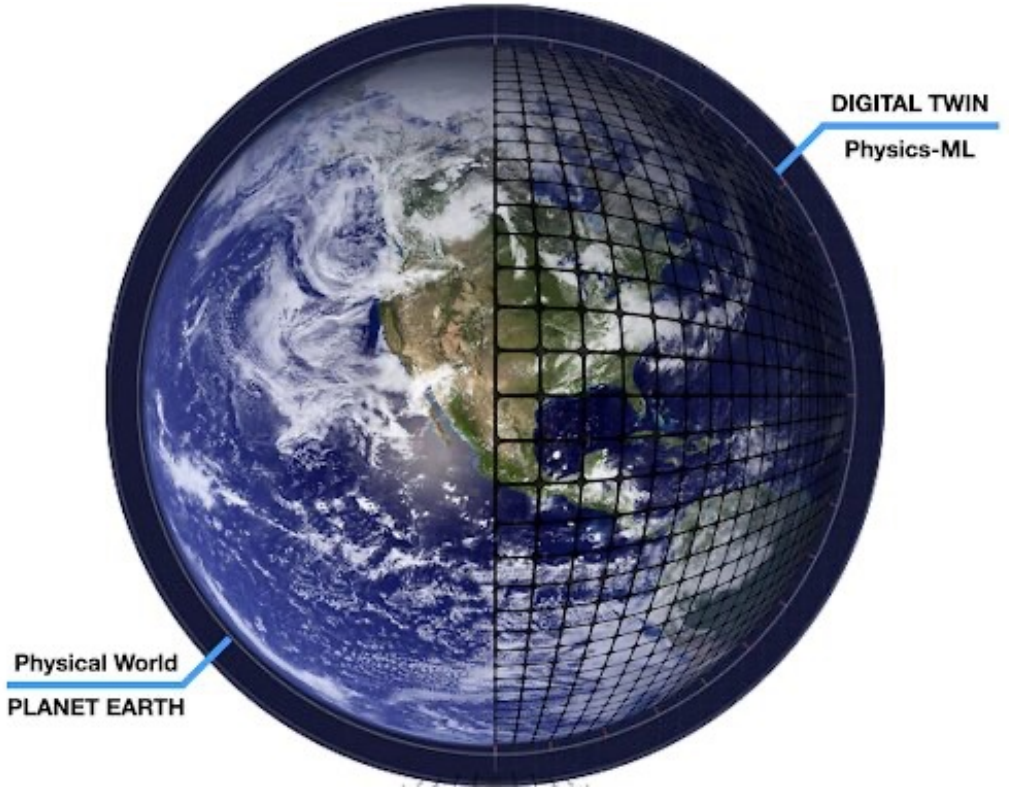
FOURCASTNET: DATA-DRIVEN HIGH-RESOLUTION ATMOSPHERIC MODELING AT SCALE



SHASHANK SUBRAMANIAN
LAWRENCE BERKELEY NATIONAL LAB
NERSC

Team

Jaideep P.
NVIDIA

Shashank S.
LBL

Peter H.
LBL

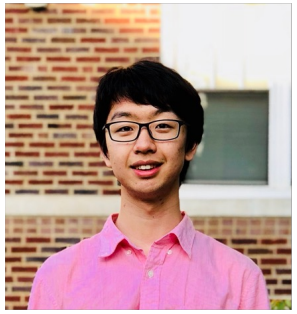
Sanjeev R.
U Michigan

Ashesh C.
Rice U.

Morteza M.
NVIDIA

Thorsten K.
NVIDIA

David H.
NVIDIA



Zongyi L.
Caltech

Kamyar A.
Purdue

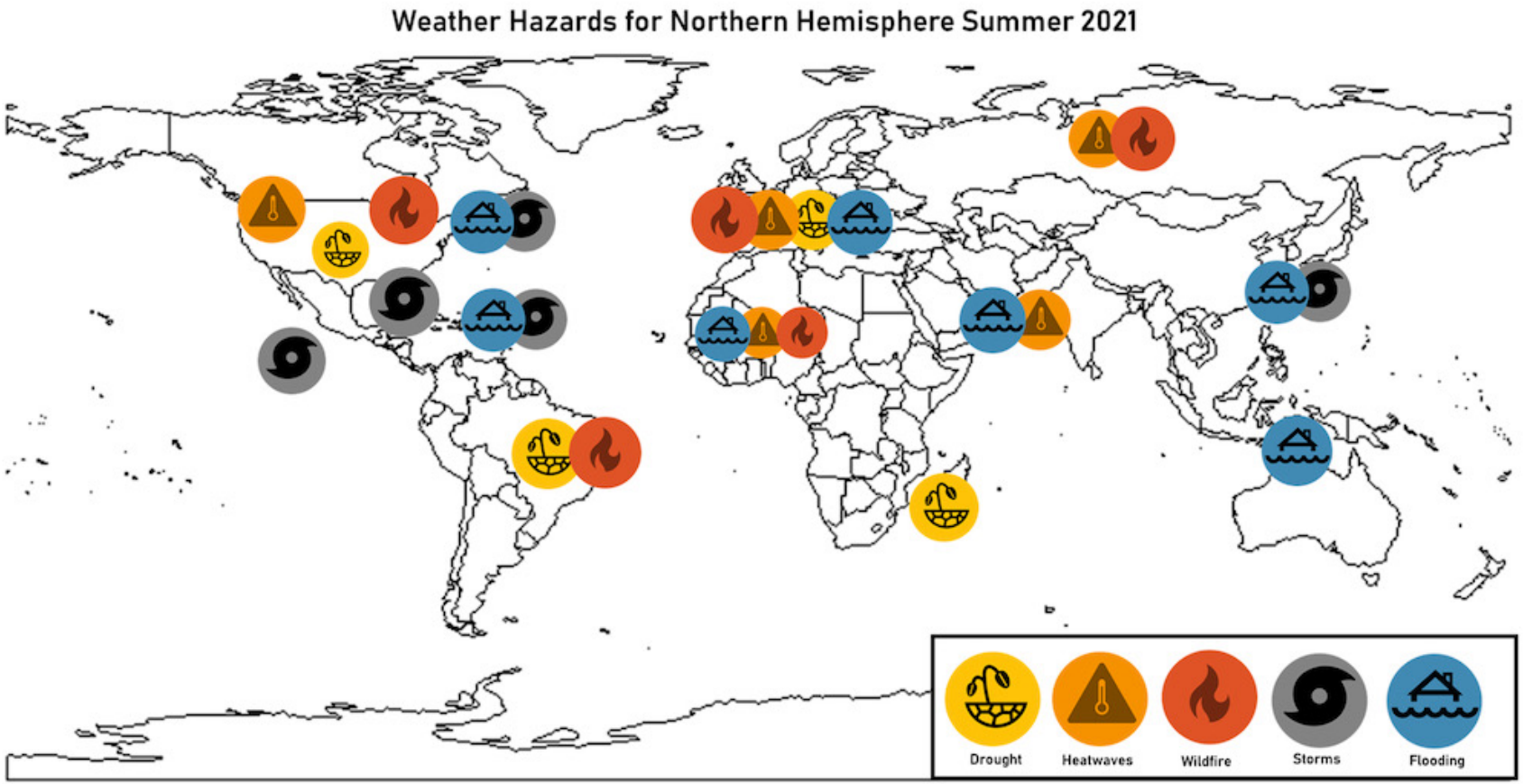
Pedram H.
Rice U.

Karthik K.
NVIDIA

Anima A.
NVIDIA / Caltech

Pathak et al. "FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators." *arXiv:2202.11214* (2022).

Dramatic rise in extreme weather events across the globe



Climate science requires million-x speedups and is challenging

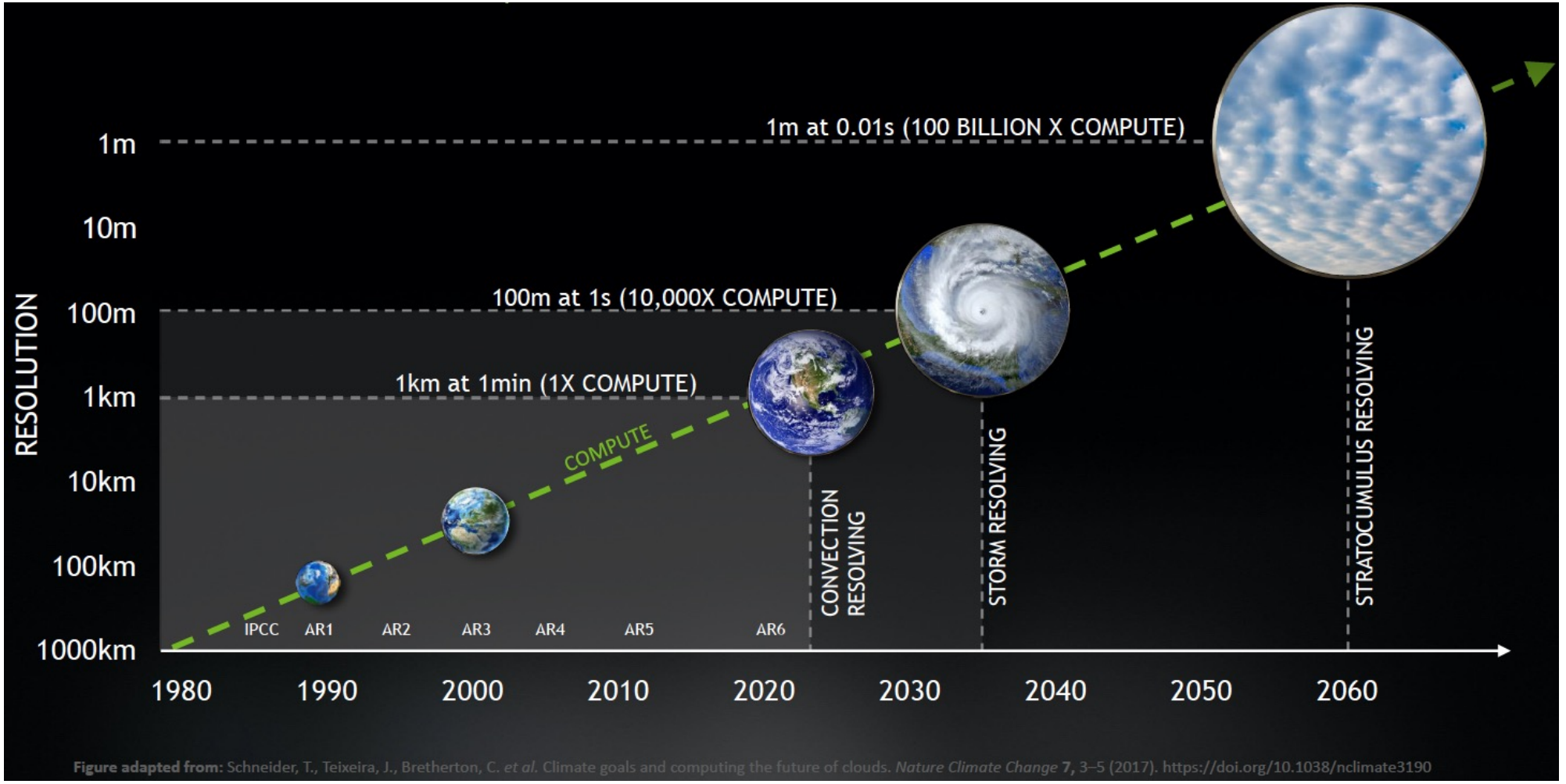


Figure adapted from: Schneider, T., Teixeira, J., Bretherton, C. *et al.* Climate goals and computing the future of clouds. *Nature Climate Change* **7,** 3–5 (2017). https://doi.org/10.1038/nclimate3190

Climate science requires million-x speedups and is challenging

Predict atmosphere dynamics and enable well-informed action

- Model complexity
  - » Multitude of physical processes: hundreds of PDEs and complex parameterizations
- Computational cost
  - » High resolution to resolve fine scales, compute scales as fourth power of resolution
  - » Large ensembles to characterize the distribution of possible outcomes
- Scalability and performance
  - » Current models not designed to exploit modern supercomputing substrates (GPUs)
  - » High energy consumption

