Pushing the Limits of High-Resolution Weather Forecasting through Data Scaling

Yang Zhao[*1,2,4], Peisong Niu[*3,4], Tian Zhou[*3,4], Ziqing Ma[4], Guanlong Ma[1,2], Rong Jin[4], Huiling Yuan[†1,2], and Liang Sun[†3,4]

[1] State Key Laboratory of Severe Weather Meteorological Science and Technology, Nanjing University, Nanjing, China
[2] School of Atmospheric Science, Nanjing University, Nanjing, China
[3] Ant Healthcare, Ant Group, Hangzhou, China
[4] DAMO Academy, Alibaba Group, Hangzhou, China
{zhaoy2024,guanlongma}@smail.nju.edu.cn, yuanhl@nju.edu.cn
{niupeisong.nps,tian.zt,ls.537724}@antgroup.com
maziqing.mzq@alibaba-inc.com, rongjinemail@gmail.com

Abstract. The development of 0.1° global weather forecasting models based on machine learning (ML) is constrained by the limited availability of high-resolution data, as decades of reanalysis are only available at 0.25° resolution. While existing approaches fine-tune 0.25° forecast models on limited 0.1° samples, we show that this transfer is hindered by the irreversible information loss inherent in coarse-resolution forecasting. Therefore, we propose **BaguanHR**, a framework that shifts the focus from transferring models to transferring data. We first show that super-resolution (SR) has lower conditional entropy and input amplification than forecasting, making it a more robust vehicle for resolution transfer. By leveraging this advantage through variable-wise SR, we synthesize extensive 0.1° data from ERA5. BaguanHR's performance on the synthetic-plus-real dataset exceeds both ML-based methods and IFS-HRES, achieving superior performance across over 85% of the lead times within 72 hours. Furthermore, our findings highlight a power-law scaling effect, as a twofold increase in data reduces RMSE by 4.6% for 72-hour forecasting and 4.9% for 120-hour forecasting. Our results demonstrate that scaling high-resolution ML-based forecasting is primarily a data bottleneck, and that variable-wise super-resolution provides a simple yet general solution to unlock long coarse-resolution reanalyses for high-resolution training.

Keywords: Weather forecasting · Scaling law · High-resolution

1 Introduction

The field of global weather forecasting is witnessing a profound change, with data-driven models now outperforming established numerical weather prediction

[*] These authors contributed equally to this work.
[†] Corresponding authors.

systems at medium ranges [3, 5, 6, 18–20, 27]. However, advancing these models toward higher spatial resolution remains fundamentally constrained by the availability of long, consistent, high-resolution global datasets. For example, the European Center for Medium-Range Weather Forecasts (ECMWF) operational analysis provides global 0.1° (roughly 9km) data only after 2016 [23], yielding at most about 10 years of training data up to 2026, which is far too limited for training large-scale ML forecast models without severe overfitting. In contrast, reanalysis data such as ERA5 [14] provides 47 years of data (up to 2026) only at a coarser 0.25° resolution, creating a resolution–data imbalance that has become a critical bottleneck for high-resolution ML weather forecasting.

A recent line of work attempts to overcome this challenge by transferring a pretrained 0.25° forecast model to 0.1° through architectural design and fine-tuning on the small amount of 0.1° high-resolution data [4, 12]. This strategy leverages prior dynamical knowledge encoded in the coarse model but is inherently limited by two factors. First, the pretrained coarse-resolution model cannot fully exploit the rich information contained in the 0.1° analysis fields because its learned representation is tied to the low-resolution grid. Second, weather prediction is a significantly more complex task than downscaling; adapting a coarse-resolution forecast model to a finer grid requires learning both fine-scale processes and the correct multiscale dynamics, which is difficult to achieve through limited fine-tuning. As a result, coarse-to-fine transfer often yields restricted accuracy improvements and depends heavily on sophisticated architectural modifications.

We propose a paradigm shift: transfer data, not models. **Our core insight is that super-resolution (SR) is a same-time conditional reconstruction task with provably lower conditional entropy than multi-step forecasting.** Following this insight, SR should be (i) intrinsically easier to learn, (ii) more stable under imperfect inputs, and (iii) suitable as a scalable mechanism to break the data bottleneck in high-resolution forecasting. We test these implications through a set of experiments that sequentially close this reasoning loop.

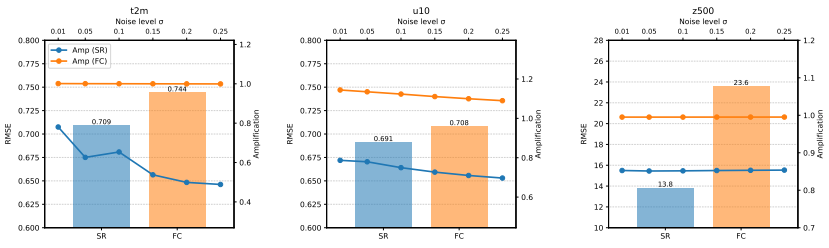

Fig. 1: Entropy-based Robustness evaluation for t2m, u10, and z500. The x-axis corresponds to different noise levels, and the y-axis is the RMSE and amplification factor. **Lower amplification factors reflect greater inherent stability.**

We investigate the mechanisms underlying our approach through two primary lenses. First, we analyze the robustness of SR and forecasting (FC) for t2m, u10 and z500, as shown in Fig. 1. By measuring the error growth per unit of injected Gaussian noise (amplification factor) for both tasks, we find that: (i) **SR is a simpler task**, yielding much lower RMSE than one-step forecasting (e.g., 13.8 vs. 23.6 for z500), which aligns with [11,24]; and (ii) **SR is more stable**, with an amplification factor (0.6–0.8×) significantly lower than that of forecasting (1.0–1.2×). Second, our **data scaling law analysis** (Sec. 5.2) demonstrates that 0.1° forecasting is strongly data-limited, exhibiting power-law improvements as training volume increases. Specifically, expanding the training data from 7 to 18 years reduces RMSE by more than 4.5% at long lead times.

