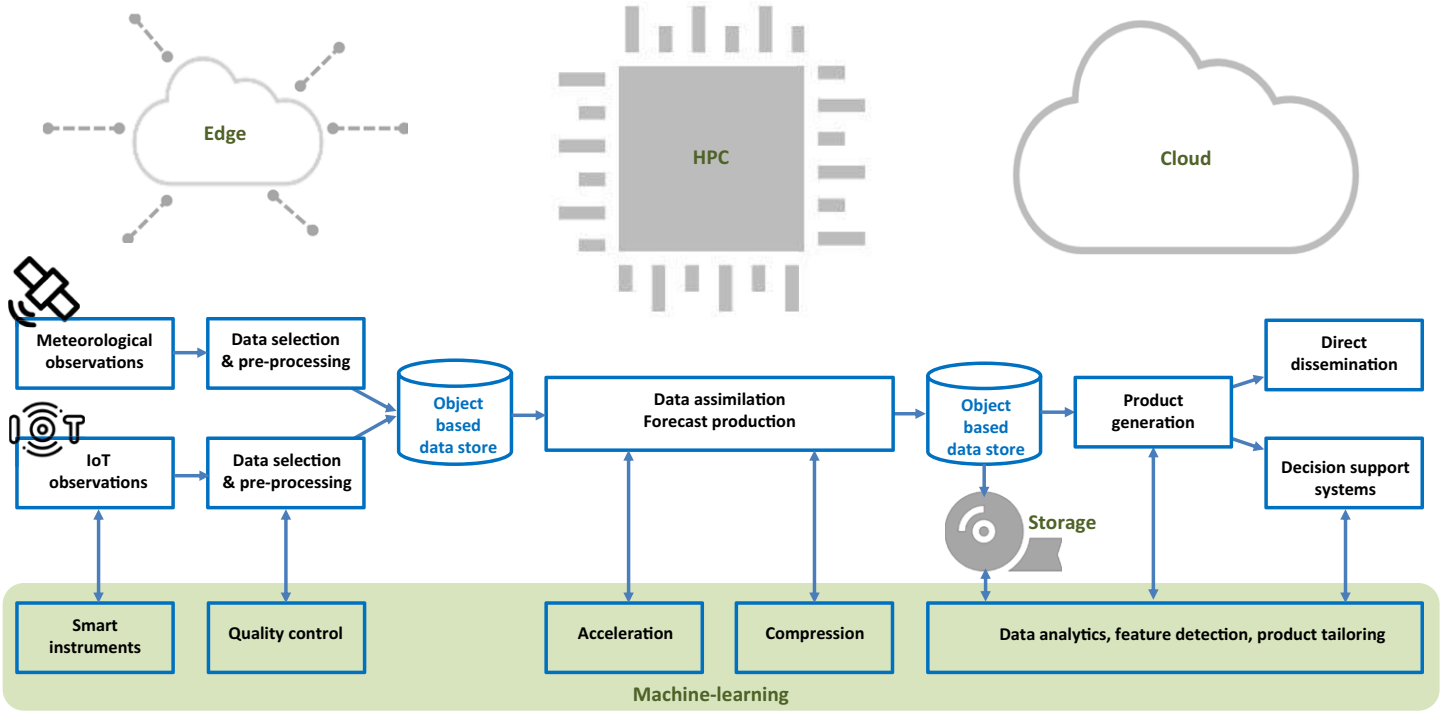

Fig. 3. Depiction of an operational workflow used for weather prediction. Workflows can be quite complex, containing hundreds to thousands of processes that are run 2–24 times per day, incorporating observation processing, data assimilation, model prediction, postprocessing, and product generation. The data may be further processed by downstream users who incorporate the data into decision support systems for specific types of guidance (e.g., fire weather, flooding, and avalanche prediction). Ideally, climate prediction shares most of the same computing/data handling components even if workflow setup and schedules are different.

that all data need to be saved is no longer possible, feasible, or cost-effective given the increasingly large share of HPC procurements devoted to data handling.[2]

[2] Climate models have been severely constrained for decades due to the complexity, simulation length, and time required to output high data volumes during a model run. Weather models have been less constrained.

3. Overview of exascale-focused activities

This section gives a survey of exascale-focused activities within the weather and climate communities. Most efforts to prepare models for exascale have focused on adapting existing codes to run on GPUs, AMD, and other types of processors. Some groups have embarked on major efforts to rewrite portions of their prediction models to address shortcomings in performance, portability, and software design.

a. Activities in the research community.

1) Europe. For many years, the European climate prediction community has invested in concerted actions to facilitate the exchange of data and models, coordinate European contributions to the Coupled Model Intercomparison Project (CMIP), and, most importantly, ensure access and support the use of European supercomputer facilities through the European Network for Earth System (ENES) modeling. Further, complementary projects have promoted the development of commonly distributed climate modeling infrastructure, shared software development, and workflow and data management to assess computing and data needs for next-generation weather and climate models.

In 2013, the European Centre for Medium-Range Weather Forecasts (ECMWF) founded its 10-yr scalability program to prepare the prediction workflow for performance, portability, and scalability challenges of the next decade (Bauer et al. 2020). Projects like energy-efficient scalable algorithms for weather and climate prediction at exascale (ESCAPE) focused on the development of novel approaches for numerical modeling, programming models for heterogeneous processor architectures, HPC benchmarks for real weather and climate prediction

workloads, and for employing machine learning to accelerate processing and support data analytics. The scalability program has spawned strong European and international collaboration on HPC, big data handling, and machine learning through 14 EU-funded projects between 2015 and 2022 and made weather and climate prediction a primary application in Europe's exascale technology roadmap. The new European Commission Horizon and Digital Europe funding programs and European High Performance Computing Joint Undertaking (EuroHPC) (delivering three European pre-exascale and two exascale HPC centers by 2021 and 2023, respectively) will provide new opportunities by 2027.

In the United Kingdom, the Met Office is midway through a next generation modeling systems (NGMS) program, which aims to reformulate and redesign its complete weather and climate research and operational/production systems for a next-generation supercomputer in the mid-2020s. The scope of the work covers atmosphere, land, marine, and ESM capabilities and includes the full processing chain from observation processing and data assimilation through the modeling components, verification, and visualization. The new atmosphere model infrastructure [U.K. Lewis Fry Richardson (LFRic)] is being developed using an approach called "separation of concerns" that relies on an in-house code generation tool called PSyclone and a domain-specific language (DSL) for the scientific code to provide performance portability (Adams et al. 2019).

