NEURAL INFORMATION PROCESSING SYSTEMS

MeLLoC: Lossless Compression with High-order Mechanism Learning

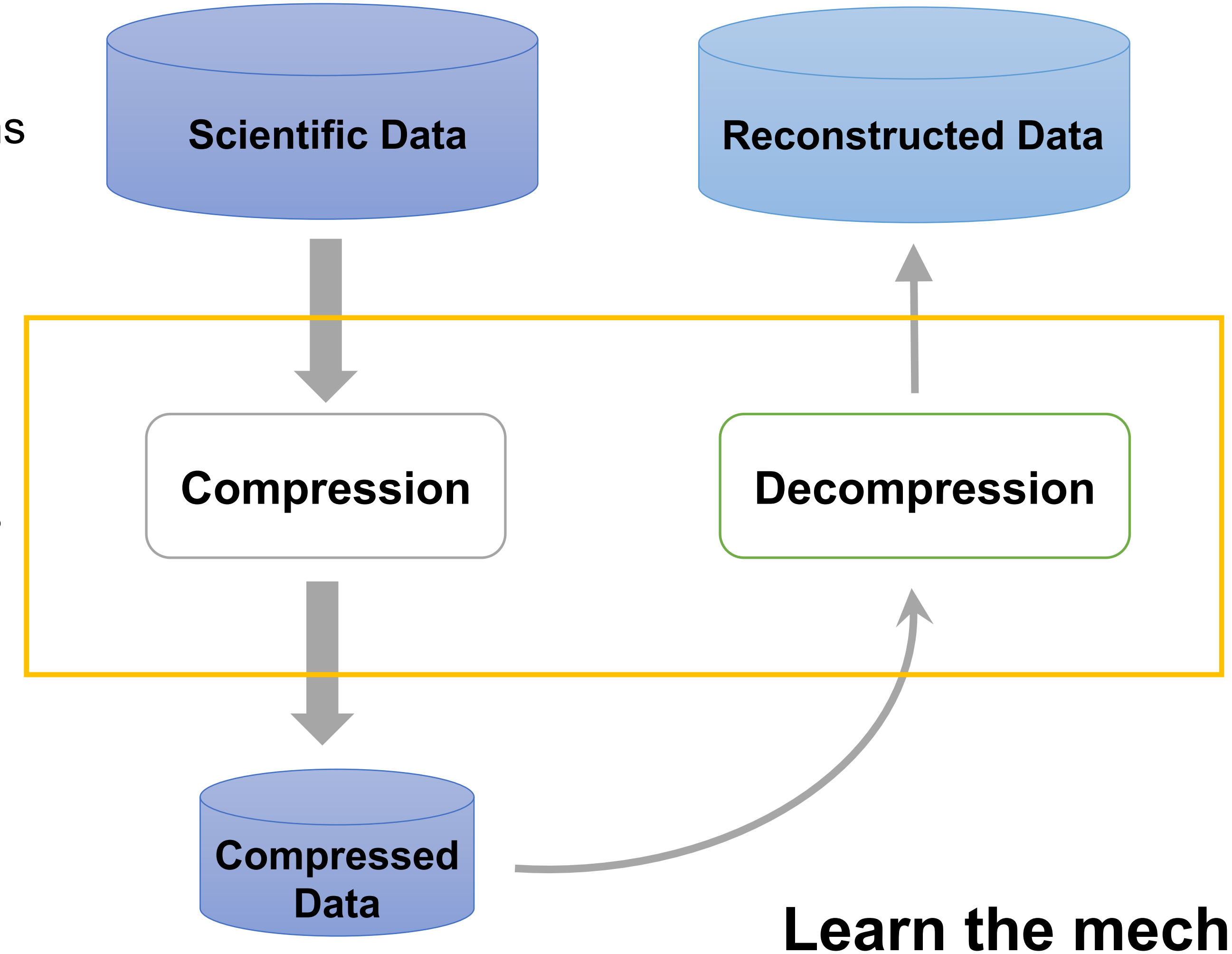
Xinyue Luo, Jin Cheng, Yu Chen*

NeurIPS 2024

