



## OPEN ACCESS

### EDITED BY

Tekin Bicer,  
Argonne National Laboratory (DOE),  
United States

### REVIEWED BY

Srutarshi Banerjee,  
Argonne National Laboratory (DOE),  
United States  
Nikolaos Tampouratzis,  
International Hellenic University, Greece

### \*CORRESPONDENCE

Dali Wang  
✉ wangd@ornl.gov

RECEIVED 31 December 2025  
REVISED 11 February 2026  
ACCEPTED 10 March 2026  
PUBLISHED 26 March 2026

### CITATION

Wang D, Gong Q, Liu Z, Wang X, Cao Q  
and Klasky S (2026) Scalable foundation  
models for numerical simulations on  
HPC platforms.  
*Front. High Perform. Comput.* 4:1778471.  
doi: 10.3389/fhpcp.2026.1778471

### COPYRIGHT

© 2026 Wang, Gong, Liu, Wang, Cao and  
Klasky. This is an open-access article  
distributed under the terms of the  
[Creative Commons Attribution License  
\(CC BY\)](#). The use, distribution or  
reproduction in other forums is  
permitted, provided the original author(s)  
and the copyright owner(s) are credited  
and that the original publication in this  
journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted  
which does not comply with these terms.

# Scalable foundation models for numerical simulations on HPC platforms

Dali Wang<sup>1\*</sup>, Qian Gong<sup>1</sup>, Zirui Liu<sup>2</sup>, Xiao Wang<sup>1</sup>, Qinglei Cao<sup>3</sup>  
and Scott Klasky<sup>1</sup>

<sup>1</sup>Oak Ridge National Laboratory, Oak Ridge, TN, United States, <sup>2</sup>Department of Computer Science,  
University of Minnesota, Minneapolis, MN, United States, <sup>3</sup>Department of Computer Science, Saint  
Louis University, Saint Louis, MO, United States

### KEYWORDS

AI, foundation model, HPC, numerical simulation, scalability

## Introduction

In recent years, foundation models (FMs) have begun to reshape numerical simulations on high-performance computing (HPC) platforms. These large, pre-trained AI models enable rapid predictions across a broad range of physical domains, including Earth system modeling, fluid dynamics, materials science, as well as complex multi-modal simulations in aerospace engineering and fusion research. By training on diverse datasets, FMs learn intricate relationships and underlying physical behavior while also enabling the quantification of uncertainty in their predictions. This capability allows simulations that once required days of numerical calculation to be completed in minutes (FM inference)<sup>1</sup>, supporting real-time design optimization, uncertainty-aware decision making, and more comprehensive exploration of complex scenarios.

Foundation models are trained on extensive simulation datasets, effectively emulating slow, iterative solvers. They serve as reusable “base models” that can be adapted to specific applications, such as inverse design and uncertainty quantification, without necessitating retraining from scratch. Consequently, FMs unlock new possibilities in engineering and scientific research, enabling swift, physics-informed optimization cycles across various domains.

The real strength of foundation models comes from their ability to learn rich physical and chemical relationships from large, diverse datasets and to assist scientists with complex, high-dimensional tasks. Rather than replacing traditional simulation or expert judgment, these models support high-throughput virtual experimentation and help researchers navigate design choices and tradeoffs that would otherwise be difficult to explore. In Earth system science, for example, models such as DeepMind’s GraphCast and NVIDIA’s FourCastNet use large spatiotemporal datasets to improve weather and climate prediction while providing actionable guidance to human forecasters. Similar approaches are emerging in other domains. In materials science, models trained on

<sup>1</sup> FMs become cost-effective when number of simulations is larger than the ratio of training time to the difference between numerical simulation and FM inference cost. It is important to note that FM inference time is usually significantly smaller than Numerical simulation time. For applications that require intensive computation, (e.g., kilometer-scale Earth system modeling may take weeks to months even on advanced computing systems), FMs serve as a practical alternative for evaluating models and quantifying uncertainties. In other scenarios, like conducting ensembles or parameter sweeps, FMs are also the preferred choice.

crystal structure databases help researchers prioritize promising candidates for further study, while in drug discovery they support protein structure prediction and molecular design. Related techniques are also being applied in fusion energy research, where data-driven models assist physicists with plasma control and scenario planning in complex operational environments.

FMs are driving a shift in numerical simulation workflows. This change parallels the long-established scientific and engineering practice of employing reduced-order or surrogate models to gain swift understanding, particularly when full-fidelity simulations are too time-intensive.