In Germany, the leading climate and weather modeling centers are the Max Planck Institute for Meteorology (MPI-M), the German Climate Computing Centre (Deutsches Klimarechenzentrum) (DKRZ), the German Weather Service (Deutscher Wetterdienst) (DWD), and the Helmholtz research centers. Collectively, these centers have driven the recent development of global storm-resolving models (Zängl et al. 2015). Kilometer-scale regional modeling and regional-scale large-eddy simulations in the project High Definition Cloud and Precipitation for Advancing Climate Prediction [HD(CP)^2] have demonstrated the capabilities of the Icosahedral Nonhydrostatic (ICON) modeling system and exposed bottlenecks that must be overcome to fully exploit exascale computing power. Further, a new generation of ocean models [ICON-O and finite-element sea ice–ocean model (FESOM)] is being tested in support of DestinationEarth and other European projects.

2) Asia. The Japanese government launched the Flagship 2020 Project (Supercomputer Fugaku) in 2014 with the mission to carry out research and development for future supercomputing. In 2018, the advancement of meteorological and global environmental predictions utilizing high-volume observational data was established as an additional priority issue and exploratory challenge. New technology is being developed to make accurate predictions of those extreme weather events using ultra-high-resolution simulations and big data obtained from satellite-based observation technologies and ground radars. In June 2021, the Japan Meteorological Agency (JMA) launched a new project to accelerate the development of its global model, high-resolution regional model, and Ensemble Prediction System (EPS) for heavy rainfall disaster prevention on Fugaku, one of the largest HPC systems in the world.

In China, the Institute of Atmospheric Physics, Sugon, Tsinghua University, and the National Satellite Meteorological Center jointly developed the Earth System Science Numerical Simulator Facility (EarthLab). The EarthLab is a numerical simulation system of the main Earth system components with matching software and hardware. The global model of EarthLab has a horizontal resolution of 10–25 km and the spatial resolution of its regional nest at 3 km over China and 1 km in key areas. The system is being designed to integrate simulations and observation data to improve the accuracy of forecasting, improve the prediction and projection skills for climate change and air pollution, provide a numerical simulation platform, and support China's disaster prevention and mitigation, climate change, and other major issues.

In India, the Ministry of Earth Sciences (MoES) enables research and development activities that improve the prediction of weather, climate, and hazard-related phenomena for societal, economic, and environmental benefits. In particular, MoES supports continued the development of the Ensemble Prediction System including running at a 12-km resolution on an 8 petaflop (PF) system currently available. Large investments in HPC planned in 2022–25 (toward a 30 PF system), combined with improvements in models and workflows, are expected to significantly improve weather and climate predictions, disaster management, and emergency response in the next decade.

3) North America. A number of U.S. exascale-focused research efforts and initiatives have the goal of advancing weather and climate prediction in the next decade. One effort, funded through the U.S. Department of Energy's Office of Science, is to develop the Energy Exascale ESM (E3SM) (Leung et al. 2020) to run multidecadal, coupled climate simulations at global, cloud-resolving (1–3 km) scales. Initiated in 2014 and building on the Community ESM (CESM), a major effort to refactor and redevelop the legacy software was undertaken to enable GPU-accelerated computing. A key component in this development is the use of the Kokkos (Edwards et al. 2014) and C++ languages to enable both computational performance and portability across vendor architectures. Algorithmic changes to numerical methods were made to improve computational performance on GPUs while preserving CPU performance. A major upgrade in 2019 was the development of the Simple Cloud Resolving E3SM Atmosphere Model (SCREAM), designed to run at cloud-resolving scales on CPU- and GPU-based exascale systems (Caldwell et al. 2021).

Other U.S. efforts are focusing on the development of prediction models to run global, storm-resolving (3 km), or finer scales on exascale systems. The NSF-funded Model for Prediction Across Scales (MPAS) developed at the National Center for Atmospheric Research (NCAR) has forged a successful collaboration with IBM Weather Company and NVIDIA to port the model to GPUs.[3] The model demonstrates good performance and scaling of atmospheric components (dynamics and physics) on the Summit system at a 3-km global scale.

Researchers at NOAA's Geophysical Fluid Dynamics Laboratory (GFDL) partnered with a private company, the Allen Institute for AI (AI2), to port the Finite-Volume Cubed-Sphere Global Forecast System (FV3GFS) climate/weather model to GPUs and other advanced architectures.[4] Significant portions of the model have been rewritten in high-level python code that are transformed via software tools into optimized, architecture-specific code (Dahm et al. 2023). Some physics parameterizations are being replaced with machine learning algorithms that are orders of magnitude faster than the traditional routines.

[3] https://ncar.ucar.edu/what-we-offer/models/model-prediction-across-scales-mpas.
[4] https://www.gfdl.noaa.gov/fv3/.
[5] https://github.com/CliMA/ClimateMachine.jl.

The U.S. Naval Research Laboratory is developing a next-generation weather prediction system called NEPTUNE, which is based on the spectral-element-based Nonhydrostatic Unified Model of the Atmosphere (NUMA), a model that demonstrated exceptional CPU and GPU performance and scaling (Abdi et al. 2019). NEPTUNE adapted the NUMA dynamical core implemented for the efficient use of CPUs and GPUs. An NSF-funded effort called EarthWorks was launched in 2020 to build an exascale-ready climate model using components from CESM and the GPU-enabled MPAS Ocean model developed by the DoE. The model will utilize a uniform grid, with a goal to run on CPUs and GPUs at storm-resolving scales.

Finally, two U.S. efforts are aimed at rewriting ESMs from the ground up to utilize exascale computing, AI, and data handling technologies more effectively. The Climate Machine,[5] developed by the Climate Modeling Alliance (CliMA), is an ESM, written in the Julia programming language, that leverages advanced computational and AI technologies,

new algorithms, and data handling approaches. NOAA's Global Systems Laboratory began the development of Geofluid Object Workbench (GeoFLOW) in 2018 to explore algorithms, software techniques, performance, and portability needed for exascale-ready models. GeoFLOW uses an object-oriented framework to evaluate scientific accuracy and computational efficiency of algorithms used in finite-element models running at global cloud-resolving scales (Rosenberg et al. 2023). The development path is to progress from simpler to more complex models using the most promising algorithms, software engineering techniques, and computing technologies.

4) Australasia, Africa, and South America. In general, the lack of resources including funding, staff, and support has made it more difficult to sustain the robust development of ESMs in these regions. Centers typically rely on collaborations and partnerships with larger centers that can provide global models, data, and computing resources. For example, Australia and New Zealand are participating in the development of the LFRic model with the Met Office. Activities in Africa include the South African Weather Service (SAWS), the Council for Scientific and Industrial Research (CSIR), and the recently launched AI Research Center that is supported by the United Nations Economic Commission for Africa (UN-ECA) (Bopape et al. 2019).