Data-driven models can help mitigate some of the challenges

Predict atmosphere dynamics and enable well-informed action

- Model complexity

1. Data-driven surrogates can overcome model bias and learn parameterizations

- Computational cost

2. Deep learning-based surrogates exploit GPU compute yielding improved performance

- Scalability and performance

3. Optimized implementations to enable scalable AI through data/model parallelism

FourCastNet is a SOTA deep-learning based weather emulator

Scalable, data-driven global weather forecasting surrogate model

- Model complexity

1. FourCastNet is a data-driven model and shows excellent skill on important variables

- Computational cost

2. FourCastNet enables 80000x faster inference and hence larger ensembles

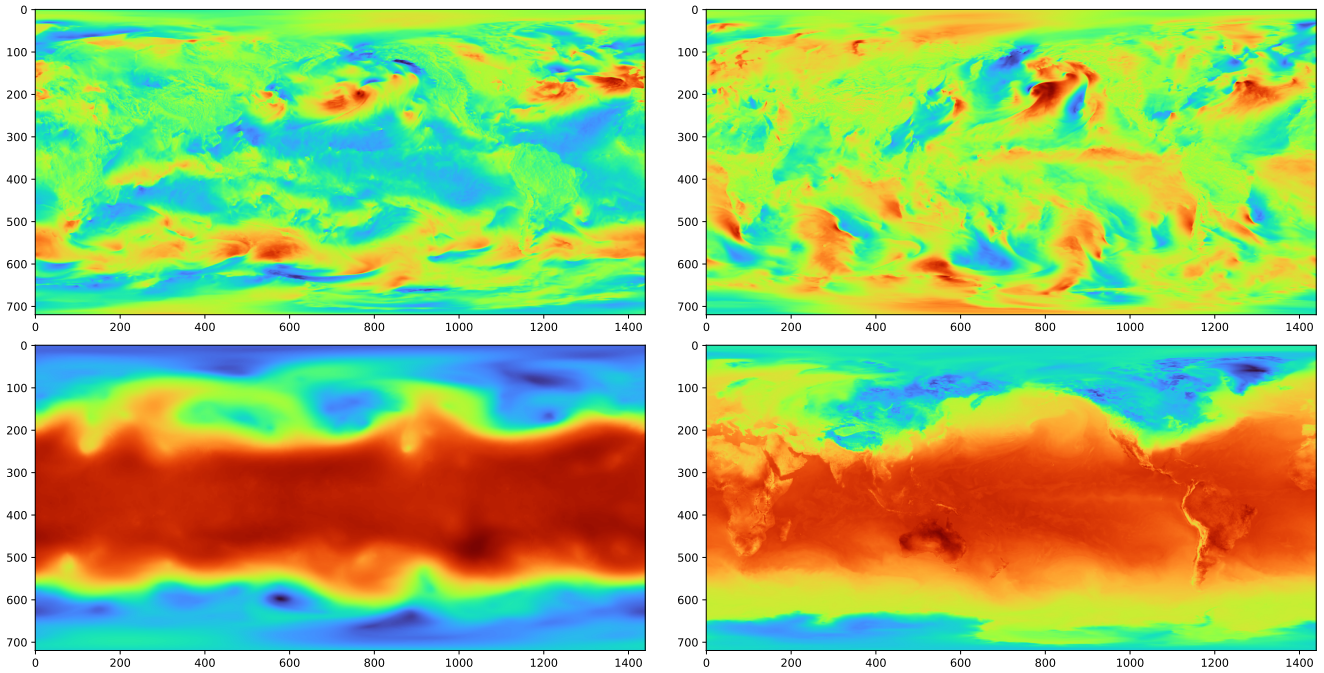
- Scalability and performance

3. FourCastNet scales to about 4000 GPUs enabling exascale weather/climate computing

Pathak et al. "FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators." *arXiv:2202.11214* (2022).

AI surrogate models are trained on ERA5 reanalysis data

- Training data from ERA5 reanalysis dataset
- 40 years (at hourly intervals) for several variables at 25km grid (720 x 1440 pixels)
- Best available estimate of the earth's atmospheric state (incorporates observations)



FourCastNet uses important global atmosphere variables

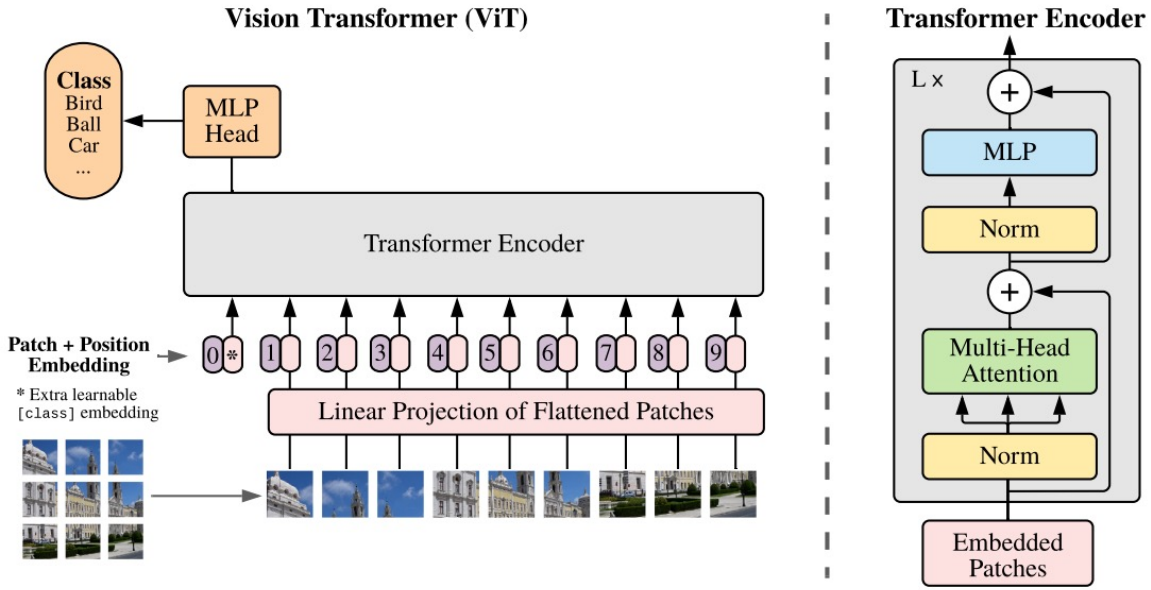
- Basic strategy: create forecasts by recursively stepping forward in time
  - » A few variables to characterize the atmospheric state
  - » Starting point for deep learning-based surrogate models
  - » Future: use all variables (like NWP)
  - » Possibly unnecessary to characterize the state
- Input vector X(t): 21 x 720 x 1440 is an input sample
  - » Training: predict target X(t + dt) from X(t)
  - » Inference: autoregressive forward step in time

Vertical Level	Variables
Surface	U_{10}, V_{10}, T_{2m}, sp, $mslp$
$1000hPa$	U, V, Z
$850hPa$	T, U, V, Z, RH
$500hPa$	T, U, V, Z, RH
$50hPa$	Z
Integrated	$TCWV$

FourCastNet is based on a vision transformer backbone

- Vision transformer: tokenization of input using patch embedding
- Scales quadratically with input

$$C \times H \times W \rightarrow N \times P^2C \rightarrow N \times D$$

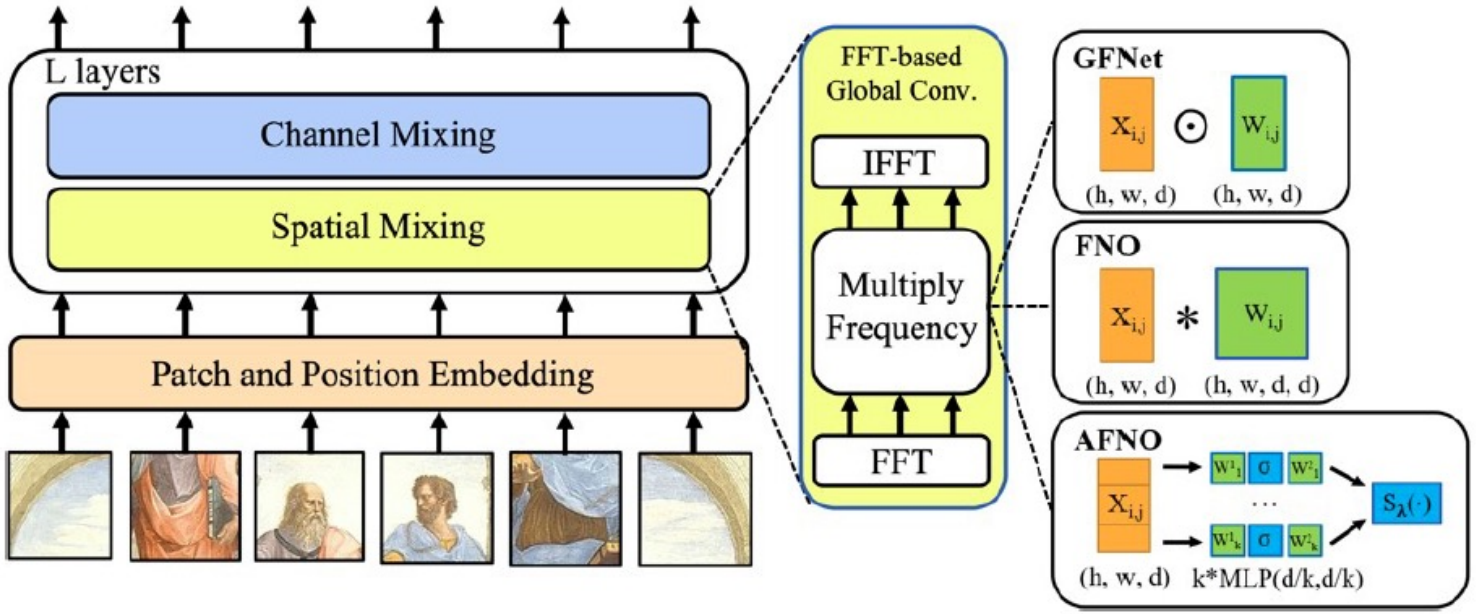


Dosovitsky et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv:2010.11929v2 (2021).

FourCastNet uses Fourier Neural Operators for token mixing

- Token mixing is done in Fourier space: deals with quadratic complexity
- Implements global convolution (global mixing of tokens)

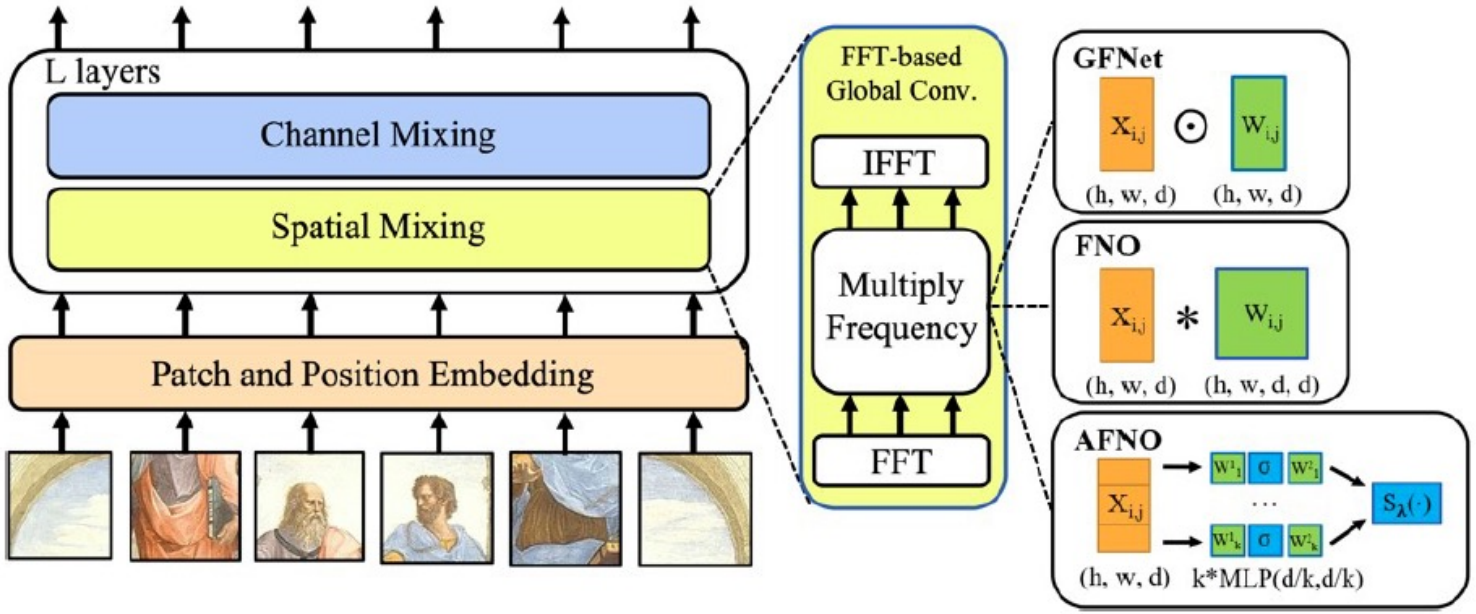
$$N \times D \rightarrow H/p \times W/p \times D$$



Gubias et al. ”Adaptive Fourier Neural Operators: Efficient Token Mixing for Transformers." arXiv:2111.13587 (2022).

