Advancing Scientific Data Compression via Cross-Field Prediction

Youyuan Liu[1], Wenqi Jia[2], Taolue Yang[1], Bo Jiang[1], Miao Yin[2], Sian Jin[1]

1. Temple University 2. University of Texas, Arlington

Introduction: High-Volume Data in Scientific Computing

Petabyte-Scale Storage and I/O Needs

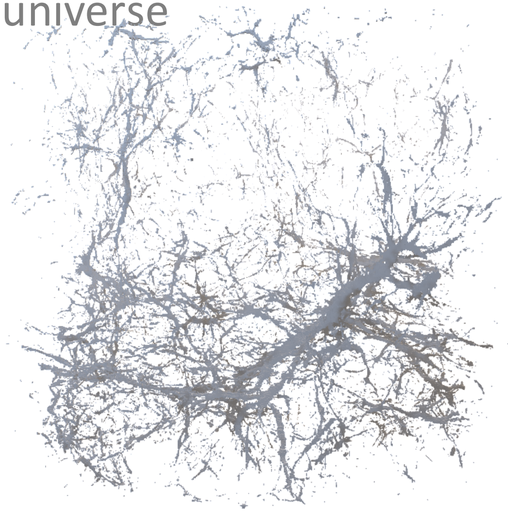
- Large-scale scientific applications generate extremely large amounts of data
- Limited storage capacity (even for large-scale parallel computers)
- The I/O bandwidth can create bottlenecks in the transmission

A Typical Scientific Simulation?

Nyx Cosmological Simulation

- Adaptive mesh, hydrodynamics code designed to model astrophysical reacting flows

- Simulate the universe and compare with our observable universe



Dark matter density of a Nyx cosmological simulation data
↑

Introduction: High-Volume Data in Scientific Computing

Petabyte-Scale Storage and I/O Needs

Application	Data scale	Passive solution	To reduce
Nyx Cosmology simulation	**2.8 TB** per snapshot	**34 GB/S** on Hopper@NERSC	**10x-100x** In need
CESM Climate simulation	20% vs **50%** of h/w budget for storage 2013 vs 2017	**5h30m** to store NSF Blue Waters, I/O at 1TBps	**10x** In need
HACC Cosmology simulation	**20PB** per one-trillian particle simulation	**use up FS** 26 PB for Mira@ANL	**10x** In need

Introduction: Optimization with Error-Bounded Lossy Compression

What is lossy compression

- Reduce data size by approximating values while allowing controlled errors
- Maintains data quality within error bounds for scientific analysis

Why lossy compression

- Significantly higher compression ratio
- Introduced error can by controlled
  - ensuring that the impact on simulation results remains minimal and within a predefined error bound

Introduction: Existing lossy compressors

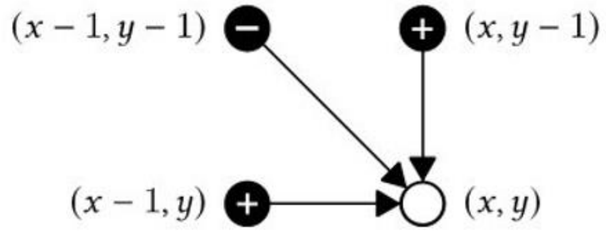
Using intra-field information to compress

SZ

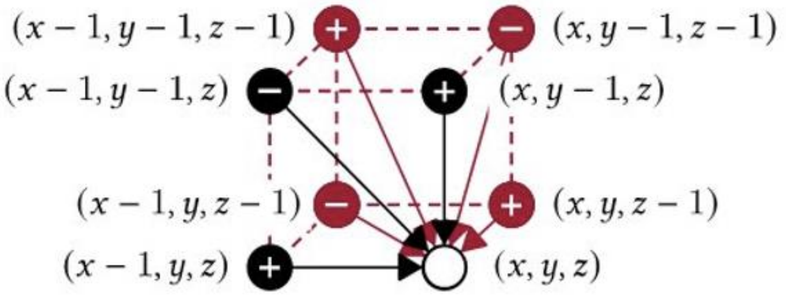
One of the most widely used prediction-based lossy compressor

- predict current value from neighbors, encode residuals
- Supports multiple predictors (e.g., Lorenzo, Interpolation)
- Effective when local correlation is strong

