Scaling Laws of Global Weather Models

Paper Code

ICML International Conference On Machine Learning

Yuejiang Yu • Langwen Huang • Alexandru Calotoiu • Torsten Hoefler

Department of Computer Science, ETH Zürich, Zürich, Switzerland

10× data → up to 3.2× lower loss
strongest data-scaling gain

Wider is better for weather models
new findings of scaling laws

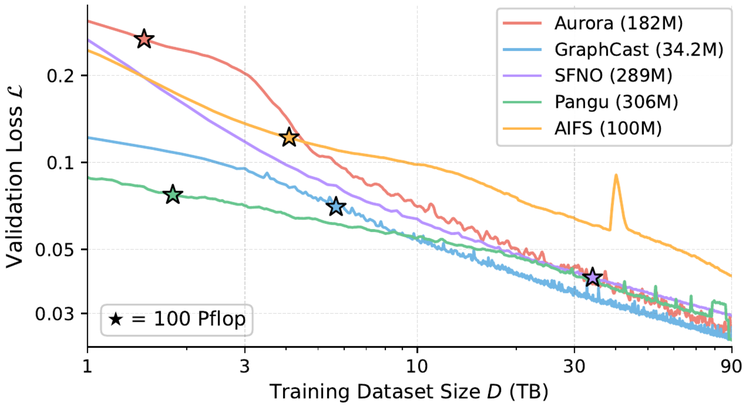
> 430,000 GPU hours
five models, one benchmark

36× higher H100 utilization
Aurora 37.2% vs. GraphCast 1.03%

1. What we scaled

- Models: Aurora, AIFS, Pangu, GraphCast, SFNO
- Scales: model size N, training data D, compute C
- Variants: model shape, validation loss, variable-level RMSE

$$\mathcal{L}(D) = \alpha D^{-\beta}, \quad \mathcal{L}(N) = \gamma N^{-\delta}, \quad \mathcal{L}(C) = \lambda C^{-\epsilon}$$

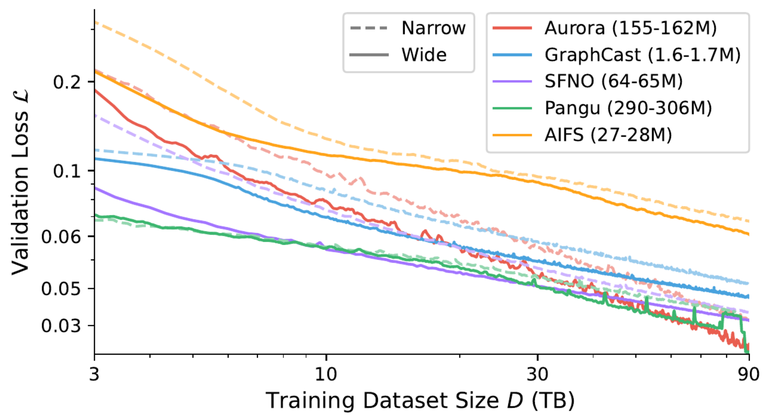


Unified benchmark

- ERA5 dataset with training: 1979–2020 and validation: 2021
- 0.25° global grid; 6-hour forecasts
- Unified weighted validation loss (MSE)
- Identical optimization and evaluation protocol
- Variable-specifis loss of 10U and 2T (RMSE)

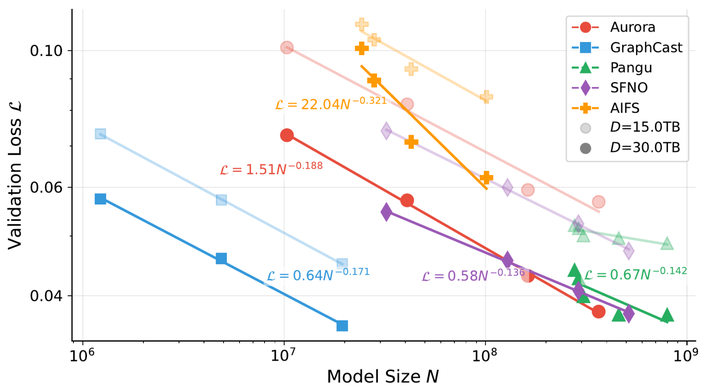
Data-scaling strength	β ↑
Aurora	0.51
AIFS	0.46
Pangu	0.43
GraphCast	0.36
SFNO	0.34

2. Width beats depth (unlike LLM)



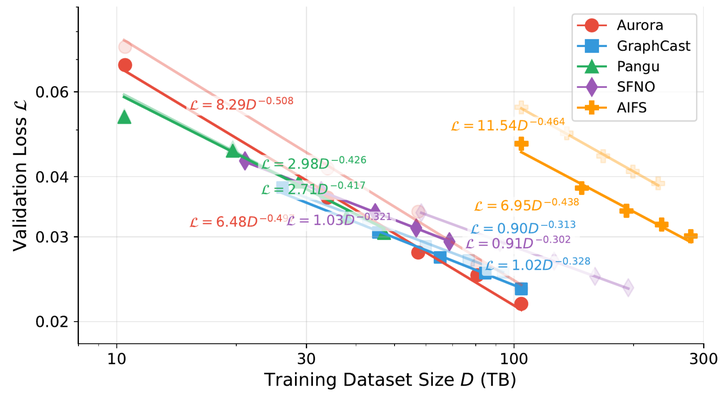
- Across all five models, wider models are better at same N
- Unlike in LLMs, where shape does not affect loss
- Reflects linearity in the latent space of weather data
- Representation power > nonlinear reasoning

3. Parameter scaling

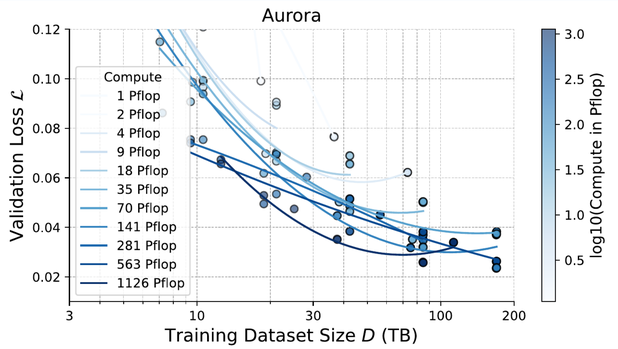


- GraphCast: low loss with fewer parameters
- Message passing: 1.03% H100 utilization (GNN)
- Aurora: 37.2% H100 utilization (Transformer)
- Doubling D shifts curves downward

4. Data scaling

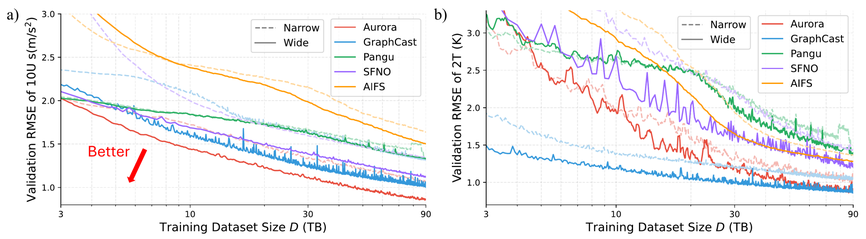


5. Compute-optimal training



- Fixed C reveals the N-D trade-off
- Larger budgets favor more data before larger N

6. Variables scale differently



We suggest future weather models should prioritize wider architectures and larger training datasets to maximize predictive performance.