

# ORBIT-2: Scaling Exascale Vision Foundation Models for Weather and Climate Downscaling

Xiao Wang  
Oak Ridge National Laboratory  
United States

Isaac Lyngaas  
Oak Ridge National Laboratory  
United States

David Pugmire  
Oak Ridge National Laboratory  
United States

Aristeidis Tsaris  
Oak Ridge National Laboratory  
United States

Mohamed Wahib  
RIKEN Center for Computational  
Science  
Japan

Prasanna Balaprakash  
Oak Ridge National Laboratory  
United States

Jong-Youl Choi  
Oak Ridge National Laboratory  
United States

Hong-Jun Yoon  
Oak Ridge National Laboratory  
United States

Ming Fan  
Oak Ridge National Laboratory  
United States

Ashwin M. Aji  
AMD Research and Advanced  
Development  
United States

Dali Wang  
Oak Ridge National Laboratory  
United States

Moetasim Ashfaq  
Oak Ridge National Laboratory  
United States

Takuya Kurihaya  
Oak Ridge National Laboratory  
United States

Xi Xiao  
University of Alabama at Birmingham  
United States

Nasik M. Nafi  
Oak Ridge National Laboratory  
United States

Maliha Hossain  
Oak Ridge National Laboratory  
United States

Peter Thornton  
Oak Ridge National Laboratory  
United States

Dan Lu  
Oak Ridge National Laboratory  
United States

## SYNOPSIS

Sparse observations and coarse-resolution climate models limit effective regional decision-making, underscoring the need for robust downscaling. However, existing AI methods struggle with generalization across variables and geographies and are constrained by the quadratic complexity of Vision Transformer (ViT) self-attention. We introduce ORBIT-2, a scalable foundation model for global, hyper-resolution climate downscaling. ORBIT-2 incorporates two key innovations: (1) Residual Slim ViT (Reslim), a lightweight architecture with residual learning and Bayesian regularization for efficient, robust prediction; and (2) TILES, a tile-wise sequence scaling algorithm that reduces self-attention complexity from quadratic to linear, enabling long-sequence processing and massive parallelism. ORBIT-2 scales to 10 billion parameters across 65,536 GPUs, achieving up to 4.1 ExaFLOPS sustained throughput and 74–98% strong scaling efficiency. It supports downscaling to 0.9 km global resolution and processes sequences up to 4.2 billion tokens. On 7 km resolution benchmarks, ORBIT-2 achieves high accuracy with  $R^2$  scores in range of 0.98–0.99 against observation data.



This work is licensed under a Creative Commons Attribution 4.0 International License. *The International Conference for High Performance Computing, Networking, Storage and Analysis (SC '25)*, November 16–21, 2025, St Louis, MO, USA, 2025.

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1466-5/2025/11

<https://doi.org/10.1145/3712285.3771989>

## CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Distributed artificial intelligence**; **Massively parallel algorithms**.

## KEYWORDS

Vision Transformer, Long Sequence Scaling, Weather and Climate Downscaling

## ACM Reference Format:

Xiao Wang, Jong-Youl Choi, Takuya Kurihaya, Isaac Lyngaas, Hong-Jun Yoon, Xi Xiao, David Pugmire, Ming Fan, Nasik M. Nafi, Aristeidis Tsaris, Ashwin M. Aji, Maliha Hossain, Mohamed Wahib, Dali Wang, Peter Thornton, Prasanna Balaprakash, Moetasim Ashfaq, and Dan Lu. 2025. ORBIT-2: Scaling Exascale Vision Foundation Models for Weather and Climate Downscaling. In *The International Conference for High Performance Computing, Networking, Storage and Analysis (SC '25)*, November 16–21, 2025, St Louis, MO, USA. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3712285.3771989>

## 1 JUSTIFICATION FOR ACM GORDON BELL PRIZE

ORBIT-2 sets new benchmarks in scalability and scientific impact, training ViT-based models with up to 10 billion parameters on 65,536 GPUs with 74–98% scaling efficiency and 4.1 ExaFLOPS sustained throughput in BF16 precision. It enables global climate downscaling at 0.9 km resolution and processes sequences up to 4.2

billion tokens—vastly surpassing prior limits. This breakthrough advances understanding of fine-scale extremes and improves climate modeling critical for risk mitigation and decision making.

Attributes	Contents
Category	<i>Time-to-solution, Scalability, Throughput</i>
Type of method	<i>Dense Vision Transformer Model</i>
Results reported on	<i>Whole application including I/O</i>
Precision reported	<i>BF16 mixed precision</i>
System scale	<i>Measured on Full-Scale System</i>
Measurement mechanism	<i>Timer, Static analysis tool, FLOP counts</i>

2 PROBLEM OVERVIEW

Many regions lack dense ground-based observational networks, hindering early warning systems, disaster risk mitigation, and climate adaptation planning. In such cases, global climate models provide a crucial alternative, simulating atmospheric processes at planetary scale. However, their coarse resolution limits the representation of fine-scale phenomena, constraining accuracy at regional levels.

Downscaling bridges this gap by translating coarse resolution global climate model into fine-scale outputs [8]. This process is critical across many sectors, including agriculture [21], water resources [12], infrastructure planning [10], energy systems [11], and extreme event forecasting [16]. Despite its importance, downscaling remains both scientifically and computationally challenging. It requires physically consistent predictions from massive, high-dimensional spatiotemporal data while maintaining accuracy across diverse regions. These challenges highlight the urgent need for scalable, high-fidelity downscaling approaches to enable effective climate services, disaster preparedness, and policy planning.

Traditional downscaling methods are either dynamical [9], which uses nested physical models to simulate fine-scale processes but is computationally intensive, requiring days or weeks on supercomputers and limited to regional domains; or statistical [48, 49], which is less computationally demanding but often lacks physical fidelity and generalizability. More recently, artificial intelligence (AI) has emerged as a powerful alternative, providing high-resolution predictions at a fraction of the computational cost—even on edge devices [23, 24, 42]. Most existing AI approaches, however, are task-specific deep learning models [1, 2, 6, 27] that map coarse to fine resolution, requiring retraining for each variable, resolution, or region, and often struggling to generalize across diverse, physically distinct climate variables [46].

To overcome the limitations of task-specific models, recent efforts have introduced foundation models such as Prithvi [36] and ClimateLearn [34], which employ multi-task Vision Transformer (ViT)-based architectures to support downscaling across variables and geographic regions. While these models represent an important step forward, they remain constrained by resolution limits, computational cost, and model scalability.

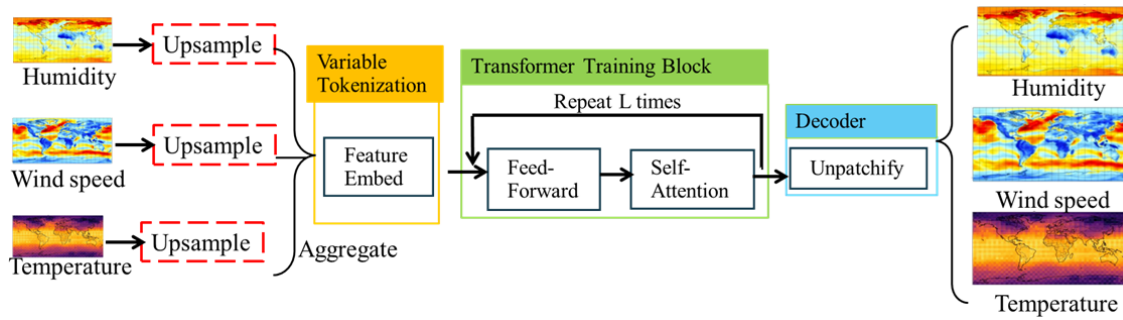
A primary bottleneck is the computational complexity of downscaling at high resolution. For instance, Prithvi achieves 12 km resolution over Europe but is restricted to 50–60 km globally due to the quadratic scaling of ViT self-attention [13]. ViTs divide spatial data into patches, treating each patch as a token, and self-attention

computes pairwise interactions among all tokens. As resolution increases, the number of token pairs grows quadratically, resulting in quadratic growth in memory and compute demands. Unlike Natural Language Processing (NLP) models, which operate on one-dimensional text sequences and scale to over one million tokens [20], ViTs handle high-dimensional spatial inputs with complex dependencies across multiple axes, making long-sequence scaling significantly more computationally intensive. As a result, the longest ViT sequence reported to date is limited to 188k tokens [40], directly constraining the maximum resolution, as sequence length scales proportionally with spatial resolution.

Another major challenge in downscaling is the inherent uncertainty, as mapping coarse-resolution data to fine scales is a highly ill-posed inverse problem: a single coarse input can correspond to many plausible fine-scale solutions, making the mapping non-unique and sensitive to small perturbations. This challenge is further amplified when downscaling multiple climate variables simultaneously. Unlike super-resolution tasks in computer vision [25, 30], where Red-Green-Blue channels represent the same physical quantity, climate variables such as temperature, humidity, and wind are governed by distinct yet interrelated physical processes. This heterogeneity increases the difficulty of learning consistent mappings and exacerbates uncertainty in predictions. A common mitigation strategy is to upsample coarse inputs prior to training [34, 36], which can help reduce uncertainty but significantly increases sequence length and, in turn, computational cost due to ViT’s quadratic complexity. Moreover, upsampling introduces artifacts that can propagate through the model, limiting its effectiveness.

Existing foundation models also face restricted model scale. For example, Prithvi [36] is constrained to 1.4 billion parameters, primarily due to the computational difficulty of scaling ViTs for high-dimensional spatiotemporal data. A major advancement is the Oak Ridge Base AI foundation model for Earth System Predictability (ORBIT) [46], which leverages hybrid sharding and orthogonal parallelisms to scale ViTs to 113 billion parameters—five times larger than previous ViTs and more than 100× larger than typical climate models. However, ORBIT is specifically designed for temporal forecasting and does not address spatial downscaling. In particular, it does not resolve the ViT long-sequence bottleneck nor mitigate the uncertainty associated with inverse downscaling problems.

