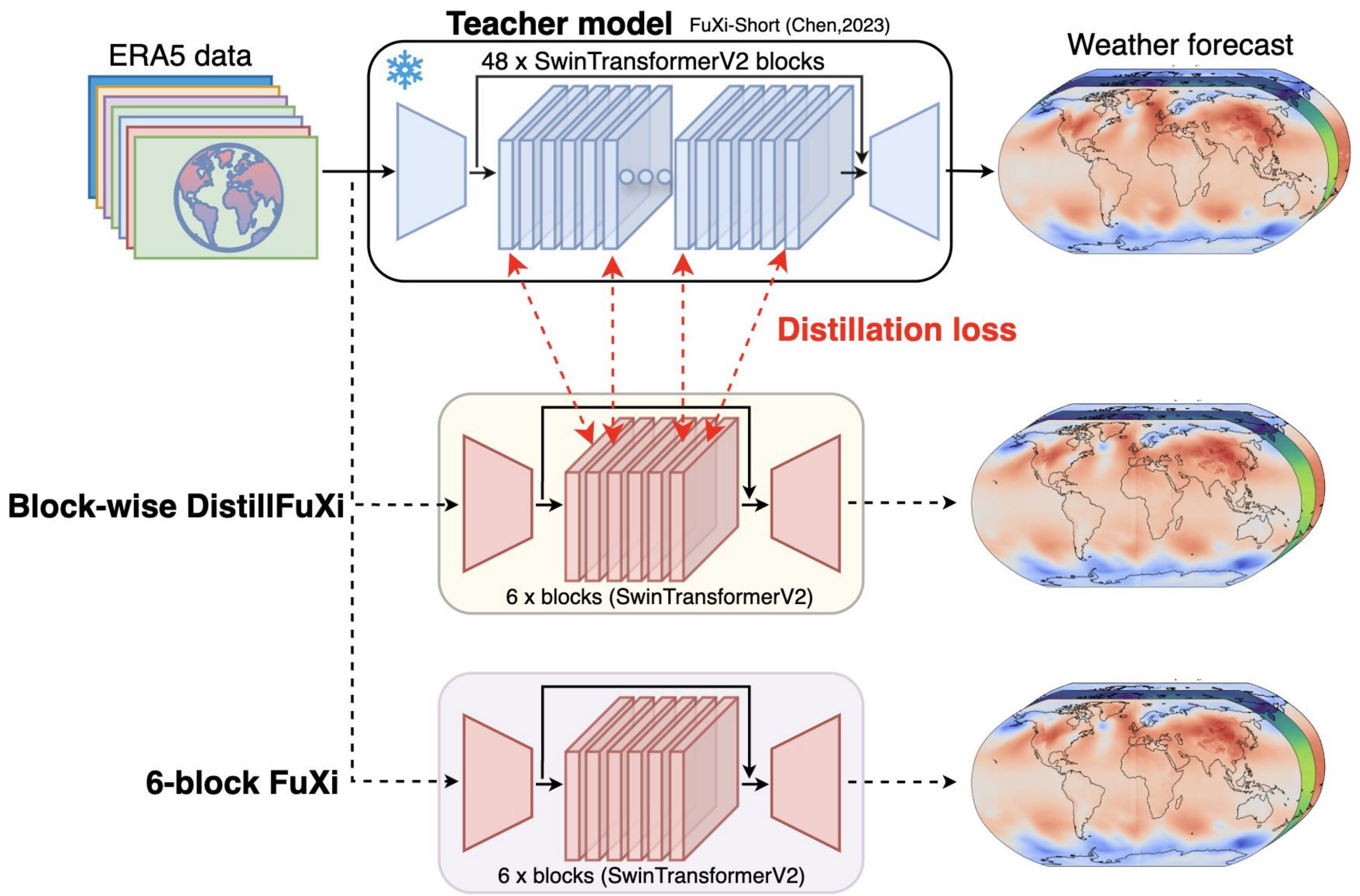
Block-wise distillation for lightweight weather models

Daniil Sukhorukov[1], Andrei Zakharov[1], Dmitry Zhevnenko[1], Vladimir Kirilin[4], Ekaterina Muravleva[2,3], Ivan Oseledets[1], Ilya Makarov[1]

[1]AIRI; [2]AI4Science, Sber; [3]Skolkovo Institute of Science and Technology; [4]Independent Researcher



Block-wise distillation pipeline: each student block mimics every 8th teacher block on ERA5 fields, combining feature and forecast losses.

