BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY
Office of Science

AI Science Application: Large-scale Transformer-based Weather Prediction

Jared D. Willard

NERSC Postdoc in Data & AI Services Group

Peter Harrington[a], Shashank Subramanian[a], Ankur Mahesh[b], Travis A. O'Brien[c], William D. Collins[b,d]

[a]LBNL, NERSC
[b]UC Berkeley, Department of Earth and Planetary Science
[c]Indiana University,, Department of Earth and Atmospheric Sciences
[d]LBNL, EESA

NERSC National Energy Research Scientific Computing Center

NERSC Data Day 2024

Berkeley UNIVERSITY OF CALIFORNIA

INDIANA UNIVERSITY

Rise in data-driven weather forecasting

Availability of large, high-quality, open, and free meteorological datasets (e.g. ERA5 Reanalysis)

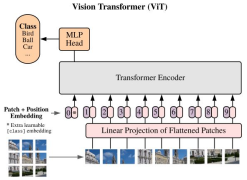
Copernicus
Europe's eyes on Earth

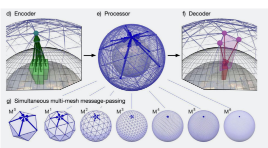
IMPLEMENTED BY ECMWF

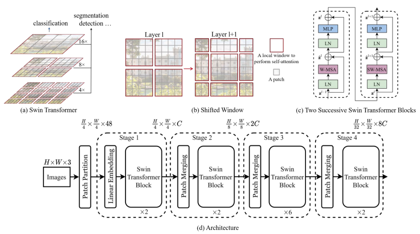
GPU-powered HPC

Advances in Deep Learning

E.g. Vision Transformers, GNN







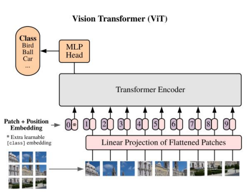
Rise in data-driven weather forecasting

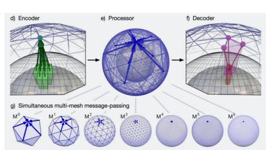
Availability of large, high-quality, open, and free meteorological datasets (e.g. ERA5 Reanalysis)

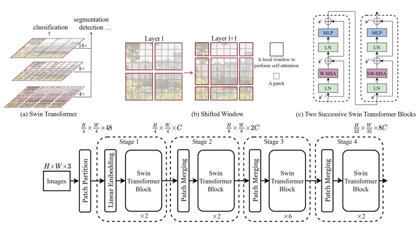
GPU-powered HPC

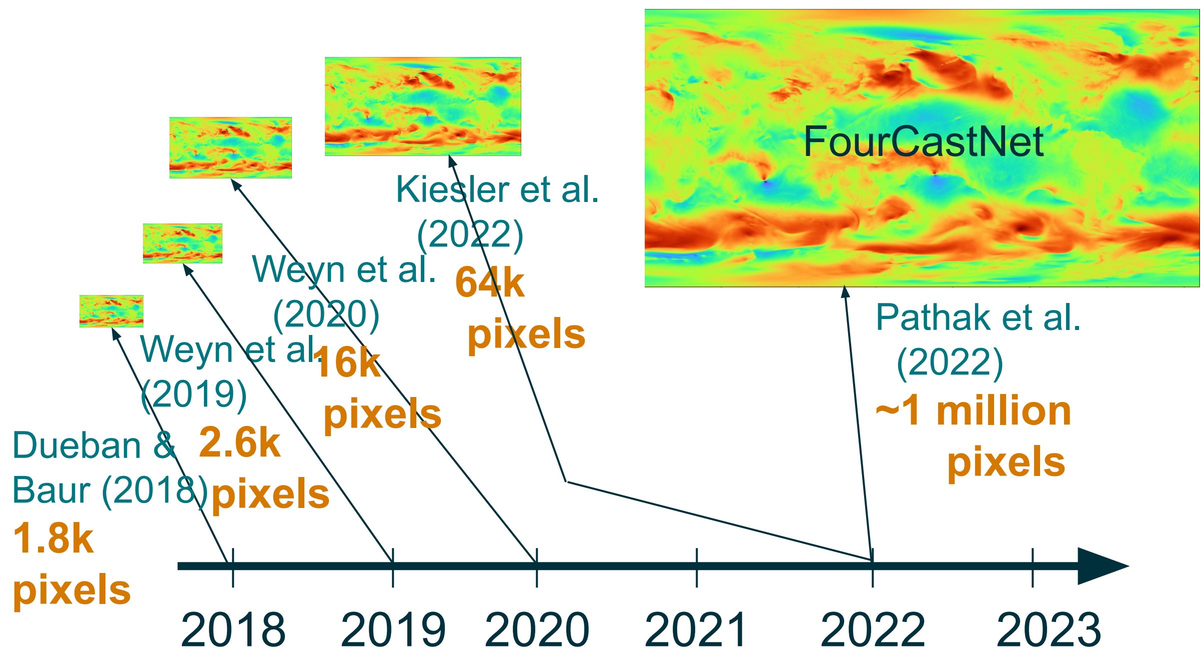
Advances in Deep Learning

E.g. Vision Transformers, GNN









Rise in data-driven weather forecasting

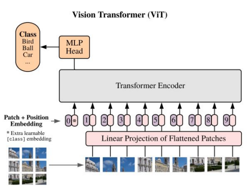
Availability of large, high-quality, open, and free meteorological datasets (e.g. ERA5 Reanalysis)

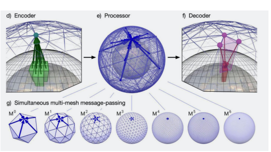
Copernicus Europe's eyes on Earth — IMPLEMENTED BY ECMWF

