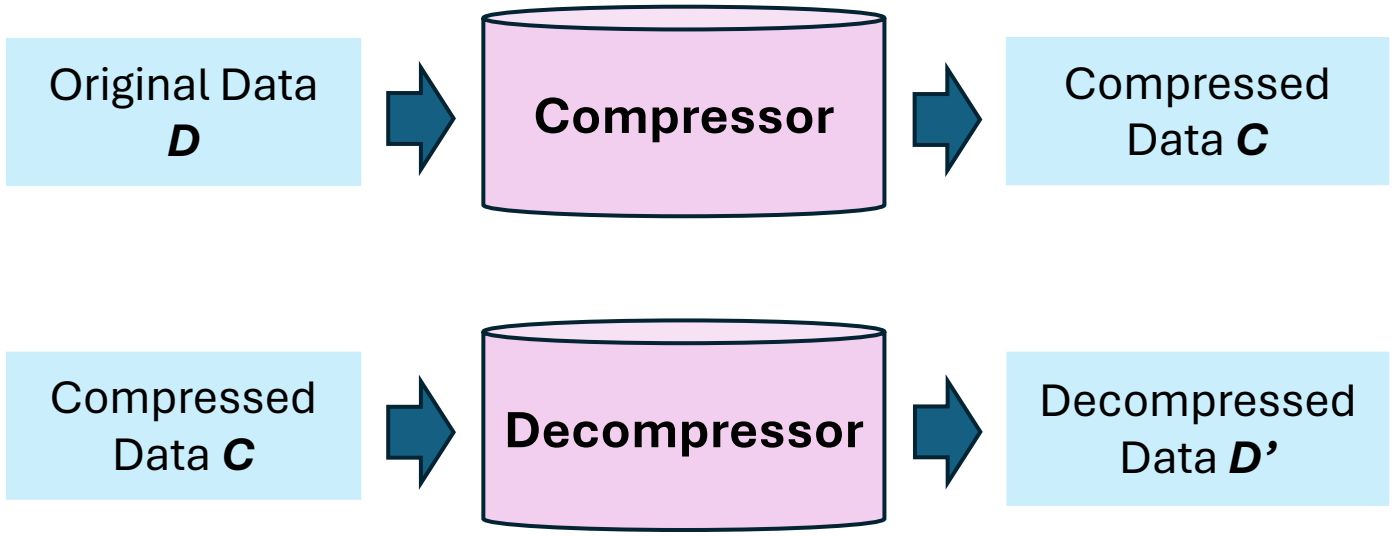
A Generic and Efficient Framework for Estimating Lossy Compressibility of Scientific Data

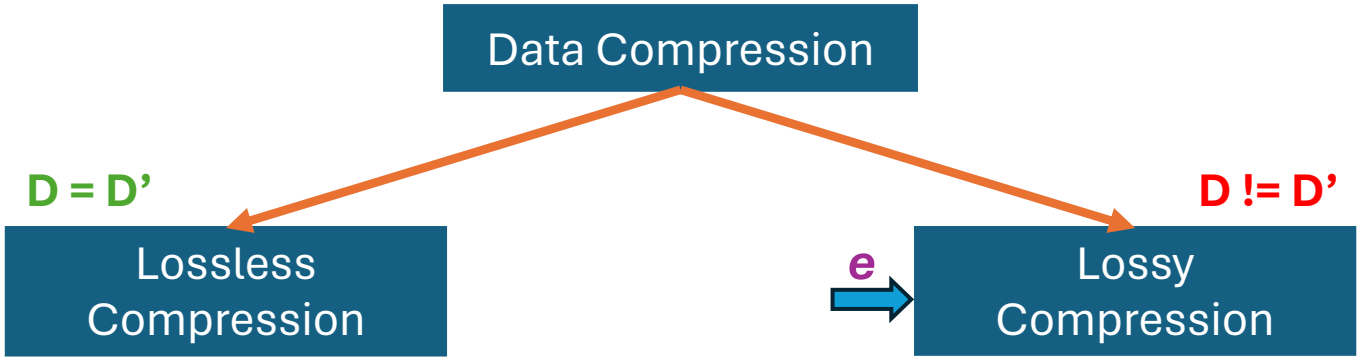
Md Hasanur Rahman, Sheng Di, Guanpeng Li, Franck Cappello

Outline

- Introduction
- Motivation
- Observations
- Methodology
- Evaluation
- Conclusion

Lossy Compression





Why Lossy Compression

- HPC applications produce vast volume of scientific data
- CESM application may produce 2.5PB of data
- User account generally has limited storage
- Lossless compression offers only ~2x compression ratio
- Lossy compression can reach ratio of 10x or higher

Metric

- Compression Ratio = Size of D / size of C
- PSNR: higher indicates D' being close to D