Similarly, in South America, centers such as Center for Weather Prediction and Climate Studies/National Institute for Space Research (CPTEC/INPE) in Brazil provide the necessary infrastructure that enable engagement in future exascale computing and modeling. While these centers lack the resources available at large centers in Europe, North America, and Asia, direct engagements have helped these regions keep up with the latest innovations including hardware technologies, models, and data processing.

b. Activities within WMO research programs. The Working Group on Numerical Experimentation (WGNE) fosters the collaborative development of ESMs for use in weather, climate, water, and environmental prediction on all time scales and includes diagnosing and resolving shortcomings. WGNE has been aware of the evolution to more massively parallel machines with alternative chip designs for more than a decade and highlighted the need to rewrite the current generation of models.

The World Climate Research Programme (WCRP) is intended to address future challenges related to ESMS that are too large and complex for a single nation to address. One such activity, called "Digital Earths," is constructing a digital and dynamic representation of the Earth system, codevelopment of high-resolution ESMs, and the exploitation of billions of observations with digital technologies from the convergence of novel HPC, big data, and artificial intelligence methodologies. In addition to the prediction and scientific aspects, this effort recognizes the importance of investment in end-to-end capabilities including orders of magnitude increases in observations, assimilation, prediction, postprocessing, and data handling needed to deliver information to diverse users to address both near-term and long-term impacts.

c. Activities within the private sector. Industry partnerships to advance climate and weather models have been robust. For example, IBM and NVIDIA provided hardware resources, technical support, and funding to support parallelization of the MPAS model in the United States. An outcome of this effort has been a GPU-enabled variant of the MPAS, called the Global High-Resolution Atmospheric Forecasting (GRAF) model, which is being used to support customers worldwide.[6] In addition to providing HPC, heavy precipitation event (HPE) has expanded its technology offerings to embrace big data, AI, and cloud computing. Intel and AMD have established Centers of Excellence at the Argonne and Oak Ridge Leadership Computing Facilities, respectively. In Europe, several HPC-oriented

[6] https://www.weathercompany.com/global-high-resolution-atmospheric-forecasting/.

projects have direct vendor involvement, but there are also bilateral activities between centers and vendor groups such as the ECMWF-Atos (e.g., NVIDIA, Mellanox, and DDN) center of excellence.

Industry has also been using AI to improve prediction capabilities. Recent large-scale AI models demonstrating weather forecasting capabilities were developed by NVIDIA, Huawei, and Google (Pathak et al. 2022; Bi et al. 2023; Lam et al. 2023). Further, NVIDIA launched Earth-2 in 2022, an HPC system dedicated to climate prediction enhanced by AI technology and the company's OMNIVERSE software.

Increasing interest in cloud computing has led to collaborations and contracts with Google, Amazon Web Services (AWS), Microsoft Azure, and other vendors to provide increasingly comprehensive HPC and data solutions for weather and climate centers. For example, the Met Office signed a $1B contract with Microsoft to provide compute (1.5 million cores) and data (4 Exabytes) services over 10 years. Further, the system will be powered by 100% renewable energy. Such an agreement suggests an increasing opportunity for further private sector engagements.

4. Technical challenges

Within the weather and climate communities, researchers have primarily focused on model development on the scientific challenges: gaining understanding and demonstrating improved accuracy of the dynamical, physical, biological, chemical, and other processes and then mapping these science problems onto computer systems through numerical methods and algorithms. However, concurrent with these science challenges are numerous technical challenges related to software, hardware, and human factors, which must be addressed for prediction models to benefit from exascale computing.

This section distills several of the most pressing and common technical challenges. While most of these challenges are not new, their difficulty and complexity are amplified in the exascale context. Further, many of the challenges outlined are not independent: addressing (or neglecting) one issue may reduce (or increase) the difficulty of another. While the relative importance may differ in weather and climate applications, the challenges are relevant, and the constraints described affect every ESM application. Additional technological challenges that arise from the introduction of AI-based models and model components are discussed in Hines et al. (2023).

a. Cost. Estimates for the computing resources needed to run weather prediction models at global 1–3-km scales operationally range broadly from 1 to 100 million CPU cores. Such estimates depend on the many factors including resolution, type and design of the model, time-to-solution requirements, and type of hardware (processors, memory, storage, etc.). Such estimates will also depend on the speed, efficiency, and scalability of the ESM applications that run on them. A million CPU cores represent the low end of an operational (8 min per forecast day) capability, sufficient for storm-resolving (3 km) weather prediction. An estimated 100 times more computing power will be needed to run at 1-km cloud-resolving scales.

Climate projection goals are much broader than weather prediction and thus harder to estimate in terms of computational requirements. In general, runtimes of climate simulations range from 1 to 20 simulated years per day (SYPD) or more. Tradeoffs between the complexity of the models (e.g., chemistry, physics, and ocean), computing requirements, and time-to-solution must be balanced to meet a broad spectrum of research, prediction, and projection requirements.

To gain insight into the cost, two systems purchased in the United States are used as a guide. The first system, NOAA's Orion computer with 72 000 cores (1750 nodes), was purchased for 22 million USD in 2018.[7] The second system, called Derecho,

[7] https://www.noaa.gov/organization/information-technology/orion.

is a hybrid CPU–GPU system purchased by NCAR's with 2488 AMD CPU nodes and 82 NVIDIA GPU nodes (328 GPUs costing approximately 35 million USD).[8] Extrapolating the hybrid system (Derecho) to a 25 000 node CPU system (1 million cores–assuming 1 GPU = 3 CPU nodes) yields roughly 310 million USD, which is similar to the estimated cost of an extrapolated million core Orion CPU system. Based on these estimates, HPC systems 100 times larger could cost 30 billion USD or more. Such estimates do not include the cost of facilities, power, and cooling needed to run them. European estimates based on running existing models at kilometer scales have also been made (Bauer et al. 2021).

[8] https://news.ucar.edu/132907/officials-inaugurate-new-nwsc-supercomputer.