Motivation

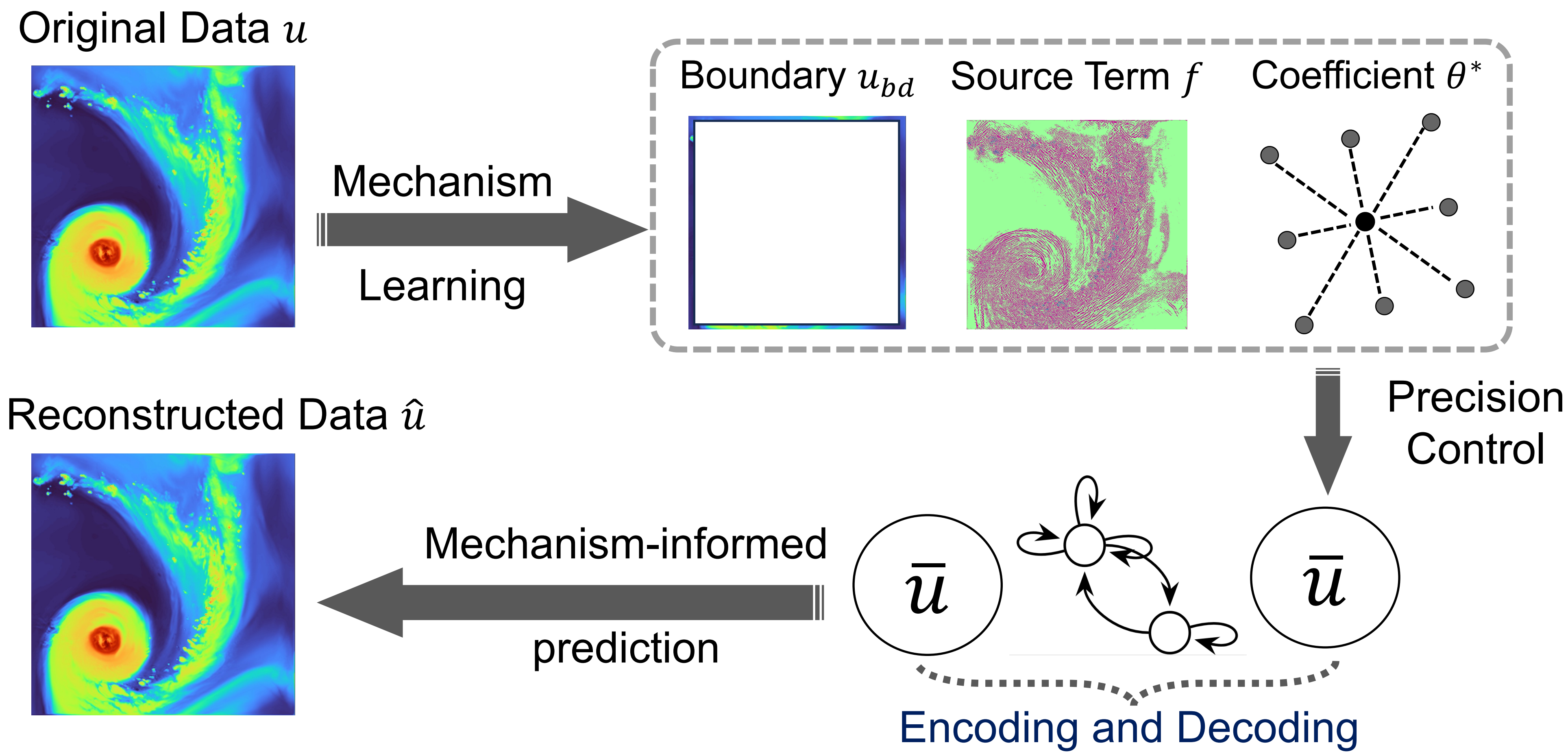
- massive
- scientific mechanisms （physical laws）

- tradeoff: accuracy vs storage efficiency



Learn the mechanism behind data!