Building on these findings, we employ a **synthetic-plus-real** strategy. Specifically, we train per-variable SR models on limited paired 0.1°–0.25° samples, apply them to decades of ERA5 to synthesize high-resolution 0.1° training data, and then train 0.1° forecast models from scratch. This data-transfer approach substantially outperforms coarse-to-fine baselines without specialized architectures or fine-tuning, indicating that scaling high-resolution ML forecasting is best addressed by data construction via variable-wise super-resolution.

The main contributions of our work are as follows:

- We propose BaguanHR, a framework leveraging the low conditional entropy of super-resolution (SR) as a robust data augmentation strategy for high-resolution weather forecasting. BaguanHR outperforms various ML-based models and IFS-HRES. Specifically, over **85%** of lead times show superior performance within 72 hours, highlighted by a **4.0%** RMSE reduction compared specifically to IFS-HRES.
- We conduct a comprehensive investigation from both experimental and theoretical perspectives to elucidate the mechanisms underlying the effectiveness of synthetic-plus-real strategy. By leveraging the inherent robustness of per-variable SR, our approach achieves up to a 20% reduction in RMSE compared to traditional coarse-to-fine methods, establishing a more stable framework for high-resolution modeling.
- We establish fundamental data scaling laws in high-resolution forecasting, revealing that performance improves following a power law as data volume increases. Specifically, expanding the training dataset from 7 to 18 years yields a **4.9%** reduction in RMSE for long-term (120-hour) forecasts and **4.6%** for 72-hour forecasts.
- Our study highlights a shift in perspective: for high-resolution ML-based weather forecasting, data construction via super-resolution proves more scalable and effective than architectural design or transfer learning. As the data availability is the primary bottleneck in high-resolution forecasting, our paradigm addresses the data directly, and can be expanded to other high-resolution tasks.

2 Related Work

2.1 Data-Driven Weather Forecasting Models

Moving beyond the resolution limits of 0.25° models [3–6, 19, 27], recent 0.1° forecasting efforts [4,12] focus on transferring knowledge from pretrained coarse-grained systems. Aurora [4] introduces bilinear upscaling for patch alignment, while Fengwu-GHR [12] employs identity-preserving mappings for high-resolution extrapolation. These methods highlight a growing trend of leveraging transfer learning to achieve operational-grade precision, but they still treat high-resolution forecasting primarily as a model-transfer problem under limited 0.1° data. In contrast, our approach tackles the same challenge from a data-centric perspective by expanding the effective 0.1° dataset via super-resolution. This data-centric route is complementary to neural-operator methods, including Fourier Neural Operators and coarse-graining operators [21, 31], which address multi-resolution physical modeling from an architecture-centric perspective.

2.2 Scaling in Weather Forecasting Task

Inspired by the scaling laws in NLP [17] and CV [7, 33], recent studies have begun exploring the relationship between compute, data, and performance in meteorology. While [27] analyzed model and data scaling at resolutions below 0.25°, [32] reported fundamental power laws across model size and compute budget. However, **whether these performance gains persist when scaling data at** 0.1° **resolution** remains an open question, especially given the scarcity of high-resolution analysis data. This motivates our investigation into effective data-scaling strategies within high-resolution regimes.

3 Rationale for Synthetic-to-Real Transfer

3.1 Robustness Evaluation: Super-Resolution vs. Forecasting

To justify the adoption of Super-Resolution as a superior data augmentation strategy, we conduct a comparative analysis between two paradigms: forecasting (temporal evolution) and super-resolution (spatial reconstruction) under varying levels of input noise. The formulations for these tasks are defined as follows:

- **High-Resolution Forecasting (FC):** BaguanHR is designed to generate global medium-range forecasts at a 0.1° spatial resolution ($H = 1801, W = 3600$). Let $X_t \in \mathbb{R}^{V \times H \times W}$ represent the atmospheric state at time t, where V denotes the number of variables. The forecasting model aims to learn a mapping $\Phi_F : X_t \to X_{t+\Delta t}$, where $\Delta t = 6$ hours. Future states are predicted in an auto-regressive manner, utilizing the previous output as input for the subsequent step.

– **Super-Resolution (SR):** To augment the training data, we define a spatial reconstruction task to bridge the gap between coarse reanalysis and high-resolution analysis data. Given a low-resolution state $Z_t \in \mathbb{R}^{V\times H'\times W'}$ (0.25°), the SR model learns a mapping $\Phi_{SR} : Z_t \to \hat{X}_t$, producing a 0.1° pseudo-label $\hat{X}_t$ that maintains physical consistency with the source data while enhancing spatial detail.

To evaluate model robustness, we perturb input fields with Gaussian noise $\delta \sim \mathcal{N}(0, \sigma^2)$ at varying intensities σ. Sensitivity is quantified by an **amplification factor**, defined as the change in prediction error relative to the baseline, normalized by the noise magnitude: $(|\mathrm{RMSE}_\sigma - \mathrm{RMSE}_0|/\sigma)$. This metric captures the marginal response of SR and FC models to input perturbations. By isolating the incremental error caused solely by input perturbations, this factor reveals the intrinsic stability of the model's mapping function. A high amplification factor suggests that small input fluctuations are significantly magnified during the inference process, whereas a low factor signifies a resilient architecture.

Our analysis reveals that the SR approach offers distinct advantages over the FC method in two key aspects. First, as illustrated in Fig. 1, SR consistently exhibits significantly lower amplification factors than forecasting across all tested meteorological variables, including t2m, u10, and z500. Second, SR maintains a substantial advantage in RMSE compared to the FC model at every noise intensity. Together, these findings suggest that SR is fundamentally less sensitive to input uncertainties and more effective at suppressing error propagation. These results demonstrate that **SR's inherent stability and resilience to input perturbations validate its application in robust data augmentation**.