Why lightweight weather models matter

Deep learning weather models such as FuXi, Aurora, or Prithvi-WxC achieve remarkable forecast accuracy but are extremely expensive to train and update.

Retraining large models on new or regional datasets is often infeasible due to GPU cost and energy use.

We therefore ask: *how far can we compress a global weather transformer without losing skill?*

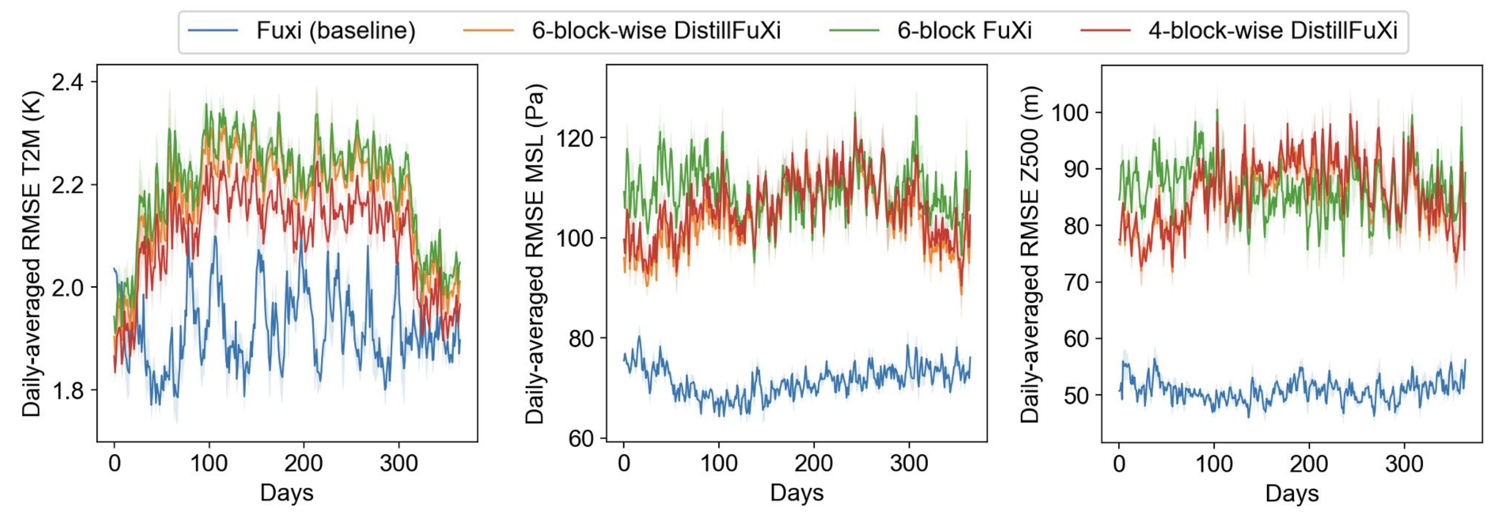
We compress **FuXi-short** (48 *SwinTransformerV2* blocks) into lightweight students with 6, 4, and 2 blocks, and study how **block-wise distillation** transfers knowledge from teacher to student.

How we compress FuXi

- Teacher: 48-block FuXi-short (SwinTransformerV2)
- Students: 6-, 4-, and 2-block versions with identical I/O
- Two training modes:
  - Baseline (from scratch with linear initialization)
  - **Block-wise distillation** — aligning every 8th teacher block using MSE loss

How much skill do we retain?

