Big Data Visual Analytics (CS 661)

Instructor: Soumya Dutta

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

email: soumyad@cse.iitk.ac.in

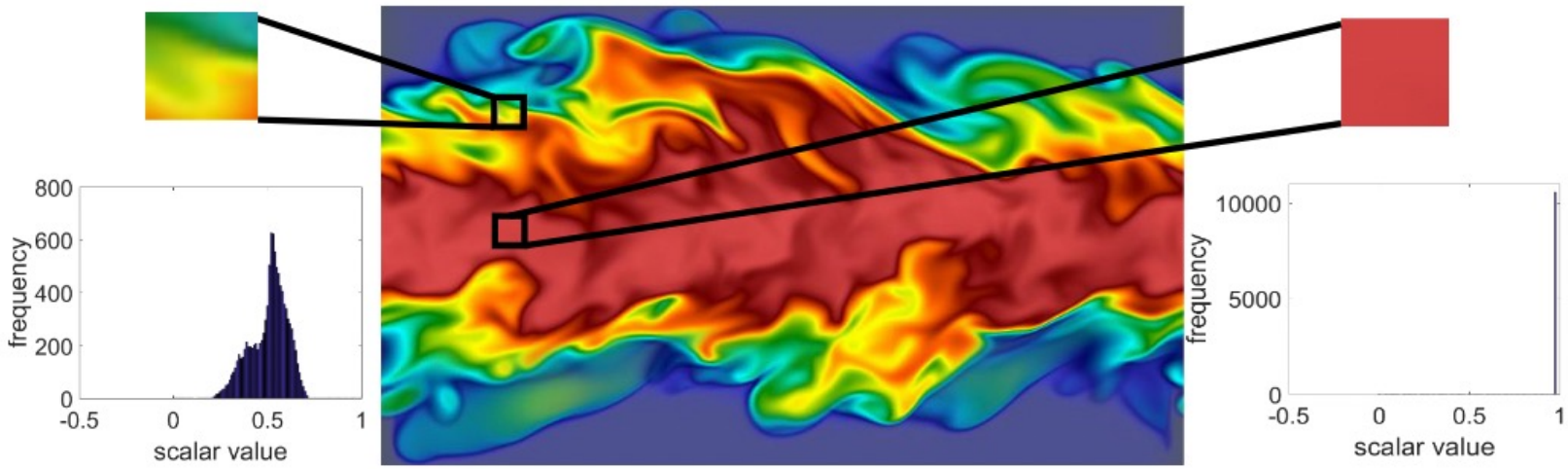
Two Make Up Classes

- Due to an official travel, I won't be able to take in person classes on April 21st and April 23rd (the last two lectures)
- So, we will have two make up classes to cover those two classes
- Make up class on April 5th (Saturday) 2:00-3:15pm at RM-101
- Make up class on April 12th (Saturday) 2:00-3:15pm at RM-101
- The quiz syllabus will remain the same, i.e., up to the class on April 2nd
- We will have final project evaluation during the final exam week
  - Possibly between 27-30th
- Get started with your projects, if clarification is needed on your proposal, I will contact you

Study Materials for Lecture 19

- Tzu-Hsuan Wei, Soumya Dutta, and Han-Wei Shen, Information Guided Data Sampling and Recovery using Bitmap Indexing, *IEEE Pacific Visualization Symposium (IEEE PacificVis)*, 2018, pp. 56-65.
- Soumya Dutta, Ayan Biswas, and James Ahrens, Multivariate Pointwise Information-driven Data Sampling and Visualization, *MDPI Entropy (Special issue in Information Theory Application in Visualization)*, 2019, Volume 21, Issue 7.

Information Guided Stratified Sampling



Comparisons between two data blocks with different data complexities. Drawing samples evenly in both regions may cause insufficient information stored in the data block shown in the left and redundant information stored in the right data block.

Information Guided Stratified Sampling

- Data is partitioned into small local regular non-overlapping blocks
- Information Entropy value guides the sample selection process from each block

