https://doi.org/10.2352/EI.2026.38.11.HPCI-200

Towards Scaling Law Analysis For Spatiotemporal Weather Data

Alexander Kiefer[1,2], Prasanna Balaprakash[2], Xiao Wang[1,2]

[1]Bredesen Center, University of Tennessee, Knoxville, Tennessee
[2]Oak Ridge National Laboratory, Oak Ridge, Tennessee

Abstract

Compute-optimal scaling laws are relatively well studied for NLP and CV, where objectives are typically single-step and targets are comparatively homogeneous. Weather forecasting is harder to characterize in the same framework: autoregressive rollouts compound errors over long horizons, outputs couple many physical channels with disparate scales and predictability, and globally pooled test metrics can disagree sharply with per-channel, late-lead behavior implied by short-horizon training. We extend neural scaling analysis for autoregressive weather forecasting from single-step training loss to long rollouts and per-channel metrics. We quantify (1) how prediction error is distributed across channels and how its growth rate evolves with forecast horizon, (2) if power law scaling holds for test error, relative to rollout length when error is pooled globally, and (3) how that fit varies jointly with horizon and channel for parameter, data, and compute-based scaling axes. We find strong cross-channel and cross-horizon heterogeneity: pooled scaling can look favorable while many channels degrade at late leads. We discuss implications for weighted objectives, horizon-aware curricula, and resource allocation across outputs.

Introduction

Large neural models for scientific spatio-temporal prediction—weather and climate emulation in particular—are now trained at substantial parameter, data, and compute budgets [1, 2, 3, 4, 5, 6, 7]. A practical question facing such projects is how to allocate those budgets: train a wider model, on more data, or spend more floating point operations (FLOPs) training, subject to a fixed resource envelope. For language and vision, empirical neural scaling laws—power laws linking loss to model size, data, and compute—have become the standard tool for answering that question at scale [8, 9, 10]

The spatio-temporal data common in scientific disciplines, such as weather, fluid dynamics, and plasma physics, differs from those benchmarks in ways that complicate a direct comparison. Quality is judged after long autoregressive rollouts, where predictions are fed back as inputs and errors compound, not from a single forward pass at the training horizon. Outputs are also high-dimensional fields with many physical channels, so scalar aggregates can hide which variables track a scaling frontier and which do not. When scaling fits are interpreted under the wrong regime, resource plans and extrapolated scaling exponents may be misleading for the error people actually care about.

Weather forecasting is a useful testbed for studying these issues: standardized reanalysis data and metrics are available [11, 12], the dynamics are challenging but well studied [13], and transformer based models are increasingly deployed in production-oriented models [6, 5, 2, 7, 14, 15]. Hardware trends, likewise, motivate understanding transformer scaling in scientific workloads, not only in natural language and computer vision.

Prior work has established neural scaling analyses for global weather emulation under a deliberately simple training objective. Subramanian et al. [15] trained a swin transformer on ERA5, compared continual training techniques, and fitted power-law relations and compute-optimal (N, D)-style allocations in the same spirit as large-language-model scaling studies. Their reported scaling summaries are based on single-step deterministic training (one forward map under mean-squared error (MSE) so that capacity–data–compute trade-offs can be measured before autoregressive effects dominate. This choice is sound for isolating pretraining scaling, but it does not answer how the same IsoFLOP family behaves when error is measured after many rollout steps on held-out years, and whether channel-wise behavior aligns with globally pooled scaling curves.

We address that gap by keeping the architecture, data splits, and IsoFLOP protocol of the preceding study and reorienting the analysis toward autoregressive test-time error. We evaluate rolled-out forecasts up to 240 h on held-out periods, characterize how error is distributed across output channels and how its growth rate varies with lead time, and refit log–log scaling relationships between test error and three scaling covariates—parameter count, training data volume, and compute—at each horizon. Relative to single-step scaling, our proposed regime directly targets the error relevant to long-horizon forecasting and exposes heterogeneity that global summaries alone would miss.

Our paper contributes the following:

- Autoregressive scaling. We report how the coeffiecient of determination (R^2) of log–log scaling laws for globally pooled test root-mean-squared error (RMSE), and the fitted scaling exponent relating error to compute, evolve with forecast horizon when models are evaluated autoregressively rather than under one-step.
- Channel- and horizon-resolved analysis. We quantify per-channel error at a fixed short lead, visualize

rollout error dynamics through the time derivative of RMSE, and provide heatmaps of scaling-law R^2 over forecast horizon and individual output channels for parameter-, data-, and compute-based frontiers.
- Evaluation protocol detail. We connect the empirical observations—substantial cross-channel and cross-horizon heterogeneity, with strong global scaling coexisting with localized breakdowns—to practical implications such as weighted training objectives and horizon-aware objective functions.

Related Work

Neural Scaling Laws

Neural scaling laws are empirical observations of models, often well approximated by power laws, that relate model size (N), training data (D), training compute (C), and predictive loss (L)[8, 9, 10]. A central use of these fits is compute-optimal allocation: for a fixed training budget C (cumulative FLOPs), one seeks to minimize the achievable loss $L(N,D)$ subject to a cost model $\mathrm{FLOPs}(N,D) = C$,

$$\big(N^*(C), D^*(C)\big) \in \underset{N,D:\,\mathrm{FLOPs}(N,D)=C}{\arg\min} L(N,D). \tag{1}$$

Writing $N^*(C)$ and $D^*(C)$ for such minimizers, one typically finds

$$N^*(C) \propto C^{\alpha}, \tag{2a}$$

$$D^*(C) \propto C^{\beta}, \tag{2b}$$

for exponents α, β that depend on the fitted law and training setup; in the common approximation $C \propto ND$, one expects $\alpha + \beta \approx 1$ up to the details of the parameterization.

