THE TIME SLICE SELECTION BAKE-OFF

by

YUYA KAWAKAMI

A THESIS

Presented to the Department of Computer and Information Science
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Master of Science

June 2022

THESIS ABSTRACT

Yuya Kawakami

Master of Science

Department of Computer and Information Science

June 2022

Title: The Time Slice Selection Bake-off

As the size of data from scientific simulations grows, the ability to identify key time steps in a simulation has emerged as a key challenge. In response, a number of time slice selection methods and algorithms have been proposed. However, no past work has performed a comparative analysis of selection methods as well as their evaluation metrics. This thesis presents results from quantitative and qualitative study of selection methods and evaluation metrics to fill this gap. Our work has three major thrusts. First, we identify similarities and dissimilarities between different time slice slice selection algorithms. Second, we evaluate conditions under which these methods may fail. Third, we also perform a comparative study with evaluation metrics to investigate how selection methods perform under the set of evaluation metrics in literature. In all, this thesis aims to understand the space of time slice selection methods, to inform future research in this area.

CURRICULUM VITAE

NAME OF AUTHOR: Yuya Kawakami

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
Grinnell College, Grinnell, IA, USA

DEGREES AWARDED:

Master of Science, Computer and Information Science, 2022, University of Oregon
Bachelor of Arts, Mathematics and Computer Science, 2020, Grinnell College

AREAS OF SPECIAL INTEREST:

High-Performance Computing
Scientific Visualization
In Situ Processing

PUBLICATIONS:

Marsaglia, N., Kawakami, Y., Schwartz, S. D., Fields, S., & Childs, H. (2021, October). An Entropy-Based Approach for Identifying User-Preferred Camera Positions. In 2021 IEEE 11th Symposium on Large Data Analysis and Visualization (LDAV) (pp. 73-83). IEEE.

Kawakami, Y., Marsaglia, N., Larsen, M., & Childs, H. (2020). Benchmarking In Situ Triggers Via Reconstruction Error. In ISAV'20 In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization (pp. 38-43).

To my Mom, Dad and Sanah.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 1

II. REVIEW OF TIME SLICE SELECTION METHODS AND EVALUATION METRICS 6

  2.1. Domain-agnostic methods 6

    2.1.1. Reconstruction-based approaches 6

    2.1.2. Information-theoretic approaches 8

    2.1.3. Other domain-agnostic approaches 8

  2.2. Domain-specific approaches 10

III. EXPERIMENTAL OVERVIEW 12

  3.1. Data sets 12

  3.2. Selection methods 12

    3.2.1. Reconstruction-based approaches 12

    3.2.2. Information-theoretic approaches 14

    3.2.3. Flow-based* method and uniform selection 15

IV. SELECTION SIMILARITY 17

  4.1. Overview 17

  4.2. Selection difference 17

  4.3. Hierarchical clustering 20

    4.3.1. Overview 20

    4.3.2. Normalized SD 22

    4.3.3. Results 23

      4.3.3.1. Results from the first clustering pass (per data set & budget clustering) 23

Chapter Page

4.3.3.2. Results from the second clustering pass (aggregate clustering) 24

4.3.3.3. Observations on the behavior of the joint entropy selection method 25

V. FAILURE CONDITIONS 27

5.1. Overview 27

5.2. Issues that data sets with high dynamic range encounter 27

5.2.1. Effects on joint entropy selection methods 30

5.2.2. Effects on reconstruction-based selection methods that use Variation of Information 31

VI. METRIC COMPARISONS 35

6.1. Overview 35

6.2. Results 36

VII. CONCLUSION 41

APPENDICES

A. ADDITINAL RESULTS FROM EACH SELECTION METHOD 43

B. EVALUATIONS COMPARING TO RANDOM SELECTIONS 49

REFERENCES CITED 52

LIST OF FIGURES

Figure Page

1. Variation of Information between two random variables, X and Y, as Venn diagrams. 13
2. Comparing two selections with Euclidean alignemnt and Dynamic Time Warping alignment 19
3. Dendrograms from the first clustering phase 20
4. Dendrograms from the second clusterig phase 21
5. Volume renderings for the `earthquakeMag` data set 28
6. Histograms of data time steps in the `earthquakeMag` data set 30
7. Results from each selection method for the `earthquakeMag` data set at a budget of 10 31
8. Pairwise mapping cost at heat maps for the `earthquakeMag` data set 33
9. Percentile of evaluations among 5000 random selections at a budget of 5. 37
10. Results from each selection method at a budget of 5 for select data sets. Each dot corresponds to the time step that each selection method makes. 38

LIST OF TABLES

Table Page

1. Terminology . . . 3
2. Data sets . . . 12
3. Selection methods . . . 16
4. Block sizes for the imporance-based approach . . . 16
5. Example of a set of evaluations . . . 35

CHAPTER I

INTRODUCTION

With ever-increasing compute resources at our disposal, the size of scientific data of have also increased, both in the spatial and temporal domains. Coupled with the comparatively slow I/O speeds, it is increasingly difficult to investigate every time step a scientific simulation may produce. As a result, a topic of interest in Scientific Visualization has been time slice selection, i.e., selecting and identifying key time steps from a scientific simulation.

The main motivation of such a system is potential time savings. If done effectively, the selected subset of time steps can be considered as a short summary of the simulation and can remove the need to manually inspect each time step of the simulation after every run. If 10 time steps of a simulation of length 100 suffices to deliver the same or similar insight as the whole time series, then a well-designed algorithm can save (1) the time to run analysis/construct visualizations for the 90 time steps (2) the time to "sort through" the 100 visualizations to extract the main insights.

However, as previous works in this area have highlighted, designing an effective time selection algorithm is difficult. In particular, domain-agnostic selection algorithms, which are the main considerations in this work, are especially difficult as the underlying physics can be vastly different among different fields like astrophysics, climate modeling and fluid dynamics. In contrast, domain-specific algorithms have the advantage that they can exploit domain-specific characteristics of simulations. Work by Bennett et al. [3] is an example of a domain-specific selection algorithm for combustion simulations that tracks and rapid increases in heat releases to identify key time steps in such simulations. However, these domain-specific methods leverage domain knowledge and are only designed for particular simulation loads.

Another area in which time slice selection is relevant is in situ processing of scientific simulations. In situ processing refers to a processing model that analyzes scientific simulations *as* they run. This contrasts with the more traditional post hoc processing model that analyzes scientific simulations after they complete. Under the post hoc model, a simulation would first dump each time step (or a subset of time steps) to storage. Once the simulation completes, the scientist would inspect the data to extract findings. While post hoc processing is still the dominant model in practice, in situ processing has been growing in interest and popularity over the last decade, primarily due to the I/O constraints on modern supercomputers [13, 7, 6, 16, 17]. To enable in situ processing, inspection routines called "triggers" have been proposed. The main idea behind triggers is to have a lightweight routine that inspects the simulation data that runs *alongside* a simulation. The task for triggers is to decide whether the current time step in a simulation warrants further action as defined by the user: visualization, analysis, further storage, etc.

The main challenge with in situ triggers is that they need to decide whether the current time step warrants further action without the knowledge of future time steps. That is, in situ triggers require time slice selection to be done *as* the simulation runs. Since a well-designed in situ trigger should "trigger" at key time steps of a simulation, one can interpret in situ trigger as a time slice selection algorithm as well - the only difference being whether the selection occurs at the end of the simulation with access to the entire data corpus or not. While the body of past work in in situ triggers is more sparse than post hoc time slice selection algorithms, there are approaches emerging, for example work by Yamaoka et al.

[35]. In all, successful time slice selection for scientific simulation is of significant interest to visualization community. Despite this, no past work has performed a comparative analysis of time slice selection methods, as well as their evaluation metrics. While the lack of comparative analysis of selection methods is troubling for simulation scientists interesting in applying one for their simulation, we believe that the lack of understanding of what each evaluation metric measures is also concerning, especially for future research in this area. Without such an analysis, how should one evaluate their new selection algorithm? Should the community be convinced if the new approach performs well on a metrics A but not on metric B? Even worse, what should one do if two metrics disagree? That is, what should one do if metrics A and B disagree whether one selection is superior to another?

Name	Description
$DATA$	Set of all $data_i$: $\{data_0, data_1..., data_{N-1}\}$
Selection	A subset of {0,1,...,N-1}, each representing a time step
Budget	The size of the selection. A selection with budget k is a Selection S such that $\|S\| = k$
Selection Method	A function $Method(DATA, k)$ that, given $DATA$ and budget k, makes a corresponding selection S
Evaluation Metric	A function $Metric(DATA, S)$ that, given $DATA$ and selection S, calculates some numeric score. Note that a higher score need not be better.

Table 1. Terminology used in this thesis. This table considers a simulation that runs for N time steps. For any $0 \leq i \leq N-1$, the data at the i^{th} time step is denoted $data_i$.

Inspired by *The Transfer Function Bake-off* by Pfister et al., [22], with this work, we present the *Time Slice Selection Bake-off*, to explore the space of time slice selection algorithms end evaluation metrics. Before we elaborate on our contributions, we clarify the terminology we employ in this work in Table 1.

In short, the contribution of this work is to illuminate similarities and differences between selection methods and evaluation metrics that are used in practice or proposed in literature. To this end, we perform both qualitative and quantitative analysis of selection methods and evaluation metrics over a range of datasets and budgets. That said, we recognize the space of selection methods, metrics, datasets, and budgets is prohibitively large to explore exhaustively, especially given the large number of selection methods and metrics. One premise of this thesis is that overall trends and characteristics can be understod by considering 10 data sets and 3 budgets. In particular, we use this approach to answer the following questions.

- **RQ1:** Which selection methods produce similar selections?
- **RQ2:** Do we observe notable *outlier* behavior?
- **RQ3:** How do selection methods compare over many metrics?

To address **RQ1**, we perform hierarchical clustering on each selection S, using a distance metric introduced in Chapter IV to identify similar and dissimilar selection methods.

RQ2 will be approached qualitatively by examining the selections that each method makes for each dataset and budget. It is important to note here that, for all of the datasets, we consider the “correct” selection for a particular budget to be unknown. Without the appropriate domain experts, it is difficult to establish an appropriate “ground truth,” even if such a selection exists. (We would argue that visualization experts and domain scientists would reasonably disagree on what an ‘optimal’ selection is for a given dataset and budget.) However, by focusing on the trends in the selections, we believe we can characterize the selection methods as well as point out clear failures conditions.

To address **RQ3**, for every selection method SM, we evaluate its selection S with every selection metric EM to identify if any method SM performs favorably across the set of metrics EM.

CHAPTER II

REVIEW OF TIME SLICE SELECTION METHODS AND EVALUATION METRICS

A number of past works have considered time slice selection methods for scientific simulations. This section reviews the past body of work in this area.

§2.1 reviews domain-agnostic approaches: reconstruction-based approaches (§2.1.1), information-theoretic approaches (§2.1.2) and other domain-agnostic approaches (§2.1.3). Finally, §2.2 reviews domain-specific approaches.

2.1 Domain-agnostic methods

2.1.1 Reconstruction-based approaches. The objective for reconstruction-based approaches is find a selection S, such that it minimizes a particular reconstruction error. The two main components for this set of methods are:

1. *Method for reconstruction:* Given some selection S, the reconstruction method generates reconstructed $data'_i$ for every $data_i \in DATA$ with access only to the data in the time steps in S.
2. *Error function between the original and reconstructed fields:* Given the original field $data_i$ and reconstructed $data'_i$, the error function quantifies the difference between the two.

Given these two components, the reconstruction-based approaches defines the *cost* of a selection S as follows. For some selection S, an error function f and the reconstructed set of $data'_i$:

$$cost = \sum_{0 \leq i \leq N} f(data_i, data'_i) \tag{2.1}$$

Finally, reconstruction-based approaches consider a selection S of budget k *optimal*, if $|S| = k$ and if it minimizes Eq. 2.1.

The most common method for reconstruction is linear interpolation, where the time slices in selection S are used to reconstruct every $data'_i$ by linearly interpolating between the closes available time steps. Zhou and Chiang used linear interpolation as the reconstruction method and proposed Variation of Information (VI) as the error function [37]. VI is a metric derived from information theory and measures the sum of *information* present in X after observing Y and in Y after observing X. In addition to VI, Zhou and Chiang also considers root-mean-square error (RMSE).

Alternatives to linear interpolation have also been proposed. Tong et al. proposed Dynamic Time Warping (DTW) as a method of reconstruction as a nonlinear time mapping method [27]. In this work, they employ Earth Mover's Distance (EMD) and isosurface similarity map[9] as the error function, but state that these were specifically chosen for the datasets that they considered (i.e., not domain agnostic). However, their DTW-based method does not depend on a particular error function; therefore, it can be used broadly with any error function.

For a given budget, the *optimal* selection for reconstruction-based approaches can be calculated through dynamic programming. That said, the time complexity for these methods grow cubically with respect to the number of time steps which can become impractical for large datasets. Hence, some have also proposed approximation methods to decrease the time complexity [37].

Reconstruction-based approaches are common in evaluation metrics as well. As an evaluation metric, the cost from Eq. 2.1 is directly used for any selection S. The following three works have used a reconstruction-based approach for evaluation, all using linear interpolation as the method for reconstruction. First, Kawakami et al. proposed an evaluation framework for in situ triggers using linear interpolation along with L1 norm of difference as the error function [11]. Second, Porter et al. evaluated their selection method using linear interpolation along with and root-mean-squared-error (RMSE) and peak signal-to-noise-ratio (PSNE) as the error function [23]. Third, Pulido et al. evaluated their selection method using linear interpolation along with image quality metrics like structural similarity index measure (SSIM) and universal quality index (UQI), among others like RMSE, as the error function [24].

2.1.2 Information-theoretic approaches. The field of scientific visualization has incorporated concepts from information theory many times to solve challenges [30, 5], including time slice selection. Previously mentioned work by Zhou and Chiang that used variation of information as the error function is one example [37]. Wang et al. formalized the notion of *importance* values, based on a sliding time window and conditional entropy values with adjacent time steps. By using applying this metric on a block-partitioned dataset, they proposed to select time slices by maximizing joint entropy of the adjacent time steps in selection S [31]. Shannon's entropy, which measures the amount of "information" in data, has also been proposed for in situ triggers, where trigger would 'fire' if the change in entropy from the last trigger instance exceeds some predefined value [13].

2.1.3 Other domain-agnostic approaches. Frey and Ertl presented a flow-based approach for time slice selection, where the difference between two time steps is quantified via a *flow-based* metric [10]. They quantify the difference between two time slices in three steps: (1) probabilistically sample from data volume, (2) construct a flow graph between the two samples, and (3) calculate the minimum-cost flow in the said graph. Finally, a selection S is made such that it minimizes the difference with respect to the original time series. Ling et al.[14] presented a time slice selection method that relied on detecting spatial deviations and temporal deviations for simulations running in a multi-process setting. Their work had each processor to maintain a KDE-estimated PDF of the variables in the simulation. They used an ensemble of decision tree regressor to detect whether (1) data in a particular processor exhibited different behavior than the rest (spatial deviations) (2) data at current time step exhibited different behavior than the previous time step (temporal deviations). Pulido et al. developed a time slice selection method based on non-negative Tucker factorization (NNTF) [24]. The key idea for their work was to use a non-negative Tucker decomposition on the 4D tensor (x,y,z,time) and extract “influential” time points as the key time steps. Their use of non-negative Tucker decomposition involved representing the simulation data as a 4D tensor (x,y,z,time) and extracting “influential” time steps as the time slices selection.

Some selection methods have specifically considered an *in situ* setting. Myers et al. presented a time step selection algorithm that fitted a piecewise linear model to the data stream as the simulation ran [19]. Each incoming time step would be evaluated on the linear model, until the precision of the previous model drops to predefined level. Once that happens, the algorithm would start a new linear model, and declare that time step as *salient.* Yamaoka et al. developed an in situ time slice selection algorithm that leveraged KL divergence to detect changes in the state of the simulation [35]. Like Ling et al., they target a multi-process simulation run, and attempt to track changes in the simulation by calculating the

KL divergence between the data in the current time step and the last time step when the algorithm made a selection.

There is also a growing body of work utilizing deep learning to time slice selection [23, 36, 15]. Porter et al. trained a encoder-decoder network to implicitly learn feature descriptors of each time step in a latent space [23]. Once the network is trained, they perform further dimensionality reduction from 1024 to 2 dimensions via t-SNE, before making selections based on the "path" that each time step forms in the projected 2D space. Zhang et al. developed a time slice selection method based on a neural network that, given data from two time slices, reconstructed the data for the intermediate time steps [36]. Based on this data reconstruction module and a predefined error bound, they performed time slice selections *in situ*. However, it should be noted that their method specifically targets ensemble simulations, where the data reconstruction module can be trained on a simulation data with similar characteristics prior to use.

2.2 Domain-specific approaches

In contrast to domain-agnostic approaches, domain-specific time slice selection approaches can incorporate prior knowledge of the simulation and its dynamics. Bennett et al. presented a time slice selection method for combustion simulation by predicting rapid heat releases in these simulations [3]. This work was continued by Salloum et al., who proposed a new metric leveraging the coefficient of variation to capture rapid heat release in combustion simulation that was more accurate and increased the robustness of trigger detection [26]. Banesh et al. presented a method to detect mesoscale ocean eddies in large simulated or observational ocean data [2]. Their method was an extension to the previously introduced work by Myers et al. [19] that utilized a piecewise linear model to model the simulation. Liu et al. utilized a Siamese network comprised of two CNNs to identify key time steps in flow field data for Computational Fluid Dynamics (CFD) simulations [15]. A key idea in their work to use a Siamese network to assess the similarity between data in two time steps. While their method outperformed the Myers' method [19] for their CFD data, NN-based methods like this method and the ones introduced in §2.1.3 need to be trained before being applied, which can be time consuming especially as the data sizes increases.

CHAPTER III

EXPERIMENTAL OVERVIEW

3.1 Data sets

This work consider 10 different datasets, which are described in Table 2.

Name	Variable of Interest	Description
Asteroid	tev, Temperature in electronvolt (eV)	Simulation of an asteroid impact with ocean $300 \times 300 \times 300$, 237 time steps [21]
Cloud	cct, Cloud top pressure in Pascals	Atmospheric simulation of clouds in Germany $1429 \times 1556 \times 1$, 481 time steps [8]
Droplet	NA, Only one field	Simulation of two droplets colliding $256 \times 256 \times 256$, 167 time steps [4]
EarthquakeMag	Magnitude of velocity	Earthquake simulation $750 \times 375 \times 100$, 227 time steps [1]
HurricanePressure	Pressure in Pascals	Atmospheric simulation of Hurricane Isabel $500 \times 500 \times 100$, 48 time steps [20]
MantleTemperature	Temperature in Kelvin	Earth mantle convection simulation $180 \times 201 \times 360$, 251 time steps [25]
Jet	NA, Only one field	Turbulent combustion simulation $480 \times 720 \times 120$, 121 time steps
Bottle	NA, Only one field	Laser pulse shooting through a bottle 900×430, 465 time steps [29]
Cloverleaf	Energy	Hydrodynamics simulation $260 \times 132 \times 260$, 91 time steps [28]
Radiation	Gas temperature in Kelvin	Cosmology simulation of radiation waves $600 \times 248 \times 1$, 200 time steps [34]

Table 2. Data sets

3.2 Selection methods

The selection methods we consider in this thesis can be broadly be separated into two categories: (1) Reconstruction-based approaches and (2) Information-theoretic approaches. In addition, §3.2.3 introduces two methods based neither on reconstruction nor information theory.