To overcome these limitations, we introduce ORBIT-2, a scalable and efficient foundation model for climate downscaling. Unlike ORBIT, which focuses on temporal weather forecasting, ORBIT-2 addresses a fundamentally different challenge: high-resolution spatial downscaling. At its core, ORBIT-2 features a novel ViT architecture, *Residual Slim ViT (Reslim)*, specifically designed to bypass the high computational cost associated with traditional upsampling-based approaches. Unlike existing models that upsample inputs to mitigate uncertainty—resulting in quadratic increases in memory and computation—Reslim operates directly on adaptively compressed spatial inputs, significantly reducing sequence length while preserving critical information. It preserves accuracy and reduces uncertainty through a lightweight residual learning architecture, enabling efficient, low-overhead predictions. Additionally, both training and inference are framed as a *Bayesian Estimation* problem, incorporating a Markov Random Field Total Variation prior to further constrain uncertainty and improve spatial consistency.



**Figure 1: A generalized AI architecture diagram for state-of-the-art downscaling foundation models. Note that upsampling is used for each channel prior to training blocks to reduce downscaling uncertainty.**

Complementing this architecture is the *Tile-Wise Sequence Scaling Algorithm (TILES)* that reduces ViT’s self-attention complexity from quadratic to linear. It works by dividing images into overlapping tiles, each processed in parallel on separate Graphical Process Units (GPUs) using localized self-attention. Each tile’s downsampled outputs are then seamlessly merged to the full image. This strategy enables efficient and scalable ViT-based downscaling, making ultra-high-resolution, global-scale applications computationally feasible.

Leveraging the above innovations, ORBIT-2 sets a new benchmark through four key breakthroughs:

- **Efficient Reslim Architecture** by operating directly on compressed inputs, achieving over 660× speedup compared to standard ViTs—without compromising accuracy.
- **Longest ViT Sequence Length** by scaling ViT sequence lengths to unprecedented levels—up to 4.2 billion tokens for a 9.5M parameter model and 671 million tokens for a 10B model—surpassing the prior state-of-the-art of 188K tokens by several orders of magnitude [40]. This eliminates the long-standing sequence bottleneck, enabling global downscaling at resolutions as fine as 0.9 Kilometer (km).
- **Scalable Large Model Training** by training models with up to 10 billion parameters across 65,536 GPUs, achieving 74–98% strong scaling efficiency and sustained throughput of up to 4.1 ExaFLOPS in BF16 precision.
- **State-of-the-Art Accuracy** achieving  $R^2$  scores of 0.98 for precipitation and 0.99 for temperature at 7 km resolution over both the continental United States and global modeling, setting a new standard in high-fidelity downscaling.

### 3 BACKGROUND & STATE OF THE ART

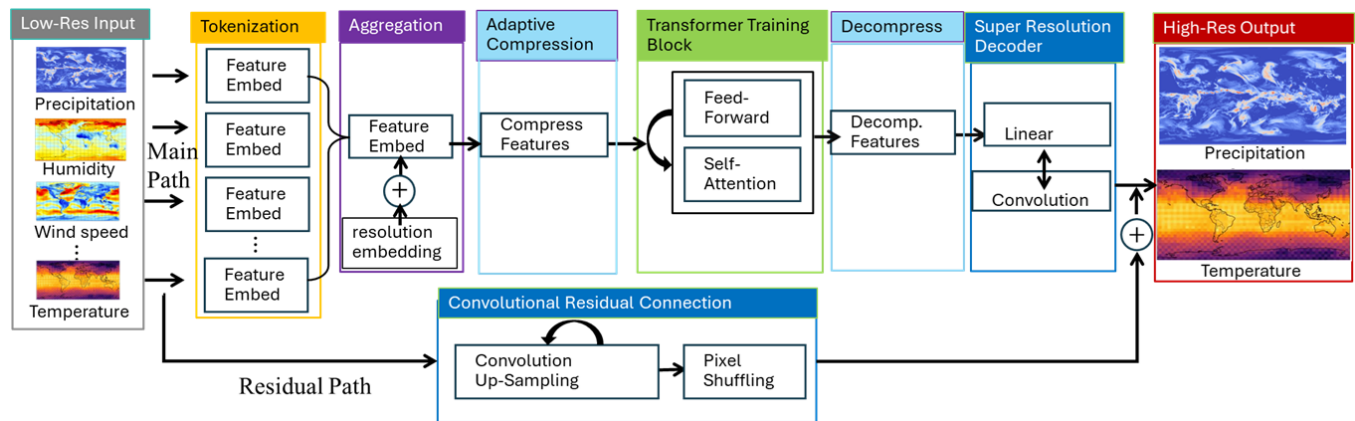
Fig. 1 illustrates the generalized architecture of leading downscaling foundation models, including Prithvi [36] and ClimateLearn [34]. The inputs consist of low-resolution data with multiple atmospheric physical variables, normalized and bias corrected, and each channel of the architecture reads data for a distinct variable. To address downscaling inverse problem uncertainty, current models upsample coarse-resolution inputs, either via interpolation [34] or convolution [36], before training. This upsampling process is crucial, as it provides a higher-resolution baseline for ViT training, mitigating uncertainty from the inherently ill-posed nature of the multi-variable downscaling problem, thereby improving accuracy and uncertainty. Once upsampled, multi-channel inputs are aggregated

into a single-channel representation in feature space, a step that can be performed using either cross-attention mechanisms [33] or shallow convolutional layers [34, 36]. This aggregated representation is then trained by the ViT training blocks, consisting of self-attention and feedforward sub-layers. Finally, the trained output is projected back from feature to image space for each individual physical variable.

This approach, however, introduces major challenges. Upsampling coarse-resolution input data before training increases the sequence length, which increases in proportion to the resolution increase, causing a quadratic increase in memory and computations due to ViT’s self-attention mechanism. This severely limits scalability and resolution, leaving the long-sequence bottleneck unresolved. Prithvi, for example, is limited to relatively coarse 50-60 km resolution for global downscaling. To address this, prior work proposed both AI architecture and scaling algorithm solutions.

**Architecture solutions.** To mitigate this, architectures like Swin Transformer alleviate some of the computational burden by introducing a hierarchical architecture with shifted window attention [28, 29]. Instead of processing the entire image at once, Swin Transformer partitions the image into smaller, non-overlapping local windows, where self-attention is computed independently within each window. To capture global spatial dependencies, features learned from local windows are aggregated into global features through an architecture hierarchy. While this reduces computing complexity, Swin Transformer has fundamental limitations and its layers of architecture hierarchy must scale proportionally with higher resolution, making it unsuitable for foundation models that needs a single model to generalize across diverse datasets with varying resolutions. Additionally, Swin Transformer’s model size grows with the architecture hierarchy, shifting the computational bottleneck from long-sequence processing to large-model scaling. Consequently, Swin Transformer can only scale up to 147K sequence length on standard 3-channel images [28], far below what is needed for high-resolution, multi-variable downscaling.

Other sparse attention architectures, such as MaxViT [41], attempt to mitigate computational cost by sampling self-attention computations. While this reduces complexity, it comes at the expense of accuracy degradation when the sampling ratio is too high, and it does not address the fundamental quadratic complexity long-sequence problem.



**Figure 2: Reslim architecture is split into main and residual paths. No upsampling is used for the main path for ViT training, leading to reduced computations. Residual path is used to condition prediction for reduced uncertainty.**

Scaling algorithm solutions. Besides architecture innovations, scaling algorithms, such as sequence parallelism [31, 40, 47], has been proposed as an alternative strategy for scaling ViT sequence length. It distributes image patch tokens across GPUs for parallel computing, alleviating memory constraints. However, because self-attention requires each token to interact with all other tokens from every other GPU, sequence parallelism incurs substantial inter-GPU communication overhead and limits its scalability. More critically, it does not resolve the fundamental quadratic complexity, which causes computational costs to grow rapidly with increased downscaling resolution. As a result, current ViT sequence parallelisms are limited to a maximum of 188K token sequence lengths [40], which remain insufficient for high-resolution multi-variable downscaling.

It is also important to note that other commonly used parallelisms—such as Fully Sharded Data Parallelism (FSDP) [51], Tensor [37], pipeline [14, 18, 22] and hybrid sharded parallelisms [46] are all designed to scale model sizes, rather than long sequences of high-resolution and high-dimensional spatial data. Consequently, none of the existing model parallelisms overcome the long-sequence bottleneck in ViTs required for high-resolution global downscaling and there is an urgent need to develop computing efficient and massively parallel architecture and scaling algorithm.

## 4 INNOVATION REALIZED

### 4.1 Reslim Architecture

Unlike existing foundation models that rely on input upsampling to establish downscaling baselines, which leads to increased sequence length and high computational cost, ORBIT-2 introduces Residual Slim ViT (Reslim), a highly efficient architecture that significantly reduces training time and memory usage without compromising accuracy. The key innovation of Reslim is its ability to operate directly on low-resolution and adaptively compressed inputs, drastically reducing sequence length and computational burden. To counteract the uncertainty typically introduced by bypassing upsampling prior to ViT training, Reslim incorporates Bayesian estimation and a residual convolutional learning path, enabling high accuracy while maintaining efficiency. Its non-hierarchical design further

promotes generalization across datasets with varying spatial resolutions, making it well-suited for scalable, foundation-level Earth system modeling.

