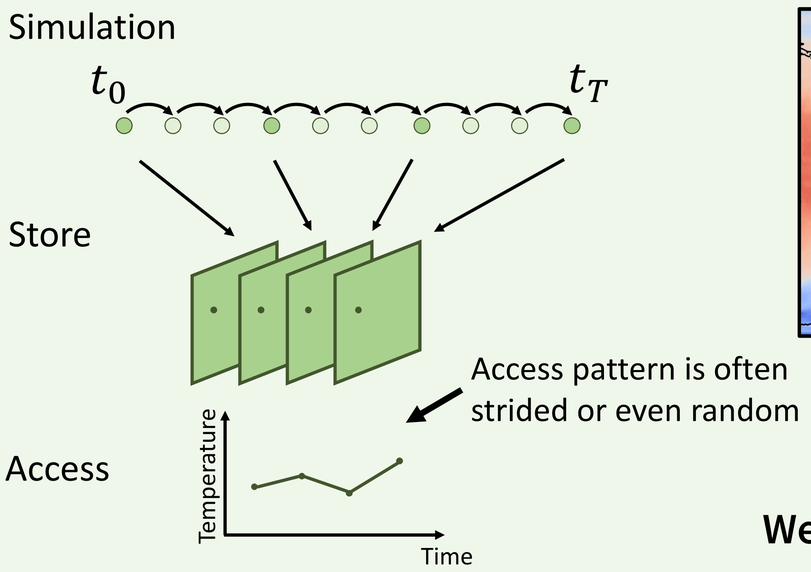
Langwen Huang
Torsten Hoefler (advisor)

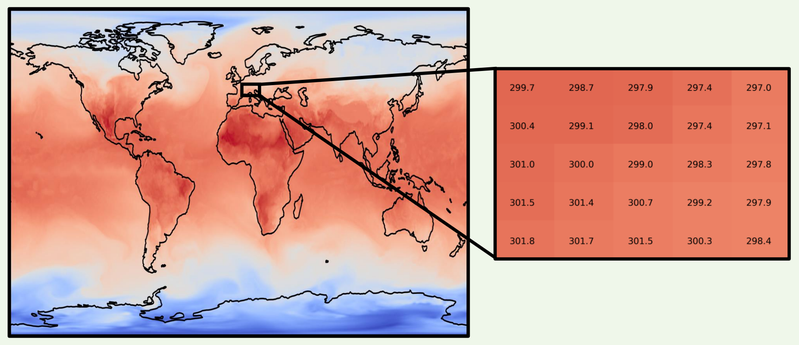
Compressing Multidimensional Weather and Climate Data into Neural Networks

Challenge: Storing and Accessing Numerical Simulation Data

Production & Consumption of Simulation Data

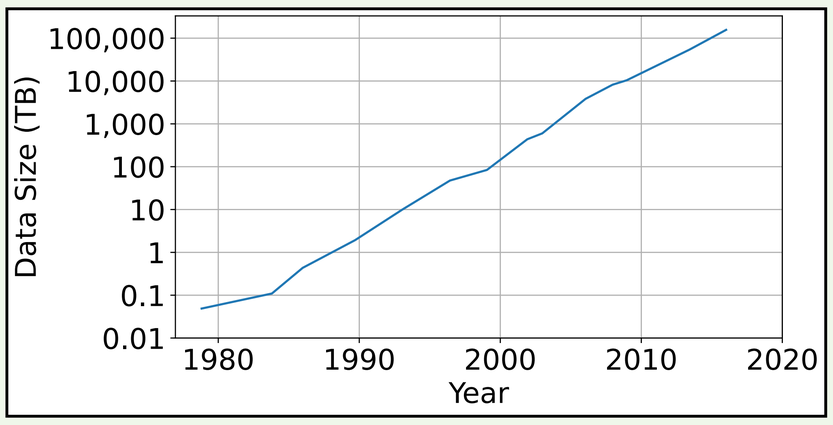


Typical Weather Data



Neighboring values are close to each other
$\Longrightarrow$ Lots of redundant information
$\Longrightarrow$ More redundancy in higher resolution data

Explosion of Data in the Weather Center



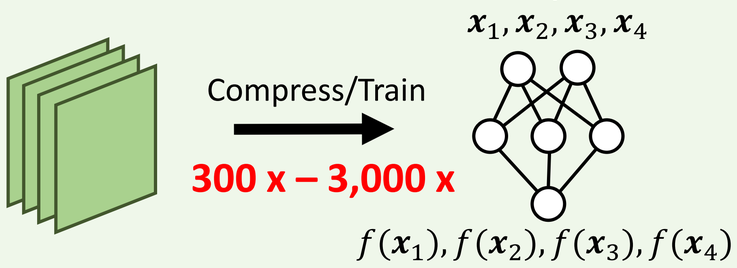
The data archive in ECMWF is growing **exponentially! [1]**

We tackle the issues by compressing weather & climatedata 300 x – 3,000 x !