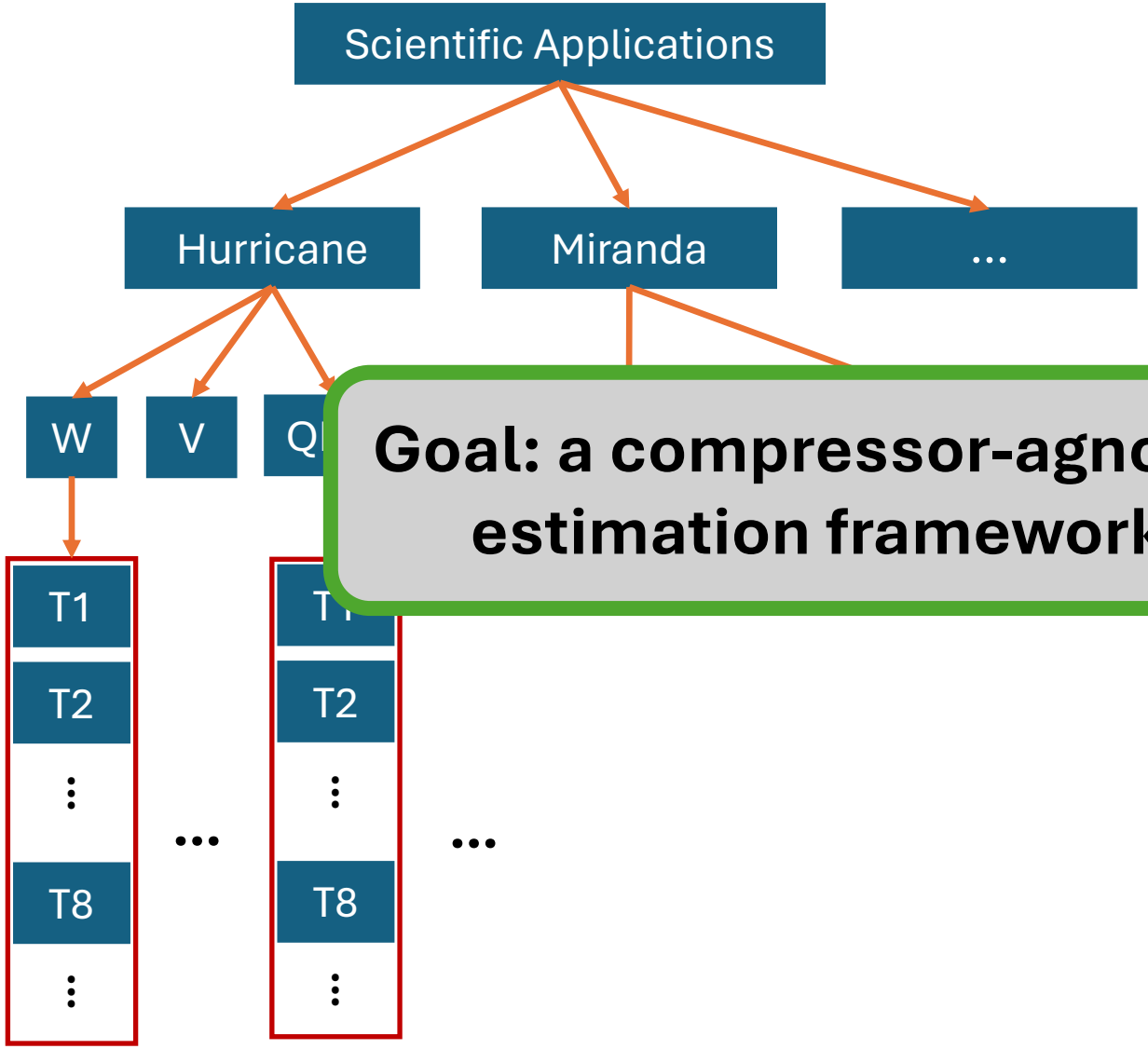
Problem & Motivation

- Lossy compressors are mainly designed on error-bound mode
- Compression ratio (compressed data size) is unknown until compression
- Compression ratio needs to be known beforehand
  - To fit
  - To de

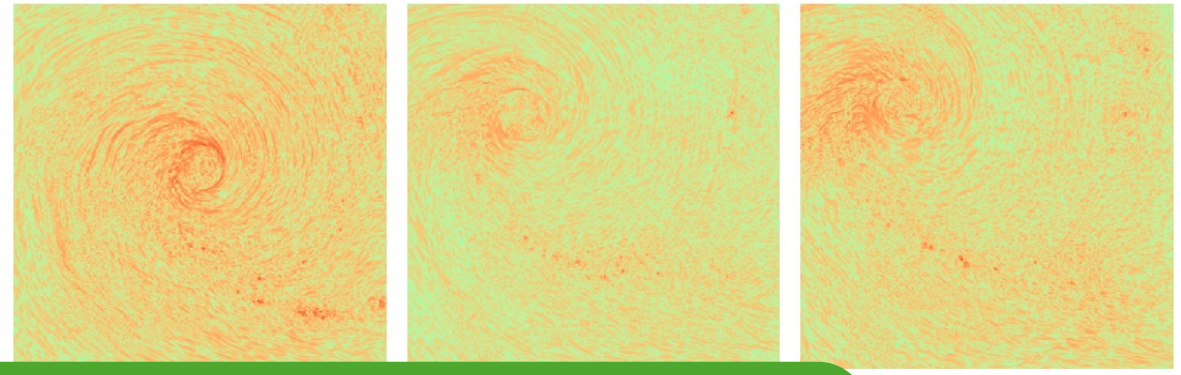
Goal: a compressor-agnostic lossy compressibility estimation framework, XTIMATE