Main ViT Path. Fig. 2 illustrates the Reslim architecture. After tokenizing each low-resolution physical variable into feature embeddings, the model proceeds along two architectural paths: the main ViT and residual paths. Crucially, the main path eliminates input upsampling, avoiding the sequence length inflation and the quadratically increased computing cost typical of ViT architectures.

First, the main path uses a cross-attention module to aggregate multi-variable embeddings into a unified representation, effectively collapsing the variable dimension. A learnable resolution embedding encodes the desired output resolution and is added to the feature embedding, enabling resolution-aware predictions—an essential capability for modeling resolution-dependent Earth system behaviors. Next, an optional adaptive spatial compression module, which will be explained further in the next paragraph, reduces the sizes of the embeddings before they are passed through ViT training blocks. When enabled, this module compresses spatial features; otherwise, it acts as an identity function. After processing, a decoder comprising convolutional layers and linear projections reconstructs the high-resolution output.

Adaptive Spatial Compression. Our objective is not only to train directly on low-resolution inputs, but also to further reduce token count and computational cost through compression. Reslim achieves this via an adaptive spatial compression technique, inspired by adaptive image patching and mesh refinement methods [50]. After aggregating multi-variable features (purple block in Fig. 2), the model projects the embedding back into image space and recursively partitions it into spatial quadrants using a quad-tree structure. Partitioning continues for any quadrant where the estimated feature density—computed via Canny edge detection—exceeds a predefined threshold, terminating when a minimum patch size is reached or below predefined threshold.

This approach enables finer-grained learning in feature-rich regions through smaller patches, and coarse-grained learning to smoother regions through larger patches, where less detail is needed. Fig. 3 illustrates an example image after variable-aggregated features are mapped back to image space. Compared to conventional



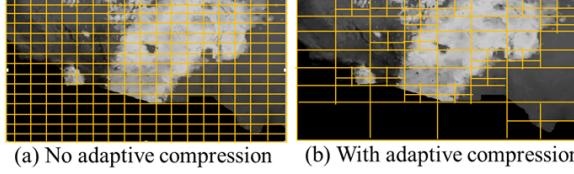


Figure 3: Comparison with and without adaptive spatial compression. Each yellow grid is an image patch.

uniform patching (Fig.3(a)), where each grid represents an image patch token, the adaptive spatial compression method (Fig.3(b)) reduces the number of patch tokens by 7x in this figure example, significantly decreasing sequence length and computing cost. After ViT training blocks, the decompression module reconstructs the high-resolution output from the compressed embeddings.

**Residual Learning.** Reslim improves computational efficiency by removing the upsampling step from the main ViT path and training directly on low-resolution, spatially compressed inputs. This design dramatically shortens sequence lengths and reduces the quadratic computational cost typically associated with ViT training. However, bypassing input upsampling introduces uncertainty, as conventional foundation models rely on upsampled inputs to provide a coarse downscaling baseline. Reslim addresses this challenge through two complementary innovations: residual convolutional learning and a Bayesian estimation objective.

The residual convolutional path reintroduces upsampling outside the main ViT path, using lightweight convolutional layers with linear complexity. This path generates a high-resolution approximation that is added to the ViT output before loss computation. Such design yields two major benefits: (1) it avoids the expensive quadratic cost of increasing the ViT sequence length due to upsampling. The upsampling is moved to the residual path, where convolutional layers have linear complexity to input size and thereby upsampling in the residual path incurs minimal computing cost. (2) it simplifies the learning task by letting the ViT focus on predicting the residual difference between the convolutional approximation and the ground truth, rather than the full downscaling transformation. This soft constraint stabilizes training, enhances physical plausibility, and significantly reduces downscaling uncertainty. As a result, Reslim achieves high downscaling accuracy with significantly reduced computations compared to conventional ViT.

**Bayesian Training Loss.** To further reduce uncertainty and improve accuracy, Reslim reformulates its training as the following Bayesian optimization problem with a Generalized Markov Random Field Total Variation prior:

$$\hat{x} \leftarrow \underset{\hat{x}}{\operatorname{argmin}} \|y - \hat{x}\|_D^2 + \sum_{k=1}^K \sum_{i=1}^N \sum_{j \in C(x_{k,i}^{\hat{x}})} b_{i,j} \|x_{k,i} - x_{k,j}\|,$$

where  $y$  is the high-resolution ground-truth,  $\hat{x}$  is the Reslim prediction, and  $D$  is a latitude weighting matrix to account for the decrease in longitudinal spacing toward the poles.  $K$  is the number of output variables, and  $N$  is the total number of pixels per variable. The neighborhood  $C(x_{k,i}^{\hat{x}})$  contains all spatial neighbors of pixel  $x_{k,i}^{\hat{x}}$ , which is the  $i^{\text{th}}$  pixel for the  $k^{\text{th}}$  variable.  $b_{i,j}$  is a spatial weighting factor inversely proportional to the euclidean distance between each pixel pair in the same neighborhood. In the above formulation, the first term,  $\|y - \hat{x}\|_D^2$ , is the Bayesian forward data likelihood

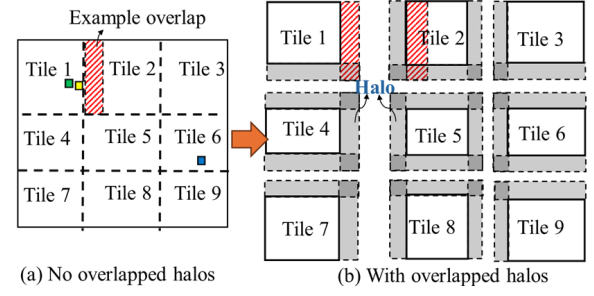


Figure 4: (a) TILES algorithm partitions inputs and outputs into tiles, each processed independently on a GPU. Yellow/green pixels denote strong local correlations, while green/blue indicate weaker long-range correlations. (b) Gray halos are added around tiles to avoid border artifacts. Boundary regions (example in red) are duplicated, appearing once in a tile’s halo and once in its neighbor’s interior.

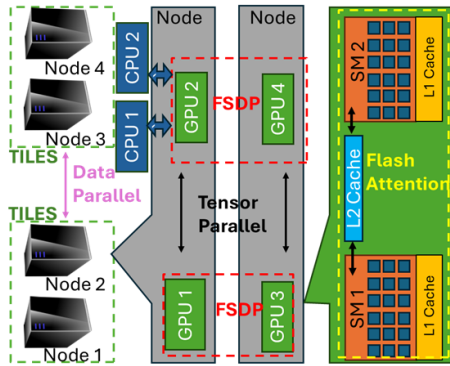
term using a latitude-weighted mean squared error. The second term is a total variation spatial prior, promoting spatial smoothness by penalizing irregularities within local neighborhoods, but also preserving edges and discontinuities. This makes it well suited for downscaling tasks for spatial coherence and structure preservation.

## 4.2 TILES: Tilewise Efficient Sequence Scaling Algorithm

While Reslim significantly reduces computation by operating on low-resolution and compressed inputs, it does not resolve the inherent quadratic complexity of self-attention. As resolution increases, this limitation becomes a bottleneck. To address this, we introduce the Tilewise Efficient Sequence Scaling Algorithm (TILES), a scalable sequence processing strategy that reduces attention complexity from quadratic to linear and enables efficient parallelization.

TILES is motivated by the spatial locality property of downscaling, where the downscaling for each high-resolution pixel is primarily influenced by spatially nearby coarse-resolution inputs. This “point spread” effect, well studied in the remote sensing literature [17, 43, 44], shows that the correlation between pixel pairs decay rapidly when the spatial distance. Consequently, long-range pixel interactions that have weak point spread effect can be safely ignored without affecting downscaling accuracy. For example, neighboring green and yellow pixels in Fig. 4(a) have high mutual influence, while distant pairs (e.g., green and blue) contribute minimally or no influence to each other’s predictions.

Leveraging this property, TILES partitions both inputs and downscaled outputs into spatial tiles, assigning each tile to a separate GPU. Each GPU then performs downscaling separately for its assigned tile, and self-attention is restricted within each tile, preserving local context while ignoring long-range dependencies across tiles. This tilewise downscaling reduces self-attention complexity from quadratic to linear. More specifically, the computation complexity is  $O(\frac{N^2}{T})$ , where  $N$  is the number of image patch tokens and  $T$  is the number of tiles. For fixed-size tiles,  $T$  increases proportionally with  $N$ , making the overall complexity linear.



**Figure 5: Orthogonal levels of parallelisms mapped to supercomputer at cluster, node, and device levels.**

However, strict tiling introduces border artifacts, as pixels near the border of each tile lack context from neighboring tiles. To mitigate this, TILES introduces halo padding, where each tile is extended with a fixed-width halo (shown in gray in Fig.4(b)) that overlaps adjacent tiles. For instance, the red boundary region in Fig.4(a) between Tiles 1 and 2 is duplicated, with one copy in Tile 1's halo and another copy in Tile 2's non-halo interior, both shaded in red in Fig.4(b). This duplication restores spatial continuity across tile boundaries, ensuring that border pixels—such as the yellow pixel—receive complete local context. Note that the halo width is tuned empirically. While in principle it can be derived directly from the data sensor's point spread function, in practice we begin with minimal halos and increase their size iteratively until both boundary artifacts and training loss are minimized.

After each GPU independently downsamples its tile, the halo regions are discarded, and the non-padded tile outputs are stitched together to form the final high-resolution output. Since each GPU processes a different tile, leading to different gradient and model parameter update, gradients from all GPUs are averaged to maintain the model consistency across GPUs and global optimization. This inter-GPU communications, however, have minimal communication overhead as it takes place only once per data batch.

