LA-UR-17-28730
Approved for public release; distribution is unlimited.

Title: Choosing the best partition of the output from a large-scale simulation

Author(s): Challacombe, Chelsea Jordan
Casleton, Emily Michele

Intended for: Report

Issued: 2017-09-26

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Choosing the best partition of the output from a large-scale simulation

Chelsea Challacombe and Emily Casleton, CCS-6

1. Introduction

Data partitioning becomes necessary when a large-scale simulation produces more data than can be feasibly stored. The goal is to partition the data, typically so that every element belongs to one and only one partition, and store summary information about the partition, either a representative value plus an estimate of the error or a distribution. Once the partitions are determined and the summary information stored, the raw data is discarded. This process can be performed in-situ; meaning while the simulation is running.

When creating the partitions there are many decisions that researchers must make. For instance, how to determine once an adequate number of partitions have been created, how are the partitions created with respect to dividing the data, or how many variables should be considered simultaneously. In addition, decisions must be made for how to summarize the information within each partition. Because of the combinatorial number of possible ways to partition and summarize the data, a method of comparing the different possibilities will help guide researchers into choosing a good partitioning and summarization scheme for their application.

In this work we will present a metric that was created to balance the tradeoff between accuracy and storage cost. These competing factors are demonstrated in the hypothetical scenario of Figure 1. Here the accuracy error of the partition, or the ability of the summarized information to recreate the raw data, is depicted on the y-axis with the size of the summarized information on the x-axis. If only size is of interest, the best partitioning scheme would be that highlighted in red on the far left, where if only accuracy is considered, the partitioning scheme represented by the point on the far right would be preferred. However, at the optimal partition for size, the accuracy is at a minimum, and optimizing with respect to only accuracy leads to the largest partitions. Therefore, if both criteria were important, the best partitioning scheme would be that which corresponds to the green circle, or size-error tradeoff. Moving away from this point in either direction on the curve represents a degradation of one of the criteria.

This idea of balancing competing criteria appears in other disciplines as well. In design of experiments, the quality of a design is measured by its ability to estimate or predict precisely but also protecting against bias from model misspecification is desired (Lu, Anderson-Cook, & Robinson, 2011). Other examples include a manufacturing company choosing a supplier by balancing cost and quality of the product (Anderson-Cook & Lu, 2012) or balancing the effectiveness of a pharmaceutical drug for treating a disease against the side effects.

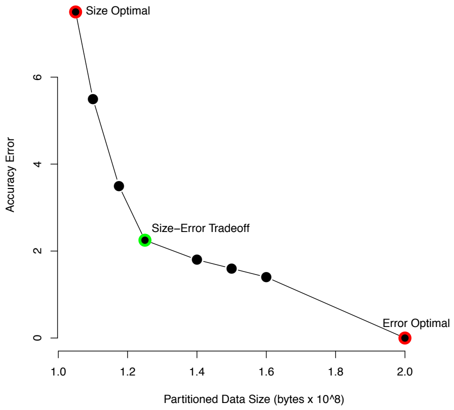


Figure 1-Demonstration of the tradeoff between accuracy error and the size of the partition.

Three datasets will be partitioned by various means, and the resulting summarizations will be compared with the pAIC to determine the most appropriate partitioning scheme. Each data set is a subset of the total output from a large-scale simulation to manage the dataset size and more easily facilitate testing. The first data set was obtained from a MC^2 (Mesh-based Cosmology Code) dark matter simulation (Woodring, Ahrens, Figg, Wendelberger, Habib, & Heitmann, 2011) and will be referred to as the *cosmology data*. The next data set will be referred to as the *ocean data* and was obtained from the Model for Prediction Across Scales-Ocean (MPAS-O) simulation (Ringler, Petersen, Higdon, Jacobsen, Jones, & Maltrud, 2013). Lastly, the *asteroid data* is a simulation of an asteroid entering the atmosphere (Gisler G. R., Weaver, Mader, & Gittings, 2004) (Gisler, Weaver, & Gittings, 2011) obtained from xRage simulation code. Table 1 describes the size of each dataset and the variable of interest.

	Variable of Interest to be summarized	**Size of raw data (number of rows)**
Cosmology data	Velocity in the x-direction	32, 768
Ocean data	Temperature of the water	57,536
Asteroid data	Temperature of asteroid in electron volts	13,253,253

Table 1-Variables of interest to be summarized for each of the three datasets.

These particular data sets were chosen because they represent a broad range of subject areas and motivations. For instance, a main goal of the cosmology data is to identify halos, or a clustering of dark-matter particles, while the ocean simulation output is used to study anthropogenic climate change, and one purpose of the asteroid simulation is to study the ablation of the asteroid as it flies through the atmosphere. In addition, each dataset contains three-dimensional spatial information, allowing the partitioning to be performed on the spatial dimensions.

The rest of the report is organized as follows. The partitioning schemes explored in this report are described in Section 2. Various error metrics, including the metric developed specifically for assessing partitions, are discussed in Section 3. Sections 4-6 demonstrate the usage of the metrics on the three data sets. Sections 7 and 8 explore further considerations for the proposed metric: the choice of weights and the conclusions drawn from the summarized data, respectively. Finally, Section 9 concludes the paper and discusses future work.

2. Partitioning Details

A brief description is provided below on how the partitions under consideration are created, and what choices need to be made within the process to create partitions representing a large amount of raw data. Note that we will refer to the combination of choices as a *partitioning scheme* and that comparing different schemes is the main goal of this work.

The algorithm used to create the partitions is a top-down kd-tree algorithm (Nouanesengsy, Woodring, Patchett, Myers, & Ahrens, 2014), where each split is axis-aligned so that the dividing line is always parallel to one axis and each split is binary, so that each node in the tree is subdivided into two leaves (see an example in Figure 4). For this report, the partitions were created using three spatial variables, and the summarization within each final partition will be on a separate variable of interest, such as temperature or pressure. Because the partitioning variables represent spatial locations, the resulting partitions form an irregular grid over the data. Figure 2 displays resulting partitions for two-dimensional spatial data, where the points represent the location of data collection, and the variable of interest is measured at each location.

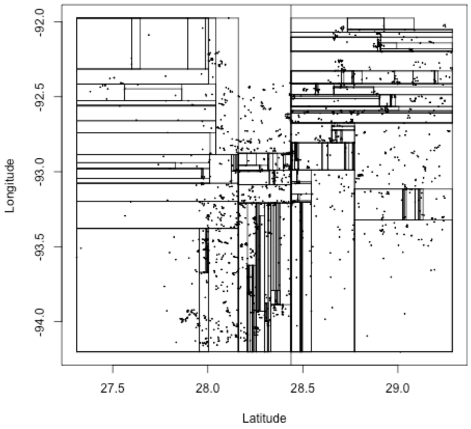


Figure 2 - Irregular grid over two-dimensional data. The points represent the location of data values and, the partitioning is performed by recursively splitting the subset with the largest variance.

The following is a list of the choices that define a partitioning scheme; a detailed description of each will follow.

1. Define the partitioning variables
2. Determine how to cycle through the partitioning variables
3. Determine where to divide the partitioning variables
4. How to decide when the final partitions have been reached
5. Choose a representative value of the final partitions
6. Choose a representation of the error in the final partitions

2.1 Partitioning Variables

In the three example datasets examined here, the partitioning variables represent spatial locations, so that the partitioning results in an irregular grid over the data. When spatial data is available, this choice is intuitive, as you would expect areas of similarity to be spatially co-located. For example, the variable of interest for the ocean data is water temperature, and it is intuitive that there are regions of water with similar temperature. Although not considered here, creating partitions over spatial locations may not always be the most informative; however, the use of a metric will quantify any differences between various choices of partitioning variables.

2.2 Cycling through partitioning variables

Given that there is more than one partitioning variable, one will need to decide which variable to partition at any given step in the algorithm. Under current consideration is to cycle through the variables, so that if the partitioning variables are x, y, and z, the algorithm will partition x, then y, then z, then x, etc. until some stopping criteria has been met.

Another possibility would be to split on whichever variable has, for example, the largest variance. After each step of the algorithm, the variance of all current partitions is calculated, and the one partition with the largest variance is then partitioned. Similarly, another possibility would be to choose the partition with the most points. Again, although only one possibility is considered here, the introduction of a metric will allow for a quantification of different options for a given set of data.

2.3 Partition Location

Once it is determined which variable will be partitioned in the algorithm, another choice is how to perform the partition. Because the algorithm performs splits that are binary and axis-aligned, this decision is equivalent to choosing where to draw a vertical or horizontal line (or plane) through the data. Consider the two dimensional projection of the ocean data, shown as points in Figure 3. If Latitude is the variable to be partitioned, the current options are to partition at the mean, median, or midpoint of the range of the splitting variable. Figure 3 shows where the partition would occur for each option. Note that if the partitioning variables are relatively symmetric, the mean and median partitioning will produce similar results.