2D Lorenzo Predictor



3D Lorenzo Predictor

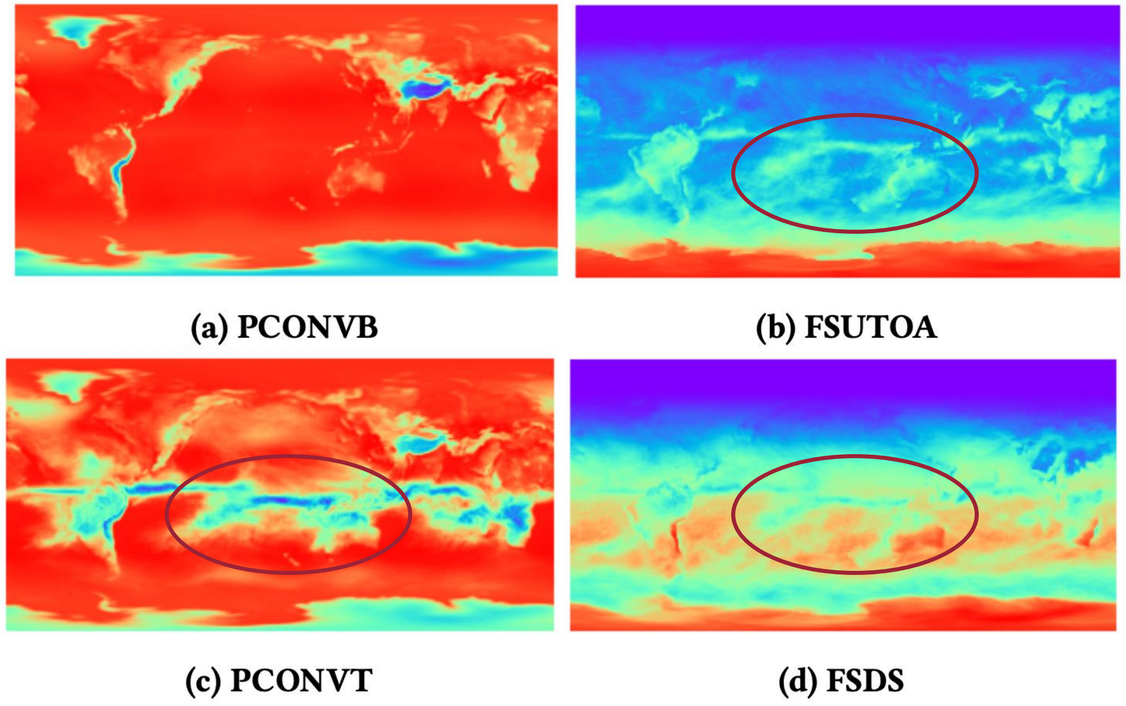


Lorenzo predictor used in SZ ↑

Introduction: Cross-Field Correlation

Missing information

- Existing compressors (e.g., SZ, ZFP) primarily exploit **intra-field information** within a single data field
- They overlook a key characteristic: Scientific data often consists of multiple, physically **inter-correlated fields**



(a) PCONVB (b) FSUTOA

(c) PCONVT (d) FSDS

4 fields in CESM-ATM dataset ↑

Our solution & Contributions

Learning-based Cross-field Prediction

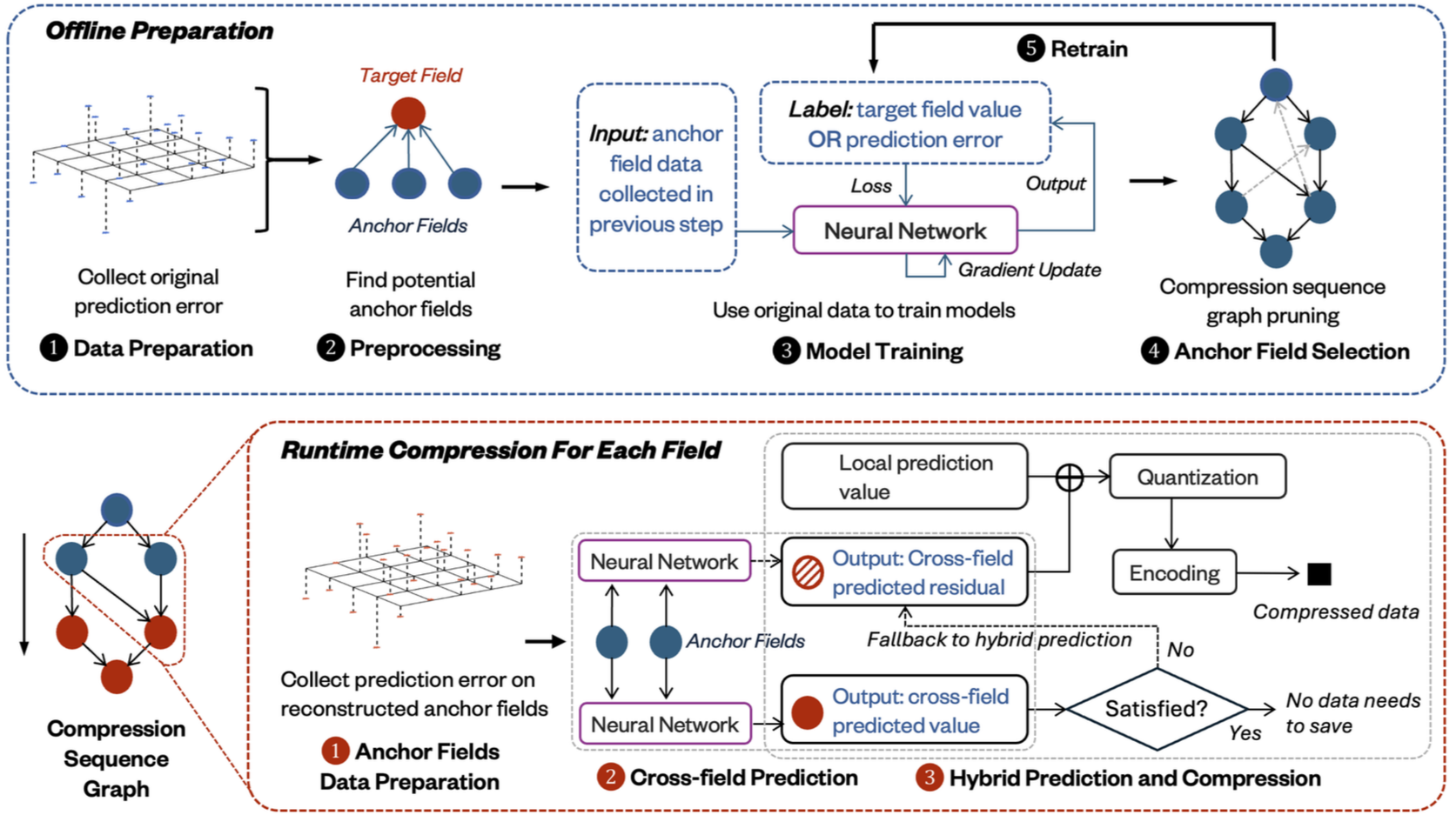
- **Cross-Field Framework:** Propose the first framework to leverage cross-field correlations
- **Automated Anchor Selection:** Design an intelligent algorithm to select optimal predictors and resolve dependencies
- **Hybrid Prediction Engine:** A flexible engine that can either enhance existing compressors (like SZ3) or fully reconstruct fields for extreme compression ratios
- **Detailed Evaluation:** Achieve up to **19.3%** overall and **103.4%** single-field compression ratio improvements on real-world datasets