3.2.1 Reconstruction-based approaches. Reconstruction-based selection methods were introduced in §2.1.1. Recall that there are two components to reconstruction-based selection methods: (1) The method to reconstruct every

$data'_i$ and (2) error function, f, used to evaluate the dissimilarity between $data_i$ and $data'_i$. This thesis considers two reconstruction methods: linear interpolation for every mesh location (LI) and Dynamic Time Warping (DTW)[27]. Furthermore, four error functions are considered: L1 norm of differernce (L1N), L2 norm of difference (L2N), mean squared error (MSE) and Variation of Information (VI) using 128 bins as used in [37]. Intuively, Variation of Information measures the difference of *information* contained between two random variables X and Y. The Venn diagram in Fig. 1 provides some graphic intuition into $VI(X, Y)$. A key property for reconstruction-based approaches is that it enforces that the first and final time step be included in the selection. This is to ensure that the reconstruction methods have the sufficient information to reconstruct the set $data'_i$ for every i. Lastly, recall from §2.1.1 that these selections can be made via dynamic programming. In total, $2 \times 4 = 8$ reconstruction-based selection methods are considered.

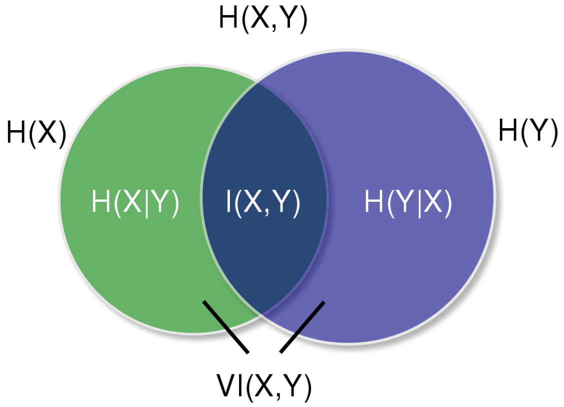


Figure 1. Variation of Information between two random variables, X and Y, as Venn diagrams.

3.2.2 Information-theoretic approaches.

The information theoretic approaches we consider are the joint entropy approach and the importance-based selection introduced by Wang et al. [31]

The joint entropy method attempts to maximize the joint entropy in selection S. For some selection $S = \{s_0, s_1, ..., s_{n-1}\}$, the joint entropy method defines the quality of S of budget n as in Eq. 3.1.

$$\text{Selection quality} = \sum_{0 \leq i \leq n-2} JE(data_{s_i}, data_{s_{i+1}}) \tag{3.1}$$

The selection S that maximizes Eq. 3.1 is considered *optimal* under the joint entropy method.

Eq. 3.1 was used as a evaluation metric in [32], but, of course, can also be used as a selection method by maximizing the joint entropy. In a similar fashion to the calculations for reconstruction-based approaches, a dynamic programming method can be employed to calculate the best selection for some budget k.

The importance-based approach by Wang et al. first subdivides the data into equal sized blocks. Next, the *importance* of each block is defined as the sum of conditional entropies of the block with those at the same spatial location in adjacent time steps. Specifically, for some data block X_j at time t, Wang et al. defines its *importance* as seen in Eq. 3.2, where $Y_{j,t}$ refers to the jth block at time step t. Finally, the *importance* of a time step is defined as the sum over all blocks as in Eq. 3.3.

$$A_{X_j,t} = 0.5 \cdot H(X_{j,t}|Y_{j,t-1}) + 0.5 \cdot H(X_{j,t}|Y_{j,t+1}) \tag{3.2}$$

$$A_t = \sum_i A_{i,t} \tag{3.3}$$

With the *importance* of each timestep, Wang et al. then partitions the time series into contiguous sections with approximately equal sum of *importance* values. Finally, a time step from each contiguous partition is chosen such that the joint entropy of every selection in S is maximized. Hence, the importance-based method can be thought of a variant of the joint entropy methods, except that it restricts the pool from which the selections can be made. The size of each block used in this paper for each data set is listed in Table 4 and details of the method can be found in [31].

Both joint entropy and importance-based selection methods requires a histogram to be constructed for each time step, thus requires the number of bins to be defined. In this paper, we consider four bin counts: 8, 64, 128, and 256 bins. Therefore, $2 \times 4 = 8$ total information-theoretic methods are considered.

3.2.3 Flow-based* method and uniform selection. In this work, we also consider the flow-based method by Frey and Ertl [10]. However, we make one key modification: we enforce that the first and last time step be included in the selection. For any budget k, recall that reconstruction-based methods effectively have $k-2$ degrees of freedom. To ensure that comparisons are valid across methods with the same budget, this paper considers a modified flow-based* method that enforces the inclusion of the endpoints. We also consider the uniform selection which places selections evenly throughout the time series.

Also note that a number of selection methods can also be directly used as an evaluation metric as well. The full set of selection methods this work considers and whether they are also valid metrics are shown in Table 3.

Name	Description	Can be a evaluation metric
LIL1N	Linear interpolation with L1 norm of difference	Yes
LIL2N	Linear interpolation with L2 norm of difference	Yes
LIMSE	Linear interpolation with mean squared error of difference	Yes
LIVI	Linear interpolation with Variation of Information	Yes
DTWL1N	Dynamic time warping with L1 norm of difference	Yes
DTWL2N	Dynamic time warping with L2 norm of difference	Yes
DTWMSE	Dynamic time warping with mean squared error of difference	Yes
DTWVI	Dynamic time warping with Variation of Information	Yes
JE8	Joint entropy with 8 bins	Yes
JE64	Joint entropy with 64 bins	Yes
JE128	Joint entropy with 128 bins	Yes
JE256	Joint entropy with 256 bins	Yes
Importance8	Importance-based selection with 8 bins	No
Importance64	Importance-based selection with 64 bins	No
Importance128	Importance-based selection with 128 bins	No
Importance256	Importance-based selection with 256 bins	No
Flow-based*	Modified version of the flow-based method [10]	Yes
Uniform	An evenly-spaced selection	No

Table 3. Selection methods

Data set	Block size
Asteroid	$20 \times 20 \times 60$
Cloud	$50 \times 50 \times 1$
Droplet	$16 \times 16 \times 16$
EarthquakeMag	$50 \times 30 \times 10$
HurricanePressure	$25 \times 25 \times 10$
MantleTemperature	$20 \times 20 \times 20$
Jet	$24 \times 26 \times 20$
Bottle	$45 \times 25 \times 1$
Cloverleaf	$23 \times 20 \times 23$
Radiation	20×20

Table 4. Block sized used for the importance-based approach [31]

CHAPTER IV

SELECTION SIMILARITY

4.1 Overview

In short, this section addresses **RQ1**: which selection methods produce similar selections? In other words, we want to cluster the selection methods in Table 3 into those that exhibit similar behavior. The importance of such an analysis is to gain a better understanding of how selection methods relate to each other across the range of data sets and budgets. To this end, we apply two passes of agglomerative clustering.

1. Agglomerative clustering on the result that each selection method makes for a particular data set *and* a particular budget.
2. Agglomerative clustering on the result that each selection method makes across *all* data sets for a particular budget.

For example, the first clustering pass would cluster selection methods based on its results on the `asteroid` data set for a budget of 5, while the second pass would cluster selection methods on its results on *all* data sets for a budget of 5. In order to quantitatively cluster selections using a hierarchical clustering model, a metric is required to capture the dissimilarity between two selections.

4.2 Selection difference

Given two different selections of budget k, S_1 and S_2, we want to define a *selection difference* $SD(S_1, S_2)$, to capture the dissimilarity between the two. In order to define SD, we borrow ideas from alignment-based metrics based on time series analysis. In particular, consider two alignment schemes: Euclidean alignment and Dynamic Time Warping alignment. (An important note: The Dynamic Time Warping (DTW) *alignment* introduced here is entirely different than the DTW-based selection method by Tong et al [27]. Though they are both based on the same idea from time series analysis, the two should not be confused.)

For two time slice selections of budget k, $S_1 : \{x_0, ..., x_{k-1}\}$ and $S_2 : \{y_0, ..., y_{k-1}\}$, an *alignment* between S_1 and S_2 is a set $P = \{p_0, ..., p_m\}$ of pairings p_i. Each p_i is pair of elements (x_i, y_j): one selection from S_1, and another from S_2.

A simple metric we can devise is to consider the difference between S_1 and S_2 as the summation of pairwise difference at each index. Aligning every i^{th} element with each other between two time series is referred to as the Euclidean alignment[33] (i.e. for any i, $p_i = (x_i, y_i)$). The difference metric that arises from this alignment can be written as follows:

$$\text{Euclidean alignment cost}(S_1, S_2) = \sum_i abs(S_1[i] - S_2[i]) \tag{4.1}$$

However, as noted by Keogh and Pazzani, the Euclidean alignment between time series is highly sensitive to small distortions in the time axis [12], thus, making it less appealing in many scenarios.

Thus, the more commonly used alternative, and the method used in this paper, is the Dynamic Time Warping (DTW) alignment [12], which allows for a more flexible and intuitive dissimilarity metric between two time series. A valid DTW alignment satisfies the following three conditions.

1. The first and last indices are paired together. In other words, $(x_0, y_0) \in P$ and $(x_{k-1}, y_{k-1}) \in P$.
2. Every x_i, y_j appears at least once in a pairing in P.
3. The pairings between S_1 and S_2 must be monotonically increasing. In other words, for any $x_i, x_j \in S_1$ with $i < j$, its respective pairing $y_s, y_t \in S_2$ must obey $s \leq t$, and vice versa. The monotonicity of pairings is enforced to preserve the time ordering in the pairings.

Finally, the optimal DTW alignment is the set of pairings P that minimizes the difference in each pairing p_i. Formally, for two selection S_1, S_2, and the set of all valid DTW alignments, denoted $\mathcal{P}$, the DTW alignment cost is defined as follows.

$$\text{DTW alignment cost}(S_1, S_2) = \min_{P \in \mathcal{P}} \sum_i abs(P_i[0] - P_i[1]) \tag{4.2}$$

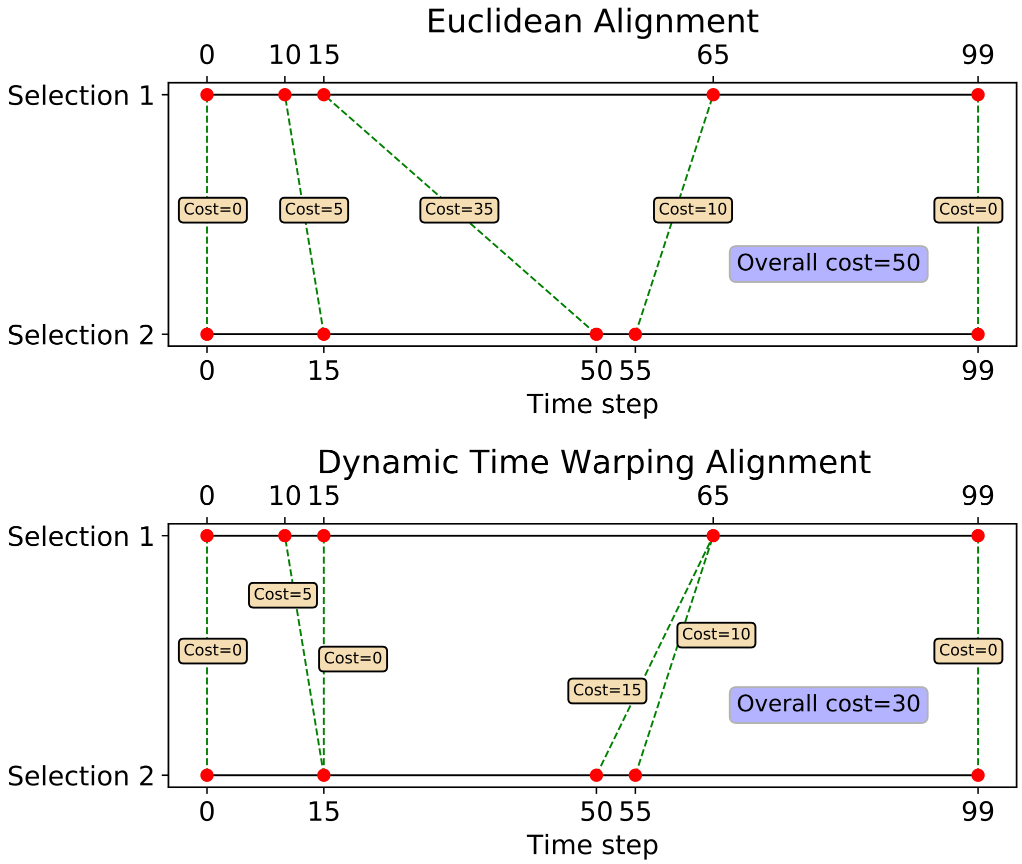
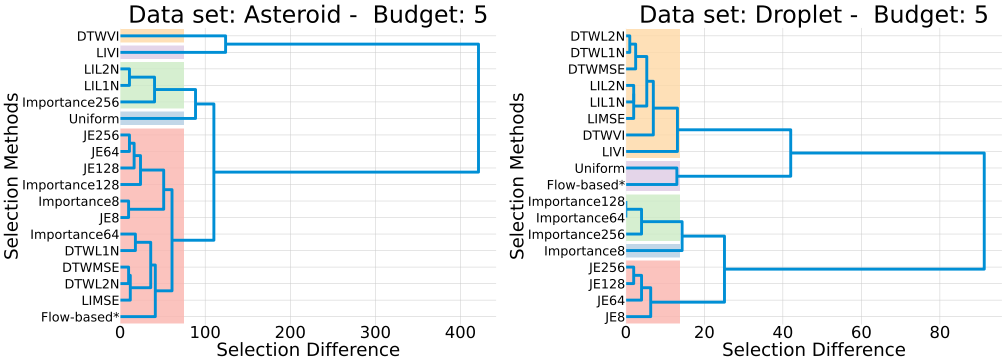


Figure 2. Euclidean alignment (top) and Dynamic Time Warping (bottom) alignment between two selections: {0, 10, 15, 65, 99} and {0, 15, 50, 55, 99}. The dotted lines denote the pairings that would be produced under the respective alignment schemes. Notice that Euclidean alignment has cost of 0+5+35+10=50, while the DTW alignment has cost of 0+5+0+15+10+0=30.

Fig. 2 plots a notational relation between the Euclidean alignment cost and DTW alignment cost. As Keogh and Pazzani pointed out, the Dynamic

Time Warping alignment better describes the difference between two time slice selections. Here, the Euclidean alignment that forces the (15, 50) pairing paints a distorted picture. As a result, as mentioned prior, in this thesis, we employ the DTW alignment cost as $SD(S_1, S_2)$.



(a) Clustering for the `asteroid` data set with a budget of 5

(b) Clustering for the `droplet` data set with a budget of 5

Figure 3. Dendrograms from the per data set/budget hierarchical clustering of selection methods (i.e. the first clustering phase). Each clustering is also color-coded based on results if 5 clusters were to be extracted.

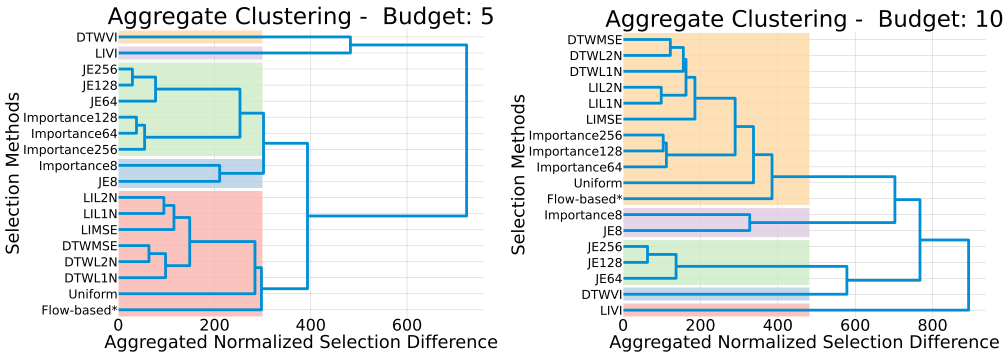
To calculate $SD(S_1, S_2)$, dynamic programming can be used, following the recurrence relation in Eq. 4.3. Let $DP[i][j]$ denote the *selection difference* between the first i elements of S_1 and first j elements of S_2. Then, the following holds.

$$\begin{aligned} DP[i][j] =& abs(S_1[i] - S_2[j]) + \\ & \min(DP[i-1][j],\ DP[i-1][j-1],\ DP[i][j-1]) \end{aligned} \tag{4.3}$$

Finally, notice that $DP[k][k]$ is equal to $SD(S_1, S_2)$ for any two selections S_1, S_2 of budget k.

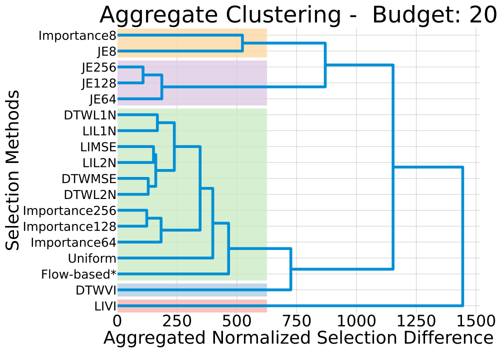
4.3 Hierarchical clustering

4.3.1 Overview. For a particular data set and budget, we determine similarities across the selection methods by applying agglomerative clustering with this SD metric. (Note that average-linkage clustering is used as the method to join



(a) Aggregate clustering using SD_{Norm} for budget of 5

(b) Aggregate clustering using SD_{Norm} for budget of 10



(c) Aggregate clustering using SD_{Norm} for budget of 20

Figure 4. Dendrograms from the aggregate hierarchical clustering of selection methods (i.e. the second clustering phase). Each clustering is also color-coded based on results if 5 clusters were to be extracted.

clusters in the hierarchical clustering process as recommended by Manning et al. [18].) However, this per data set/budget clustering only highlights local clusterings. Therefore, as noted in §4.1, in order to cluster the behavior of selection methods across data sets for a particular budget, we use agglomerative clustering on the aggregated selection differences across all data sets.

There are number of ways to aggregate selection differences across data sets. The most simple method, and perhaps obvious, is a simple addition of SD values. In other words, we can assess the difference between two selection methods for some budget k by summing SD value for each of the 10 data sets. However, simple addition of SD across data sets can be a problem, since the number of time steps in each data set varies. The range of values that SD can take on increases for data sets with longer time series; hence, simple addition of raw SD values will inflate contributions from data sets with more time steps. Therefore, for the aggregate hierarchical clustering phase, we devise a normalized SD metric, to overcome this issue.

4.3.2 Normalized SD.

In essence, for the normalized SD, or SD_{Norm}, we consider each element in a selection, not as a time slice index, but as a percentage of the whole time series. By way of example, consider a data set with 50 time steps, and a selection for this data set $\{0, 10, 15, 49\}$. In the SD_{Norm} calculation, we instead treat this selection as $\{0\%, 20.4\%, 30.6\%, 100\%\}$ or $\{0, 20.4, 30.6, 100\}$, and proceed with the same calculation as in Eq. 4.2. This transformation is permissible for a several reasons. First, notice that this linear transformation does not affect the local per data set/budget clustering from the first clustering phase, since SD and SD_{Norm} would bring rise to the same clusters in the first per data set/budget clustering phase. Second, the number of time steps in a particular data set is often arbitrary. Simulation scientists will often decide the rate to dump the simulation data to disk (e.g. every 5 simulation cycles). Therefore, we argue the actual time step index of the selection is less interesting than where it is temporally located in the time series.