Solution: Compression with a Neural Representation Approach

General Idea

Multidimensional Data → Compress/Train → **Neural Representation**



- ✅ High compression ratio
- ✅ Decompress on-demand
- ✅ Mesh-free
- ✅ GPU friendly
- ❌ Lossy compression
- ❌ Slow training process

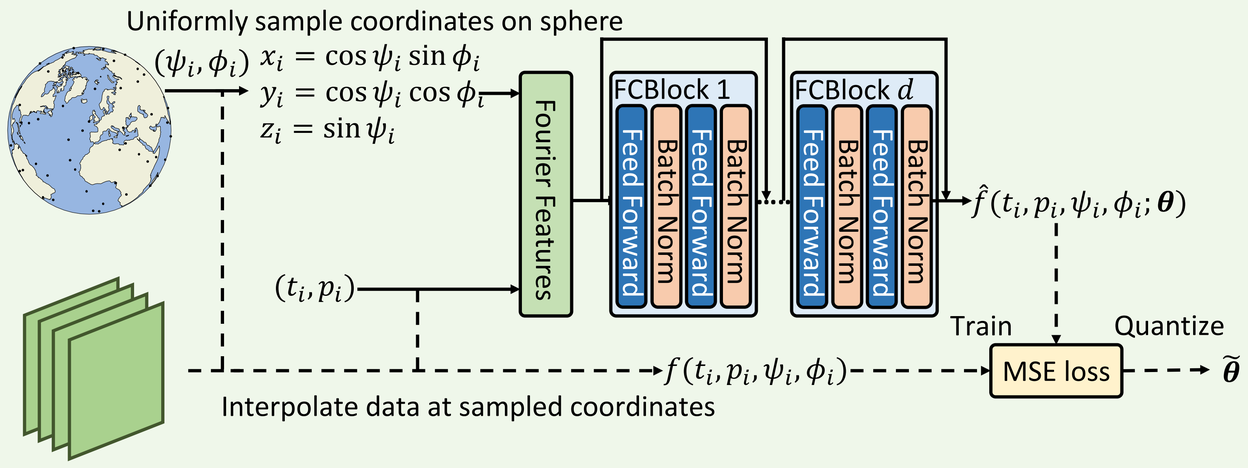
Properties of Data

- Continuous & smooth in every dimensions
- Stratified: variations between levels are much larger than inside levels
- Defined on a sphere
- Random access is preferred

Our Treatment

- Smooth activation function (GELU)
- Level-wise de-normalization
- Uniformly sampling on the sphere
- $(\psi, \phi) \rightarrow (x, y, z)$
- Random access does not lead to overhead by construction

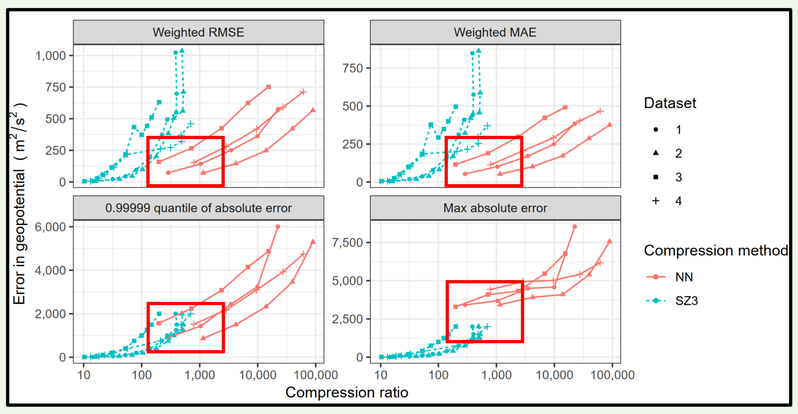
Neural Network Structure



Method	Compression Ratio	Compression Speed	Continuous Decomp.	Random Decomp.
SimFS [2]	Arbitrary	▶▶▶	▶▶	▶
ZFP [3]	< 10 x	▶▶▶	▶▶▶	▶▶▶
TTHRESH [4]	< 300 x	▶▶	▶▶▶	▶▶
SZ3 [5]	< 400 x	▶	▶▶▶	▶▶
NN (Ours)	**300 x – 3,000 x**	▶	▶▶▶	▶▶▶

Evaluation

Compression Errors vs Compression Ratios

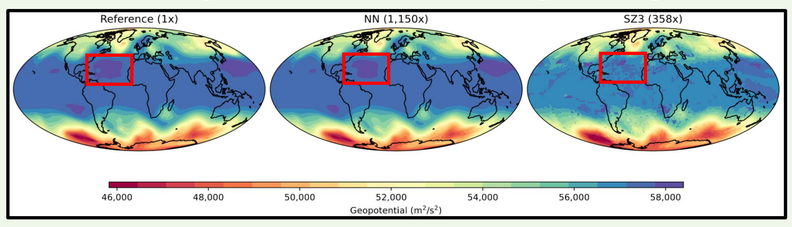


Dataset	Size	#Levels	Spatial Res.
1	3.9 GB	11	0.5°
2	15.6 GB	11	0.25°
3	2.7 GB	1	5.625°
4	10.7 GB	1	2.8125°

- Much lower RMSE & MAE at CR 300 x – 3,000 x
- Higher max abs. error but only occurring seldomly (<0.001%)
- Free C.R. when increase spatial res.!

