

EVALUATING NEURAL OPERATOR ARCHITECTURES FOR DATACENTER DIGITAL TWIN SURROGATE MODELS

Presented by:

Sai Ranga Reddy Nukala
Shaila Reddy Kankanala
Sadhvi Singh
Akshith Reddy Jonnalagadda

Motivation & Problem Statement

Datacenters consume ~1.5% of global electricity (2024), projected to double by 2030 with rising AI workloads. A single hotspot can trigger emergency shutdowns and hardware damage.

CFD simulations (e.g. OpenFOAM) yield high-fidelity thermal predictions but take up to 15 hours per geometry - impractical for real-time management.

Neural surrogate models promise a 15,000× speedup, but the critical unanswered question is:

Which neural architecture should power the predictive engine of a Datacenter Digital Twin?

No prior study has evaluated FNO on NVIDIA's public PhysicsNeMo-Datacenter-CFD benchmark with comprehensive multi-field evaluation in physical units.

Dataset & Experimental Setup

960
CFD Simulations

7.37M
Grid Points / Sample

15 hrs
CFD Runtime

6
Models Benchmarked

PhysicsNeMo-Datacenter-CFD

- 960 structured CFD simulations, grid: 960×96×80
- 3 physical fields: T (°C), U (m/s), p (Pa)
- Input features: coordinates, wall distance, server geometry
- Split: 768 train / 192 test (fixed seed)

Hardware Platform

- NVIDIA GB10 (Grace Blackwell, 128 GB unified memory)
- Alienware RTX 5080 (prototyping)

Six Models Benchmarked

- Spatial** 3D U-Net, PI-U-Net
- Spectral** FNO, PI-FNO
- Transformer** Transolver
- Baseline** POD+MLP

Architecture Overview

✓ 3D U-Net (Recommended)

Params: 22.6M | Type: CNN

Key: Encoder-decoder with skip connections preserving fine-scale spatial detail on structured grids
FP16 support: Full (standard convolutions)
Inference: 390 ms FP32 → 236 ms FP16



3D FNO

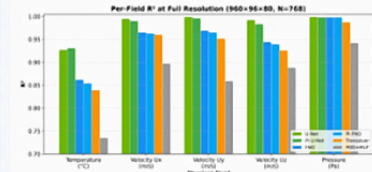
Params: 28.3M | Type: Neural Operator

Key: Global spectral convolutions via FFT; resolution-invariant in theory
FP16 support: cuFFT requires FP32 for non-power-of-2 dims
Inference: 680 ms FP32 only

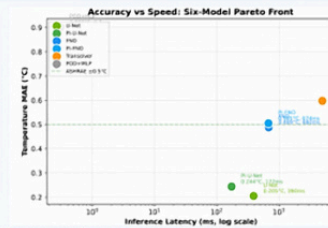


Results: Full-Resolution Per-Field Accuracy (All 6 Models)

Field	U-Net	PI-U-Net	FNO	PI-FNO	Transolver	POD+MLP
T (°C) MAE	0.205	0.244	0.489	0.505	0.598	0.902
T R ²	0.927	0.931	0.862	0.854	0.839	0.735
Ux (m/s) MAE	0.023	0.034	0.074	0.080	0.077	0.144
Uy (m/s) MAE	0.027	0.045	0.143	0.161	0.189	0.342
p (Pa) MAE	0.068	0.099	0.099	0.104	0.249	0.573
Latency (ms)	390	172	680	674	4972	0.3



Per-Field R² Bar Chart

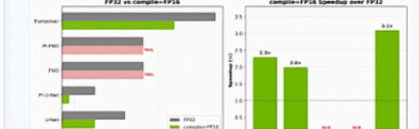


Accuracy-Speed Pareto Front

Key Finding: U-Net achieves 2.4× lower temperature MAE (0.20°C vs 0.49°C) and R² = 0.927 vs FNO's 0.862.

Key Discovery: Architecture Ranking Reversal

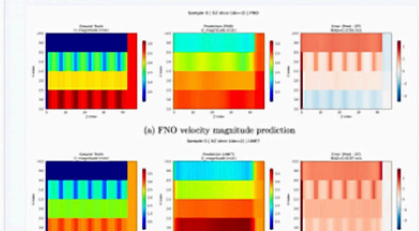
FNO outperforms U-Net at downsampled resolution (×20), but U-Net dominates at full resolution - across all fields and all training set sizes.



Inference Optimization & Deployment

PI-U-Net FP16	87 ms
U-Net FP16	222 ms ✓
U-Net FP32	408 ms
FNO FP32	680 ms
FNO FP16	FAIL

Temperature Prediction Visualization



References

- [1] IEA, Electricity 2024, Paris, 2024.
- [2] Yue et al., "FNO/U-Net/VIT on datacenter CFD," 2025 (private dataset).
- [3] Cao et al., "GNOT for datacenter thermal prediction," NeurIPS 2023.
- [4] Rahaman et al., "Spectral bias in neural networks," ICLR 2019.
- [5] OpenFOAM Foundation, OpenFOAM v11, 2023.
- [6] NVIDIA / Wistron, PhysicsNeMo Datacenter CFD, 2024.
- [7] Li et al., "Fourier Neural Operator," ICLR 2021.
- [8] Herde et al., "Poseidon / Transolver," ICLR 2024.

Interested in partnering with us on a capstone project or exploring other ways to collaborate with the College of IDS? Contact CIDS_CAP@sjsu.edu to discuss potential projects and partnership opportunities.