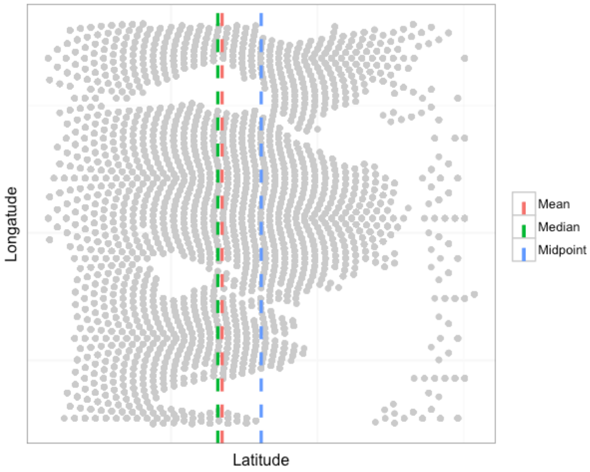


Figure 3-Demonstration of the difference between splitting at the mean, median, or midpoint of the latitude variable for the ocean data.

2.4 Stopping Criteria

Determining when to terminate the partitioning algorithm is the last decision to be made with respect to creating partitions. The stopping criteria will define when the algorithm has arrived at the final partitions. In this work there are four criteria that will be considered for stopping the partitioning algorithm, and individual criterion can also be used together in combination. Note that for the first two criteria listed, it is the variable of interest that is examined, not the partitioning variables.

1. *Variable Range*—This method will examine the variable of interest, (e.g., temperature) within the current partitions and compare each range value to some specified number, say 5. After each iteration of the algorithm, the temperature range is computed within each partition, and if there is at least one with a range greater than 5, the algorithm will continue.
2. *Variable variance*—similar to 1., but with the variance.
3. *Cell count*—This stopping criterion defines the maximum number data values in each final partition. Before each split is performed the number of data values in the partition is checked, and if this number is more than the user-defined cell count, the algorithm will continue.
4. *Number of levels*—This is easiest to describe with a tree, like the one shown in Figure 4. Each branch represents a partition, so the circle at the top denotes all the data. A binary partition is defined on variable x at 0.1, so the orange circle to the left is data less than 0.1, and the pink circle on the right is all the data greater than 0.1. The number of levels criterion examines the deepest branch within the tree, so in this example there are 6 levels.

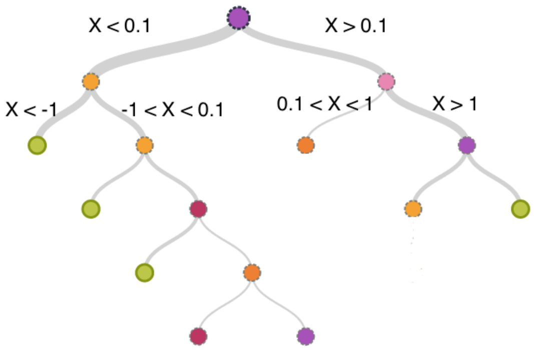


Figure 4-An example tree used to describe the number of levels stopping criterion.

2.5 Partition Representation

Once the final partitions are created, the variable of interest will be summarized within each partition with a single, representative value. In the current work, the representative value will be the mean, median, or midpoint. Future work includes preserving a distributional representation of the variable of interest, rather than a single value; however, this will require an extension of the metrics defined in the following section.

2.6 Error Representation

An important aspect to consider when discarding the raw data is the amount of information lost by representing a collection of potentially variable data with only a single, representative value. The error representation is a summary of the distribution of errors between the chosen representative value and each value of the raw data within the corresponding partition. Under consideration are four representations of the error distribution: the mean or median of the error distribution, the maximum error, or a percentile from the error distribution (1-99%), which represents the value for which p% of the errors are smaller (100% percentile is the same as the maximum error, 50% is equivalent to the median error).

Figure 5 displays a flow chart summarizing the steps of the partitioning approach used in this work. Not all decisions discussed above are displayed in this figure.

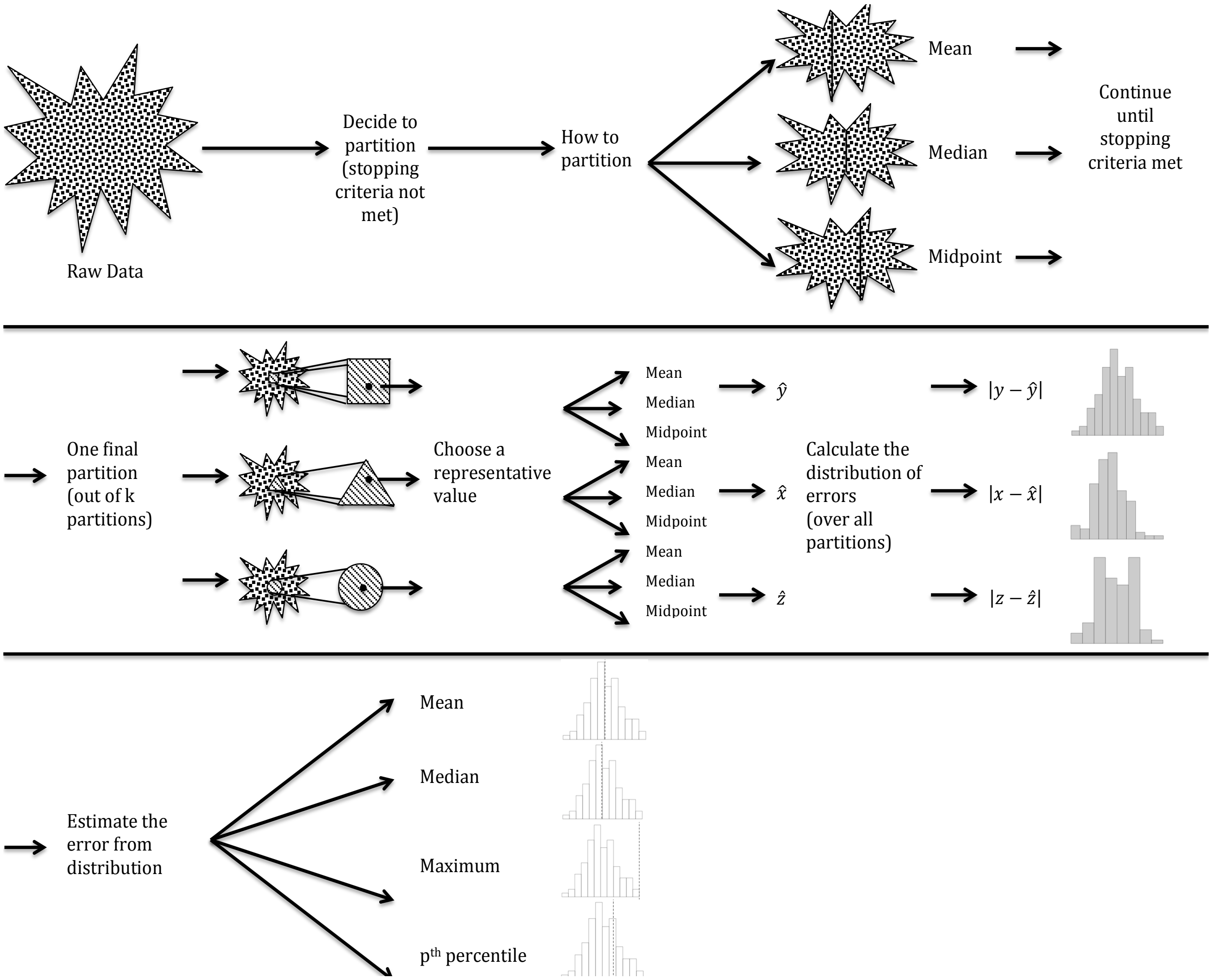

Raw Data
Decide to partition (stopping criteria not met)
How to partition
Mean
Median
Midpoint
Continue until stopping criteria met
One final partition (out of k partitions)
Choose a representative value
Mean
Median
Midpoint
Mean
Median
Midpoint
Mean
Median
Midpoint
$\hat{y}$
$\hat{x}$
$\hat{z}$
Calculate the distribution of errors (over all partitions)
$|y - \hat{y}|$
$|x - \hat{x}|$
$|z - \hat{z}|$
Estimate the error from distribution
Mean
Median
Maximum
p^th percentile

Figure 5- Flow chart displaying the partitioning approach.

3. Metrics

3.1 pAIC

The metric developed to balance the competing criteria of size and accuracy was inspired by the AIC (Akaike Information Criteria), a common measure used in statistics to choose an appropriate model for a given dataset. The AIC balances the tradeoff between goodness of fit of a model and model complexity, and thus has a similar goal in selecting an appropriate fit. In addition, the partitioning scheme can be interpreted as a model of the raw data.

Our metric is called the pAIC (partition-AIC) and is computed with the following formula:

$$pAIC = \omega_1\left(\frac{k}{\sqrt{N}}\right) + \omega_2\left(\frac{\sum_{p=1}^{k} n_p\left[\sum_{i\in p}\frac{(y_{ip}-\hat{y}_p)^2}{n_p}\right]}{\sigma^2}\right) + \omega_3\left(\frac{\sum_{p=1}^{k} n_p\left[\sum_{i\in p}\frac{[|y_{ip}-\hat{y}_p|-e_p]^2}{n_p}\right]}{(0.5*range)^2}\right)$$

Equation 1-pAIC

where the N values of the raw data are represented as $y_{ip\,,}\ p = 1, \dots, k$ where the partitioning scheme has p partitions; the n_p raw values in partition p are summarized with representative value $\hat{y}_p$ and error e_p. The $\omega_1, \omega_2, \omega_3$ are weights that sum to 1 and place relative importance on each term. In words, the first term is a penalty for the number of partitions scaled by the square root of N; the second term is the mean of squared errors (MSE), or a measure of the average accuracy of the representative value within partition p, which is then scaled by the variance. The final term is a measure of the average accuracy of the estimated error, which is scaled by half of the range squared. The denominators in each term are intended to scale the term by the worst-case scenario so that each term contributes to the overall metric on a comparable scale.