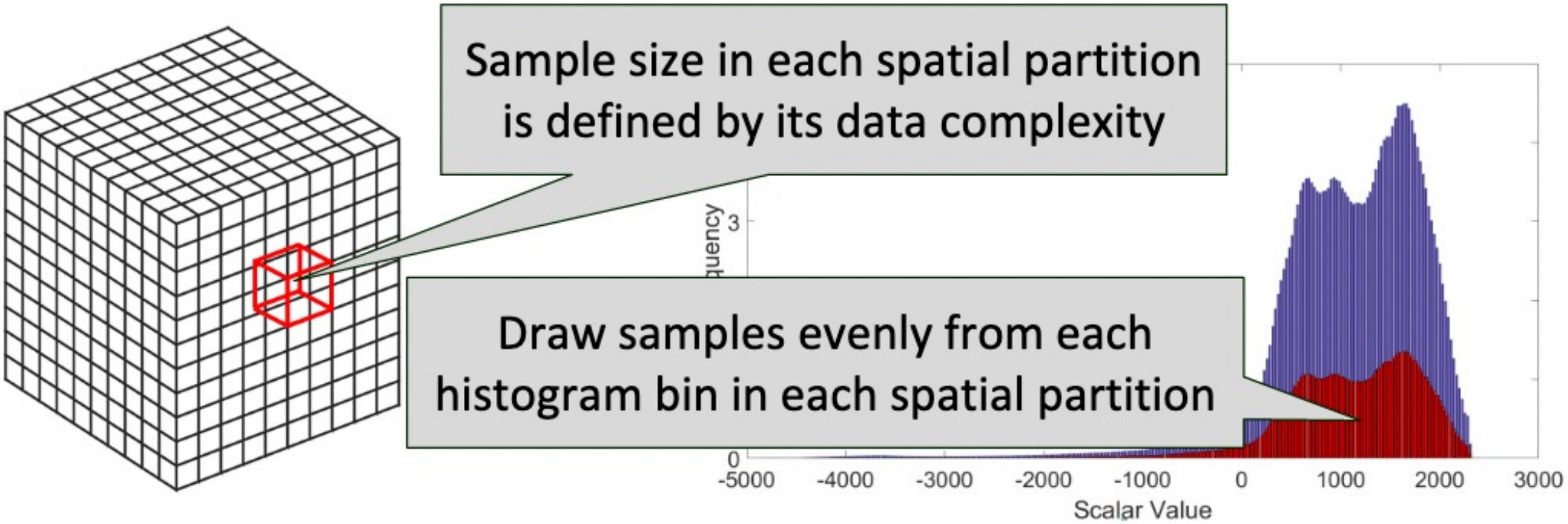
$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i)$$

- Determine the sampling percentage *s* as:

$$s = \frac{2^{H(X)}}{n} \times f$$

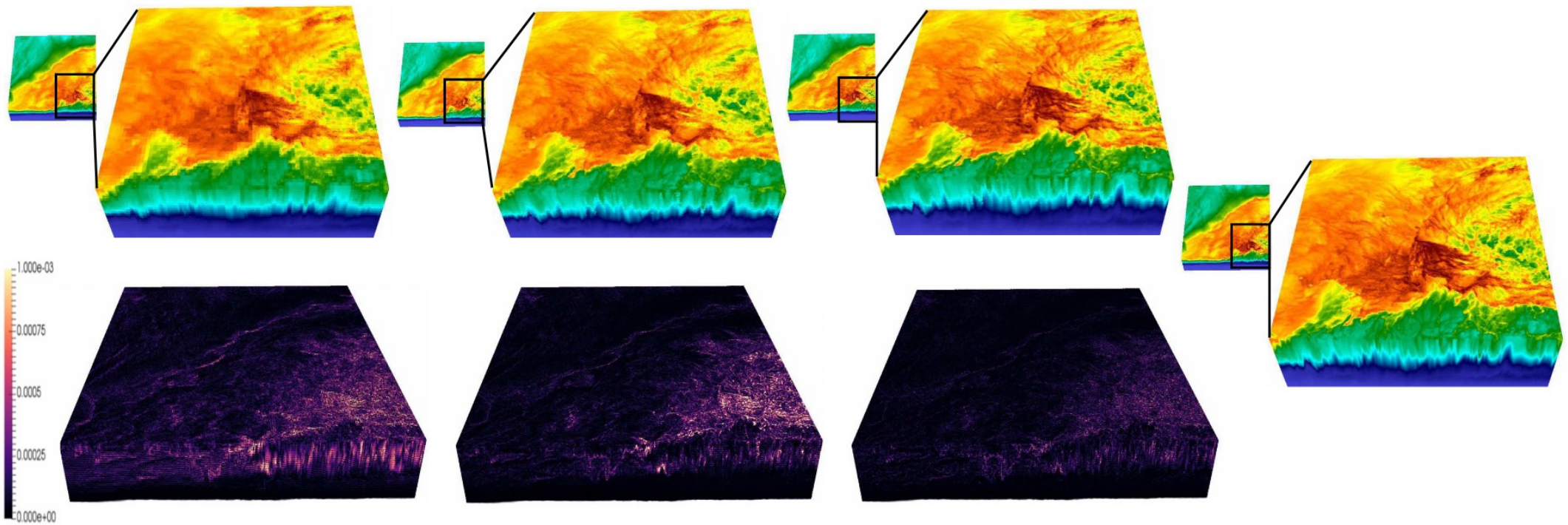
f = sampling fraction user given
H(X) = Entropy
n = number of histogram bins