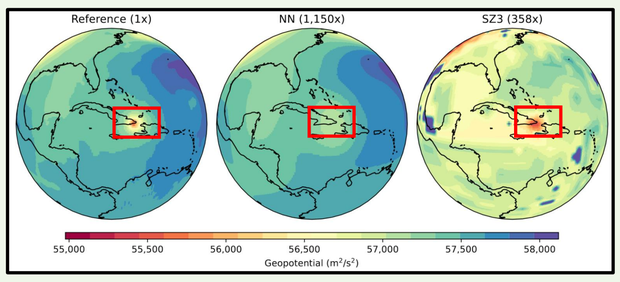
Case Study: 5th Oct, 2016 (Hurricane Matthew)

Geopotential at 500 hPa (Global)



- Our method well preserves general shape and average value while 1,150 x smaller.
- SZ3 introduces many artifacts that breaks important weather structure, it also badly preserves the average value.

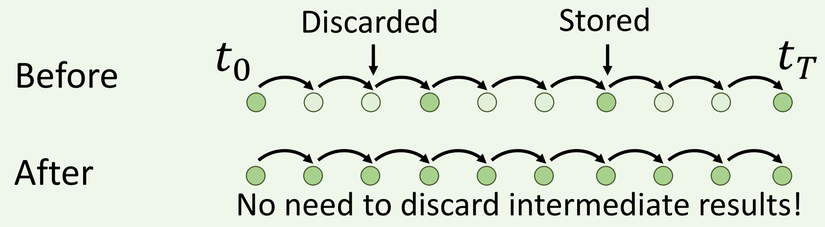
Geopotential at 500 hPa (Regional)



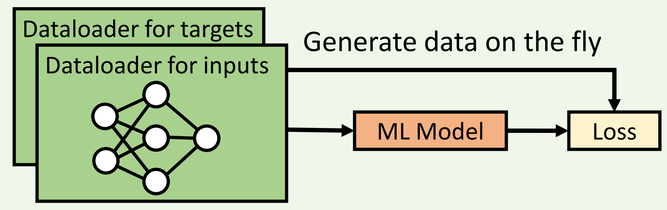
- Our method may hardly capture the extreme values in a small area like a hurricane center

Applications

Store More Simulation Data



Compressed Dataloader for Training ML Models



Experiment: replacing dataloader when training a CNN

Dataset 3	C.R.	Weighted RMSE error (test set)	
		Z at 500 hPa (m^2/s^2)	T at 850 hPa (K)
Original	1 x	632.9	2.906
NN Compressed	**198 x**	**637.3 (+0.7%)**	**2.944 (+1.3%)**
SZ3 Compressed	71 x	650.6 (+2.8%)	2.985 (+2.7%)
Dataset 4			
Original	1 x	688.8	2.834
NN Compressed	**790 x**	**697.3 (+1.2%)**	**2.888 (+1.9%)**
SZ3 Compressed	106 x	702.9 (+2.0%)	2.887 (+1.9%)

Reference

[1] “ECMWF's Vision for Big Data, AI and Cloud Computing,” 2019.
[2] Girolamo, Salvatore Di, P. Schmid, Thomas Schulthess, and Torsten Hoefler. ‘SimFS: A Simulation Data Virtualizing File System Interface’. In *IPDPS’19*. Rio de Janeiro, Brazil: IEEE, 2019.
[3] Lindstrom, Peter. ‘Fixed-Rate Compressed Floating-Point Arrays’. *IEEE Transactions on Visualization and Computer Graphics* 20, no. 12 (2014): 2674–83.
[4] Ballester-Ripoll, Rafael, Peter Lindstrom, and Renato Pajarola. ‘TTHRESH: Tensor Compression for Multidimensional Visual Data’. *IEEE Transactions on Visualization and Computer Graphics* 26, no. 9 (2019): 2891–2903.
[5] Liang, Xin, Kai Zhao, Sheng Di, Sihuan Li, Robert Underwood, Ali M. Gok, Jiannan Tian, et al. ‘SZ3: A Modular Framework for Composing Prediction-Based Error-Bounded Lossy Compressors’. *IEEE Transactions on Big Data*, 2022.

ETH zürich

Arxiv

GitHub repo

PASC 23 Davos Switzerland 26-28 June 2023