As with the original AIC, smaller values of the pAIC indicate a better tradeoff of accuracy and size, while the magnitude of the metric values are irrelevant. Thus, the pAIC is not appropriate for comparing partitioning schemes across datasets or variables within the same dataset.

The performance of the pAIC will be compared to four other commonly used metrics within the statistics and computer science literature: root mean squared error (RMSE), signal to noise ration (SNR), JPEG precision, and the correlation coefficient. Each metric will be described in turn.

3.2 RMSE

The RMSE is a commonly used metric in computer science literature to compare differences between two sets of data; often, one of which is observed and the other is estimated. RMSE is a good estimate of accuracy, but depends on the scale of the original data. So, as with the pAIC, RMSE should not be used to compare across different datasets. The formula for RMSE:

$$RMSE = \sqrt{\sum_{p=1}^{k} n_p\left[\sum_{i\in p}\frac{(y_{ip}-\hat{y}_p)^2}{n_p}\right]}$$

Equation 2-RMSE

which is the square root of the second term from the pAIC. Smaller values of RMSE indicate a more accurate partitioning scheme. Notice that the RMSE, as well as all other metrics discussed in this section, do not include a penalty term for size, and thus take only one aspect of the size-accuracy tradeoff into consideration. The RMSE will tend to be a non-increasing function of the number of partitions as a result.

3.3 SNR

Another popular metric, particularly in engineering, is the signal to noise ratio, which compares the amount of discernable signal in the data to the amount of background noise. The comparison is accomplished through a proportion, so values greater than 1 indicate more signal than noise, while SNR values less than 1 imply more noise than signal. Particular applications will compute signal and noise differently, but for the partitioning scheme, we will compute SNR as

$$SNR = \frac{\sigma^2}{\sum_{p=1}^{k} n_p \left[\sum_{i \in p} \frac{(y_{ip} - \hat{y}_p)^2}{n_p}\right]}$$

Equation 3-SNR

where σ^2, the signal, or numerator, is the amount of variability in the raw data and the noise, or denominator, is represented by the MSE. The numerator will be constant across all partitioning schemes of a given dataset, thus the SNR is a scaled inverse of the RMSE. Unlike the previous two metrics, a larger value of SNR indicates a more accurate representation. This is also the inverse of the second term of the pAIC.

3.4 JPEG Precision

The next metric was developed in conjunction with work on JPEG 2000 compression (Woodring, Mniszewski, Brislawn, DeMarle, & Ahrens, 2011) and was named *precision*. Because precision typically refers to the inverse of the variance in statistics, this metric will be referred to as JPEG precision in this work. This metric is computed as follows:

$$JPEG\ precision = \frac{\sum_{p=1}^{k} n_p \left[\sum_{i \in p} \frac{|y_{ip}|}{n_p}\right]}{\max_{p}\left\{\max_{i \in p}\left|y_{ip} - \widehat{y_p}\right|\right\}}$$

Equation 4-JPEG precision.

Thus, the metric represents a ratio of the average magnitude of the data to maximum error over all partitions. The JPEG precision metric differs from those previously discussed because it contains the maximum error, rather than an average error. The advantage of using an average error is that it is a summary of the entire error distribution, where the maximum error represents only one value in the extreme. However, an extreme error may be diminished and not evident if examining an average error, particularly when summarizing a large number of partitions, while the maximum error will highlight if this extreme value exists. As with SNR, more desirable partitioning schemes will have larger values of JPEG precision.

3.5 Pearson Correlation Coefficient

The last metric examined in this work is the Pearson product-moment correlation coefficient, or correlation coefficient, $r \in [0,1]$. This metric measures the strength of the linear relationship between two sets of data. Here, we take those sets of data to be the raw data and corresponding predicted value. The formula is as follows

$$r = \frac{\sum_{p=1}^{k}\left[\sum_{i\in p}(y_{ip} - \bar{y})(\widehat{y_p} - \bar{\hat{y}})\right]}{\sqrt{\sum_{p=1}^{k}\left[\sum_{i\in p}(y_{ip} - \bar{y})^2\right]}\sqrt{\sum_{p=1}^{k}\left[\sum_{i\in p}(\widehat{y_p} - \bar{\hat{y}})^2\right]}}$$

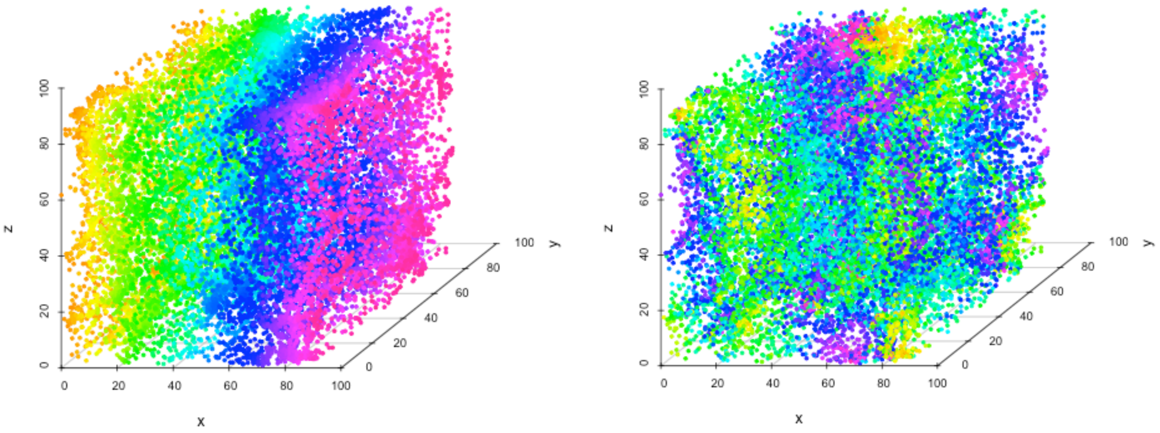
Equation 5-Pearson correlation coefficient

where $\bar{y}$ is the average of the raw data, and $\bar{\hat{y}}$ is the average of the predicted values over all partitions. If all the representative values, $\widehat{y_p}$, were exactly equal to all the raw data values, y_{ip}, then the correlation coefficient will be its maximum value of 1. However, if there is a systematic additive bias in the representative values, this maximum could also be attained. In general, the closer the correlation coefficient is to 1, the more accurately the representative values are estimating the raw data.

As previously mentioned, the metrics described in sections 4.2-4.5 do not include a penalty term for the size of the partition. In addition, none of these metrics consider the quality of the error estimate within each partition. Each examines only how well the representative value estimates the raw data.

4. Cosmology Data Results

The cosmology datasets consists of a large-N body simulation of dark matter physics from the Road-Runner Universe MC^3 (Habib, et al., 2009). One time step consists of 4000^3, or about 64 billion, particles with 36 bytes per particle, resulting in a dataset of 2.3 TB per time step. The specific dataset analyzed here was obtained through in-situ subsampling (Woodring, Ahrens, Figg, Wendelberger, Habib, & Heitmann, 2011) and consists of only 32,768 points in space and one time step. The three-dimensional spatial locations of the subsampled points are displayed in Figure 6. The plot on the left displays the particles colored by their value on the x-axis. Even with only a small fraction of the data, the space is still dense. Coloring of the plot on the right corresponds to the variable of interest, the velocity of x, or vx. This plot indicates that there are regions of similar values of velocity and that this dataset could benefit from spatial partitioning on the velocity. It should be noted that if this method were to be used in practice, the partitioning would be performed on the 64 billion particles; however, for demonstration and testing purposes, the small subset is examined here.



(a) Spatial locations of cosmology points colored by x.

(b) Spatial locations of cosmology points colored by the variable of interest, vx.

Figure 6-Three-dimensional spatial locations of the particle data from the cosmology dataset.

Through the pAIC, the effect of splitting location and how to summarize the resulting partitions can be examined. These correspond to decisions number 3 and 5. In addition, different partition numbers result from using various maximum cell count as stopping criteria (decision 4). All decisions that were made to arrive at these partitions are summarized in **Table 2**. As the maximum cell count increases, the number of partitions decreases, as demonstrated in Figure 7 for the partitioning schemes. The number of partitions is plotted on a log scale to better highlight the trend. Notice that for the median, there are only five distinct partitions (i.e., multiple values of the maximum cell count produced identical resulting partitions), in contrast to the mean and midpoint, which resulted in 19 unique partitionings.

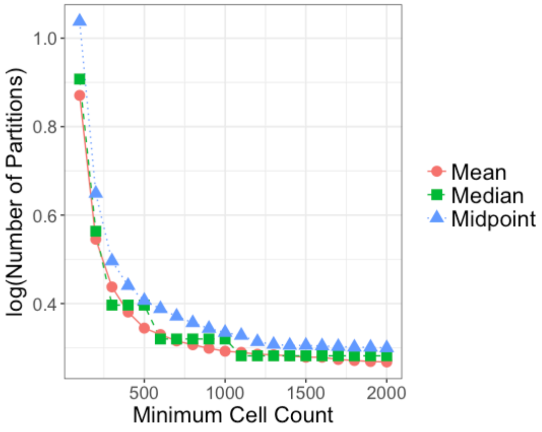


Figure 7-Plot of minimum cell count against the log of the number of partitions.