3.2 Synthetic-plus-Real vs. Coarse-to-Fine Adaptation

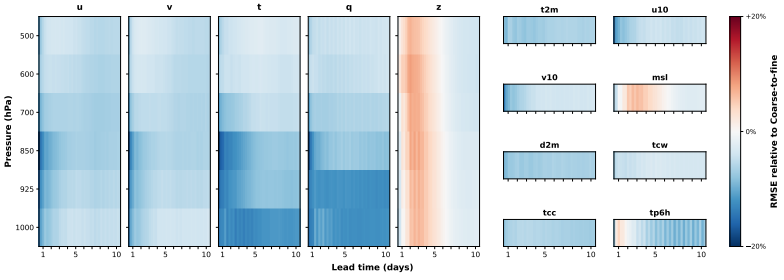

Fig. 2: Comparative performance of the "synthetic-plus-real" training-from-scratch paradigm versus the "coarse-to-fine transfer" approach. The x-axis is the lead time (days), and different colors show different values of relative RMSE difference.

At the onset of our study, we conduct a comprehensive comparison between two technical paradigms for high-resolution adaptation: the conventional "coarse-

to-fine" approach and our proposed "synthetic-plus-real" strategy. This comparison serves to identify **whether model-level transfer or data-driven augmentation provides a more effective route for** $0.1°$ **modeling**.

Experimental Setup Current $0.1°$ resolution models, represented primarily by [12] and [4], both follow a 'coarse-to-fine' technical route. In this work, we choose to implement our scheme by reproducing the methodology established by [12], which adapts a pre-trained $0.25°$ model to $0.1°$ resolution through decompositional and combinational transfer learning. In contrast, the "synthetic-plus-real" paradigm, as described in Sec. 4, augments the training set by generating high-resolution synthetic data and trains the $0.1°$ model from scratch. To ensure a fair comparison, both paradigms are evaluated on 2025 real-time analysis data.

Training Datasets The "synthetic-plus-real" training set integrates two key components:

- **Real-Time Analysis Data:** We utilize $0.1°$ ECMWF 4D-Var analysis data (2017–2024). For accumulated variables (fdir, ssrd, and tp) unavailable in the analysis archive, we source them from zero-hour IFS-HRES forecasts.
- **Synthetic Data:** To supplement the limited $0.1°$ archive, we apply a per-variable super-resolution (SR) model to $0.25°$ ERA5 reanalysis (2007–2016). This yields 10 years of $0.1°$ pseudo-labels that maintain the physical consistency of the original ERA5 records.

The complete dataset encompasses 13 standard pressure levels (50–1000 hPa) for upper-air variables (z, t, u, v, q) and a comprehensive set of surface parameters, including standard meteorological fields (t2m, d2m, u10, v10, msl, sp), specialized variables (u100, v100, lcc, tcc, sst), radiation metrics (ssrd, fdir), and moisture metrics (tcw, tcwv, tp). This dataset serves as the foundation for the experimental results presented in Sec. 5.

Results As illustrated in Fig. 2, the synthetic-plus-real paradigm significantly outperforms the coarse-to-fine baseline across the majority of variables. Notably, RMSE reductions reach as much as 20% (highlighted by the prominent blue regions). These findings provide empirical evidence for **why data transfer works**: training directly on high-fidelity synthetic data facilitates the learning of more robust spatial representations than parameter-level fine-tuning on interpolated inputs. Thus, **we adopt the synthetic-plus-real paradigm as our core methodology for all subsequent stages**.

3.3 Theoretical Justification of Synthetic Data Transfer

While including clean data can mitigate *model collapse* in iterative learning [9, 29], the consensus remains that synthetic data generally impairs generalization [2]. We challenge this view by demonstrating that an appropriate synthetic

generation procedure can significantly reduce generalization error in linear regression.

Consider a multi-cluster distribution where samples (x, y) are generated via $x = u \otimes e_k$ with $u \sim \mathcal{N}(\mu_k, \sigma^2 I)$ and $y = \langle w, x \rangle + z$ with noise $z \sim \mathcal{N}(0, \gamma^2)$. Let $D_a = \{(x_i, y_i)\}_{i=1}^n$ denote the training set. Under assumptions on μ_k (normalization, orthogonality, and separation) and $\sigma = o(1)$, we obtain an estimation of w(denoted as $\widehat{w}$) using the standard linear regression model:

$$\widehat{w} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i y_i \right). \tag{1}$$

The following theorem shows the estimation error for $\widehat{w}$ in Eq. (1).

Theorem 1. *With a probability* $\geq 1 - \delta$*, we have*

$$|w - \widehat{w}| \leq O\left(\gamma \sqrt{\frac{dm}{n} \log \frac{1}{\delta}} \right). \tag{2}$$

Suppose that we have N additional noisy examples $D_b = \{(u_i, y_i)\}_{i=1}^N$ ($N \gg n$), where each u_i is generated via the same process but lacks its cluster index k_i. Because this key information is latent, we refer to D_b as noisy data. The following theorem indicates that a linear regression model learned directly from D_b will suffer a substantially larger error compared to $\hat{w}$ defined in Eq. (1).

Theorem 2. *Let* $w_u \in \mathbb{R}^d$ *be the linear regression model we learned from* $\mathcal{D}_b$*. Assume* $N = +\infty$*. We have*

$$\mathrm{E}_{k \in [m]} \mathrm{E}_{u \sim \mathcal{N}(\mu_k, \sigma^2 I_d)} \left[|\langle u, w_u \rangle - \langle u, w_k \rangle|^2 \right] \geq \sigma^2 \, VAR\left(w_1, \ldots, w_m\right), \tag{3}$$

where

$$VAR\left(w_1, \ldots, w_m\right) = \mathrm{E}_{k \in [m]} \left[|w_k - \mathrm{E}_j \left[w_j \right]|^2 \right]. \tag{4}$$

Compared to $\widehat{w}$ in Eq. (1), we will have $\widehat{w}$ yields a smaller regression error than w_u, even when $N \to +\infty$. Our second approach is to learn from $\mathcal{D}_a$ a model to predict the index of Gaussian distribution, and apply the learned prediction model to augment the input patterns with the predicted index. The following theorem bounds the recovery error for $\widetilde{w}$.

