Block-wise distillation for lightweight weather models

Daniil Sukhorukov
MIRAI
Industrial AI Lab

Andrei Zakharov
Industrial AI Lab

Dmitry Zhevnenko
Industrial AI Lab

Vladimir Kirilin
Independent Researcher

Ekaterina Muravleva
AI4Science Center
Institute of Numerical Mathematics RAS

Ivan Oseledets
Industrial AI Lab
Institute of Numerical Mathematics RAS

Ilya Makarov
Industrial AI Lab
ISP RAS

Abstract

State-of-the-art machine learning weather forecasting systems, such as FuXi, achieve skillful global predictions but at the cost of large model sizes and high training demands. In this work, we investigate how far such architectures can be reduced without significant loss of accuracy. Specifically, we compress FuXi-short by replacing its 48 SwinTransformerV2 blocks with only 6 (additionally evaluating 4- and 2-block variants), and probe two training strategies: (i) training the reduced model from scratch, including an efficient one-step regression initialization to quickly adapt the architecture to weather data, and (ii) block-wise distillation, where each reduced block is trained to approximate every 8th block of the original model using MSE loss. Despite the eightfold reduction in depth, accuracy on the most critical variables remains effectively unchanged. For example, mean sea level pressure RMSE increases by 0.026% relative to the mean and 1.95% relative to the standard deviation, while temperature RMSE changes only by 0.068% and 0.87%, respectively. Importantly, with one-eighth the depth, the model is substantially faster to train, enabling more agile adaptation to changing climate data. These results highlight the importance and limits of architectural compression of large models where forecasting skill can be retained even under drastic reduction. In this ongoing work we will quantify benefits of such approach, explore compression strategies, and assess robustness across seasons and longer horizons.

1 Introduction

As climate change drives more extreme weather, forecasting systems must be updated rapidly and run at high spatial resolution. Conventional numerical weather prediction (NWP) is extremely compute-intensive, often requiring high-performance clusters and hours per run [1]. In contrast, DL-based models can leverage GPUs for much faster inference. In recent years a variety of such models has been developed to more efficiently tackle the weather problem through employing advanced training techniques and architectures, including CNNs [2], GNNs [3, 4, 5], Transformers [6, 7, 8, 9, 10], and hybrid models [11]. For example, Price et al. [12] show that an ensemble model (GenCast) produces a 15-day global forecast (0.25° resolution) in on the order of 8 minutes. These speedups suggest DL forecasting could support rapid forecast cycles, but training state-of-the-art models at scale remains costly and time-consuming.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Recent DL weather systems have indeed achieved remarkable accuracy. For example, Chen et al. introduced FuXi [13], a cascaded Swin Transformer V2 model [14] delivering 15-day global forecasts at 0.25° (6-hour steps). Its core architecture is a ”U-Transformer” with 48 sequential SwinV2 blocks, trained on 39 years of ERA5 reanalysis data [15]. In evaluation, FuXi's forecasts match the ECMWF 15-day ensemble mean in skill and significantly extend deterministic forecast lead times. However, achieving this performance is computationally intensive. Authors report roughly 30 hours of pretraining on 8 A100 GPUs (plus days of fine-tuning) just to train FuXi-short. The substantial compute demands of such models make frequent retraining or real-time updates nearly impractical.

To address this, we investigate compressing a version of the FuXi model into a lightweight model with minimal skill loss. We compress the FuXi-short (0–5 day) model from 48 blocks to 6,4,2 blocks and compare their forecasts to FuXi-short on a set of key variables (2m temperature, humidity, MSLP, wind and other parameters), and evaluate for that two distinct strategies. First, we train the 6-block network from scratch, initializing its first layer by a one-step linear regression to mimic the teacher's one-step forecast. Second, we use block-wise distillation [16] where each student block is trained to mimic the output of a corresponding teacher block, aligning for example student blocks 1–6 with teacher blocks 8, 16, 24, 32, 40, 48, and using a mean-squared-error loss. This block-level supervision is intended to transfer intermediate features from the teacher model. Training and evaluation are performed on held-out ERA5 data, the same reanalysis used to develop FuXi, providing us with the direct comparison of the compressed model's forecast skill to the FuXi-short teacher.