Overview


Original Data u
Mechanism
Learning
Boundary u_{bd}
Source Term f
Coefficient θ^*
Precision
Control
Reconstructed Data $\hat{u}$
Mechanism-informed
prediction
$\bar{u}$
$\bar{u}$
Encoding and Decoding

Mechanism Learning

Mechanism assumption: cross-sectional 2nd order linear differential equations

$$\mathcal{L}u = f \ in \ D.$$

Local representation:

$$C_1 u_{i-1,j-1} + \cdots + C_9 u_{i+1,j+1} = f_{i,j} \ ,$$

In the following part, we denote $\theta = \{C_i\}_{i=1}^{9}$.

$$\underset{\theta}{\operatorname{argmin}} \ F\left(\theta; u_d^{bd}, u_d^{in}\right)$$

where $F = \|\mathcal{L}_\theta u_d\|_2^2 = \sum_{i,j} \left(C_1 u_{i-1,j-1} + \cdots + C_9 u_{i+1,j+1} - u_{i,j}\right)^2$

$\boldsymbol{C_4 u_{i-1,j+1}}$	$\boldsymbol{C_8 u_{i,j+1}}$	$\boldsymbol{C_9 u_{i+1,j+1}}$
$\boldsymbol{C_3 u_{i-1,j}}$	$\boldsymbol{C_5 u_{i,j}}$	$\boldsymbol{C_7 u_{i+1,j}}$
$\boldsymbol{C_1 u_{i-1,j-1}}$	$\boldsymbol{C_2 u_{i,j-1}}$	$\boldsymbol{C_6 u_{i+1,j-1}}$

Various mechanisms, unique minimizer, direct calculation due to linearity

Mechanism Learning

For $u \in C^4(\Omega), \Omega \subset \mathbb{R}^2$,

$$\sum_{k,l=-1}^{1} C_{k,l}u(x+kh, y+lh) = \Big[2(c_1+c_5-c_6)h\partial_x + 2(c_2+c_5+c_6)h\partial_y + (c_3+c_7+c_8)h^2\partial_{xx}^2$$

$$+(c_4+c_7+c_8)h^2\partial_{yy}^2 + 2(c_7-c_8)h^2\partial_{xy}^2 + c_9\Big]u(x,y) + o(h^2)$$

The relationship between $C_{k,l}$ and c_n :

$$\mathbf{C} = c_1\mathbf{A}_1 + c_2\mathbf{A}_2 + c_3\mathbf{A}_3 + c_4\mathbf{A}_4 + c_5\mathbf{A}_5 + c_6\mathbf{A}_6 + c_7\mathbf{A}_7 + c_8\mathbf{A}_8 + c_9\mathbf{A}_9$$

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{A}_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}, \mathbf{A}_3 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{A}_4 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

$$\mathbf{A}_5 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \mathbf{A}_6 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \mathbf{A}_7 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -2 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \mathbf{A}_8 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{A}_9 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

$$\{\boldsymbol{C}_{k,l}\}^* = \underset{\boldsymbol{C}_{k,l}}{\operatorname{argmin}} \boldsymbol{F}(\boldsymbol{C}_{k,l}; \boldsymbol{u}) = \underset{\boldsymbol{C}_{k,t}}{\operatorname{argmin}} \sum_{\boldsymbol{i},\boldsymbol{j}} \left(\sum_{\boldsymbol{k},\boldsymbol{l}=-1}^{1} \boldsymbol{C}_{k,l}\boldsymbol{u}(\boldsymbol{i}+\boldsymbol{kh}, \boldsymbol{j}+\boldsymbol{lh}) \right)^2$$