Theorem 3. *Assume* $n \geq md \log \frac{N}{\delta}$*. Then, with a probability at least* $1 - \delta$*, we have*

$$|\widetilde{w} - w| \leq O\left(\sqrt{\frac{dm}{N} \log \frac{1}{\delta}} \right). \tag{5}$$

Comparing Theorem 3 to Theorem 1, we can see that the solution estimated from the synthetical dataset $\mathcal{D}_c$ is significantly closer to the oracle w than the one that is learned from the clean dataset $\mathcal{D}_a$. **This result theoretically validates the feasibility of employing super-resolution models for robust data augmentation**. Further details and proofs are provided in the supplementary material.

4 System Architecture and Training Strategies

4.1 Framework of BaguanHR

As illustrated in Fig. 3, BaguanHR comprises a per-variable super-resolution generator and a high-resolution forecasting model. To mitigate the scarcity of 0.1° EC analysis data, we employ the SR generator to synthesize 0.1° pseudo-labels from 0.25° ERA5 data.

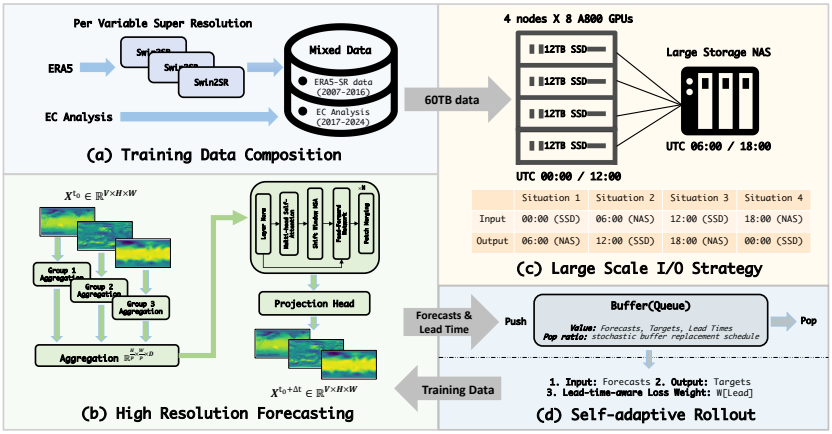

Fig. 3: Overview of BaguanHR's architecture. (a) Training Data Composition. (b) High-resolution forecasting model. (c) Large-scale I/O strategy. (d) Self-adaptive rollout strategy.

Per-Variable Super-Resolution Unlike weather forecasting, which is formulated as a physics-constrained initial value problem, super-resolution emphasizes topographic-guided spatial reconstruction. Given that variable-specific training yields higher fidelity [1,13,30], we adopt Swin2SR [8] to downscale 0.25° ERA5 data, treating each atmospheric variable independently. Each SR model is trained on paired 0.25°–0.1° fields using an MAE reconstruction objective.

High-Resolution Forecasting Model As illustrated in Fig. 3(b), the forecasting model processes inputs through variable-specific tokenization and hierarchical weather embedding, followed by multiple Swin Transformer blocks [22], and a linear projection head for spatial restoration.

Variable-Specific Tokenization and Hierarchical Embedding. Given an input of shape $V \times H \times W \times D$, where V is 80 (tp solely as an output), H is 1801 and W is 3600. Then each variable is independently tokenized via patchification with a patch size p, resulting in a reshaped tensor of $V \times h \times w \times D$ (where $h = H/p$ and $w = W/p$). However, the "brute-force" cross-attention aggregation common in prior works [25–27] becomes computationally prohibitive for such

high-resolution atmospheric data. To mitigate this, we propose a **hierarchical weather embedding**. By partitioning variables into 12 groups, our model first distills them into group-specific queries and then fuses them into a global summary. This two-level aggregation strategy effectively minimizes GPU memory overhead while streamlining the training process.

Swin Transformer Blocks and Projection Head. The embeddings are processed by Swin Transformer blocks to facilitate patch interactions. Finally, a linear projection head ($\mathbb{R}^{D\times Vp^2}$) restores the representation to its original spatial dimensions.

4.2 Self-Adaptive Rollout Strategy

To mitigate error accumulation in auto-regressive forecasting, recent models [4, 12] have adopted RL-inspired replay buffers for long-range stability. However, over-emphasizing long-lead horizons often degrades short-term accuracy due to a forgetting effect. To resolve this, we introduce two strategies for optimized replay buffer utilization that balance long-range stability with short-term fidelity.

- **Lead-Time-Aware Loss Weighting:** We employ lead-time-dependent weights to ensure short-lead predictions retain sufficient gradient influence. This prevents long-lead cumulative errors from dominating the optimization, thereby preserving foundational accuracy and mitigating error propagation during auto-regression.
- **Stochastic Buffer Replacement** To enhance data diversity, we implement a randomized buffer update strategy rather than a standard FIFO approach. By maintaining a heterogeneous mixture of lead times per batch, this method stabilizes distributed training and mitigates the instability typically caused by small local batch sizes.

4.3 Training and I/O Optimization

We also implement a scalable training framework for BaguanHR, leveraging advanced I/O partitioning and multi-stage optimization to ensure high-fidelity forecasting and high-throughput training.

- **Training Pipeline** BaguanHR utilizes 32 NVIDIA A800 GPUs in three stages: (1) **Pretraining**, for 250k steps (14 days) on 8-year EC analysis data; (2) **Synthetic Fine-tuning**, for 70k steps (4 days) incorporating 10-year super-resolution data; and (3) **Rollout Training**, for 2 days with a self-adaptive replay buffer to suppress long-range error accumulation.
- **I/O Strategies** To address the I/O bottleneck caused by 32-bit data files of 4–5 GB each, we *achieve a* 4× *speedup* via: (1) **Data Partitioning**, sharding data across local 12 TB SSDs per node; (2) **Hybrid Storage**, caching UTC 00:00/12:00 data on SSDs while keeping 06:00/18:00 data on NAS.