While reducing deep models for efficiency is a well-studied ML problem, common approaches mostly include designing smaller architectures, pruning parameters, quantization, neural architecture search, and knowledge distillation. In knowledge distillation (KD), a small student network is trained to mimic a large teacher network's behavior (e.g. matching its output logits or hidden-layer features). Introduced by Hinton et al.[17], KD enables deploying neural networks on resource-limited hardware without major performance loss. In practice, KD is valued for significantly reducing inference cost and model size while preserving accuracy.

Beyond matching only final outputs, layer-wise or block-wise distillation methods align intermediate representations. Such techniques help bridge the capacity gap between deep teachers and shallow students. For example, Wang et al.[18] propose Progressive Block-wise KD (PBKD), which gradually replaces teacher subnetworks with student subnetworks block by block, aligning their features at each stage. Blakeney et al.[19] describe a parallel block-level strategy where multiple layers are distilled in parallel and then merged. These studies demonstrate that aligning student blocks to teacher blocks, for example by minimizing an MSE on their outputs, can effectively compress deep transformer models.

Weather forecasting has advanced rapidly using large ConvNets or Transformer models trained on reanalysis data. Standard benchmarks, such as WeatherBench [20, 21], use ERA5 reanalysis as ground truth. To date, most research has focused on maximizing accuracy of large models, with little attention to compression. To our knowledge no prior work has distilled a global-scale weather transformer into a much smaller model. This work bridges DL compression methods with weather forecasting by applying block-distillation to the FuXi transformer, aiming to retain forecast skill under aggressive model downscaling.

2 Methods

Data. We use the ERA5 reanalysis dataset as the training (2016-2019 years) and evaluation source (2020-2021 years). Input fields include total of 70 meteorological variables both at the surface and within 13 different atmospheric layers. For consistency with FuXi, we use six-hourly data at 0.25° horizontal resolution. The dataset is split into training, validation, and test periods by year, ensuring no overlap in time between train and evaluation. All experiments are conducted using the same preprocessing as in [13] to allow direct comparison with FuXi-short.

Teacher model. Our teacher is the FuXi-short model [13], designed to forecast the 0–5 day horizon using a Swin Transformer V2–based architecture with 48 sequential transformer blocks and total of 1556M parameters. It was pretrained on 39 years of ERA5 and fine-tuned specifically for short-range prediction. As reported, this model requires extensive compute for training (tens of GPU-days), but serves as a high-skill reference and the source of intermediate features for distillation.

Student model. We construct a compressed 6-block student network, preserving FuXi's input and output structure but reducing depth from 48 to 6 Swin Transformer V2 blocks of total 640M parameters (4-block with 470M and 2-block with 300M parameters, respectively). The channel dimensions, embedding layers, and positional encodings are kept identical to the teacher. This allows the student to operate as a drop-in replacement while being substantially smaller and faster to train, while remaining architecturally compatible with FuXi.

Training strategies. We evaluate two approaches for training the student (Fig.1):

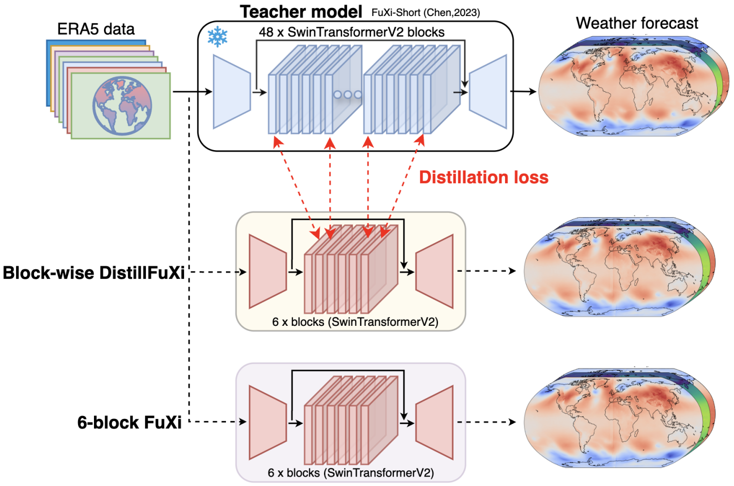
ERA5 data
Teacher model FuXi-Short (Chen,2023)
48 x SwinTransformerV2 blocks
Weather forecast
Distillation loss
Block-wise DistillFuXi
6 x blocks (SwinTransformerV2)
6-block FuXi
6 x blocks (SwinTransformerV2)