Improvements to the prediction models, including the use of AI, represent the best opportunity to improve the computational efficiency of the models and thereby reduce the cost of HPC. Deployment of cloud computing may offer benefits but does not appear to fundamentally alter the expected cost.

b. Environmental impact. Costs, power consumption, and environmental footprint, or stated differently, economic and social affordability are driving efforts to reduce emissions. The environmental impact of large-scale HPC systems must be considered, specifically CO_2 emissions associated with the generation of electricity required to power them. Clearly, this impact is highly dependent on the means of energy production. For example, using the U.S. EPA Greenhouse Gas Equivalencies Calculator,[9] the carbon footprint of a 29-MW supercomputer is over 100 000 t yr^{-1}. However, reduced or zero-emission data centers are being deployed that use cleaner sources of energy. For example, a EuroHPC pre-exascale system deployed in 2023 in Finland benefits from local hydropower generation, dry air cooling, and excess heat injection to nearby communities.

[9] https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator.

c. Software investment. The cost of designing, developing, deploying, and maintaining the software used on HPC systems is significant and often overlooked. This can include scientific software, such as applications, libraries, and visualization tools; development tools, such as compilers, profilers, and debuggers; and systems software, such as operating systems, job schedulers, and monitoring tools. Figure 4 illustrates software approaches that require investment in languages, libraries, and frameworks that are designed to improve performance, portability, and productivity.

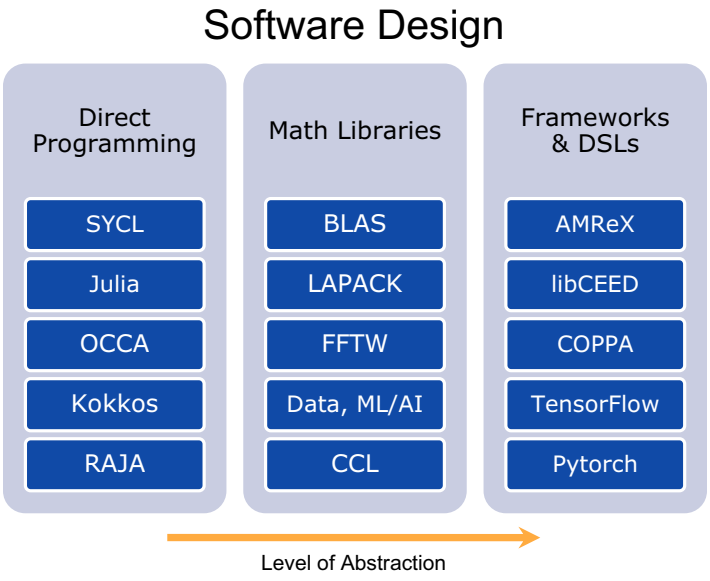

Fig. 4. Languages, libraries, frameworks, and DSLs can be deployed to improve application portability. Direct languages were designed to support CPU, GPU, and hybrid architectures at the language level. Libraries, frameworks, and DSLs increase the level abstraction (orange arrow) in the application, simplifying development and potentially improving portability and usability.

Funding required for a team of dedicated research software engineers can easily run into tens of millions of dollars per year. When this type of funding is not available, the burden of software development often falls to scientific and early career staff, including graduate researchers and postdocs. Training—and career tracks—for

professional staff whose skill sets lie along the continuum between software engineering and applied science is critical. A focused effort is needed to support software institutes, strengthen undergraduate education, and offer workshops, hackathons, and summer programs to further develop research software engineers with domain knowledge and computational skills. Training and workshops offered by ECMWF are one example of such events.[10]

[10] https://www.ecmwf.int/en/learning.

Cost estimates to adapt models to exascale systems are based on assumptions that the software has been adequately prepared. However, unless the model has been carefully designed with the most efficient algorithms and approaches, it will gain little or no benefit from additional computing resources. This is why most exascale efforts described above also investigate the best algorithmic approaches including spatial and temporal discretization, numerical solvers, and process coupling with computational efficiency and data centricity in mind.

d. Performance and scalability. Performance refers to how fast an application will run with a specific amount of compute resources. For example, operational weather models are expected to produce a 10-day forecast in 75–80 min or 7.5–8 min per forecast day. Climate models are expected to run at least five SYPD, which means century runs can be completed in 20 days and millennial runs in 200 days. Given the massive estimated computing requirements, researchers have recently suggested that one SYPD may be sufficient for short-duration (20–100 years), global, 1-km cloud-resolving climate predictions (Neumann et al. 2019).

Scalability refers to how the application behaves when more (or fewer) computing resources are used. Two types of scaling are commonly used: weak scaling and strong scaling (Hager and Wellein 2010). These metrics can be used to make realistic estimates of computing requirements if model resolution is increased from 10 to 1 km for example.

Informally, weak-scaling metrics answer the question, "will using twice the computing resources allow a problem double the size of the current one to be solved in the same amount of time?" It is particularly useful for understanding interprocess communication behavior as the model scales to higher numbers of processors. Models that require no global communications often demonstrate close to a 100% weak scaling efficiency.

Similarly, strong-scaling metrics answer the question, "Can the same problem be solved in half the time using double the computing resources?" This measure defines the term perfect strong scaling (100% efficiency) and is often used to estimate future compute requirements. However, models do not scale perfectly. In fact, as models are run at higher resolutions, scaling efficiencies will decline due to decreasing amounts of work per processor, limited parallelism, and a relative increase in interprocess communications. Overcoming such scaling issues usually requires more compute power. For example, a 50% scaling efficiency means a further doubling of compute resources (4× total) is needed to run the application in half the time. Further increases will eventually lead to performance "roll over" (0% efficiency), where more compute provides no additional benefit.

e. Model I/O. The quantity of data produced by increasingly high-resolution models and assimilation highlights problems including storage requirements, speed of I/O operations, and availability of data needed to support weather and climate workflows. Increases in model resolution, frequency of output, and number of ensemble members are key factors that drive storage requirements. For example, model output for a 3-km resolution weather model, with 192 vertical levels and output every 3 h for a 10-day forecast, is estimated to be 0.5 petabytes per model run. Increasing model resolution to 1 km would produce 64–100 times more data. Similarly, with increasing simulation length and number of fields, climate model output could easily exceed 50 PB per run.

Further, the speed in which data can be written to disk is already a major bottleneck. The classical simulation workflow, where a large set of model fields are dumped to disk and analyzed later (postprocessing), has reached bandwidth and storage capacity limits. So far, modeling groups and weather centers react by restricting ensemble sizes and limiting the number of output fields but this is not sustainable.