$C_{-1,1}$	$C_{0,1}$	$C_{1,1}$
$C_{-1,0}$	$C_{0,0}$	$C_{1,0}$
$C_{-1,-1}$	$C_{0,-1}$	$C_{1,-1}$

Data Composition

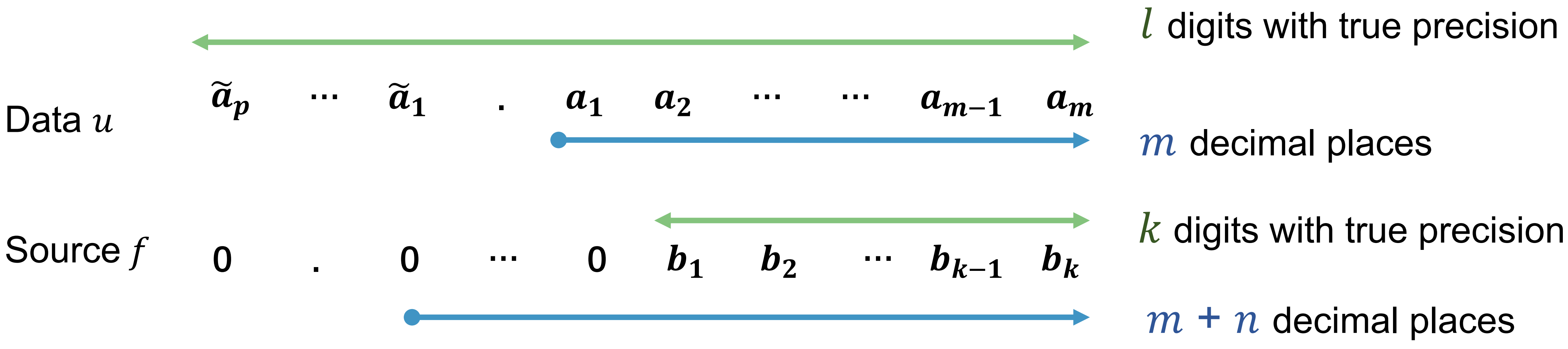
Let us consider the decomposition of data by linearity as u

$$u = \mathcal{L}^{-1}f + u_0 + u_{err}$$

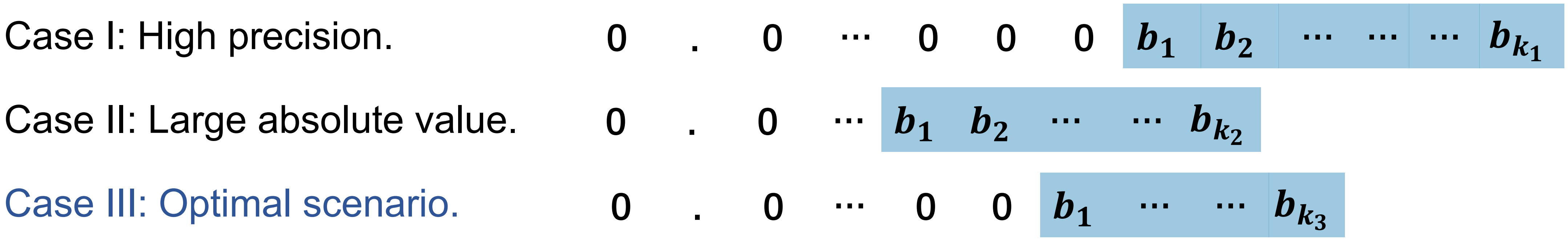
where

- $\mathcal{L}^{-1}f(= G * f)$ with G being the Green's function for the domain. This part corresponds to solution to the non-homogeneous equation with homogeneous boundary conditions, determined by the source $f(2\mathrm{D})$.
- $u_0 \left(= \frac{\partial G}{\partial v} * u^{bd}\right)$ denotes the solution to the homogeneous equation is determined only by the boundary data (1D).
- u_{err} denotes the residual part (2D).

Precision Control



Trade-off between the absolute value and the precision digits of the source term .