Figure 1: Overview of the training pipeline. ERA5 reanalysis fields are fed to three models: the pretrained FuXi-short teacher (48 SwinTransformerV2 blocks, frozen), a compressed 6-block FuXi student trained from scratch (optionally warm-started by a one-step regression), and a compressed 6-block student trained with block-wise distillation. Importantly, the distillation student receives intermediate-feature supervision from the teacher (student blocks are aligned to every 8th teacher block) via an MSE distillation loss, while both students are trained with an output forecast MSE. All models produce single-step weather forecasts.

(i) Baseline training from scratch, where the student is initialized randomly, except for the input projection layer, which is fitted via one-step linear regression to match the teacher's immediate forecast. The model is then trained end-to-end on ERA5 with a latitude-weighted MSE loss on forecast targets.

(ii) Block-wise distillation. In this setting, the student is trained not only on the forecast target but also to mimic the teacher's intermediate representations. Specifically, for 6-block-wise DistillFuxi version we align each of the 6 student blocks with every 8th teacher block [1,2,3,4,5,6] ↔ [8,16,24,32,40,48], where for each aligned pair, we minimize the MSE between the student and teacher block outputs after layer normalization. Similar approach is used for 4- and 2-block-wise students. The overall loss is a weighted sum $\mathcal{L} = \mathcal{L}_{forecast} + \lambda \mathcal{L}_{distill}$, where λ is dynamic weight gradually changing with training. The forecast loss is defined as $\mathcal{L}_{forecast} = \frac{1}{B\times H\times W\times C}\sum_{b,h,w,c}^{B,H,W,C} a_h w_c(\hat{Y}^{b,h,w,c} - Y^{b,h,w,c})^2$, where B is the batch size, H, W are latitude and longitude grid points, and C represents weather variables with latitude weights a_h and channel weights w_c. The distillation loss is defined as $\mathcal{L}_{distill} = \frac{1}{B\times H_{emb}\times W_{emb}\times C_{emb}}\sum_{l,b,h,w,c}^{L,B,H_{emb},W_{emb},C_{emb}} w_l(\hat{Y}_{emb}^{b,h,w,c} - Y_{emb}^{b,h,w,c})^2$, where L is the number of aligned block pairs, $H_{emb}, W_{emb}, C_{emb}$ are the dimensions of intermediate embeddings in the teacher model, and w_l are block-specific weights set to $[1.0, 0.8, 0.6, 0.4, 1.0, 1.0]$.

Training setup. We optimize all models with AdamW using a learning rate $5e-6$. The distillation loss weight decays as $\lambda(k) = (1e-2)\cdot 0.995^k$, where k is the iteration index. Unless specified, hyperparameters match those of FuXi to ensure comparability.

Evaluation. We evaluate the student and teacher models on held-out ERA5 data. Metrics include root mean square error (RMSE) and mean absolute error (MAE). The FuXi-short teacher serves as the skill reference.