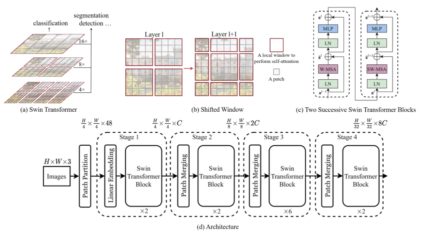
GPU-powered HPC

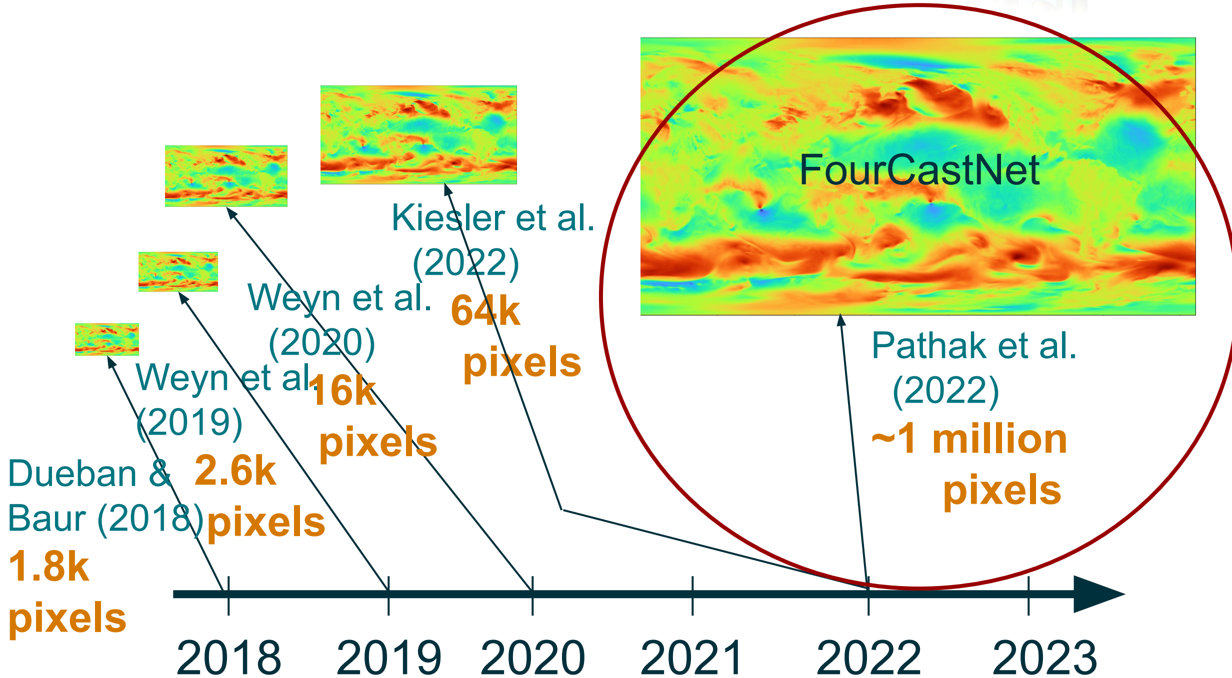
Advances in Deep Learning

E.g. Vision Transformers, GNN





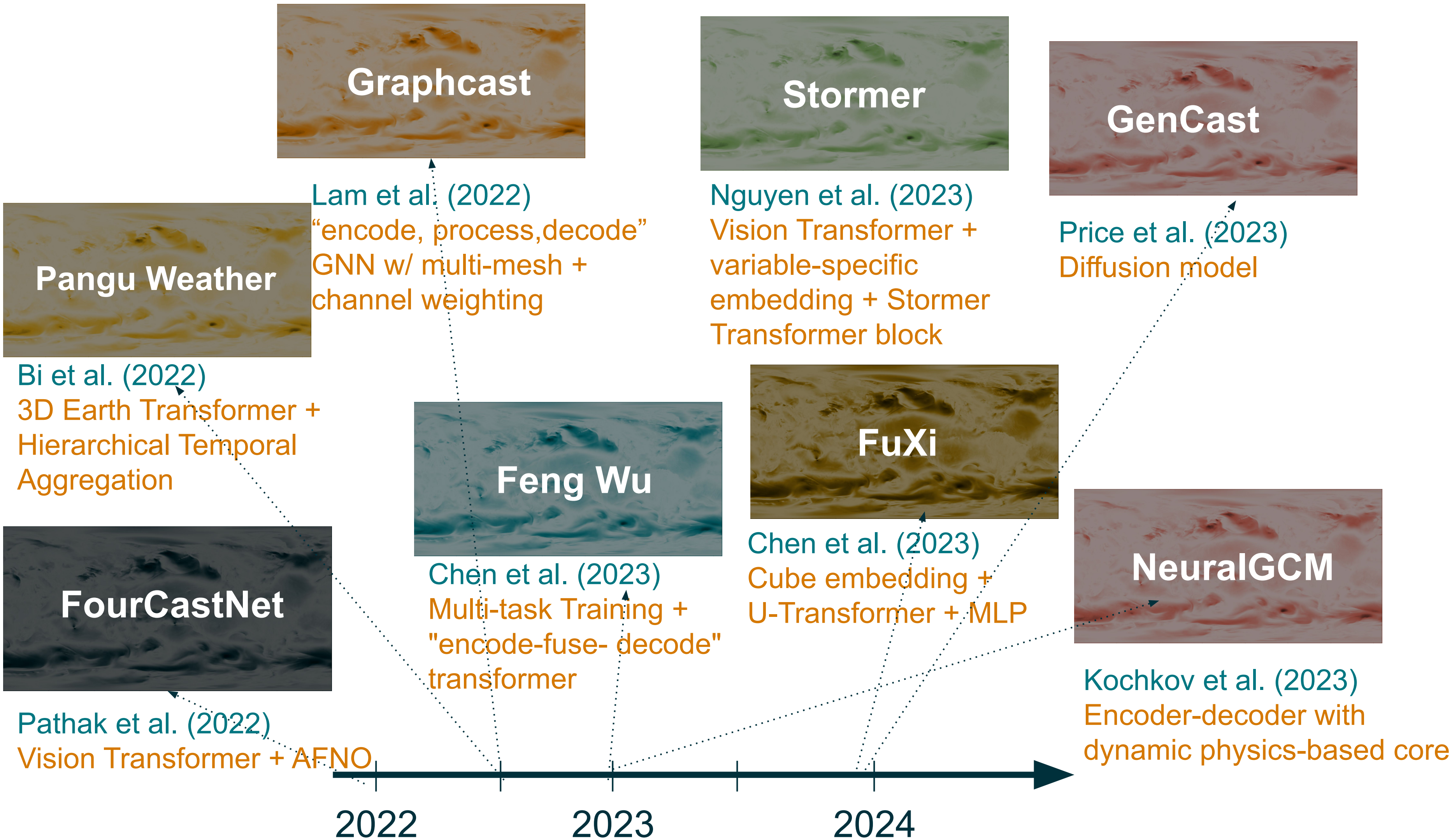




Dilemma: Tons of Models and Deep Learning Techniques

Pathak et al. (2022)
Vision Transformer + AFNO

Dilemma: Tons of Models and Deep Learning Techniques



Gap of studies comparing these different techniques

Need for ablation!