- Existing studies [28,29,30] do not provide
  - Compressor agnostic solution
  - Accurate estimation capability under different compression configurations

Research Challenges



Goal: a compressor-agnostic lossy compressibility estimation framework across different fields



pression ratio: 128.19
(c) W timestep 48
f Same Field.

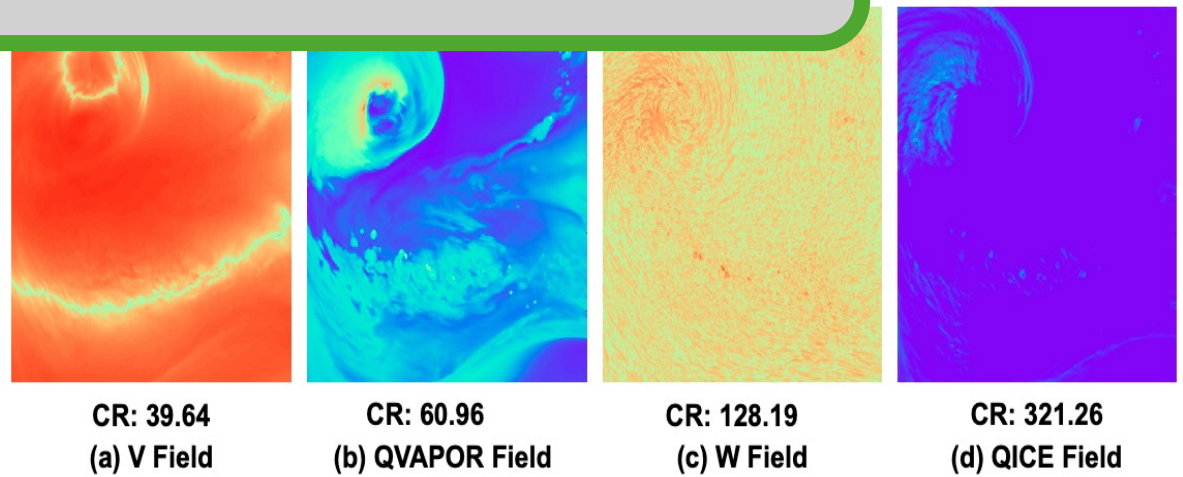


Figure: Illustrating Different Field Datasets.

Key Observations

- Scientific datasets across different fields exhibit various form of data textures
- Different attributes of data textures have notable impact on data compressibility
- Characterizing these data textures is novel
- Sobel kernel filter-based convolution is used to characterize data textures

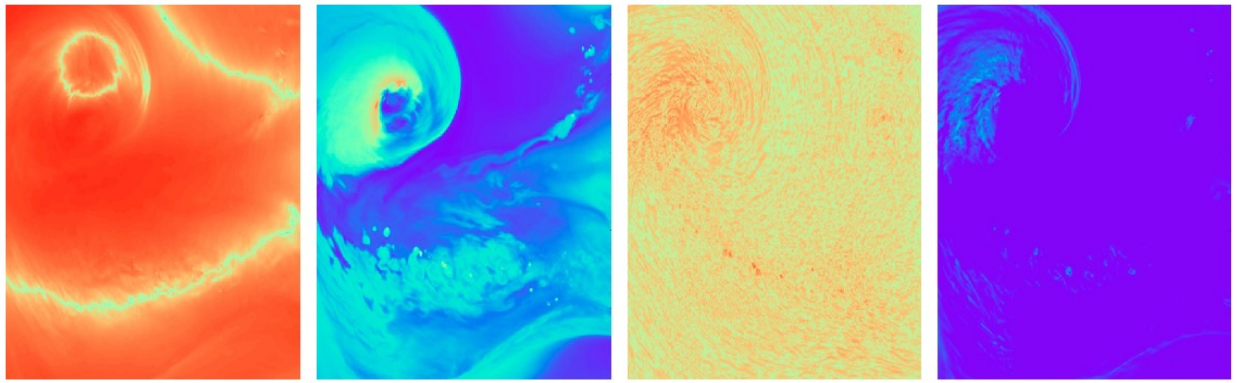


Figure: Illustrating Different Datasets Fields That Have Different Form of Data Textures.