3 Results and discussion

3.1 Forecast skill of distilled models

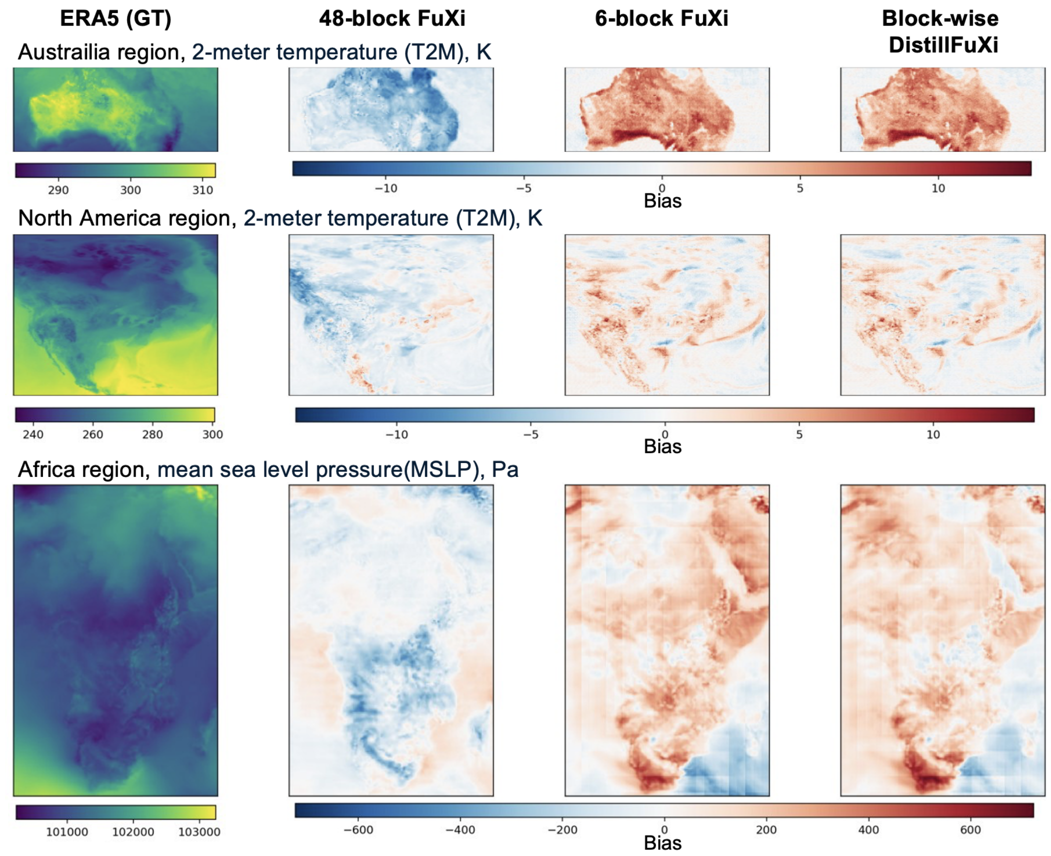

Figure 2: Spatial bias patterns in single-step forecasts on January 1, 2021 across three different geographic regions (rows). The first column shows ERA5 ground-truth fields, while the next three columns depict the forecast bias (prediction minus ERA5) for the 48-block FuXi, the 6-block FuXi, and the block-wise 6-block DistillFuXi. Negative and positive values indicate under- and overestimation relative to ERA5, respectively.

We begin by quantifying single-step forecast skill of the compressed students relative to the full teacher. Table 1 compares single-step RMSE and MAE (averaged over January 2021) for the full 48-block FuXi teacher, the 6-block forecast-only student (essentially it is a reduced 6-block FuXi), and the 6-block block-wise student (Block-wise DistillFuXi). As expected, the teacher attains the lowest errors, whereas DistillFuXi shows notable degradation. Importantly, adding block-wise distillation reduces this gap for every one of the total 70 evaluated parameters, including key fields such as MSLP and T2M in both RMSE and MAE metrics. However, relative to variability of each field, the remaining differences are still small. For example accuracy on thermodynamic fields (T2M, MSLP) degrades only by 0.9% and 1.9% of their STD, respectively, but larger for wind components. Overall, these results demonstrate that block-wise distillation effectively transfers intermediate feature knowledge from the teacher, substantially improving accuracy relative to a simple forecast-only student without increasing model depth.