Information Guided Stratified Sampling



The concept of information guided stratified sampling

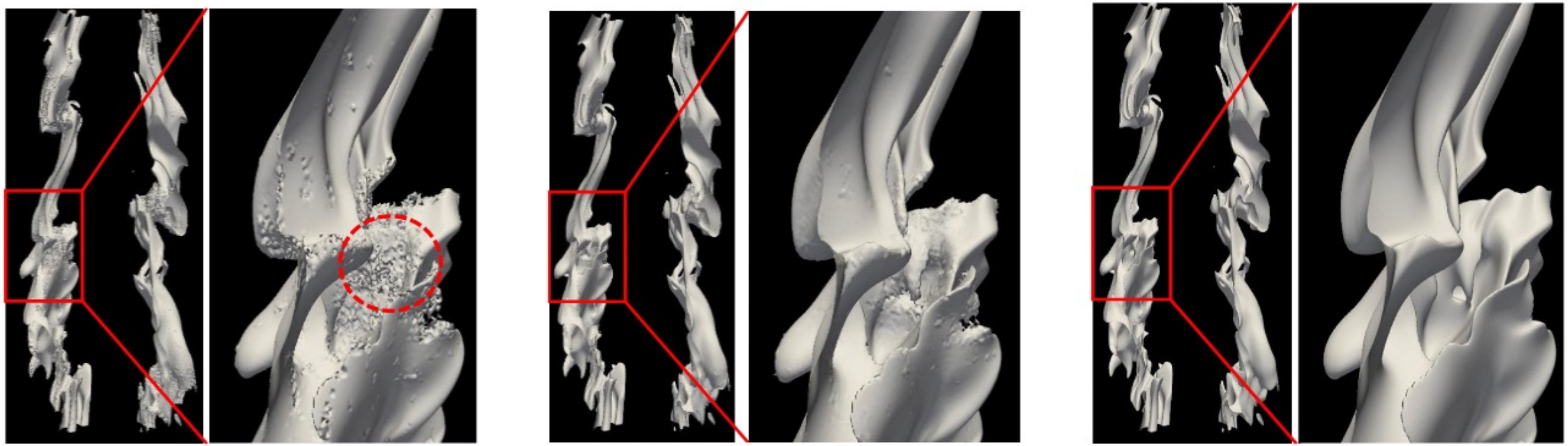
Reconstruction Results: Volume Rendering



(a) IGStS + RC_{mean} (38.3 MB) (b) AStS + $RC_{nearest}$ (34.7 MB) (c) IGStS + RC_{cost} (37.0 MB) (d) Ground Truth (1.67 GB)

The reconstruction using cost function is proposed as an assignment problem and Hungarian algorithm is used to solve it.

Reconstruction Results: Isosurface



IGStS + RC_{mean} | IGStS + RC_{cost} | ground truth

Sampling percentage set to 0.16

Multivariate Importance-based Sampling

How about Sampling Multiple Variables?

- So far, we have looked at methods that sample one variable at a time
  - Systematic, Random, Stratified sampling is trivially extended to multivariate domain
  - How do we perform importance-based multivariate sampling?
    - How to define multivariate importance?
- Example:
  - Sample points from Pressure and Temperature field simultaneously
    - How do we identify points that are important for both Pressure and Temperature?
  - **Idea**: Use information theoretic measure Pointwise Mutual Information (PMI) to quantify the association of each point and same points that have higher association value

Mutual Information (MI)

- Mutual information in information theory measures the mutual dependence between two random variables. It indicates the amount of information shared between two random variables.
- Mutual information is also interpreted as a measure of nonlinear dependence

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

$p(x)$ is the probability of a particular occurrence x of X
$p(y)$ is the probability of y of variable Y
$p(x,y)$ is their joint probability

$MI(X;Y)$ >= 0 and $MI(X;Y)$= $MI(Y;X)$

Pointwise Mutual Information (PMI)

- Given two random variables X and Y, if x is an observation of X and y for Y, then the PMI value for the value pair (x, y) is expressed as

$$PMI(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

$p(x)$ is the probability of a particular occurrence x of X
$p(y)$ is the probability of y of variable Y
$p(x,y)$ is their joint probability

- When $p(x,y) > p(x)p(y)$, PMI$(x,y) > 0$,
- When $p(x,y) < p(x)p(y)$, PMI$(x,y) < 0$,
- When $p(x,y) \approx p(x)p(y)$, PMI$(x,y) \approx 0$

- PMI is defined for two variables only

Specific Correlation for More Than Two Variables

$$SI(x_1, x_2, .., x_n) = log \frac{p(x_1, x_2, ..., x_n)}{p(x_1)p(x_2)...p(x_n)}$$

$p(x_1, x_2, \ldots xn)$ is the joint probability

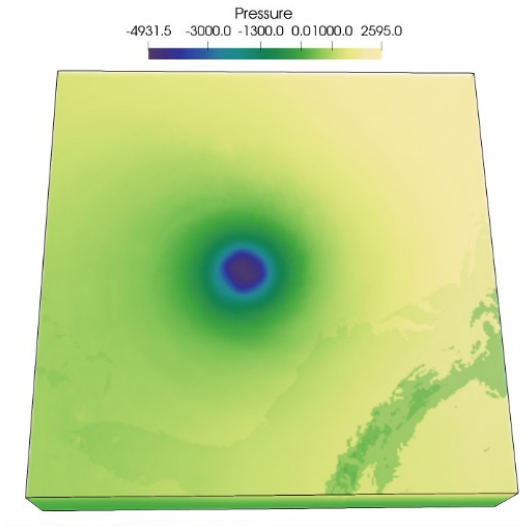
$SI(x_1, x_2, .., x_n)$ = Specific Correlation, pointwise association measure for more than two variables