An earlier line of work on learning curves in economics and related fields studied how performance scales with data, often holding model capacity fixed [16]. Hestness et al. [8] extended this perspective with a broad study of deep networks, observing power-law relations between error or loss and both model and data scale across machine translation, language modeling, image classification, and speech recognition. A key takeaway was that the scaling exponents depend strongly on the task, whereas many architectural choices mainly rescale prefactors rather than the exponent.

For autoregressive transformer language models, Kaplan et al. [9] fitted joint scaling of loss with parameter count N, token count D, and compute, and derived the compute-optimal bounds for their experimental regime. Their analysis implied that, at a fixed FLOP budget, performance was improved by training larger models on smaller token budgets than was the norm at the time. Hoffmann et al. [10] also explored compute-optimal training at larger scale and with more refined accounting of model size. They found a substantially more data-heavy optimum. In their setting, compute-optimal training uses many more tokens per parameter than the Kaplan study. In short, training tokens and non-embedding parameters should scale together, with on the order of twenty tokens per parameter for the one-epoch, cosine-schedule setups they emphasize. This finding, called Chinchilla scaling after their reference model, does not negate power-law scaling itself, but refines the empirically inferred balance between N and D at fixed C. Related scaling analyses have been reported for vision transformers across classification, captioning, VQA, and transfer settings [17].

Weather Forecasting

Weather forecasting is a critical problem with real-world impact, from predicting extreme weather phenomena to quantifying actuarial risk. The objective is to predict the future state of multiple atmospheric variables that jointly describe the atmosphere.

The traditional and most trusted approach is numerical weather prediction (NWP). In NWP, the three-dimensional atmospheric state is modeled on a global or regional grid, augmented by boundary conditions and external forcings, before integrating the fluid and thermodynamic equations forward in time. The initial state is generated by data assimilation, which combines observations with a short-range prior forecast so that the model starts from physically accurate initial conditions. These systems couple large-scale dynamics, subgrid physics, and assimilation, and their computational cost grows quickly with spatial resolution, forecast length, and ensemble size [13, 18, 19].

This expensive cycle of numerical simulation motivates neural surrogates. Instead of directly integrating the primitive equations, one learns an approximation to the mapping from recent observations to future states as a multivariate spatio-temporal regression problem.

Let $\mathbf{x}_t$ denote the observed state vector of atmospheric variables at time t. The forecasting task aims to learn a function f_θ such that

$$\hat{\mathbf{x}}_{t+\tau} = f_\theta(\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-k+1}), \tag{3}$$

where k is the input history length and τ is the lead time of the forecast. The model parameters θ are learned by minimizing a loss L over a dataset $\mathscr{D}$ of historical observations:

$$\min_\theta \frac{1}{|\mathscr{D}|} \sum_{(\mathbf{x}_{t-k+1:t}, \mathbf{x}_{t+\tau}) \in \mathscr{D}} L\big(f_\theta(\mathbf{x}_{t-k+1:t}), \mathbf{x}_{t+\tau}\big). \tag{4}$$

Neural surrogate models have proven to be highly effective at trading the extremely high cost of NWP models for relatively modest penalties in accuracy [20, 21, 14, 2, 6, 22]. Additionally, these deterministic forecasting models lend themselves well to extensions as probabilistic forecasting models, where an ensemble of deterministic models with slightly varying initial conditions samples from the distribution and minimizes a scoring function, such as the continuous ranked probability score (CRPS) [23], as in FourCastNet [24].

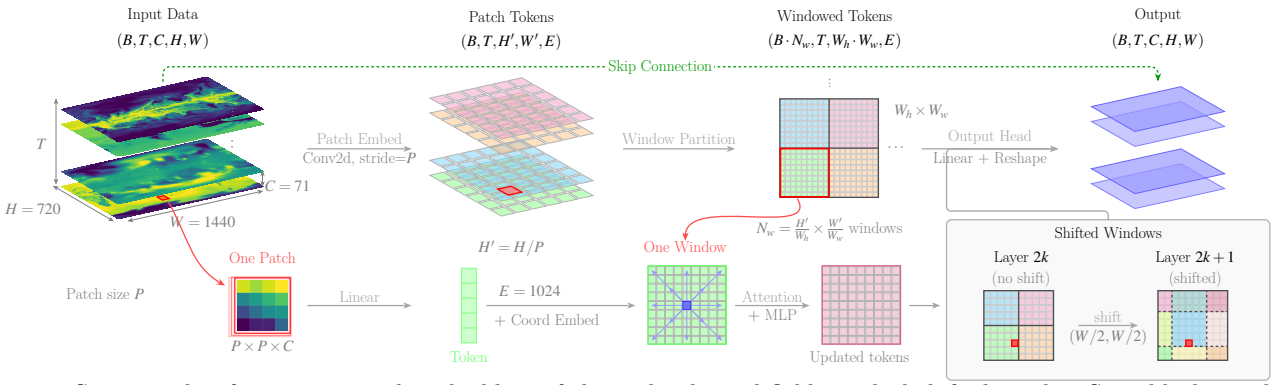


Figure 1: Swin weather forecaster: patch embedding of the multi-channel field, stacked shifted-window Swin blocks, and patch unembedding to the predicted state.

Scaling Laws for Weather

Subramanian et al. [15] presented one of the first rigorous study of neural scaling laws for global weather emulation. They trained a family of minimal, scalable Swin Transformers on ERA5 and compared continual training with a constant learning rate and periodic cooldowns against repeated from-scratch runs under cosine decay. The continual recipe matches or improves validation loss while substantially reducing the cost of sweeping many compute budgets. Across model sizes, dataset sizes, and FLOP budgets they constructed IsoFLOP frontiers and identified compute-optimal (N,D)-style allocations in the same form as scaling analyses for large language models, but for a deterministic weather forecasting task.