Types of Choices

Base Architecture

- Vision Transformer
- Graph Neural Network
- Swin Transformer

⋮

Architectural Modifications

- Spherical Fourier Neural Operator
- Weather-specific embedding layer
- Physics-based Core

⋮

Training Techinques

- Multi-step fine-tuning
- Multi-task loss function
- Randomized Forecast Interaval
- Hierarchical Temporal Aggregation

⋮

- **Initial Study**
  - Nguyen et al. (2023) attempt initial ablations, though at coarser resolution (~1.41°) using "Stormer" model

2023-12-8

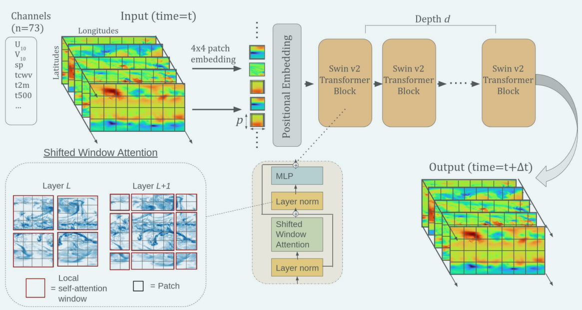
Scaling transformer neural networks for skillful and reliable medium-range weather forecasting

Tung Nguyen[1], Rohan Shah[1,2], Hritik Bansal[1], Troy Arcomano[3], Romit Maulik[3,4], Veerabhadra Kotamarthi[3], Ian Foster[3], Sandeep Madireddy[3], and Aditya Grover[1]
[1]UCLA, [2]CMU, [3]Argonne National Laboratory, [4]Penn State University

Research Objectives:

Showcase off-the-shelf model performance

Swin v2 Transformer



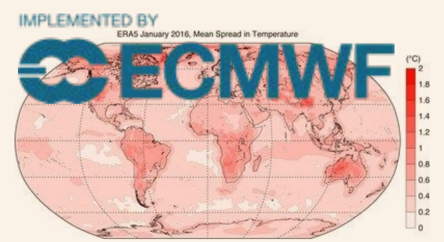
- Parameterized for moderate compute budget (0.5-2 days on 16 A100 GPU nodes)
- Showcase superior performance relative to ECMWF's Integrated Forecasting System (IFS)

Perform ablations using techniques from recently published literature:

Key Ablations

- Graphcast-inspired channel weighting and invariants
- Multi-step fine-tuning used in FourCastNet and Graphcast
- Variable tokenization and aggregation layer from the Stormer model study
- Multi-task learning loss function from the Feng Wu

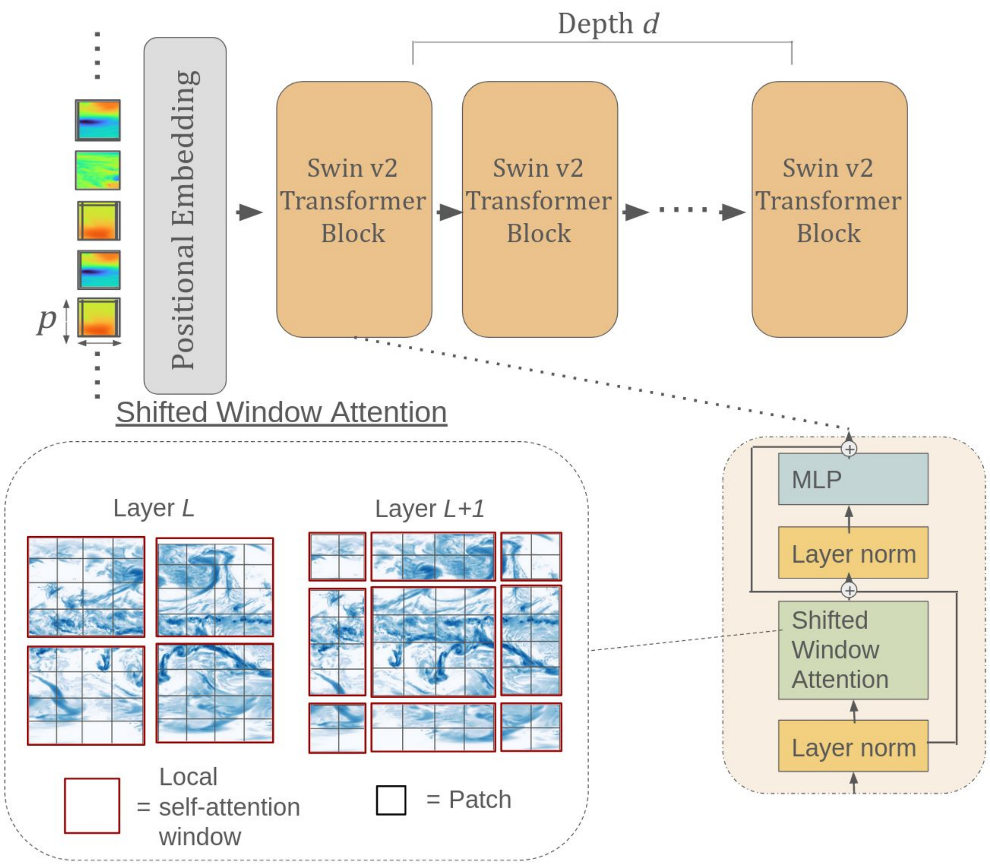
Using large-scale high resolution data (ERA5)



Baseline Model: Swin v2 Transformer

Using an off-the-shelf model for autoregressive prediction

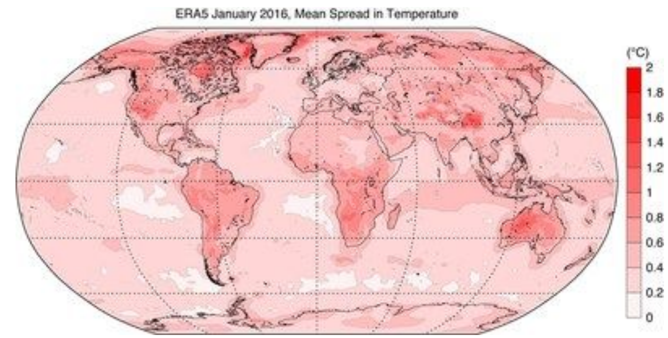
- Main idea
  - Shifting window partitioning scheme for computing self-attention
- Benefits
  - Scalability and efficiency being applied to high resolution prediction
  - Comparable performance to other more complicated or experimental architectures



Dataset: ERA5 Reanalysis

- **<u>ERA5 73 channel reanalysis dataset</u>**
  - Data from 1979 - 2018
    - Train: 1979-2015
    - Validate: 2016-2017
    - Test: 2018
  - 0.25° x 0.25° resolution
  - 6 hr timestep
  - Regridded to 2D field of shape (721 × 1440)