Historically, methods like Reynolds-averaged simulations have been utilized for quick engineering analysis by simplifying the underlying physics, thereby avoiding the substantial cost of repeatedly solving extensive, high-resolution models. FMs share this foundational motivation but provide a more flexible and expressive contemporary alternative. By learning from large collections of simulation and experimental data, and by leveraging architectures such as transformers and graph neural networks, they can capture complex relationships across regimes while still enabling fast, adaptable emulation. This makes them a powerful complement to traditional solvers, supporting quicker iteration and more informed decision making without sacrificing necessary fidelity.

FMs are trained predominantly with self-supervised objectives on large simulation and observational data, improving data efficiency by learning structure directly from unlabeled data. We incorporate physics-aware inductive biases (e.g., conservation, symmetries) so that the models remain consistent with governing laws and suitable for scientific use. FMs are treated as reusable emulators that complement—rather than replace—high-fidelity numerical simulations.

The remainder of this article is organized as follows: (i) AI-ready data management; (ii) FM architecture and design considerations; (iii) parallelization experiences for scaling ViT-based FMs; and (iv) training and inference acceleration on HPC; and finally a discussion that formalizes trust via a verification-and-validation protocol and presents a quantitative breakeven analysis under matched accuracy/compute budgets.

## AI-ready data management for foundation models

As FMs scale to production on HPC systems, AI-ready data is key to achieving the speedups and generalization needed. Building high-quality datasets requires programmatic pipelines for collection, quality control, de-duplication, consistent labeling, and machine-readable provenance, moving beyond generic FAIR to a campaign-oriented governance model. By recording simulation campaigns with standardized workflows and ontologies, data curation can be driven by impact metrics like time-to-insight and downstream model quality, aligning data work with FM utility rather than mere storage volume.

At exascale, storing all high-order fields is infeasible, making compression a first-class design choice. Modern error-bounded lossy methods can significantly reduce footprint (Di et al., 2025).

For FMs, fidelity must be defined by physics-aware criteria—for example, controlling errors in energy or gradients—and by quantifying how compression perturbations affect surrogates and emulators (Gong et al., 2025). Integrated *in situ* or in transit, compression also acts as a learnable encoder, yielding compact features that minimize I/O and memory during pre-training and fine-tuning while maintaining strict error controls for diagnostics (Goldman et al., 2024).

Scalable FMs also demand a unified description of scientific data across domains. We advocate a token-centric view of “AI-ready” description mirroring FM tokenization. Descriptor tokens capture meshes, coordinates, units, and provenance; feature tokens represent compressed field content. Fusion mechanisms bind tokens across simulations, experiments, and modalities into coherent scientific narratives. Anchoring these tokens in community standards (e.g., netCDF-like schemas for Earth system modeling) provides model teams with a stable, extensible interface to simulation data.

By integrating campaign governance, physics-aware compression, and token-centric schemas, data management establishes a robust feedback loop from simulation to FM pre-training and inference on HPC (Gong et al., 2025). This foundational work directly optimizes the end-to-end objective: accelerating time-to-solution and ensuring trustworthy predictions, which underpins the necessary architectural and parallelization strategies.

## Foundation model architecture and design considerations

Many numerical simulations are developed based on spatiotemporal physics laws; consequently, FMs for scientific simulation must accurately capture spatiotemporal physics while also mapping efficiently to HPC hardware. Generally, this involves utilizing graph-based models that respect irregular meshes or Vision Transformer (ViT)-based models (Dosovitskiy, 2020) that operate on regular grids. For example, on regular latitude-longitude grids, ViT-style encoders employ patch embeddings that align with GPU tensor cores and batched I/O. This alignment facilitates the computation of large-scale spatial correlations with predictable memory layouts. For spherical or unstructured domains, graph neural networks encode locality and adjacency through message passing, thereby preserving mesh topology without the need for computationally expensive remeshing. Temporal structure is commonly modeled using autoregressive or time-conditioned transformers; however, long temporal windows significantly increase attention memory, communication overhead, and checkpoint sizes—all of which represent key bottlenecks on distributed HPC systems.

Two principles align architectures with the end-to-end objectives of utilization, time-to-solution, and trustworthy predictions highlighted earlier. First, physics-aware inductive biases constrain models to respect invariants, conservation, and symmetries, improving data efficiency and reliability across regimes. Second, token- and compute-adaptivity control cost by scaling with solution complexity rather than nominal grid size.

Practical mechanisms include windowed or hierarchical attention for long-range dependencies, sparsity and mixture-of-experts for conditional computation, and adaptive token pruning or region-of-interest refinement that focuses capacity where dynamics are active. Linear- or subquadratic-attention variants further reduce the quadratic growth of standard self-attention, enabling extreme resolutions on modern accelerators.