That work is focused on single-step deterministic training, where the network maps the state at time t to the state at $t+\Delta t$ under a mean-squared-error objective, so that capacity/data/compute trade-offs can be measured before compounding effects of long autoregressive rollouts dominate. We reuse the architecture, ERA5 splits, and IsoFLOP training design, but our analyses, motivated in the introduction, target autoregressive test-time error, per-channel structure, and horizon-dependent scaling diagnostics.

Methodology

This section isolates what is new in our paper. We reuse the architecture, ERA5 preprocessing, and IsoFLOP training family from [15], but change the target of analysis from single-step training loss to autoregressive rollout error and its scaling behavior over lead time and output channel.

Rollout Evaluation Framework

At inference, the network is applied recursively: the prediction at $t+\Delta t$ is fed back into the input window for the next six-hour step. We evaluate on the held-out years in Table 1. Initial conditions (ICs) are sampled every 12 h in the test window, and each IC is rolled out to fixed lead times up to 240 h. Metrics are logged per lead and per channel, then reduced across data-parallel workers.

Rollout Scaling-law Analysis

We reuse the same IsoFLOP frontiers and compute-optimal allocations from training, but replace single-step training loss with inference error at each rollout horizon. For each lead time h (and, where applicable, channel c), we fit

$$\log \varepsilon_{h,c} = a_{h,c} + b_{h,c} \log s, \tag{5}$$

where s is the scaling covariate (optimal parameter count, optimal data size, or compute along the frontier) and ε is RMSE. The slope $b_{h,c}$ is the empirical scaling exponent at that horizon/channel, and $R^2_{h,c}$ quantifies how well a linear relation in log–log space explains the observed frontier. This is a two-stage procedure. For each FLOP budget we first estimate an IsoFLOP optimum from an order-2 polynomial fit in log–log space, then we fit the linear power law above to those extracted optima across budgets. For each horizon/channel fit, we report

$$R^2_{h,c} = 1 - \frac{\sum_i \left(\log \varepsilon_{h,c,i} - \widehat{\log \varepsilon}_{h,c,i}\right)^2}{\sum_i \left(\log \varepsilon_{h,c,i} - \overline{\log \varepsilon}_{h,c}\right)^2}, \tag{6}$$

where i indexes FLOP budgets (or extracted optima), $\widehat{\log \varepsilon}_{h,c,i}$ is the fitted log-error, and $\overline{\log \varepsilon}_{h,c}$ is the sample mean log-error for that (h,c) slice.

Metrics

Unless otherwise noted, we report area-weighted RMSE per channel and lead time, where latitude weights account for grid-cell area on the latitude–longitude mesh. We also report globally pooled RMSE (the square root of channel-uniform global MSE used in training) as a single scalar baseline.

Experimental Setup

Model Architecture

We use a swin transformer [25] adapted to latitude–longitude fields as in [15]: patch embedding maps the multi-channel atmospheric state to a latent grid; stacked

swin blocks apply shifted-window multi-head self-attention and an MLP with pre-RMSNorm, residual connections, and QK-normalization for stability. Alternate layers shift the window partition so information mixes across windows. The mask respects east–west periodicity. A final head maps latents back to physical channels.

Distributed Training

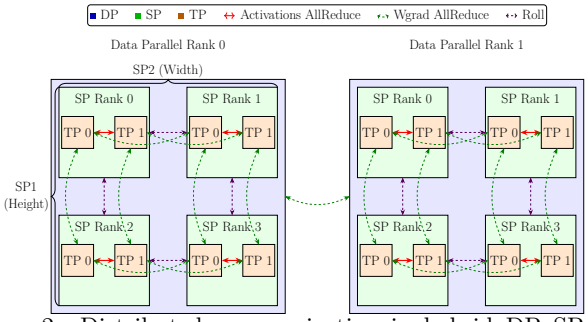


Figure 2: Distributed communication in hybrid DP–SP training: gradient synchronization across data-parallel replicas, halo and roll traffic for shifted-window attention under spatial decomposition, and partial-sum reductions for sharded attention and MLP blocks.

We use the hybrid data-spatial-tensor parallel distributed training framework from [15]. Global fields at 0.25° yield very large spatial tensors. With small batch sizes, activation memory often dominates over weight storage, especially when fine-tuning or analyzing autoregressive unrolls. We therefore rely on hybrid parallel training that combines data parallel (DP) replication over batches with two-dimensional spatial domain decomposition (SP1 along latitude, SP2 along longitude). This model also supports additional decomposition dimensions (for example, depth-parallel SP3, for 3D spatial inputs); they are not required for our runs as we treat our ERA5 inputs as 2D images. Tensor parallel (TP) can shard attention heads and MLP hidden dimensions with AllReduce collectives. A typical layout uses $dp \times sp_1 \times sp_2$ GPUs (and optional sp_3, tp), with DP groups outside and SP–TP nested within. Local patch-grid extents scale inversely with sp_1, sp_2. Shifted-window attention requires halo exchanges and a distributed cyclic roll along decomposed latitude–longitude axes so each rank sees correct window boundaries—Figure 2 sketches these communication patterns.

Per-device activation footprint scales roughly as $O(BH_\ell W_\ell E)$ for local height–width H_ℓ, W_ℓ, batch B, and embed dimension E, so increasing $sp_1 \cdot sp_2$ reduces memory in proportion to the subdomain area until communication or surface-to-volume effects dominate. The Swin window size (in patches) further limits how far sp_1 and sp_2 can be pushed; each rank must contain atleast one window.

Data

The network takes macroscopic variables and time-invariant fields (land–sea mask, orography, cosine latitude). The model input has shape $B \times T \times C \times H \times W$, where B is batch size, T is context length, C is the per-step channel count, and H, W are spatial dimensions. The autoregressive prediction target is the next state with shape $B \times 1 \times C \times H \times W$. In our experiments, we limit the size of T to 1 across all experiments.

We use the ERA5 reanalysis [26] with the 71-channel subset and pressure levels listed in Table 1.