IMPLEMENTED BY ECMWF Climate Change Service

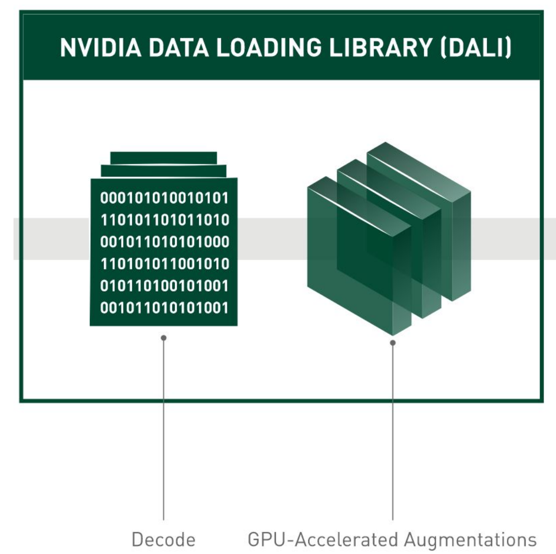


- **<u>Storage Details</u>**
  - HDF5 files for fast performance
  - ~20 Terabytes on Scratch

Computing Details

Efficient, Scalable

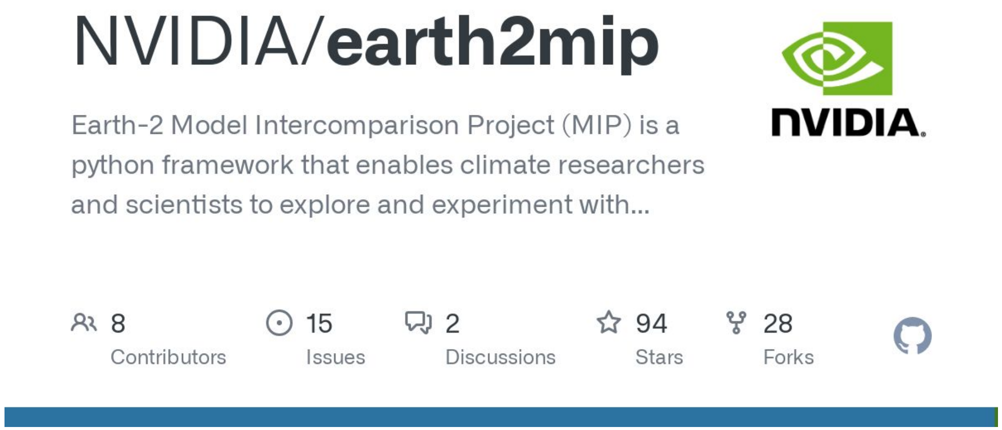
- Pytorch shifter container
- Trained on 64 GPUs using data parallelism
  - Training time **less than a day**
- DALI data loader for overlapped IO and compute
- Experiment tracking and visualization with Weights & Biases



developer.nvidia.com/dali

Evaluation: Earth-2 Model Intercomparison Project

- Python library from Nvidia
- Scores averaged over 11 initial lead times evenly spaced throughout 2018 and forecasts are rolled out 7 days at 6 hour intervals.

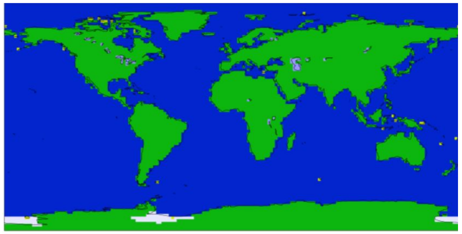


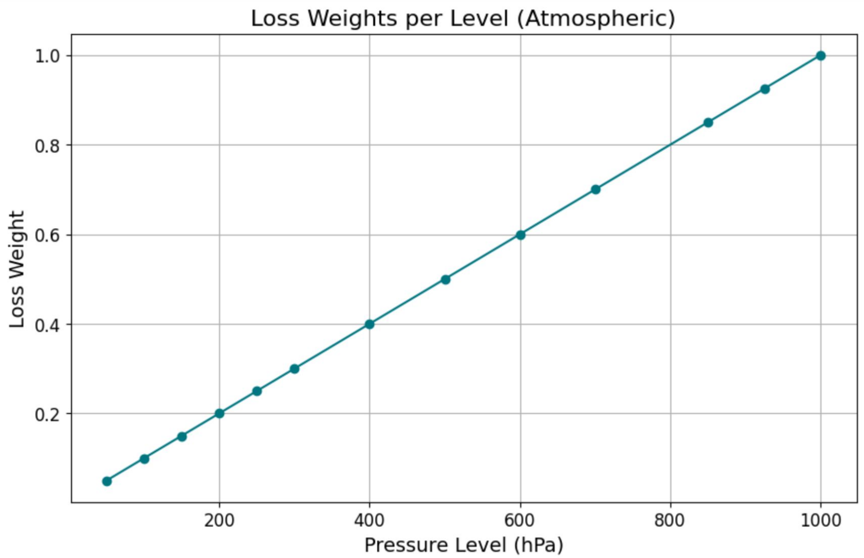
12

Ablation #1: Graphcast-inspired channel weighting + invariants

Surface importance and additional static information

- Includes 2 “static” inputs
  - 1. Orography (surface geopotential)
  - 2. Land-sea mask
- Weights prioritize weather variables closer to the surface
  - Also used in “Stormer” paper (Nguyen et al. (2022))





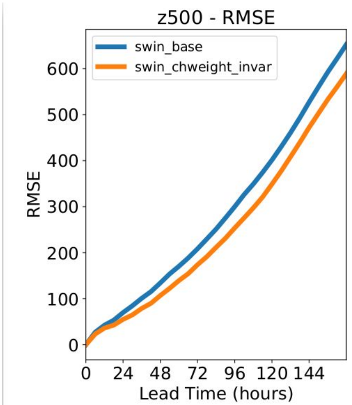
Ablation #1: Results for channel weighting + invariants

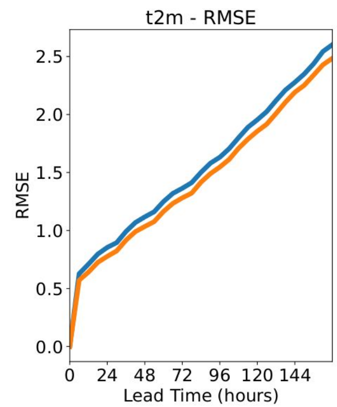
Forecast skill up to 7 days

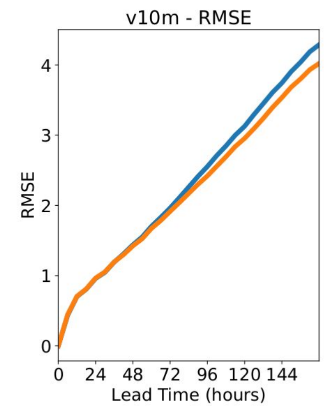
Baseline

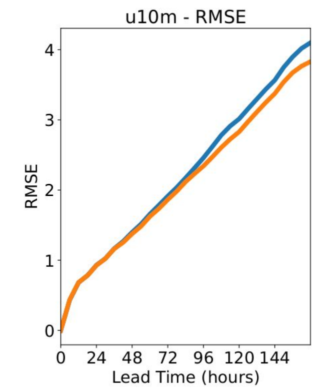
Channel-weighting and invariants

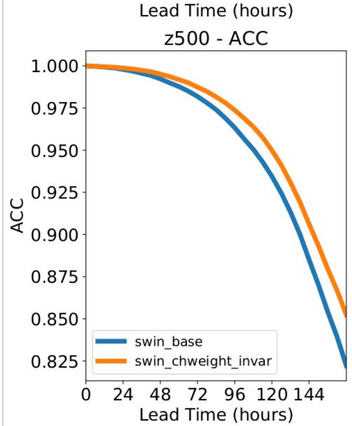
500 hPa Geopotential Height	2m Temperature	10m Northward Wind (v10m)	10m Eastward Wind (u10m)