Background: Boosting Learning

- **What is Boosting?** An ensemble learning technique that **combines multiple simple models** ("weak learners") to create a single, powerful model
- **How does it work?** It works through **iterative error correction**: each new model in the sequence is trained to fix the prediction errors made by the previous ones
- **Relevance to Data Compression:** While originally for classification, this principle can be adapted for prediction. One could use a secondary predictor (like an NN) to **learn and correct the residual errors** from a primary predictor (like traditional predictor)
- **Key Advantage:** This two-stage approach is promising because predicting residual is often a **simpler and more efficient** task for a model than predicting the original, complex data values

Design: A Two-Stage Framework

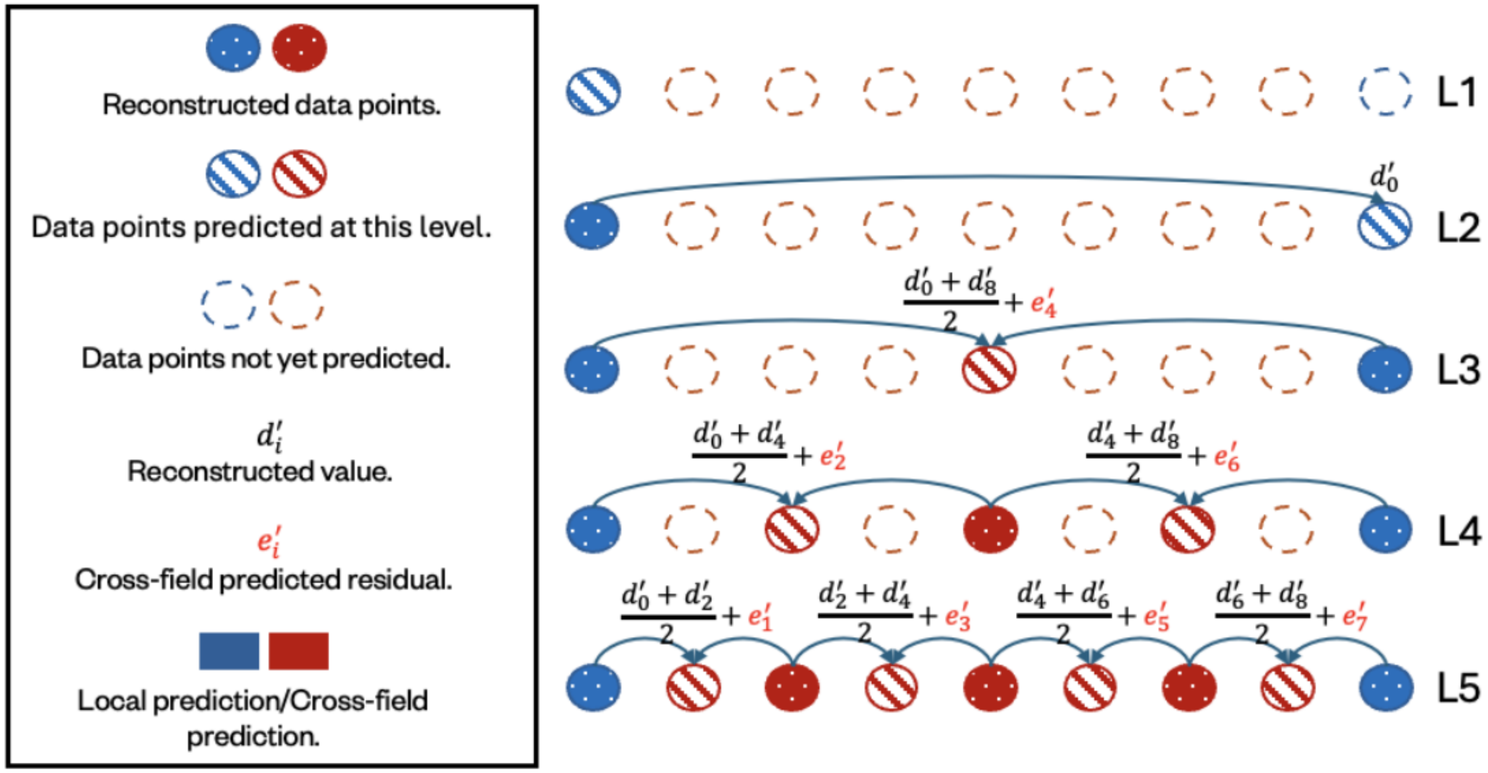
An overview of our offline preparation and runtime compression pipeline.