Looking ahead, diffusion-based generative models (Song et al., 2020) provide a principled path to probabilistic forecasting, ensemble generation, and uncertainty quantification in chaotic systems, complementing deterministic surrogates. Combined with multi-resolution training curricula and schema-driven data pipelines, these choices tie model design to the same data and workflow principles used for AI-ready datasets.

## Parallelization and linear-time alternatives for foundation models

Parallelizing FMs for numerical simulation on HPC is essential for efficiency and scale, yet challenging. Increasing resolution (e.g., for climate) stresses the quadratic cost of ViT self-attention; multi-variable fields also raise memory and communication pressure. Making these models practical requires mixed-precision training and hybrid parallelization that respect interconnect bandwidth, memory residency, and checkpointing constraints.

Scaling FMs to an extreme level presents significant challenges. Experience from ORBIT-class efforts highlights four levers for ViT-based FMs on HPC: (i) linear-complexity attention (e.g., TILES) reduces self-attention from  $O(n^2)$  to  $O(n)$ , enabling multi-billion-token contexts at high resolution; (ii) hybrid parallelization (e.g., tensor/model sharding plus data parallelism such as STOP) scales to 100B+ parameters while controlling communication and memory; (iii) lightweight architectural efficiency (e.g., Reslim) improves utilization without sacrificing accuracy; and (iv) domain specialization with robust uncertainty quantification maintains physical consistency in multi-variable predictions (Wang et al., 2024, 2025). These techniques trade exact global attention for sparse/global routing and introduce modest auxiliary indices/buffers; we evaluate their accuracy–cost impact explicitly.

Complementing ViT-focused scaling, emerging linear-time generative surrogates pair diffusion models with state-space architectures (Mamba) to target  $O(N)$  complexity and lower VRAM (Mo, 2025). Mapping 1D Mamba to 2D/3D fields uses multi-directional and zigzag scanning; hybrids retain a small fraction of self-attention to preserve global coherence. These designs reduce activation/KV traffic and interconnect pressure, enable longer contexts on fixed Host Memory Buffer (HBM) budgets, and naturally improve the scalability of diffusion-model based FMs.

## Training and inference acceleration on HPC platforms

Performance of FMs is often limited more by memory capacity/bandwidth and collective communication than by peak FLOP/s. Costs for processing multi-channel 2D/3D fields over

long contexts are dominated by activation/KV-cache residency, optimizer/state footprints, tensor layout conversions, and synchronization. Training emphasizes throughput (time-to-train), while inference emphasizes latency (time-to-first-token and response time). Both must be assessed by end-to-end utilization within the simulation-to-AI loop (Kwon et al., 2023).

Several techniques mitigate memory traffic and replication. IO-aware and fused attention kernels [e.g., FlashAttention-2 Dao, 2023] reduce high-bandwidth memory (HBM) movement. Distributed training uses hybrid parallelism (data, tensor/model, and pipeline) and sharding parameters, gradients, and optimizer states to reduce replication and communication. Inference uses KV-cache management like PagedAttention to improve batching efficiency. Topology-aware placement, tensor fusion, and overlapping collectives with compute convert theoretical peak performance into sustained throughput.

Cross-layer co-design further improves efficiency. Sequence parallelism and tiling constrain activation footprints. Linear/subquadratic attention variants reduce quadratic growth. Activation checkpointing trades compute for memory, while gradient accumulation increases effective batch size. Conditional computation, such as sparse activations and mixture-of-experts, scales cost with solution complexity. Reliability is critical: mixed/low-precision training requires scaling and clipping for numerical stability, and deterministic kernels support fast recovery and reproducibility.

HPC-native runtimes, such as task-based systems [e.g., PaRSEC; Bouteiller et al., 2025], provide orchestration by expressing the workflow as dependency graphs that overlap collectives with computation, prioritizing latency-critical inference alongside throughput-driven training. Integrating these techniques with schema-driven I/O and *in-situ*/in-transit compression keeps data movement low and the simulation-to-inference loop active. By refining these strategies, we can harness the full potential of FMs in numerical simulations, delivering trustworthy predictions on production supercomputers.

## Discussion

Despite their potential, the adoption of foundation models for numerical simulations faces significant challenges, particularly regarding trustworthiness and interpretability in industrial and scientific settings.

Adoption ultimately depends on trust. Robust validation is necessary to compare FM outputs with traditional numerical solvers and observed data, quantify uncertainty through methods like ensembles or diffusion techniques, and enforce physical constraints. Establishing a continuous feedback loop in simulation workflows to detect, diagnose, and correct regime-specific errors will enhance reliability over time. Moreover, evaluation should focus on end-to-end time-to-quality and sustained utilization rather than isolated metrics. Ensuring stability under mixed or low precision requires gradient scaling and clipping, while reproducible, shard-aware checkpointing facilitates rapid recovery. Such alignment allows FMs to provide swift, credible scientific insights while optimizing supercomputing resources.