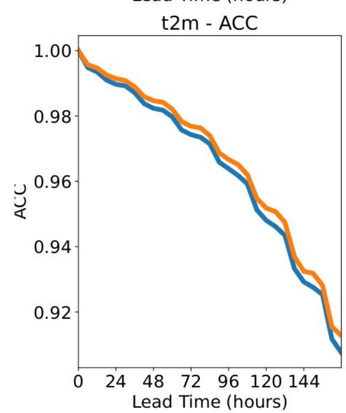


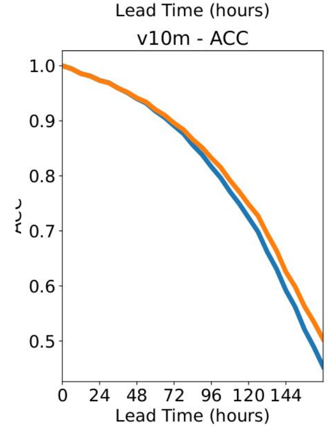


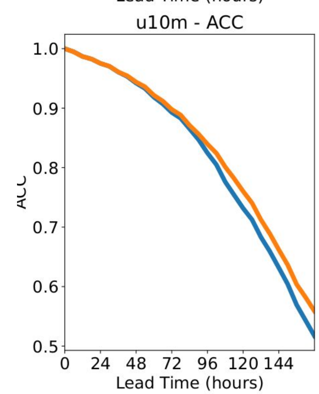












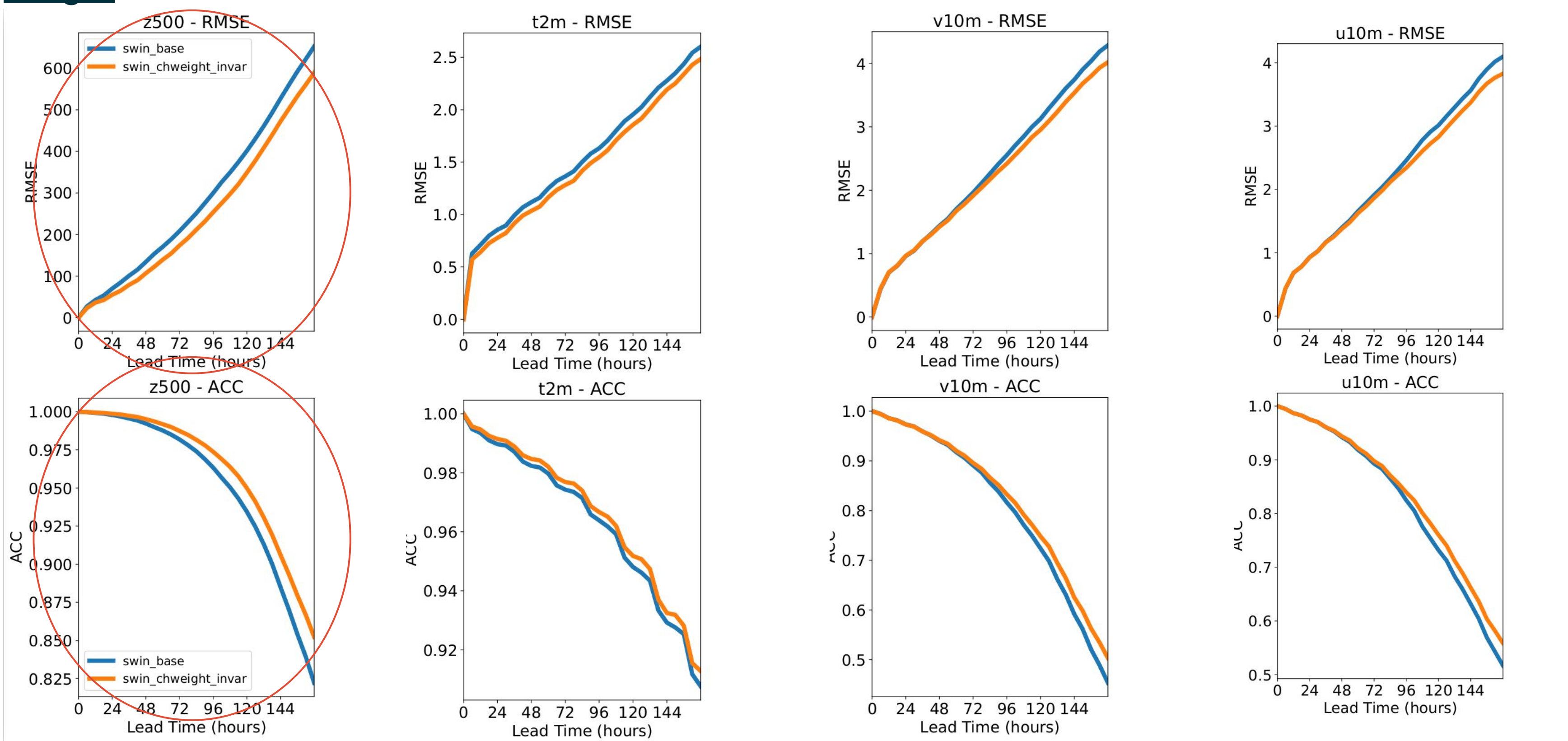
Ablation #1: Results for channel weighting + invariants