Overall, the training pipeline takes about 20 days on 32 NVIDIA A800 GPUs and uses about 60 TB of storage for 18 years of 0.1° synthetic-plus-real data; detailed compute and I/O costs are provided in Supplementary Sec. C.4.

5 Experiments

5.1 Dataset & Metrics

Detailed data descriptions are provided in Sec. 3.2. We adopt the standard weather evaluation methodology outlined in WeatherBench [28], focusing on forecasts initialized at 00/12 UTC for the year 2025. We use weighted Root Mean Square Error (RMSE) and weighted Anomaly Correlation Coefficient (ACC), which have also been widely used in AI-based weather models [3, 6, 19, 27], as our main metrics.

5.2 The Scaling of Data Volume

Fig. 4 reveals consistent power-law-like improvements in forecast skill as the volume of training data increases. This confirms the pivotal role of data scaling in ML-based weather models. Beyond this overall trend, we observe:

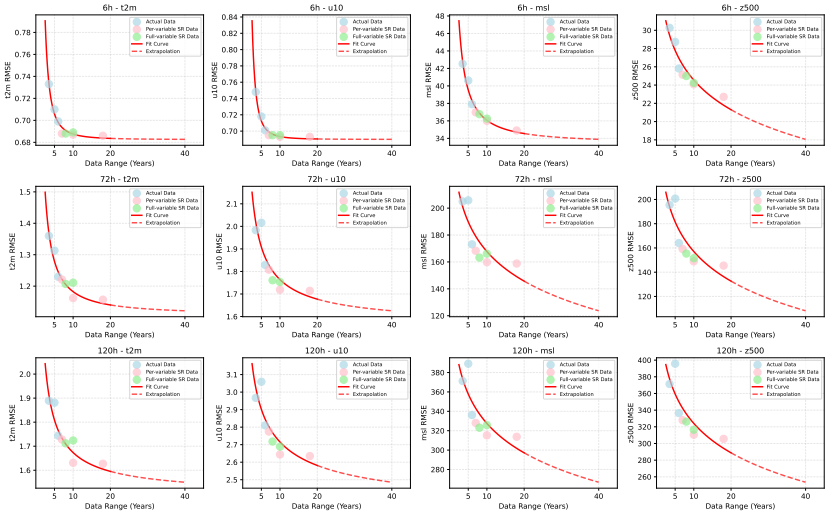

Fig. 4: Scaling law curves for different variables and lead times. Different subplots correspond to different variables and forecasting horizons. We plot the x-axis in log scale to reflect the scaling law.

1. **Scaling Consistency Across Lead Times:** The power-law behavior of error reduction remains remarkably stable across different forecast lead times. This consistent scaling slope suggests that increasing data volume is a universally robust strategy for improving both short-term deterministic and long-term atmospheric predictions. Notably, for long-term forecasting, RMSE is reduced by **4.6%** at 72 hours and **4.9%** at 120 hours.

2. **Theoretical Saturation and Diminishing Returns:** To provide practical guidance for large-scale training, we conducted an extrapolation analysis beyond the 18-year data threshold. While the full 47-year ERA5 archive could be leveraged for data augmentation via SR, this extrapolation should be interpreted as a trend analysis rather than a definitive performance estimate. The fitted curves suggest that marginal performance gains may gradually diminish beyond 18 years, indicating a practical trade-off where further data expansion yields sub-linear improvements relative to the increasing storage and I/O cost.
3. **Variable-Specific Sensitivity to Data Volume:** Sensitivity to data scaling varies significantly by physical field: while most variables (e.g., z500, t2m) plateau early, others like msl exhibit sustained headroom for improvement. This heterogeneity informs our decision to standardize on an 18-year period to optimize overall storage and I/O efficiency.
4. **Efficacy of Granular Data Augmentation:** Comparative analysis shows that models trained on data generated by per-variable SR outperform those trained on full-variable SR output at a 10-year scale. This advantage is particularly pronounced for surface variables (e.g., t2m), highlighting the value of decoupled strategies in data synthesis.

5.3 Performance Evaluation and Benchmarking

To evaluate the model's efficacy, we adopt a dual-perspective approach: first, a holistic global assessment across multiple variables to establish baseline predictive skill, followed by a rigorous investigation into model robustness under diverse extreme scenarios.

General Comparison of Variables To show the superiority of BaguanHR, we evaluate it against both physics-based and ML-based baselines. IFS-HRES [10] serves as the operational physics-based 0.1° reference. For ML baselines, we use Baguan+swin2sr as the primary comparison, where 0.25° Baguan forecasts are adapted to the 0.1° grid using Swin2SR. Since the first 3-day forecast is more accurate, directly actionable, and therefore especially important, we further include GraphCast+swin2sr, Pangu+swin2sr, and Baguan+GHR for short-term lead times. As Fengwu-GHR [12] cannot be directly replicated, Baguan+GHR is re-implemented based on Baguan [27] as a coarse-to-fine high-resolution adaptation baseline.

BaguanHR surpasses the baselines (Fig. 5), improving over **85%** of lead times within 72 hours and reducing RMSE by **4.0%** compared to IFS-HRES. Specifically, BaguanHR achieves an average RMSE reduction of **5.8%** at 24 hours and **9.7%** at 72 hours relative to the IFS-HRES baseline. Furthermore, BaguanHR demonstrates a clear performance edge over 0.25° outputs upscaled via Swin2SR [8], particularly in short-range windows. While auto-regressive systems often suffer from variance loss and converge toward the ensemble mean over time, post-processing techniques are inherently limited by the information bottleneck of coarse-grained inputs. In contrast, **BaguanHR's end-to-end**