Design: Mode 1 - Error-Bounded Hybrid Compression

Enhancing existing compressors by predicting the error with cross-field info

- **Goal:** Improve the compression ratio of compressors like SZ3 while respecting strict error bounds
- **Method:** uses ANCHOR FIELDS to predict the **residual error** of the local predictor (e.g., interpolation)



How we enhance the interpolation prediction

Design: Mode 2 - Fully Cross-field Prediction

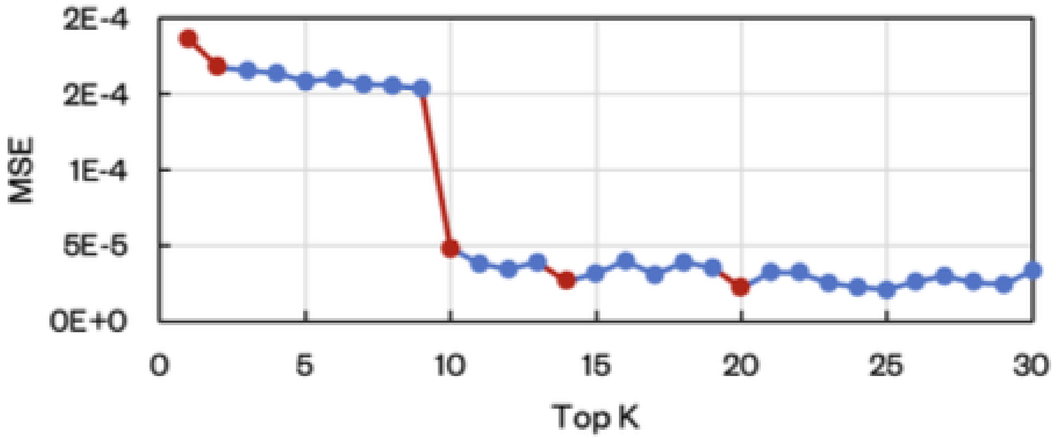
Achieving extreme compression by reconstructing fields entirely from anchors.

- **Goal**: highest possible compression ratio under scenarios where strict error bounds are not essential (e.g., visualization)
- **Method**: Reconstructs the target field entirely from its anchor fields
- **Two-Stage NN**: It uses two networks: an initial predictor generates a first estimate, and a second U-Net-like model corrects the residual error
- **Result**: Theoretically infinite compression ratio, limited only by model size

Design: Automated Anchor Field Selection

Finding the most informative predictors without expert knowledge.

- **Challenge**
  - Exhaustive anchor search is too costly
  - Manual selection needs domain knowledge
- **Our Solution: A Dual-Sorting Process**.
  - **Stage 1 (Pre-sort):** Pearson correlation-based sorting
  - **Stage 2 (Refine):** Use a lightweight NN to evaluate the top-K candidates, selecting those that cause the largest drop in Mean Squared Error (MSE)