Forecast skill up to 7 days

Baseline

Channel-weighting and invariants

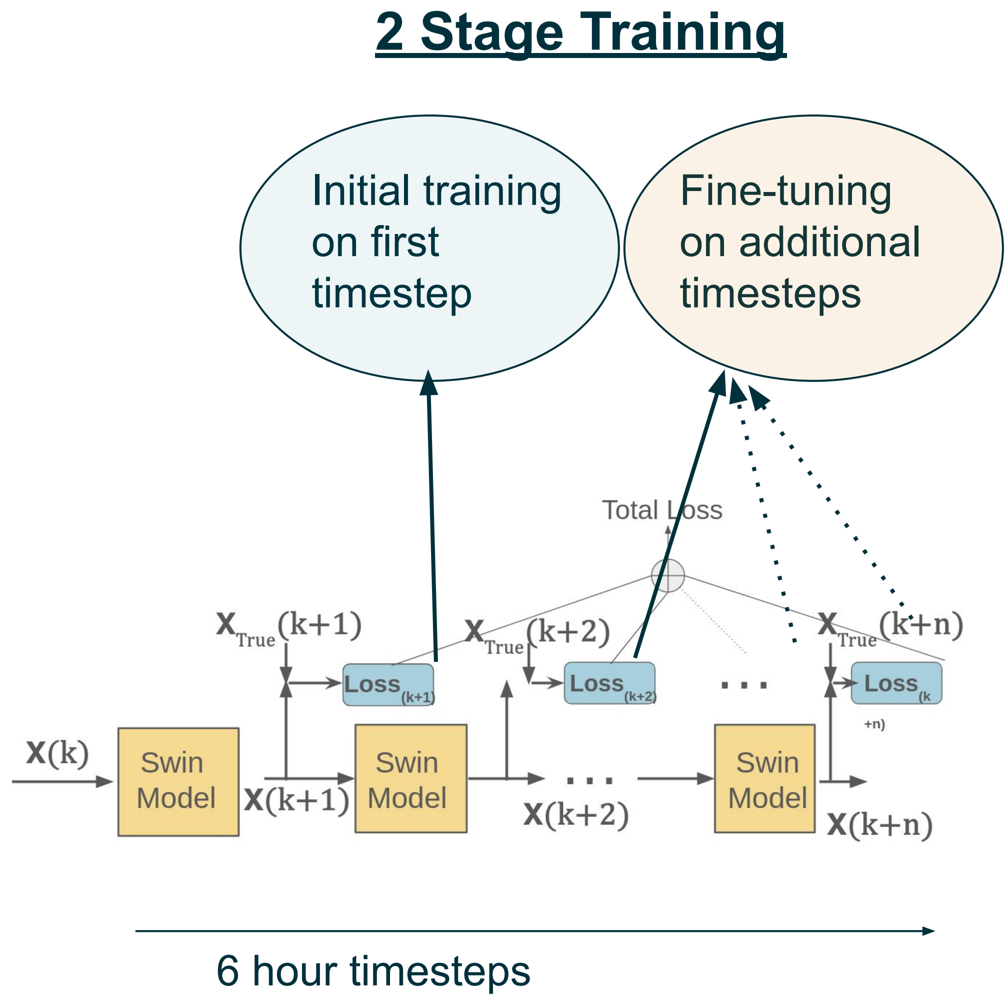
500 hPa Geopotential Height | 2m Temperature | 10m Northward Wind (v10m) | 10m Eastward Wind (u10m)



Ablation #2: Multi-step fine-tuning

Improved Performance and Rollout Stability

- Pioneered by FourCastNet and used in many newer models
- Addresses possible drawback of autoregressive models of rapid error accumulation



Results of multi-step fine tuning

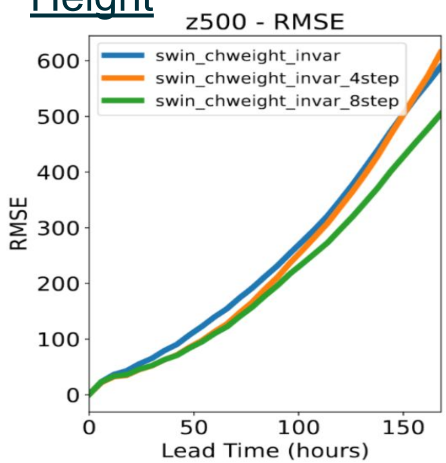
Forecast skill up to 7 days

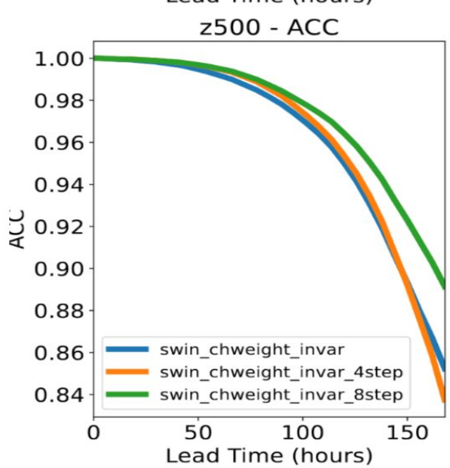
Baseline+Channel Weighting

4 step fine-tune

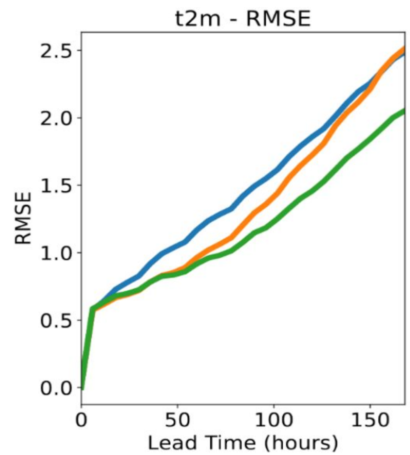
8 step fine-tune

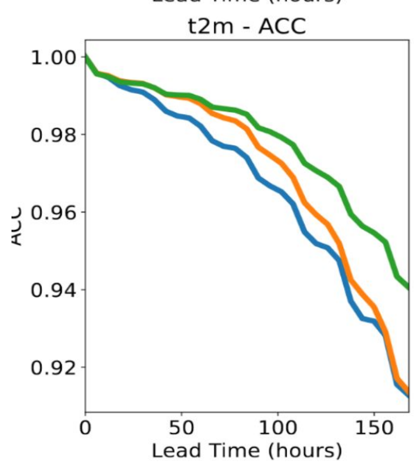
500 hPa Geopotential Height



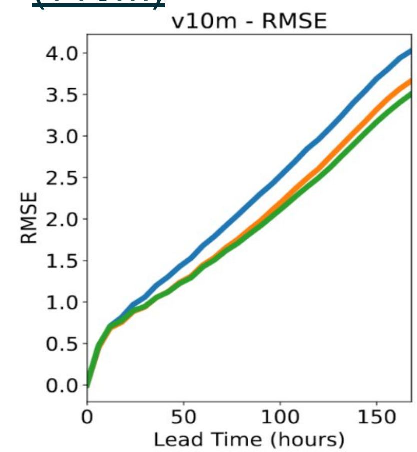


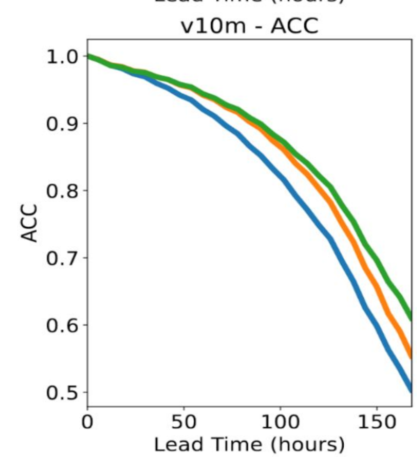
2m Temperature





10m Northward Wind (v10m)