The key, desirable property of SD_{Norm} is that it sidesteps the issue that SD had under simple addition across data sets. Since the values are all within [0, 100], simple addition of SD_{Norm} can be used to cluster the behavior of selection methods across all data sets. (We cannot, however, use SD_{Norm} to cluster across different budgets, since the range of SD_{Norm} increases as the budget increases.)

With SD_{Norm} established, we define the pairwise difference between two selection methods SM_1, SM_2 for a budget k in the aggregate clustering phase in Eq. 4.4. Let D denote the set of all considered data sets. Then,

$$\text{Aggregated difference}(SM_1, SM_2, k) = \sum_{dset \in D} SD_{Norm}(SM_1(dset, k), SM_2(dset, k)) \tag{4.4}$$

4.3.3 Results.

4.3.3.1 Results from the first clustering pass (per data set & budget clustering). For this selection similarity analysis, we consider three budgets: 5, 10, and 20. This analysis does not consider larger budgets for two reasons: (1) As described in the introduction, time slice selection algorithm are often used to generate a short summary of the data set. Therefore, comparisons at lower budgets are more meaningful. (2) Differences between selection methods are more accentuated at lower budgets. Now, since there are 10 data set under consideration, this yields $10 \times 3 = 30$ per data set/budget clusterings, and 3 aggregate clustering using SD_{Norm} for each budget. Instead of analyzing all 33 clustering results and dendrograms, we highlight five clustering results and show their dendrograms in Figs. 3 and 4.

Figs. 3a and 3b consider the `asteroid` and `droplet` data set, respectively, at a budget of 5. Notice that the clustering results between the two share some similarities, but also have key differences. Some observations follows.

- The joint entropy methods (i.e. *JE8*, *JE64*, *JE128*, and *JE256*) make similar selections in both data sets.
- DTW-based methods except *DTWVI* (i.e. *DTWL1N*, *DTWL2N*, *DTWMSE*) make similar selections for both data sets.

- The relative behavior of the importance-based method (i.e. *Importance8*, *Importance64*, *Importance128*, *Importance256*) is different. While their behavior are similar for the `droplet` data set, the behavior is is more varied for the `asteroid` data set.
- LI-based methods except *LIVI* (i.e. *LIL1N*, *LIL2N*, *LIMSE*) make similar selections for the `droplet` data set, but for the `asteroid` data set, *LIMSE* performs differently than the other two.

4.3.3.2 Results from the second clustering pass (aggregate clustering). With the 30 local clusterings as mentioned prior, however, exploring each clustering and their differences is (1) time consuming and (2) less interesting than characterizing general behavior over all datasets. Instead, this section will mainly focus on the results from Fig. 4 to answer the question posed in §4.1 - which selection methods produce similar results? In the aggregate, we make the following observations.

- *LIL1N*, *LIL2N*, *LIMSE*, *DTWL1N*, *DTWL2N*, and *DTWMSE* all produce similar results for all data sets for all budgets.
- *JE64*, *JE128*, *JE256* produce similar results for all data sets for all budgets.
- *Importance64*, *Importance128*, *Importance256* produce similar results for all data sets for all budgets.
- Flow-based* generally performs most similar to *LIL1N*, *LIL2N*, *LIMSE*, *DTWL1N*, *DTWL2N*, and *DTWMSE*.
- *JE8* and *Importance8* are most similar to each other, and generally dissimilar to the rest.
- VI-based methods (i.e. *LIVI*, and *DTWVI*) are outliers. LIVI forms a single leaf at budgets 10 and 20. At budget 5, though it forms a cluster with

DTWVI, LIVI is still significantly different that the rest of the selection methods. Despite it following the same reconstruction-based approach as methods like *LIL1N*, their behaviors present a significant departure from them. The reason for this departure is explored in Chapter V.

4.3.3.3 Observations on the behavior of the joint entropy selection method. A key difference that separates reconstruction-based and joint entropy selection method is whether they enforce some level of temporal spacing in their selections. Reconstruction-based selection methods, by way of their design, tend to prefer selections that are temporally spaced out. In fact, this naturally follows from how the cost for reconstruction-based methods are defined. Regardless of the specific reconstruction paradigm, these methods hinge on finding time steps in the simulation from which reconstruction of other time steps are best achieved. Therefore, skipping many intermediate time steps tends to introduce high error in the cost calculation, leading it to prefer to spread out its selections. In contrast, joint entropy methods pay little attention to how selections are "spread" out temporally, since no reconstruction is done. Thus, there is no inherent penalty for skipping many intermediate time steps in adjacent selections - they only serve to maximize the joint entropy of adjacent selections. This is seen in Fig. 7 and Figs. 10b and 10c. Notice that the joint entropy methods, in the `earthquakeMag`, `asteroid`, and `jet` data sets, tend to cluster their selection more in comparison to the reconstruction-based methods. This is not to say, reconstruction-based selection methods never clump their selections. Especially, at higher budgets, (e.g. `earthquakeMag` with budget 20) reconstruction-based methods do make clumped selections. We merely observe that the tendency to clump selections is higher for the joint entropy methods, providing insight into why the clustering analysis often classified them as dissimilar.

CHAPTER V

FAILURE CONDITIONS

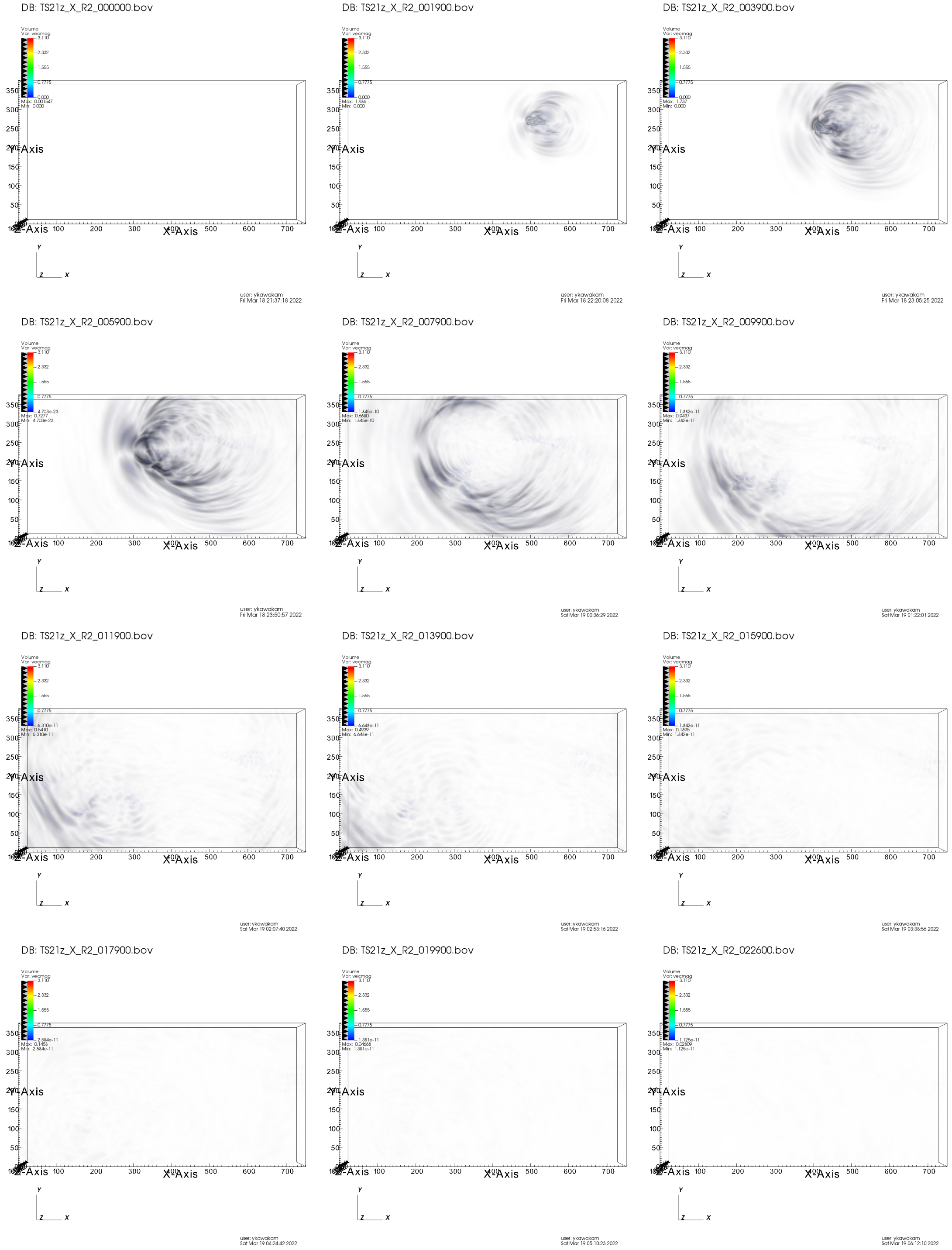
5.1 Overview

This section mainly addresses **RQ2**: do we observe notable *outlier* behavior? Now, as mentioned in the introduction, establishing the “correct” selections for a particular combination of data set of budget is difficult without consulting a domain scientist. Therefore, this section points out *clear* cases where we observe selection methods and evaluation metrics misbehave.

5.2 Issues that data sets with high dynamic range encounter

In the context of this work, data sets with high dynamic range refers to data sets where the local minimum and maximum values of the each time step change drastically over the course of the time series. Among the data sets that considered, two data sets, `asteroid` and `earthquakeMag`, have high dynamic range. `asteroid`, for example, models an asteroid collision with a ocean that occurs at time step 4 in our data set. After this initial impact, the remaining time steps track the aftermath of the collision and how the energy disperses through the air and ocean. As a result, the range of values at each time step are drastically different. For examples the range of values (temperature in eV) at time step 4 is [0.01870, 2.350], while the range of values at time step 200 is [0.01075, 0.2057].

The `earthquakeMag` data set is a similar case. In this simulation of a magnitude 7.7 earthquake on the Southern San Andreas Fault, the rupture begins at time step 0, and continues until time step 54. After the fault rupture stops, the first waves reach the bounding box of the simulation by around time step 100. The waves continue to propagate the until the end of the time series. Fig. 5 shows volume rendering of the `earthquakeMag` simulation. Due to this nature of the



DB: TS21z_X_R2_000000.bov DB: TS21z_X_R2_001900.bov DB: TS21z_X_R2_003900.bov

DB: TS21z_X_R2_005900.bov DB: TS21z_X_R2_007900.bov DB: TS21z_X_R2_009900.bov

DB: TS21z_X_R2_011900.bov DB: TS21z_X_R2_013900.bov DB: TS21z_X_R2_015900.bov

DB: TS21z_X_R2_017900.bov DB: TS21z_X_R2_019900.bov DB: TS21z_X_R2_022600.bov

Figure 5. Volume renderings of magnitude of velocity in the `earthquakeMag` data set. Shown are time steps (right-to-left, top-to-bottom) 0, 19, 39, 79, 99, 119, 139, 159, 179, 199, and 226

simulation, the range of values (magnitude of velocity) at time step 39 is [0, 3.119], while the range of values at time step 200 is [0, 0.07701].

These drastic differences in ranges presents a challenge for any selection methods using histogram to represent data - namely the joint-entropy methods, *DTWVI* and *LIVI*. There are two options when constructing the histograms at each time step: (1) Use the local maximum and minimum values or (2) Use the global maximum and minimum values. Now, using local maximum and minimums for constructing the per time step histogram is problematic since this would effectively change resolution of each bin over the time series. Under this paradigm, entropy measures are can become practically useless since small variation of data can be distorted to represent high "information".

As a consequence, using the global maximum and minimum values for the histograms is the only reasonable option as this ensures that the Shannon entropy numbers are meaningful in the context of the whole time series. However, using global maximum and minimum values faces one key issue. In many cases, this leads to heavy underutilization of histogram bins.

Fig. 6 shows the histogram at each time step using the global maximum and minimum values for the `earthquakeMag` data set. Note that the histograms use a log-scaled y-axis. Further, with the exception of some of the time steps that correspond to the rupture expansion and the initial wave propagation (i.e. time steps 0-54), many histogram bins are unused, or have comparatively very little count. For selection methods and evaluation metrics that rely on histograms to represent the data, this proves to be a significant challenge.

Fig. 7 plots the result from each selection method for the `earthquakeMag` data set at a budget of 10.

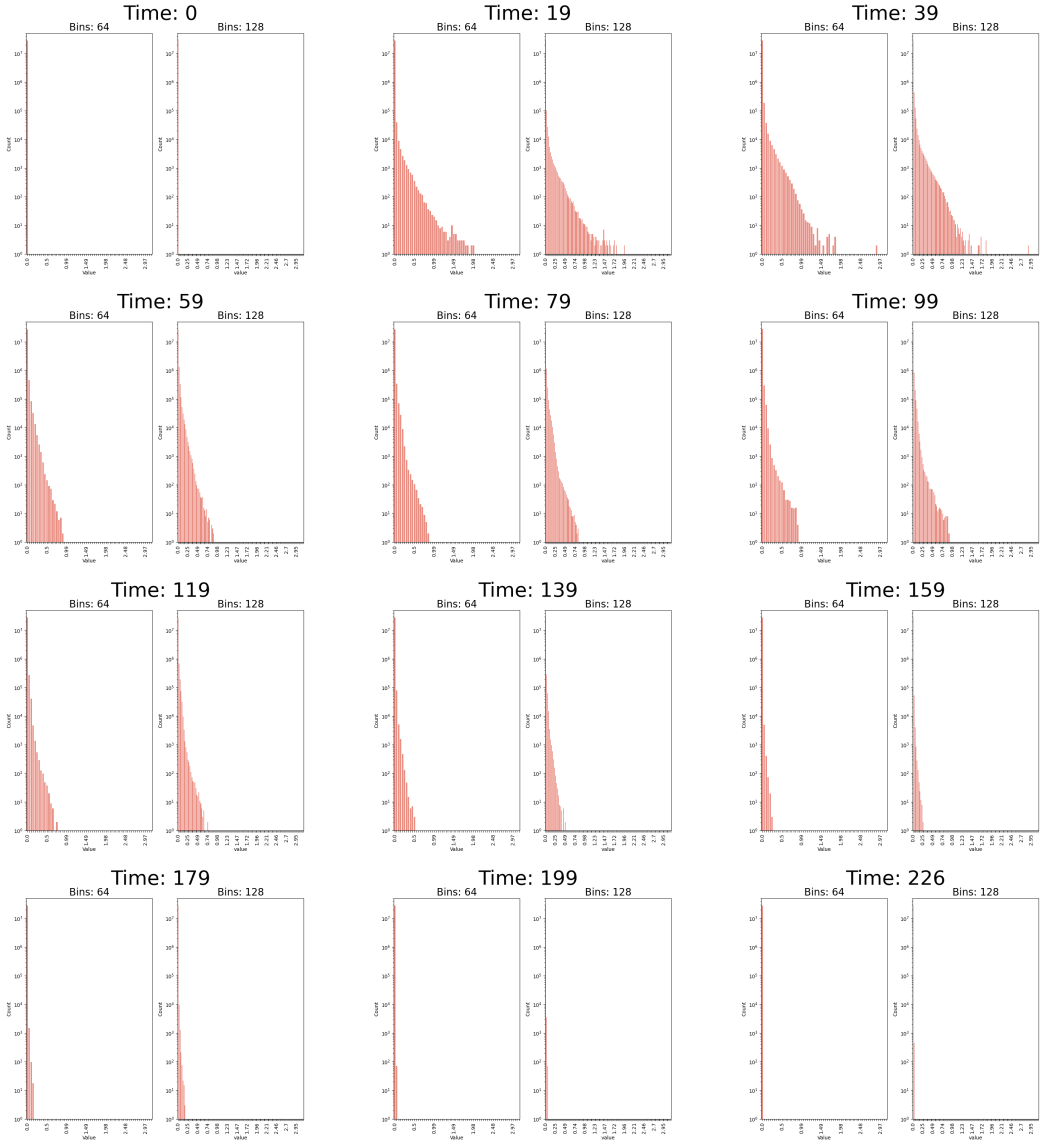


Figure 6. Histogram of time steps from the `earthquakeMag` data set using 64 and 128 bins. Shown are time steps (right-to-left, top-to-bottom) 0, 19, 39, 79, 99, 119, 139, 159, 179, 199, and 226.

5.2.1 Effects on joint entropy selection methods. Notice first the outlier behavior of the joint entropy approaches: *JE8*, *JE64*, *JE128*, and *JE256*. While other methods maintain some temporal resolution, purely maximizing the joint entropy of the selection yields selections that are heavily clumped around where the ruptures ends and starts to spread throughout the data set.

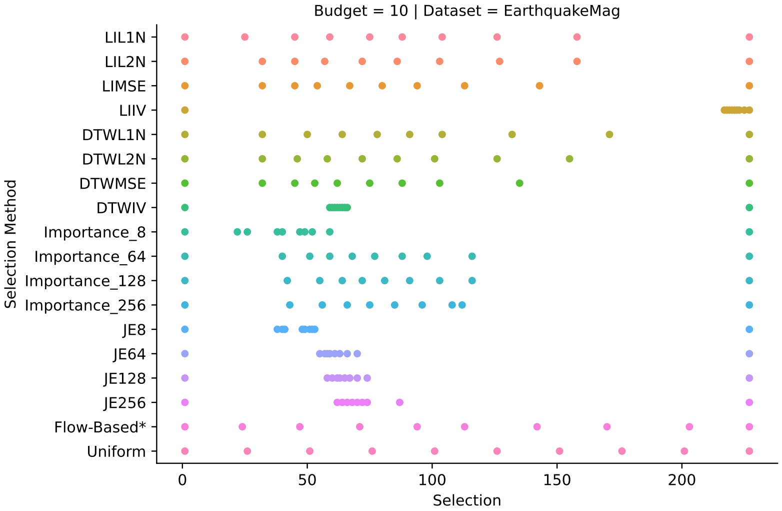


Figure 7. Result from each selection method at a budget of 10 for the `earthquakeMag` data set. Each dot corresponds to the time slice that each selection method makes.

(i.e. time steps 50-100) The underutilized bins in the other time steps, and the consequent empty joint histograms, imply that the joint entropy methods do not perceive these time steps as "worthy" for selection. These qualities lead to the selections by the joint entropy approaches as seen in Fig. 7. This may not be a true "'failure" condition, i.e., it may be the case that this is the only region of interest to domain scientists. That said, from the perspective of understanding the simulation as a whole, this selection certainly underdelivers, and it is a notable effect of underutilized bins for data set with high dynamic range.

5.2.2 Effects on reconstruction-based selection methods that use Variation of Information. Another notable case of the effects of underutilized bins is seen with the reconstruction-based methods that use Variation of Information as the error metric. Consider *LIVI*'s selection for the `earthquakeMag` data set in Fig. 7 and notice that it is clearly another outlier in relation to the rest. In fact, the selection method *LIVI* selects $\{0, 216, 217, 218, 219, 220, 221, 222, 224, 226\}$ as its selection for the `earthquakeMag` data set at budget of 10. The cause of this behavior can be seen by revisiting how Variation of Information is defined.