Property	Description
Training years	1979–2016.
Validation year	2017.
Test years	2018–2022.
Spatial resolution	$0.25° \times 0.25°$.
Grid dimensions	1440×721 (longitude × latitude).
Surface variables	TCWV [kg/m^2]; u10m, v10m [m/s]; t2m [K]; sp, msl [Pa].
Pressure-level variables	u, v [m/s]; geopotential z [m^2/s^2]; temperature t [K]; specific humidity q [kg/kg].
Pressure levels (hPa)	50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000.
Total channels	6 surface + 5 fields × 13 levels = 71.
Temporal sampling	Hourly in ERA5; the model uses a 6 h forecast step (inputs and targets six hours apart).

Table 1: ERA5 setup and data splits.

Training Protocol

Our IsoFLOP construction matches [15]. Models are trained to predict the next six-hour state under uniform MSE over channels, with normalization statistics computed from the training split. Different IsoFLOP points are trained through constant learning rate (LR) schedule with periodic cooldowns to multiple compute budgets.

Results

Six-hour per-channel error structure

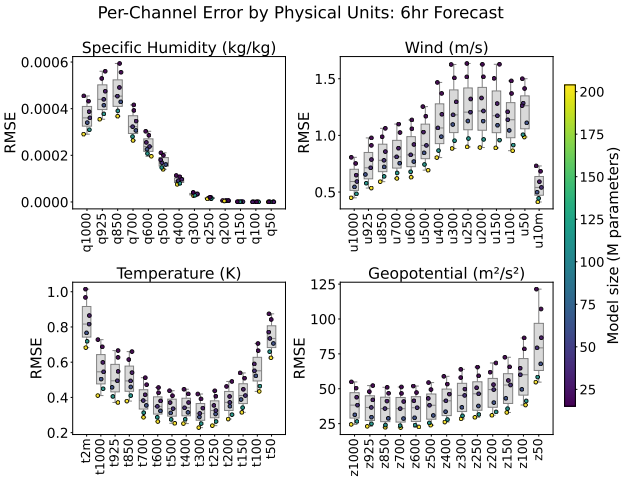


Figure 3: Per-channel area-weighted RMSE at six hours. Panels group channels by unit; each channel shows a box plot and one RMSE value per IC

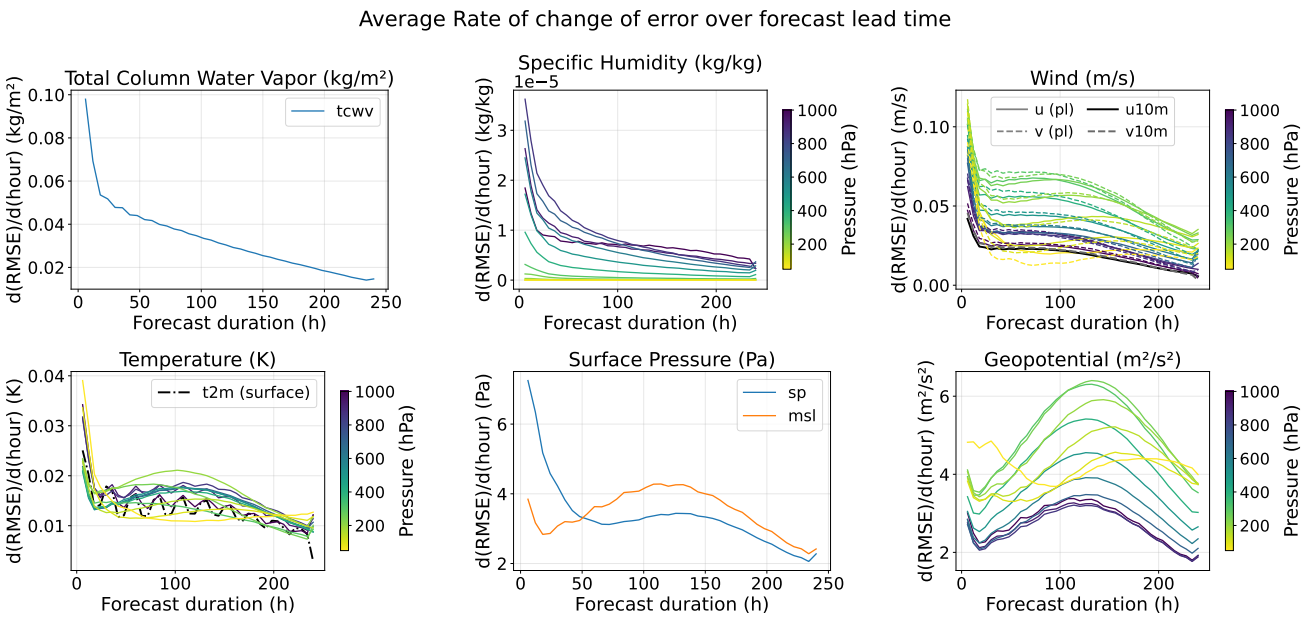


Figure 4: Average time derivative of area-weighted RMSE, $\mathrm{d}(\mathbf{RMSE})/\mathrm{d}t$, versus forecast lead time (hours), through 240 h. Curves are averaged over ICs; panels group channels by unit (color encodes pressure level where applicable).

Figure 3 shows how prediction error is distributed across targets at the first rollout step. Each component of the 71-dimensional output is one scalar channel; we plot area-weighted RMSE per channel, with one point per IC and a box summary on top. Channels are split into six panels by measurement unit so that values within a panel live on comparable numeric scales.