In the future, more flexibility may be required—for example, interacting with ongoing simulations to turn certain output fields on or off for live (during model execution) visualization or select processing of the fields necessary to run a regional flood or fire impact model. In addition, concepts under discussion include the exploration of new compression methods (Baker et al. 2016) and the use of AI to regenerate model results from archived data with lower information content (Wang et al. 2021). Within a decade, output generation is expected to become too slow, requiring new approaches such as postprocessing data in situ while the model is running.

f. Data handling. New strategies are needed to overcome expected 1000-to-10 000-fold increases from increasingly dense observations and high-resolution data assimilation, model, and ensemble output. Increasingly, high-volume data must be stored where it is generated and accessed by applications that extract, analyze, visualize, and distribute only information needed to serve application and user requests. Such a data-in-place strategy will require collocation of HPC and data storage and support for flexible, scalable mechanisms for access by automated and interactive processes.

Advanced AI systems have shown the ability to perform analysis "in-flight," which may help alleviate some of the challenges currently faced with the exponential increase in I/O. Some supercomputing centers host community filesystems, which allow the secure and seamless sharing of big data generated by large-scale simulations or experimental facilities. The Petrel data service (Allcock et al. 2019) at the Argonne Leadership Computing Facility (ALCF), for example, provides access to a 3.2 petabyte high-speed file system that can be integrated into automated workflows using Python, JavaScript, or other data science tools.

g. Productivity. Software productivity describes the ease in which users develop, test, share, maintain, and document code. Historically, scientists have led model development: making decisions about code structure, algorithms, and testing sufficient to meet project objectives. However, due to increasing scientific and computational requirements, heterogeneous computing platforms, and complex software ecosystems, model development is now an often-arduous effort.

Codesign is a more robust approach, where domain scientists and research software engineers collaborate closely on all aspects of model design and development. Figure 5 illustrates the importance of codesign and development to enable more robust applications in terms of software productivity, portability, and

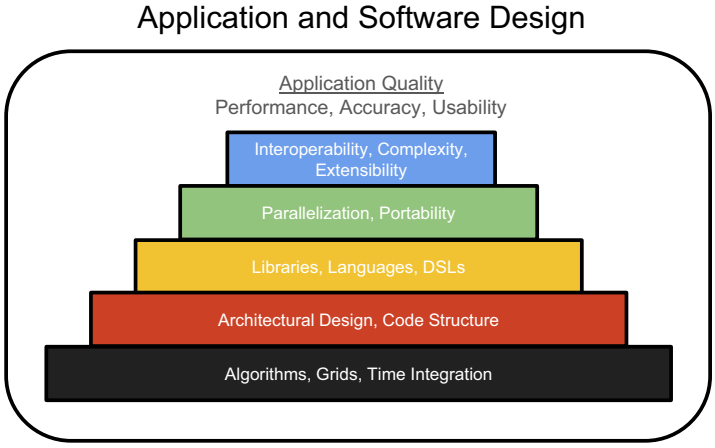

Fig. 5. An illustration of the design and software development layers within an application. The lowest layers (algorithms, design, code structure) will enable or limit the quality of the application in terms of computational performance, scientific accuracy, and usability across diverse modeling and computational systems. Quality metrics are nominally listed as computational performance, scientific accuracy, and usability of the application by the development team and community of users.

performance across diverse hardware and system architectures. The lower layers in the figure are the foundation upon which capabilities of layers above are enabled or limited. Careful selection of algorithms and software design must be considered as equally important to the languages and frameworks that are used. These decisions should be made collectively by the codevelopment teams who must balance scientific and computational requirements.

h. Portability. HPC systems are being designed with increasingly diverse hardware, combining CPUs, GPUs, and other accelerators from a variety of vendors. There are several approaches modeling teams are using to achieve performance and portability across CPU, GPU, hybrid, and other systems. The simplest approach is to use directives that inform the compiler where parallelism exists and how it can be exploited. OpenACC or OpenMP directives are inserted that minimally impact the original code. However, to get good performance, modest to substantial changes may be required. In some cases, performance portability is not possible due to the underlying algorithms, code structure, or organization of the calculations that is incompatible with the CPU, GPU, FPGAs, or other processors.

Another approach is the use of cross-platform abstraction layers—such as SYCL, Kokkos (Edwards et al. 2014), RAJA (Beckingsale et al. 2019), and OCCA (Medina et al. 2014). These require more changes to the existing application code than directive-based programming; however, code divergence is still minimal.

An extreme approach to application portability is to develop separate implementations of an application for each platform: in most cases, the cost of developing and maintaining such software makes this solution infeasible. Code divergence—which quantifies the number of lines of source code that differ between two implementations of an application that target different platforms (Harrell et al. 2018)—is a useful metric when comparing different approaches to portability since lower code divergence is associated with lower human and capital costs.

An alternative to the direct programming approaches above is the use of libraries and frameworks. Most major computer vendors provide free implementations of math libraries—such as basic linear algebra subprogram (BLAS), linear algebra package (LAPACK), and Fastest Fourier Transform in the West (FFTW)—that are highly optimized for their architectures. Since the API remains the same, few or no changes to source code are required to run on new platforms. Specialized application frameworks, such as AMReX (Zhang et al. 2021) and libCEED (Brown et al. 2021), target specific classes of discretization techniques and are also being employed to achieve portability goals.

Finally, the development of DSLs applied to the weather and climate domains is being used as a means to improve application portability, reduce complexity, and improve application performance. DSLs are often tightly linked to specific modeling centers, where support by the institution is assured. Notable use of DSLs to improve portability and productivity includes PSyclone used with the LFRric model and GridTools used with the ICON, COSMO, and FV3 models.

5. Critical gaps

This section highlights critical gaps needed to significantly advance weather and climate prediction capabilities.

a. Improvements to prediction systems. Researchers worldwide increasingly believe that new approaches are needed to gain significant improvement in prediction capabilities, not only to the model codes themselves but also include the entire prediction workflow. Transformational changes to the models must address fundamental limitations in the processors, HPC systems, prediction models, and data requirements. These challenges are highlighted

in the 2017 paper "*Position Paper on High Performance Computing Needs in Earth System Prediction*" with a call to action for vendors and model developers (Carman et al. 2017).

Reaching the same conclusion in 2018, a European consortium of Earth system and computing scientists as well as socioeconomic impact domain experts put forward the ExtremeEarth proposal aimed at a radical reformulation of Earth system simulation and data assimilation workflows to allow extreme-scale computing, data management, and machine learning on emerging and future digital technologies. The main components of ExtremeEarth have now been included in the DestinationEarth[11] project that is part of the European Commission's Green Deal.[12]

[11] https://digital-strategy.ec.europa.eu/en/policies/destination-earth.

[12] https://www.sciencemag.org/news/2020/10/europe-building-digital-twin-earth-revolutionize-climate-forecasts.