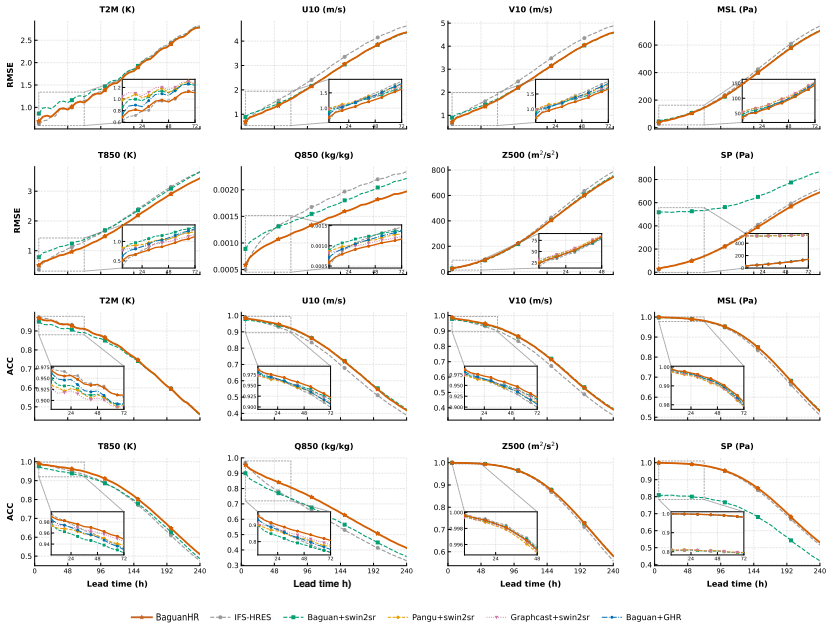

Fig. 5: Overall performance of BaguanHR against several baselines. 'Baguan+GHR' denotes the version replicated from [12]. We zoom in the first 72 hours for better comparison among different methods.

high-resolution framework circumvents these limitations by capturing fine-grained dynamics natively, yielding a more physically consistent and detailed atmospheric representation.

Predictive Skill in Extreme Scenarios To evaluate BaguanHR's reliability in high-impact scenarios, we assess its performance across three categories of extremes: categorical detection of intense moisture events, predictive accuracy for both relative (percentile-based) and absolute (threshold-based) extremes.

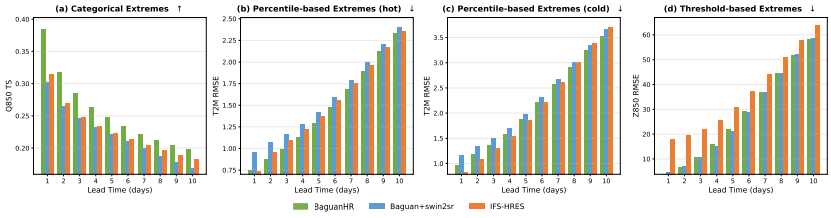

Fig. 6: Extreme event verification. Higher TS represents better categorical skill; lower RMSE signifies enhanced accuracy for percentile and threshold-based extremes.

Categorical Extreme Event Detection. Using q850 $\geq$ 14 g/kg as a thermodynamic proxy for heavy precipitation [15], we evaluate detection skill via the Threat Score (TS; Fig. 6(a)). BaguanHR consistently outperforms both Baguan+swin2sr and IFS-HRES, achieving gains exceeding 10% in most cases. This confirms its superior ability to resolve the intense moisture convergence required for high-impact events.

Percentile-based Relative Extremes. To identify statistical anomalies across latitudes, we define extreme heat and cold using the local 90th and 10th percentiles of t2m [34] (Fig. 6(b–c)). BaguanHR outperforms competitors in more than 70% of cases, surpassing Baguan+swin2sr by 15% in the short term and maintaining a 3% edge in long-term extremes. This highlights its precision in capturing the "heavy-tail" thermal distributions often smoothed by lower-resolution models.

Threshold-based Absolute Extremes. To evaluate dynamical intensity, we identify mid-latitude low-pressure systems using an absolute threshold (z850 $<$ 13,500 m^2/s^2) [16] (Fig. 6(d)). BaguanHR delivers the best results, outperforming IFS-HRES by over 30% and maintaining a significant lead over Baguan+swin2sr at both short (1–2 days) and long (9–10 days) lead times. These results demonstrate its capacity to reconstruct the sharp pressure gradients of severe storms.

5.4 Controlled Comparisons

We perform controlled comparisons to justify the core components of BaguanHR across four dimensions: variable modeling (independent vs. joint), physical consistency of SR fields, power spectral analysis, and optimization stability (replay buffer and adaptive weighting).

Full-Variable vs. Per-Variable Super-Resolution Fig. 7 (a) shows the RMSE improvement of per-variable SR over full-variable SR across variables. It shows that per-variable SR significantly improves surface variables (e.g., 90.7% for sp, >20% for t2m and msl) by capturing local specificities. Upper-air fields achieve modest improvements (1.6%–4.7% for temperature and geopotential). This suggests that unlike forecasting, SR favors intra-variable textures over cross-variable coupling, making variable-independent modeling more effective.

Physical Consistency of SR Fields We diagnose the physical consistency of SR-generated pseudo-labels using three diagnostics based on relative humidity (RH), geostrophic wind speed $|\mathbf{V}_g|$, and potential temperature differences $\Delta\theta$. Their latitude-weighted correlations against the 0.1° analysis remain high across pressure levels ($\mathbf{RH} \geq 0.94$, $|\mathbf{V}_g| \geq 0.82$, $\Delta\theta \geq 0.93$), indicating that per-variable SR preserves key moisture, dynamical, and thermodynamic structures, alleviating concerns about physically inconsistent high-frequency artifacts. Per-level numbers and the power-spectrum analysis are in Supplementary Sec. E.3.