Table 1: Comparison of single-step prediction performance between the full 48-block FuXi model and a reduced 6-block FuXi, and its block-wise 6-block DistillFuXi variant. We report root mean squared error and mean absolute error averaged over Jan 2021 on a set of key weather parameters. For each parameter, the mean and standard deviation of the target values are also shown, providing a reference scale to interpret relative error magnitudes. Lower RMSE/MAE indicates better accuracy.

Model/Parameter	T2M	MSLP	T850	Z850	Z500	U1000	V1000
Parameter mean	278.21	100958.7	274.36	13739.4	54080.3	−0.034	0.186
STD	21.432	1328.67	15.709	1470.77	3365.26	6.031	5.208
				RMSE			
Full 48-block FuXi	1.867	73.124	0.899	49.794	52.994	0.844	0.972
6-block FuXi	2.089	108.372	1.154	71.337	86.233	1.928	2.215
Block-wise DistillFuXi	2.057	98.946	1.119	65.677	80.261	1.906	2.181
				MAE			
Full 48-block FuXi	0.920	42.113	0.499	29.436	30.999	0.468	0.538
6-block FuXi	0.965	60.818	0.645	40.932	48.223	1.004	1.139
Block-wise DistillFuXi	0.958	55.677	0.627	37.705	44.555	1.000	1.114

To examine where the student models lose skill and to diagnose spatial patterns behind the numeric differences, we visualize bias maps. In Figure 2 we show ERA5 truth and model bias maps (model - ERA5) for 2-m temperature over Australia and North America and MSLP over Africa. Two clear patterns are emerging here. First, the teacher tends to underestimate temperature while both students tend to overestimate it, and second, for MSLP the teacher shows an ocean/land sign transition that the students do not reproduce. Importantly, student biases concentrate at sharp spatial gradients, especially sea–land transition zones, whereas the teacher's errors are smoother and more spatially coherent. We attribute this behavior to reduced capacity and receptive-field depth in the students in addition to feature-level distillation. The deep teacher captures multi-scale context and surface-coupling effects that produce smoother, state-dependent corrections, resulting in a more pronounced ocean/land transition and milder regression toward the mean. In contrast, shallow students, even when distilled, lack some high-frequency, long-range interactions, so residual errors tend to localize and can flip sign (over/underestimate) near strong gradients. Block-wise MSE distillation also tends to transfer lower-frequency features more readily than fine-scale corrections, which can lead to gradient-localized bias. We expect that enhancing high-frequency or surface-conditioned losses, adding projection heads or multi-scale distillation, or targeting anchors near gradient regions to recover transition-zone structure while retaining the training-speed can strongly improve the compressed models.

3.2 Robustness and generalization

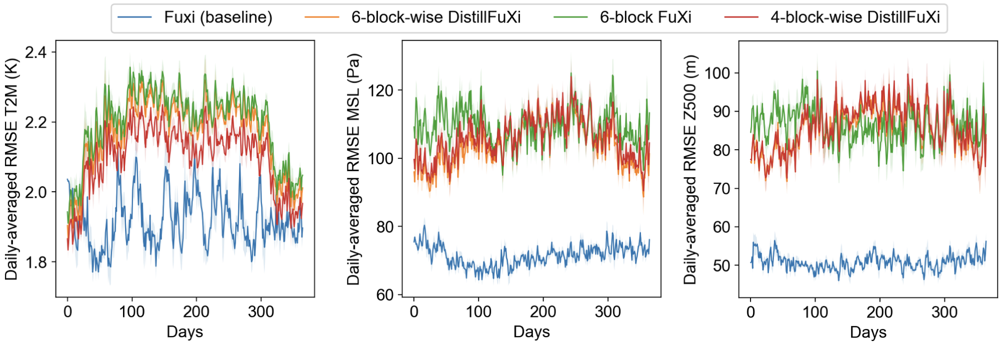

Figure 3: Daily-averaged RMSE over a one-year evaluation period for four models. Each panel reports a different weather parameter. Lower RMSE indicates more accurate forecasts.