### 4.3 Orthogonal Parallelisms

TILES efficiently scales sequence length, enabling high-resolution downscaling. However, it does not address model size scaling. To support both large foundation models and high-resolution downscaling, TILES must be integrated with complementary model-parallel strategies. Since TILES and model parallelism target orthogonal goals, with TILES for sequence length and model parallelisms for model size, they can be combined seamlessly. This results in a unified framework with four distinct parallelism strategies:

- **TILES sequence parallelism:** Distributes long sequence lengths for ViTs for tilewise approximation as discussed before. Requires least communication overhead.
- **Fully Sharded Data Parallelism (FSDP)** [51]: shards both data and model parameters across GPUs, but requires temporarily gathering the full model during forward and backward passes. Requires moderate communication overhead.

- **Tensor Model Parallelism** [37]: Only shards model parameters and keep parameters sharded throughout training. Requires most communication overhead.
- **Distributed Data Parallelism (DDP)** [26]: distributes only the training data without sharding model parameters. Requires least communication overhead.

Fig. 5 shows how these orthogonal parallelisms map to a supercomputer hardware. Two adjacent nodes form a TILES sequence parallel group (green dashed boxes), responsible for scaling sequence lengths. Multiple sequence parallel groups form a DDP group, distributing data batches across the system.

Within each sequence parallel group, GPUs participate in both tensor model and FSDP parallelisms for model scaling. Tensor model parallelism operates within a node, leveraging its low-latency interconnect to mitigate communication overhead for hidden dimension partitioning. FSDP (red dashed boxes) spans GPUs across neighboring nodes within the same TILES group, enabling parameter and data sharding.

Once all parallelisms are established, each GPU receives a subset of the model and data that can be optimized with Flash Attention [5] to reduce memory use and cache misses, as detailed in the next subsection. Within each GPU, streaming multiprocessors (SMs) are mapped to Flash Attention cache blocks, executing vector operations in parallel within each block. Meanwhile, CPUs asynchronously load data and construct quad-trees to track the spatial layout of adaptively compressed patches described in Fig. 3.

Note that this multi-level strategy aligns the parallelism hierarchy with the hardware architecture to optimize performance. Neither DDP nor TILES sequence parallelisms requires frequent communication, and are therefore mapped to cluster nodes with slower network communication. Tensor and FSDP model parallelisms requires more frequent communication, and are therefore mapped to GPUs within the same node and across neighboring nodes to utilize their faster in-node and neighboring-node network communications. The Flash Attention require the most frequent communication and are therefore mapped to SMs within the same GPU, which has the fastest network through shared L2 cache.

### 4.4 Optimizations

To further boost performance, we applied these optimizations:

**Hybrid-OP Parallelism.** We adopt the Hybrid-OP optimization technique from ORBIT [46], which leverages the mathematical structure of matrix chain multiplication to shard model parameters in alternating row and column dimensions. This optimization combines tensor model parallelism with FSDP, achieving superior scalability with reduced communication overhead and frequency compared to without Hybrid-OP.

**Flash Attention.** To accelerate self-attention computation, we use Flash Attention [5], which applies a cache-blocking technique to minimize memory access to GPU global memory. By maximizing data reuse from high-bandwidth on-chip cache, Flash Attention significantly improves compute throughput through higher cache hit rates and faster memory access.

**Mixed Precision and Layer Wrapping.** We further utilize BFLOAT16 mixed-precision to speed up training while reducing memory usage. To address numerical instability—where gradients with extreme

Dataset	Type	Interval	Region	Res. (km)	In/Out Vars	Sample Dim. (in→out)	# Pairs	Size (GB)
<b>Pretraining</b>								
ERA5 → ERA5	Reanalysis	Hourly	Global	622→156	23 / 3	[32,64,23]→[128,256,3]	367,920	200
ERA5 → ERA5	Reanalysis	Hourly	Global	112→28	23 / 3	[180,360,23]→[720,1440,3]	367,920	6,328
PRISM → PRISM	Observation	Daily	U.S.	16→4	7 / 3	[180,360,7]→[720,1440,3]	14,235	189
DAYMET → DAYMET	Observation	Daily	U.S.	16→4	7 / 3	[180,360,7]→[720,1440,3]	14,946	200
<b>U.S. Regional Fine-Tuning for Tmin, Tmax and Precipitation</b>								
[ERA5, DAYMET] → DAYMET	Reanalysis & Observation	Daily	U.S.	28→7	23 / 3	[180,360,23]→[720,1440,3]	14,946	254
<b>Global Precipitation Fine-Tuning</b>								
[ERA5, IMERG] → IMERG	Reanalysis & Observation	Daily	Global	28→7	23 / 1	[720,1440,23]→[2880,5760,3]	8,401	745

**Table 1: Datasets used for pretraining, fine-tuning, and inference. Each entry specifies the downscaling resolution, input/output variable counts, sample dimensions, number of training samples, and total storage size.**

magnitudes may underflow or overflow in BFLOAT16—we apply PyTorch’s dynamic gradient scaling [35]. This technique automatically rescales gradients into a representable range and reverses the scaling during parameter updates, ensuring numerical stability.

To further reduce communication cost, we apply FSDP in a layer-wise fashion [46]. Instead of sharding all model layers in a single instance, parameters are sharded one layer at a time. This reduces synchronization overhead and memory use.

## 5 HOW PERFORMANCE WAS MEASURED

**Model Configuration.** All experiments in Secs. 6 and 7 use four model configurations: 9.5M (256-dim embedding, 6 layers, 4 heads), 126M (1024-dim, 8 layers, 16 heads), 1B (3072-dim, 8 layers, 24 heads), and 10B (8192-dim, 11 layers, 32 heads) parameters.

**System Details.** All experiments were conducted on the Frontier supercomputer from Oak Ridge National Lab. Each node consists of one 64-core AMD EPYC CPU and 8 GPUs (64 GB memory each), organized into 4 MI250X cards with two GPUs per card. GPUs on the same card communicate via Infinity Fabric CPU-GPU, while all four MI250X cards are connected via 50 GB/s GPU-GPU Infinity Fabric. Nodes are interconnected using 100 GB/s Slingshot-11. Software stack includes PyTorch v2.7, ROCm v6.3.1, and libfabric v1.22.

The only exception is Table 4, which compares sustained throughput on Frontier and Alps supercomputer. Alps is built on NVIDIA GH200 Grace Hopper Superchips. Each node features 4 Hopper GPUs (96 GB memory each), with every GPU paired to a 72-core Grace CPU via a 900 GB/s NVLink-C2C interface, enabling unified CPU-GPU memory. Nodes are interconnected using NVIDIA Quantum-2 InfiniBand.

**Datasets.** Table 1 summarizes the datasets used for pretraining and fine-tuning. ORBIT-2 can perform downscaling at arbitrary resolution but as a use case demonstration, our model is trained on paired input → output datasets for 4x spatial refinement.

For global pretraining, we use the ERA5 reanalysis dataset (1980–2020) [15], which integrates historical observations with numerical simulations. Two ERA5 resolution pairs are used: 622 km → 156 km and 112 km → 28 km. The data are split into 38 years for training, 2 years for validation, and 1 year for testing. 23 input variables are used, including 5 static fields, 12 atmospheric variables (humidity, wind speed, and temperature at 200, 500, and 850 hPa), and 6 surface variables. For United States (US)-focused pretraining, we utilize the PRISM and DAYMET observation datasets (from 1980 to

2022) [4, 39], which derived from ground-based weather stations. We perform 4x downscaling from 16 km to 4 km.

For continental U.S. fine-tuning, inputs include both ERA5 and DAYMET at 28 km, with 7 km DAYMET as output ground truth. The fine-tuning dataset is split into training, validation, and testing in the same way as pretraining. The prediction output variables are daily minimal temperature (Tmin), maximal temperature (Tmax), and total precipitation. For global fine-tuning, we target precipitation only and inputs include both ERA5 and IMERG precipitation data at 28 km, downsampled to 7 km resolution. IMERG [19] is a satellite-based global precipitation observations from NASA. Model predictions are daily aggregated and compared to IMERG.

**Performance Metrics.** The total number of floating point operations (FLOPs) of the systems was collected via the Microsoft Deepspeed Profiler [32] and we only gathered the FLOPs on GPUs. Only the mixed-precision BFLOAT16 results were reported. We measured following performance:

- *Time-to-solutions.* Defined as the average wall time to downscale each sample during training. Equivalent to epoch runtime divided by the total sample count. We reported numbers for both 622 → 156 and 112 → 28 km resolutions.
- *Strong scaling efficiency.* Measured speedup per epoch relative to GPU count, with the runtime at 512 GPUs (64 nodes) as the 100% baseline.
- *Sustained throughput.* It is defined as the total FLOPs to downscale each data point, divided by the average wall clock time in seconds. The performance includes the whole application with IO. Reported in ExaFLOPs.
- *Accuracy.* We use both scientific and image-based metrics for downscaling accuracy against observations: Coefficient of determination ( $R^2$ ), Root-Mean-Square-Error (RMSE), RMSE for Quantiles, Structural Similarity Index (SSIM), and Peak-Signal-Noise-Ratio (PSNR). Higher  $R^2$ , SSIM, and PSNR scores represent higher fidelity downscaling, while lower RMSE represents higher fidelity downscaling.