Power Spectral Analysis We compute power spectral density (PSD; variance/energy as a function of spatial wavelength) for 6–72h forecasts over the full 2025 validation dataset; EC analysis denotes the corresponding year-mean PSD

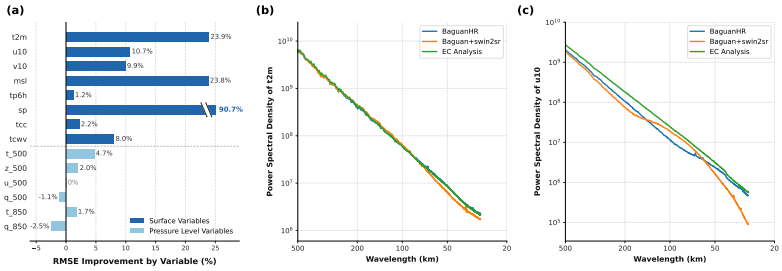

Fig. 7: (a) Relative RMSE improvement of per-variable versus full-variable super-resolution. (b-c) Power spectral analysis of t2m and u10.

of analysis data. The results in Fig. 7 (b-c) demonstrate that BaguanHR maintains higher short-wavelength spectral power, effectively preserving mesoscale structures. Conversely, the Baguan+swin2sr pipeline exhibits systematic energy attenuation, suggesting that post-hoc SR tends to smooth fields rather than recovering fine-scale variability from coarse inputs.

Self-Adaptive Rollout Strategy Tab. 1 validates our enhanced replay buffer strategy. Loss weighting suppresses auto-regressive error propagation, as its removal leads to degraded performance on long-range forecasts. Stochastic replacement further boosts performance by ensuring lead-time diversity. Together, BaguanHR harmonizes divergent loss magnitudes across horizons, ensuring stable optimization and superior metrics across all variables.

Table 1: Performance comparison of different strategies on replay buffer.

Methods	T2M-72h		U10-72h		MSL-72h		TCC-72h	
	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑
w/o Random & loss weight	1.12	0.90	1.64	0.91	147.80	0.980	0.289	0.61
w/o Random	1.12	0.91	1.63	0.91	147.41	0.980	0.288	0.62
w/o loss weight	1.11	0.91	1.63	0.92	145.32	0.981	0.289	0.62
BaguanHR	**1.09**	**0.92**	**1.62**	**0.92**	**144.57**	**0.981**	**0.287**	**0.64**

5.5 Case Studies

The aforementioned capabilities of BaguanHR are further demonstrated through the case studies of tropical cyclone and cold-air outbreak prediction.

Tropical Cyclone Prediction Tropical cyclones are among Earth's most hazardous weather events. While prior ML methods typically lag behind the IFS-HRES in intensity [3, 4], BaguanHR achieves comparable intensity while maintaining superior track accuracy. For the double-landfall cyclone CO-MAY, BaguanHR precisely identifies two landfall sites, outperforming the IFS-HRES track (Fig. 8(a)). Furthermore, as shown in Fig. 8(b), BaguanHR substantially outperforms Baguan+swin2sr in both intensity and track errors.

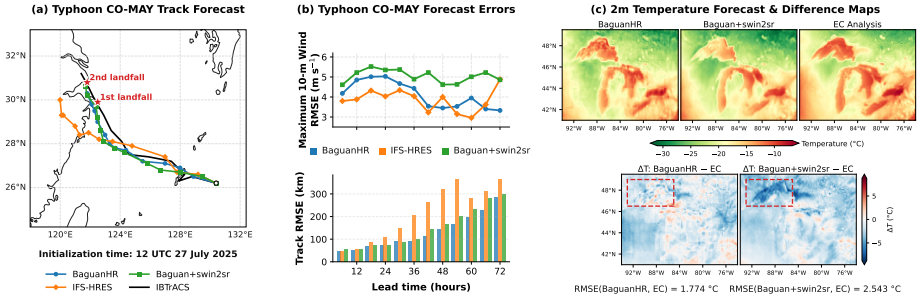

Fig. 8: (a) Typhoon CO-MAY tracking forecast, which is initialized at 12 UTC, 27 July 2025. (b) Forecast errors of track and intensity. (c) Case study of 2m temperature forecast over the U.S. Great Lakes region, initialized at 00 UTC, 20 January 2025.

Cold-Air Outbreak Fig. 8(c) presents a 2m temperature (t2m) forecast case study in a cold-air outbreak over the U.S. Great Lakes region. BaguanHR successfully captures the spatial structure characteristics, e.g., the distinct temperature characteristics over lake surfaces compared to surrounding land areas, a level of physical detail that the post-hoc Baguan+swin2sr baseline fails to resolve, demonstrating its superior performance in extreme value prediction.

6 Conclusion

In this paper, we present BaguanHR, a 0.1° global weather forecasting model built on a data-transfer framework that leverages super-resolution-generated synthetic data. We demonstrate fundamental data scaling laws in high-resolution forecasting, showing that model performance follows a power-law improvement with training data volume. Using synthetic 0.1° data, BaguanHR achieves superior performance across nearly all variables, especially in the first 72 hours, outperforming both IFS-HRES and SR-enhanced 0.25° models. At 10–15 day lead times, BaguanHR brings limited gains as forecasts are increasingly controlled by large-scale atmospheric patterns and fine-scale details become weakly predictable. Building on Baguan, we are developing the NJU-Earth series as an evolving model family for global weather forecasting, with BaguanHR as its high-resolution component. This work will be extended toward ensemble forecasting and operational applications such as tropical cyclone intensity prediction.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (U2342218), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM907), the DAMO Academy Research Intern Program, and the High Performance Computing Center, Nanjing University (NJU).