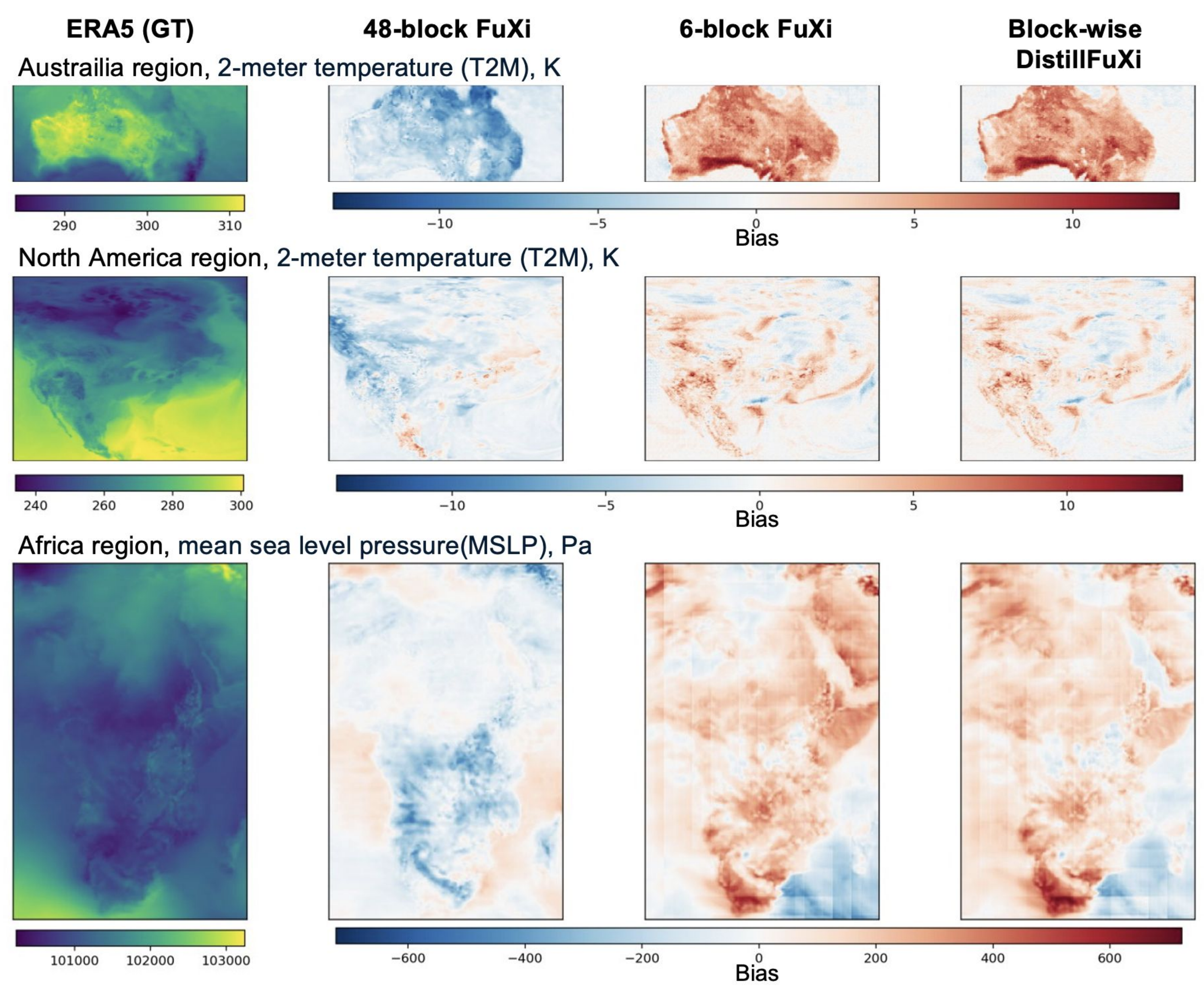
Block-wise distillation effectively transfers intermediate representations and narrows the accuracy gap to the full 48-block FuXi.
Mean degradation on thermodynamic fields (2 meter temperature, mean sea level pressure) remains under 2 % of variable STD.



Daily RMSE over the 2020 test year. DistillFuXi achieves the most stable accuracy among lightweight variants.