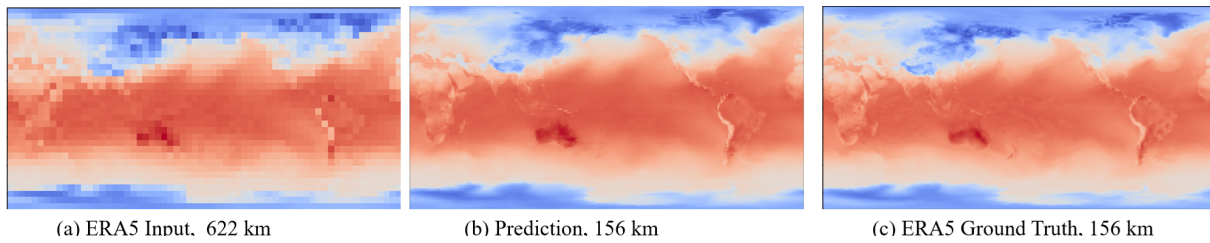
**Dataset, Code and Model Checkpoint Release.** Datasets, source code, and model checkpoints from this work are made publicly available through GitHub repository [45].

## 6 COMPUTING PERFORMANCE RESULTS

**Speedup Ablation Studies.** Table 2(a) presents an ablation study comparing the training performance of vanilla conventional ViT

Arch	Model Size	Resolution (km)	Seq. Length	Compression	Tiles	Time/sample (s)	Speedup	PSNR	SSIM
<b>(a) Reslim Architecture Speedup Comparison with ViT</b>									
ViT	9.5M	622→156	24,576	1×	1	7.3e-4	1	35.0	0.94
Reslim	9.5M	622→156	24,576	1×	1	1.1e-6	660	36.7	0.96
ViT	9.5M	112→28	777,660	1×	1	OOM	NA	NA	NA
Reslim	9.5M	112→28	777,660	1×	1	1.2e-3	NA	37.6	0.96
<b>(b) Adaptive Compression &amp; Tiling Speedup Comparison with Reslim Baseline</b>									
Reslim	9.5M	112→28	777,660	8×	1	3.6e-4	3.3	37.7	0.96
Reslim	9.5M	112→28	777,660	16×	1	1.8e-4	6.6	37.8	0.96
Reslim	9.5M	112→28	777,660	32×	1	1.7e-4	7.1	37.9	0.96
Reslim	9.5M	112→28	777,660	1×	4	8.0e-4	1.5	37.7	0.96
Reslim	9.5M	112→28	777,660	1×	16	6.3e-4	1.9	37.7	0.96
Reslim	9.5M	112→28	777,660	1×	36	7.4e-4	1.6	37.7	0.96

**Table 2: (a) Computing performance comparison between vanilla conventional ViT and Reslim at 128 GPUs. For 622→156 km downscaling, Reslim achieves a 660× speedup over conventional ViT while maintaining accuracy. (b) Illustrates Reslim’s performance gains at varying adaptive compression rates and tile counts, relative to a Reslim baseline without compression or tiling.**



**Figure 6: Illustration of ERA5 surface temperature downscaling: (a) input at 622 km resolution, (b) ORBIT-2’s downscaled prediction at 156 km, and (c) ground-truth at 156 km.**

and the proposed Reslim architectures, using models with 9.5M parameters and at 128 GPUs. Two ERA5→ERA5 downscaling tasks were used: 622→156 km and 112→28 km (details in Table 1). Each hourly output sample is tokenized into image patches. For the 622→156 km task, outputs of shape [128, 256, 3] and 2×2 patch size yield sequence length of 24,576; for 112→28 km, output size [720, 1440, 3] produces 777,660 tokens. No adaptive spatial compression or tiling was applied in this comparison. For visualization, Fig. 6 illustrates an example ERA5 hourly sample for surface temperature: (a) ERA5 input at 622 km, (b) downscaled prediction at 156 km using Reslim architecture, and (c) ERA5 ground truth at 156 km.

The seventh column of Table 2(a) reports the average time to downscale each hourly sample. The eighth column shows the speedup from Reslim relative to the ViT baseline. Notably, the Reslim architecture avoids expensive upsampling operations by operating directly on low-resolution inputs, resulting in significant computational savings. For the smaller 622→156 km task, Reslim achieves a 660× speedup over ViT at the same number of GPUs while maintaining comparable accuracy, as measured by PSNR and SSIM. This demonstrates the effectiveness of Reslim’s residual learning design and Bayesian training loss in maintaining predictive accuracy while reducing computational cost. For the larger 112→28 km resolution task, the ViT model fails due to out-of-memory (OOM) errors. Consequently, a direct speedup comparison is not available, although Reslim completes the task efficiently and maintains high accuracy.

Table 2(b) explores further speedup gains from adaptive spatial compression and sequence tiling, compared to the Reslim baseline (Table 2(a), row 4), all using 128 GPUs. Adaptive compression with a 32× sequence length reduction yields up to a 7.1× speedup with no loss in PSNR or SSIM. Further compression yields diminishing

returns due to quad-tree overhead. Tiling provides up to 1.9× with 16 tiles per sample. Further tiling introduces excessive halo padding overhead and degrades computing performance. Accuracy remains the same across all settings.

**Maximal Sequence Length Scaling.** Table 3 presents sequence length and resolution scaling performance of various model architectures and strategies, demonstrating how the combination of spatial compression, tiling, and the Reslim architecture enables extreme sequence lengths. We achieve sequence lengths of up to 4.2 billion tokens (global downscaling resolution of 0.9 km) for a 9.5M parameter model and up to 671 million tokens (global resolution 2.3 km) for a 10B parameter model. These results surpass the state-of-the-art in sequence scaling by more than 22,000×, compared to state-of-the-art sequence parallelism of 188K tokens [40], and the Swin Transformer at 147K tokens [28].

All experiments utilize 23 input variables (12 atmospheric, 6 surface, and 5 static) and produce 18 output variables (excluding static inputs). Using a conventional ViT with 9.5M parameters, the maximum sequence length is limited to 25K tokens (coarse 156 km global resolution) when using 8 GPUs. Scaling this ViT model to 10B parameters results in an out-of-memory (OOM) error, making global downscaling infeasible.

In contrast, Reslim demonstrates significantly better scaling. With just 8 GPUs, a 9.5M parameter Reslim model scales to 298M tokens at a 3.5 km global resolution. This corresponds to an output tensor of shape [5760, 4520, 18], assuming a 2×2 image patch size. Increasing the number of GPUs to 32, we achieve 466M tokens at 2.7 km resolution.



Architecture	Model Size	Compression	Tiles	GPUs	Max Seq. Length	Output Size	Global Resolution (km)
ViT	9.5M	1×	1	8	25K	[128, 256, 18]	156
ViT	10B	1×	1	8	OOM	—	—
Reslim	9.5M	1×	1	8	298M	[5760, 11520, 18]	3.5
Reslim	9.5M	1×	1	32	466M	[7200, 14400, 18]	2.7
Reslim	9.5M	4×	16	8	1.1B	[11520, 23040, 18]	1.7
Reslim	9.5M	4×	16	128	4.2B	[21600, 43200, 18]	0.9
Reslim	10B	1×	1	8	18M	[1440, 2880, 18]	14
Reslim	10B	4×	16	8	74M	[2880, 5760, 18]	6.9
Reslim	10B	4×	16	512	671M	[8640, 17280, 18]	2.3

Table 3: Maximum sequence length scaling across architectures, model sizes, compression, tiles and GPU count.

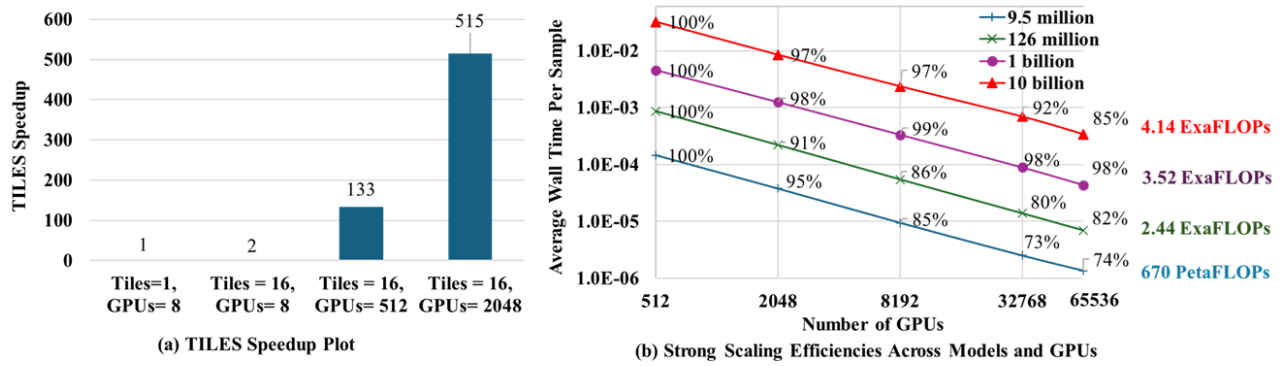


Figure 7: (a) TILES sequence scaling algorithm speedup across GPUs, compared to an 8-GPU baseline that does not utilize tiling. (b) Strong scaling efficiencies up to 8192 nodes (65,536 GPUs) for various model sizes, maintaining a strong scaling efficiencies of 74–98% at 8192 nodes.

Model Size	9.6M	126M	1B	10B
Frontier (MI250x)	23	76	122	166
Alps (GH200)	72	212	251	368

Table 4: Sustained computing throughput (PFLOPs) comparison of Frontier and Alps on 2048 GPUs.

Model Size	9.5M	126M	1B	10B
Inference Time (s)	4.5e-3	2.6e-2	7.2e-2	5.5e-1

Table 5: Model inference speed on a single node (8 GPUs).

When combining Reslim with both spatial tiling (16 tiles per sample) and adaptive spatial compression (4×) techniques, substantial improvements are obtained. With these methods, the model achieves 1.1B tokens on only 8 GPUs, corresponding to down-scaled output of size [11520, 23040, 18]. This result is made possible through several key compression techniques:

- Channel aggregation in Reslim (see Fig. 2) reduces the sequence length by 18× by aggregating channels.
- Spatial tiling divides the sample into 16 tiles, reducing the sequence length per GPU by 16×.
- Adaptive spatial compression reduces sequence by 4×.
- Reslim processes directly on low-resolution input, reducing the effective sequence length by 60×.