XTIMATE Design Overview

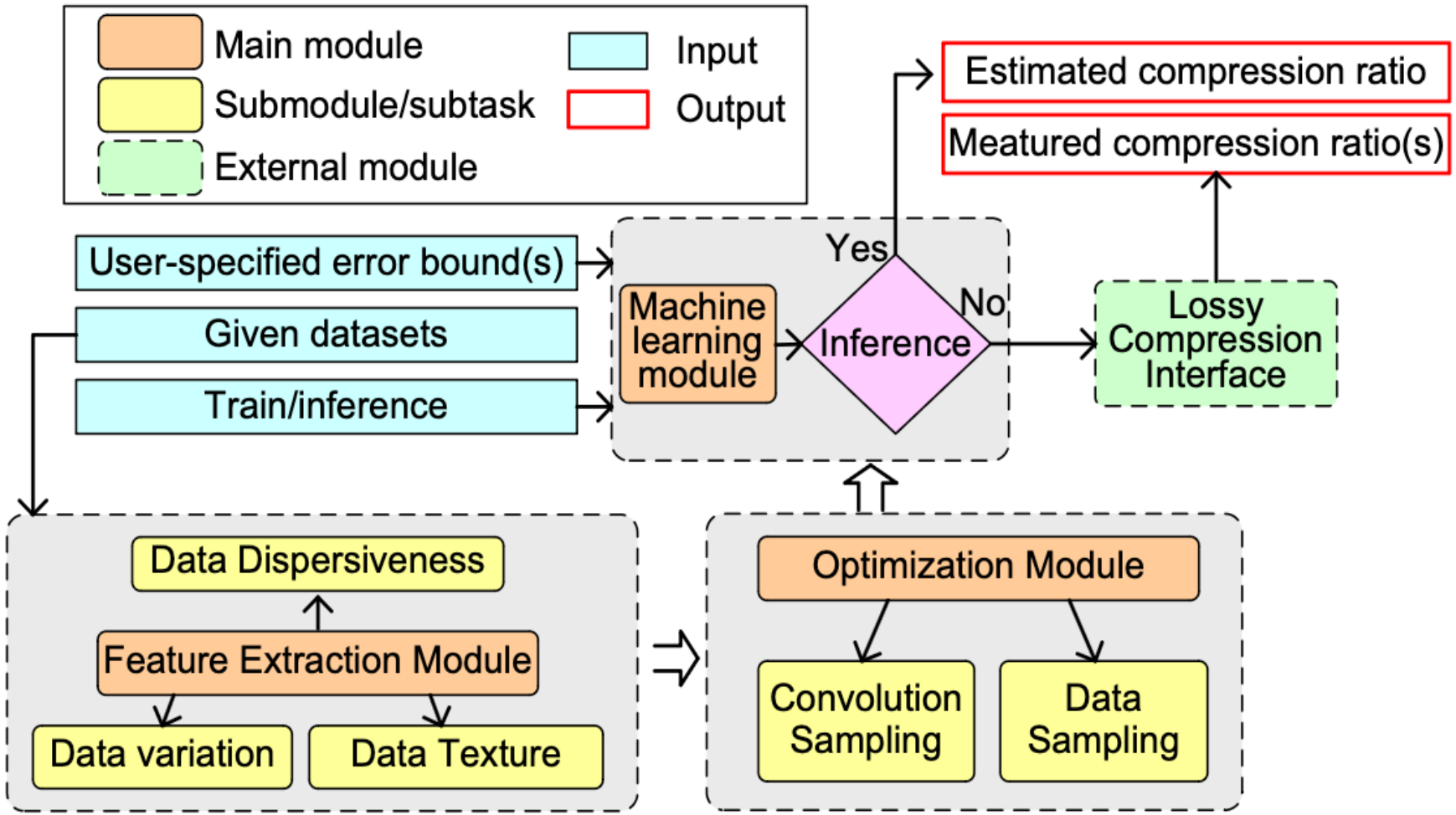
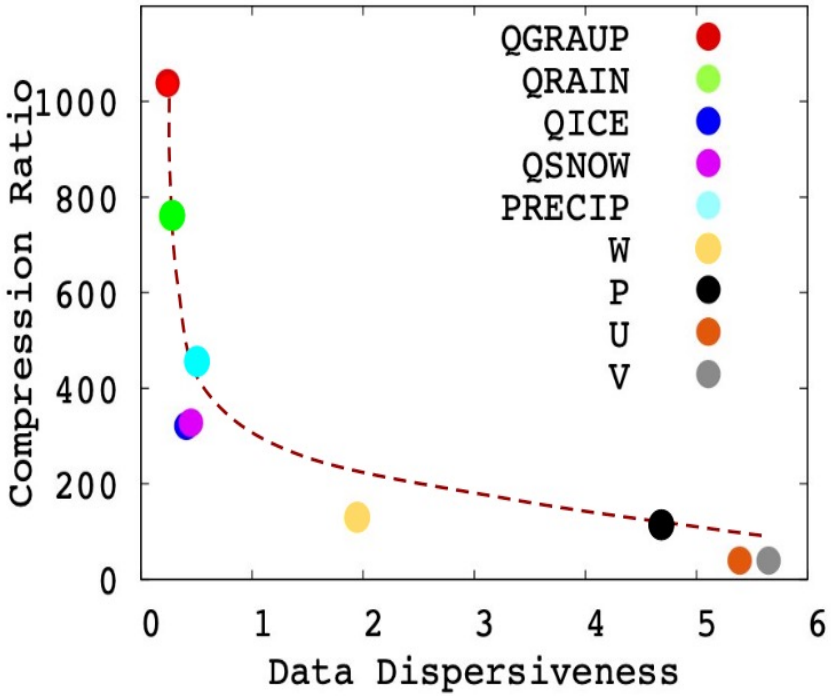