Several patterns are useful for downstream analysis. First, error is not evenly spread across channels. Even with a single global MSE loss during training, the implied per-channel RMSE ranges differ widely (both medians and spreads). Second, for variables repeated at multiple vertical indices (pressure levels), error often changes smoothly along that index: it is not i.i.d. across levels. For example, humidity channels tend to show larger RMSE at low levels and smaller RMSE at high levels, with a visible change of slope partway up the stack. Wind components show RMSE that grows toward the middle of the vertical index, then tapers slightly at the top. Temperature error is high for the dedicated surface channel, smaller through a band of mid-level channels, and rises again for the highest (sparsest) level we retain. Geopotential is relatively stable across most indexed levels and increases only toward the top of the stack. Among a few non-level channels in the same unit family, one pressure-related surface field is much harder than another (wider boxes and more scatter across ICs).

In short, the model induces a structured error profile over its output vector: targets differ in both typical error and IC-to-IC variability. That heterogeneity matters for any scaling-law fit that is computed per channel or pooled across channels, which we examine next.

Error Growth Over Rollout

Figure 4 summarizes the rate of error accumulation, not the error level: the vertical axis is the derivative of RMSE with respect to time. When this derivative is positive, RMSE is still rising; when the curve falls, the rise is slowing (error growth is decelerating), even if RMSE itself has not peaked within the window.

RMSE at a chosen lead is only a snapshot: the same value can arise from different temporal paths, and comparing two leads does not say how aggressively error is still being added in between. The derivative separates regime—whether each extra hour is still buying a large increment of error (high derivative), or whether the system is entering a saturation-like phase where RMSE may rise only slowly (small derivative). It also surfaces short windows where growth briefly speeds up again, which a few sparse RMSE evaluations can miss.

Across targets, the derivative curves are not simple translations of one another. Many channels show a rapid drop in the first day of integration: the largest incremental RMSE gains occur early, after which additional hours add error more slowly. That pattern is clearest for the single column-integrated target in the first panel and for several families of indexed channels, but the initial magnitude and decay rate differ by target.

For multi-level fields, the ordering of curves often mirrors the static RMSE stratification from Fig. 3: channels that were “hard” at six hours also tend to carry larger short-horizon derivatives, while some upper-level traces sit close to zero for long stretches (error barely climbs per hour). Wind and temperature panels illustrate that mid-stack indices can dominate both the level and the shape of the derivative, including modest mid-rollout structure (plateaus or secondary bumps) rather than monotone exponential decay.

The two surface pressure variables are a useful sanity check that the dynamics are target-specific: their deriva-

tive traces diverge in shape—one settles toward an approximate plateau after roughly four days while the other shows a broad mid-rollout maximum before easing. Geopotential is the starkest departure from an "early shock then decay": derivatives rise into a mid-horizon hump and only then decline, so error growth accelerates before it slows. That non-monotone velocity means a single functional form for "rollout error vs. time" will not fit all channels equally well.

Overall, Fig. 4 argues that autoregressive error should be analyzed as a multivariate time series over lead time: pooling channels or horizons without accounting for these profiles can hide which targets drive long-horizon drift. The next subsection holds the forecast horizon fixed at many values and asks how well a single power-law description of compute versus error still holds when error is measured after longer rollouts.

Global scaling-law fit vs. forecast horizon

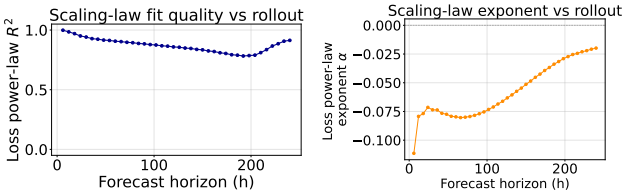


Figure 5: Global scaling-law diagnostics for log–log fits of compute vs. global test RMSE, plotted as a function of forecast horizon (hours). (a) shows the coefficient of determination R^2 as in the original analysis; (b) shows the fitted scaling exponent for each horizon. Each point re-fits the frontier using the same IsoFLOP design as in training.

Figure 5 does not plot RMSE itself. Panel (a) shows how well a straight line in log–log space explains the relationship between compute and globally test RMSE when the latter is evaluated at each forecast horizon. Panel (b) shows the slope of that same fit—the empirical scaling exponent linking log-compute to log-error at each horizon. Together, they explain "is the frontier still a line?" and "how steep is the frontier?"

Interpreting R^2 as linearity after log transforms, panel (a) shows that loss-based scaling is an excellent local approximation at short leads (R^2 near one at the first step) but not uniformly tight across the rollout. R^2 drifts down through a mid-window, reaches a trough near 190–200 h, and then rises again toward 240 h. We do not over-interpret the trough. It is consistent with a shift in which error components dominate overall RMSE after many autoregressive steps, and with the non-monotone error-velocity patterns in Fig. 4.

Panel (b) shows complementary information about sensitivity. The fitted exponent is most negative at the shortest horizons and drifts toward zero as the evaluation lead grows, with modest non-monotone structure in the first day or two before a steadier approach to the origin at long leads. In the usual sign convention for "more compute → lower RMSE," a slope closer to zero means that each additional factor of compute buys a smaller proportional reduction in error at the IsoFLOP frontier: scaling remains meaningful, but its apparent strength weakens once rollouts are long. Together with panel (a), the mid-horizon band is where linearity is worst and the slope is still far from its long-lead value. The late rise in R^2 does not reverse that trend toward a flatter exponent—by 240 h the exponent is still much closer to zero than at six hours—so late-rollout error is both relatively easier to summarize with a line and relatively less responsive to compute.

The global panels are coarse summaries: they average over all channels through pooled RMSE. Figure 6 resolves log–log goodness-of-fit along both forecast horizon and individual output channel, and repeats the analysis for three scaling covariates from the IsoFLOP design (parameter count, training data, and compute).