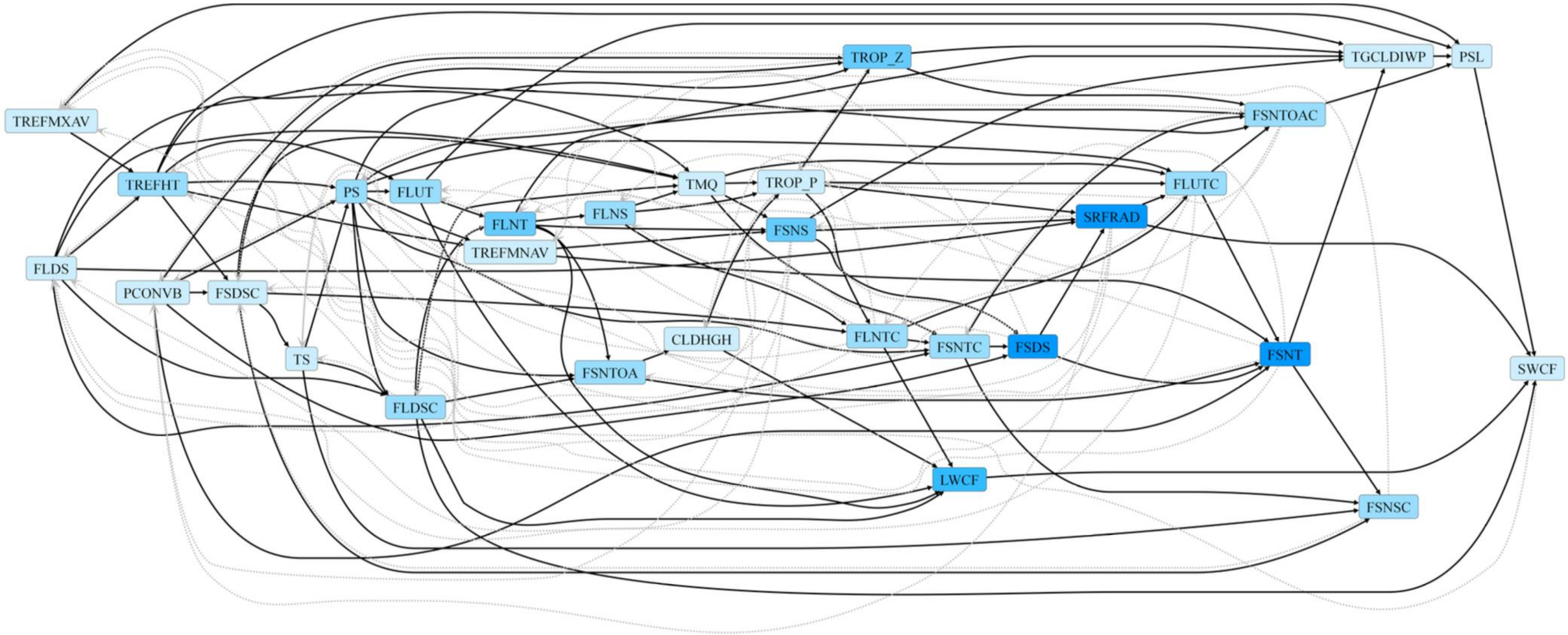
Red line highlights the significant drops

Design: Resolving Circular Dependencies

Creating a valid compression order from the dependency graph.

- **Problem:** Anchor selection creates a directed graph that may contain cycles (e.g., A predicts B, B predicts A), causing deadlocks
- **Step 1 (Detect Cycles):** We use **Tarjan's algorithm** to efficiently find all Strongly Connected Components (SCCs), which are guaranteed to contain any and all cycles in the graph
- **Step 2 (Break Cycles):** Within each SCC, a **greedy algorithm** prunes the least important edges to break the cycles, transforming the graph into a Directed Acyclic Graph (DAG)
- **Result:** A valid, topological order for compression

Design: Resolving Circular Dependencies



A sub compression graph of CESM-ATM

Design: Our Neural Network Models

Network structure

- **For Error-Bounded Mode:**
  - A lightweight **CNN with several residual blocks**
  - Kernel size is set to 5 to **align with the SZ3 cubic interpolation predictor**
- **For Fully Cross-Field Mode:**
  - A two-stage design inspired by boosting learning
    - **Initial Predictor:** A CNN-based model for the first estimate
    - **Residual Model:** A **U-Net-like architecture** to capture global context and correct the initial prediction's error

Evaluation: Experiment Setup

- **Baseline:** We compare against **SZ3 with its interpolation predictor**, a state-of-the-art, widely-used lossy compressor
- **Metrics:** Performance is measured by:
  - Compression Ratio (CR)
  - Data Quality (PSNR)
- **Hardware**
- 2x Intel Xeon E5-2620v4 with 4xV100

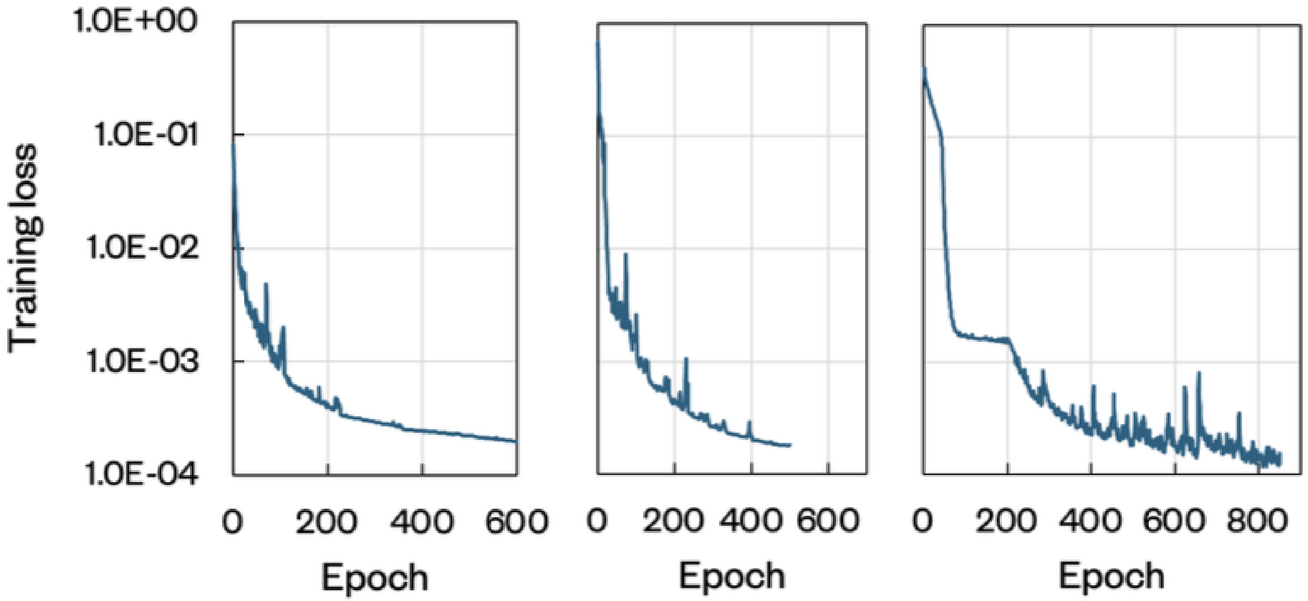
Name	Dims	Fields	Description
Nyx	512x512x512	6	Cosmology simulation
CESM-ATM	1800x3600	79	Climate simulation