Combining all four, the effective per-GPU sequence length becomes only 17,280 tokens, despite the global output representing 1.1 billion tokens. Finally, by scaling to 128 GPUs, we achieve our largest configuration: 4.2 billion tokens at 0.9 km resolution.

For the 10B parameter model, Reslim still scales efficiently. Without compression or tiling, it reaches 18 million tokens. With 4× compression, 16 tiles, and 512 GPUs, the model handles 671 million tokens at 2.3 km resolution.

**TILES Sequence Scaling Speedup.** Fig. 7(a) demonstrates the scalability of the TILES algorithm. With 16 tiles per sample, TILES

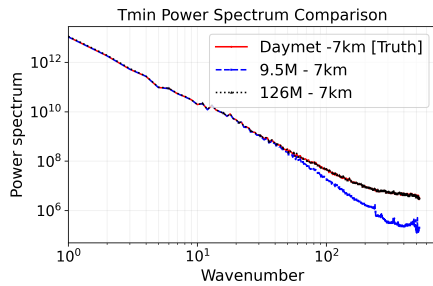
achieves a 1.9× speedup over the non-tiling baseline, both at 8 GPUs, using a 9.5M parameter model on the ERA5→ERA5 112→28 km downscaling task. As GPU count increases, speedup scales nearly linearly, reaching 515× at 2048 GPUs relative to the 8-GPU baseline without tiling. This highlights the scalability and minimal overhead of the TILES approach for distributed training.

**Strong Scaling Efficiencies & Sustained Throughput.** Fig. 7(b) presents strong scaling performance across model sizes (9.5M to 10B parameters) using the same dataset as in Fig. 7(a). Experiments were conducted at scales of 64, 256, 1024, 4096 and 8192 nodes with 8 GPUs for each node at Frontier supercomputer, employing all four forms of orthogonal parallelism (see Sec. 4). Each point in the figure reports the average runtime in second per hourly sample with a data label for corresponding strong scaling efficiency, relative to each model’s baseline performance at 64 nodes (512 GPUs).

All model sizes maintain high strong scaling efficiencies between 74–98%. The smallest model (9.5M) underutilizes hardware at large scales due to insufficient computing with small model, with 1.3e-6 seconds per sample and a sustained computing throughput at 670 PetaFLOPs at 8192 nodes (65,536 GPUs). In contrast, larger models saturate compute resources: the 126M, 1B, and 10B models reach sustained throughputs of 2.4, 3.5, and 4.1 ExaFLOPs, respectively, at 8192 nodes. These results demonstrate the strong scalability of

Fine-Tuning Task	Model Size	$R^2$	RMSE	RMSE $\sigma_1 > 68\%$	RMSE $\sigma_2 > 95\%$	RMSE $\sigma_3 > 99.7\%$	SSIM	PSNR
U.S. Tmin (Kelvin)	9.5M	0.991	3.812	4.652	9.704	15.497	0.958	29.02
	126M	0.999	0.505	0.630	1.025	1.491	0.987	45.96
U.S. Precipitation (mm/day)	9.5M	0.975	0.146	0.166	0.344	0.449	0.931	29.03
	126M	0.979	0.135	0.154	0.296	0.365	0.932	30.20
Global Precipitation (mm/day)	9.5M	0.986	0.098	0.120	0.136	0.133	0.923	35.09
	126M	0.986	0.099	0.122	0.137	0.138	0.913	35.04

**Table 6: Comparison of downscaling accuracy for fine-tuning tasks for U.S. temperature (Tmin), U.S. precipitation, and global precipitation using models with 9.5M and 126M parameters. The table includes  $R^2$ , RMSE, RMSE at different percentiles ( $\sigma_1$ ,  $\sigma_2$ ,  $\sigma_3$ ), SSIM, and PSNR.**



**Figure 8: Power spectra of downscaled minimum temperature using 9.5M and 126M parameter models, showing improved high-frequency fidelity with increased model capacity.**

Reslim and the effectiveness of orthogonal parallelism for exascale climate downscaling.

**Cross Platform Throughput Comparison.** Table 4 compares sustained computing throughput on Frontier (AMD MI250x) and Alps (NVIDIA GH200) supercomputers, using 2,048 GPUs across varying model sizes. The training dataset is the same as Fig. 7, downscaling ERA5 from 112 km to 28 km resolution. On Frontier, throughput scales from 23 PFLOPs for a 9.6M-parameter model to 166 PFLOPs for a 10B-parameter model. On Alps, throughput is higher, reaching 72 PFLOPs for the 9.6M-parameter model and 368 PFLOPs for the 10B-parameter model.

**Inference Speed.** All the results above focus on computing performance for training, but a key advantage of AI foundation models is their efficiency at inference. Once trained, they can be deployed on edge devices with limited resources and deliver near real-time predictions. We evaluate inference performance on a single Frontier node with 8 GPUs (Table 5), using the same ERA5→ERA5 dataset as in Fig. 7 and Table 4. For the 9.5M-parameter model, downscaling each global sample requires only 4 millisecond. For the 10B-parameter model, it takes 0.55 second. In contrast, the non-AI numerical approaches require days or weeks of computation on a large supercomputer [7, 38]. This highlights the unmatched prediction speed of AI, enabling deployment in resource-limited environments with near real-time performance.

## 7 SCIENCE PERFORMANCE RESULTS

**U.S. Regional Fine-Tuning.** Following pretraining on the datasets in Table 1, we fine-tune ORBIT-2 models with 9.5M and 126M parameters on two tasks for science demonstration: (1) US-specific downscaling of ERA5 and DAYMET data (1980 to 2022) from 28 km to 7 km, evaluated against 7 km DAYMET observations for both daily total precipitation and Tmin; and (2) global precipitation downscaling of ERA5 and IMERG (1998 to 2022) from 28 km to 7 km,

evaluated against 7 km IMERG observations. These tasks evaluate model capability in downscaling at both regional and global scale.

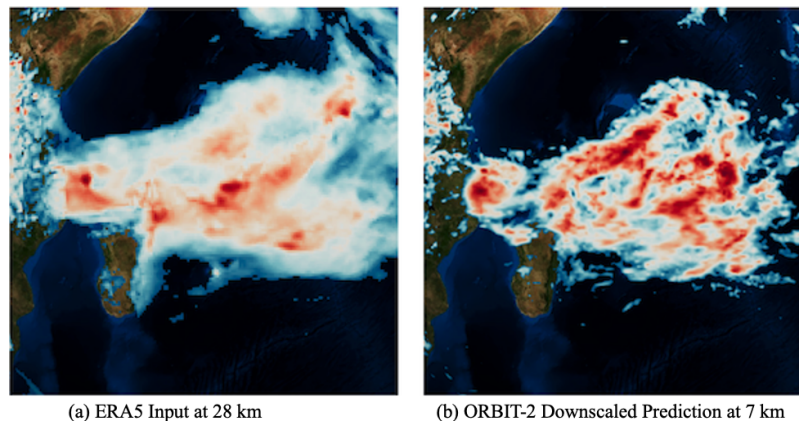
Table 6’s second row summarizes U.S. regional downscaling results for Tmin at the unseen testing years. Both 9.5M and 126M models accurately reconstruct high-resolution temperature at 7 km resolution and are capable of capturing extremes, with the larger 126M model consistently outperforming the smaller 9.5M model across all metrics with lower RMSE, and higher  $R^2$ , SSIM and PSNR. Notably, the 126M model achieves an  $R^2$  of 0.999 and SSIM of 0.987, establishing a new benchmark for temperature downscaling at 7 km resolution.

Fig. 8 provides corresponding spectral analysis for Tmin downscaling results in Table 6’s row two, by comparing the spatial power spectra of the two model sizes. The 126M model accurately captures high-frequency content, closely matching the high frequency spectral characteristics of the DAYMET observation ground truth. In contrast, the 9.5M model deviates from the ground truth at high frequencies. This demonstrates the larger model’s ability to resolve fine-scale spatial variability, emphasizing the value of increased model capacity for high-fidelity climate downscaling.

Table 6’s third row presents downscaling results for daily total precipitation—one of the most challenging variables due to its high spatial variability and localized extremes. ORBIT-2 demonstrates strong performance, closely matching observations when downscaling to 7 km resolution. The 126M model consistently outperforms the 9.5M model across all evaluation metrics, achieving an  $R^2$  of 0.979 and an overall RMSE of 0.135 mm/day. Notably, the large model also accurately captures extreme precipitation events, with RMSE values of 0.365 mm/day at the 99.7th percentile and 0.525 mm/day at the 99.99th percentile. All RMSE values for precipitation are computed in log-transformed space using  $\log(x + 1)$ , where  $x$  denotes daily precipitation in millimeters.

**Global Precipitation Fine-Tuning.** ORBIT-2’s fine-tuning performance for global precipitation downscaling task is summarized in Table 6 row 4. Both model sizes (9.5M and 126M parameters) achieve similar accuracy, with  $R^2 = 0.98$ , SSIM = 0.92, PSNR = 35.0, and RMSE = 0.09 mm/day (in  $\log(x + 1)$  space). These results underscore ORBIT-2’s robustness across both regional and global downscaling tasks. For visualization, Fig. 9 shows a one-day snapshot of an example region, comparing ERA5 precipitation at 28 km with ORBIT-2 downscaling at 7 km. A full-year animation is available online [Click Here](#), illustrating the spatial resolution gains and improve climate pattern for year 2020.