Steps for Multivariate Importance-Driven Sampling

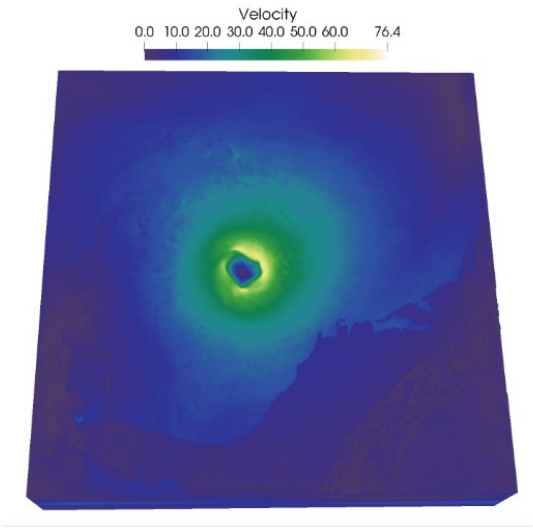
- Steps for 2 variable:

1. Compute 2D Histogram from data values of two variables
2. Estimate pointwise mutual information value for each 2D bin
3. Normalize the PMI values
4. Use the normalized PMI value for each bin as the acceptance probability for that bin
5. To determine whether to select a data point belonging to bin (i, j) having normalized PMI value PMI(i, j), first we generate a sample s drawn from a standard Uniform distribution U(0, 1).
6. If s < PMI(i, j) then the data point is selected.
7. This sample selection process is repeated for all the data points for each bin

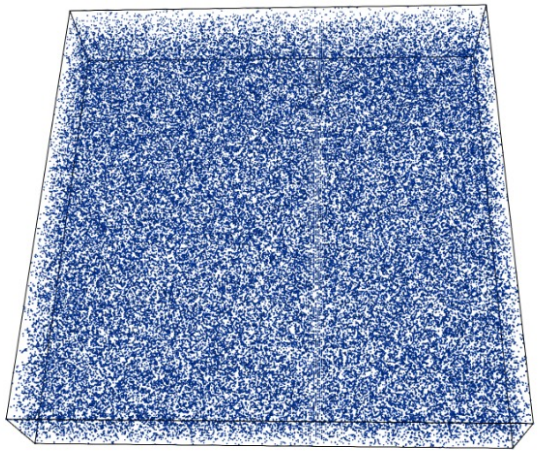
Results of Multivariate Importance-Driven Sampling



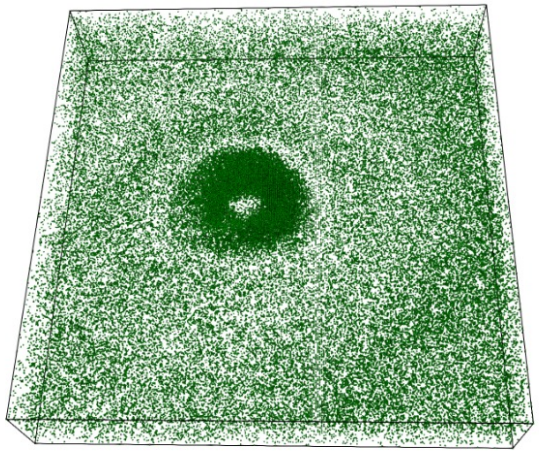
Pressure variable



Velocity variable

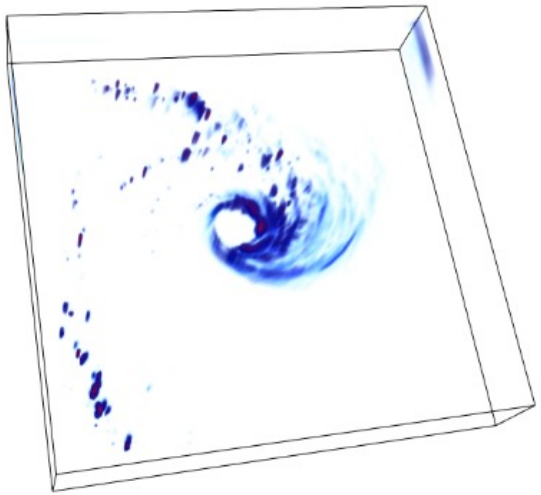


Random sampled

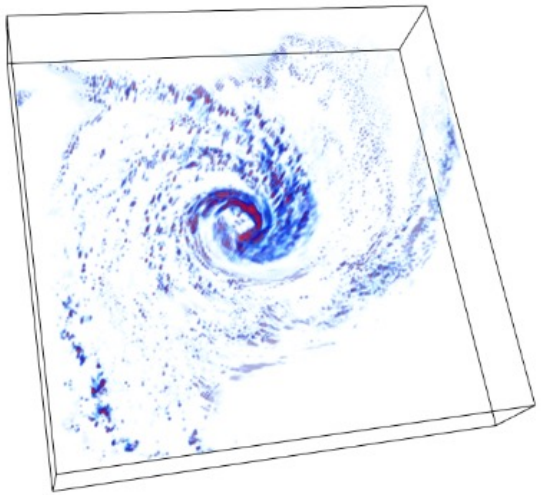


PMI sampled

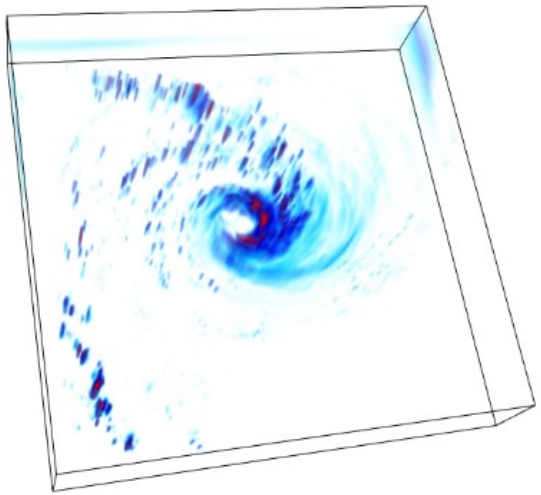
Results of Multivariate Importance-Driven Sampling