Adaptive Fourier Neural Operator controls channel complexity

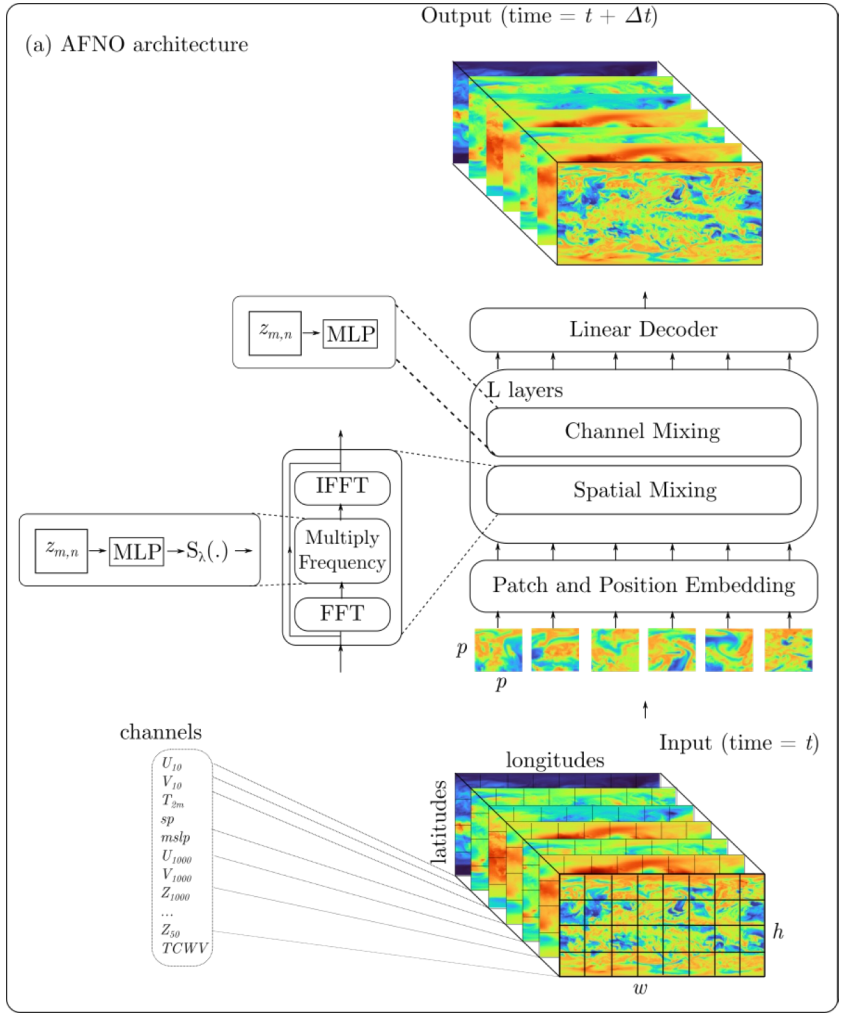
- Block diagonal structure on channel-mixing weights
- Weights are shared across tokens
- Sparse prior on frequencies via soft thresholding



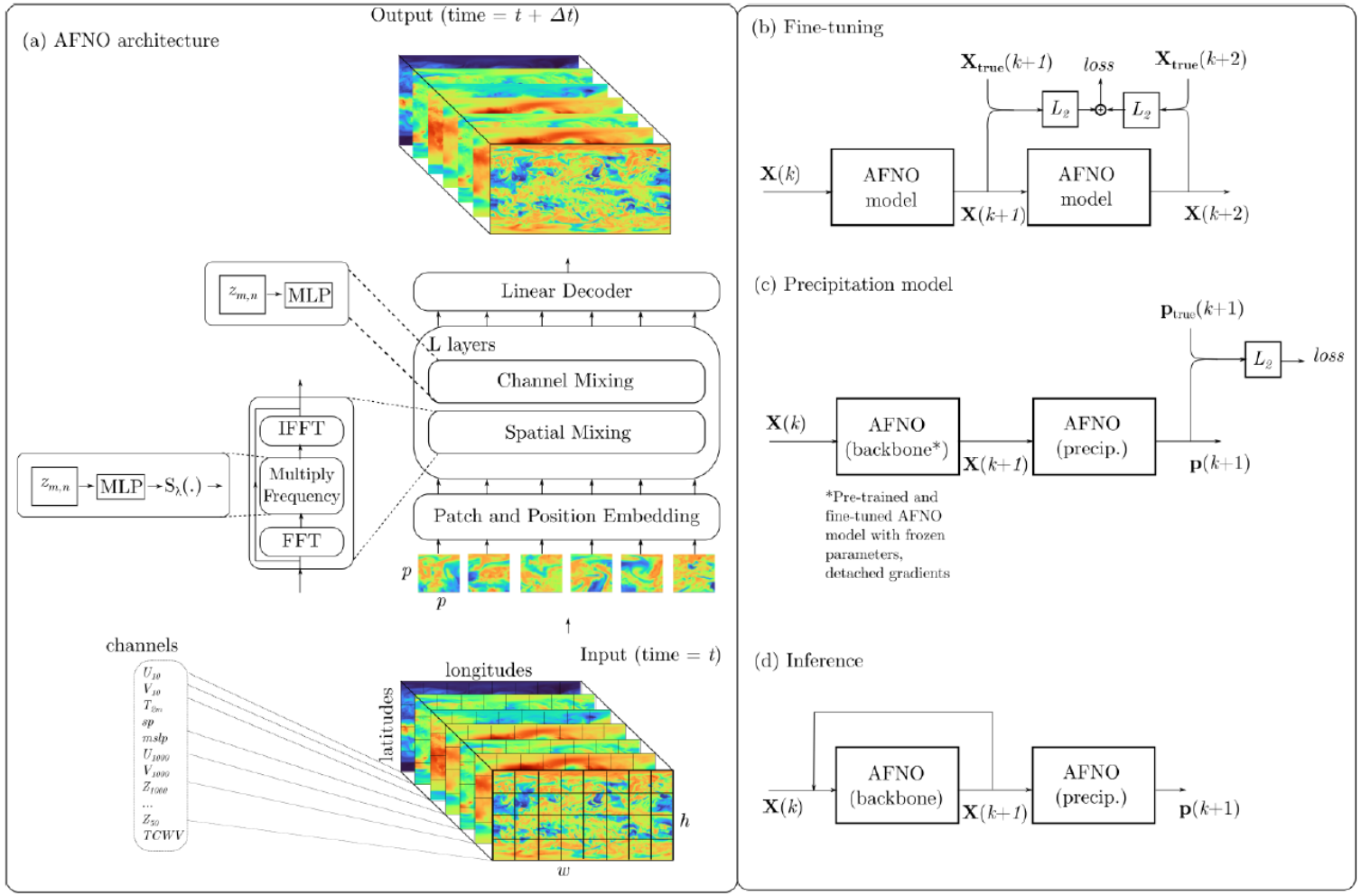
Gubias et al. ”Adaptive Fourier Neural Operators: Efficient Token Mixing for Transformers." arXiv:2111.13587 (2022).

FourCastNet architecture is stacked AFNO blocks

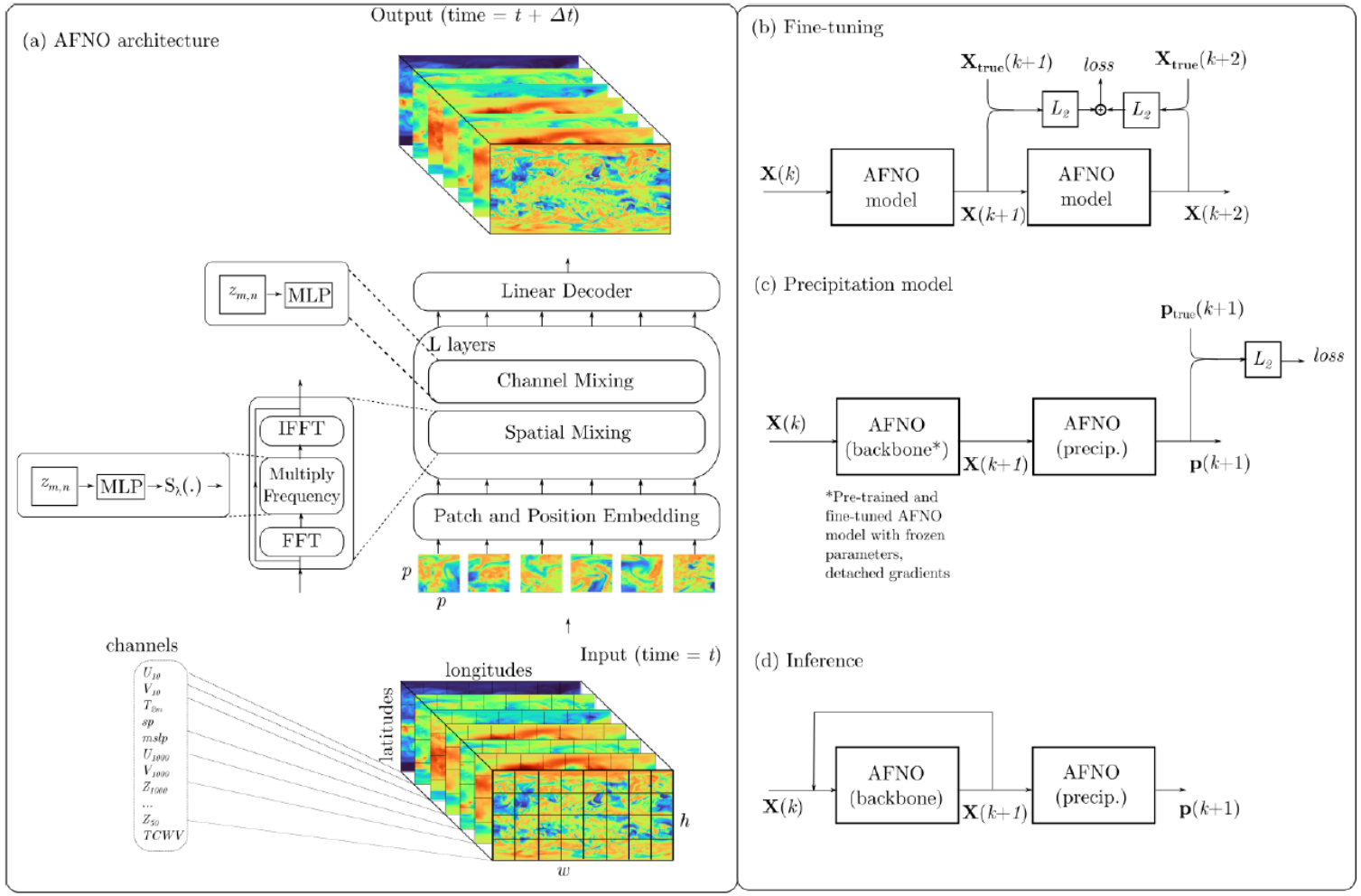
- AFNO blocks repeated (12x)
- Outputs are reshaped into input image tensor size at future time
- Trained on full, high-resolution inputs
- Training on 1979 – 2015
- Validation on 2016, 2017
- Testing on 2018+
- Total data size ~ 5TB



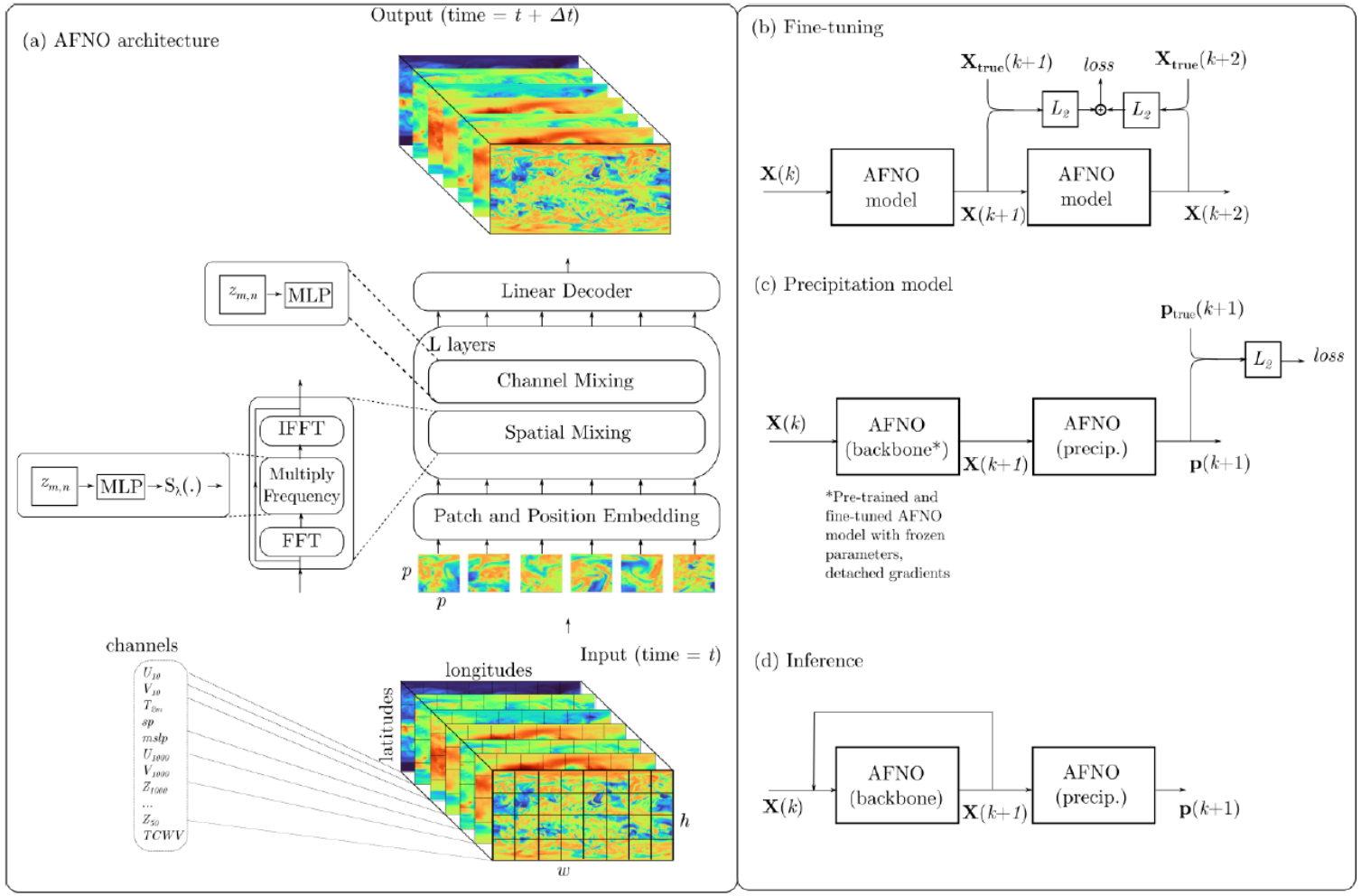
Model is pretrained and finetuned with two time steps