It should be noted that rapid advances in AI-driven prediction models (e.g., from NVIDIA, Huawei, and Microsoft) are demonstrating competitive skill and may fill some of these gaps.

b. Access to sufficient HPC resources. Computing requirements needed to run cloud-resolving weather and storm-resolving climate models will require systems 100 times larger than leadership class systems in use today. The cost and environmental impact of systems of this magnitude suggest that shared modeling centers dedicated to weather and climate prediction may become a necessity in the future. The distinction between research and operational centers will likely continue given the specialized requirements and critical need to produce reliable, timely weather forecasts and access to data resources, storage, and analysis tools.

Access to significant HPC resources is limited by the increasing cost of the systems and data centers themselves. Figure 6 illustrates the geographic disparity of HPC, with the majority of the TOP500 list of high-end HPC centers located in the United States, China, Europe, and Japan. Such disparities limit access to and engagement by countries that may be most affected by climate changes, for example.

c. Access to data resources, storage, and analysis tools. Development and improvement of a prediction system require large computing and storage to run simulations, evaluate results, and improve capabilities. Large centers with shared access to such resources are the most effective way for the community to collaborate and make improvements in all aspects of the prediction system. Cloud computing represents a viable technology capable of storing and sharing large amounts of data. However, more robust mechanisms are needed to organize, discover, analyze, mine, and generate information from such data.

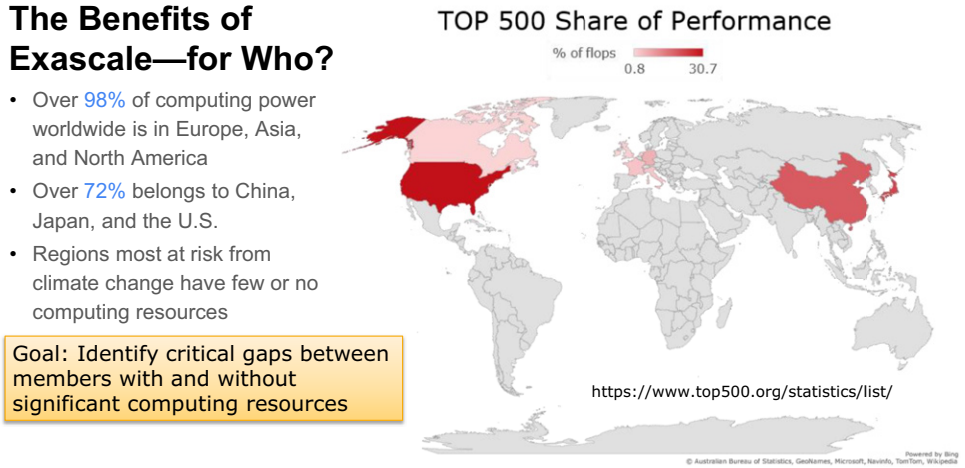

Fig. 6. The location of the top 500 HPC computing centers worldwide is illustrated. Over 98% of computing power worldwide is located in Europe, Asia, and North America. Further, over 72% belong to China, Japan, and the United States.

Given the volume of data consumed and produced by Earth system prediction models, collocation of HPC with data is expected to be essential. Shared access to such facilities will permit more effective collaborations between national and international groups. However, the cost to access high-volume data may limit open access to regions including South America and Africa that have limited resources, skills, and tools.

d. Access to highly specialized knowledge and skills. Significant and coordinated efforts are needed to address the critical shortage of qualified software professionals. Developing prediction models for increasingly diverse computing environments and leadership class HPC systems requires expertise in the science domain, applied mathematics, computer science, and software engineering. Given the disruptive changes with the HPC, AI, and associated software environments, stronger and coordinated actions with the ESM community are needed to recruit, train, and retain a workforce able to compete with the industry for the best and brightest. Coordinated actions could include the establishment of scientific software institutes, university curricula, and certifications that are specific to the needs of the ESM community.

6. Summary and next steps

Earth system modeling and prediction stands at a crossroad. Exascale computing and artificial intelligence offer powerful new capabilities to advance Earth system predictions. However, models, assimilation, and data processing systems are increasingly unable to exploit these technologies due to workforce, scientific, software, and computational limitations. Development of new prediction models is needed that incorporate the rapid and disruptive changes in HPC and the widening role of AI in models, data processing, and workflows.

This paper builds upon findings of a 2023 WMO report on exascale computing and data handling. Urgent actions are needed to overcome challenges including the enormous cost of future HPC, a 1000× projected increase in data (observations, model output), and increasing scientific and software complexity of models and applications that inhibit portability, performance, and user productivity. Technical and budgetary challenges identified are becoming too large to be addressed individually.

Comprehensive, collaborative, and sustained national-scale efforts are recommended to meet critical needs at a time of increasing societal risks. Figure 7 highlights recommendations and actions in four areas:

1) Advocate to leaders, sponsors, and stakeholders the need to address fundamental limitations in HPC, prediction models, and data handling systems that threaten continued improvements in weather and climate prediction capabilities. Urgent need for immediate action and investment is based on both societal needs for significantly improved predictions and the excellent return on investment (ROI) of such actions. In the United Sates alone, increasingly severe weather and climate disasters are costing over $100B annually.[13] Doubling or tripling funding to significantly improve prediction capabilities in the United States would represent a fraction of those costs.
2) Assess capabilities of current prediction systems to understand gaps and deficiencies and thus drive collaborations and actions by the ESM community. Estimates of computing and data requirements targeting specific weather and climate configurations (based on societal benefit) will serve to focus and justify investments.
3) Develop an action plan that brings the worldwide community together to address and collaborate on solutions that benefit the ESM community and stakeholders. As computing,

[13] According to NOAA's National Centers for Environmental Information (NCEI), the total cost of billion dollar disasters was $595B over the last 5 years (2018–22) and $1.1T in the last 10 (2013–22).