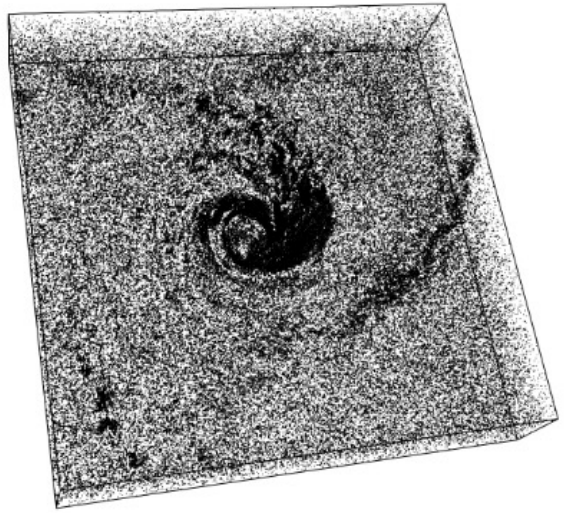
Variable 1: QGraup



Variable 2: QCloud

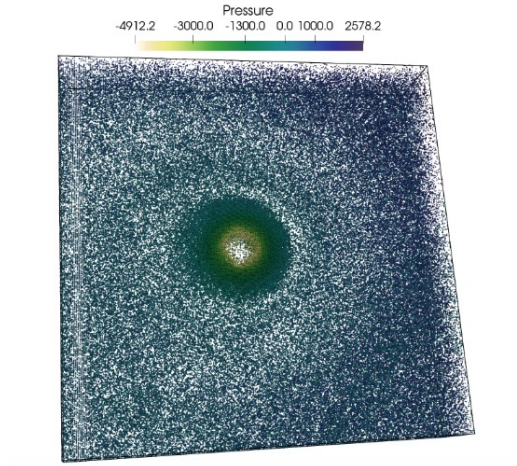


Variable 3: Precipitation

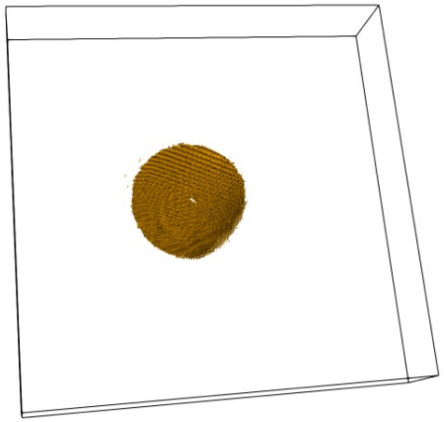


PMI Sampled Points

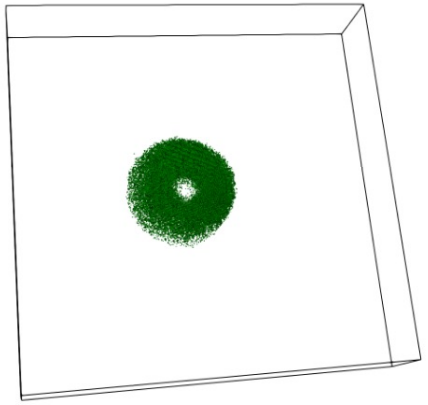
Multivariate Query-driven Analysis



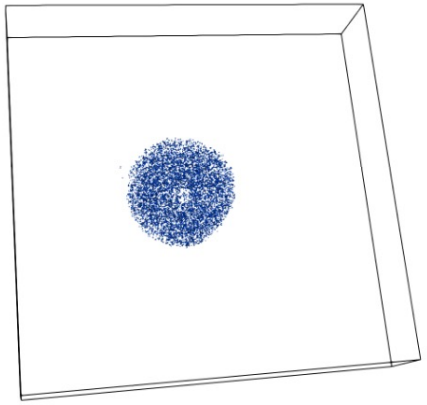
(a) All the data points sampled by the proposed sampling method.



(b) Result of the query when applied on raw data.