Precipitation is diagnostic with the same model architecture

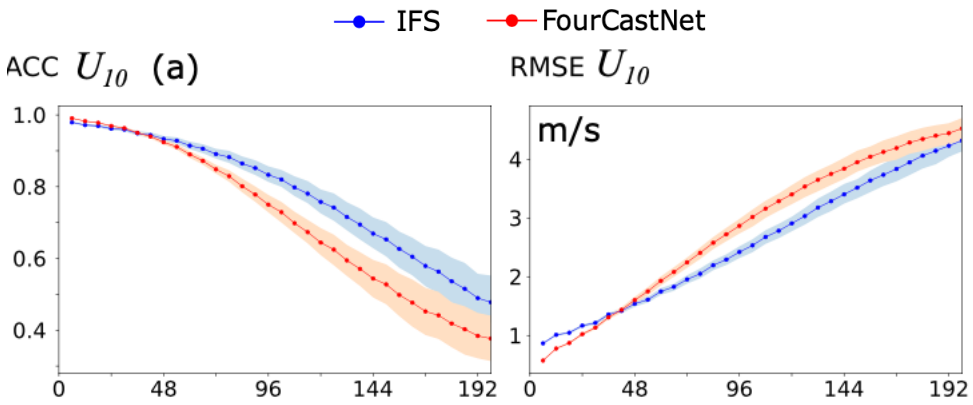


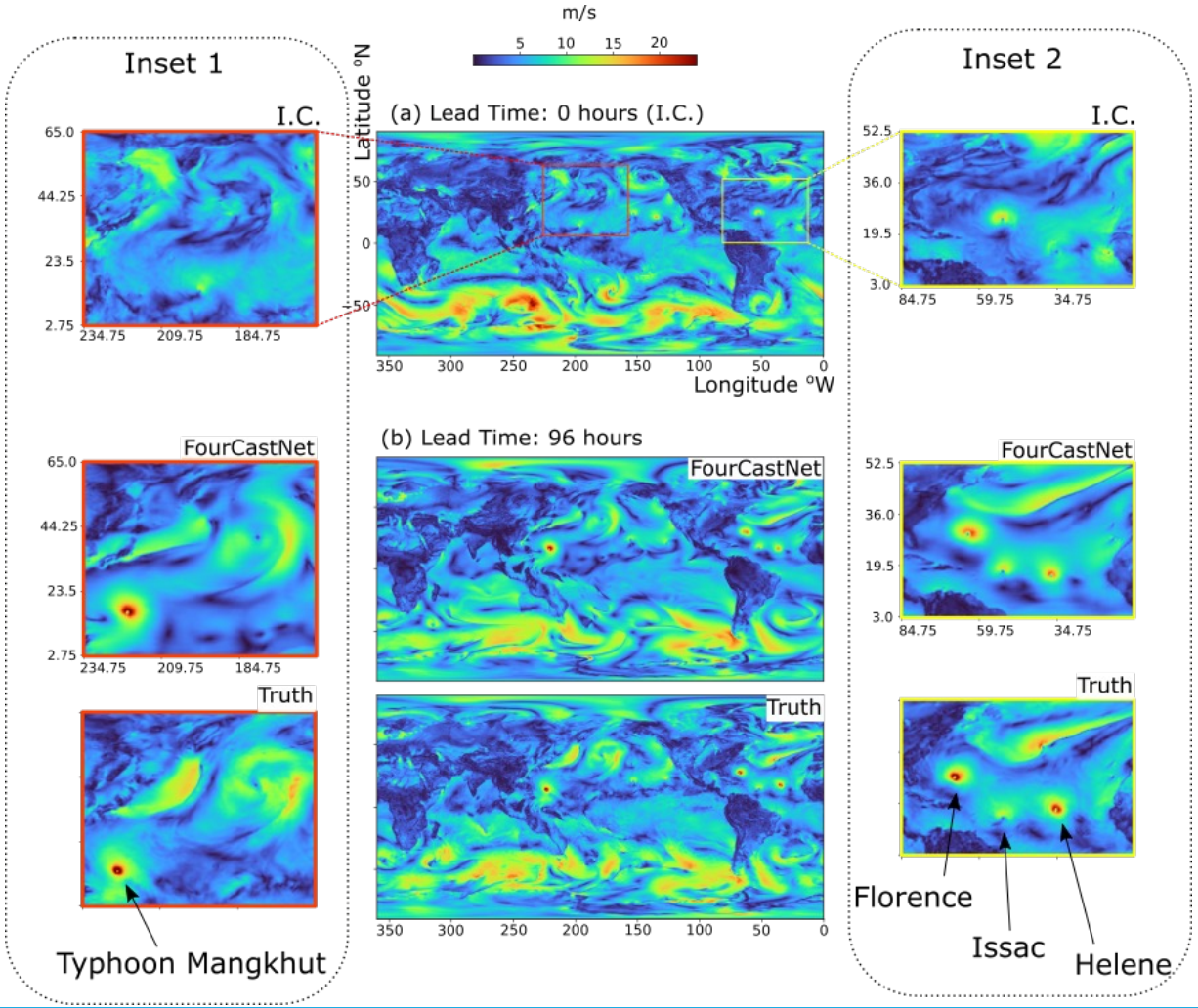
Inference is autoregressive



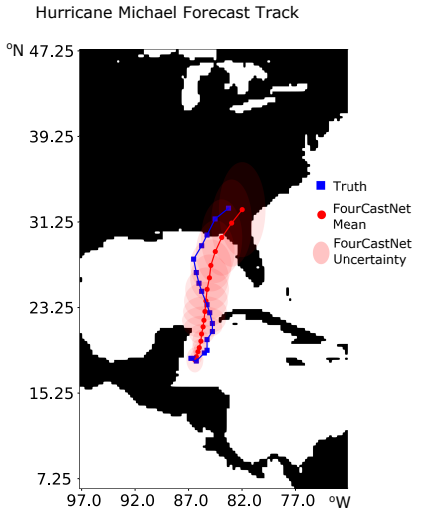
FourCastNet shows excellent skill in predicting surface winds

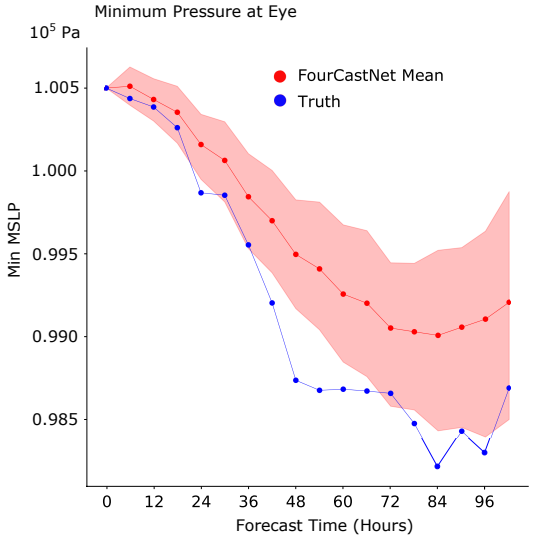
- Accuracy metrics used are ACC and RMSE
- Comparable to the operational IFS at high resolution