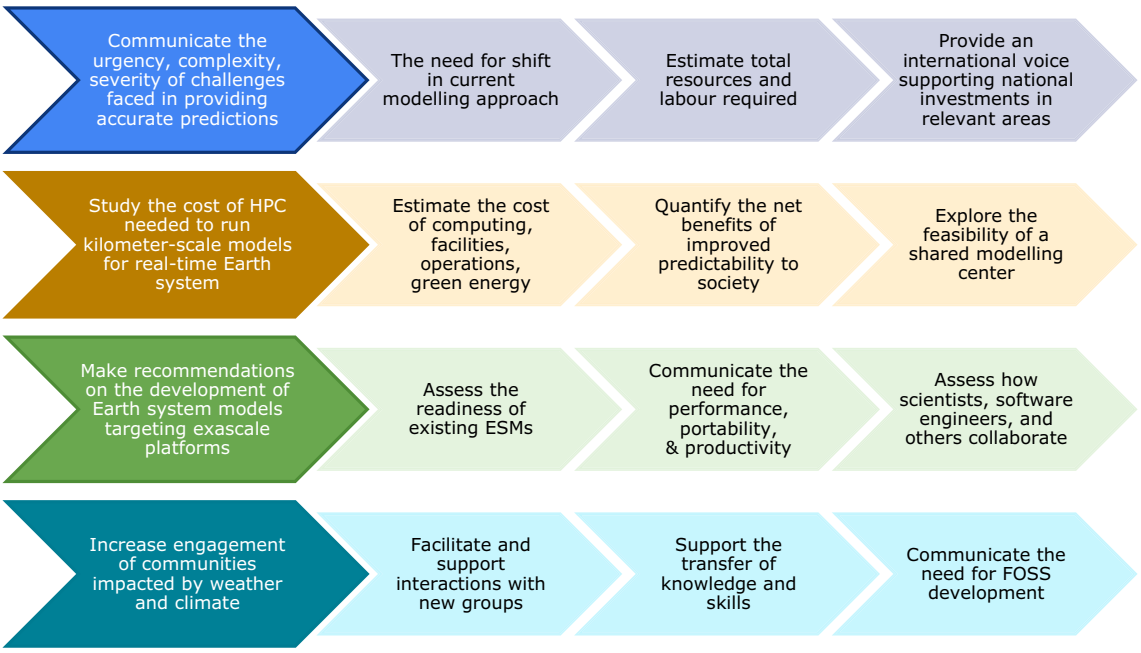

Fig. 7. An illustration summarizing an action plan on exascale computing and data handling proposed to the WMO in 2021.

data, and software complexity grow, few institutions or countries will be able to overcome the challenges alone. Strong collaborations and codesign on computing, science, software, and data will be essential.

4) Engage with industry, academic, and government partners on computing, model development, data systems to enable cost sharing, enhance data use, and improve system efficiencies. For example, foundation models and public sector datasets (e.g., reanalysis data, observations, and model data) could enable strong public–private sector partnerships on AI developments (Bauer 2023).

Assessment of current modeling and prediction systems is an important first step to both understand the capabilities and limitations of current models and determine the cost of future computing. Such computational assessments for exascale have already begun at some centers. For example, the exascale readiness assessment of United States global prediction models is being conducted as part of the Interagency Council for Advancing Meteorological Services (ICAMS) by HPC experts at the DoE, NOAA, NASA, NCAR, and the Navy. The goal is to "*determine the current state of ESMs including performance, scalability, portability, and their ability to run at fine spatial scales being targeted by leading weather and climate modeling centers*." The outcome of such comparisons could help reduce duplication, increase focus, and reduce costs.

Scientific assessments are also needed to quantify the benefit of increases in model resolution at fine scales. Such efforts are both limited by the lack of computing and the need to meet timeliness constraints. For example, Giorgetta et al. (2022) ported the ICON model to GPUs and demonstrated improvements in performance, portability, and predictability. Researchers determined that the climate model could run at a 1.25-km resolution but could not achieve a minimum time constraint of 1 SYPD even with the most advanced CPU and GPU processors. The effort helped identify where further improvements are needed.

Using exascale computing and AI effectively will require sustained efforts to design, build, and revitalize prediction models and data systems. Leadership, funding, long-term commitment, and strong collaborations will be needed to significantly improve predictions and mitigate the risks associated with extreme weather and climate change.

Acknowledgments. This research used resources of the Argonne Leadership Computing Facility, a U.S. Department of Energy (DOE) Office of Science user facility at Argonne National Laboratory and is based on research supported by the U.S. DOE Office of Science-Advanced Scientific Computing Research Program, under Contract Number DE-AC02-06CH11357.

Data availability statement. Data sharing is not applicable as the paper surveyed and referenced papers within the ESM community. No datasets were generated or analyzed.

References

Abdi, D. S., L. C. Wilcox, T. C. Warburton, and F. X. Giraldo, 2019: A GPU-accelerated continuous and discontinuous Galerkin non-hydrostatic atmospheric model. *Int. J. High Perform. Comput. Appl.*, **33**, 81–109, https://doi.org/10.1177/1094342017694427.

Adams, S. V., and Coauthors, 2019: LFRic: Meeting the challenges of scalability and performance portability in Weather and Climate models. *J. Parallel Distrib. Comput.*, **132**, 383–396, https://doi.org/10.1016/j.jpdc.2019.02.007.

Allcock, W. E., and Coauthors, 2019: Petrel: A programmatically accessible research data service. *PEARC'19: Proc. Practice and Experience in Advanced Research Computing on Rise of the Machines (learning)*, Chicago, IL, Association for Computing Machinery, 1–7, https://doi.org/10.1145/3332186.3332241.

Baker, A. H., and Coauthors, 2016: Evaluating lossy data compression on climate simulation data within a large ensemble. *Geosci. Model Dev.*, **9**, 4381–4403, https://doi.org/10.5194/gmd-9-4381-2016.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

——, and Coauthors, 2020: The ECMWF scalability programme: Progress and plans. ECMWF Tech. Memo. 857, 112 pp., https://www.ecmwf.int/en/elibrary/81155-ecmwf-scalability-programme-progress-and-plans.

——, P. D. Deuben, 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.

——, P. Dueben, M. Chantry, F. Doblas-Reyes, T. Hoefler, A. McGovern, and B. Stevens, 2023: Deep learning and a changing economy in weather and climate prediction. *Nat. Rev. Earth Environ.*, **4**, 507–509, https://doi.org/10.1038/s43017-023-00468-z.

Beckingsale, D. A., and Coauthors, 2019: RAJA: Portable performance for large-scale scientific applications. *2019 IEEE/ACM Int. Workshop on Performance, Portability and Productivity in HPC (P3HPC)*, Denver, CO, Institute of Electrical and Electronics Engineers, 71–81, https://doi.org/10.1109/P3HPC49587.2019.00012.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

Bonavita, M., 2024: On some limitations of current Machine Learning weather prediction models. *Geophys. Res. Lett.*, **51**, e2023GL107377, https://doi.org/10.1029/2023GL107377.

Bopape, M.-J. M., and Coauthors, 2019: A regional project in support of the SADC cyber-infrastructure framework implementation: Weather and climate. *Data Sci. J.*, **18**, 1–10, https://doi.org/10.5334/dsj-2019-034.