Figure 8 shows the pAIC against the log of the number of partitions resulting from a particular partition scheme. Again, the log is used for visualization purposes only. Because the splitting is done on the spatial locations, the differences between the various curves of Figure 8 result from a non-symmetric distribution of points in the raw data in each dimension. Figure 9 displays the distributions of the three dimensions of spatial locations. Notice how the x-dimension distribution is relatively symmetric, but points in the y and z dimensions are not, which leads to differing partitions based on where the splitting is performed.

	Mean Partitioning Scheme	**Median Partitioning Scheme**	**Midpoint Partitioning Scheme**
1. Splitting variables	x,y,z	x,y,z	x,y,z
2. Cycle through	Round robin	Round robin	Round robin
3. Partition Location	Mean	Median	Midpoint
4. Stopping Criteria	Cell Count (100, 200, . . ., 2000)	Cell Count (100, 200, . . ., 2000)	Cell Count (100, 200, . . ., 2000)
5. Partition Representation	Mean	Median	Midpoint
6. Error Representation	Mean	Mean	Mean
7. ω_1, ω_2, ω_3	0.25, 0.25, 0.5	0.25, 0.25, 0.5	0.25, 0.25, 0.5

Table 2-Decisions from section 2, as well as the pAIC parameters, used to create the partitions summarized in **Figure 7**, **Figure 8**, and **Figure 10**.

The most appropriate partition, as indicated by the pAIC, is that which produces the minimum pAIC. Therefore, if the mean partition and representation is desired, the minimum pAIC occurs for 27 partitions, which results from setting the minimum cell count stopping criteria to 1700. For the median representation and partition location the most desirable partitioning scheme results from 32 partitions or when the stopping criteria is between 1100 and 2000, and for midpoint, 36 partitions which result with a minimum cell count of 1400 is preferred. Note that for mean and midpoint, the minimum pAIC does not occur with the smallest partition size, but rather the fourth and sixth smallest number of partitions. If a particular representation and partition location is not specified, the most appropriate partitioning scheme according to pAIC is to use the mean because it produced the smallest pAIC of the three options.

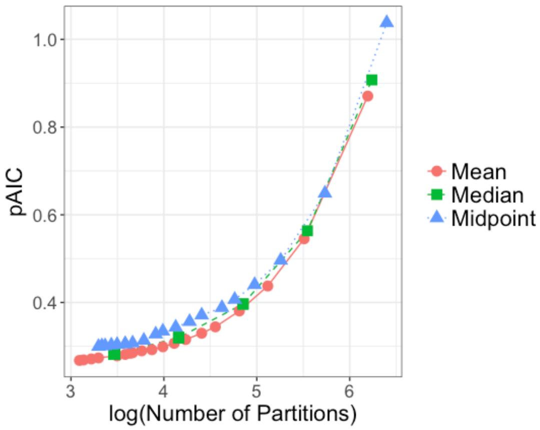


Figure 8-Effect of varying how to split and how to summarize on the pAIC when applied to the cosmology data.

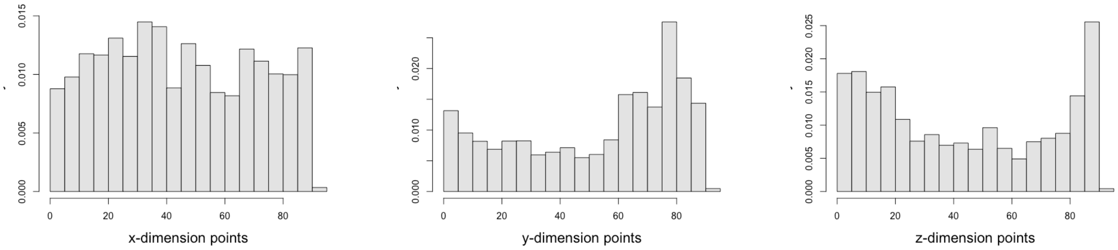
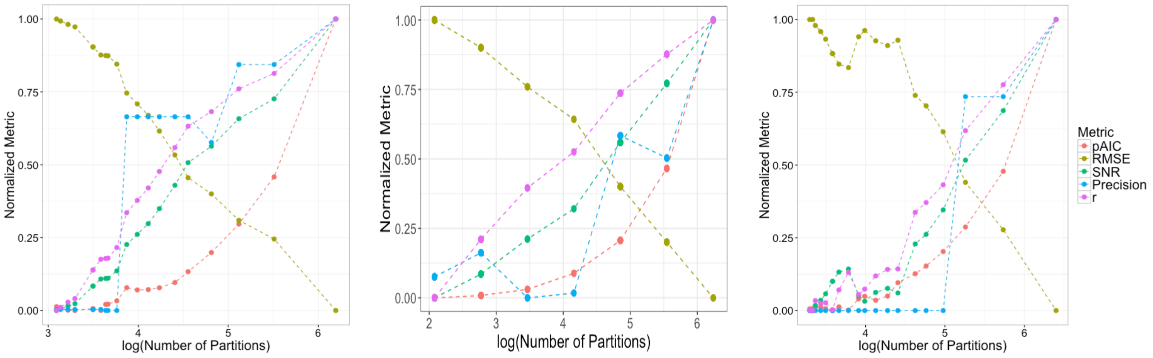


Figure 9-Distribution of the three dimensions of spatial locations for the cosmology data

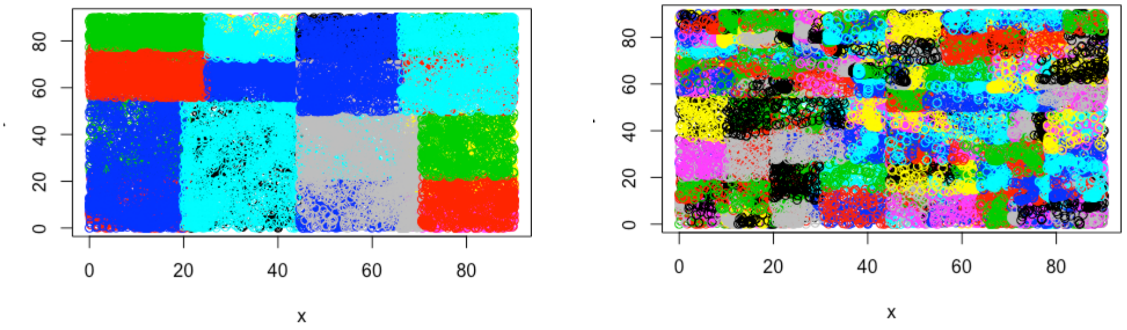
A comparison between the various partitioning metrics is displayed in Figure 10. Because some metrics are on a different scale, the values within each metric and for each partitioning scheme have been standardized so that all values lie between 0 and 1. The standardization occurs by dividing the metrics by the maximum, so that for each scheme and each metric, there is at least one value of 1. From Figure 10 it can be seen that the partitioning scheme that minimizes the RMSE and maximizes the SNR, precision, and correlation, is the scheme with the largest number of partitions. The only exception is the SNR for the median partitioning scheme, which favors a smaller number of partitions. Because of the penalty of the number of partitions, the pAIC actually choses a partitioning scheme that would be much smaller to store.



(a) Mean Partitioning Scheme (b) Median Partitioning Scheme (c) Midpoint Partitioning Scheme

Figure 10-Number of partitions plotted against the five normalized metrics for the three partition criteria summarized in **Table 2**.

Figure 11 compares the partitioning schemes determined to be the best by pAIC and the other, non-penalized metrics. The points on the left are colored by partition and represent only 27 partitions. On the right the points are colored by the 490 partitions that were most desirable given the more traditional metrics. Both plots show the points projected into the x-y plane because it resulted in a better visualization.



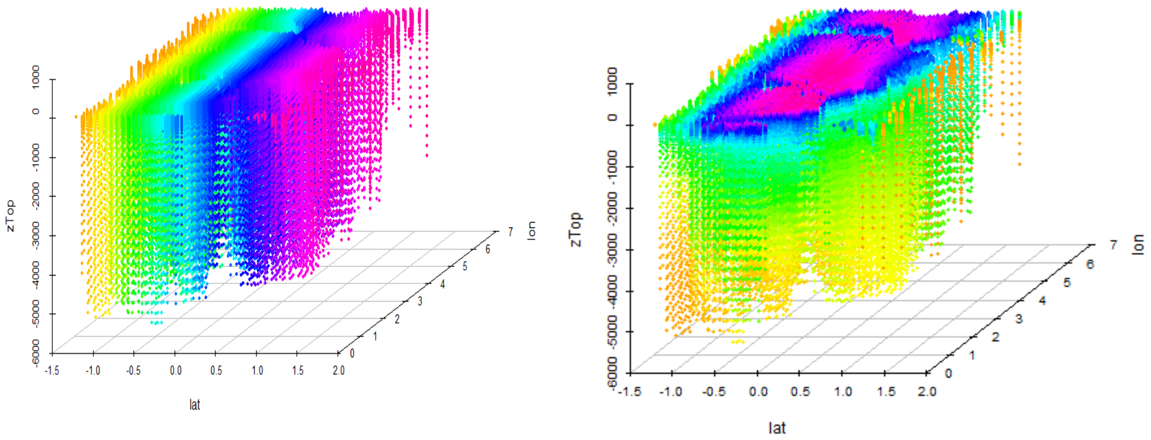
(a) Points projected onto x-y plane and colored according to most desirable partitioning scheme with respect to pAIC.