As seen in Fig. 1, at its core, Variation of Information between two random variables X and Y is the sum of conditional entropies, $H(X|Y)$ and $H(Y|X)$. Since reconstruction-based methods try to minimize

$$\sum_i VI(data_i, data_i')$$

where $data_i$ is the original data and $data_i'$ is the reconstructed data, it follows that this is equivalent to minimizing

$$\sum_i H(data_i \,|\, data_i') + H(data_i' \,|\, data)$$

A key property of conditional entropy between two random variables X and Y is that it always obeys $H(X) \geq H(X|Y)$. In other words, if $H(X)$ is very small, then $H(X|Y)$ must also be a very small quantity. If X contains little information itself, then the information in X *given* the information of Y is naturally very small.

Returning to the `earthquakeMag` data set and its histograms in Fig. 6, notice that the first and last time step have little *information* since most data is in one bin. (i.e. $H(data_0) \approx 0$ and $H(data_{226}) \approx 0.$). This in turns means that any linearly interpolated data between time step 0 and time step 226 also must have very little *information*. (This follows since linear interpolation will never introduce values outside of what it sees at the endpoints for each mesh location.) The effect of this behavior is that, under evaluation metric EM_{LIVI}, $\{0, 226\}$ is evaluated as a better selection than $\{0, 100, 150, 226\}$. The two selections are given scores 38.041 and 55.003 respectively. (Lower is better.) The reason for this behavior is as outlined above. Since the entropy of $data_i'$ is small if linearly interpolated between time step 0 and 226, $H(data'_i|data_i)$ also tends to be to a small amount, causing this odd effect to preferring no intermediate selections between 0 and 226. Minimization of Variation of Information can be achieved, if the reconstructed data had little to no *information* altogether. This, of course, is a clear failure condition. $\{0, 100, 150, 226\}$ as a selection contains all the information of $\{0, 226\}$; therefore, it should be evaluated at least as well as the former - not worse.

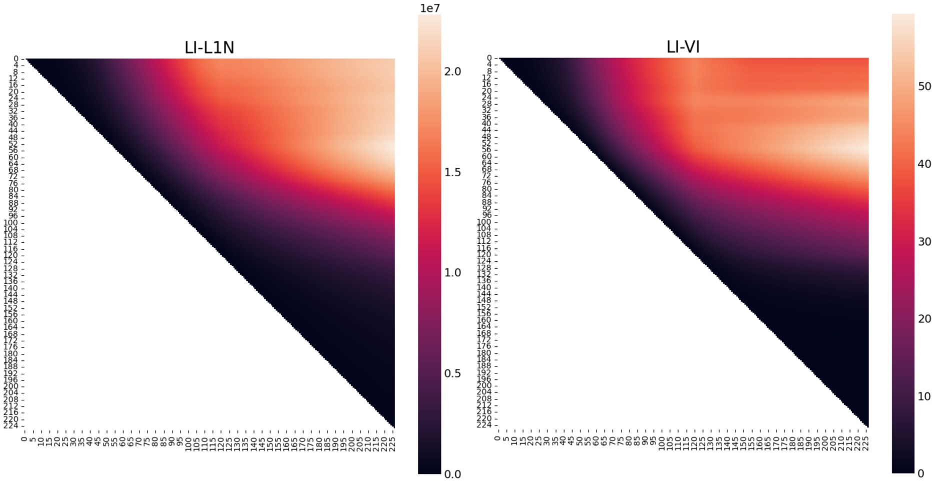


Figure 8. Pairwise mapping cost for the `earthquakeMag` data set: *LIL1N* on the right and *LIVI* on the left.

Fig. 8 shows the mapping cost for *LIL1N* and *LIVI* evaluation metric between any two time slices between 0 and 226 for the `earthquakeMag` data set. Every $[i, j]^{th}$ entry of the upper triangle heat map shows the cost between time step i and time step j under EM_{LIL1N} and EM_{LIVI}. (i.e. Linear interpolating every time step between time step i and time step j and summing the differences via *LIL1N* or *LIVI*.) The cause of outlier behavior from SM_{LIVI} can be seen in Fig. 8 as well. Notice the difference between the upper right regions of the two plots in Fig. 8. (i.e. Mapping costs between lower (0-15) time steps and higher (200-227) time steps.) While *LIL1N* assesses the comparatively high error in that region, notice that *LIVI* does not. This provides another explanation as to why the *LIVI* selection methods makes its selection for `earthquakeMag`. Similar outlier behavior is seen with the `asteroid` data set with *LIVI* and *DTWVI*.

CHAPTER VI

METRIC COMPARISONS

6.1 Overview

This section mainly addresses **RQ3**: how do selection methods compare over many metrics? To this end, for every selection method SM in Table 3, we evaluate its selection with every evaluation metric EM in Table 3. This analysis results in the following: for every selection method SM, a budget k, and a data set, we calculate 13 metric scores, each corresponding the score that evaluation metric assigns to the selection that a particular selection method makes. However, reporting on the raw metric scores loses some key information as it relates to what the scores mean in a relative sense. For example, consider two selections S_1, S_2, and two evaluation metrics EM_1, EM_2. (Assume that lower is better for these two metrics.) Suppose that their evaluations are as shown in Table 5. The proper interpretation of these raw results can be difficult. For example, is the difference between 1900 and 2000 for EM_2 a meaningful difference? Or, in the eyes of EM_1, is S_1 twice as good a selection as S_2? The key issue here is that raw scores from evaluation metrics lack context. Raw scores from each EM do not deliver any insight as to how each EM behaves over a range of different selections. To address this issue, for the purposes of metric comparison, this section considers instead the percentile of score among 5000 random selections rather than the raw scores itself. Specifically, when evaluating a selection S with an evaluation metric EM, we first

Evaluation Metric	Evaluation of S_1	Evaluation of S_2
EM_1	10	20
EM_2	1900	2000

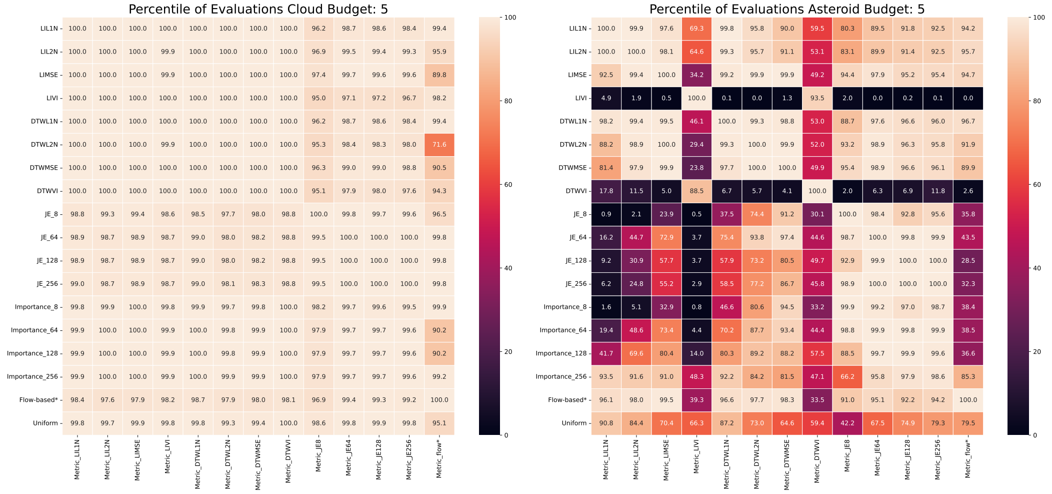
Table 5. Example of a set of evaluations

randomly generate 5000 selections of the same budget, evaluate each 5000 selections with EM and report the percentile of $EM(S)$ among the random 5000 selections. Note that a higher percentile is better for all metrics. For example, if a metric EM evaluates a selection S in the 75^{th} percentile, this means that, in the eyes of EM, the selection S is better than 75% of the 5000 random selections considered.

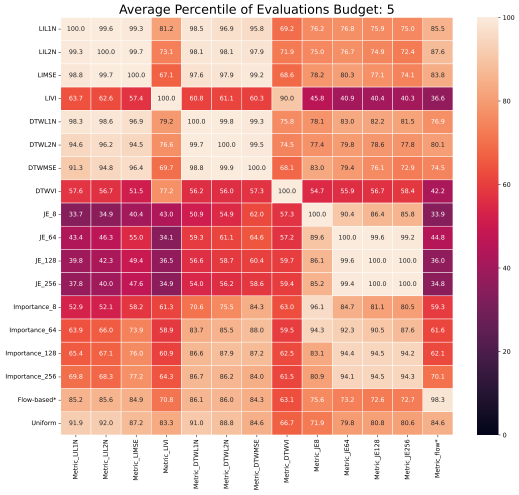
6.2 Results

Figs. 9a and 9b plot the results from the `asteroid` and `cloud` data sets at a budget of 5 as heatmaps. Fig. 9c reports the average percentile over all 10 data sets at a budget of 5. Each row in Fig. 9 corresponds to a selection method and each column corresponds to a evaluation metric. Therefore, each cell (i, j) in the heatmaps corresponds to SM_i's percentile among 5000 random selection under the metric EM_j.

Comparing Figs. 9a and 9b, the first obvious difference between the two is lack of poor evaluations for the `cloud` data set in Fig. 9a. Notice that every cell with the exception of $SM_{DTWL2N} - EM_{flow*}$ scores higher than a 90^{th} percentile. In contrast, for the `asteroid` data set in Fig. 9b, the heatmap shows many more combinations with poor evaluations. This difference is primarily due to the nature of the data set. The `asteroid` data set captures a asteroid collision with a ocean, while the `cloud` data set tracks cloud over a regular day in central Europe. As a result, the `cloud` data is far more static (i.e. less changes in the data) than the `asteroid` data set. Consequently, for the `cloud` data set, all selection methods make selection very similar to an uniform selection. Since the data stays fairly similar over time, the methods fall back to placing selections mostly evenly across the time series. This is seen in Fig. 10a. The takeaway from Fig. 9a is that for



(a) `cloud` data set (b) `asteroid` data set

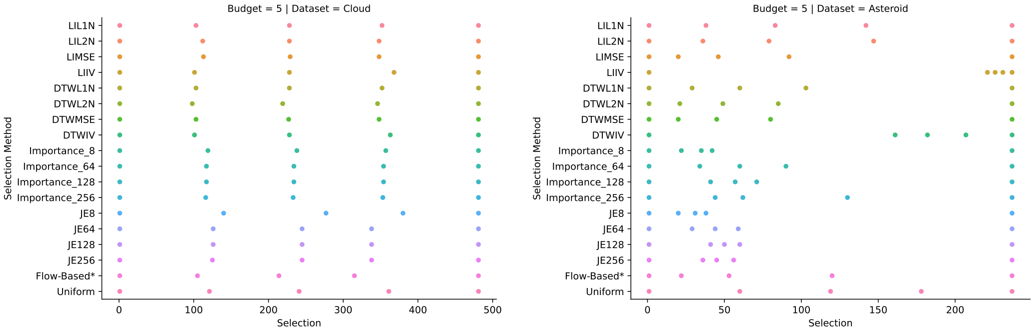


(c) Average of evaluation percentiles for all data sets

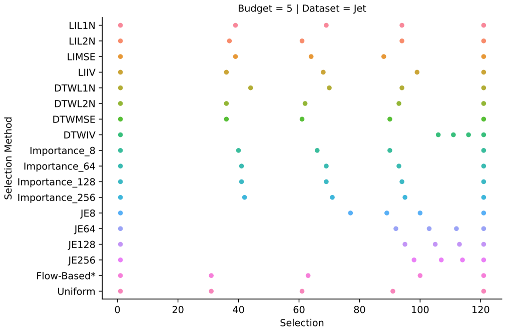
Figure 9. Percentile of evaluations among 5000 random selections at a budget of 5.

static dataset like `cloud`, all methods perform similarly, and since all methods make similar selections, all evaluation methods rate them similarily as well.

On the other hand, the `asteroid` data set tells a very different story. Not only do methods make different selections (as seen in Fig. 10b), but there are notable differences in how methods fare under the range of evaluation metrics.



(a) `Cloud` data set (b) `Asteroid` data set



(c) `Jet` data set

Figure 10. Results from each selection method at a budget of 5 for select data sets. Each dot corresponds to the time step that each selection method makes.

First, and perhaps most notably, notice the poor evaluation for SM_{LIVI} and SM_{DTWVI} across the board. Of course, since SM_{LIVI} and SM_{DTWVI} optimize for their respective metrics, EM_{LIVI} and EM_{DTWVI}, they perform at the 100^{th} percentile under them. However, under any other metric, they perform poorly, around the 5-10^{th} percentile range. The reason for this disparity becomes obvious in Fig. 10b: SM_{LIVI} and SM_{DTWVI} focus on a entirely different part of the time series than the other selection methods. The cause for this was discussed in Chapter V on failure conditions.

Second, we also notice an asymmetry between the non-VI reconstruction-based approaches (i.e. *LIL1N*, *LIL2N*, *LIMSE*, *DTWL1N*, *DTWL2N*, and *DTWMSE*) and the joint entropy approaches. While the reconstruction-based metrics poorly evaluates the selections that the joint entropy methods make, the same is not true for the converse. In fact, the joint entropy metrics favorably evaluates selections that reconstruction-based methods make as well as the joint entropy selection methods. Notice that for the `asteroid` data set in Fig. 9b, with the exception of *LIVI* and *DTWVI*, all combinations of reconstruction-based methods and joint entropy metrics yield a higher than 80^{th} percentile. But, several combinations of joint entropy methods and joint entropy metrics yield poor percentiles, some in the single digits. This trend holds across data sets as well, as seen in Fig. 9c. Notice that while the joint entropy metric evaluates all methods favorably, reconstruction-based metrics are more critical, especially the linear-interpolation based metrics.

Another observation we make is on the performance of uniform selection. For `cloud`, `asteroid` and overall, notice that uniform selection does fairly well across all metrics. In fact, in many instances, uniform selection is *preferred* by metrics over other methods that involve computations to generate. This result in both encouraging and discouraging. On the one hand, as a domain scientist, Fig. 9 shows that in many cases using a uniform selection will net better or equally as good results than applying a more sophisticated selection method. But, as a visualization researcher, this demonstrates that there is substantial room for improvement - beating a uniform selection should be a low bar considering its simplicity.

Finally, from Fig. 9c, we observe that the flow* metric evaluates similarily, compared to the non-VI reconstruction-based approaches. On average, the flow* metric favorably evaluates non-VI reconstruction-based approaches ($\sim 80^{th}$ percentile), poorly evaluates joint entropy approaches ($\sim 45^{th}$ percentile) and evaluates the importance-based methods in between the aforementioned two. These results are consistent with the clustering results from §4.3.3 that found the flow* selection method to make similar selections as non-VI reconstruction-based approaches.

Finally, examining the set of selection methods and their average performance over the set of evaluation metrics, we find the following

1. Non-VI reconstruction-based and flow-based* selection methods are favorably evaluated over our set of evaluation metrics, with the exception of EM_{LIVI} and EM_{DTWVI}.
2. The average performance of SM_{LIVI} and SM_{DTWVI} is poor due, in particular, to its issues outlined in Chapter V.
3. Joint entropy selection methods perform well under joint entropy evaluation metrics, but perform poorly under reconstruction-based metrics.

CHAPTER VII

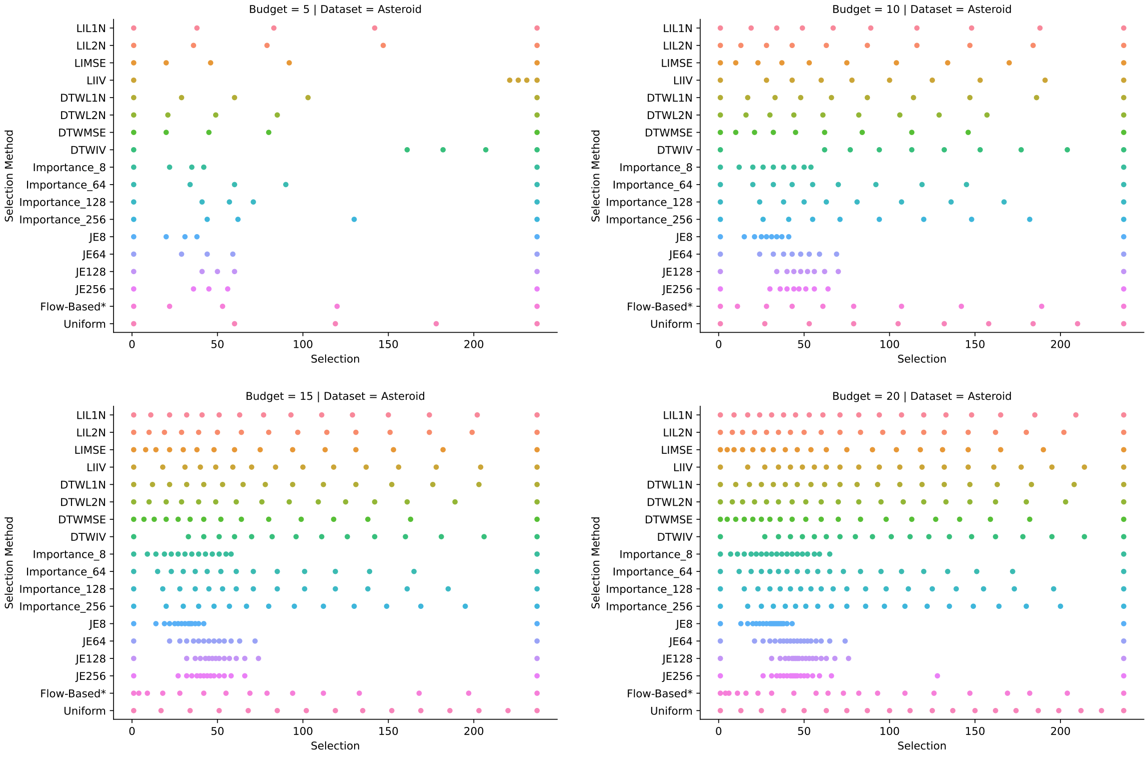
CONCLUSION

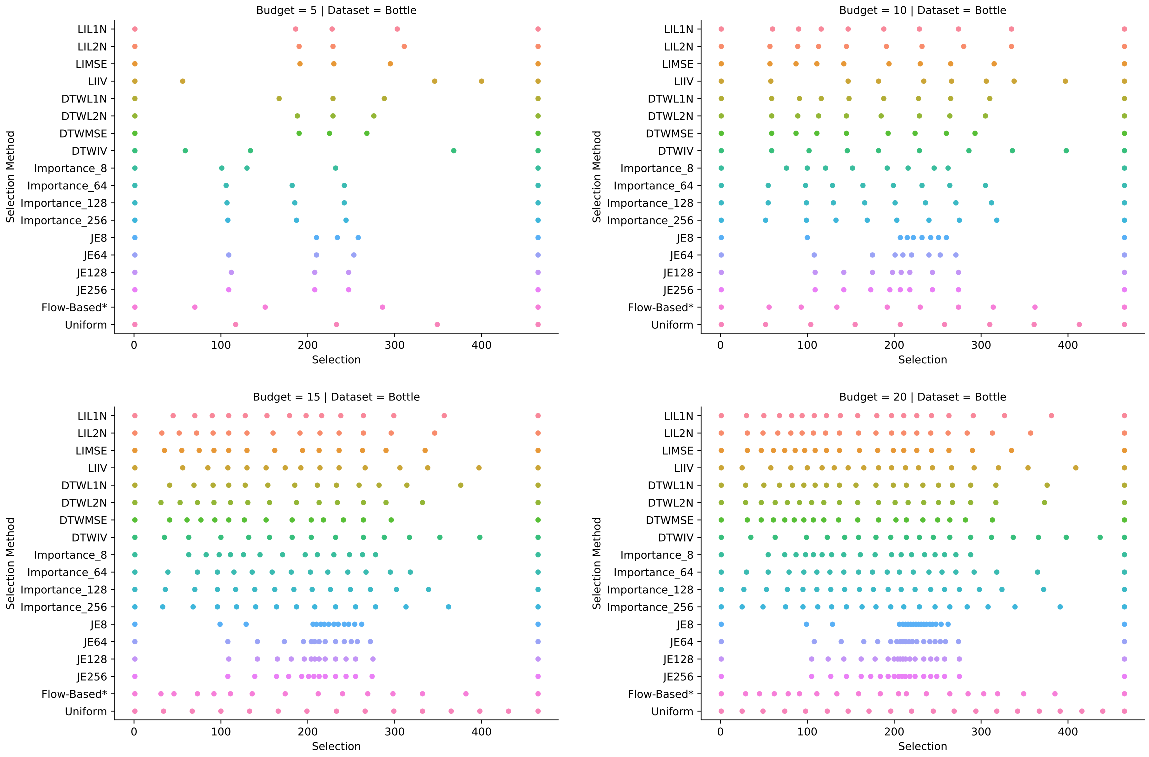
This thesis considered 10 data sets and 18 selection methods and explored the behavior of time slice selection methods over a diverse range of scientific simulations. Through a Dynamic Time Warping-alignment based selection difference metric and agglomerative clustering of selection methods, we identified similarities and dissimilarities between time slice selection algorithms in literature. Furthermore, we investigated failure conditions that some methods face. We identified that selection method that rely on histograms to represent data struggles when simulations have high dynamic range, since this leads to significantly underutilized bins in many cases. Lastly, by evaluating every selection method with every other evaluation metric, we illuminated the behavior of evaluation metrics, and identified selection methods that perform well across the set of all evaluation metrics.