Brown, J., and Coauthors, 2021: libCEED: Fast algebra for high-order element-based discretizations. *J. Open Source Software*, **6**, 2945, https://doi.org/10.21105/joss.02945.

Caldwell, P. M., and Coauthors, 2021: Convection-permitting simulations with the E3SM global atmosphere model. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002544, https://doi.org/10.1029/2021MS002544.

Carman, J., and Coauthors, 2017: Position paper on high performance computing needs in Earth system prediction. National Earth System Prediction Capability, https://doi.org/10.7289/V5862DH3.

Dahm, J., and Coauthors, 2023: Pace v0.2: A Python-based performance-portable atmospheric model. *Geosci. Model Dev.*, **16**, 2719–2736, https://doi.org/10.5194/gmd-16-2719-2023.

Ebert-Uphoff, I., and K. Hilburn, 2023: The outlook for AI weather prediction. *Nature*, **619**, 473–474, https://doi.org/10.1038/d41586-023-02084-9.

Edwards, H. C., C. R. Trott, and D. Sunderland, 2014: Kokkos: Enabling many-core performance portability through polymorphic memory access patterns. *J. Parallel Distrib. Comput.*, **74**, 3202–3216, https://doi.org/10.1016/j.jpdc.2014.07.003.

Gan, L., H. Fu, W. Luk, C. Yang, W. Xue, X. Huang, and Y. Zhang, 2013: Accelerating solvers for global atmospheric equations through mixed-precision data flow engine. *2013 23rd Int. Conf. on Field programmable Logic and Applications*, Porto, Portugal, Institute of Electrical and Electronics Engineers, 1–6, https://doi.org/10.1109/FPL.2013.6645508.

Giorgetta, M. A., and Coauthors, 2022: The ICON-A model for direct QBO simulations on GPUs (version icon-cscs:baf28a514). *Geosci. Model Dev.*, **15**, 6985–7016, https://doi.org/10.5194/gmd-15-6985-2022.

Hager, G., and G. Wellein, 2010: *Introduction to High Performance Computing for Scientists and Engineers.* CRC Press, 356 pp.

Harrell, S. L., and Coauthors, 2018: Effective performance portability. *2018 IEEE/ACM Int. Workshop on Performance, Portability and Productivity in HPC (P3HPC)*, Dallas, TX, Institute of Electrical and Electronics Engineers, 24–36, https://doi.org/10.1109/P3HPC.2018.00006.

Heavens, N. G., D. S. Ward, and M. M. Natalie, 2013: Studying and projecting climate change with Earth system models. *Nat. Educ. Knowl.*, **4**, 4.

Hines, A., and Coauthors, 2023: WMO concept note on data handling and the application of artificial intelligence in environmental modeling. WMO Library Doc. 11573, 34 pp., https://library.wmo.int/idurl/4/66272.

Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382**, 1416–1421, https://doi.org/10.1126/science.adi2336.

Lawrence, B. N., and Coauthors, 2018: Crossing the chasm: How to develop weather and climate models for next generation computers. *Geosci. Model Dev.*, **11**, 1799–1821, https://doi.org/10.5194/gmd-11-1799-2018.

Leung, L. R., D. C. Bader, M. A. Taylor, and R. B. McCoy, 2020: An introduction to the E3SM special collection: Goals, science drivers, development, and analysis. *J. Adv. Model. Earth Syst.*, **12**, e2019MS001821, https://doi.org/10.1029/2019MS001821.

Maynard, C. M., and D. N. Walters, 2019: Mixed-precision arithmetic in the ENDGame dynamical core of the Unified Model, a numerical weather prediction and climate model code. *Comput. Phys. Commun.*, **244**, 69–75, https://doi.org/10.1016/j.cpc.2019.07.002.

Medina, D., A. St.-Cyr, and T. Warburton, 2014: OCCA: A unified approach to multi-threading languages. arXiv, 1403.0968v1, https://doi.org/10.48550/arXiv.1403.0968.

Neumann, P., and Coauthors, 2019: Assessing the scales in numerical weather and climate predictions: Will exascale be the rescue? *Philos. Trans. Roy. Soc.*, **A377**, 20180148, https://doi.org/10.1098/rsta.2018.0148.

Palmer, T., and B. Stevens, 2019: The scientific challenge of understanding and estimating climate change. *Proc. Natl. Acad. Sci. USA*, **116**, 24390–24395, https://doi.org/10.1073/pnas.1906691116.

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.

Price, I., and Coauthors, 2023: GenCast: Diffusion-based ensemble forecasting for medium-range weather. arXiv, 2312.15796v2, https://doi.org/10.48550/arXiv.2312.15796.

Rosenberg, D., B. Flynt, M. Govett, and I. Jankov, 2023: GeoFluid Object Workbench (GeoFLOW) for atmospheric dynamics in the approach to exascale: Spectral element formulation and CPU performance. *Mon. Wea. Rev.*, **151**, 2521–2540, https://doi.org/10.1175/MWR-D-22-0250.1.

Satoh, M., B. Stevens, F. Judt, M. Khairoutdinov, S.-J. Lin, W. M. Putnam, and P. Dueben, 2019: Global cloud-resolving models. *Curr. Climate Change Rep.*, **5**, 172–184, https://doi.org/10.1007/s40641-019-00131-0.

Schulthess, T. C., P. Bauer, N. Wedi, O. Fuhrer, T. Hoefler, and C. Schär, 2019: Reflecting on the goal and baseline for exascale computing: A roadmap based on weather and climate simulations. *Comput. Sci. Eng.*, **21**, 30–41, https://doi.org/10.1109/MCSE.2018.2888788.

Slater, L. J., and Coauthors, 2023: Hybrid forecasting: Blending climate predictions with AI models. *Hydrol. Earth Syst. Sci.*, **27**, 1865–1889, https://doi.org/10.5194/hess-27-1865-2023.

Wang, J., Z. Liu, I. Foster, W. Chang, R. Kettimuthu, and V. R. Kotamarthi, 2021: Fast and accurate learned multiresolution dynamical downscaling for precipitation. *Geosci. Model Dev.*, **14**, 6355–6372, https://doi.org/10.5194/gmd-14-6355-2021.

Zängl, G., D. Reinert, P. Ripodas, and M. Baldauf, 2015: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quart. J. Roy. Meteor. Soc.*, **141**, 563–579, https://doi.org/10.1002/qj.2378.

Zhang, W., A. Myers, K. Gott, A. Almgren, and J. Bell, 2021: AMReX: Block-structured adaptive mesh refinement for multiphysics applications. *Int. J. High Perform. Comput. Appl.*, **35**, 508–526, https://doi.org/10.1177/10943420211022811.