(b) Points projected onto x-y plane and colored according to most desirable partitioning scheme with respect to RMSE, SNR, Precision, and r.

Figure 11-Comparison of the "best" partitioning schemes with respect to different metrics by projecting the 3d points into the x-y plane.

5. Ocean Data Results

The ocean data was simulated using the Models for Prediction Across Scales-Ocean (MPAS-O). MPAS is set of environmental simulation codes jointly developed at the National Center for Atmospheric Research and Los Alamos National Laboratory for the purpose of studying anthropogenic climate change. Simulation of the ocean is done across many spatial and time scales. The advance MPAS-O provides over other global ocean models is its ability to resolve various resolutions in a single simulation, allowing it to accurately reproduce mesoscale ocean activity (Ringler, Petersen, Higdon, Jacobsen, Jones, & Maltrud, 2013). The points of interest that we will be partitioning are displayed in Figure 12 colored on the left by value of latitude (x-dimension) and colored on the right by the variable of interest, the temperature. The z-dimension is ocean depth, and this plot displays that there is more variation in water surface temperature than temperature along the depth.



(a) Spatial locations of points from the ocean simulation colored by the x variable, latitude.

(b) Spatial locations of ocean data points colored by the variable of interest, temperature.

Figure 12-Three-dimensional spatial locations of the particle data from the ocean dataset.

As with the cosmology data, the points analyzed here represent only a subset of a single simulation. The 57,536 points are located around the equator in the Gulf of Guinea off the coast of Central Africa (see Figure 13).

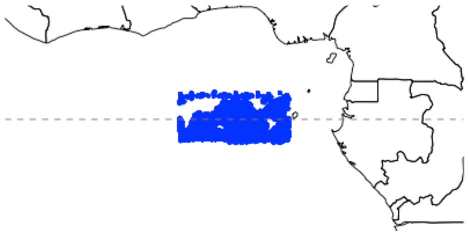


Figure 13-Spatial location of points analyzed as the ocean data. Land mass picture is Central Africa. Horizontal line represents the equator.

The effect of where to split and how to summarize the resulting partitions (decisions 3 and 5) can be examined over varying minimum cell count stopping criteria (decision 4) creating a varying number of partitions. The specific decisions used in the analysis are shown in **Table 3**.

	Mean Partitioning Scheme	**Median Partitioning Scheme**	**Midpoint Partitioning Scheme**
1. Splitting variables	latitude, longitude, ocean depth	latitude, longitude, ocean depth	latitude, longitude, ocean depth
2. Cycle through	Round robin	Round robin	Round robin
3. Partition Location	Mean	Median	Midpoint
4. Stopping Criteria	Cell Count: (100,150, . . ., 500,1000,1500, . . .,5000)	Cell Count (100,150, . . ., 500,1000,1500, . . .,5000)	Cell Count (100,150, . . ., 500,1000,1500, . . .,5000)
5. Partition Representation	Mean	Median	Midpoint
6. Error Representation	Mean	Mean	Mean
7. ω_1, ω_2, ω_3	0.1, .01, 0.8	0.1, .01, 0.8	0.1, .01, 0.8

Table 3 - Decisions from section 2, as well as the pAIC parameters, used to create the partitions summarized in Figure 14 and Figure 15.

The resulting pAIC against the log of the number of partitions is shown in Figure 14. Note the J shape for all curves. This pattern indicates the pAIC decreases (and thus, performance of the partitions increases) as more partitions are added up to a point. After this point, the cost of adding more partitions does not outweigh the added precision, and thus the pAIC begins to increase. For the mean curve, the minimum pAIC occurred for 1500 minimum cell count, which resulted in 61 partitions. Minimum for the median curve occurred with the 16 partitions created when the minimum cell count was 4000, and the midpoint occurred when the minimum cell count was set to 2500 for 42 partitions. Note that for the various partition location and representation, different values of the stopping criteria led to different number of partitions. Again, splitting with the mean of the distribution and summarizing the resulting partitions with the mean lead to the minimum pAIC.

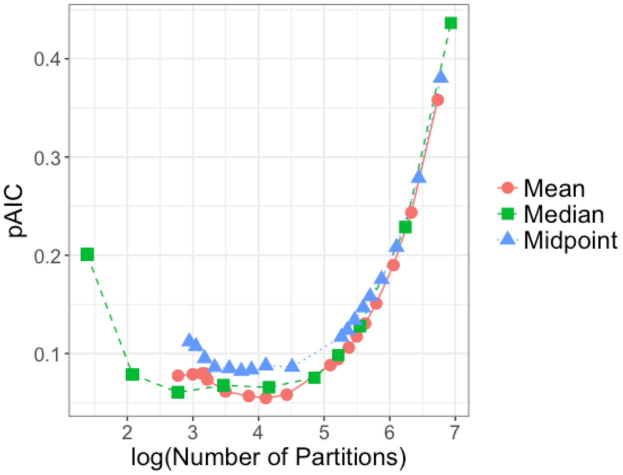
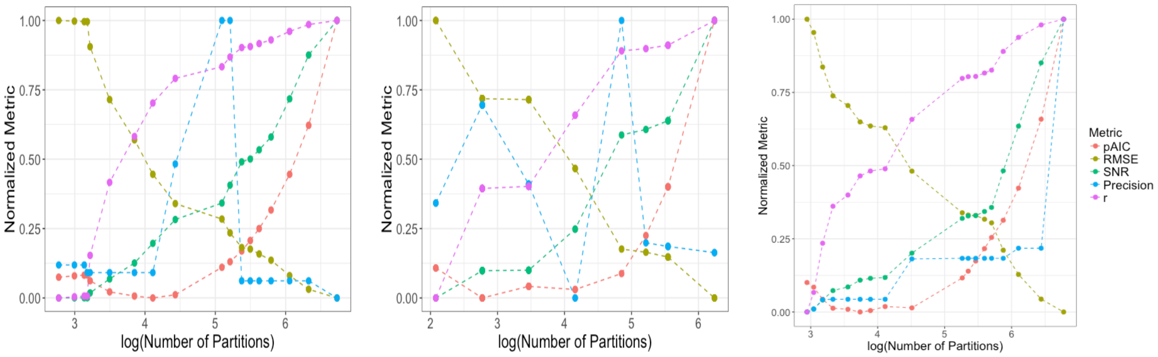


Figure 14 - Effect of varying how to split and how to summarize on the pAIC when applied to the ocean data.

Figure 15 compares all the metrics under consideration for the three partitioning scenarios described in **Table 3**. A similar pattern as was seen in the cosmology data is seen here in that most metrics in all three plots, other than the pAIC, are either maximized or minimized by the scheme with the most partitions. The one exception is JPEG precision, which does not indicate that the largest number of partitions is the most desirable. This metric is also the only one to consider the maximum error.



(a) Mean Partitioning Scheme

(b) Median Partitioning Scheme

(c) Midpoint Partitioning Scheme

Figure 15- Number of partitions plotted against the five metrics for the three partition criteria summarized in **Table 3**.

Note that the value of the ω$_1$, ω$_2$, ω$_3$ in the last line of **Table 3**. These numbers were chosen through trial and error with the goal that each of the three terms contributed to the pAIC and that one term did not dominate the metric. To further explore this concept, consider Figure 16. This plots the percentage of the pAIC metric each of the three terms contributed to the pAIC metric against the number of partitions for the mean partitioning scheme(first column in **Table 3** with results displayed in Figure 15 (a)). As the number of partitions increases, the first term, which represents the number of partitions, begins to dominate the metric. The black, dashed vertical line indicates the point at which the pAIC is minimized. At this location the percentage contribution for each term is 0.46, 0.32, and 0.20, respectively. Preventing the dominance of a single term more generally is an area of future work and the choice of weights will be discussed further in Section 7.