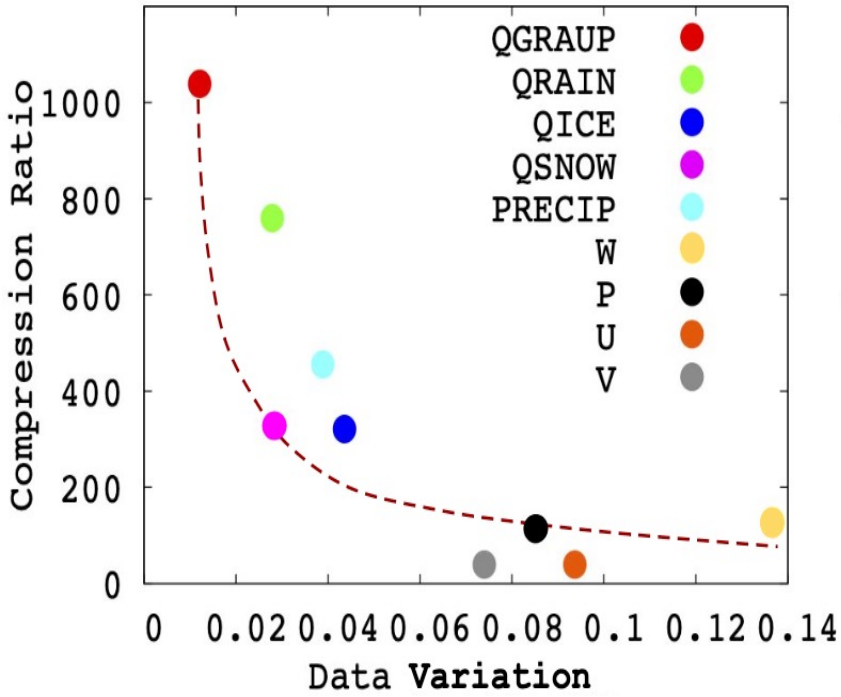
Figure: XTIMATE Design Workflow.

Feature Extraction Module

Data Dispersiveness



Data Variation



Textures Sharpness

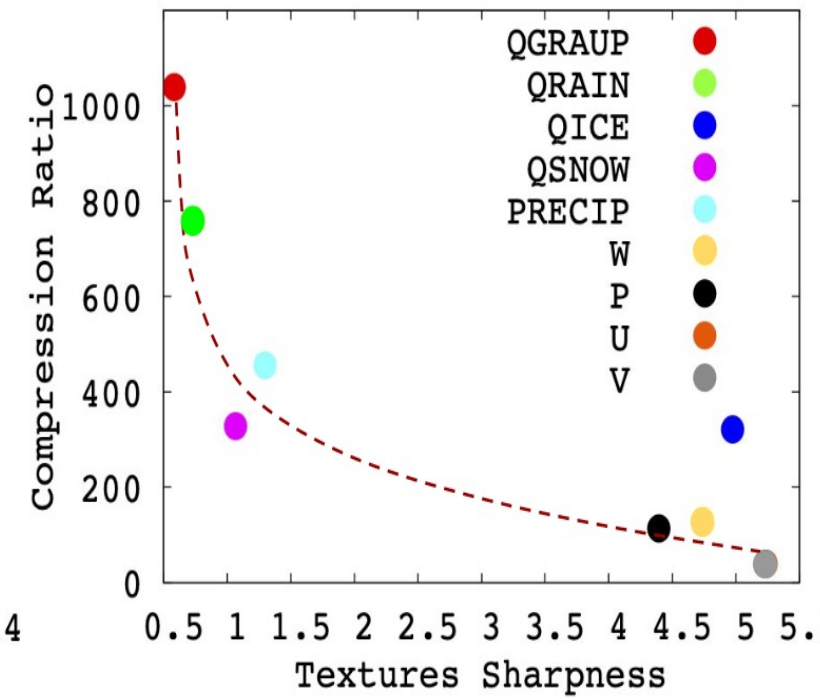


Figure: Illustrating The Impact of Our Data Features on Data Compressibility.