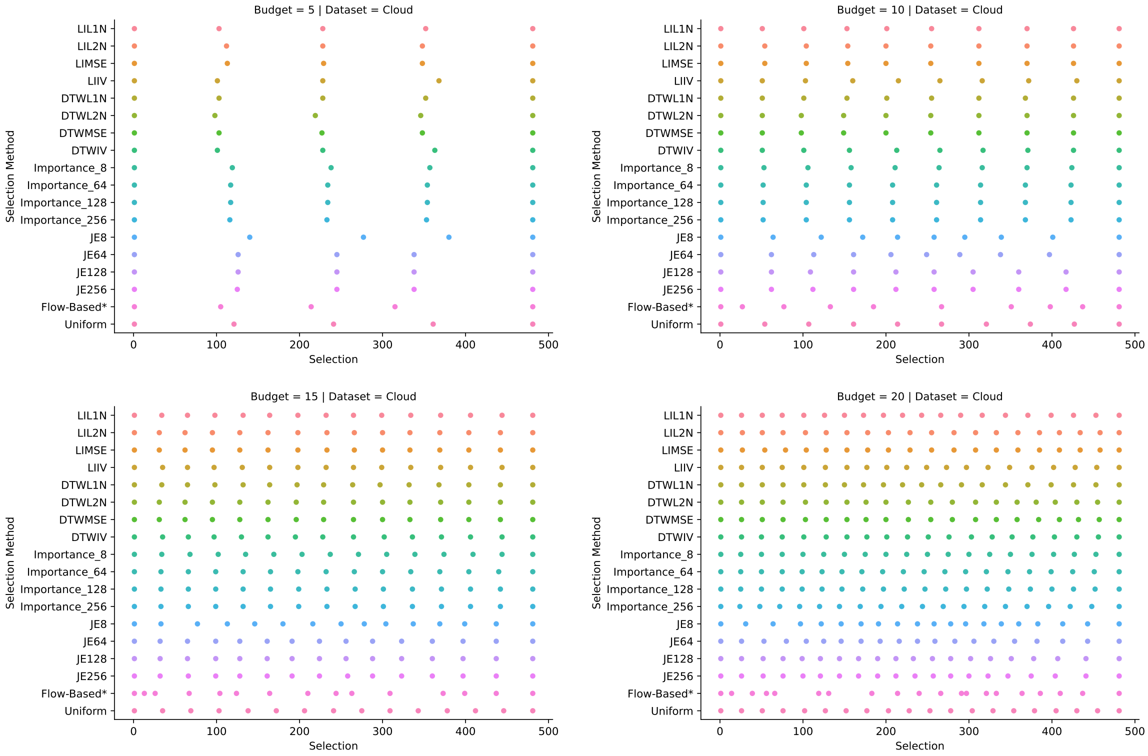
We believe this work can serve as a launching pad for further investigation into the behavior of selection methods. For example, the stability of selection methods is an important consideration. If the behavior of a selection method were significantly different for spatially subsampled data compared to the original data, this would be a cause of concern. Considering the stability of selection methods under temporal and spatial subsampling and other perturbations of the data, is an area of future work. Deep-learning based selection methods that were introduced in Chapter II may also be a worthy topic of investigation, considering the demonstrated power of deep learning in other domains of computer science. That said, this thesis has focused on domain-agnostic approaches, and deep-learning methods will need to innovate to become domain agnostic. Future work should include the aforementioned stability work, investigating deep-learning based selection methods, assessing how selection methods behave over different budgets, as well as expanding the set of considered data sets. Finally, we note that a user survey targeting domain scientists is critical to gain the full picture of the space of time slice selection methods. Recall from Chapter I that we currently consider the “correct” selection for any data set and budget to be unknown. Clearly, this is a significant obstacle in designing time slice selection methods since the real quality of selection methods can only be defined in relation to a known ground truth. In our case, the ground truth is what a domain scientist would select based on their knowledge of their domain and dynamics of a particular simulation. As such, we consider this user survey a critical future work as well.

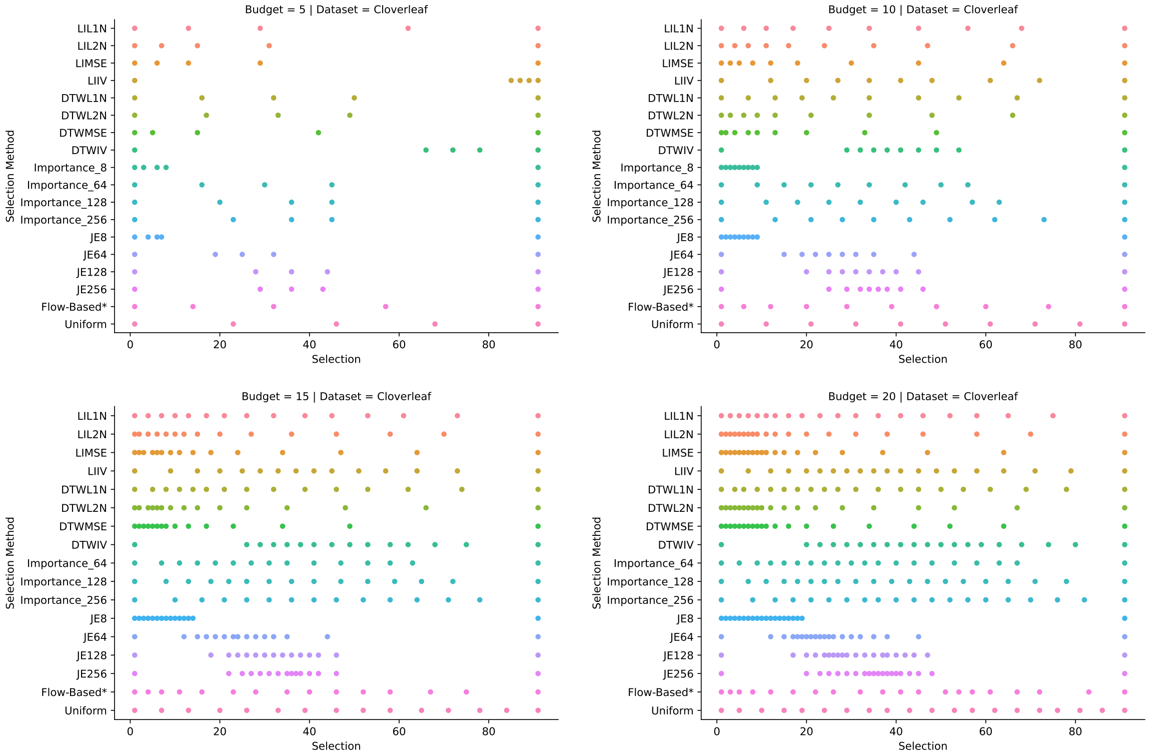
APPENDIX A

ADDITINAL RESULTS FROM EACH SELECTION METHOD

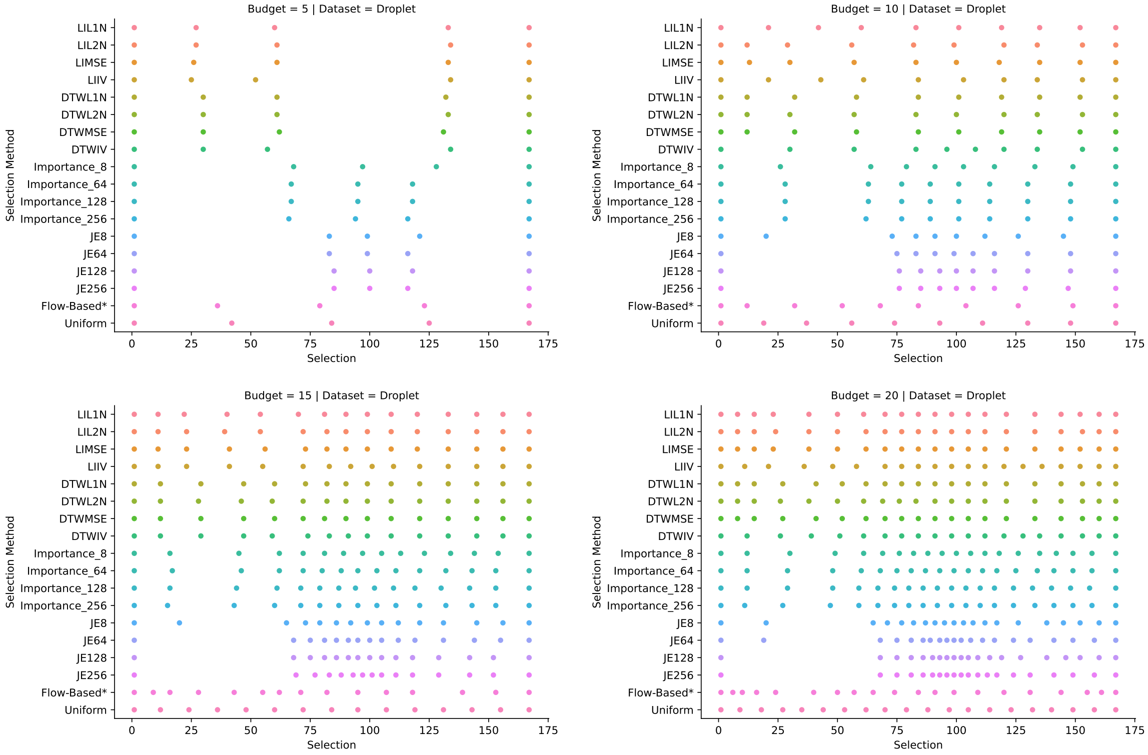



Budget = 5 | Dataset = Bottle
Budget = 10 | Dataset = Bottle
Budget = 15 | Dataset = Bottle
Budget = 20 | Dataset = Bottle
Selection Method
LIL1N
LIL2N
LIMSE
LIIV
DTWL1N
DTWL2N
DTWMSE
DTWIV
Importance_8
Importance_64
Importance_128
Importance_256
JE8
JE64
JE128
JE256
Flow-Based*
Uniform
0
100
200
300
400
Selection

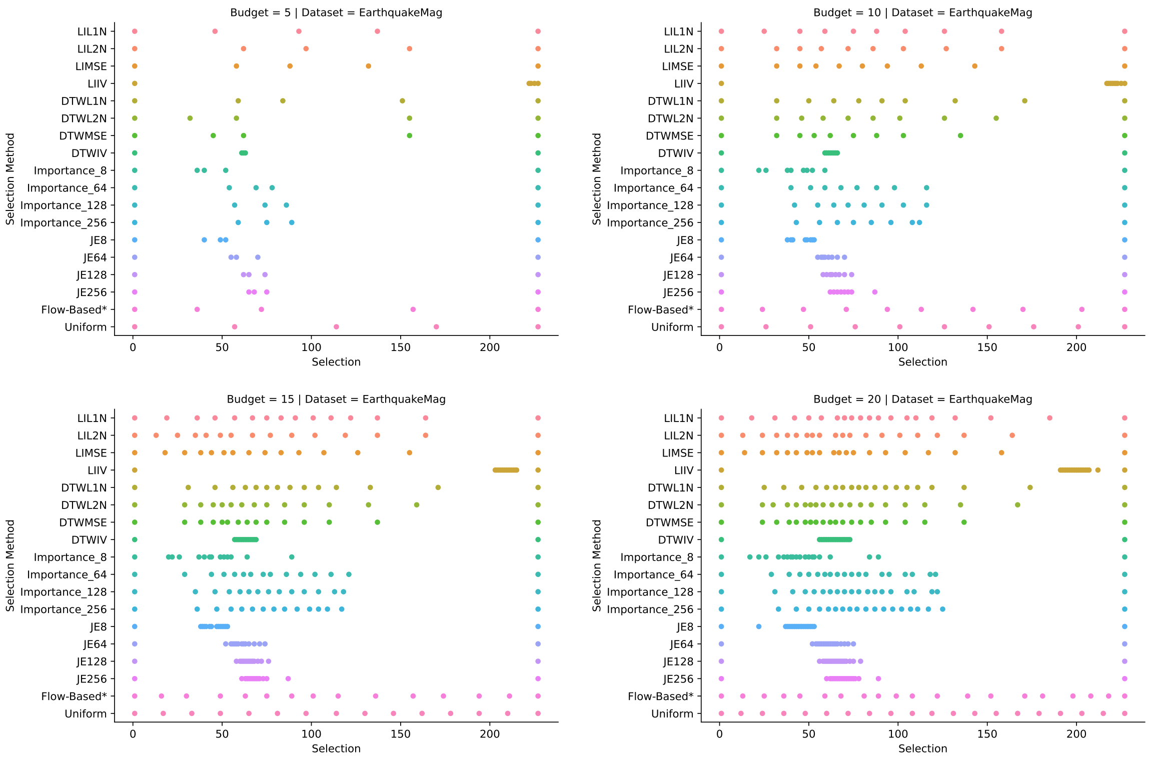

Budget = 5 | Dataset = Cloud
Budget = 10 | Dataset = Cloud
Budget = 15 | Dataset = Cloud
Budget = 20 | Dataset = Cloud
Selection Method
LIL1N
LIL2N
LIMSE
LIIV
DTWL1N
DTWL2N
DTWMSE
DTWIV
Importance_8
Importance_64
Importance_128
Importance_256
JE8
JE64
JE128
JE256
Flow-Based*
Uniform
0
100
200
300
400
500
Selection

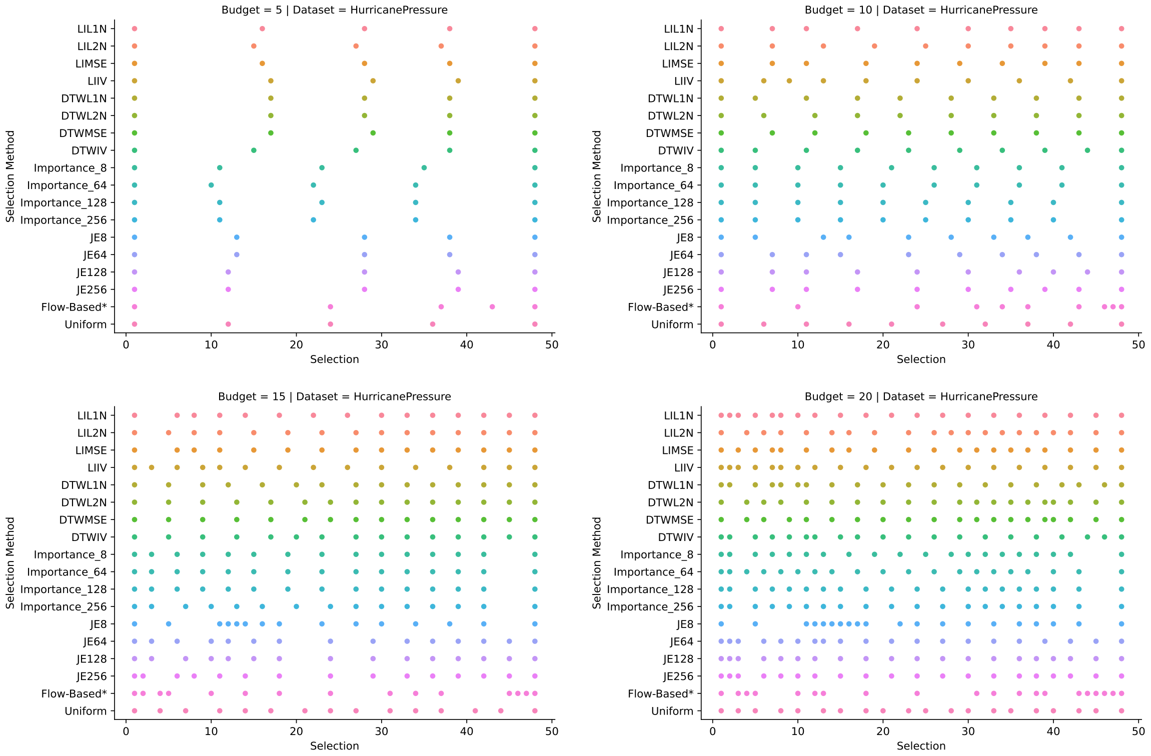


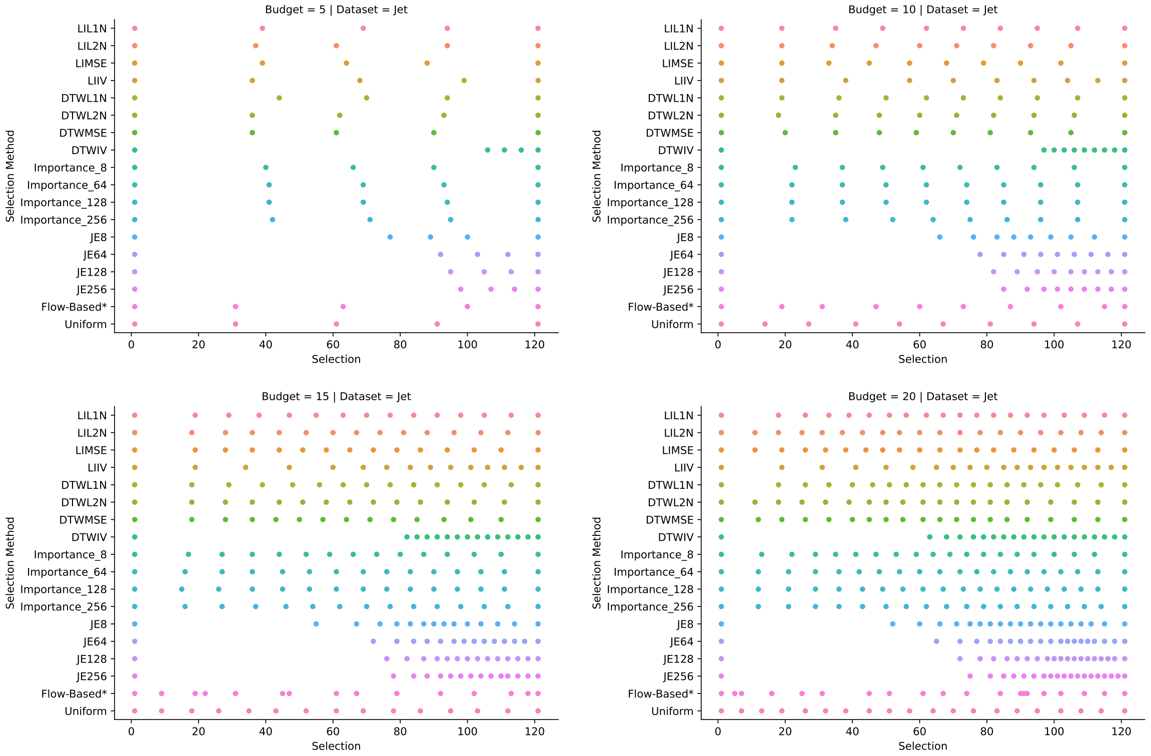
Budget = 5 | Dataset = Cloverleaf
Budget = 10 | Dataset = Cloverleaf
Budget = 15 | Dataset = Cloverleaf
Budget = 20 | Dataset = Cloverleaf
Selection Method
LIL1N
LIL2N
LIMSE
LIIV
DTWL1N
DTWL2N
DTWMSE
DTWIV
Importance_8
Importance_64
Importance_128
Importance_256
JE8
JE64
JE128
JE256
Flow-Based*
Uniform
Selection
0
20
40
60
80