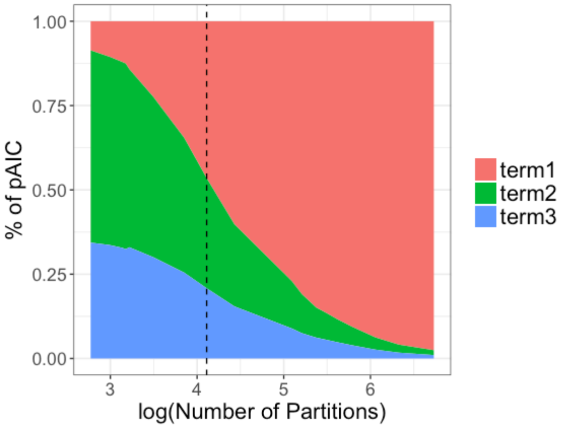


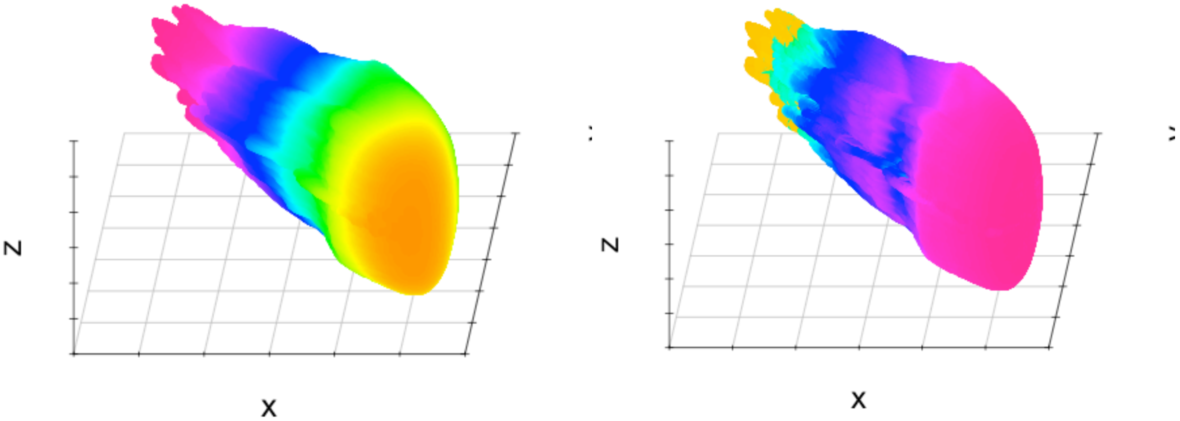
Figure 16-Percentages each of the three terms in the pAIC contributes to the overall metric, plotted against the log of the number of resulting partitions.

6. Asteroid Data Results

Seventy-one percent of the earth is covered by water and of that, 96.5% of the water is contained in an ocean (United States Geological Survey, 2016). If an asteroid is to enter the earth's atmosphere, it will most likely crash into an ocean. The large amount of kinetic energy of the asteroid is transferred to the atmosphere and water, forcing water and water vapor into the air (Patchett, et al., 2016) in addition to hurricane-force winds and air temperatures in excess of 100 C. Tsunamis are also likely to result from such an event, although worldwide devastation is not a probable as once thought (Gisler, Weaver, & Gittings, 2011). Costal communities near a point of impact would be in grave danger as the resulting waves could reach heights that are significant fractions of the total ocean depth (Gisler, Weaver, & Gittings, 2011). Long-term, global weather and climate consequences could occur if water were to reach the stratosphere. The residence time in this atmospheric level is decades (in contrast to weeks in the lower, troposphere), but the resulting effects are undetermined to be warming because of greenhouse effects or cooling from the formation of ice clouds (Gisler, Weaver, & Gittings, 2011).

The mechanism of asteroids crashing into oceans of most interest is the transfer of kinetic energy from asteroid to atmosphere and water (Patchett, et al., 2016). Parameters that affect this exchange is asteroid size and mass, the angle the asteroid enters the atmosphere, and if the asteroid burst and at what elevation the airburst occurred (Samsel, Rogers, Patchett, & Tsai, 2017). The entry of the asteroid under various parameter settings can be simulated using xRage, a multi-physics, parallel Eulerian hydrodynamics code (Patchett, Nouanesengsy , Gisler, Ahrens, & Hagen, 2017), which was developed and is maintained by the Advanced Scientific Computing program at Los Alamos National Laboratory (Patchett, et al., 2016). The simulation is run on a computational mesh, where the cell size and placement are adaptive so that at each time step more, smaller cells are placed where more is occurring in the simulation.

The parameter of asteroid size, angle impact, and airburst or no, were initially varied to determine the lower bound of dangerous asteroids that NASA needed to track (Samsel, Rogers, Patchett, & Tsai, 2017). A secondary mechanism is to study the ablation of the asteroid before it hits the water (Patchett, Nouanesengsy , Gisler, Ahrens, & Hagen, 2017). Because the behavior of the asteroid before it hits the water is of interest the output of the simulation has been subsetted to only those cells with a partial asteroid density greater than 0. Therefore, at time steps before the asteroid impacts the water, this eliminates the cells that include only water or atmosphere a distance away from the asteroid. The cells of the output that will be partitioned in this study are shown in Figure 17. The plot on the left is colored by location in the y direction, while the plot on the right is colored by the variable of interest, which is temperature. Note that the asteroid is hottest at the front as it is blasting through the atmosphere and the tail that is ablating off is cooler.



(a) Spatial locations colored by y.

(b) Spatial locations colored by the variable of interest, temperature.

Figure 17- Three-dimensional spatial locations of the particle data from the asteroid dataset.

As in the previous examples, the effect of where to split and how to summarize the resulting partitions (decisions 3 and 5) are set at the three possibilities of mean, median, and midpoint. Instead of minimum cell count, the algorithm is terminated with a stopping criteria of a minimum range (decision 4) over the sequence of (0.7, 0.75, . . ., 1.7). All decisions used to create the partitions are shown in Table **4**.

	Mean Partitioning Scheme	**Median Partitioning Scheme**	**Midpoint Partitioning Scheme**
1. Splitting variables	x,y,z	x,y,z	x,y,z
2. Cycle through	Round robin	Round robin	Round robin
3. Partition Location	Mean	Median	Midpoint
4. Stopping Criteria	Maximum Range (0.7, 0.75, . . ., 1.7)	Maximum Range (0.7, 0.75, . . ., 1.7)	Maximum Range (0.7, 0.75, . . ., 1.7)
5. Partition Representation	Mean	Median	Midpoint
6. Error Representation	Mean	Mean	Mean
7. ω_1, ω_2, ω_3	0.1, 0.1, 0.8	0.1, 0.1, 0.8	0.1, 0.1, 0.8

Table 4- Decisions from section 2, as well as the pAIC parameters, used to create the partitions summarized in Figure 19 and Figure 20.

The various values of the maximum range within a single partition leads to a varying number of partitions. The resulting number of partitions is plotted against the maximum range value in Figure 18. As the maximum range increases, the number of partitions decreases, so as the maximum range within each partition must be smaller, more partitions are needed. In contrast to the analysis of the cosmology data in Section 5, each partitioning scheme on the asteroid data created a unique number of partitions, although this could be a result of the fact that the asteroid dataset is roughly 400 times larger than the cosmology dataset.

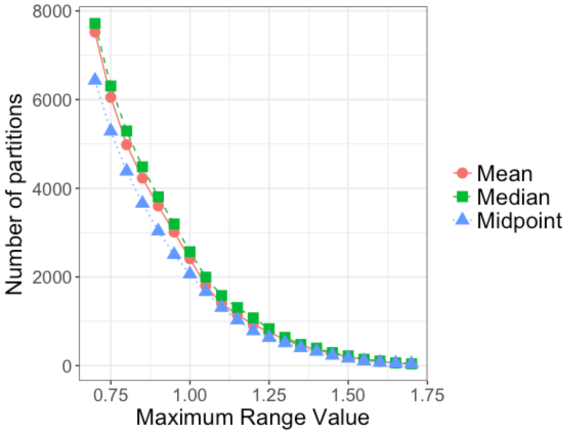


Figure 18-The effect of changing the maximum range on the resulting number of partitions for the mean, median, and midpoint partitioning schemes on the asteroid data.

The resulting pAIC values for each of the partitioning scenarios are plotted in Figure 19. Again, for each partitioning scheme there is the J-shape, indicating that at some number of partitions, the increase in accuracy does not outweigh the increase in storage. For the mean partitioning scheme, the minimum occurs when the maximum range is set to 1.35 resulting in 455 partitions, the median is close with a maximum range of 1.4 and 394 partitions, and the minimum of the midpoint partitioning scheme occurs at the maximum range value of 1 for 2065 partitions. As can be clearly seen, the overall minimum occurs for the mean partitioning scheme.

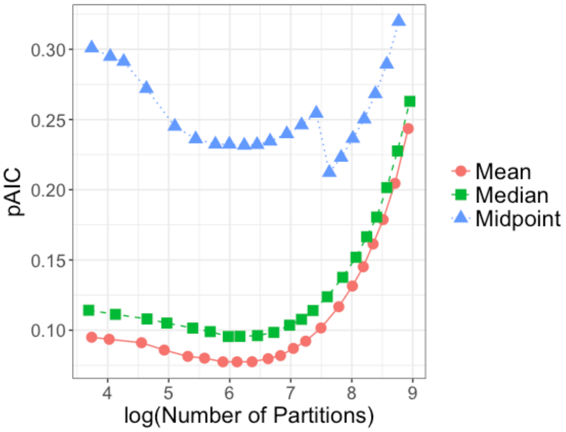
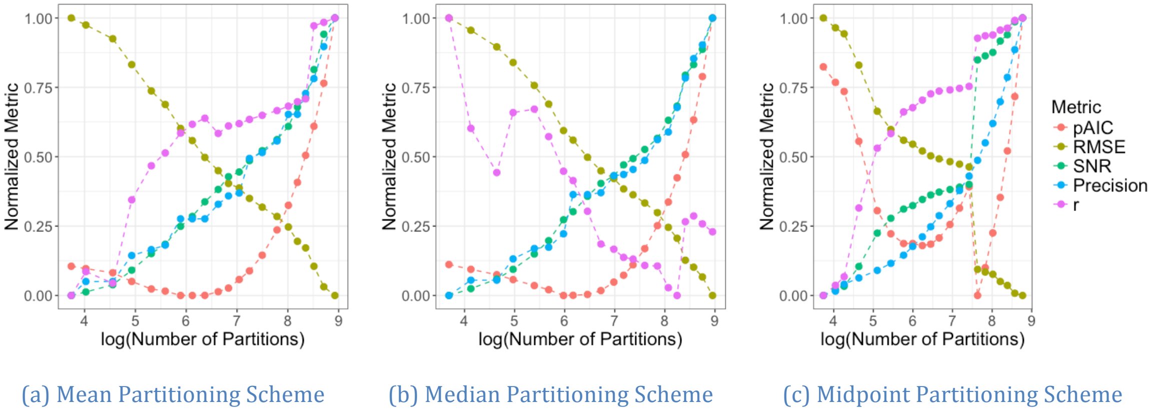


Figure 19 -Effect of varying how to split and how to summarize on the pAIC when applied to the asteroid data.