In Figure 3, we show the day-to-day evolution of single-step forecast errors over the 2020 test year, providing insight into temporal variability and model stability beyond aggregate RMSE values. We also include a 4-block distillation variant alongside the 6-block forecast-only and 6-block block-wise students to probe how anchor count and placement affect stability. Although all student models preserve mean single-step skill close to the teacher, they start to show larger daily RMSE variability being noticeably more volatile and ”juggling” from day to day rather than smoothly following the teacher. This pattern suggests increasing capacity and representation mismatch. A shallow student cannot fully reproduce the teacher’s richer features, and multiple intermediate-matching terms can introduce competing gradients or transfer fine-grained teacher components that are unhelpful for a single-step objective. In addition, these effects are amplified in weather forecasting because atmospheric fields span many scales and include strong seasonal and surface-coupled signals. MSLP and Z500 variables, dominated by large-scale synoptic structure, show little seasonal RMSE modulation, while T2M is sensitive to surface processes and exhibits a mid-year RMSE rise for the students. These observations point us to practical high-leverage strategies that can be useful, including better-placed anchors, lightweight projection heads to align feature spaces, more tolerant feature losses, and tuned distillation schedules.

3.3 Ablation study

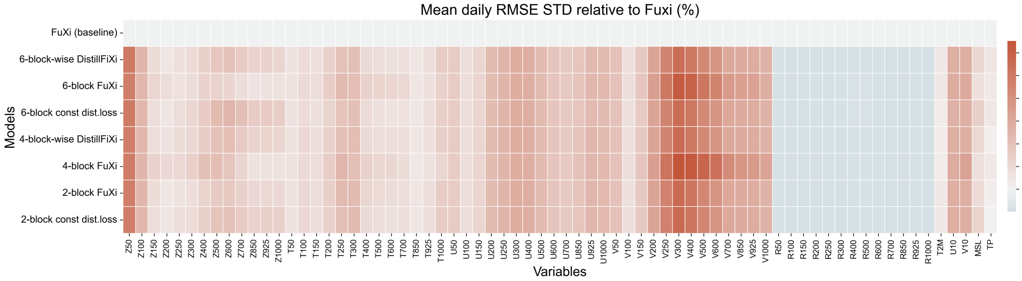

Figure 4: Relative variability of student models compared to teacher FuXi, reported as the percentage ratio of their standard deviations. Positive values indicate higher variability than FuXi, while negative values correspond to smoother forecasts. Rows correspond to different models, and columns are weather parameters.

We perform a compact ablation over seven different student variants: 6-, 4-, and 2-block compressed FuXi (no distillation), plus 6- and 2-block DistillFuXi at constant distillation weight ($\lambda = 5e - 5$), and a 6-, 4-block-wise DistillFuXi, and compare each model to the FuXi teacher using the percentage ratio of their per-variable standard deviations (Figure 4). The heatmap shows a clear pattern, where block-wise distillation consistently reduces short-term variability relative to undistilled students, and this stabilizing effect becomes more pronounced as student depth decreases (2-block distilled models are the smoothest in many variables). In contrast, simple model compression without distillation often produces an unbalanced profile, improving stability for large-scale geopotential (Z variables) fields while increasing variability on small-scale-sensitive fields such as the wind components (V variables). However, a few isolated variables (R variables) show unexpected negative ratios (student smoother than teacher), probably related to preprocessing subtleties that we will investigate further. Overall, the ablation supports our findings that block-wise distillation acts like a targeted regularizer that guides low-capacity students toward the teacher’s useful representations, improving temporal stability without sacrificing the efficiency gains of compression.