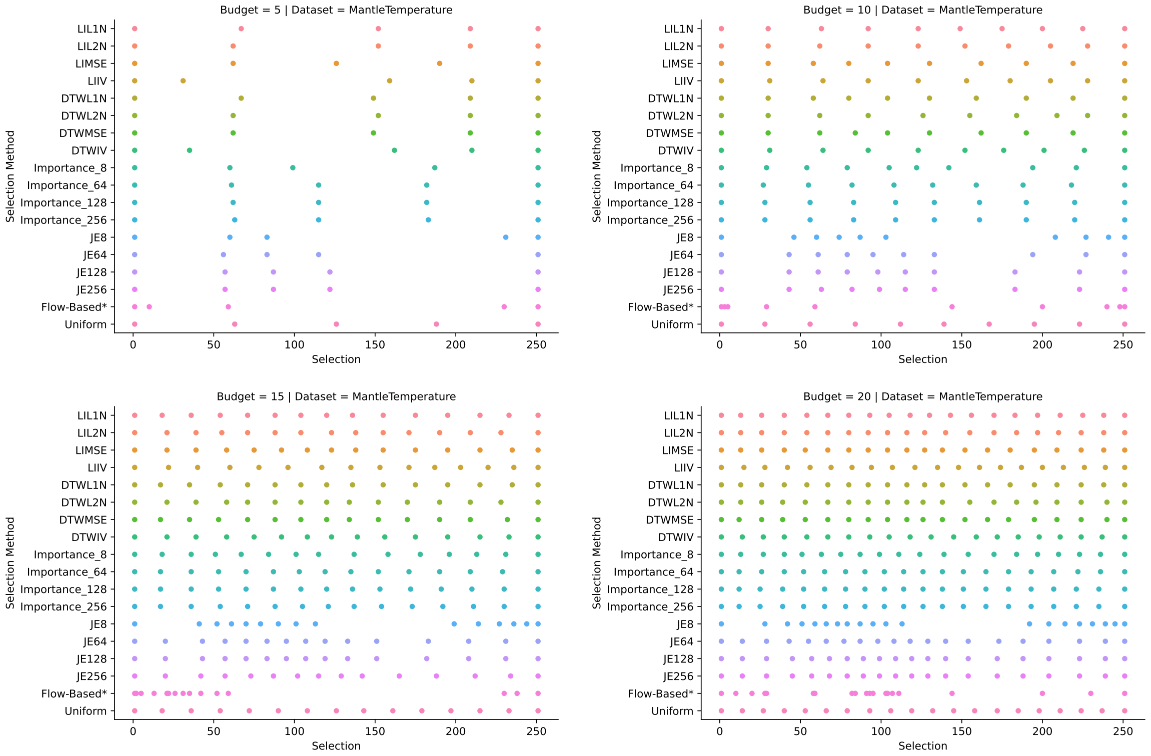


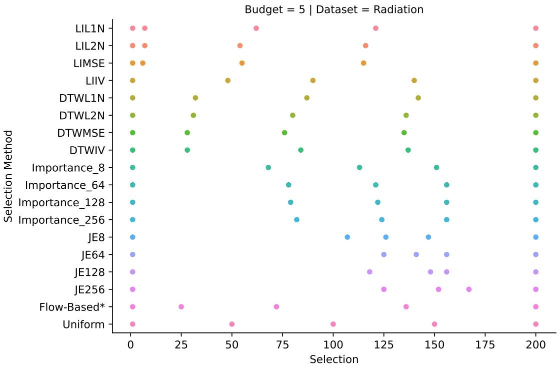
Budget = 5 | Dataset = Droplet
Budget = 10 | Dataset = Droplet
Budget = 15 | Dataset = Droplet
Budget = 20 | Dataset = Droplet
Selection Method
LIL1N
LIL2N
LIMSE
LIIV
DTWL1N
DTWL2N
DTWMSE
DTWIV
Importance_8
Importance_64
Importance_128
Importance_256
JE8
JE64
JE128
JE256
Flow-Based*
Uniform
Selection
0
25
50
75
100
125
150
175

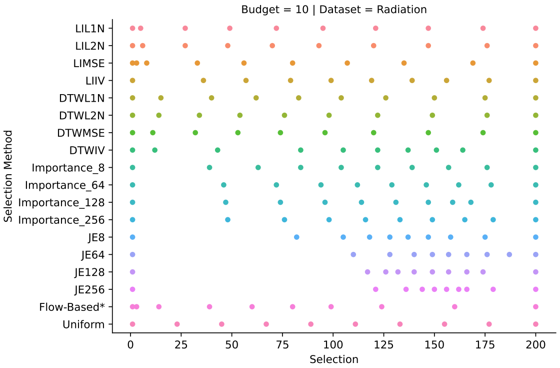

Budget = 5 | Dataset = EarthquakeMag
Budget = 10 | Dataset = EarthquakeMag
Budget = 15 | Dataset = EarthquakeMag
Budget = 20 | Dataset = EarthquakeMag
Selection Method
LIL1N
LIL2N
LIMSE
LIIV
DTWL1N
DTWL2N
DTWMSE
DTWIV
Importance_8
Importance_64
Importance_128
Importance_256
JE8
JE64
JE128
JE256
Flow-Based*
Uniform
Selection
0
50
100
150
200

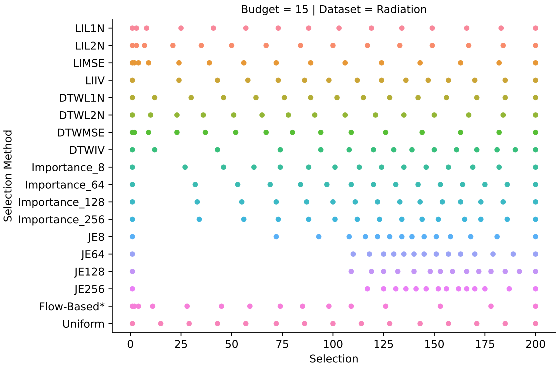

Budget = 5 | Dataset = HurricanePressure
Budget = 10 | Dataset = HurricanePressure
Budget = 15 | Dataset = HurricanePressure
Budget = 20 | Dataset = HurricanePressure
Selection Method
LIL1N
LIL2N
LIMSE
LIIV
DTWL1N
DTWL2N
DTWMSE
DTWIV
Importance_8
Importance_64
Importance_128
Importance_256
JE8
JE64
JE128
JE256
Flow-Based*
Uniform
Selection
0
10
20
30
40
50

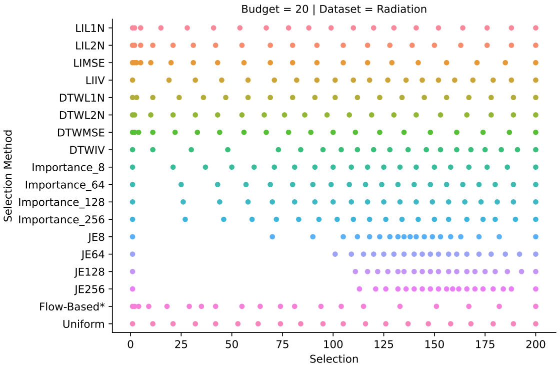

Budget = 5 | Dataset = Jet
Budget = 10 | Dataset = Jet
Budget = 15 | Dataset = Jet
Budget = 20 | Dataset = Jet
Selection Method
LIL1N
LIL2N
LIMSE
LIIV
DTWL1N
DTWL2N
DTWMSE
DTWIV
Importance_8
Importance_64
Importance_128
Importance_256
JE8
JE64
JE128
JE256
Flow-Based*
Uniform
Selection
0
20
40
60
80
100
120


Budget = 5 | Dataset = MantleTemperature
Budget = 10 | Dataset = MantleTemperature
Budget = 15 | Dataset = MantleTemperature
Budget = 20 | Dataset = MantleTemperature
Selection Method
LIL1N
LIL2N
LIMSE
LIIV
DTWL1N
DTWL2N
DTWMSE
DTWIV
Importance_8
Importance_64
Importance_128
Importance_256
JE8
JE64
JE128
JE256
Flow-Based*
Uniform
Selection
0
50
100
150
200
250


Budget = 5 | Dataset = Radiation
Selection Method
LIL1N
LIL2N
LIMSE
LIIV
DTWL1N
DTWL2N
DTWMSE
DTWIV
Importance_8
Importance_64
Importance_128
Importance_256
JE8
JE64
JE128
JE256
Flow-Based*
Uniform
0
25
50
75
100
125
150
175
200
Selection

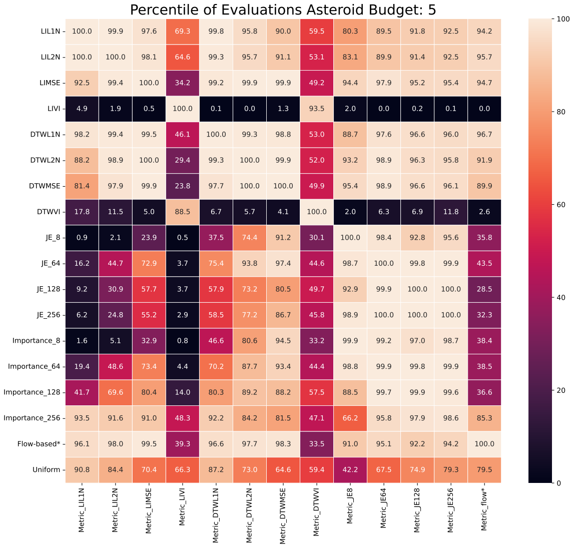

Budget = 10 | Dataset = Radiation
Selection Method
LIL1N
LIL2N
LIMSE
LIIV
DTWL1N
DTWL2N
DTWMSE
DTWIV
Importance_8
Importance_64
Importance_128
Importance_256
JE8
JE64
JE128
JE256
Flow-Based*
Uniform
0
25
50
75
100
125
150
175
200
Selection


Budget = 15 | Dataset = Radiation
Selection Method
LIL1N
LIL2N
LIMSE
LIIV
DTWL1N
DTWL2N
DTWMSE
DTWIV
Importance_8
Importance_64
Importance_128
Importance_256
JE8
JE64
JE128
JE256
Flow-Based*
Uniform
0
25
50
75
100
125
150
175
200
Selection


Budget = 20 | Dataset = Radiation
Selection Method
LIL1N
LIL2N
LIMSE
LIIV
DTWL1N
DTWL2N
DTWMSE
DTWIV
Importance_8
Importance_64
Importance_128
Importance_256
JE8
JE64
JE128
JE256
Flow-Based*
Uniform
0
25
50
75
100
125
150
175
200
Selection

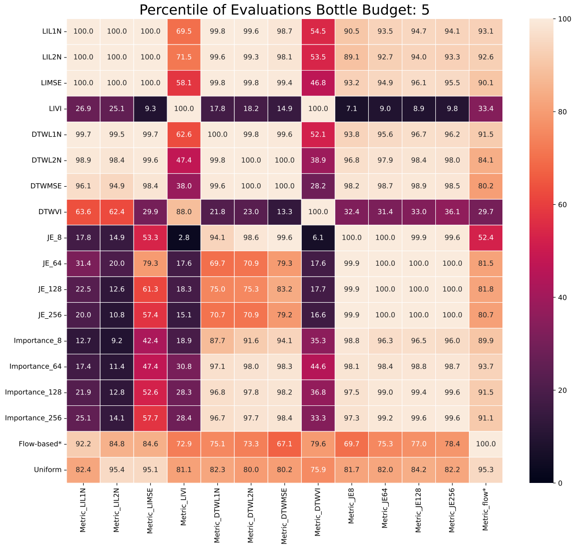
APPENDIX B

EVALUATIONS COMPARING TO RANDOM SELECTIONS



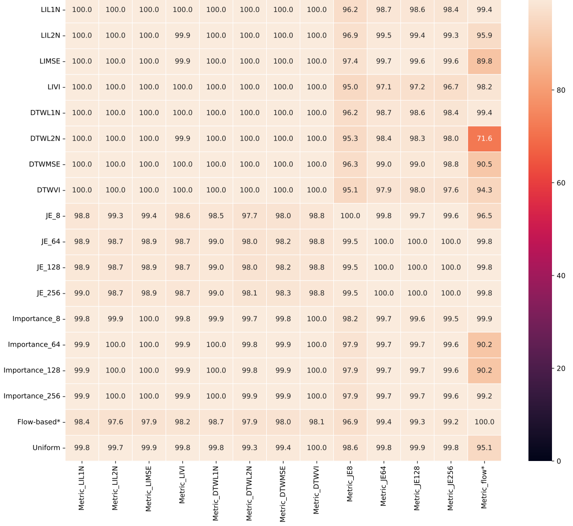
Percentile of Evaluations Asteroid Budget: 5

	Metric_LIL1N	Metric_LIL2N	Metric_LIMSE	Metric_LIVI	Metric_DTWL1N	Metric_DTWL2N	Metric_DTWMSE	Metric_DTWVI	Metric_JE8	Metric_JE64	Metric_JE128	Metric_JE256	Metric_flow*
LIL1N	100.0	99.9	97.6	69.3	99.8	95.8	90.0	59.5	80.3	89.5	91.8	92.5	94.2
LIL2N	100.0	100.0	98.1	64.6	99.3	95.7	91.1	53.1	83.1	89.9	91.4	92.5	95.7
LIMSE	92.5	99.4	100.0	34.2	99.2	99.9	99.9	49.2	94.4	97.9	95.2	95.4	94.7
LIVI	4.9	1.9	0.5	100.0	0.1	0.0	1.3	93.5	2.0	0.0	0.2	0.1	0.0
DTWL1N	98.2	99.4	99.5	46.1	100.0	99.3	98.8	53.0	88.7	97.6	96.6	96.0	96.7
DTWL2N	88.2	98.9	100.0	29.4	99.3	100.0	99.9	52.0	93.2	98.9	96.3	95.8	91.9
DTWMSE	81.4	97.9	99.9	23.8	97.7	100.0	100.0	49.9	95.4	98.9	96.6	96.1	89.9
DTWVI	17.8	11.5	5.0	88.5	6.7	5.7	4.1	100.0	2.0	6.3	6.9	11.8	2.6
JE_8	0.9	2.1	23.9	0.5	37.5	74.4	91.2	30.1	100.0	98.4	92.8	95.6	35.8
JE_64	16.2	44.7	72.9	3.7	75.4	93.8	97.4	44.6	98.7	100.0	99.8	99.9	43.5
JE_128	9.2	30.9	57.7	3.7	57.9	73.2	80.5	49.7	92.9	99.9	100.0	100.0	28.5
JE_256	6.2	24.8	55.2	2.9	58.5	77.2	86.7	45.8	98.9	100.0	100.0	100.0	32.3
Importance_8	1.6	5.1	32.9	0.8	46.6	80.6	94.5	33.2	99.9	99.2	97.0	98.7	38.4
Importance_64	19.4	48.6	73.4	4.4	70.2	87.7	93.4	44.4	98.8	99.9	99.8	99.9	38.5
Importance_128	41.7	69.6	80.4	14.0	80.3	89.2	88.2	57.5	88.5	99.7	99.9	99.6	36.6
Importance_256	93.5	91.6	91.0	48.3	92.2	84.2	81.5	47.1	66.2	95.8	97.9	98.6	85.3
Flow-based*	96.1	98.0	99.5	39.3	96.6	97.7	98.3	33.5	91.0	95.1	92.2	94.2	100.0
Uniform	90.8	84.4	70.4	66.3	87.2	73.0	64.6	59.4	42.2	67.5	74.9	79.3	79.5



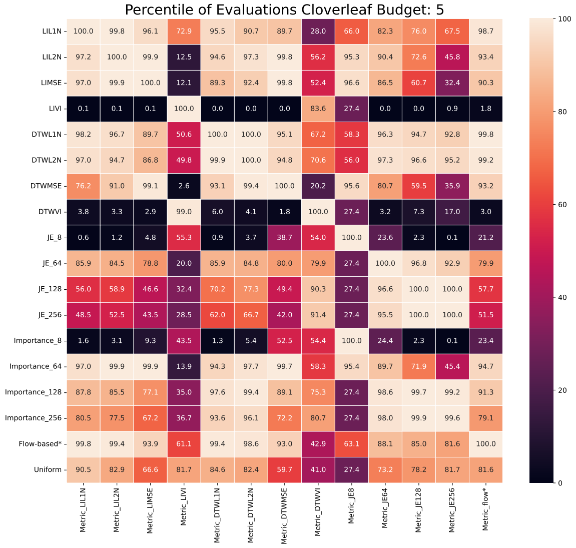
Percentile of Evaluations Bottle Budget: 5