Per-channel R^2 across horizons and scaling axes

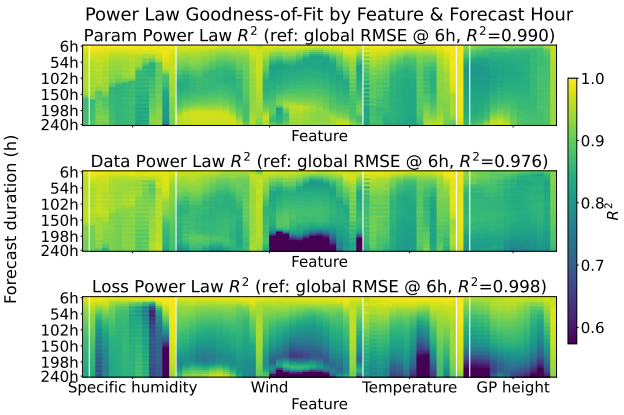


Figure 6: Heatmaps of R^2 for per-channel log–log scaling-law fits. Top: error vs. model size; middle: error vs. training data; bottom: error vs. compute. Brighter colors indicate a tighter linear fit in log–log space. Annotations give global test RMSE R^2 at 6 h for each scaling axis.

Figure 6 turns each cell of the tensor "(channel, lead, scaling variable)" into a single diagnostic. Reading down a column shows how well a fixed power-law survives as the evaluation rollout lengthens; reading across columns at a fixed lead shows which targets are easy or hard to describe with the same exponent family as the rest of the frontier.

The three rows answer different practical questions. Parameter scaling (top) is the most forgiving: R^2 stays high for most channels and leads, with only modest cooling (lower R^2) for some humidity and meridional-wind columns after roughly one week of integration. Data scaling (middle) is broadly similar but develops a sharper failure mode—a contiguous block of poor fits for low-level meridional wind past ∼180 h—showing that "more data" and "larger model" are not interchangeable as predictors of long-horizon error once channels are separated. Compute / loss scaling (bottom) tracks the training-aligned budget most directly: it yields the strongest short-horizon linearity (consistent with the strong linearity of the global reference at six hours) but also the most heterogeneous long-horizon map, with isolated pockets of low R^2 for upper-level humid-

ity, near-surface meridional wind, selected mid-to-upper temperature levels, and several geopotential columns at the longest leads.

A scalar scaling law fit to pooled loss (Fig. 5) can look excellent even while many individual channels violate the same template at late leads. The parametric row is a better "single exponent for everyone" summary if robustness across channels matters; the compute row is the right diagnostic if the scientific claim is tied to FLOPs or training budget. None of the three rows is uniformly dominant, which is expected once autoregressive evaluation mixes targets with different error velocities (Fig. 4) and different static difficulty (Fig. 3).

Discussion and Conclusions

We explore neural scaling for a global autoregressive weather transformer, specifically in the rollout regime. The primary message of our work is that scaling behavior depends strongly on what is being measured and when it is measured. A globally pooled metric can still exhibit strong log–log structure over compute at many horizons, yet channel-resolved diagnostics reveal substantial heterogeneity in both error magnitude and scaling-law fit quality.

From a modeling perspective, this matters because long-horizon forecasting quality is not governed by a single homogeneous error process. Per-channel RMSE at short lead is already structured across variables and pressure levels, and the derivative analysis shows that error accumulation rates also differ across targets over rollout time. These differences help explain why a single exponent summary can be informative at the aggregate level but incomplete for variable-level conclusions, especially at later horizons where some channels deviate from a common frontier while others remain well behaved.

The horizon-dependent global exponent provides a complementary interpretation to R^2: as rollout length increases, the fitted loss–compute slope moves closer to zero, suggesting weaker proportional returns from additional compute on pooled error at long leads, even when a log–log line remains a reasonable summary. In this sense, fit quality and scaling strength are related but distinct diagnostics, and both are needed for robust conclusions about compute allocation in autoregressive scientific forecasting.

These findings suggest practical directions for training and evaluation. First, weighted or channel-aware objectives may better align optimization with long-horizon error sensitivities than uniform MSE alone. Second, horizon-aware curricula or multi-objective training could reduce the mismatch between short-horizon fit quality and late-rollout robustness. Third, future scaling studies should report both global and channel-resolved diagnostics to avoid over-generalizing from pooled metrics.

This study has limitations. Our conclusions are conditioned on the current model family, data splits, and IsoFLOP design. Extending the same analysis to alternative architectures, larger compute regimes, and probabilistic objectives is a natural next step. Overall, the results support a nuanced view: scaling laws remain useful in autoregressive weather emulation, but their reliability and practical meaning are strongly horizon and channel-dependent.

Acknowledgments

This manuscript was co-authored by Oak Ridge National Laboratory (ORNL), operated by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. Any subjective views or opinions expressed in this paper do not necessarily represent those of the U.S. Department of Energy or the United States Government.

This research used the Perlmutter supercomputing resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231. An award of computer time was provided by the ASCR Leadership Computing Challenge (ALCC) program at NERSC under ERCAP0038267. Our conclusions do not necessarily reflect the position or the policy of our sponsors, and no official endorsement should be inferred.

References

[1] D. Szwarcman, S. Roy, P. Fraccaro, T. E. Gíslason, B. Blumenstiel, R. Ghosal, P. H. de Oliveira, J. L. de Sousa Almeida, R. Sedona, Y. Kang, S. Chakraborty, S. Wang, A. Kumar, M. Truong, D. Godwin, H. Lee, C.-Y. Hsu, A. Akbari Asanjan, B. Mujeci, T. Keenan, P. Arévolo, W. Li, H. Alemohammad, P. Olofsson, C. Hain, R. Kennedy, B. Zadrozny, G. Cavallaro, C. Watson, M. Maskey, R. Ramachandran, and J. Bernabe Moreno, "Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications," arXiv preprint arXiv:2412.02732, 2024. [Online]. Available: https://arxiv.org/abs/2412.02732