Datasets used

Evaluation: Model Training & Convergence

Verifying that the neural network models learn effectively.

- **Model 1:** Predicts residual error in the error-bounded mode
- **Models 2 & 3:** Work as a two-stage predictor (initial prediction + residual correction) in the fully cross-field mode

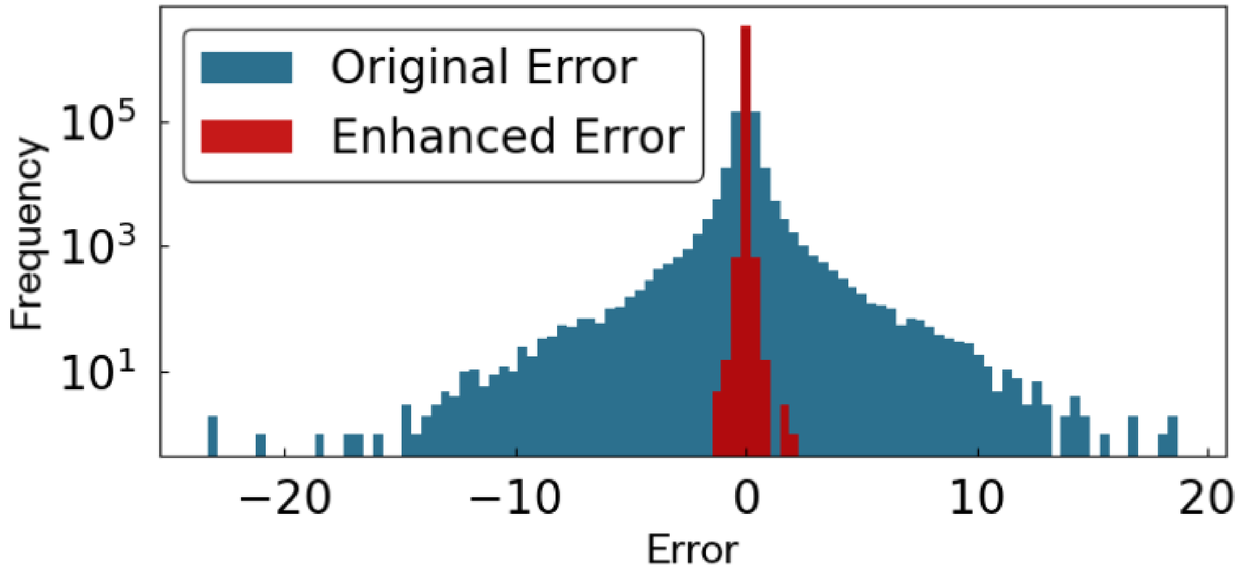


(a) Model 1. (b) Model 2. (c) Model 3.

Evaluation: Improved Error Distribution at Similar Compression Ratio

Our method concentrates the prediction error closer to zero.

- Enhanced prediction reduces the error value range
- Error distribution becomes more concentrated, aiding entropy encoding