	Metric_LIL1N	Metric_LIL2N	Metric_LIMSE	Metric_LIVI	Metric_DTWL1N	Metric_DTWL2N	Metric_DTWMSE	Metric_DTWVI	Metric_JE8	Metric_JE64	Metric_JE128	Metric_JE256	Metric_flow*
LIL1N	100.0	100.0	100.0	69.5	99.8	99.6	98.7	54.5	90.5	93.5	94.7	94.1	93.1
LIL2N	100.0	100.0	100.0	71.5	99.6	99.3	98.1	53.5	89.1	92.7	94.0	93.3	92.6
LIMSE	100.0	100.0	100.0	58.1	99.8	99.8	99.4	46.8	93.2	94.9	96.1	95.5	90.1
LIVI	26.9	25.1	9.3	100.0	17.8	18.2	14.9	100.0	7.1	9.0	8.9	9.8	33.4
DTWL1N	99.7	99.5	99.7	62.6	100.0	99.8	99.6	52.1	93.8	95.6	96.7	96.2	91.5
DTWL2N	98.9	98.4	99.6	47.4	99.8	100.0	100.0	38.9	96.8	97.9	98.4	98.0	84.1
DTWMSE	96.1	94.9	98.4	38.0	99.6	100.0	100.0	28.2	98.2	98.7	98.9	98.5	80.2
DTWVI	63.6	62.4	29.9	88.0	21.8	23.0	13.3	100.0	32.4	31.4	33.0	36.1	29.7
JE_8	17.8	14.9	53.3	2.8	94.1	98.6	99.6	6.1	100.0	100.0	99.9	99.6	52.4
JE_64	31.4	20.0	79.3	17.6	69.7	70.9	79.3	17.6	99.9	100.0	100.0	100.0	81.5
JE_128	22.5	12.6	61.3	18.3	75.0	75.3	83.2	17.7	99.9	100.0	100.0	100.0	81.8
JE_256	20.0	10.8	57.4	15.1	70.7	70.9	79.2	16.6	99.9	100.0	100.0	100.0	80.7
Importance_8	12.7	9.2	42.4	18.9	87.7	91.6	94.1	35.3	98.8	96.3	96.5	96.0	89.9
Importance_64	17.4	11.4	47.4	30.8	97.1	98.0	98.3	44.6	98.1	98.4	98.8	98.7	93.7
Importance_128	21.9	12.8	52.6	28.3	96.8	97.8	98.2	36.8	97.5	99.0	99.4	99.6	91.5
Importance_256	25.1	14.1	57.7	28.4	96.7	97.7	98.4	33.3	97.3	99.2	99.6	99.6	91.1
Flow-based*	92.2	84.8	84.6	72.9	75.1	73.3	67.1	79.6	69.7	75.3	77.0	78.4	100.0
Uniform	82.4	95.4	95.1	81.1	82.3	80.0	80.2	75.9	81.7	82.0	84.2	82.2	95.3



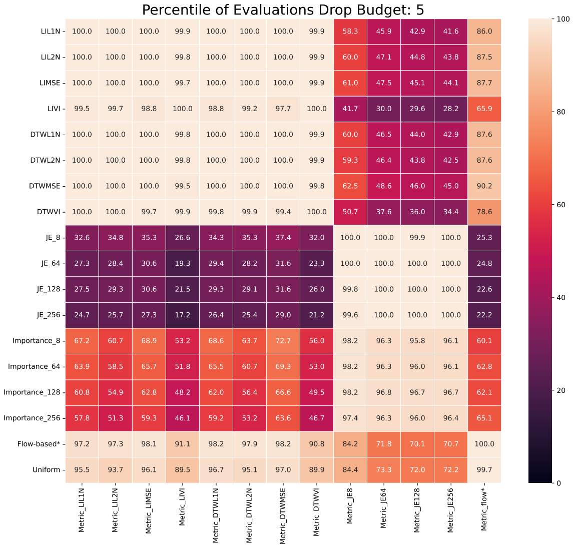
Percentile of Evaluations Cloud Budget: 5

	Metric_LIL1N	Metric_LIL2N	Metric_LIMSE	Metric_LIVI	Metric_DTWL1N	Metric_DTWL2N	Metric_DTWMSE	Metric_DTWVI	Metric_JE8	Metric_JE64	Metric_JE128	Metric_JE256	Metric_flow*
LIL1N	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.2	98.7	98.6	98.4	99.4
LIL2N	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	96.9	99.5	99.4	99.3	95.9
LIMSE	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	97.4	99.7	99.6	99.6	89.8
LIVI	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.0	97.1	97.2	96.7	98.2
DTWL1N	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.2	98.7	98.6	98.4	99.4
DTWL2N	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	95.3	98.4	98.3	98.0	71.6
DTWMSE	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.3	99.0	99.0	98.8	90.5
DTWVI	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.1	97.9	98.0	97.6	94.3
JE_8	98.8	99.3	99.4	98.6	98.5	97.7	98.0	98.8	100.0	99.8	99.7	99.6	96.5
JE_64	98.9	98.7	98.9	98.7	99.0	98.0	98.2	98.8	99.5	100.0	100.0	100.0	99.8
JE_128	98.9	98.7	98.9	98.7	99.0	98.0	98.2	98.8	99.5	100.0	100.0	100.0	99.8
JE_256	99.0	98.7	98.9	98.7	99.0	98.1	98.3	98.8	99.5	100.0	100.0	100.0	99.8
Importance_8	99.8	99.9	100.0	99.8	99.9	99.7	99.8	100.0	98.2	99.7	99.6	99.5	99.9
Importance_64	99.9	100.0	100.0	99.9	100.0	99.8	99.9	100.0	97.9	99.7	99.7	99.6	90.2
Importance_128	99.9	100.0	100.0	99.9	100.0	99.8	99.9	100.0	97.9	99.7	99.7	99.6	90.2
Importance_256	99.9	100.0	100.0	99.9	100.0	99.9	99.9	100.0	97.9	99.7	99.7	99.6	99.2
Flow-based*	98.4	97.6	97.9	98.2	98.7	97.9	98.0	98.1	96.9	99.4	99.3	99.2	100.0
Uniform	99.8	99.7	99.9	99.8	99.8	99.3	99.4	100.0	98.6	99.8	99.9	99.8	95.1



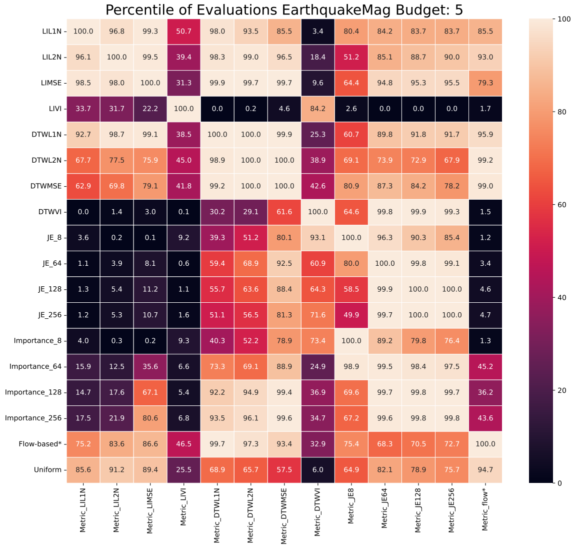
Percentile of Evaluations Cloverleaf Budget: 5

	Metric_LIL1N	Metric_LIL2N	Metric_LIMSE	Metric_LIVI	Metric_DTWL1N	Metric_DTWL2N	Metric_DTWMSE	Metric_DTWVI	Metric_JE8	Metric_JE64	Metric_JE128	Metric_JE256	Metric_flow*
LIL1N	100.0	99.8	96.1	72.9	95.5	90.7	89.7	28.0	66.0	82.3	76.0	67.5	98.7
LIL2N	97.2	100.0	99.9	12.5	94.6	97.3	99.8	56.2	95.3	90.4	72.6	45.8	93.4
LIMSE	97.0	99.9	100.0	12.1	89.3	92.4	99.8	52.4	96.6	86.5	60.7	32.4	90.3
LIVI	0.1	0.1	0.1	100.0	0.0	0.0	0.0	83.6	27.4	0.0	0.0	0.9	1.8
DTWL1N	98.2	96.7	89.7	50.6	100.0	100.0	95.1	67.2	58.3	96.3	94.7	92.8	99.8
DTWL2N	97.0	94.7	86.8	49.8	99.9	100.0	94.8	70.6	56.0	97.3	96.6	95.2	99.2
DTWMSE	76.2	91.0	99.1	2.6	93.1	99.4	100.0	20.2	95.6	80.7	59.5	35.9	93.2
DTWVI	3.8	3.3	2.9	99.0	6.0	4.1	1.8	100.0	27.4	3.2	7.3	17.0	3.0
JE_8	0.6	1.2	4.8	55.3	0.9	3.7	38.7	54.0	100.0	23.6	2.3	0.1	21.2
JE_64	85.9	84.5	78.8	20.0	85.9	84.8	80.0	79.9	27.4	100.0	96.8	92.9	79.9
JE_128	56.0	58.9	46.6	32.4	70.2	77.3	49.4	90.3	27.4	96.6	100.0	100.0	57.7
JE_256	48.5	52.5	43.5	28.5	62.0	66.7	42.0	91.4	27.4	95.5	100.0	100.0	51.5
Importance_8	1.6	3.1	9.3	43.5	1.3	5.4	52.5	54.4	100.0	24.4	2.3	0.1	23.4
Importance_64	97.0	99.9	99.9	13.9	94.3	97.7	99.7	58.3	95.4	89.7	71.9	45.4	94.7
Importance_128	87.8	85.5	77.1	35.0	97.6	99.4	89.1	75.3	27.4	98.6	99.7	99.2	91.3
Importance_256	80.5	77.5	67.2	36.7	93.6	96.1	72.2	80.7	27.4	98.0	99.9	99.6	79.1
Flow-based*	99.8	99.4	93.9	61.1	99.4	98.6	93.0	42.9	63.1	88.1	85.0	81.6	100.0
Uniform	90.5	82.9	66.6	81.7	84.6	82.4	59.7	41.0	27.4	73.2	78.2	81.7	81.6



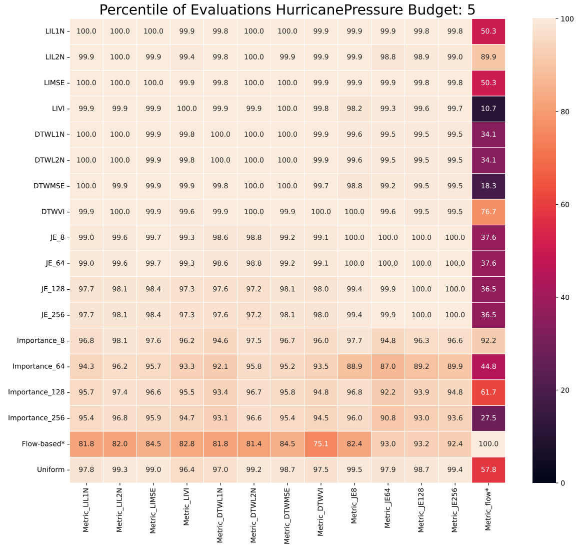
Percentile of Evaluations Drop Budget: 5

	Metric_LIL1N	Metric_LIL2N	Metric_LIMSE	Metric_LIVI	Metric_DTWL1N	Metric_DTWL2N	Metric_DTWMSE	Metric_DTWVI	Metric_JE8	Metric_JE64	Metric_JE128	Metric_JE256	Metric_flow*
LIL1N	100.0	100.0	100.0	99.9	100.0	100.0	100.0	99.9	58.3	45.9	42.9	41.6	86.0
LIL2N	100.0	100.0	100.0	99.8	100.0	100.0	100.0	99.9	60.0	47.1	44.8	43.8	87.5
LIMSE	100.0	100.0	100.0	99.7	100.0	100.0	100.0	99.9	61.0	47.5	45.1	44.1	87.7
LIVI	99.5	99.7	98.8	100.0	98.8	99.2	97.7	100.0	41.7	30.0	29.6	28.2	65.9
DTWL1N	100.0	100.0	100.0	99.8	100.0	100.0	100.0	99.9	60.0	46.5	44.0	42.9	87.6
DTWL2N	100.0	100.0	100.0	99.8	100.0	100.0	100.0	99.9	59.3	46.4	43.8	42.5	87.6
DTWMSE	100.0	100.0	100.0	99.5	100.0	100.0	100.0	99.8	62.5	48.6	46.0	45.0	90.2
DTWVI	100.0	100.0	99.7	99.9	99.8	99.9	99.4	100.0	50.7	37.6	36.0	34.4	78.6
JE_8	32.6	34.8	35.3	26.6	34.3	35.3	37.4	32.0	100.0	100.0	99.9	100.0	25.3
JE_64	27.3	28.4	30.6	19.3	29.4	28.2	31.6	23.3	100.0	100.0	100.0	100.0	24.8
JE_128	27.5	29.3	30.6	21.5	29.3	29.1	31.6	26.0	99.8	100.0	100.0	100.0	22.6
JE_256	24.7	25.7	27.3	17.2	26.4	25.4	29.0	21.2	99.6	100.0	100.0	100.0	22.2
Importance_8	67.2	60.7	68.9	53.2	68.6	63.7	72.7	56.0	98.2	96.3	95.8	96.1	60.1
Importance_64	63.9	58.5	65.7	51.8	65.5	60.7	69.3	53.0	98.2	96.3	96.0	96.1	62.8
Importance_128	60.8	54.9	62.8	48.2	62.0	56.4	66.6	49.5	98.2	96.8	96.7	96.7	62.1
Importance_256	57.8	51.3	59.3	46.1	59.2	53.2	63.6	46.7	97.4	96.3	96.0	96.4	65.1
Flow-based*	97.2	97.3	98.1	91.1	98.2	97.9	98.2	90.8	84.2	71.8	70.1	70.7	100.0
Uniform	95.5	93.7	96.1	89.5	96.7	95.1	97.0	89.9	84.4	73.3	72.0	72.2	99.7



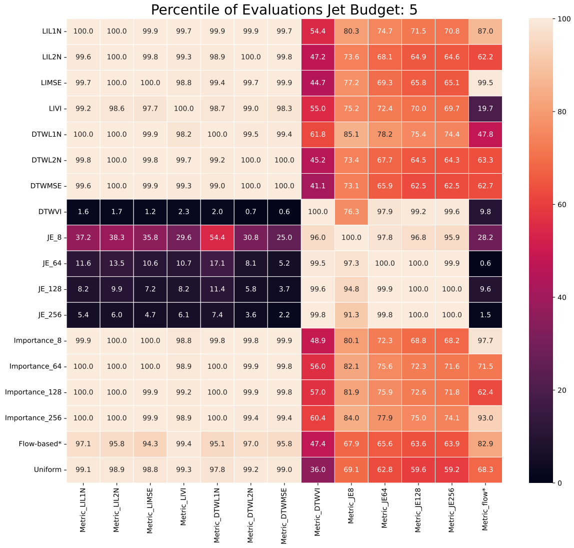
Percentile of Evaluations EarthquakeMag Budget: 5

	Metric_LIL1N	Metric_LIL2N	Metric_LIMSE	Metric_LIVI	Metric_DTWL1N	Metric_DTWL2N	Metric_DTWMSE	Metric_DTWVI	Metric_JE8	Metric_JE64	Metric_JE128	Metric_JE256	Metric_flow*
LIL1N	100.0	96.8	99.3	50.7	98.0	93.5	85.5	3.4	80.4	84.2	83.7	83.7	85.5
LIL2N	96.1	100.0	99.5	39.4	98.3	99.0	96.5	18.4	51.2	85.1	88.7	90.0	93.0
LIMSE	98.5	98.0	100.0	31.3	99.9	99.7	99.7	9.6	64.4	94.8	95.3	95.5	79.3
LIVI	33.7	31.7	22.2	100.0	0.0	0.2	4.6	84.2	2.6	0.0	0.0	0.0	1.7
DTWL1N	92.7	98.7	99.1	38.5	100.0	100.0	99.9	25.3	60.7	89.8	91.8	91.7	95.9
DTWL2N	67.7	77.5	75.9	45.0	98.9	100.0	100.0	38.9	69.1	73.9	72.9	67.9	99.2
DTWMSE	62.9	69.8	79.1	41.8	99.2	100.0	100.0	42.6	80.9	87.3	84.2	78.2	99.0
DTWVI	0.0	1.4	3.0	0.1	30.2	29.1	61.6	100.0	64.6	99.8	99.9	99.3	1.5
JE_8	3.6	0.2	0.1	9.2	39.3	51.2	80.1	93.1	100.0	96.3	90.3	85.4	1.2
JE_64	1.1	3.9	8.1	0.6	59.4	68.9	92.5	60.9	80.0	100.0	99.8	99.1	3.4
JE_128	1.3	5.4	11.2	1.1	55.7	63.6	88.4	64.3	58.5	99.9	100.0	100.0	4.6
JE_256	1.2	5.3	10.7	1.6	51.1	56.5	81.3	71.6	49.9	99.7	100.0	100.0	4.7
Importance_8	4.0	0.3	0.2	9.3	40.3	52.2	78.9	73.4	100.0	89.2	79.8	76.4	1.3
Importance_64	15.9	12.5	35.6	6.6	73.3	69.1	88.9	24.9	98.9	99.5	98.4	97.5	45.2
Importance_128	14.7	17.6	67.1	5.4	92.2	94.9	99.4	36.9	69.6	99.7	99.8	99.7	36.2
Importance_256	17.5	21.9	80.6	6.8	93.5	96.1	99.6	34.7	67.2	99.6	99.8	99.8	43.6
Flow-based*	75.2	83.6	86.6	46.5	99.7	97.3	93.4	32.9	75.4	68.3	70.5	72.7	100.0
Uniform	85.6	91.2	89.4	25.5	68.9	65.7	57.5	6.0	64.9	82.1	78.9	75.7	94.7



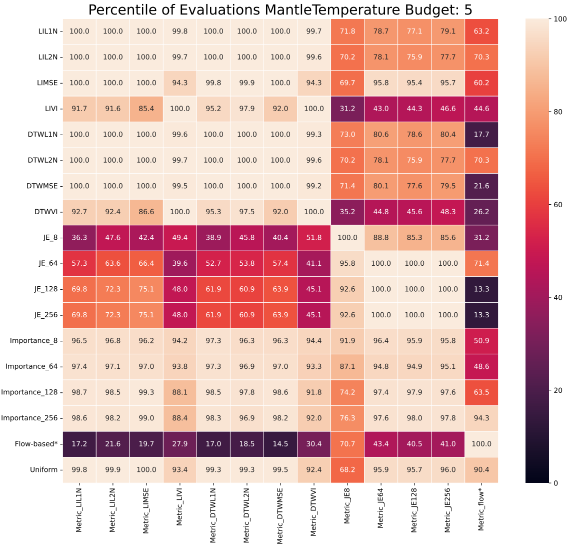
Percentile of Evaluations HurricanePressure Budget: 5