References

1. Addison, H., Kendon, E., Ravuri, S., Aitchison, L., Watson, P.A.: Machine learning emulation of a local-scale uk climate model. arXiv preprint arXiv:2211.16116 (2022)
2. Amin, K., Bie, A., Kong, W., Syed, U., Vassilvitskii, S.: Learning from synthetic data: Limitations of erm (2026), https://arxiv.org/abs/2601.15468
3. Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q.: Accurate medium-range global weather forecasting with 3d neural networks. Nature **619**(7970), 533–538 (2023)
4. Bodnar, C., Bruinsma, W.P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J.A., Dong, H., Gupta, J.K., Thambiratnam, K., Archibald, A.T., Wu, C.C., Heider, E., Welling, M., Turner, R.E., Perdikaris, P.: A foundation model for the earth system. Nature **641**(8065), 1180–1187 (May 2025). https://doi.org/10.1038/s41586-025-09005-y
5. Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J., Chen, X., Ma, L., Zhang, T., Su, R., Ci, Y., Li, B., Yang, X., Ouyang, W.: The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time. Communications Earth & Environment (2025). https://doi.org/10.1038/s43247-025-02502-y
6. Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., Li, H.: Fuxi: a cascade machine learning forecasting system for 15-day global weather forecast. npj Climate and Atmospheric Science (1), 190. https://doi.org/10.1038/s41612-023-00512-1
7. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2818–2829 (2023)
8. Conde, M.V., Choi, U.J., Burchi, M., Timofte, R.: Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In: European Conference on Computer Vision. pp. 669–687. Springer (2022)
9. Dey, A., Donoho, D.: Universality of the $\pi^2/6$ pathway in avoiding model collapse (2024), https://arxiv.org/abs/2410.22812
10. European Centre for Medium-Range Weather Forecasts: ECMWF Integrated Forecasting System (IFS) High Resolution (HRES) Forecasts. https://www.ecmwf.int/ (2024), accessed: 2025-01-01
11. Guen, V.L., Thome, N.: Disentangling physical dynamics from unknown factors for unsupervised video prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11474–11484 (2020)
12. Han, T., Guo, S., Ling, F., Chen, K., Gong, J., Luo, J., Gu, J., Dai, K., Ouyang, W., Bai, L.: Fengwu-ghr: Learning the kilometer-scale medium-range global weather forecasting. arXiv preprint arXiv:2402.00059 (2024)
13. Harris, L., McRae, A.T., Chantry, M., Dueben, P.D., Palmer, T.N.: A generative deep learning approach to stochastic downscaling of precipitation forecasts. Journal of Advances in Modeling Earth Systems **14**(10), e2022MS003120 (2022)
14. Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The era5 global reanalysis. Quarterly Journal of the Royal Meteorological Society **146**(730), 1999–2049 (2020)
15. Holton, J.R., Hakim, G.J.: An introduction to dynamic meteorology, vol. 88. Academic press (2013)

16. Hoskins, B.J., McIntyre, M.E., Robertson, A.W.: On the use and significance of isentropic potential vorticity maps. Quarterly Journal of the Royal Meteorological Society **111**(470), 877–946 (1985)
17. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
18. Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., Anandkumar, A.: Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In: Huebl, A., Silvano, C., Robinson, T. (eds.) Proceedings of the Platform for Advanced Scientific Computing Conference, PASC 2023, Davos, Switzerland, June 26-28, 2023. pp. 13:1–13:11. ACM (2023). https://doi.org/10.1145/3592979.3593412
19. Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., Battaglia, P.: Learning skillful medium-range global weather forecasting. Science **382**(6677), 1416–1421 (2023). https://doi.org/10.1126/science.adi2336
20. Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M.C.A., Lessig, C., Maier-Gerber, M., Magnusson, L., Bouallègue, Z.B., Nemesio, A.P., Dueben, P.D., Brown, A., Pappenberger, F., Rabier, F.: Aifs – ecmwf's data-driven forecasting system (2024), https://arxiv.org/abs/2406.01465
21. Li, Z., Kovachki, N.B., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A.M., Anandkumar, A.: Fourier neural operator for parametric partial differential equations. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=c8P9NQVtmnO, accessed: 2026-06-28
22. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
23. Malardel, S., Wedi, N., Deconinck, W., Diamantakis, M., Kühnlein, C., Mozdzynski, G., Hamrud, M., Smolarkiewicz, P.: A new grid for the ifs. ECMWF newsletter **146**(23-28), 321 (2016)
24. Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.Y., Liu, C.C., Vahdat, A., Kashinath, K., Kautz, J., Pritchard, M.: Generative residual diffusion modeling for km-scale atmospheric downscaling. arXiv preprint arXiv:2309.15214 **10** (2023)
25. Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J.K., Grover, A.: Climax: A foundation model for weather and climate. In: International Conference on Machine Learning (2023), https://api.semanticscholar.org/CorpusID:256231457
26. Nguyen, T., Shah, R., Bansal, H., Arcomano, T., Madireddy, S., Maulik, R., Kotamarthi, V.R., Foster, I.T., Grover, A.: Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. Advances in Neural Information Processing Systems **37**, 13689–13723 (2024)
27. Niu, P., Ma, Z., Zhou, T., Chen, W., Shen, L., Jin, R., Sun, L.: Utilizing strategic pre-training to reduce overfitting: Baguan-a pre-trained weather forecasting model. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. pp. 2186–2197 (2025)
28. Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russel, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Bouallegue, Z.B., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., Sha, F.: Weatherbench 2: A benchmark for the next generation of data-driven global weather models (2024), https://arxiv.org/abs/2308.15560

29. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y.: Ai models collapse when trained on recursively generated data. Nature **631**, 755–759 (2024)
30. Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., Ganguly, A.R.: Deepsd: Generating high resolution climate change projections through single image super-resolution. In: Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining. pp. 1663–1672 (2017)
31. Wang, C., Berner, J., Li, Z., Zhou, D., Wang, J., Bae, J., Anandkumar, A.: Coarse graining with neural operators for simulating chaotic systems. arXiv preprint arXiv:2408.05177 (2024)
32. Yu, Y., Huang, L., Calotoiu, A., Hoefler, T.: Scaling laws of global weather models (2026), `https://arxiv.org/abs/2602.22962`
33. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12104–12113 (2022)
34. Zhang, X., Alexander, L., Hegerl, G.C., Jones, P., Tank, A.K., Peterson, T.C., Trewin, B., Zwiers, F.W.: Indices for monitoring changes in extremes based on daily temperature and precipitation data. Wiley Interdisciplinary Reviews: Climate Change **2**(6), 851–870 (2011)