Where do lightweight models fail?

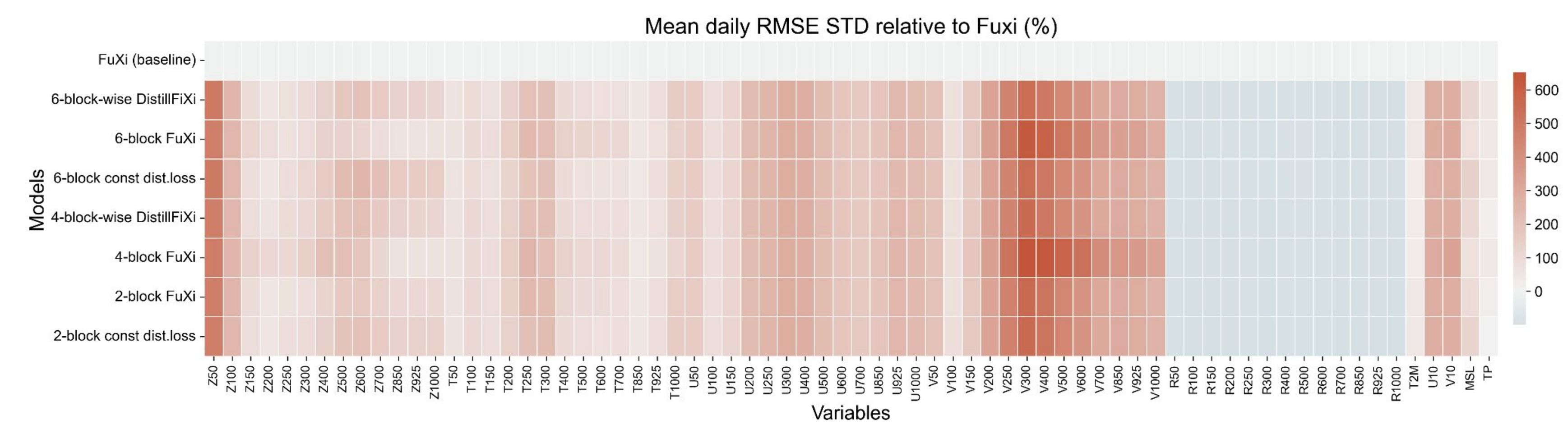
Bias maps highlight where compressed models diverge from ERA5 truth.
FuXi-short underestimates surface temperature, while small students tend to overestimate it.
DistillFuXi reproduces smoother, more coherent bias structures and partially restores ocean-land gradients.



ERA5, FuXi (48), 6-block FuXi, and Block-wise DistillFuXi — forecast bias for T2M and MSLP on January 1, 2021.

How stable are the distilled models?

Block-wise distillation acts as a regularizer for shallow students, reducing day-to-day RMSE volatility.
Without distillation, students exhibit larger temporal fluctuations, especially on small-scale surface variables.



Daily RMSE over the 2020 test year. DistillFuXi achieves the most stable accuracy among lightweight variants.

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

RMSE and MAE values averaged over January 2021. Lower is better.

What did we learn?

- FuXi can be compressed 8× (48 → 6 blocks) with minimal skill loss.
- Block-wise distillation successfully transfers teacher features and improves both accuracy and stability.
- Compressed models are **faster to train**, **cheaper to retrain**, and suitable for **regional adaptation**.
- Future work: multi-step forecasts, optimized block weights, and robustness under climate shifts.

References

- Chen, L. et al. (2023). *FuXi: A cascade machine learning forecasting system for 15-day global weather forecast.* npj Climate and Atmospheric Science, 6(1): 190.
- Wang, H. et al. (2018). *Progressive block-wise knowledge distillation for neural network acceleration.* IJCAI 2018.
- Schmude, J. et al. (2024). *Prithvi-WxC: Foundation model for weather and climate.* arXiv:2404.03972.
- Rasp, S. et al. (2024). *WeatherBench 2: A benchmark for the next generation of data-driven global weather models.* arXiv:2407.06494.