[2] J. Schmude, S. Roy, W. Trojak, J. Jakubik, D. S. Civitarese, S. Singh, J. Kuehnert, K. Ankur, A. Gupta, C. E. Phillips, R. Kienzler, D. Szwarcman, V. Gaur, R. Shinde, R. Lal, A. D. Silva, J. L. G. Diaz, A. Jones, S. Pfreundschuh, A. Lin, A. Sheshadri, U. Nair, V. Anantharaj, H. Hamann, C. Watson, M. Maskey, T. J. Lee, J. B. Moreno, and R. Ramachandran, "Prithvi WxC: Foundation Model for Weather and Climate," Sep. 2024. [Online]. Available: http://arxiv.org/abs/2409.13598

[3] T. Nguyen, A. Koneru, S. Li, and A. grover, "PhysiX: A Foundation Model for Physics Simulations," Jun. 2025. [Online]. Available: http://arxiv.org/abs/2506.17774

[4] S. Subramanian, P. Harrington, K. Keutzer, W. Bhimji, D. Morozov, M. W. Mahoney, and A. Gholami, "Towards Foundation Models for Scientific Machine Learning: Characterizing Scaling and Transfer Behavior," Advances in

Neural Information Processing Systems, vol. 36, pp. 71 242–71 262, Dec. 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/e15790966a4a9d85d688635c88ee6d8a-Abstract-Conference.html

[5] M. Herde, B. Raonić, T. Rohner, R. Käppeli, R. Molinaro, E. de Bézenac, and S. Mishra, "Poseidon: Efficient Foundation Models for PDEs," May 2024. [Online]. Available: http://arxiv.org/abs/2405.19101

[6] T. Nguyen, J. Brandstetter, A. Kapoor, J. K. Gupta, and A. Grover, "ClimaX: A foundation model for weather and climate," Dec. 2023. [Online]. Available: http://arxiv.org/abs/2301.10343

[7] X. Qin, D. Wang, J. Zhang, F. Wang, X. Su, B. Du, and L. Zhang, "TiMo: Spatiotemporal Foundation Model for Satellite Image Time Series," May 2025. [Online]. Available: http://arxiv.org/abs/2505.08723

[8] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, "Deep Learning Scaling is Predictable, Empirically," Dec. 2017. [Online]. Available: http://arxiv.org/abs/1712.00409

[9] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling Laws for Neural Language Models," Jan. 2020. [Online]. Available: http://arxiv.org/abs/2001.08361

[10] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. Rae, and L. Sifre, "An empirical analysis of compute-optimal large language model training," Advances in Neural Information Processing Systems, vol. 35, pp. 30 016–30 030, Dec. 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html

[11] S. Rasp, S. Hoyer, A. Merose, I. Langmore, P. Battaglia, T. Russell, A. Sanchez-Gonzalez, V. Yang, R. Carver, S. Agrawal, M. Chantry, Z. Ben Bouallegue, P. Dueben, C. Bromberg, J. Sisk, L. Barrington, A. Bell, and F. Sha, "WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models," Journal of Advances in Modeling Earth Systems, vol. 16, no. 6, p. e2023MS004019, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1029/2023MS004019

[12] V.-H. Peuch, R. Engelen, M. Rixen, D. Dee, J. Flemming, M. Suttie, M. Ades, A. Agustí-Panareda, C. Ananasso, E. Andersson, D. Armstrong, J. Barré, N. Bousserez, J. J. Dominguez, S. Garrigues, A. Inness, L. Jones, Z. Kipling, J. Letertre-Danczak, M. Parrington, M. Razinger, R. Ribas, S. Vermoote, X. Yang, A. Simmons, J. Garcés De Marcilla, and J.-N. Thépaut, "The Copernicus Atmosphere Monitoring Service: From Research to Operations," Bulletin of the American Meteorological Society, vol. 103, no. 12, pp. E2650–E2668, Dec. 2022. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/103/12/BAMS-D-21-0314.1.xml

[13] L. Harris, X. Chen, W. Putman, L. Zhou, and J.-H. Chen, "A Scientific Description of the GFDL Finite-Volume Cubed-Sphere Dynamical Core," 2021. [Online]. Available: https://doi.org/10.25923/6nhs-5897

[14] X. Wang, J.-Y. Choi, T. Kurihaya, I. Lyngaas, H.-J. Yoon, M. Fan, N. M. Nafi, A. Tsaris, A. M. Aji, M. Hossain, M. Wahib, D. Wang, P. Thornton, P. Balaprakash, M. Ashfaq, and D. Lu, "ORBIT-2: Scaling Exascale Vision Foundation Models for Weather and Climate Downscaling," May 2025. [Online]. Available: http://arxiv.org/abs/2505.04802

[15] S. Subramanian, A. Kiefer, A. Nigmetov, A. Gholami, D. Morozov, and M. W. Mahoney, "On Neural Scaling Laws for Weather Emulation through Continual Training," in ICLR 2026 Workshop on Foundation Models for Science: Real-World Impact and Science-First Design, Mar. 2026. [Online]. Available: https://openreview.net/forum?id=6xmHT5vO9P

[16] F. Mohr and J. N. v. Rijn, "Learning Curves for Decision Making in Supervised Machine Learning: A Survey," Machine Learning, vol. 113, no. 11-12, pp. 8371–8425, Dec. 2024. [Online]. Available: http://arxiv.org/abs/2201.12150

[17] I. M. Alabdulmohsin, X. Zhai, A. Kolesnikov, and L. Beyer, "Getting ViT in Shape: Scaling Laws for Compute-Optimal Model Design," Advances in Neural Information Processing Systems, vol. 36, pp. 16 406–16 425, Dec. 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/3504a4fa45685d668ce92797fbbf1895-Abstract-Conference.html