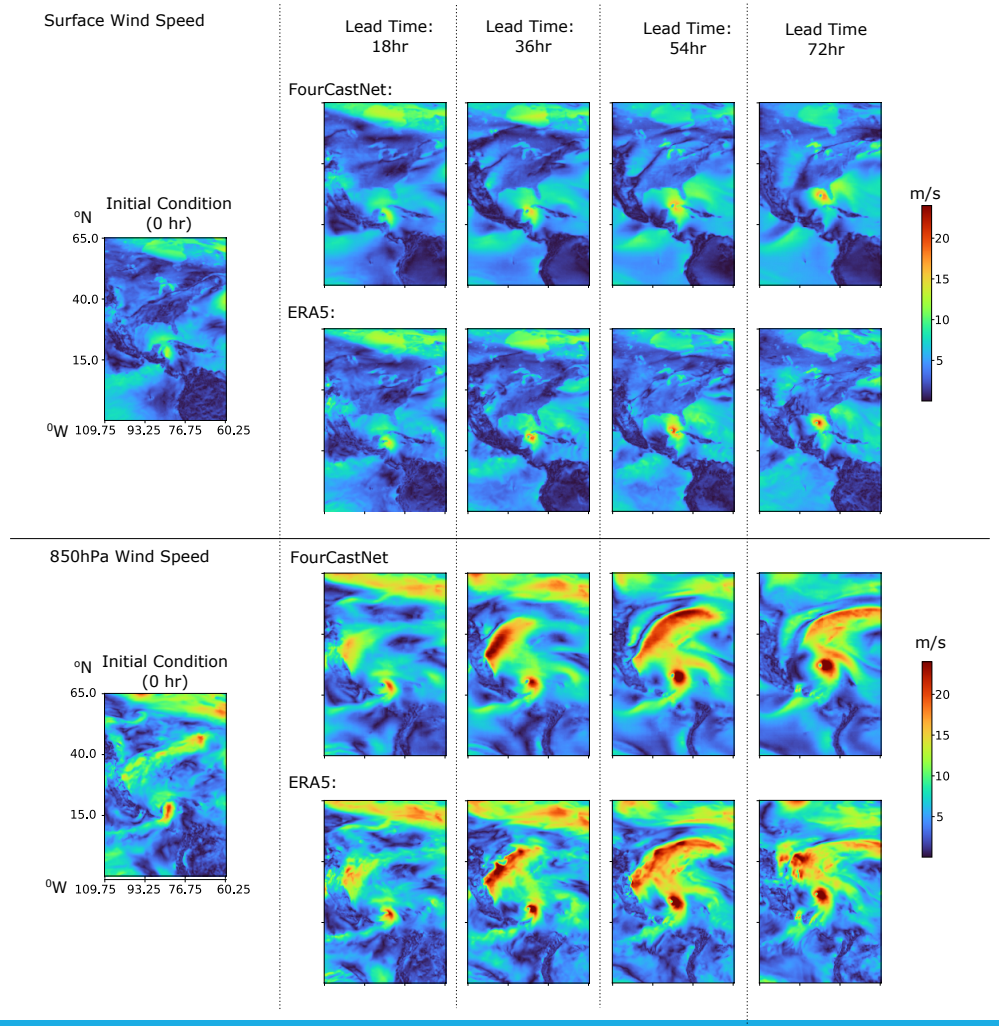




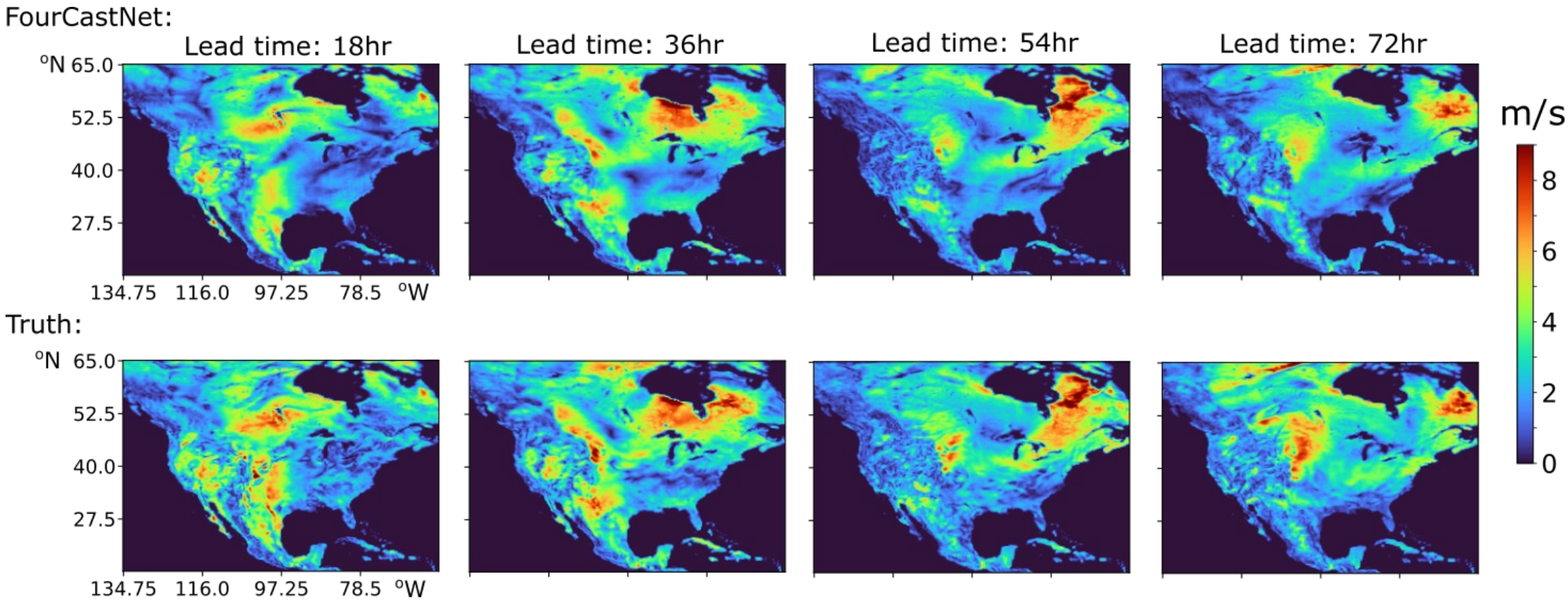
FourCastNet predicts hurricane paths and intensities





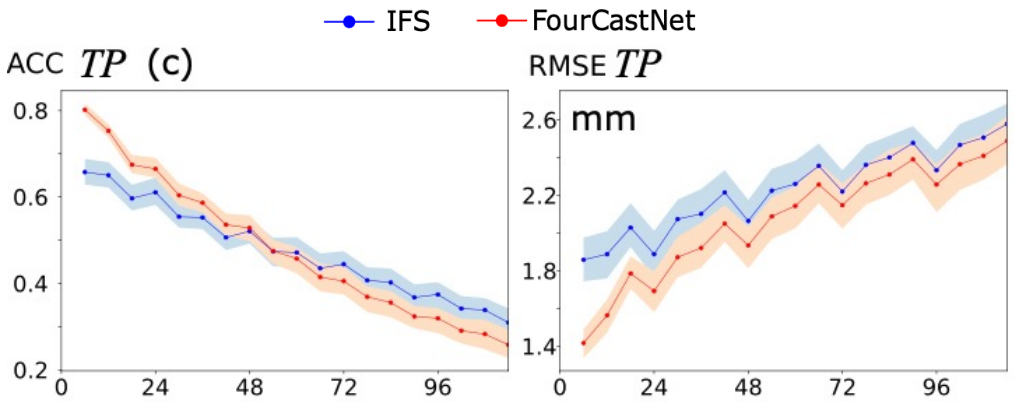


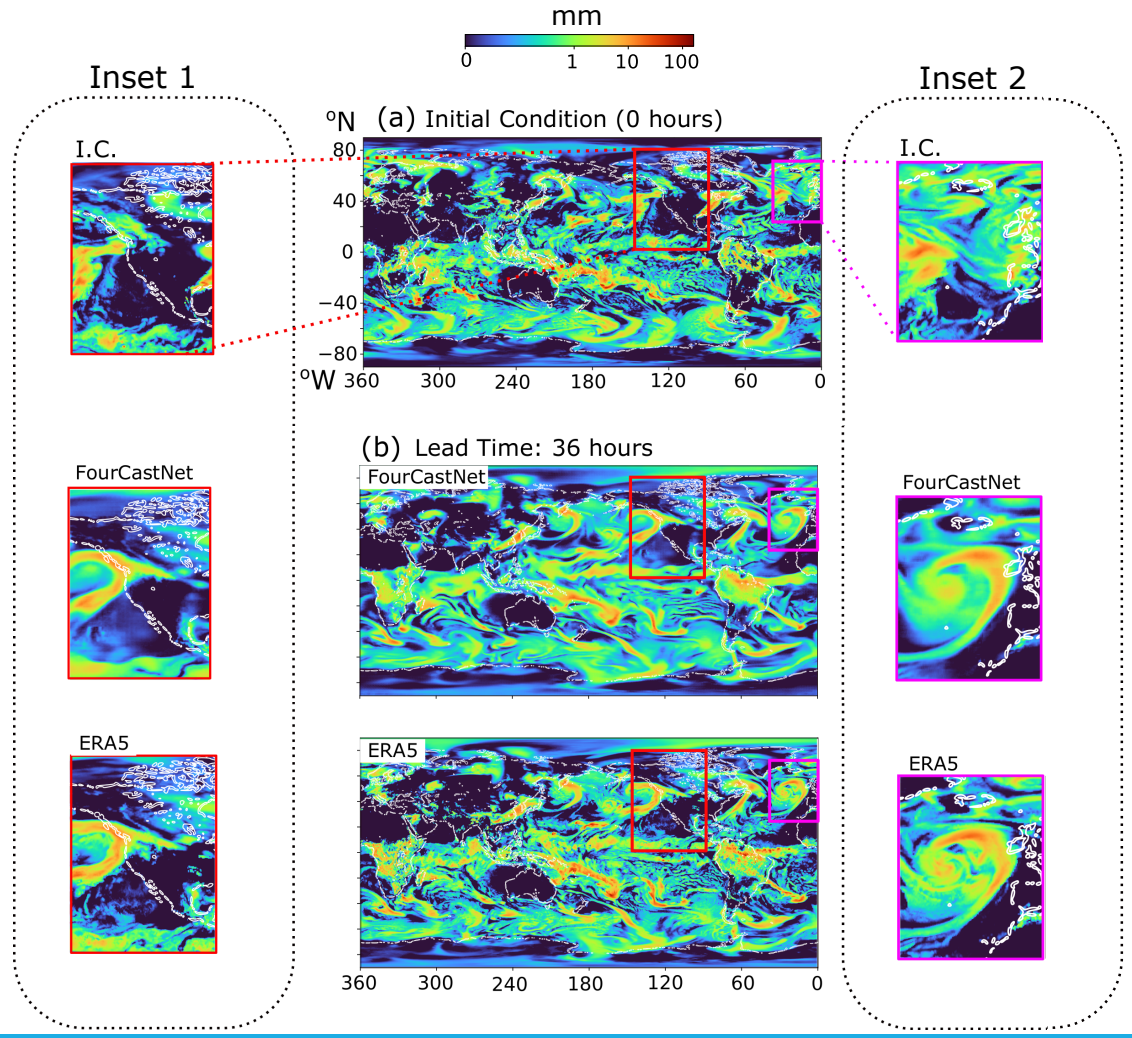
Near surface wind predictions over land show good accuracy



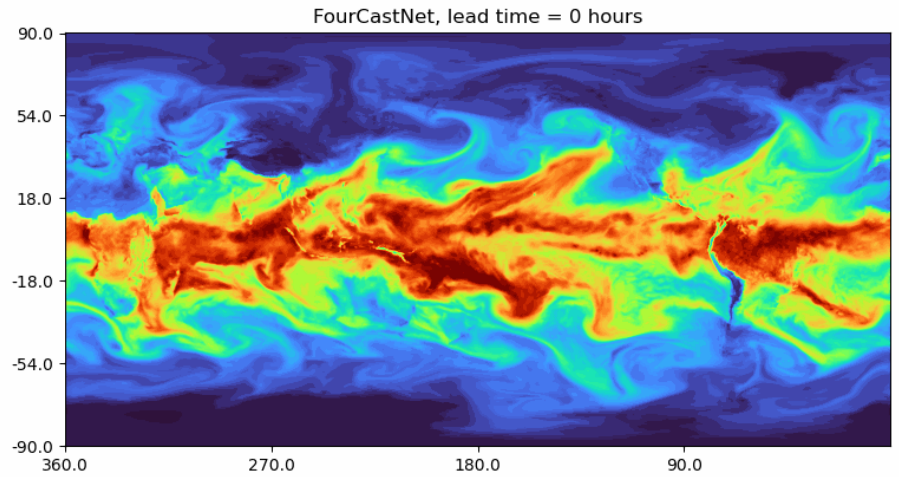
FourCastNet excellent skill on total precipitation

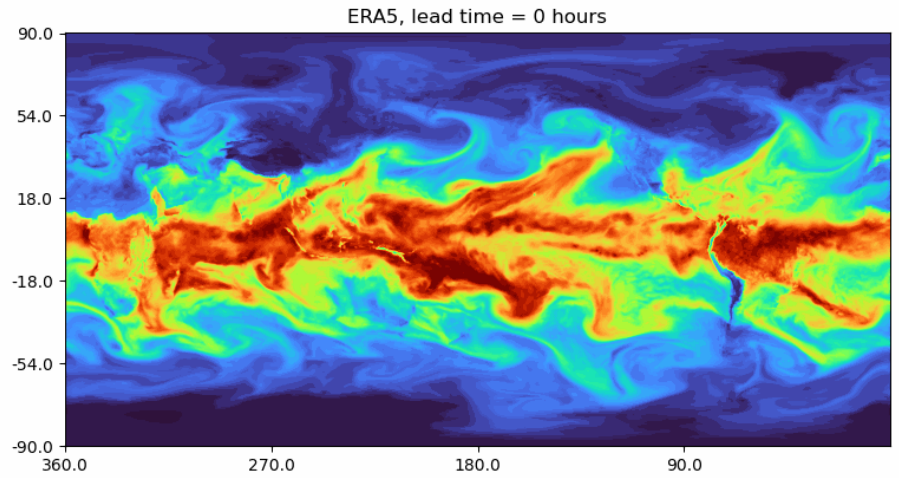
- Shows comparable skill to the IFS
- Better skill at short time horizons
- Captures fine-scale features well