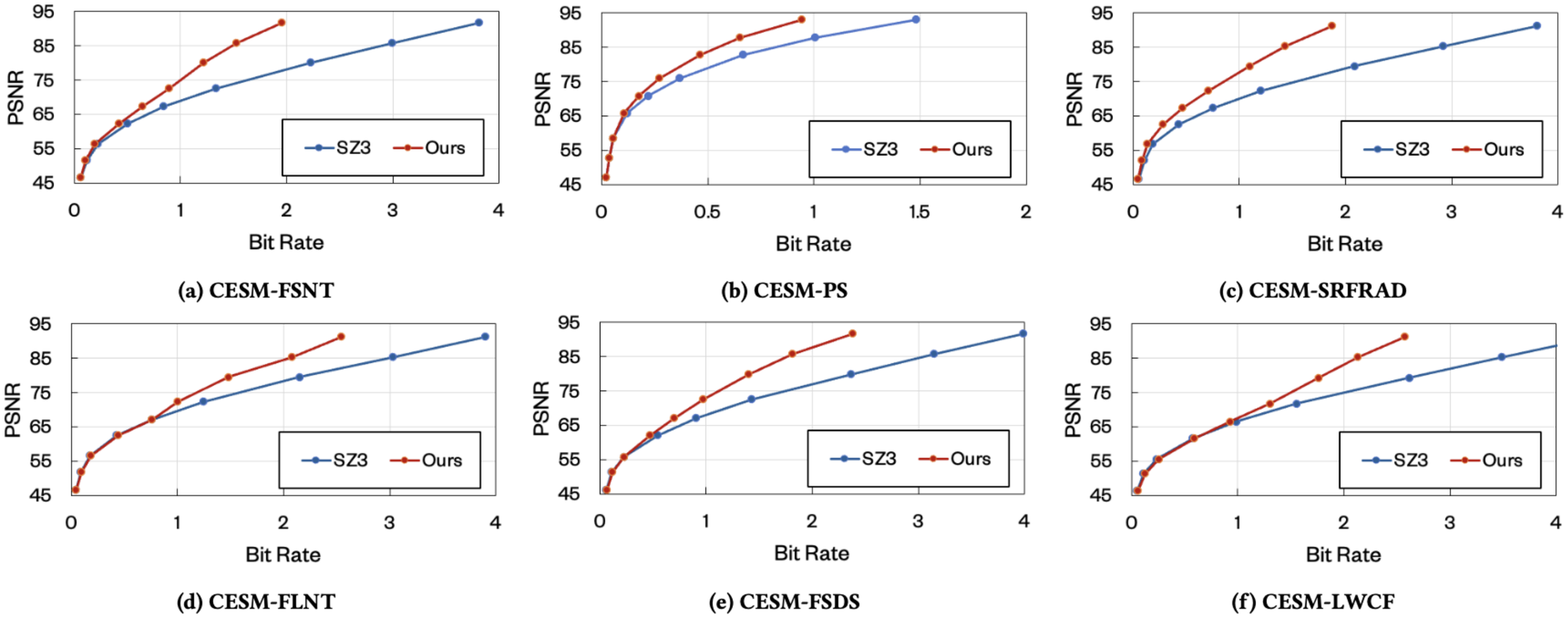
The distribution of prediction error of FSNT field in CESM-ATM

Evaluation: Detailed Result Using Mode 1

Improvement on single field

Error Bound			2E-3	1E-3	5E-4	2E-4	1E-4	5E-5
SZ3	CESM-ATM	PS	265.36	143.96	86.58	47.85	31.79	21.58
		FSNT	63.51	37.87	23.99	14.38	10.69	8.38
		SRFRAD	74.21	42.58	26.53	15.34	10.96	8.4
		FSDS	58.81	35.23	22.4	13.52	10.16	8.02
		FLNT	74.56	42	25.66	14.83	10.56	8.2
		LWCF	55.25	32.43	20.53	12.22	9.16	7.29
Ours	CESM-ATM	PS	295.34(+11.3%)	178.78(+24.19%)	117(+35.14%)	69.01(+44.22%)	48.98(+54.07%)	33.89(+57.04%)
		FSNT	76.05(+19.74%)	49.72(+31.29%)	35.94(+49.81%)	26.23(+82.41%)	20.92(+95.7%)	16.36(+95.23%)
		SRFRAD	112.95(+52.2%)	68.38(+60.59%)	45.03(+69.73%)	29.03(+89.24%)	22.3(+103.47%)	17.07(+103.2%)
		FSDS	68.11(+15.81%)	45.38(+28.81%)	32.81(+46.47%)	22.86(+69.08%)	17.66(+73.82%)	13.44(+67.58%)
		FLNT	72.13(-3.26%)	42.25(+0.60%)	31.84(+24.08%)	21.54(+45.25%)	15.38(+45.64%)	12.57(+53.29%)
		LWCF	53.96(-2.33%)	34.31(+5.80%)	24.44(+19.05%)	18.11(+48.20%)	15(+63.76%)	12.42(+70.37%)

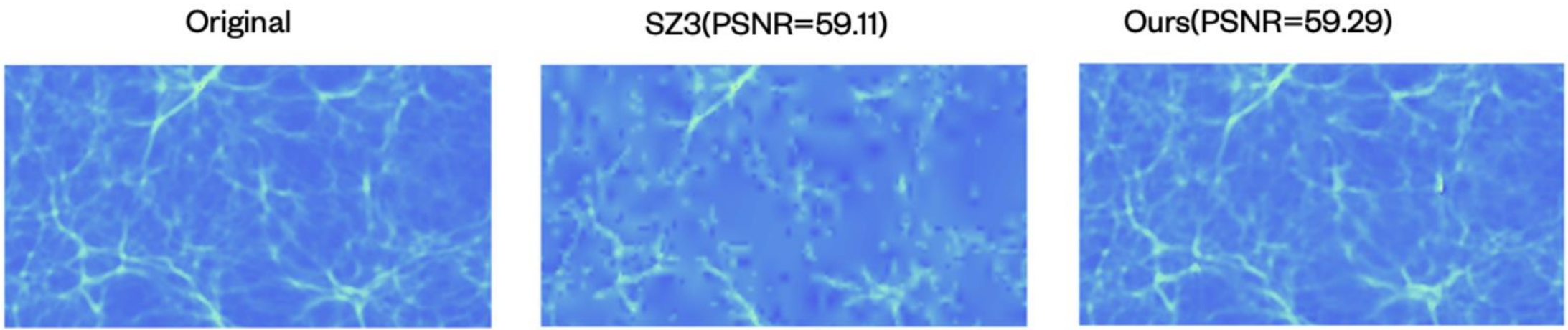
Evaluation: Rate-Distortion Comparison Under Mode 1