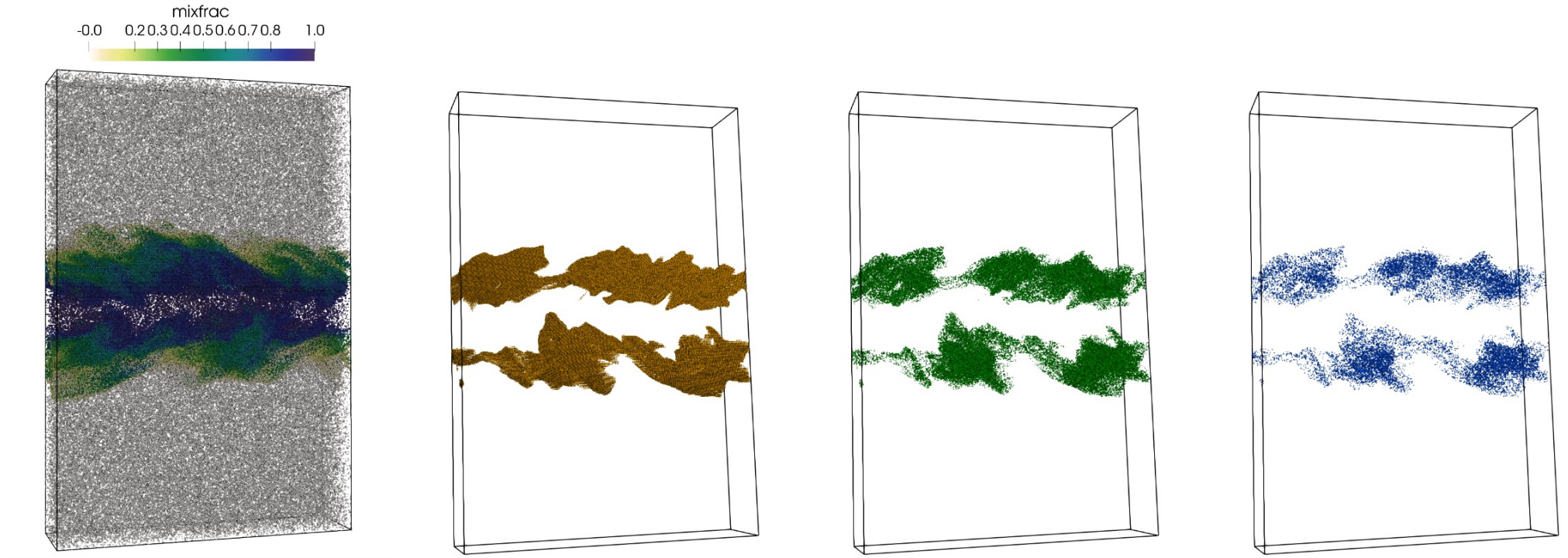
(c) Result of the query when applied on PMI-based sampled data.



(d) Result of the query applied to randomly sampled data.

Multivariate query: −100 < Pressure < −4900 AND Velocity > 10

Multivariate Query-driven Analysis



(a) All the data points sampled by the proposed sampling method.

(b) Result of the query when applied on raw data.

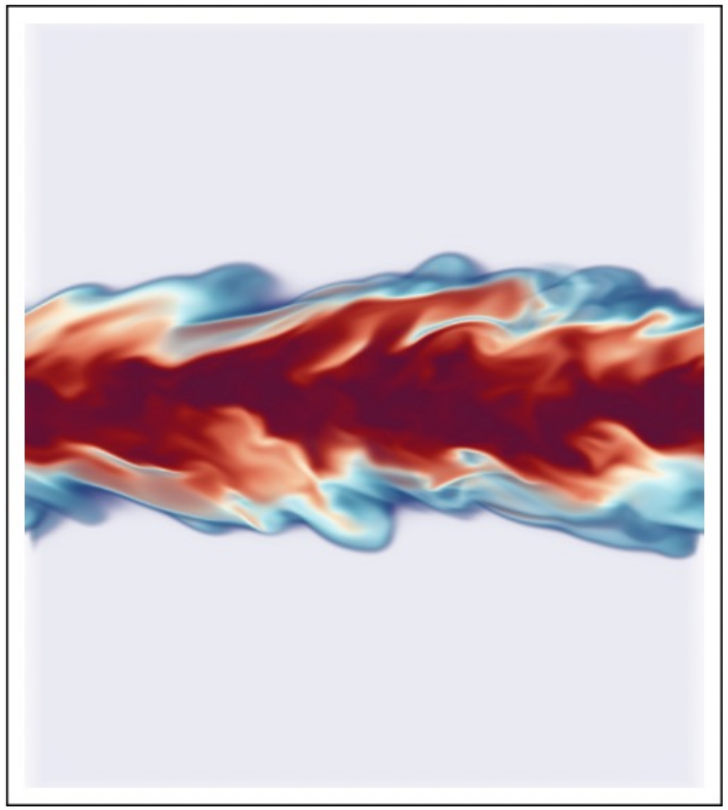
(c) Result of the query when applied on PMI-based sampled data.

(d) Result of the query applied to randomly sampled data.