Optimization: Reducing Convolution Space

- Sobel kernel filter convolution is used to characterize data textures
- Total convolution space is NxMxd; N data points, M convolution points, d dimension data
- Convolution space would be huge for large datasets
- Our estimation should be faster than the actual compression time
- XTIMATE prunes space by only focusing on the critical convolution regions

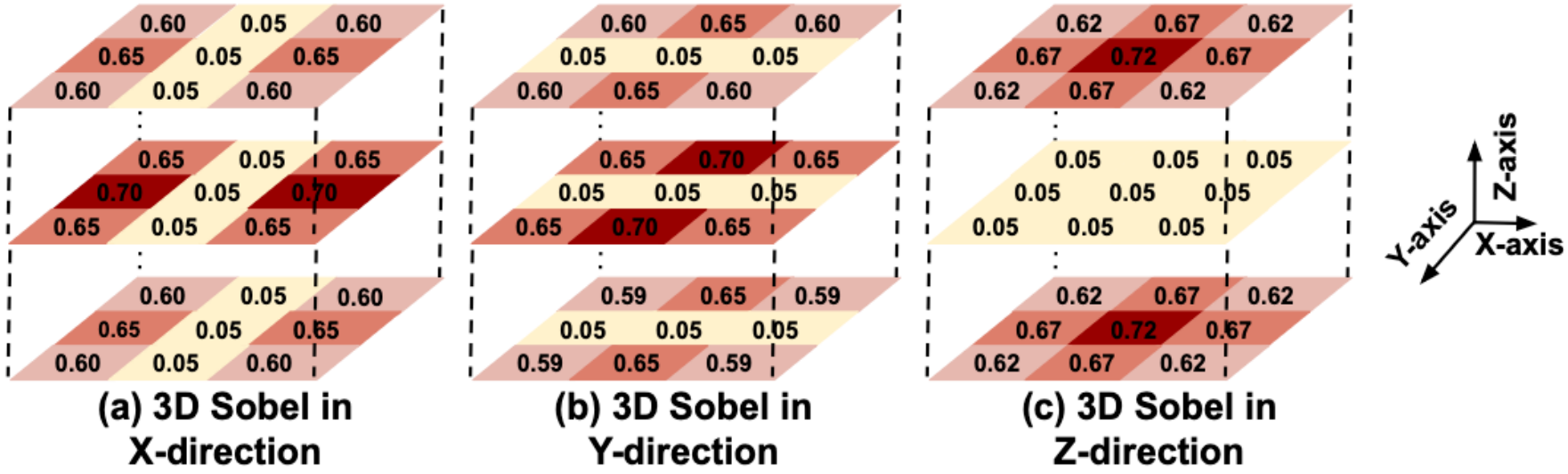
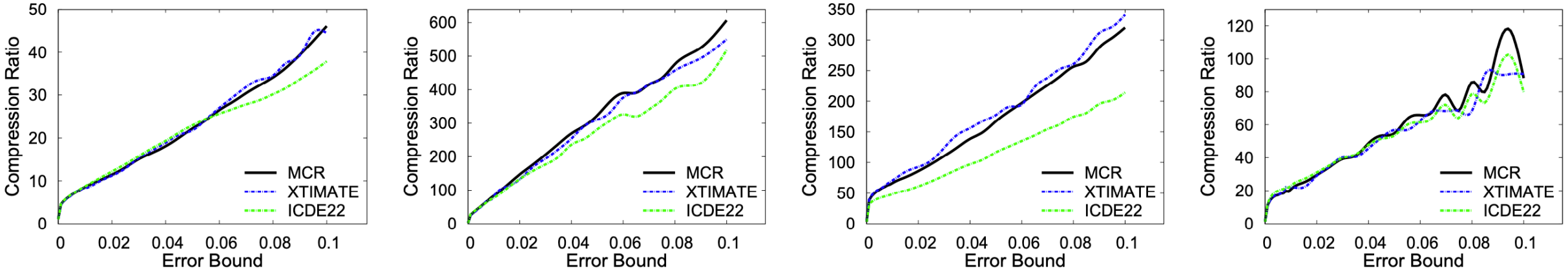


Figure: Heat Map to Identify Significance Level of Each Convolution Space in Sobel Filter.

Experimental Setup

- Scientific datasets
  - 43 real-world scientific dataset fields from 5 HPC simulations/applications
- Evaluated lossy compressors
  - SZ : prediction-based lossy compression scheme
  - ZFP: transformation-based lossy compression scheme
  - MGARD: multigrid adaptive reduction approach
- Baselines
  - ICDE22[28]: a compression ratio estimation framework specific to SZv3 compressor
  - IDDPS18[29]: estimates compression ratio with SZv1.4 and ZFPv0.5.0 compressors

Accuracy Evaluation



(a) Exaalt Y (with SZ3) (b) CESM FLNT (with SZ3) (c) Miranda Velocityy (with SZ3) (d) Nyx Temperature (with SZ3)