Rate-Distortion comparison of 6 fields selected from CESM-ATM

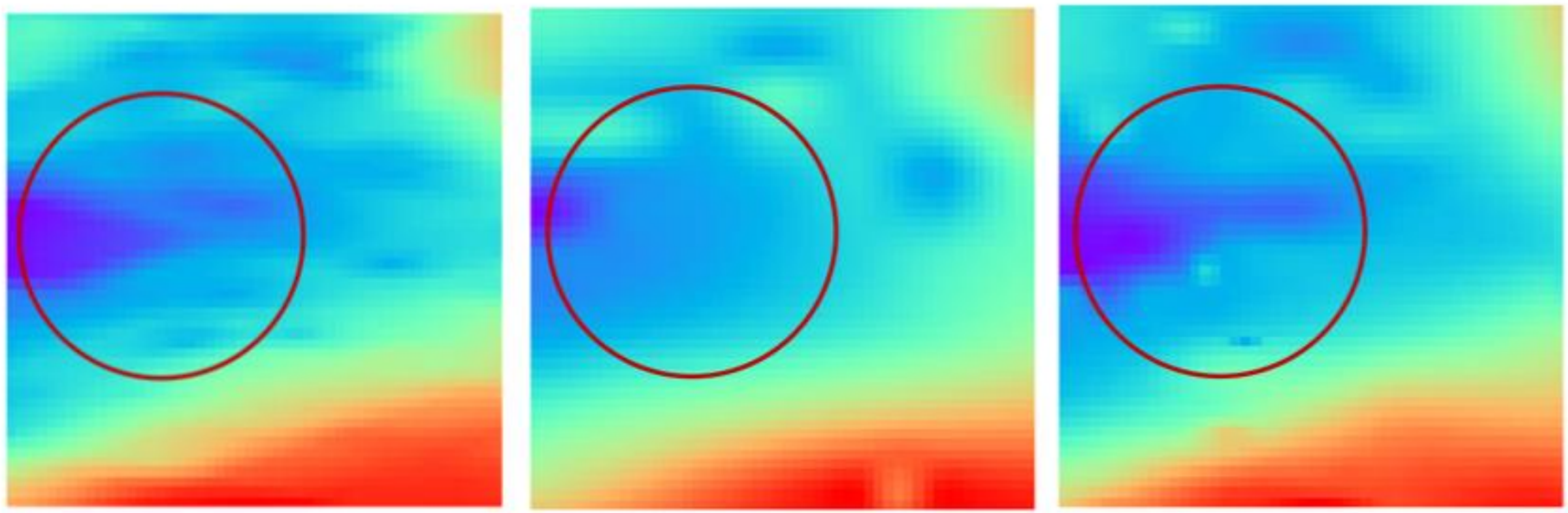
Evaluation: Enhanced Visual Quality

Preserving fine-grained details and reducing compression artifacts.



Evaluation: Enhanced Visual Quality

Preserving fine-grained details and reducing compression artifacts.



(a) Original

(b) SZ3
(CR=112.93, PSNR=59.55)

(c) Our Framework
(CR=112.95, PSNR=63.23)

Evaluation: Overall Performance on Full Datasets

Error Bound		2E-2	1E-2	5E-3	2E-3	1E-3	5E-4	2E-4	1E-4	5E-5
SZ3	CESM-ATM	630.35	353.15	198.13	98.98	61.79	40.61	24.96	18.23	13.91
	Nyx	116.77	69.59	45.61	27.95	19.91	14.54	9.97	7.79	6.33
Ours	CESM-ATM	752.01(+19.3%)	374.99(+6.19%)	209.22(+5.60%)	104.69(+5.76%)	66.07(+6.94%)	44.15(+8.73%)	27.58(+10.48%)	20.15(+10.51%)	15.34(+10.31%)
	Nyx	129.97(+11.3%)	79.18(+13.78%)	46.19(+1.3%)	27.95	19.91	14.54	9.97	7.79	6.33

Conclusion

- **Propose a novel framework** that leverages overlooked cross-field correlations to enhance compression
- **Design a fully automated** anchor selection method to make the framework practical and robust
- **Extend and validate the framework** across a broader range of scientific applications and datasets
- **Provide** up to **19.3% overall** compression ratio improvement and superior visual quality

Future work

- **Explore more architectures** to further improve both compression ratio and throughput
- **Extend the framework** across a broader range of scientific applications and datasets

Thank you!

Any questions are welcome!

Contact: Youyuan Liu youyuan.liu@temple.edu