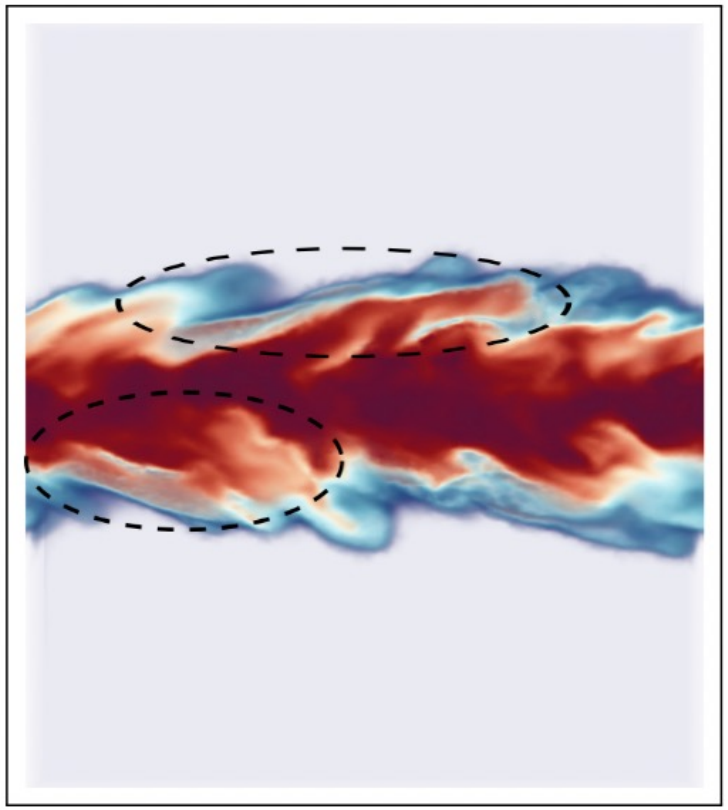
Multivariate query: 0.3 < mixfrac < 0.7 AND 0.0006 < Y_OH 0.1

Reconstruction and Visualization

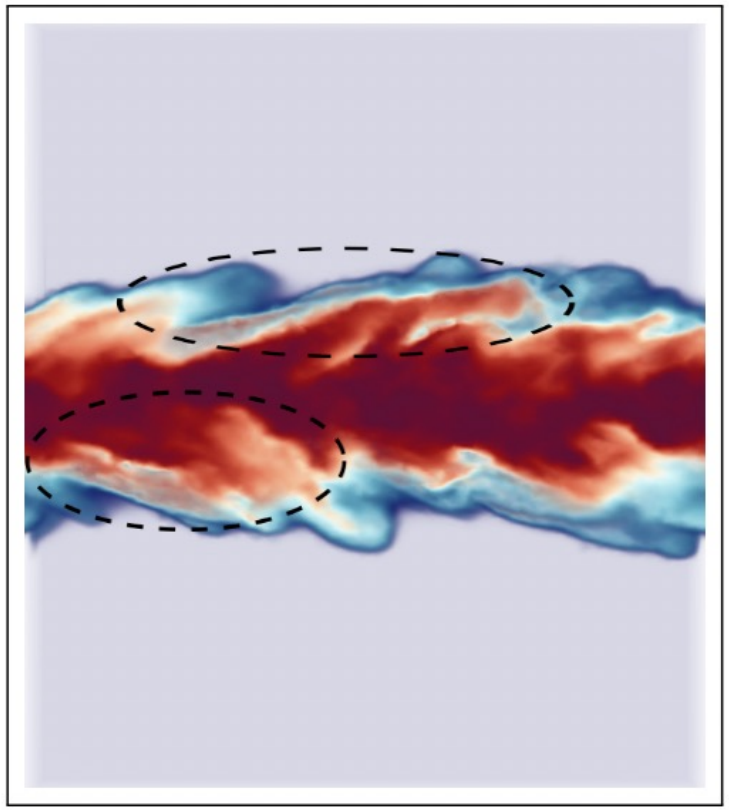
Linear interpolation-based reconstruction