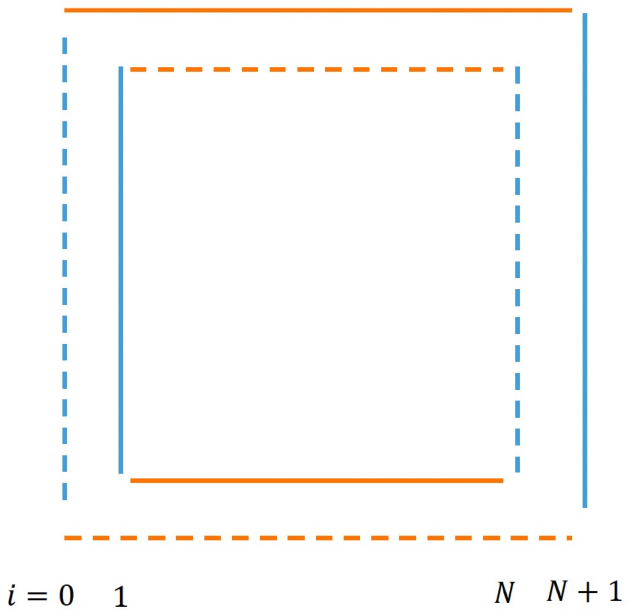


Fast Fourier-based Solver

- Main idea: Utilizes periodic extension FFT to accelerate computations
- Key steps:
  1. Periodically extend discrete field data
  2. Apply 2D Fourier series expansion

$$u(m,n) = \sum_{k=1}^{N_2} \sum_{j=1}^{N_1} \frac{1}{N_1 N_2} \hat{u}_{jk} e^{\frac{2\pi i}{N_1}(j-1)(m-1)} e^{\frac{2\pi i}{N_2}(k-1)(n-1)}$$

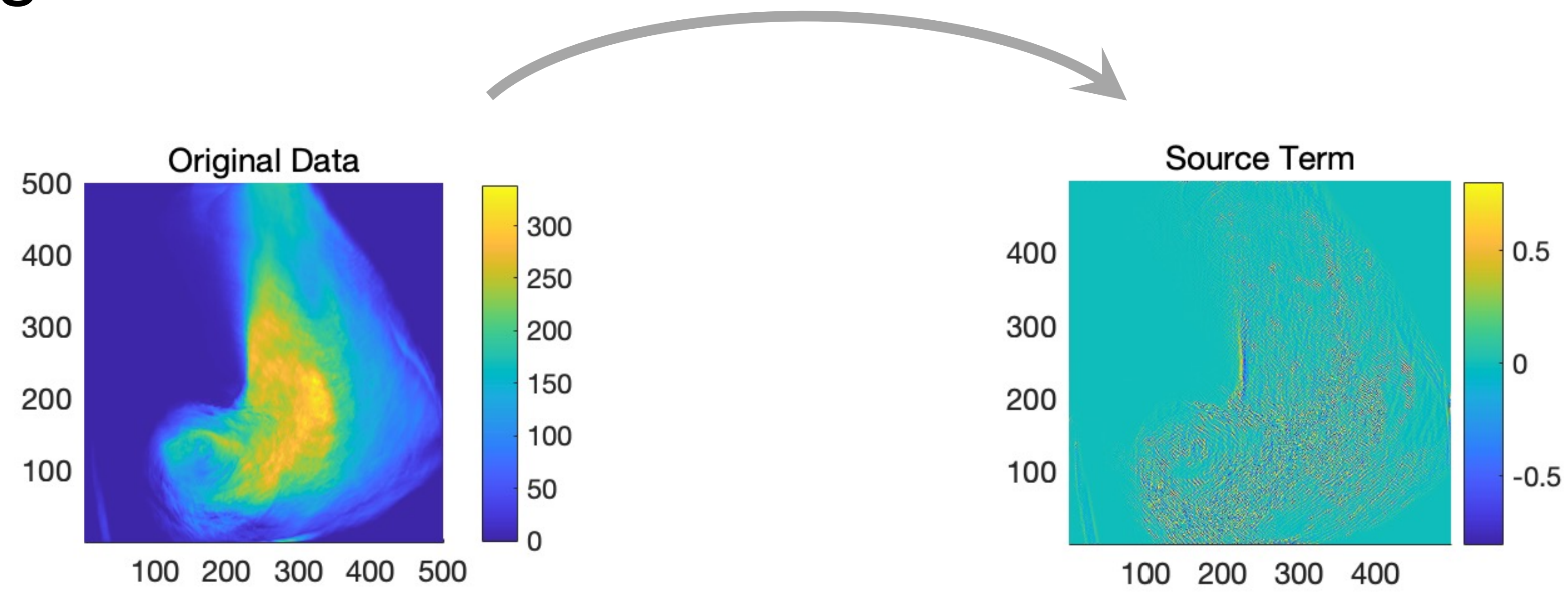
  3. FFT and direct solving frequency equation: $\hat{u}_{j,k} = \frac{\hat{f}_{j,k}}{B_{j,k}}$
  4. Process zero frequency terms and apply iFFT