[18] M. Zhao, J.-C. Golaz, I. M. Held, H. Guo, V. Balaji, R. Benson, J.-H. Chen, X. Chen, L. J. Donner, J. P. Dunne, K. Dunne, J. Durachta, S.-M. Fan, S. M. Freidenreich, S. T. Garner, P. Ginoux, L. M. Harris, L. W. Horowitz, J. P. Krasting, A. R. Langenhorst, Z. Liang, P. Lin, S.-J. Lin, S. L. Malyshev, E. Mason, P. C. D. Milly, Y. Ming, V. Naik, F. Paulot, D. Paynter, P. Phillipps, A. Radhakrishnan, V. Ramaswamy, T. Robinson, D. Schwarzkopf, C. J. Seman, E. Shevliakova, Z. Shen, H. Shin, L. G. Silvers, J. R. Wilson, M. Winton, A. T. Wittenberg, B. Wyman, and B. Xiang, "The GFDL Global Atmosphere and Land Model AM4.0/LM4.0: 1. Simulation Characteristics With Prescribed SSTs," Journal of Advances in Modeling Earth Systems, vol. 10, no. 3, pp. 691–734, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/2017MS001208

[19] M. Eshagh, "Earth System Modeling, Data Assimilation, and Predictability: Atmosphere, Oceans, Land and Human Systems, 2nd Edition, Eugenia Kalnay, Sofa Mote and Cheng Da, Cambridge University

Press, 2024, Paperback £49.99, 340 pp., ISBN 978-1-10-740146-4," Weather, vol. 80, no. 9, pp. 319–319, 2025. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/wea.7729

[20] C. Bodnar, W. P. Bruinsma, A. Lucic, M. Stanley, A. Allen, J. Brandstetter, P. Garvan, M. Riechert, J. A. Weyn, H. Dong, J. K. Gupta, K. Thambiratnam, A. T. Archibald, C.-C. Wu, E. Heider, M. Welling, R. E. Turner, and P. Perdikaris, "A foundation model for the Earth system," Nature, vol. 641, no. 8065, pp. 1180–1187, May 2025. [Online]. Available: https://www.nature.com/articles/s41586-025-09005-y

[21] P. B. Holden, N. R. Edwards, P. H. Garthwaite, and R. D. Wilkinson, "Emulation and interpretation of high-dimensional climate model outputs," Journal of Applied Statistics, vol. 42, no. 9, pp. 2038–2055, Sep. 2015. [Online]. Available: https://doi.org/10.1080/02664763.2015.1016412

[22] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia, "Learning skillful medium-range global weather forecasting," Science, vol. 382, no. 6677, pp. 1416–1421, 2023. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.adi2336

[23] J. E. Matheson and R. L. Winkler, "Scoring Rules for Continuous Probability Distributions," Management Science, vol. 22, no. 10, pp. 1087–1096, Jun. 1976. [Online]. Available: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.22.10.1087

[24] B. Bonev, T. Kurth, A. Mahesh, M. Bisson, J. Kossaifi, K. Kashinath, A. Anandkumar, W. D. Collins, M. S. Pritchard, and A. Keller, "FourCastNet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale," Jul. 2025. [Online]. Available: http://arxiv.org/abs/2507.12144

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," 2021, pp. 10 012–10 022. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper

[26] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, and others, "The ERA5 global reanalysis," Quarterly Journal of the Royal Meteorological Society, vol. 146, no. 730, pp. 1999–2049, 2020. [Online]. Available: https://doi.org/10.1002/qj.3803

Author Biography

Alex Kiefer is a PhD student at the University of Tennessee - Knoxville and graduate research assistant at Oak Ridge National Laboratory. He received his BS and MS in computer science in 2022 from Indiana University. His work focuses on the efficient scaling of AI models on HPCs for spatio-temporal data. He is a member of ACM and SIGHPC.

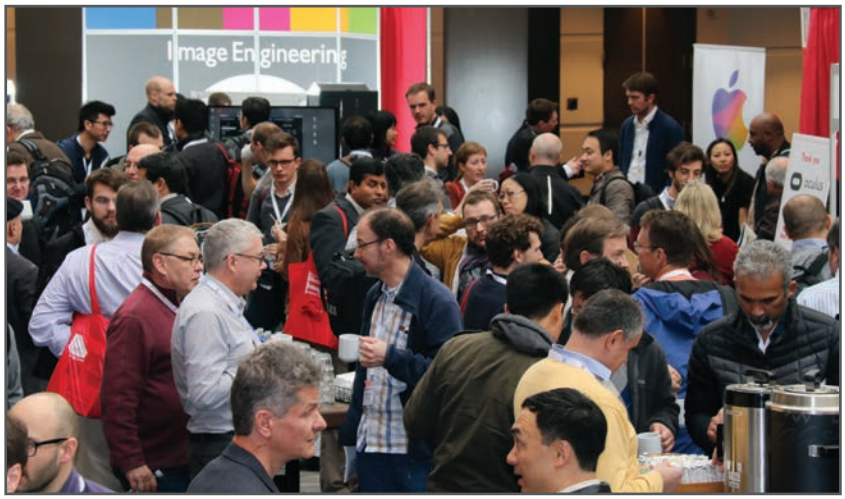
Dr. Prasanna Balaprakash is the Director of AI at Oak Ridge National Laboratory. He leads research efforts at Oak Ridge in launching transformational AI model consortium (ModCon), which is a basis for DOE Genesis Mission to accelerate science through AI. He received his PhD in 2009 from the Université libre de Bruxelles.

Dr. Xiao Wang is a research staff scientist in the Computational Science and Engineering Division at Oak Ridge National Laboratory (ORNL). He holds dual Bachelor's degrees in Mathematics and Computer Science, graduating with honors from Saint John's University, MN, in 2012. He completed his M.S. and Ph.D. in Electrical and Computer Engineering at Purdue University in 2016 and 2017, under the mentorship of Dr. Charles Bouman and Dr. Samuel Midkiff. Following his Ph.D., Dr. Wang pursued postdoctoral research at Harvard Medical School and Boston Children's Hospital, specializing in medical imaging, before joining ORNL in 2021. Dr. Wang's research focuses on the intersection of artificial intelligence (AI), high-performance computing (HPC), and computational imaging.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org