Ground truth



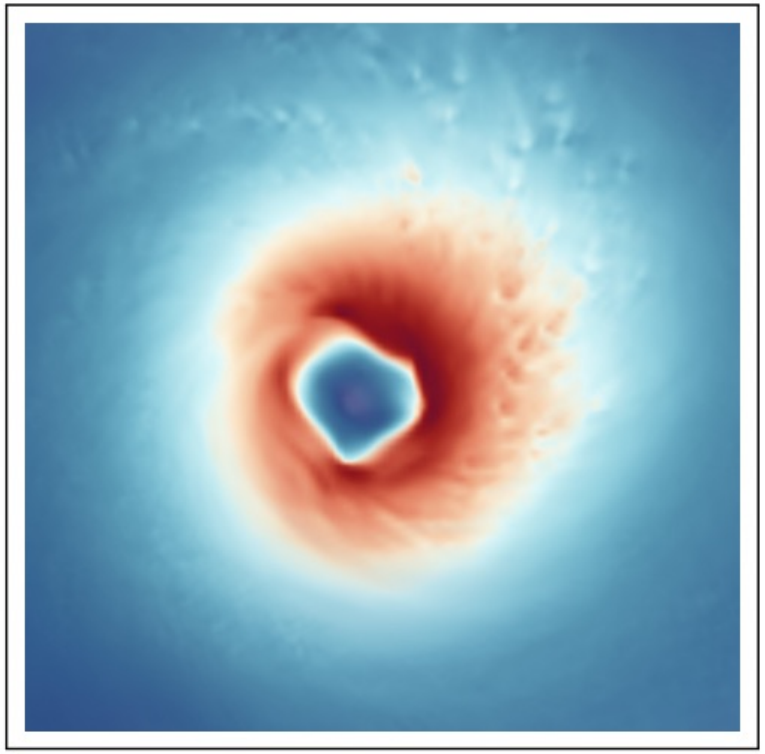
PMI-based Method



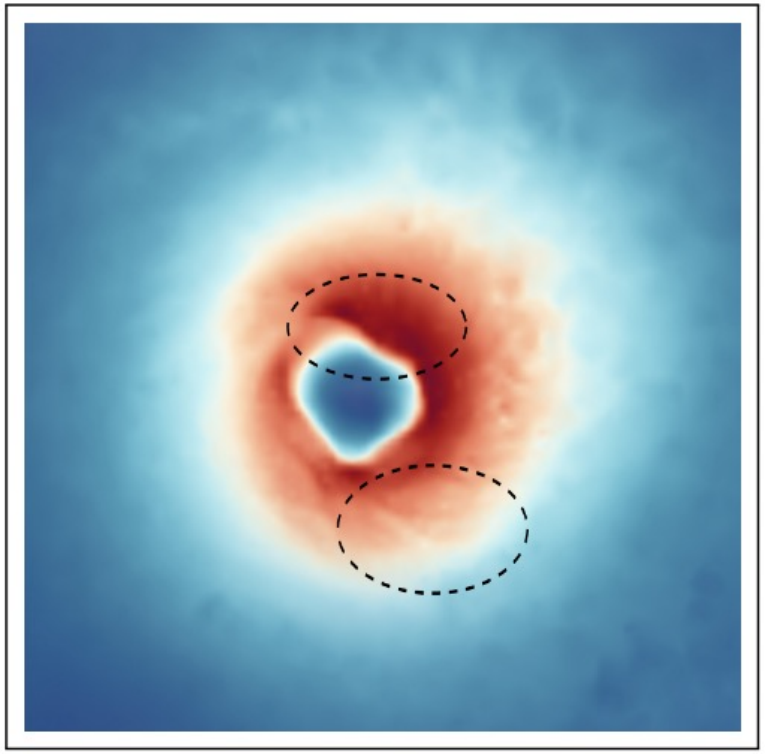
Random Sampling

Reconstruction and Visualization

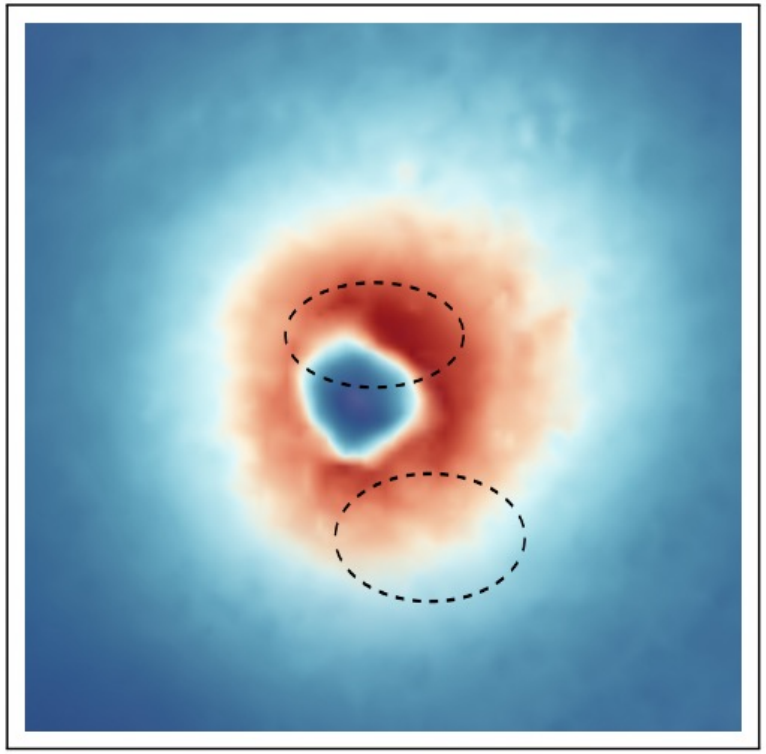
Linear interpolation-based reconstruction



Ground truth



PMI-based Method



Random Sampling

Image-based Quality comparison

- Structural Similarity (SSIM)

$$SSIM(I_1, I_2) = l(I_1, I_2)^a \cdot c(I_1, I_2)^b \cdot s(I_1, I_2)^c$$

l(I1, I2) represents luminance similarity
c(I1, I2) denotes contrast similarity
s(I1, I2) is the structural information
a, b, c are exponential weights

Image-based Quality comparison

Image-based comparison of Pressure and Velocity field of Isabel data set.

Isabel Pressure Field	samp. frac: 0.01		samp. frac: 0.03		samp. frac: 0.05	
	Random	Proposed	Random	Proposed	Random	Proposed
SSIM	0.9844	0.9915	0.9916	0.9931	0.9926	0.9939
MSE	6.5563	1.9267	2.5239	1.2559	2.0576	0.8961
Isabel Velocity Field	**samp. frac: 0.01**		**samp. frac: 0.03**		**samp. frac: 0.05**	
	Random	Proposed	Random	Proposed	Random	Proposed
SSIM	0.9234	0.9559	0.9427	0.9649	0.9516	0.9702
MSE	13.9638	8.492	10.6865	6.0452	8.1166	5.0213