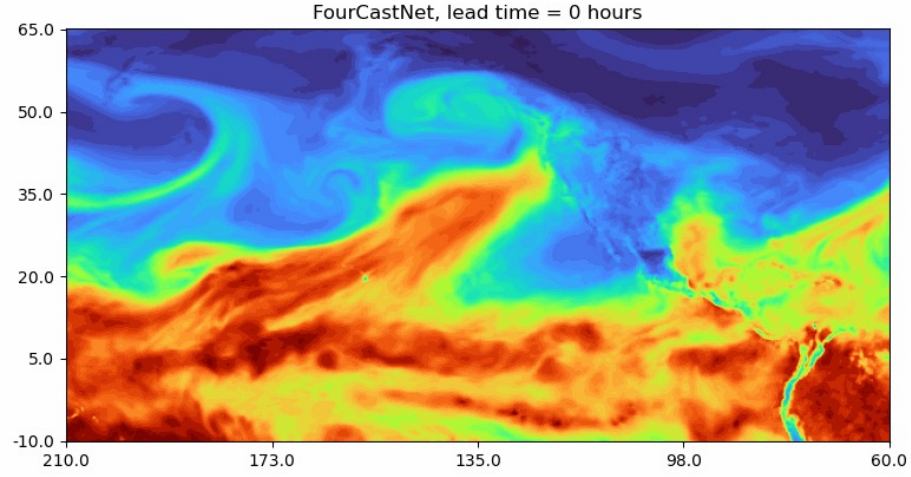


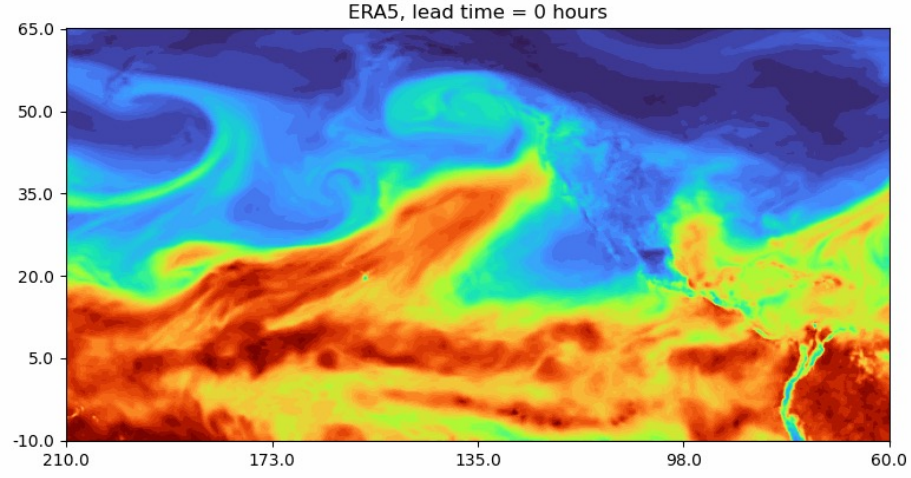


Moisture variable dynamics are captured well

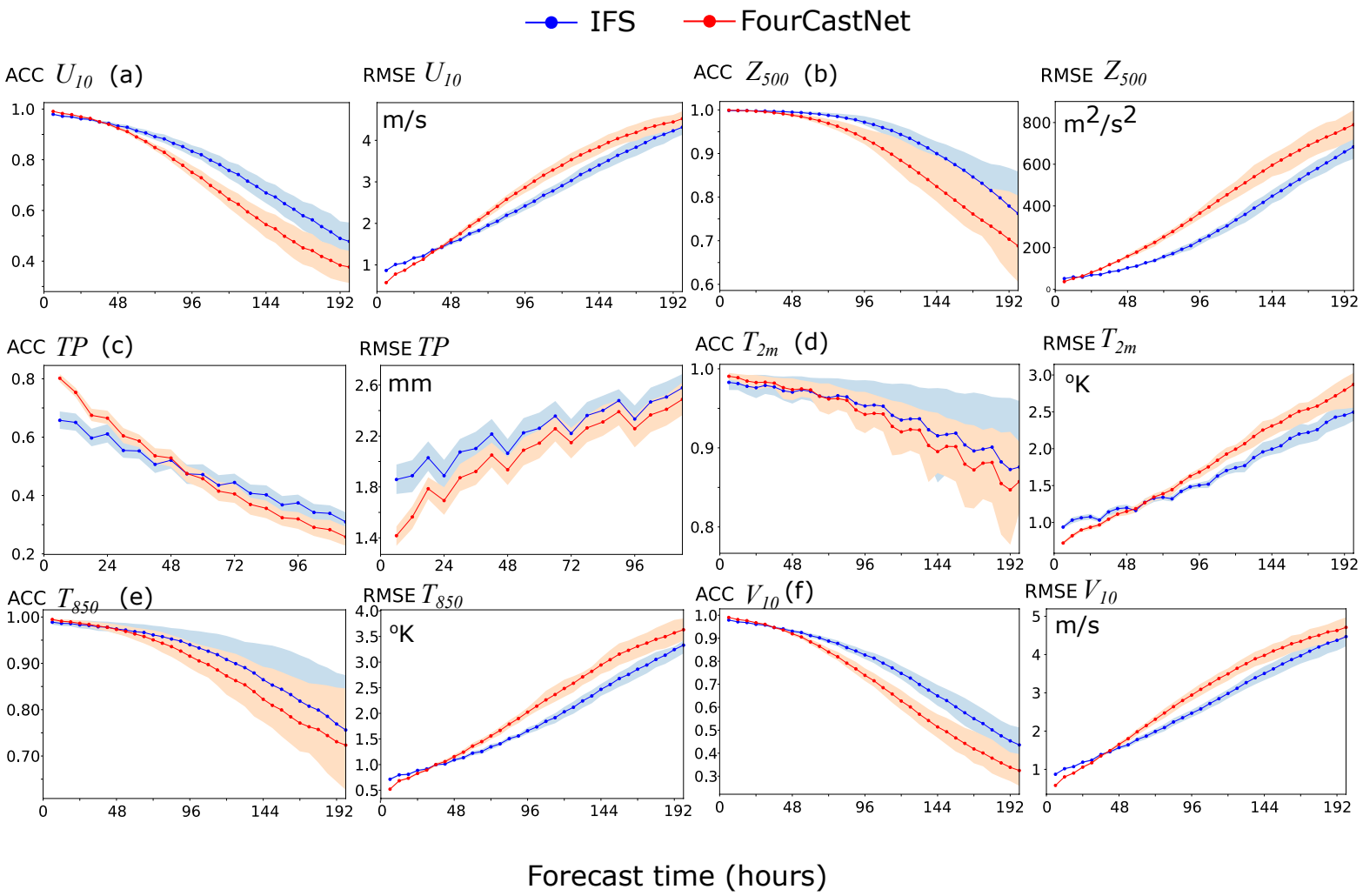




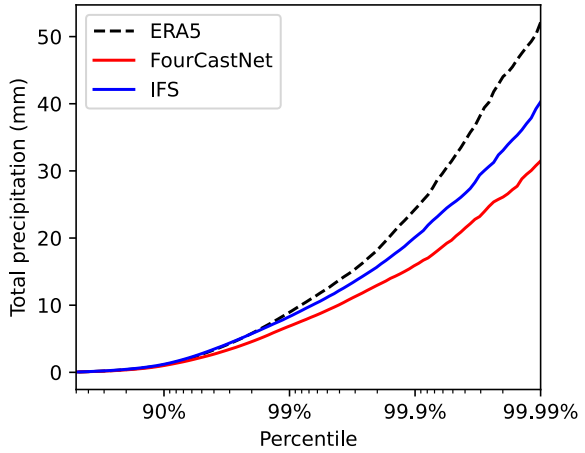


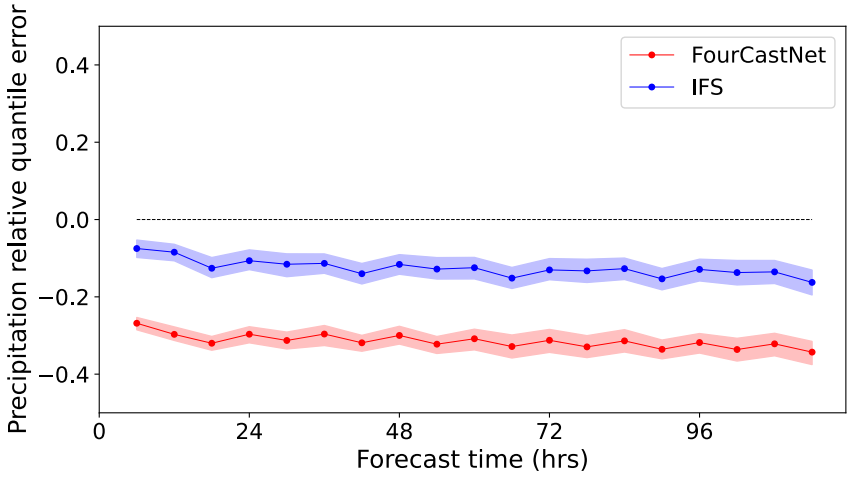


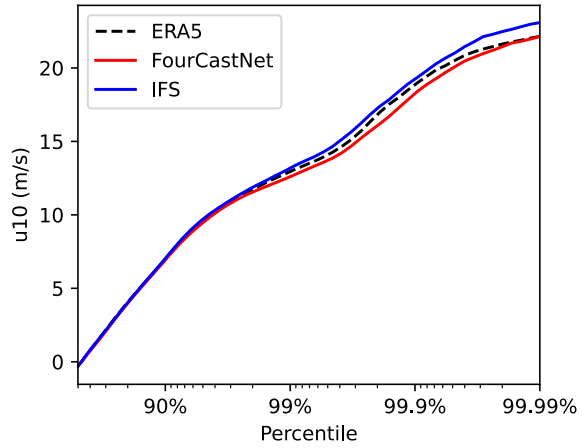
Short-term ACC and RMSE metrics are close to IFS

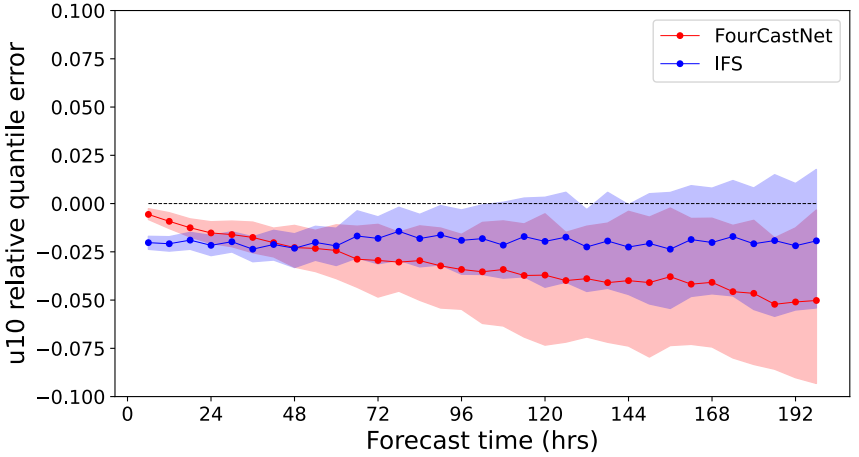


Good performance on extremes but model misses rare events



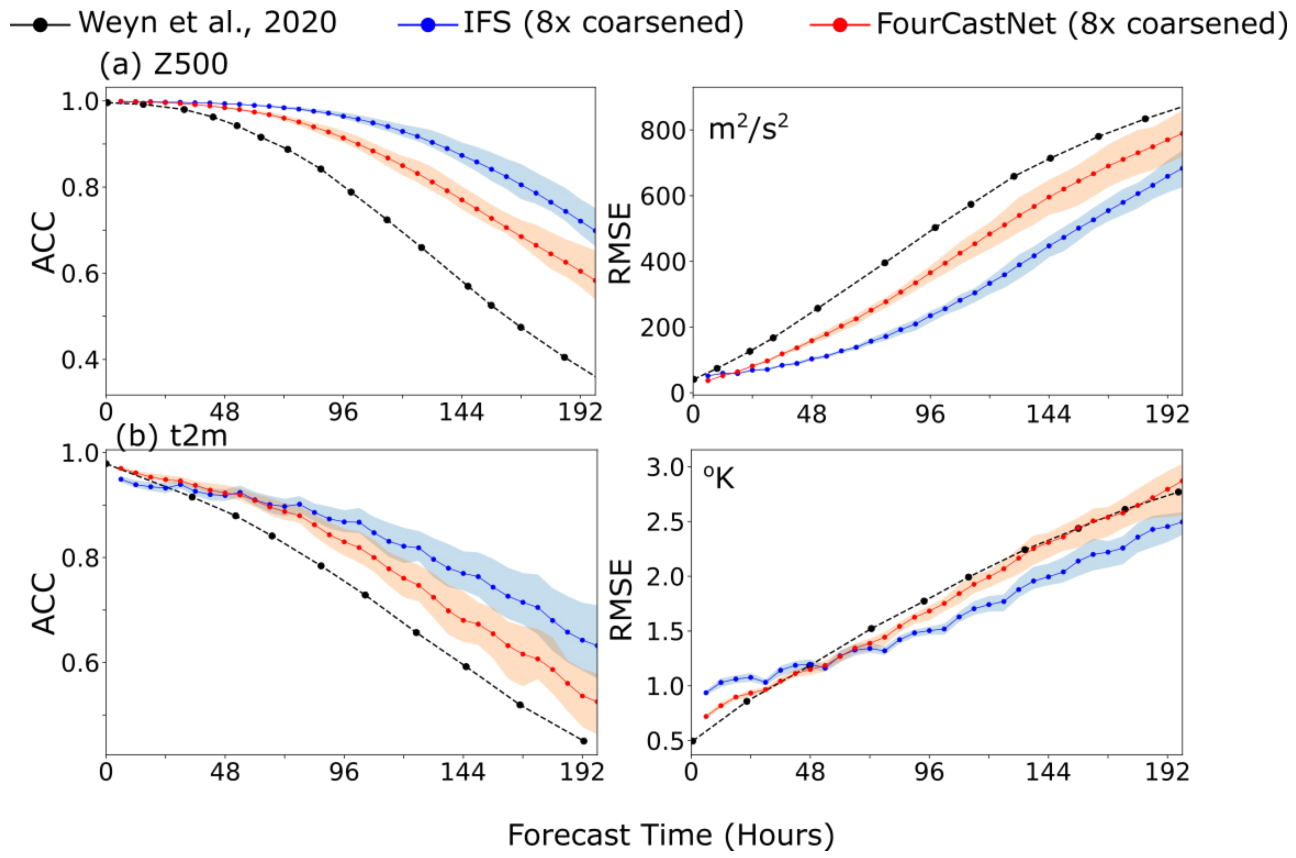






FourCastNet beats the current SOTA for DL weather models

- Current SOTA is DLWP
- Operates at 8x coarser resolution
- Downsample FourCastNet predictions for comparison
- Higher skill on weather timescales
- Important note: DLWP can predict well at S2S timescales



Weyn et al. ”Improving data-driven weather prediction using deep convolutional networks on a cubed sphere." arXiv:2003.11927 (2020).