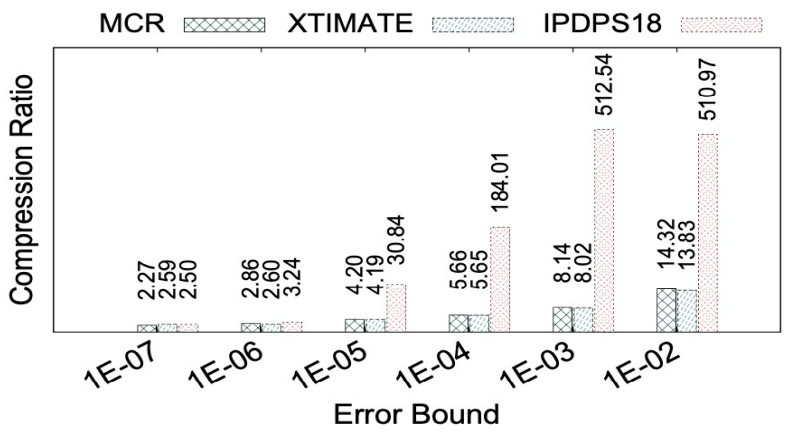
Figure: Comparison of Estimation Error between XTIMATE and ICDE22 Models across Different Error Bounds.

Comp.	Exaalt (2D)		CESM (2D)		Hurricane (3D)	
	Vy	Y	CLDMED	FLNT	V	PRECIP
MGARD+	2.10%	1.65%	6.31%	5.42%	6.82%	13.01%

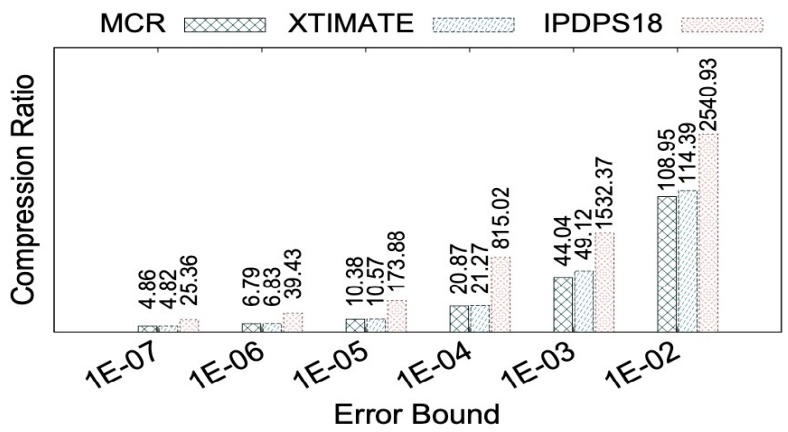
Comp.	Miranda (3D)		Nyx (3D)		Average
	Viscocity	Velocityy	Temperature	Velocityy	
MGARD+	2.07%	2.79%	8.91%	2.42%	5.15%

Table: Average Estimation Error of XTIMATE with MGARD+ across Different Error Bounds .

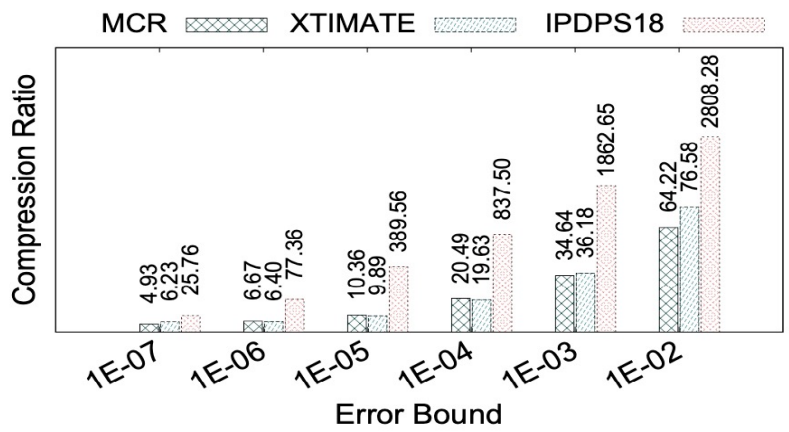
Accuracy Evaluation



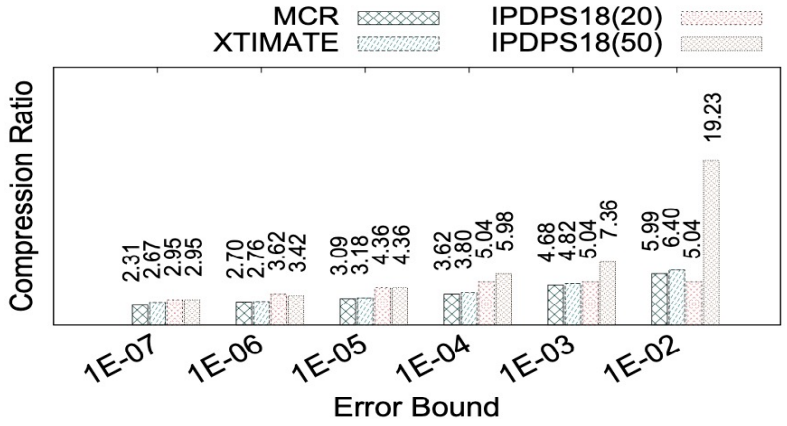
(a) Exaalt Y (with SZ1.4)