To evaluate long-term fidelity and climate pattern, Fig. 10 reports performance of high-resolution downscaling in capturing key precipitation characteristics for 1998–2021 across 58 climatically



**Figure 9: Example region for global precipitation downscaling during fine-tuning. (a) ERA5 input at 28 km resolution; (b) ORBIT-2 downscaled output at 7 km. Click on the link here for an interactive visualization of global precipitation downscaling for the year 2020 [Click Here](#).**

West Africa	76.32	86.75	92.82	88.96	91.84
South Asia	88.47	93.37	92.46	92.11	90.39
South America	97.47	94.71	90.18	95.80	92.35
Southeast Africa	93.62	95.21	93.35	89.01	92.19
Australia	93.86	90.36	95.72	86.99	91.11
East Africa	67.65	61.09	91.44	82.47	92.39
East Asia	59.60	64.94	80.13	82.98	83.98
Southeast Asia	91.99	87.24	76.37	84.37	91.44
North America	81.79	50.63	87.99	88.72	92.48
	Onset	Withdrawal	Seasonality	Entropy	Hovmoller

**(a) ERA5 at 28 km Resolution**

West Africa	99.11	99.52	99.11	99.34	99.70
South Asia	99.15	99.93	98.66	99.32	99.74
South America	99.75	99.71	98.60	99.45	99.65
Southeast Africa	99.47	99.51	95.71	99.43	99.80
Australia	99.64	99.73	96.67	99.28	99.82
East Africa	98.29	98.39	99.74	99.23	99.62
East Asia	99.59	99.74	99.31	99.06	99.75
Southeast Asia	99.38	99.68	97.91	98.49	99.77
North America	99.04	97.30	96.97	99.02	99.40
	Onset	Withdrawal	Seasonality	Entropy	Hovmoller

**(b) ORBIT-2 Downscaled Output at 7 km Resolution**

**Figure 10: (a) ERA5 skill scores for global precipitation at 28 km resolution, evaluated using averaged  $R^2$  and SSIM, over 24 years (1998–2021). Analysis metrics include monsoon onset and withdrawal, seasonality, entropy, and Hovmöller diagnostics. (b) ORBIT-2 skill scores for global precipitation at 7 km resolution, showing improved fidelity across the same metrics.**

homogeneous land and ocean regions defined by the Intergovernmental Panel on Climate Change (IPCC). For each region, we report a composite skill score, defined as the averaged  $R^2$  correlation and SSIM (scaled between 0 to 100), with higher values indicating stronger agreement with the IMERG 7 km observational reference.

Performance is further analyzed in nine monsoon regions, where precipitation is governed by strong seasonality and land-ocean atmosphere interactions, and significant hydrological and economic impacts. Following [3], we evaluate monsoon onset and withdrawal timing, precipitation seasonality and entropy, and the northward progression of rainfall (via Hovmöller diagnostics). Results show substantial improvements with ORBIT-2. As shown in Fig.10(a), ERA5 at 28 km resolution exhibits moderate skill scores across metrics and regions, whereas ORBIT-2 7 km downscaling (Fig.10(b)) consistently achieves much higher skill scores across all metrics and monsoon regions. These results highlight the effectiveness of ORBIT-2 in enhancing the spatiotemporal fidelity of precipitation, especially in regions governed by complex climatic processes.

## 8 IMPLICATION

ORBIT-2 represents a significant leap forward in the convergence of artificial intelligence, HPC, and Earth system science. By overcoming fundamental challenges in scalability, resolution, and uncertainty, ORBIT-2 sets a new standard for climate and earth system foundation models.

**Impact on HPC.** ORBIT-2 pushes the frontiers of HPC by enabling ViTs at unprecedented scale. Through its novel TILES algorithm, ORBIT-2 reduces the self-attention complexity from quadratic to linear, allowing for efficient processing of ultra-long sequences. Complementing this, the Reslim architecture introduces a lightweight, uncertainty-aware learning framework that leverages residual learning and Bayesian regularization to improve efficiency and training robustness.

Crucially, ORBIT-2 achieves breakthrough scalability in ViT training, scaling up to 10 billion parameters model size across 65,536 GPUs, and scale up to 4.2 billion token sequence length, several magnitudes longer than the state-of-the-art long sequence implementation that scales to 188K tokens. It sustains up to 4.1 ExaFLOPs of performance with 74–98% scaling efficiency at 65,536 GPUs, setting a new benchmark for exascale AI. ORBIT-2 serves as a

blueprint for next-generation exascale foundation models, enabling transformative applications across domains including genomics, fluid dynamics, astrophysics, and Earth system modeling.

**Impact on Climate Science.** ORBIT-2 enables hyper-resolution, global-scale downscaling with state-of-the-art accuracy and efficiency. Against observations, it achieves an  $R^2$  of 0.999 for temperature and 0.979/0.986 for precipitation at 7 km resolution over the continental United States and globally, respectively. Unlike traditional approaches, ORBIT-2 generalizes across variables and regions using a single foundation model, improving climate projection fidelity, resolving fine-scale processes, and detecting localized extremes. This unified capability makes ORBIT-2 a valuable tool for supporting climate adaptation and mitigation strategies. Besides the above capability, this paper also shows ORBIT-2 foundation AI model's capability to perform global downscaling in 4 millisecond for 9.5M parameter model and 0.5 second for a 10B parameter model with only 8 GPUs through model inferencing, showing the AI model's capability for near-real-time climate and weather prediction on edge devices. This is transformative compared to the conventional numerical simulation methods, which often takes hours and days to compute on a large supercomputer.

**Limitations and Practical Considerations.** Like all AI-based methods, ORBIT-2 may inherit biases from training data source and lacks strict enforcement of physical conservation laws, requiring caution in applications sensitive to energy or mass balance. Practical deployment will require careful consideration on application-based bias correction, physics-informed integration, and domain-specific fine-tuning to ensure reliability across diverse climates and regimes. Finally, while near-real-time inference on limited hardware is promising, robust evaluation in operational settings is needed before widespread deployment in climate services and decision-making.

## ACKNOWLEDGMENTS

We thank Jonathan Coles for helping us get scalability numbers on the Alps cluster. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). This research was primary supported by the ORNL's AI Initiative sponsored by the Director's Research and Development Program at ORNL, additionally supported by DOE Early Career Project sponsored by the BER program. It was also supported as part of the Energy Exascale Earth System Model (E3SM) project, funded by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research, and Earth Systems Model Development Program. An award of computer time was provided by the INCITE program through the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility.

## REFERENCES

- [1] Michael Aich, Philipp Hess, Baoxiang Pan, Sebastian Bathiany, Yu Huang, and Niklas Boers. 2024. Conditional diffusion models for downscaling & bias correction of Earth system model precipitation. *arXiv preprint arXiv:2404.14416* (2024).
- [2] Nicolaas J Annau, Alex J Cannon, and Adam H Monahan. 2023. Algorithmic hallucinations of near-surface winds: Statistical downscaling with generative adversarial networks to convection-permitting scales. *Artificial Intelligence for the Earth Systems* 2, 4 (2023), e230015.
- [3] Moetasim Ashfaq, Tereza Cavazos, Michelle Simões Reboita, José Abraham Torres-Alavez, Eun-Soon Im, Christiana Fumilola Olusegun, Lincoln Alves, Kesondra Key, Mojisola O Adeniyi, Moustapha Tall, et al. 2021. Robust late twenty-first century shift in the regional monsoons in RegCM-CORDEX simulations. *Climate Dynamics* 57, 5 (2021), 1463–1488.
- [4] Christopher Daly and Kirk Bryant. 2013. The PRISM climate and weather system—an introduction. *Corvallis, OR: PRISM climate group 2* (2013).
- [5] Tri Dao. 2023. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. *arXiv:2307.08691* [cs.LG]
- [6] Veronika Eyring, Pierre Gentine, Gustau Camps-Valls, David M Lawrence, and Markus Reichstein. 2024. AI-empowered next-generation multiscale climate modelling for mitigation and adaptation. *Nature Geoscience* 17, 10 (2024), 963–971.
- [7] O. Fuhrer, T. Chadha, T. Hoefler, G. Kwasniewski, X. Lapillonne, D. Leutwyler, D. Lüthi, C. Osuna, C. Schär, T. C. Schulthess, and H. Vogt. 2018. Near-global climate simulation at 1 km resolution: establishing a performance baseline on 4888 GPUs with COSMO 5.0. *Geoscientific Model Development* 11, 4 (2018), 1665–1681.
- [8] Filippo Giorgi. 2019. Thirty Years of Regional Climate Modeling: Where Are We and Where Are We Going next? *Journal of Geophysical Research: Atmospheres* 124, 11 (2019), 5696–5723. <https://doi.org/10.1029/2018JD030094> *arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018JD030094*
- [9] Filippo Giorgi and William J Gutowski Jr. 2015. Regional dynamical downscaling and the CORDEX initiative. *Annual review of environment and resources* 40, 1 (2015), 467–490.
- [10] Evan H Girvetz, Edwin P Maurer, Philip B Duffy, Aaron Ruesch, Bridget Thrasher, and Chris Zganjar. 2013. Making climate data relevant to decision making: the important details of spatial and temporal downscaling. *World Bank* (2013).
- [11] Iratxe González-Aparicio, Fabio Monforti, Patrick Volker, Andreas Zucker, Francesco Careri, Thomas Huld, and Jake Badger. 2017. Simulating European wind power generation applying statistical downscaling to reanalysis data. *Applied Energy* 199 (2017), 155–168.
- [12] Ethan Gutmann, Tom Pruitt, Martyn P Clark, Levi Brekke, Jeffrey R Arnold, David A Raff, and Roy M Rasmussen. 2014. An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resources Research* 50, 9 (2014), 7167–7186.
- [13] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. 2023. FLatten Transformer: Vision Transformer using Focused Linear Attention. *arXiv:2308.00442* [cs.CV] <https://arxiv.org/abs/2308.00442>
- [14] Chaoyang He, Shen Li, Mahdi Soltanolkotabi, and Salman Avestimehr. 2021. PipeTransformer: Automated Elastic Pipelining for Distributed Training of Transformers. *arXiv:2102.03161* [cs.LG]
- [15] Hans Hersbach et al. 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* 146, 730 (2020), 1999–2049. <https://doi.org/10.1002/qj.3803> *arXiv:https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803*
- [16] Bruce C Hewitson and Robert George Crane. 1996. Climate downscaling: techniques and application. *Climate Research* 7, 2 (1996), 85–95.
- [17] Chengquan Huang, John R.G. Townshend, Shunlin Liang, Satya N.V. Kalluri, and Ruth S. DeFries. 2002. Impact of sensor's point spread function on land cover characterization: assessment and deconvolution. *Remote Sensing of Environment* 80, 2 (2002), 203–212. [https://doi.org/10.1016/S0034-4257\(01\)00298-X](https://doi.org/10.1016/S0034-4257(01)00298-X)
- [18] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Mia Xu Chen, Dehao Chen, Hyoun Joong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. 2019. GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism. *arXiv:1811.06965* [cs.CV]
- [19] George J Huffman et al. 2020. Integrated multi-satellite retrievals for the global precipitation measurement (GPM) mission (IMERG). *Satellite precipitation measurement: Volume 1* (2020), 343–353.
- [20] Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. 2023. DeepSpeed Ulysses: System Optimizations for Enabling Training of Extreme Long Sequence Transformer Models. *arXiv:2309.14509* [cs.LG] <https://arxiv.org/abs/2309.14509>
- [21] Peter G Jones and Philip K Thornton. 2013. Generating downscaled weather data from a suite of climate models for agricultural modelling applications. *Agricultural Systems* 114 (2013), 1–5.
- [22] Chiheon Kim, Heungsung Lee, Myungryong Jeong, Woonhyuk Baek, Boogeon Yoon, Ildoo Kim, Sungbin Lim, and Sungwoong Kim. 2020. torchgpipe: On-the-fly Pipeline Parallelism for Training Giant Models. *arXiv:2004.09910* [cs.DC]
- [23] Nikolay Koldunov, Thomas Rackow, Christian Lessig, Sergey Danilov, Suvarchal K Cheedela, Dmitry Sidorenko, Irina Sandu, and Thomas Jung. 2024. Emerging AI-based weather prediction models as downscaling tools. *preprint arXiv:2406.17977* (2024).
- [24] Bipin Kumar, Rajib Chattopadhyay, Manmeet Singh, Niraj Chaudhari, Karthik Kodari, and Amit Barve. 2021. Deep learning-based downscaling of summer monsoon rainfall data over Indian region. *Theoretical and Applied Climatology* 143 (2021), 1145–1156.
- [25] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* 479 (2022), 47–59.