FourCastNet is faster and more energy efficient than NWP

Latency and Energy consumption for a 24-hour 100-member ensemble forecast				
	IFS	FCN - 30km (actual)	FCN - 18km (extrapolated)	IFS / FCN(18km) Ratio
Nodes required	3060	1	2	**1530**
Latency (Node-seconds)	984000	7	22	**44727**
Energy Consumed (kJ)	271000	7	22	**12318**

- 100 member ensemble forecast in 7 seconds with 7 kJ energy consumed
- 4 to 5 orders of magnitude speedup
- 4 orders of magnitude smaller energy footprint

FourCastNet is faster and more energy efficient than NWP

Latency and Energy consumption for a 24-hour 100-member ensemble forecast				
	IFS	FCN - 30km (actual)	FCN - 18km (extrapolated)	IFS / FCN(18km) Ratio
Nodes required	3060	1	2	**1530**
Latency (Node-seconds)	984000	7	22	**44727**
Energy Consumed (kJ)	271000	7	22	**12318**

- FourCastNet is *not* physics constrained
- Fewer variables and levels than IFS
- Cannot perform data assimilation and is reliant on IFS for an IC

Scaling challenges exist as we go to higher resolutions

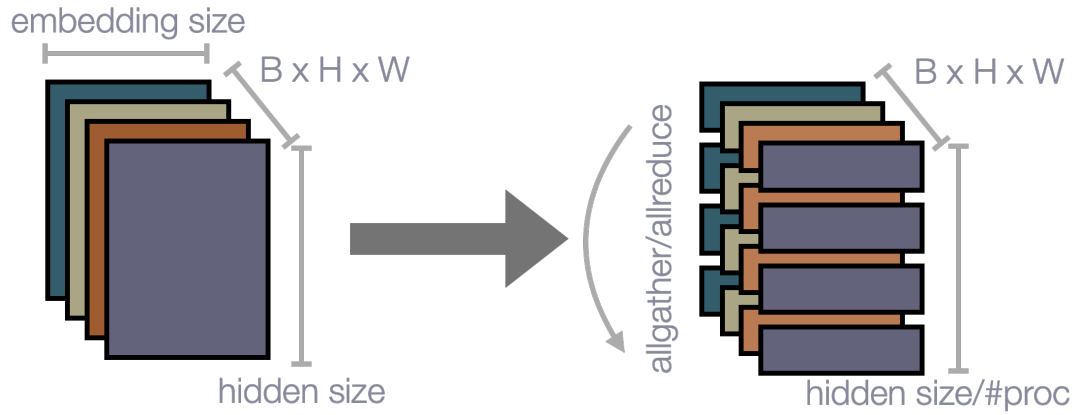
Models	Complexity (FLOPs)	Parameter Count
AFNO (ours)	$Nd^2/k + Nd \log N$	$(1 + 4/k)d^2 + 4d$

	Current (25 km)	Intermediate (5 km)	Large (1 km)
N (p = 1)	1M	25M	625M
FFTs	720 x 1440 (d of them)	3600 x 7200 (d of them)	18k x 36k (d of them)
Matmul	[4d x d] * [d] (N of them)	[4d x d] * [d] (N of them)	[4d x d] * [d] (N of them)

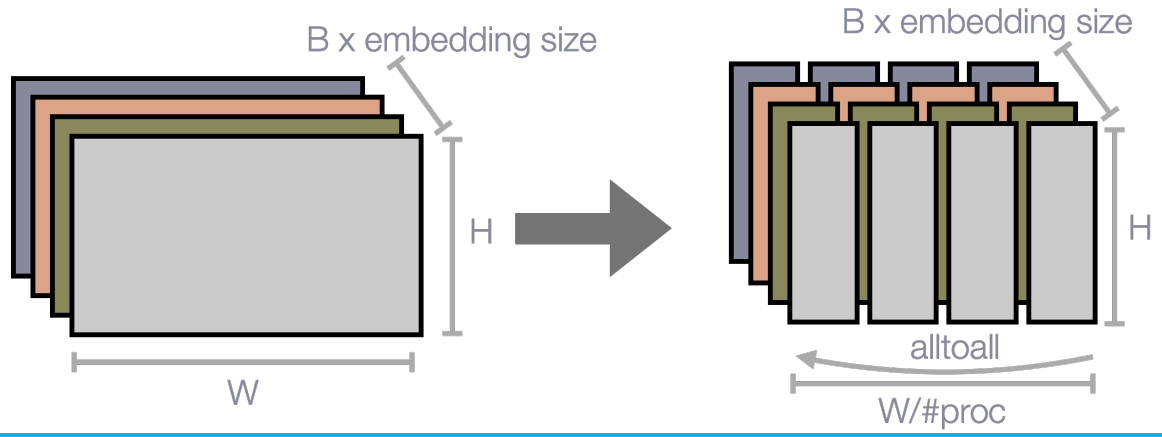
- Have not reached the target resolution: would require much higher compute
- Memory footprint also increases with resolution (hundreds of GB at intermediate resolutions)
- Temporal resolution can be increased as well
- Different parallelization strategies needed

Different parallelization strategies to think about

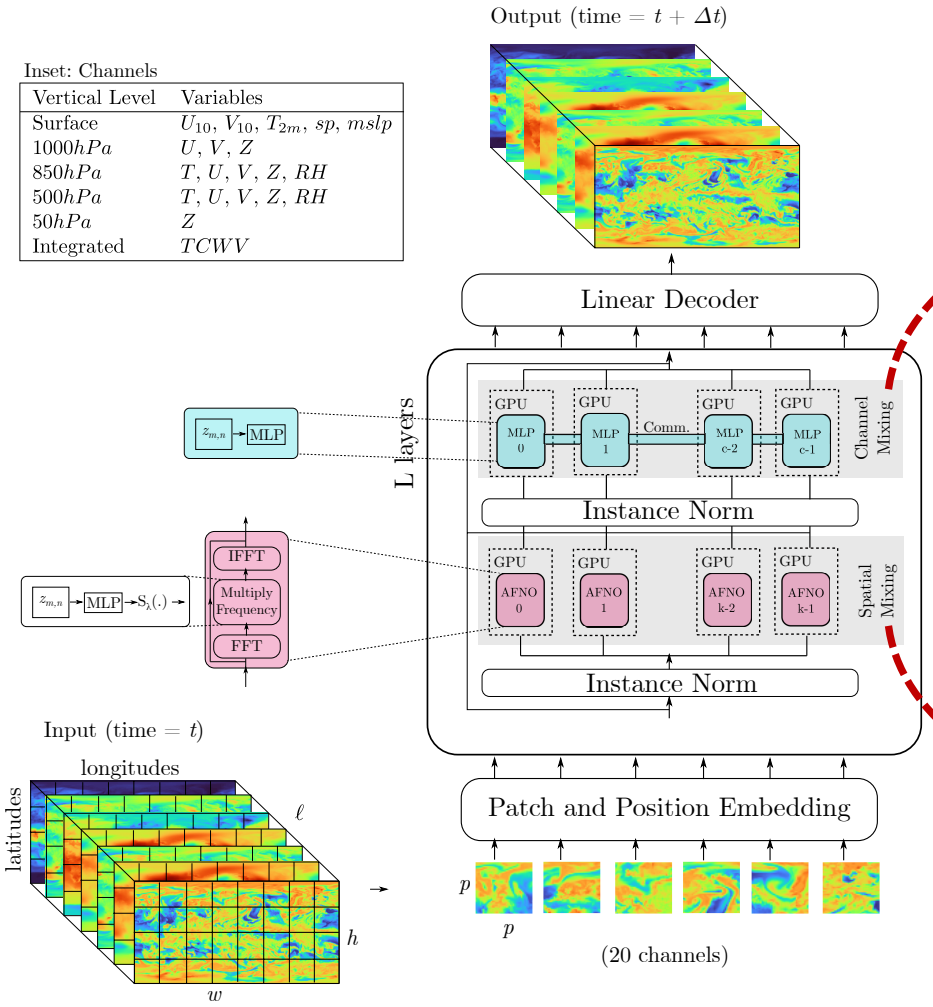
- Data parallelism: standard, global batch is split across procs (GPUs)
- Feature parallelism: split channel dim, dense layers become distributed matrix multiply



- Spatial parallelism: split H/W and FFTs become distributed FFTs, need to exchange stats for layer norms



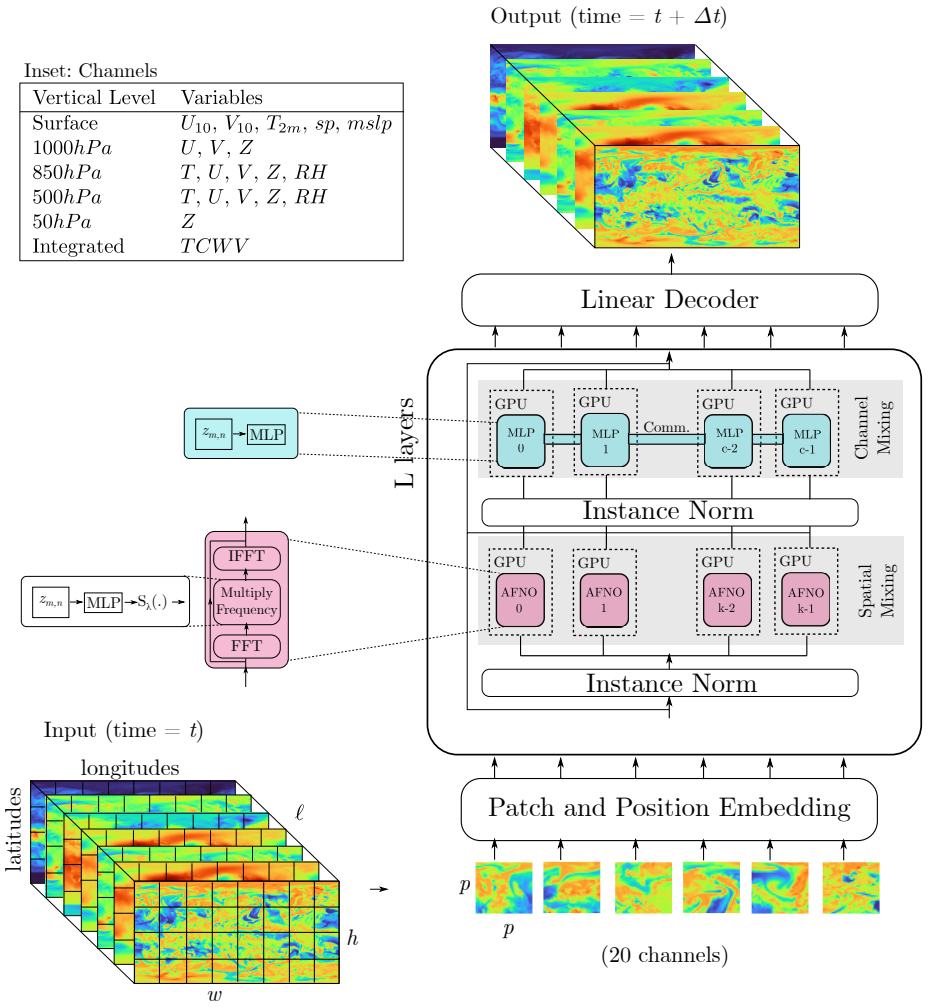
FourCastNet employs feature parallelism and data parallelism



Channel mixing MLPs split hidden dimension across GPUs and require communication in every fwd/bwd pass

FFT-based spatial mixing operate on disjoint blocks and are embarrassingly parallel

FourCastNet employs feature parallelism and data parallelism



- Model parallelism across GPUs within a node: model instance
  - » Exploit NVLink interconnect
- Data parallelism across model instances
  - » Gradient reduction across model instances
- Independent communicators for model parallel and data parallel

Need to think about forward and backward pass for MLPs

- Operations need different communication in forward and backward passes
  - » Row-parallel matrix multiply has no comm in forward but AllReduce in backward pass
- Hidden dim is 4x and split
  - » NCHW format, matmul with pointwise conv layer
  - » Employ row-parallelism for the first matrix multiply
  - » Employ column-parallelism for the second matrix multiply
- Custom autograd functions to implement these cases

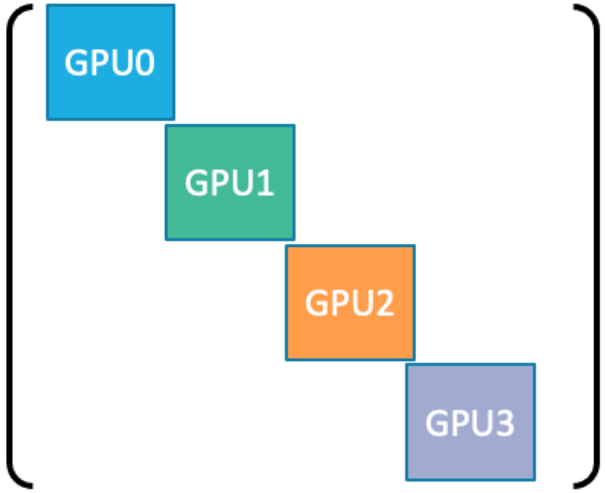
```
def forward(self, x):
    x = copy_to_matmul_parallel_region(x)

    # do the mlp
    x = F.conv2d(x, self.w1, bias=self.b1)
    x = self.act(x)
    x = self.drop(x)
    x = F.conv2d(x, self.w2, bias=None)
    x = reduce_from_matmul_parallel_region(x)
    x = x + torch.reshape(self.b2, (1, -1, 1, 1))
    x = self.drop(x)

    return x
```