An interesting characteristic of the midpoint plot in Figure 19 is the drop in the curve that occurs when the number of partitions is 2065 (or $\log(2065) = 7.63$). This interesting behavior is also seen in the other metrics for the midpoint at this location, as shown in Figure 20. The general interpretation of the three plots in Figure 20 is the same as before: The pAIC balances competition criteria, while the other metrics consider only a single criterion and thus always recommend adding more partitions.



(a) Mean Partitioning Scheme (b) Median Partitioning Scheme (c) Midpoint Partitioning Scheme

Figure 20 -Number of partitions plotted against the five metrics for the three partition criteria summarized in Table **4**.

7. Weights

When comparing partitioning schemes with the pAIC, one choice the researcher must make is the choice of the three weights, {ω_1, ω_2, ω_3} on the terms that represent number of partition or size of resulting summarization, accuracy through the mean square error, and precision through the accuracy of the estimated error. This choice should be driven by which of the three criteria is important to the researcher, as the weights represent the relative importance of each term of the pAIC. However, choosing the weights appropriately requires expert knowledge about the process and a prior knowledge about the possible values each term can take (Lu, Anderson-Cook, & Robinson, 2011). In addition, multiple experts could have multiple opinions on ranking the importance the terms.

The pAIC metric is similar to a desirability function in design of experiments. Both are linear combinations of multiple criteria, where the intent is to make a decision based on one collection of user-defined weights (Lu & Anderson-Cook, 2012). One disadvantage identified for the desirability function approach is the sensitivity of the solution (here chosen design) to the specified weights. One computationally expensive resolution is to minimize the desirability function for many sets of weights. Lu, Anderson-Cook and Robinson 2011 present the Pareto Front approach, which adds rigor and structure to decision-making when constrained by a choice in weights (Chapman, Lu, & Anderson-Cook, 2014), and is directly applicable to the pAIC discussion.

The Pareto frontier approach is a two-step procedure developed to determine an appropriate design for an experiment. The first step is the objective step (Chapman, Lu, & Anderson-Cook, 2014) where inferior designs are removed from consideration. Those still considered are designs on the Pareto front, which represent designs such that no design can improve a single criterion without decreasing at least one other criterion (Lu, Anderson-Cook, & Robinson, 2011). This concept is demonstrated for two criteria in Figure 21, which was directly copied from (Lu, Anderson-Cook, & Robinson, 2011). In this example, the goal is to maximize both criteria, so larger values are better. The Utopia point represents the best possible solution, as both criteria are simultaneously maximized; however, this point is often unattainable. The Criterion Space demonstrates all possible solutions. The Pareto Front is shown as the set of points within the criterion space that are equidistant to the Utopia point. Moving away from the Pareto front represents a decrease in at least one of the criteria. Once the Pareto front has been identified, the second, subjective, step of the Pareto front approach involves a detailed examination of each contending solutions in terms of the trade-offs between criteria and sensitivity to weights to finally arrive at one single solution.

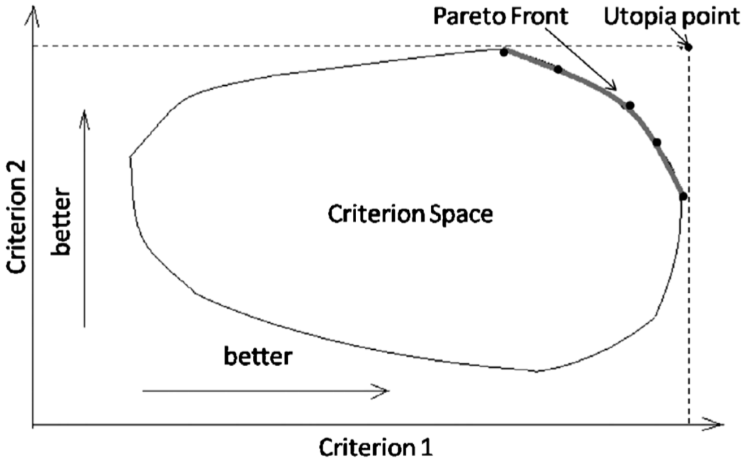


Figure 21-Pareto front in two-dimensions where two criteria are to be maximized. The Utopia point is the best possible point, but often unattainable. Figure copied from (Lu, Anderson-Cook, & Robinson, 2011).

One way to examine and compare the Pareto front solutions proposed by (Lu, Anderson-Cook, & Robinson, 2011) is with a mixture plot. An example for the asteroid analysis is shown in Figure 22. Because the sum of the three weights is 1, each point in a mixture plot represents one combination of the three weights. The vertices of the mixture plot represent only one term, where the edges represent only two terms. The larger the area a specific partitioning scheme represents in the mixture plot, the more robust the scheme is to changes in the weights.

In the application of the Pareto front approach to the asteroid dataset, the partitioning approaches under consideration are the three partitioning schemes (Mean-Median-Midpoint) summarized in Table **4** for the stopping criteria of maximum range from (1, 1.05, 1.1, . . ., 1.7) and the same three partitioning schemes for the maximum cell count of (5k, 10k, . . ., 100k). Thus, there are 105 (= 3*(20 cell counts values) + 3*(15 range values)) possible partitioning schemes under consideration. The criterion space is in three dimensions; representing the three, unweight terms of the pAIC, with the goal of each to be minimized. Therefore, the Utopia point is the point (0,0,0). The Pareto front approach identified 29 of the 105 schemes as plausible and that should be considered further.

The partitioning scenario with the largest area of the mixture plot is scenario 86, which represents 15% of the triangle. This partitioning scheme was obtained by setting the stopping criteria to a maximum value count of 5000 with a midpoint partitioning scheme. In fact, the midpoint-cell count partitioning schemes (labeled 86-105, representing the various values of maximum cell count) represent 51% of the triangle. This result is interesting in contrast to those found in Section 7, where the midpoint showed the least favorable results for one specific weight combination; although, the results of Section 7 and for when range was used as a stopping criteria. In general, the midpoint-cell count partitioning schemes perform well when terms 1 and 3 are weighted more heavily, but other partitioning schemes are more appropriate if term2 has larger weights (scenarios 47-51 are from a maximum cell count, mean partitioning scheme and 1-11 are maximum range, mean partitioning scheme).

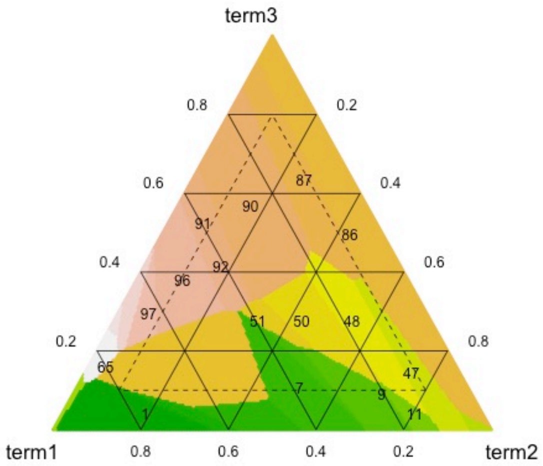


Figure 22 -Mixture plot for the asteroid analysis showing which partitioning scenarios minimize the pAIC for different combinations of weights, {ω_1, ω_2, ω_3}.

8. Hypothesis testing

A main area of concern for the research scientists interested in analyzing the cosmology, ocean, and asteroid datasets is, "will the summarized data give me the same conclusions to questions of interest, as the raw data would have?" The pAIC attempts to quantify an answer to this question by balancing precision, error estimation, and subsample size. Researchers are hesitant to commit to any specific set of research questions, as they fear they may discover an interesting feature during an analysis and would like the ability to dig further into any potential future considerations. However, until raw storage increases at the same pace as computation ability, data will need to be discarded. In addition, the ability to process the data and digest and plot it to arrive at meaningful conclusions becomes more difficult the larger the dataset. Therefore, the pAIC allows researchers to choose an intelligent subset of the data while retaining the ability to draw conclusions from the data.

For the ocean data, three of the six top diagnostics identified by subject matter experts are averages over space, (e.g., Laborador Sea average temperature), time (e.g., average ocean temperature in June), or weighted global averages. Therefore, we will explore the conclusions drawn from tests of the mean on the summarized data and the raw data.