- [26] Shen Li et al. 2020. PyTorch Distributed: Experiences on Accelerating Data Parallel Training. *Proceedings of the VLDB Endowment* 13 (2020), 12 pages. <https://doi.org/10.14778/3415478.3415530>
- [27] Yuhao Liu, James Doss-Gollin, Guha Balakrishnan, and Ashok Veeraraghavan. 2024. Generative Precipitation Downscaling using Score-based Diffusion with Wasserstein Regularization. *arXiv preprint arXiv:2410.00381* (2024).
- [28] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2022. Swin Transformer V2: Scaling Up Capacity and Resolution. *arXiv:2111.09883 [cs.CV]* <https://arxiv.org/abs/2111.09883>
- [29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv:2103.14030 [cs.CV]* <https://arxiv.org/abs/2103.14030>
- [30] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tiejiong Zeng. 2022. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 457–466.
- [31] Isaac Lyngaas, Murali Gopalakrishnan Meena, Evan Calabrese, Mohamed Wahib, Peng Chen, Jun Igarashi, Yuankai Huo, and Xiao Wang. 2024. Efficient Distributed Sequence Parallelism for Transformer-Based Image Segmentation. *Electronic Imaging* 36, 12 (2024), 199–1–199–1. <https://doi.org/10.2352/EL.2024.36.12.HPCI-199>
- [32] Microsoft. 2024. Microsoft DeepSpeed. <https://github.com/microsoft/DeepSpeed>. Accessed: 2024-04-05.
- [33] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. 2023. ClimateX: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343* (2023).
- [34] Tung Nguyen, Jason Jewik, Hritik Bansal, Prakhar Sharma, and Aditya Grover. 2023. ClimateLearn: Benchmarking Machine Learning for Weather and Climate Modeling. *arXiv:2307.01909 [cs.LG]*
- [35] Pytorch. 2024. Automatic Mixed Precision Package - TORCH.AMP. <https://pytorch.org/docs/stable/amp.html#torch.cuda.amp.GradScaler>. Accessed: 2024-04-05.
- [36] Johannes Schumde et al. 2024. Prithvi WxC: Foundation Model for Weather and Climate. *arXiv:2409.13598 [cs.LG]* <https://arxiv.org/abs/2409.13598>
- [37] Mohammad Shoeibi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *arXiv:1909.08053 [cs.CL]*
- [38] Mark A. Taylor, Peter M. Caldwell, Luca Bertagna, Conrad Clevenger, Aaron S. Donahue, James G. Foucar, Oksana Guba, Benjamin R. Hillman, Noel Keen, Jayesh Krishna, Matthew R. Norman, Sarat Sreepathi, Christopher R. Terai, James B. White, Danying Wu, Andrew G. Salinger, Renata B. McCoy, L. Ruby Leung, and David C. Bader. 2023. The Simple Cloud-Resolving E3SM Atmosphere Model Running on the Frontier Exascale System. In *SC23: International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–11.
- [39] Peter E. Thornton, Michele M. Thornton, Benjamin W. Mayer, Nate Wilhelmi, Yaxing Wei, Ranjeet Devarakonda, and Robert B. Cook. 2014. *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2*. Technical Report. Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States).
- [40] Aristidis Tsaris, Chengming Zhang, Xiao Wang, Junqi Yin, Siyan Liu, Moetasim Ashfaq, Ming Fan, Jong Youl Choi, Mohamed Wahib, Dan Lu, Prasanna Balaprakash, and Feiyi Wang. 2024. Sequence Length Scaling in Vision Transformers for Scientific Images on Frontier. *arXiv:2405.15780 [cs.CV]* <https://arxiv.org/abs/2405.15780>
- [41] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. 2022. MaxViT: Multi-Axis Vision Transformer. *arXiv:2204.01697 [cs.CV]* <https://arxiv.org/abs/2204.01697>
- [42] Fang Wang, Di Tian, Lisa Lowe, Latif Kalin, and John Lehrter. 2021. Deep learning for daily precipitation and temperature downscaling. *Water Resources Research* 57, 4 (2021), e2020WR029308.
- [43] Qunming Wang, Yijie Tang, and Peter M. Atkinson. 2020. The effect of the point spread function on downscaling continua. *ISPRS Journal of Photogrammetry and Remote Sensing* 168 (2020), 251–267. <https://doi.org/10.1016/j.isprsjprs.2020.08.016>
- [44] Qunming Wang, Chengyuan Zhang, Xiaohua Tong, and Peter M. Atkinson. 2020. General solution to reduce the point spread function effect in subpixel mapping. *Remote Sensing of Environment* 251 (2020), 112054. <https://doi.org/10.1016/j.rse.2020.112054>
- [45] Xiao Wang, Jong-Youl Choi, Isaac Lyngaas, Hong-Jun Yoon, Ashwin Aji, Takuya Kurihana, Ming Fan, Prasanna Balaprakash, and Dan Lu. [n. d.]. *ORBIT-2 Weather and Climate Downscaling Software Repository*. <https://github.com/XiaoWang-Github/ORBIT-2>
- [46] Xiao Wang, Siyan Liu, Aristidis Tsaris, Jong-Youl Choi, Ashwin Aji, Ming Fan, Wei Zhang, Junqi Yin, Moetasim Ashfaq, Dan Lu, and Prasanna Balaprakash. 2024. ORBIT: Oak Ridge Base Foundation Model for Earth System Predictability. *arXiv:2404.14712 [physics.ao-ph]* <https://arxiv.org/abs/2404.14712>
- [47] Xiao Wang, Isaac Lyngaas, Aristidis Tsaris, Peng Chen, Sajal Dash, Mayanka Chandra Shekar, Tao Luo, Hong-Jun Yoon, Mohamed Wahib, and John Gouley. 2023. Ultra-Long Sequence Distributed Transformer. *arXiv:2311.02382 [cs.DC]* <https://arxiv.org/abs/2311.02382>
- [48] Robert L. Wilby and Christian W. Dawson. 2013. The Statistical Downscaling Model: insights from one decade of application. *International Journal of Climatology* 33, 7 (2013).
- [49] Robert L. Wilby, TML Wigley, D. Conway, P. D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks. 1998. Statistical downscaling of general circulation model output: A comparison of methods. *Water resources research* 34, 11 (1998), 2995–3008.
- [50] Enzhi Zhang, Isaac Lyngaas, Peng Chen, Xiao Wang, Jun Igarashi, Yuankai Huo, Mohamed Wahib, and Masaharu Munetomo. 2024. Adaptive Patching for High-resolution Image Segmentation with Transformers. *arXiv:2404.09707 [cs.CV]* <https://arxiv.org/abs/2404.09707>
- [51] Yanli Zhao et al. 2023. PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel. *Proceeding of VLDB Endowment* 16, 12 (August 2023), 3848–3860. <https://doi.org/10.14778/3611540.3611569>