Spectral convolutions are embarrassingly parallel

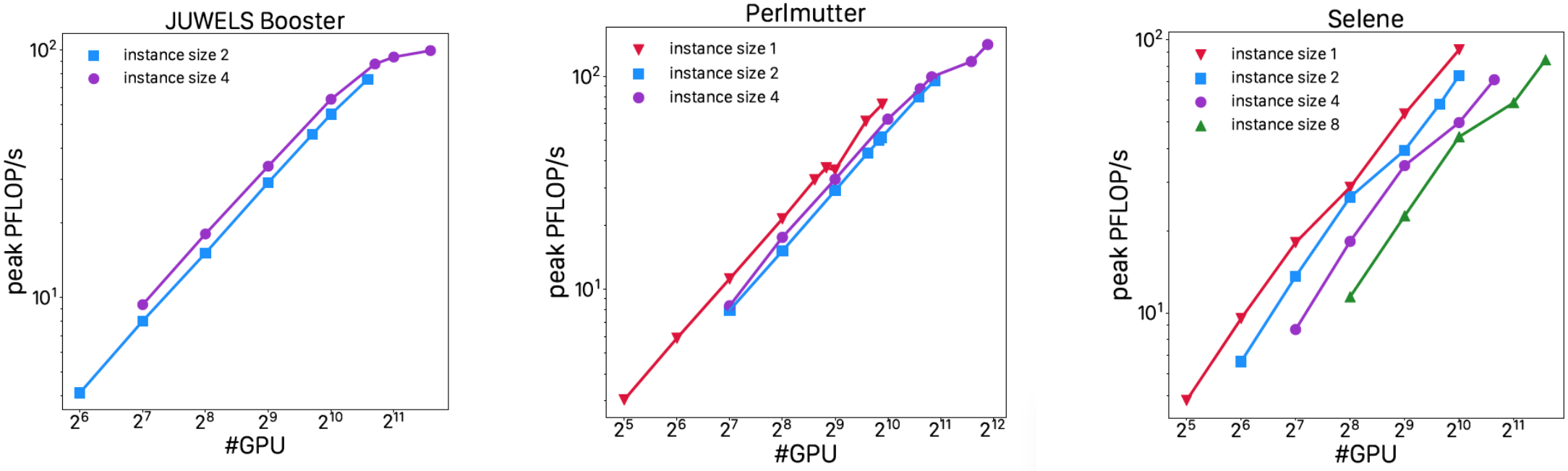
- No specialized PyTorch layer: implement using R2C and C2R FFT layers
- FFTs for channels are split (feature parallelism)
- Blocks are distributed to different GPUs
- Both are embarrassingly parallel



DALI for I/O, CUDA graphs to keep CPU out, JIT for metrics

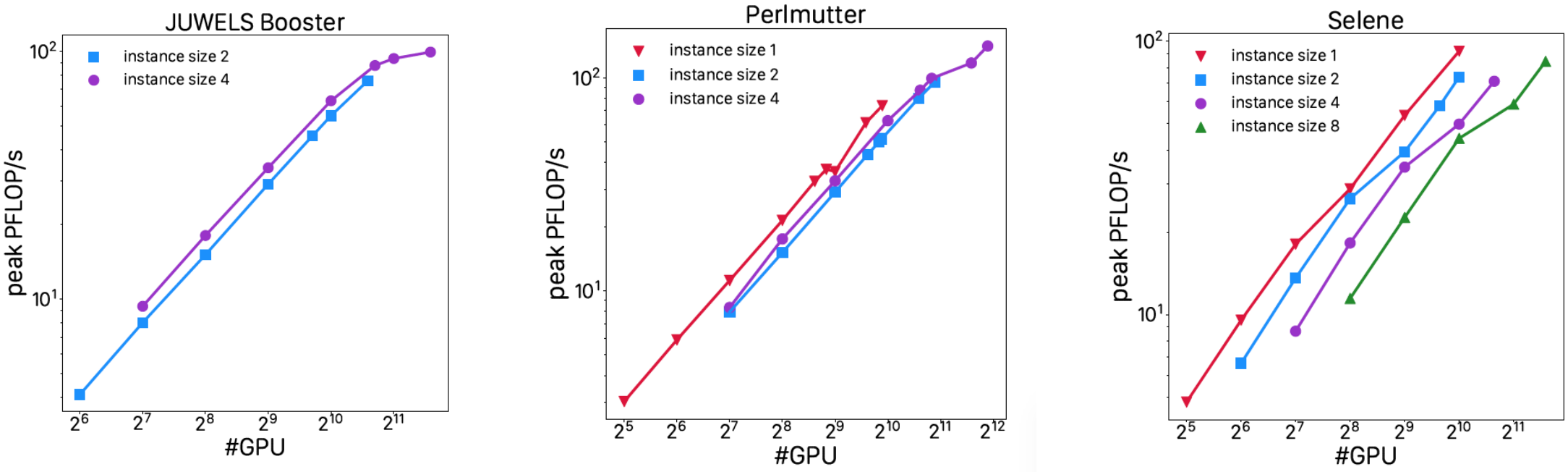
- NVIDIA DALI for I/O
  - Efficient I/O ops that run concurrent to GPU computations
  - Preprocessing steps on the GPU (normalization, adding noise)
- CUDA graphs to take CPU out of the loop as much as possible
  - CPU determines what kernels to launch next and can stall GPU exec
  - Graphs record a sequence of kernel launches and replay it as often as needed
- Metrics like ACC and RMSE are lightweight ops but generate separate kernel calls
  - JIT from torch to fuse ops into larger kernels to reduce launch latency
- NVIDIA Nsight Compute to measure floating point ops
  - FLOPs generating instructions multiplied by weighting factors (ex: FP32 FMA weight is 2)
  - Single GPU FLOPs x #GPUs x #iterations (median and max)

FourCastNet scales to 3808 GPUs with 140.8 PFLOP/s peak



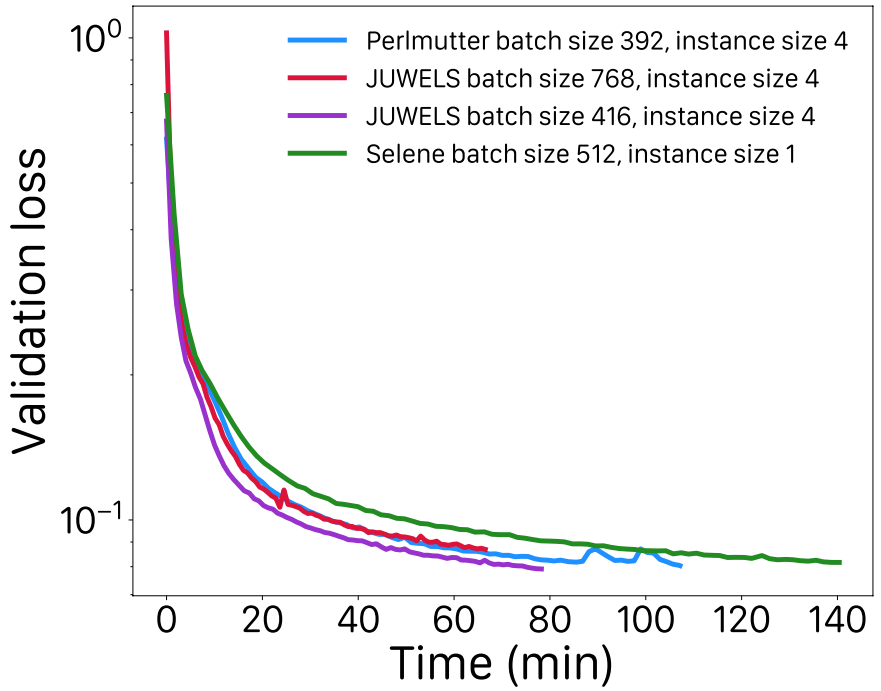
- PM scales till ~3800 GPUs, Selene ~3100 GPUs, JUWELS starts to taper off at ~1024 GPUs
- Peak performance of PM is 140.8 PFLOP/s on 3808 GPUs

System dependent performance for model and data parallel



- Model instance 4 perform better than 2 for PM and JUWELS
- Higher performance gain when data parallelism is increased on Selene
  - » Selene has balanced intranode and internode bandwidth and does not suffer from large scale collectives

Large resources enable short training times



- Would take 40 hours to train a batch size 32 model with no model parallelism
- Faster time-to-solutions with greater parallelism

FourCastNet holds great promise towards DL weather models

- Model complexity

1. FourCastNet is a data-driven model and shows excellent skill on important variables

- Computational cost

2. FourCastNet enables 80000x faster inference and hence larger ensembles

- Scalability and performance

3. FourCastNet scales to about 4000 GPUs enabling exascale weather/climate computing

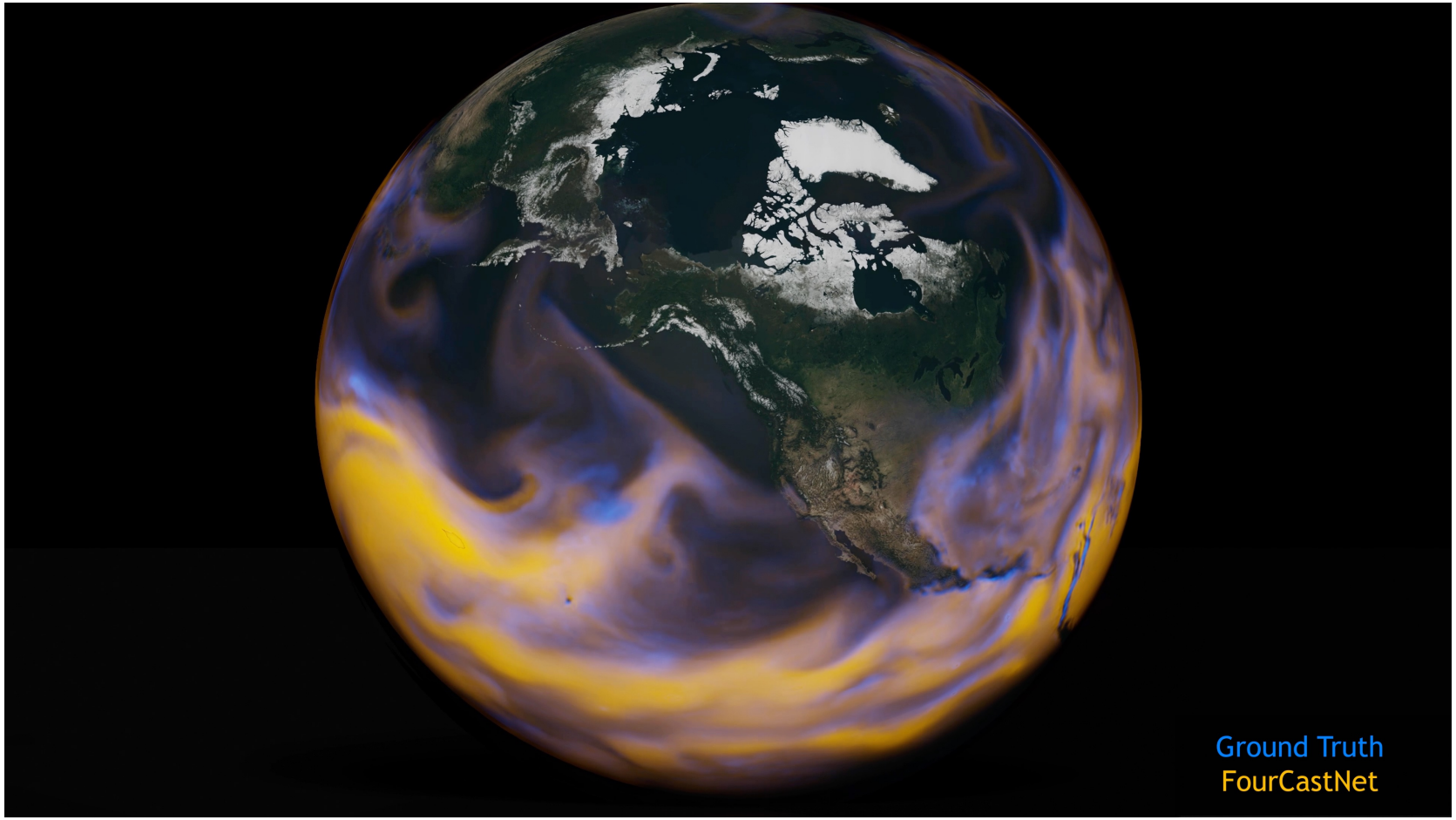
Scalable, data-driven global weather model surrogate enabling interactivity at scale

Pathak et al. "FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators." *arXiv:2202.11214* (2022).

Looking ahead

- Higher resolutions (9 km, 5 km, 1 km)
- Full state vector
- Physics constraints
- Observational ground truth / diagnostics
- Calibrated ensembles
- Weather to climate scales
- Data assimilation

Thank you



Pathak et al. "FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators." *arXiv:2202.11214* (2022).