	Metric_LIL1N	Metric_LIL2N	Metric_LIMSE	Metric_LIVI	Metric_DTWL1N	Metric_DTWL2N	Metric_DTWMSE	Metric_DTWVI	Metric_JE8	Metric_JE64	Metric_JE128	Metric_JE256	Metric_flow*
LIL1N	100.0	100.0	100.0	99.9	99.8	100.0	100.0	99.9	99.9	99.9	99.8	99.8	50.3
LIL2N	99.9	100.0	99.9	99.4	99.8	100.0	99.9	99.9	99.9	98.8	98.9	99.0	89.9
LIMSE	100.0	100.0	100.0	99.9	99.8	100.0	100.0	99.9	99.9	99.9	99.8	99.8	50.3
LIVI	99.9	99.9	99.9	100.0	99.9	99.9	100.0	99.8	98.2	99.3	99.6	99.7	10.7
DTWL1N	100.0	100.0	99.9	99.8	100.0	100.0	100.0	99.9	99.6	99.5	99.5	99.5	34.1
DTWL2N	100.0	100.0	99.9	99.8	100.0	100.0	100.0	99.9	99.6	99.5	99.5	99.5	34.1
DTWMSE	100.0	99.9	99.9	99.9	99.8	100.0	100.0	99.7	98.8	99.2	99.5	99.5	18.3
DTWVI	99.9	100.0	99.9	99.6	99.9	100.0	99.9	100.0	100.0	99.6	99.5	99.5	76.7
JE_8	99.0	99.6	99.7	99.3	98.6	98.8	99.2	99.1	100.0	100.0	100.0	100.0	37.6
JE_64	99.0	99.6	99.7	99.3	98.6	98.8	99.2	99.1	100.0	100.0	100.0	100.0	37.6
JE_128	97.7	98.1	98.4	97.3	97.6	97.2	98.1	98.0	99.4	99.9	100.0	100.0	36.5
JE_256	97.7	98.1	98.4	97.3	97.6	97.2	98.1	98.0	99.4	99.9	100.0	100.0	36.5
Importance_8	96.8	98.1	97.6	96.2	94.6	97.5	96.7	96.0	97.7	94.8	96.3	96.6	92.2
Importance_64	94.3	96.2	95.7	93.3	92.1	95.8	95.2	93.5	88.9	87.0	89.2	89.9	44.8
Importance_128	95.7	97.4	96.6	95.5	93.4	96.7	95.8	94.8	96.8	92.2	93.9	94.8	61.7
Importance_256	95.4	96.8	95.9	94.7	93.1	96.6	95.4	94.5	96.0	90.8	93.0	93.6	27.5
Flow-based*	81.8	82.0	84.5	82.8	81.8	81.4	84.5	75.1	82.4	93.0	93.2	92.4	100.0
Uniform	97.8	99.3	99.0	96.4	97.0	99.2	98.7	97.5	99.5	97.9	98.7	99.4	57.8

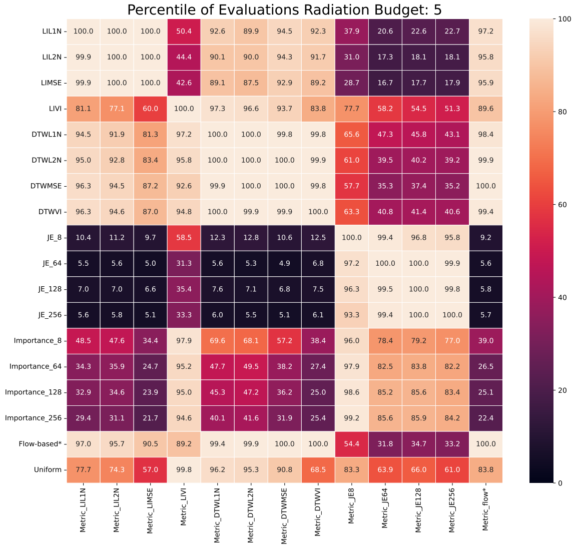


Percentile of Evaluations Jet Budget: 5

	Metric_LIL1N	Metric_LIL2N	Metric_LIMSE	Metric_LIVI	Metric_DTWL1N	Metric_DTWL2N	Metric_DTWMSE	Metric_DTWVI	Metric_JE8	Metric_JE64	Metric_JE128	Metric_JE256	Metric_flow*
LIL1N	100.0	100.0	99.9	99.7	99.9	99.9	99.7	54.4	80.3	74.7	71.5	70.8	87.0
LIL2N	99.6	100.0	99.8	99.3	98.9	100.0	99.8	47.2	73.6	68.1	64.9	64.6	62.2
LIMSE	99.7	100.0	100.0	98.8	99.4	99.7	99.9	44.7	77.2	69.3	65.8	65.1	99.5
LIVI	99.2	98.6	97.7	100.0	98.7	99.0	98.3	55.0	75.2	72.4	70.0	69.7	19.7
DTWL1N	100.0	100.0	99.9	98.2	100.0	99.5	99.4	61.8	85.1	78.2	75.4	74.4	47.8
DTWL2N	99.8	100.0	99.8	99.7	99.2	100.0	100.0	45.2	73.4	67.7	64.5	64.3	63.3
DTWMSE	99.6	100.0	99.9	99.3	99.0	100.0	100.0	41.1	73.1	65.9	62.5	62.5	62.7
DTWVI	1.6	1.7	1.2	2.3	2.0	0.7	0.6	100.0	76.3	97.9	99.2	99.6	9.8
JE_8	37.2	38.3	35.8	29.6	54.4	30.8	25.0	96.0	100.0	97.8	96.8	95.9	28.2
JE_64	11.6	13.5	10.6	10.7	17.1	8.1	5.2	99.5	97.3	100.0	100.0	99.9	0.6
JE_128	8.2	9.9	7.2	8.2	11.4	5.8	3.7	99.6	94.8	99.9	100.0	100.0	9.6
JE_256	5.4	6.0	4.7	6.1	7.4	3.6	2.2	99.8	91.3	99.8	100.0	100.0	1.5
Importance_8	99.9	100.0	100.0	98.8	99.8	99.8	99.9	48.9	80.1	72.3	68.8	68.2	97.7
Importance_64	100.0	100.0	100.0	98.9	100.0	99.9	99.8	56.0	82.1	75.6	72.3	71.6	71.5
Importance_128	100.0	100.0	99.9	99.2	100.0	99.9	99.8	57.0	81.9	75.9	72.6	71.8	62.4
Importance_256	100.0	100.0	99.9	98.9	100.0	99.4	99.4	60.4	84.0	77.9	75.0	74.1	93.0
Flow-based*	97.1	95.8	94.3	99.4	95.1	97.0	95.8	47.4	67.9	65.6	63.6	63.9	82.9
Uniform	99.1	98.9	98.8	99.3	97.8	99.2	99.0	36.0	69.1	62.8	59.6	59.2	68.3


Percentile of Evaluations MantleTemperature Budget: 5

	Metric_LIL1N	Metric_LIL2N	Metric_LIMSE	Metric_LIVI	Metric_DTWL1N	Metric_DTWL2N	Metric_DTWMSE	Metric_DTWVI	Metric_JE8	Metric_JE64	Metric_JE128	Metric_JE256	Metric_flow*
LIL1N	100.0	100.0	100.0	99.8	100.0	100.0	100.0	99.7	71.8	78.7	77.1	79.1	63.2
LIL2N	100.0	100.0	100.0	99.7	100.0	100.0	100.0	99.6	70.2	78.1	75.9	77.7	70.3
LIMSE	100.0	100.0	100.0	94.3	99.8	99.9	100.0	94.3	69.7	95.8	95.4	95.7	60.2
LIVI	91.7	91.6	85.4	100.0	95.2	97.9	92.0	100.0	31.2	43.0	44.3	46.6	44.6
DTWL1N	100.0	100.0	100.0	99.6	100.0	100.0	100.0	99.3	73.0	80.6	78.6	80.4	17.7
DTWL2N	100.0	100.0	100.0	99.7	100.0	100.0	100.0	99.6	70.2	78.1	75.9	77.7	70.3
DTWMSE	100.0	100.0	100.0	99.5	100.0	100.0	100.0	99.2	71.4	80.1	77.6	79.5	21.6
DTWVI	92.7	92.4	86.6	100.0	95.3	97.5	92.0	100.0	35.2	44.8	45.6	48.3	26.2
JE_8	36.3	47.6	42.4	49.4	38.9	45.8	40.4	51.8	100.0	88.8	85.3	85.6	31.2
JE_64	57.3	63.6	66.4	39.6	52.7	53.8	57.4	41.1	95.8	100.0	100.0	100.0	71.4
JE_128	69.8	72.3	75.1	48.0	61.9	60.9	63.9	45.1	92.6	100.0	100.0	100.0	13.3
JE_256	69.8	72.3	75.1	48.0	61.9	60.9	63.9	45.1	92.6	100.0	100.0	100.0	13.3
Importance_8	96.5	96.8	96.2	94.2	97.3	96.3	96.3	94.4	91.9	96.4	95.9	95.8	50.9
Importance_64	97.4	97.1	97.0	93.8	97.3	96.9	97.0	93.3	87.1	94.8	94.9	95.1	48.6
Importance_128	98.7	98.5	99.3	88.1	98.5	97.8	98.6	91.8	74.2	97.4	97.9	97.6	63.5
Importance_256	98.6	98.2	99.0	88.4	98.3	96.9	98.2	92.0	76.3	97.6	98.0	97.8	94.3
Flow-based*	17.2	21.6	19.7	27.9	17.0	18.5	14.5	30.4	70.7	43.4	40.5	41.0	100.0
Uniform	99.8	99.9	100.0	93.4	99.3	99.3	99.5	92.4	68.2	95.9	95.7	96.0	90.4


Percentile of Evaluations Radiation Budget: 5

	Metric_LIL1N	Metric_LIL2N	Metric_LIMSE	Metric_LIVI	Metric_DTWL1N	Metric_DTWL2N	Metric_DTWMSE	Metric_DTWVI	Metric_JE8	Metric_JE64	Metric_JE128	Metric_JE256	Metric_flow*
LIL1N	100.0	100.0	100.0	50.4	92.6	89.9	94.5	92.3	37.9	20.6	22.6	22.7	97.2
LIL2N	99.9	100.0	100.0	44.4	90.1	90.0	94.3	91.7	31.0	17.3	18.1	18.1	95.8
LIMSE	99.9	100.0	100.0	42.6	89.1	87.5	92.9	89.2	28.7	16.7	17.7	17.9	95.9
LIVI	81.1	77.1	60.0	100.0	97.3	96.6	93.7	83.8	77.7	58.2	54.5	51.3	89.6
DTWL1N	94.5	91.9	81.3	97.2	100.0	100.0	99.8	99.8	65.6	47.3	45.8	43.1	98.4
DTWL2N	95.0	92.8	83.4	95.8	100.0	100.0	100.0	99.9	61.0	39.5	40.2	39.2	99.9
DTWMSE	96.3	94.5	87.2	92.6	99.9	100.0	100.0	99.8	57.7	35.3	37.4	35.2	100.0
DTWVI	96.3	94.6	87.0	94.8	100.0	99.9	99.9	100.0	63.3	40.8	41.4	40.6	99.4
JE_8	10.4	11.2	9.7	58.5	12.3	12.8	10.6	12.5	100.0	99.4	96.8	95.8	9.2
JE_64	5.5	5.6	5.0	31.3	5.6	5.3	4.9	6.8	97.2	100.0	100.0	99.9	5.6
JE_128	7.0	7.0	6.6	35.4	7.6	7.1	6.8	7.5	96.3	99.5	100.0	99.8	5.8
JE_256	5.6	5.8	5.1	33.3	6.0	5.5	5.1	6.1	93.3	99.4	100.0	100.0	5.7
Importance_8	48.5	47.6	34.4	97.9	69.6	68.1	57.2	38.4	96.0	78.4	79.2	77.0	39.0
Importance_64	34.3	35.9	24.7	95.2	47.7	49.5	38.2	27.4	97.9	82.5	83.8	82.2	26.5
Importance_128	32.9	34.6	23.9	95.0	45.3	47.2	36.2	25.0	98.6	85.2	85.6	83.4	25.1
Importance_256	29.4	31.1	21.7	94.6	40.1	41.6	31.9	25.4	99.2	85.6	85.9	84.2	22.4
Flow-based*	97.0	95.7	90.5	89.2	99.4	99.9	100.0	100.0	54.4	31.8	34.7	33.2	100.0
Uniform	77.7	74.3	57.0	99.8	96.2	95.3	90.8	68.5	83.3	63.9	66.0	61.0	83.8

REFERENCES CITED

[1] 2006 IEEE Visualization Design Contest. Terashake 2.1 earthquake simulation data set. `http://sciviscontest.ieeevis.org/2006/index.html`, 2006.

[2] Banesh, D., Wendelberger, J., Petersen, M. R., Ahrens, J. P., and Hamann, B. Change point detection for ocean eddy analysis. In *EnvirVis@ EuroVis* (2018), pp. 27–33.

[3] Bennett, J. C., Bhagatwala, A., Chen, J. H., Pinar, A., Salloum, M., and Seshadhri, C. Trigger detection for adaptive scientific workflows using percentile sampling. *SIAM Journal on Scientific Computing 38*, 5 (2016), S240–S263.

[4] C. Miester / Institute of Aerospace Thermodynamics, University of Stuttgart. Droplet collision.

[5] Chen, M., and Jäenicke, H. An information-theoretic framework for visualization. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1206–1215.

[6] Childs, H., Ahern, S. D., Ahrens, J., Bauer, A. C., Bennett, J., Bethel, E. W., Bremer, P.-T., Brugger, E., Cottam, J., Dorier, M., et al. A terminology for in situ visualization and analysis systems. *The International Journal of High Performance Computing Applications 34*, 6 (2020), 676–691.

[7] Childs, H., Bennett, J., Garth, C., and Hentschel, B. In situ visualization for computational science. *IEEE Computer Graphics and Applications 39*, 6 (2019), 76–85.

[8] DKRZ/MPI-M. Visualization of clouds and atmospheric processes. https://scivis2017.dkrz.de, 2017.

[9] Fang, Z., Moeller, T., Hamarneh, G., and Celler, A. Visualization and exploration of time-varying medical image data sets. In *Proceedings of Graphics Interface 2007* (2007), pp. 281–288.

[10] Frey, S., and Ertl, T. Flow-based temporal selection for interactive volume visualization. In *Computer Graphics Forum* (2017), vol. 36, Wiley Online Library, pp. 153–165.

[11] Kawakami, Y., Marsaglia, N., Larsen, M., and Childs, H. Benchmarking in situ triggers via reconstruction error. In *ISAV'20 In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization.* 2020, pp. 38–43.

[12] Keogh, E. J., and Pazzani, M. J. Scaling up dynamic time warping to massive datasets. In *European Conference on Principles of Data Mining and Knowledge Discovery* (1999), Springer, pp. 1–11.

[13] Larsen, M., Harrison, C., Turton, T. L., Sane, S., Brink, S., and Childs, H. Trigger happy: Assessing the viability of trigger-based in situ analysis. In *2021 IEEE 11th Symposium on Large Data Analysis and Visualization (LDAV)* (2021), IEEE, pp. 22–31.

[14] Ling, J., Kegelmeyer, W. P., Aditya, K., Kolla, H., Reed, K. A., Shead, T. M., and Davis, W. L. Using feature importance metrics to detect events of interest in scientific computing applications. In *2017 IEEE 7th Symposium on Large Data Analysis and Visualization (LDAV)* (2017), IEEE, pp. 55–63.

[15] Liu, Y., Lu, Y., Wang, Y., Sun, D., Deng, L., Wan, Y., and Wang, F. Key time steps selection for cfd data based on deep metric learning. *Computers & Fluids 195* (2019), 104318.

[16] Ma, K.-L. In situ visualization at extreme scale: Challenges and opportunities. *IEEE Computer Graphics and Applications 29*, 6 (2009), 14–19.

[17] Malakar, P., Vishwanath, V., Munson, T., Knight, C., Hereld, M., Leyffer, S., and Papka, M. E. Optimal scheduling of in-situ analysis for large-scale scientific simulations. In *SC'15: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (2015), IEEE, pp. 1–11.

[18] Manning, C. D., Raghavan, P., and Schütze, H. *Hierarchical clustering.* Cambridge University Press, 2018.

[19] Myers, K., Lawrence, E., Fugate, M., Bowen, C. M., Ticknor, L., Woodring, J., Wendelberger, J., and Ahrens, J. Partitioning a large simulation as it runs, 2014.

[20] National Center for Atmospheric Research. Hurricane isabel data. `http://vis.computer.org/vis2004contest`, 2004.

[21] Patchett, J. M., and Gisler, G. R. Deep water impact ensemble data set. In *Technical Report LA-UR-17-21595* (2017).

[22] Pfister, H., Lorensen, B., Bajaj, C., Kindlmann, G., Schroeder, W., Avila, L. S., Raghu, K., Machiraju, R., and Lee, J. The transfer function bake-off. *IEEE Computer Graphics and Applications 21*, 3 (2001), 16–22.

[23] Porter, W. P., Xing, Y., von Ohlen, B. R., Han, J., and Wang, C. A deep learning approach to selecting representative time steps for time-varying multivariate data. In *2019 IEEE Visualization Conference (VIS)* (2019), IEEE, pp. 1–5.

[24] Pulido, J., Patchett, J., Bhattarai, M., Alexandrov, B., and Ahrens, J. Selection of optimal salient time steps by non-negative tucker tensor decomposition.

[25] Pysklywec Lab/University of Toronto. Earth's mantle convection. `https://scivis2021.netlify.app`, 2021.

[26] Salloum, M., Bennett, J. C., Pinar, A., Bhagatwala, A., and Chen, J. H. Enabling adaptive scientific workflows via trigger detection. In *Proceedings of the first workshop on in situ infrastructures for enabling extreme-scale analysis and visualization* (2015), pp. 41–45.

[27] Tong, X., Lee, T.-Y., and Shen, H.-W. Salient time steps selection from large scale time-varying data sets with dynamic time warping. In *IEEE Symposium on Large Data Analysis and Visualization (LDAV)* (2012), pp. 49–56.

[28] Uk-Mac. UK-MAC/cloverleaf ref. `https://github.com/UK-MAC/CloverLeafref`.

[29] Velten, A., Wu, D., Jarabo, A., Masia, B., Barsi, C., Joshi, C., Lawson, E., Bawendi, M., Gutierrez, D., and Raskar, R. Femto-photography: Capturing and visualizing the propagation of light. *ACM Trans. Graph. 32*, 4 (jul 2013).

[30] Wang, C., and Shen, H.-W. Information theory in scientific visualization. *Entropy 13*, 1 (2011), 254–273.

[31] Wang, C., Yu, H., and Ma, K.-L. Importance-driven time-varying data visualization. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1547–1554.

[32] Wang, C., Yu, H., and Ma, K.-L. Importance-driven time-varying data visualization. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1547–1554.

[33] Wang, W., Lyu, G., Shi, Y., and Liang, X. Time series clustering based on dynamic time warping. In *2018 IEEE 9th international conference on software engineering and service science (ICSESS)* (2018), IEEE, pp. 487–490.

[34] Whalen, D., and M.L.Norman. Competition data set and description. `http://sciviscontest.ieeevis.org/2008/index.html`, 2008.

[35] Yamaoka, Y., Hayashi, K., Sakamoto, N., and Nonaka, J. In situ adaptive timestep control and visualization based on the spatio-temporal variations of the simulation results. In *Proceedings of the Workshop on In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization* (2019), pp. 12–16.

[36] Zhang, Y., Guo, H., Shang, L., Wang, D., and Peterka, T. A multi-branch decoder network approach to adaptive temporal data selection and reconstruction for big scientific simulation data. *IEEE Transactions on Big Data* (2021).

[37] Zhou, B., and Chiang, Y. Key time steps selection for large-scale time-varying volume datasets using an information-theoretic storyboard. *Computer Graphics Forum 37*, 3 (June 2018), 37–49.