Issues with multi-step fine-tuning

- Blurriness, poor fine-scale detail
- Especially problematic in ensemble context for numerical weather prediction (Brenowitz et al. (2024)

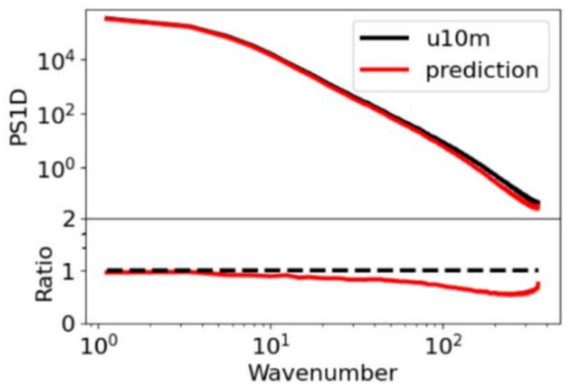
Issues with multi-step fine-tuning

- Blurriness, poor fine-scale detail
- Especially problematic in ensemble context for numerical weather prediction (Brenowitz et al. (2024))

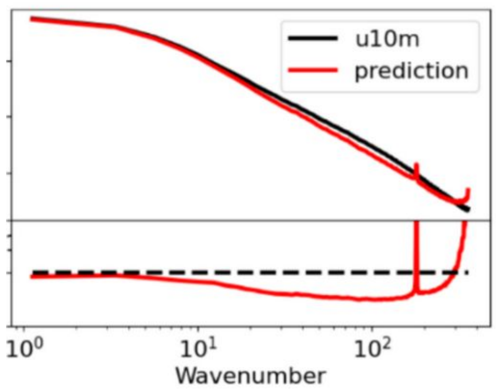
Do we see blurriness in our models?

Can examine spectra!

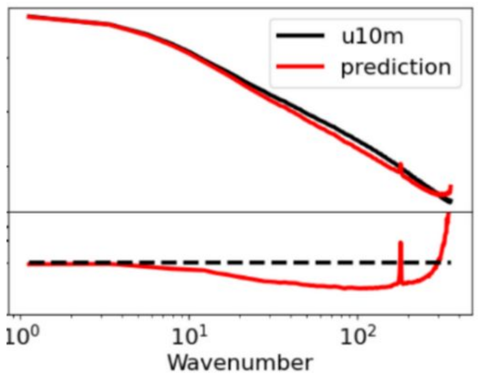
Base Model



4 step fine-tune

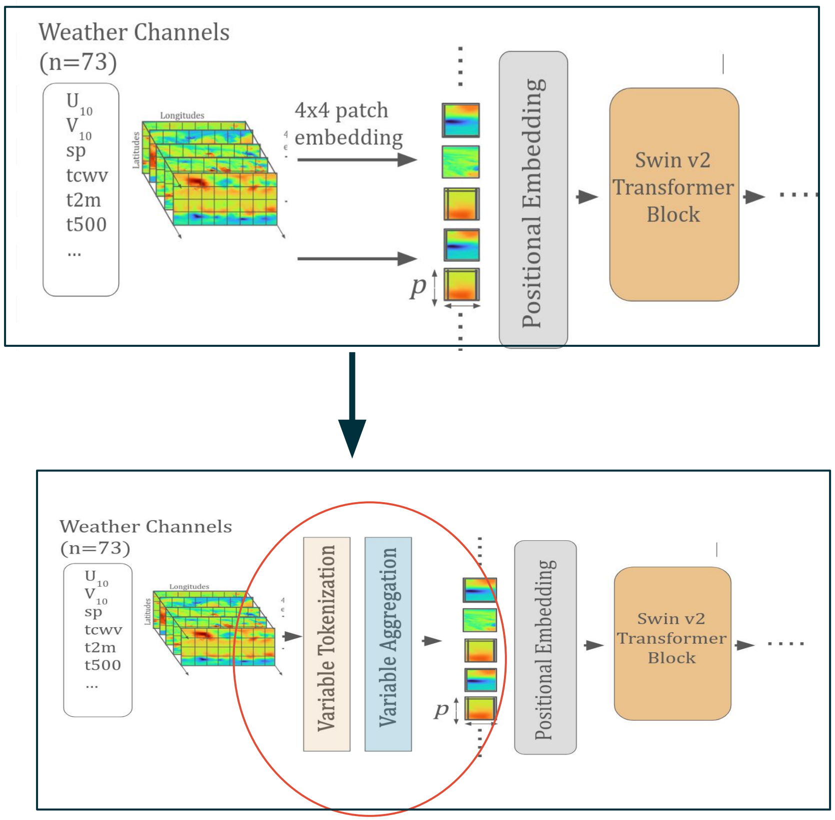


8 step fine-tune



Unsuccessful Ablations

Weather-specific embedding layer (Nguyen et al. (2023))



Effects:

- Minimal improvements at the cost of large computational expense and memory footprint
- Infeasible for high resolution, aggregation done sequentially due to high memory pressure from the cross-attention

Unsuccessful Ablations

Uncertainty-based Multi-task Loss Function (Chen et al. (2023))

New Prediction Capability

$$(\hat{\mu}^{k+1}, \hat{\sigma}^{k+1}) = \mathrm{Swin}(\mathcal{X}^{k})$$

Loss Function Change

Mean Squared Error (MSE) → Negative Log Likelihood

Effects:

- Consistently decreased performance in the validation alongside increased blurring.
- Continued even when model complexity was increased.