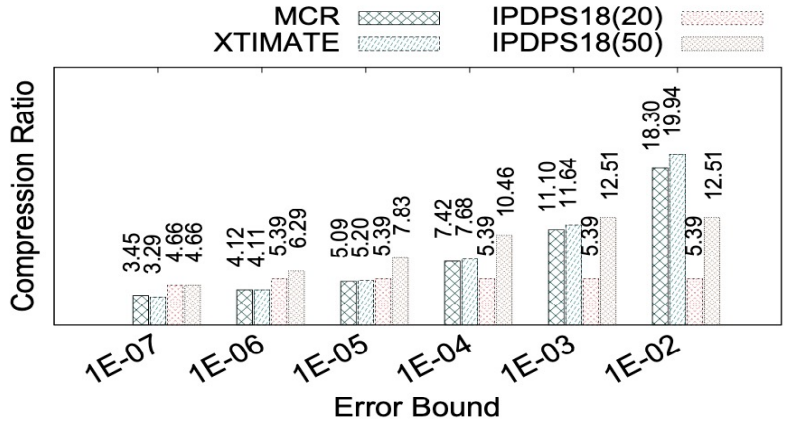
(b) CESM FLNT (with SZ1.4)



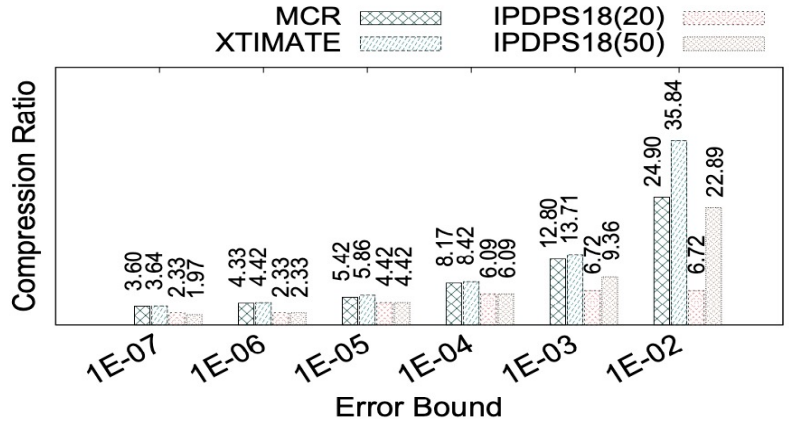
(c) Hurricane V (with SZ1.4)



(d) Exaalt Y (with ZFP0.5.0)



(e) CESM FLNT (with ZFP0.5.0)



(f) Hurricane V (with ZFP0.5.0)

Figure: Comparison of Estimation Error between XTIMATE and IPDPS18 Models across Different Error Bounds. IPDPS18(20) and IPDPS18(50) denote the model execution under 20 and 50 max-iterations respectively.

Efficiency Evaluation

Application	Testing Field Datasets	SZ3		SZ1.4		ZFP0.5.0		MGARD+
		XTIMATE	ICDE22	**XTIMATE**	IPDPS18	**XTIMATE**	IPDPS18	**XTIMATE**
Exaalt(1D)	Vy	0.195x	0.301x	0.405x	7.668x	0.398x	5.416x	0.414x
	Y	0.181x	0.309x	0.340x	7.187x	0.338x	4.593x	0.405x
CESM(2D)	CLDMED	0.028x	0.390x	0.071x	10.404x	0.063x	3.692x	0.078x
	FLNT	0.025x	0.372x	0.064x	8.951x	0.059x	4.088x	0.071x
Hurricane(3D)	V	0.017x	0.663x	0.026x	6.377x	0.022x	1.359x	0.041x
	PRECIP	0.011x	0.666x	0.022x	8.367x	0.014x	0.943x	0.032x
Miranda(3D)	Viscocity	0.012x	0.757x	0.029x	6.782x	0.031x	1.134x	0.034x
	Velocityy	0.015x	0.672x	0.024x	5.508x	0.026x	1.058x	0.038x
Nyx(3D)	Temperature	0.025x	0.718x	0.056x	4.274x	0.058x	0.796x	0.045x
	Velocity-y	0.023x	0.723x	0.059x	5.153x	0.065x	0.914x	0.045x
Average	Across All Fields	0.053x	0.557x	0.126x	7.342x	0.122x	2.743x	0.120x

Table: Average Execution Time Compared to Corresponding Compressor's Compression Time for XTIMATE, ICDE22 and IPDPS18 (20 Iterations) Models across Different Error Bounds.

Conclusion

- Propose a compressor-agnostic ratio estimation framework
- Key insights: data textures have notable impact on data compressibility
- Propose two optimization strategies to improve accuracy and efficiency
- XTIMATE provides efficient data orchestration during runtime
- XTIMATE incurs only 6.77% average ratio estimation error
- Performance gain up to 50× compared to related studies