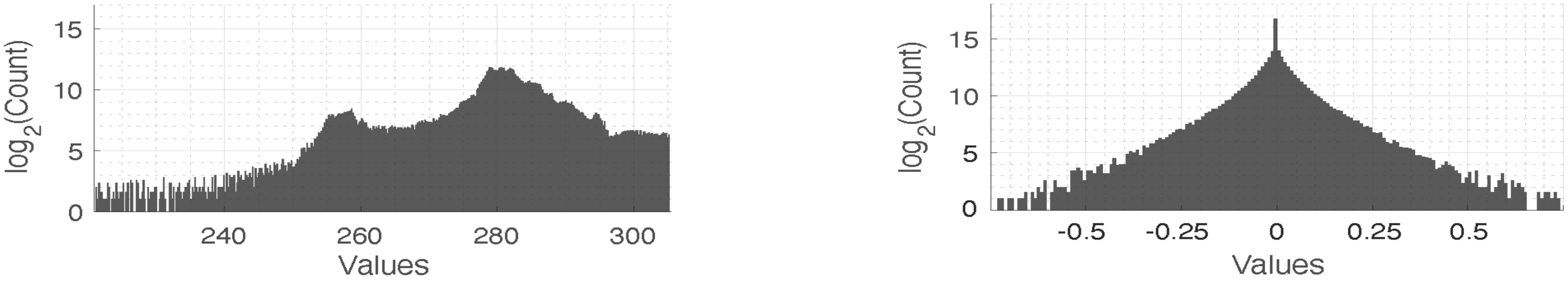


- Computational complexity: $O(N^2 \log N)$

Experiments

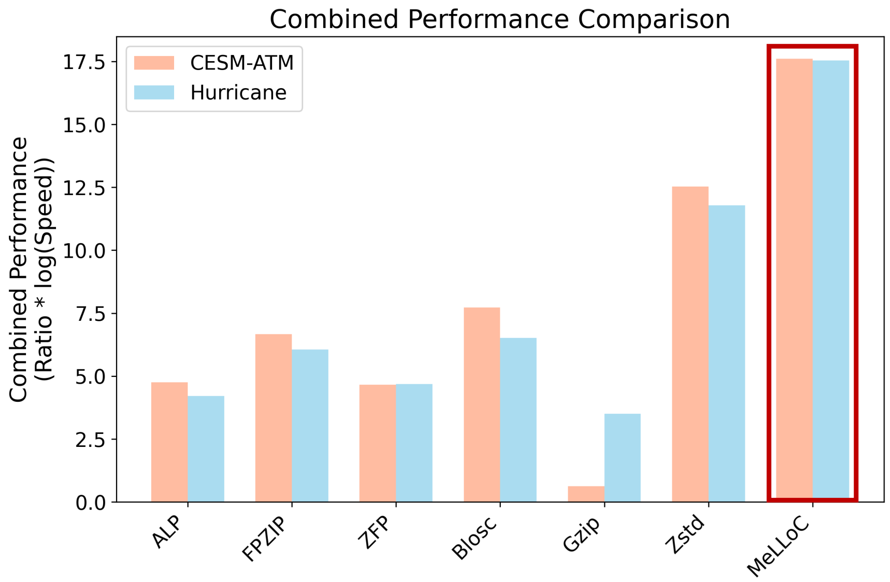
CESM_ATM dataset

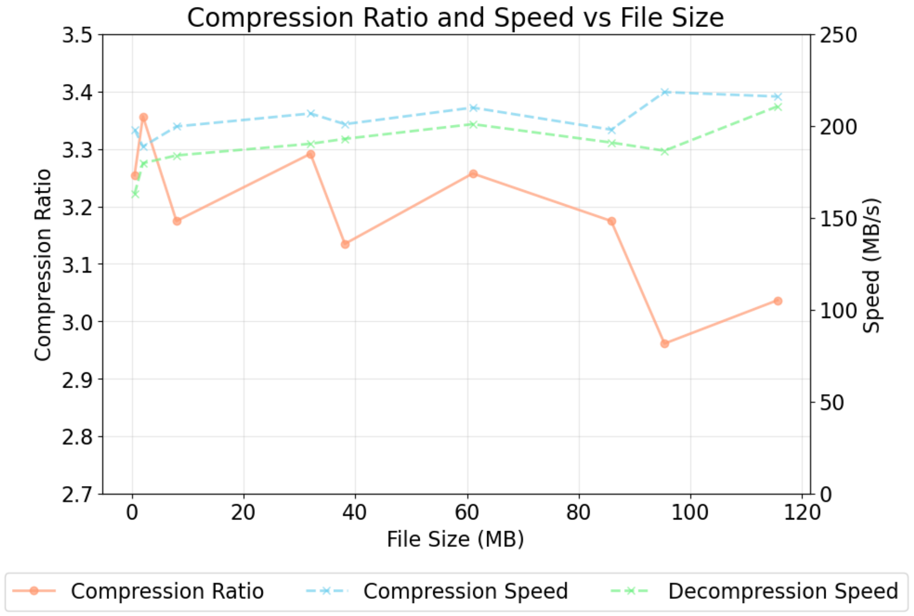




Experiments

	CESM-ATM				Hurricane		
Method	Ratio	Compression	Decompression	Method	Ratio	Compression	Decompression
ALP	1.16 X	46.93 Mb/s	1054.95 Mb/s	ALP	1.11 X	45.74 Mb/s	973.63 Mb/s
FPZIP	1.63 X	59.68 Mb/s	70.94Mb/s	FPZIP	1.63 X	41.22Mb/s	53.95Mb/s
ZFP	1.02 X	96.17 Mb/s	81.97 Mb/s	ZFP	1.01 X	102.95 Mb/s	68.06 Mb/s
Blosc	1.30 X	293.71 Mb/s	632.76 Mb/s	Blosc	1.12 X	888.65 Mb/s	6516.29 Mb/s
Gzip	1.89 X	1.40 Mb/s	266.94 Mb/s	Gzip	1.00 X	33.25 Mb/s	212.35 Mb/s
Zstandard	2.69 X	105.51Mb/s	152.81Mb/s	Zstandard	2.78 X	69.51Mb/s	271.32Mb/s
MeLLoC	**3.36 X**	**188.77 Mb/s**	**179.76Mb/s**	**MeLLoC**	**3.29 X**	**206.80Mb/s**	**190.35Mb/s**





Summary

- **Innovative Compression:** MeLLoC combines high-order mechanism learning with classical encoding for accurate data reconstruction.
- **Performance**: MeLLoC consistently achieves high compression ratios and competitive throughput.
- **Stability and Uniqueness**: The approach ensures stable, unique reconstructions through periodic extension and fast Fourier-based solutions.
- **Broad Applicability**: Effective for compressing scientific datasets like CESM-ATM and Hurricane data, showcasing its potential across various scientific fields.

Thank you!