Comparing best models with IFS

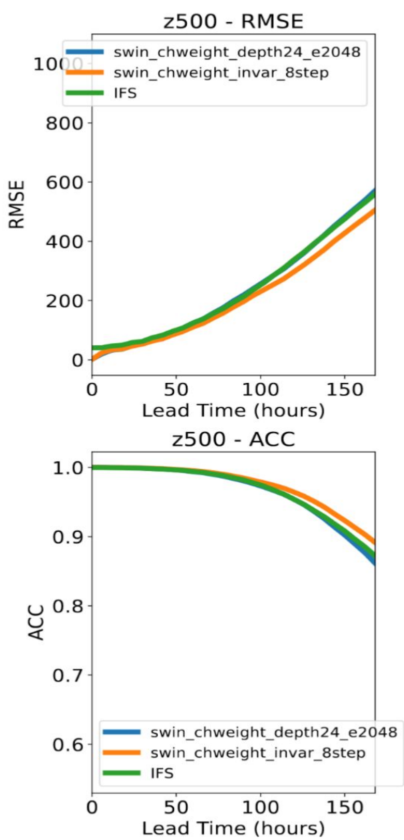
Forecast skill up to 7 days

Channel Weighting + High Complexity

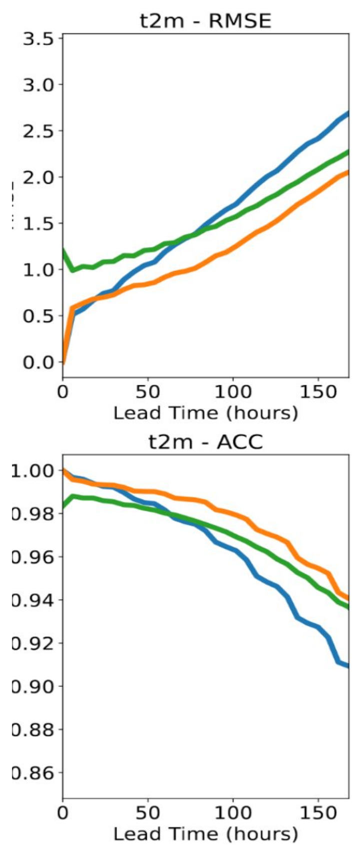
8 step fine-tune

IFS (Integrated Forecast System)

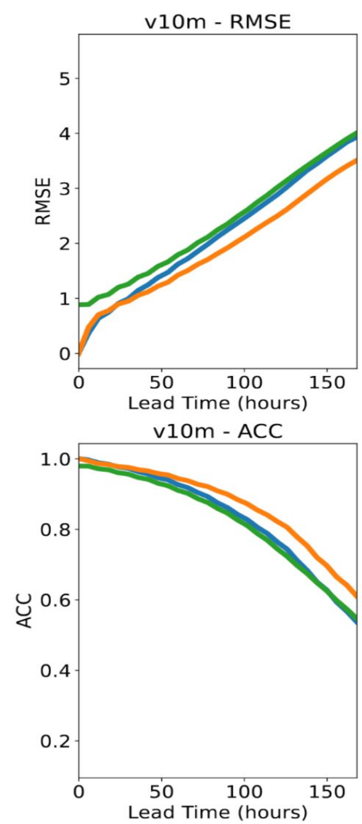
500 hPa Geopotential Height



2m Temperature



10m Northward Wind (v10m)



P

Comparing best models with IFS

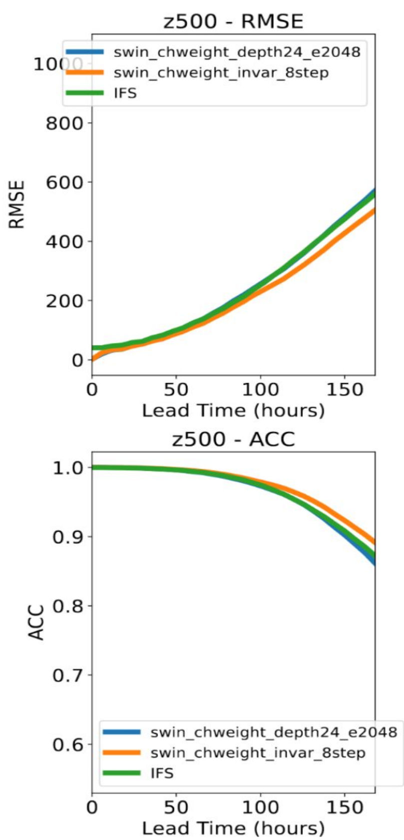
Forecast skill up to 7 days

Channel Weighting + High Complexity

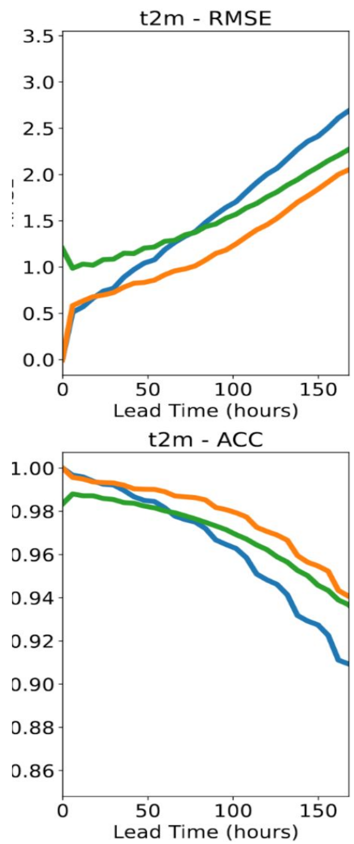
8 step fine-tune

IFS (Integrated Forecast System)

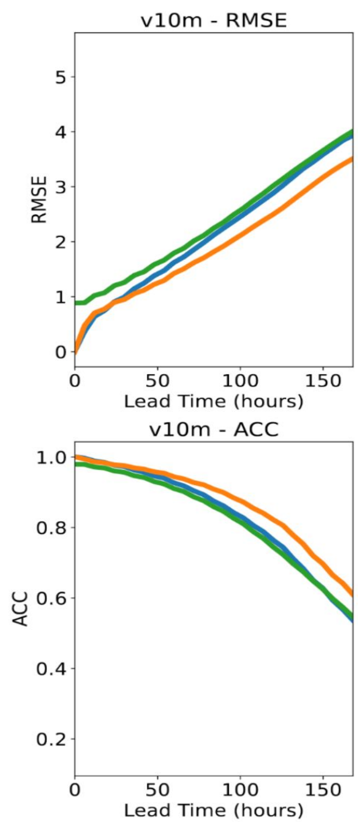
500 hPa Geopotential Height



2m Temperature



10m Northward Wind (v10m)



Takeaways

- 8 step fine-tuned model outperforms IFS at all lead times
- 1 step high complexity model outperforms only at shorter lead times

P

Conclusions

Off-the-shelf models do very well!

- We demonstrate that an off-the-shelf SwinV2 Transformer model can surpass the Integrated Forecasting System's (IFS) performance with minimal modification.
  - Training on < 1% of perlmutter for less than a day, ~10,000x speedup compared to operational numerical weather prediction.

- Of the ablations tested, the channel weighting was effective for single-step prediction compared to the baseline whereas the uncertainty loss and variable aggregation strategies did not help.

- Though multi-step fine tuning helps significantly in rollout RMSE, it still exhibits blurring effects in important fields like u10m/v10m for this resolution and model architecture. This shows the tradeoff between better RMSE and the loss of high frequency information.

Thank you!