A standardized evaluation framework for assessing FM performance could greatly enhance their deployment in scientific domains. This framework should incorporate metrics such as accuracy, efficiency, and trustworthiness, mirroring the validation processes in traditional numerical simulations, which include code verification, physics consistency checks, and cross-solver benchmarking with confidence intervals (Oberkampf and Roy, 2010). Additionally, a hardware-aware simulation platform can rigorously evaluate FM pipelines across various computational architectures, establishing pre-registered acceptance thresholds (Tampouratzis et al., 2025). Understanding the tradeoff between accuracy and computational time will enable practitioners to allocate resources effectively based on their specific needs and available capacity, ultimately facilitating informed decision-making.

## Author contributions

DW: Writing – review & editing, Writing – original draft. QG: Writing – original draft, Writing – review & editing. ZL: Writing – original draft, Writing – review & editing. XW: Writing – review & editing, Writing – original draft. QC: Writing – review & editing, Writing – original draft. SK: Writing – original draft, Writing – review & editing.

## Funding

The author(s) declared that financial support was not received for this work and/or its publication.

## References

- Bouteiller, A., Hernaut, T., Cao, Q., Schuchart, J., and Bosilca, G. (2025). PaRSEC: scalability, flexibility, and hybrid architecture support for task-based applications in ECP. *Int. J. High Perform Comput Appl.* 39, 147–166. doi: 10.1177/10943420241290520
- Dao, T. (2023). Flashattention-2: faster attention with better parallelism and work partitioning. *arXiv*. [preprint]. arXiv:2307.08691. doi: 10.48550/arXiv.2307.08691
- Di, S., Liu, J., Zhao, K., Liang, X., Underwood, R., Zhang, Z., et al. (2025). A survey on error-bounded lossy compression for scientific datasets. *ACM Comput. Surv.* 57, 1–38. doi: 10.1145/3733104
- Dosovitskiy, A. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv*. [preprint]. arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929
- Goldman, O., Caciularu, A., Eyal, M., Cao, K., Szpektor, I., and Tsarfaty, R. (2024). Unpacking tokenization: evaluating text compression and its correlation with model performance. *arXiv* [preprint]. arXiv:2403.06265. doi: 10.48550/arXiv.2403.06265
- Gong, Q., Ainsworth, M., Chen, J., Liang, X., Zhu, L., Klasky, E., et al. (2025). “Stability-preserving lossy compression for large-scale partial differential equations,” in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (St. Louis: <https://impact.ornl.gov/en/publications/stability-preserving-lossy-compression-for-large-scale-partial-di/Association-for-Computing-Machinery, Inc>), 1992–2005. doi: 10.1145/3712285.3759878
- Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., et al. (2023). “Efficient memory management for large language model serving with pagedattention,” in

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author DW declared that they were an editorial board member of *Frontiers*, at the time of submission. This had no impact on the peer review process and the final decision.

## Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by *Frontiers* with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Proceedings of the 29th Symposium on Operating Systems Principles* (New York, NY: Association for Computing Machinery), 611–626. doi: 10.1145/3600006.3613165

Mo, S. (2025). “Scaling diffusion mamba with bidirectional SSMs for efficient 3D shape generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39 (AAAI Press), 19475–83. doi: 10.1609/aaai.v39i18.34144

Oberkampf, W. L., and Roy, C. J. (2010). *Verification and Validation in Scientific Computing*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511760396

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv* [preprint]. arXiv:2011.13456. doi: 10.48550/arXiv.2011.13456

Tampouratzis, N., Papaefstathiou, I., Gomez-Lopez, G., et al. (2025). Distributed fast and accurate simulation platform for advanced ARM- and RISC-V-based HPC systems. *J. Supercomputing* 81:1484. doi: 10.1007/s11227-025-07972-7

Wang, X., Choi, J.-Y., Kurihaya, T., Lyngaas, I., Yoon, H.-J., Xiao, X., et al. (2025). “ORBIT-2: Scaling exascale vision foundation models for weather and climate downscaling,” in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC 25)* (New York, NY: Association for Computing Machinery), 86–98. doi: 10.1145/3712285.3771989

Wang, X., Liu, S., Tsaris, A., Choi, J. Y., Aji, A. M., Fan, M., et al. (2024). “ORBIT: Oak ridge base foundation model for earth system predictability,” in *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis* (Atlanta, GA: IEEE), 1–11. doi: 10.1109/SC41406.2024.00007