4 Conclusion and future directions

In this work we showed that a large FuXi-short weather forecasting transformer model can be aggressively compressed from 48 SwinTransformerV2 blocks down to a handful, while retaining most single-step forecast skill on key variables. Block-wise distillation, in particular, consistently narrows the gap to the teacher compared with a simply smaller student and acts as an effective, lightweight regularizer for low-capacity models. At the same time, the experiments reveal a consistent caveat that compression can increase short-term volatility in errors for some variables, especially those tied to small-scale or surface-coupled processes. This trade-off between preserving mean skill and reducing temporal stability in a few diagnostics highlights the limits of architectural compression for climate-scale forecasting. Several directions follow naturally from this work. First, we will quantify actual training and inference speed-ups (wall-clock, GPU-hours and energy) to translate parameter reductions into operational cost estimates. Second, we plan to extend student models to recursive multi-step forecasting and evaluate whether distillation benefits propagate or degrade under autoregressive rollout. Third, we will refine the distillation recipe by optimizing anchor placement and per-block weights, by adding lightweight projection heads or tolerant feature losses, and testing hybrid compression, for example pruning/quantization plus distillation. Fourth, we will probe robustness under distribution shift for different years, extreme events, and climate-shift scenarios, and investigate uncertainty quantification and ensembling as mitigations for day-to-day volatility. Finally, by measuring compute and carbon footprints alongside accuracy, we aim to provide practical guidance for deploying compact weather forecasting models in settings that require rapid retraining or limited resources, for example operational centers, regional modeling, and on-premises research clusters.

Acknowledgments and Disclosure of Funding

The work of D. Sukhorukov was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-15-2025-013, dated June 20, 2025, subsidy identifier 000000C313925P4B0002).

References

[1] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

[2] Sebastian Scher. Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, 45(22):12–616, 2018.

[3] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, December 2023.

[4] Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana C. A. Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, Zied Ben Bouallègue, Ana Prieto Nemesio, Peter D. Dueben, Andrew Brown, Florian Pappenberger, and Florence Rabier. Aifs – ecmwf's data-driven forecasting system, 2024.

[5] Simon Lang, Mihai Alexe, Mariana C. A. Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D. Dueben, Sara Hahner, Pedro Maciel, Ana Prieto-Nemesio, Cathal O'Brien, Florian Pinault, Jan Polster, Baudouin Raoult, Steffen Tietsche, and Martin Leutbecher. Aifs-crps: Ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score, 2024.

[6] Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. 2022.

[7] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nat.*, 619(7970):533–538, 2023.

[8] Tao Han, Song Guo, Fenghua Ling, Kang Chen, Junchao Gong, Jingjia Luo, Junxia Gu, Kan Dai, Wanli Ouyang, and Lei Bai. Fengwu-ghr: Learning the kilometer-scale medium-range global weather forecasting. 2024.

[9] Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, Kumar Ankur, Aman Gupta, Christopher E Phillips, Romeo Kienzler, Daniela Szwarcman, Vishal Gaur, Rajat Shinde, Rohit Lal, Arlindo Da Silva, Jorge Luis Guevara Diaz, Anne Jones, Simon Pfreundschuh, Amy Lin, Aditi Sheshadri, Udaysankar Nair, Valentine Anantharaj, Hendrik Hamann, Campbell Watson, Manil Maskey, Tsengdar J Lee, Juan Bernabe Moreno, and Rahul Ramachandran. Prithvi wxc: Foundation model for weather and climate. 2024.

[10] Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Vaughan, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. Aurora: A foundation model for the earth system, 2024.

[11] Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, July 2024.

[12] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, pages 1–7, 2024.

[13] Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):190, 2023.

[14] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

[15] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[16] Changlin Li, Jiefeng Peng, Liuchun Yuan, Guangrun Wang, Xiaodan Liang, Liang Lin, and Xiaojun Chang. Block-wisely supervised neural architecture search with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1989–1998, 2020.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 2015.

[18] Hui Wang, Hanbin Zhao, Xi Li, and Xu Tan. Progressive blockwise knowledge distillation for neural network acceleration. In *IJCAI*, pages 2769–2775, 2018.

[19] Cody Blakeney, Xiaomin Li, Yan Yan, and Ziliang Zong. Parallel blockwise knowledge distillation for deep neural network compression. *IEEE Transactions on Parallel and Distributed Systems*, 32(7):1765–1776, 2020.

[20] Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11), November 2020.

[21] Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models, 2024.