Figure 23 shows the resulting p-value when performing a test of the mean for various null values of the mean, μ_0, when testing with the raw data and separately the summarized data. The summarized data was obtained from the partitioning scheme chosen by the pAIC as the most appropriate: the mean partitioning scheme with a maximum cell count of 1500. The black horizontal line indicates a p-value of 0.05, the most common level of significance. For the summarized data, the result of the hypothesis test would be to fail to reject for all hypothesized means in the range because all points fall above the 0.05 cut-off. The hypothesis test for the raw data would be rejected for values that are far away from sample mean of 8.976. Therefore, different conclusions are reached for the majority of the examined range when using the raw versus the summarized data.

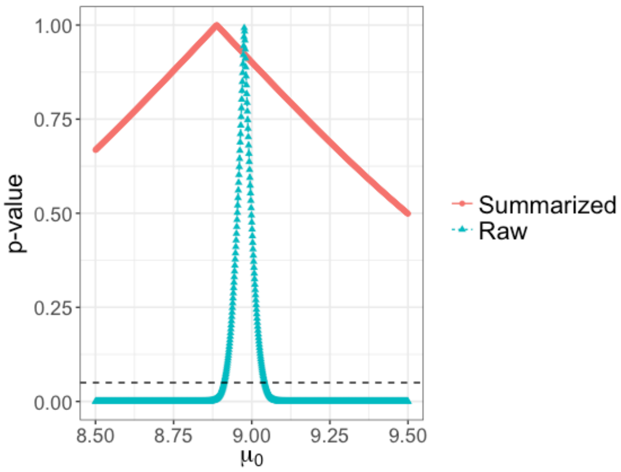


Figure 23-Resulting p-value from the test of H0: μ = μ0 for various values of μ0 for the raw data and summarized data. The black horizontal line is at 0.05.

The results shown in Figure 23 are discouraging, but are largely driven by the difference in sample size, rather than the difference in the dataset's ability to describe the temperature of the ocean. The sample size, n, of the raw data is 57,536, and the summarized data contains only 61 data values. The test statistic for the hypothesis test of the mean is shown in Equation 6, where s is the sample standard deviation.

$$Z = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Equation 6-Test statistic for the hypothesis test of the mean.

Because the sample size appears in the denominator of the test statistic, large values of n will cause the test statistic to be small, resulting in a small p-value and thus the conclusion to reject the null hypothesis. This does not imply that the summarized data is a poor indication of the raw data, because any large sample size will cause a significant difference. Sullivan and Feinn, 2012 recommend also examining the effect size, which they define to be the "magnitude of the difference between groups". This definition is directed at medical education research studies, and should be amended for our situation. Here we are not directly interested in the difference between the raw and the summarized data, but rather, if we had only the summarized data and not the raw, would we make the same conclusions?

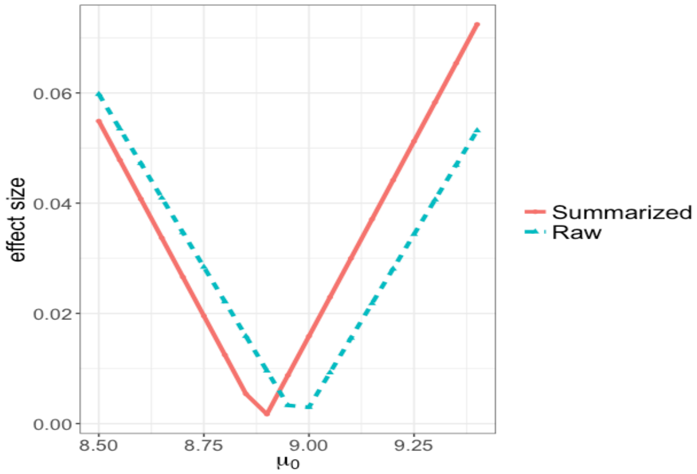


Figure 24- Resulting effect size for various values of μ0 for the raw data and summarized data.

Instead of the p-value, the effect size is examined in Figure 24, for the same sequence of possible hypothesized means. The effect size is computed as:

$$\frac{|\bar{X} - \mu_0|}{s}$$

The sharp elbow of both curves occurs at the sample mean for each, which is 8.887 and 8.976 for the summarized and raw data, respectively. This plot indicates that the effect size is similar for the raw and summarized values, and that the summarized data, with only 0.1% of the size, is able to describe the mean behavior of the data.

9. Conclusions and Future Work

In conclusion, we have developed and presented a novel approach to quantifying the size-precision trade-off for comparing partitioning schemes. This approach was demonstrated on three example datasets of output from large simulation codes. Small subset examples of simulations of the universe, the ocean, and an asteroid, were used to illustrate the procedure. Although each example demonstrated partitioning of three-dimensional spatial variables, the pAIC can be used to compare partitioning in any number and type of dimensions.

Further considerations related to the pAIC were also discussed. First was the issue of how to choose the weights on the terms of the pAIC, which will directly affect the choice of partitioning scheme. The recommendation is to examine each plausible (as defined by a Pareto frontier) partitioning scheme's robustness to the choice of the three weights using the approach presented in (Lu, Anderson-Cook, & Robinson, 2011). Effect size was also examined to explore what results would be drawn from the summarized values, and compared to the results from the raw data.

The pAIC is a good start to comparing partitioning schemes in a quantitative manner; however, there are many areas of future work. First, the summarized and raw data conclusions were compared with the single value of effect size. The representative value from each partition was used to compute the effect size of the summarized data. In addition to saving the representative value for each partition, an estimate of the error is also saved; however, incorporating this information into the effect size is not straightforward. One task is to modify the effect size to incorporate an error for each data value. Also, the work presented here tested results for the mean of the distribution. Other possibilities include the variance or different quantiles of interest.

The partitioning schemes described considered in this work and the pAIC only summarize one variable of interest (e.g., water temperature in the ocean dataset). However, the simulations often produce many outputs (e.g., water salinity, displaced density, potential density, kinetic energy, etc.) that are most likely correlated. Therefore, partitioning schemes (and comparison metrics) that consider multiple variables simultaneously are also an area of future work.

Lastly, the partitioning is to be performed in-situ, or while the simulation is running. Producing and comparing multiple schemes in-situ is most likely not tractable, as creating the partitions and computing the pAIC is computationally expensive. The current thought for how to include the pAIC into this framework is to run a small suite of runs of the simulation and use the pAIC to choose an appropriate partitioning scheme off-line (in a manner similar to what was presented here). As the simulation is running, the pAIC could be computed for the chosen partitioning scheme, as a check to ensure the partitioning scheme is still appropriate, in a manner similar to how control charts ensure manufacturing processes are in-line with expected behavior. Work is in progress to speed up the pAIC computation and understand the distribution of the pAIC metric, both of which are necessary to implement this procedure in practice.

Bibliography

Anderson-Cook, C., & Lu, L. (2012). Weighing Your Options. *Quality Progress , 45* (10), 50-52.

Chapman, J. L., Lu, L., & Anderson-Cook, C. M. (2014). Incorporating response variability and estimation uncertainty into Pareto front optimization . *Computers & Industrial Engineering , 76*, 253-267.

Gisler, G. R., Weaver, R. P., Mader, C. L., & Gittings, M. L. (2004). Two-and three-dimensional asteroid impact simulations. *Computing in Science & Engineering , 6* (3), 45-55.

Gisler, G., Weaver, R., & Gittings, M. (2011). Calculations of asteroid impacts into deep and shallow water. *Pure and applied geophysics , 168* (6-7), 1187-1198.

Habib, S., Pope, A., Lukic, Z., Daniel, D., Fasel, P., Desai, N., et al. (2009). *Hybrid petacomputing meets cosmology: The Roadrunner Universe project.* Techinical , Los Alamos National Laboratory.

Lu, L., & Anderson-Cook, C. M. (2012). Rethinking the Optimal Response Surface Design for a First-Order Model with Two-Factor Interactions, When Protecting against Curvature . *Quality Engineering , 24* (3), 404-422.

Lu, L., Anderson-Cook, C. M., & Robinson, T. J. (2011). Optimization of Designed Experiments Based on Multiple Criteria Utilizing a Pareto Frontier . *Technometrics , 53* (4), 353-365.

Nouanesengsy, B., Woodring, J., Patchett, J., Myers, K., & Ahrens, J. (2014). Adr visualization: A generalized framework for ranking large-scale scientific data using analysis-driven refinement. *2014 IEEE 4th Symposium on Large Data Analysis and Visualization (LDAV)* (pp. 43-50). IEEE.

Patchett, J. M., Nouanesengsy , B., Gisler, G., Ahrens, J., & Hagen, H. (2017). In Situ and Post Processing Workflows for Asteroid Ablation Studies . *Eurographics Conference on Visualization (EuroVis) .* Barcelona.

Patchett, J. M., Samsel, F. J., Tsai, K. C., Gisler, G. R., Rogers, D. H., Abram, G. D., et al. (2016). Visualization and Analysis of Threats from Asteroid Ocean Impacts . *Supercomputing.*

Ringler, T., Petersen, M., Higdon, R. L., Jacobsen, D., Jones, P. W., & Maltrud, M. (2013). A multi-resolution approach to global ocean modeling. *Ocean Modeling , 69*, 211-232.

Samsel, F., Rogers, D. H., Patchett, J. M., & Tsai, K. (2017). Employing Color Theory to Visualize Volume-rendered Multivariate Ensembles of Asteroid Impact Simulations. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1126-1134). ACM.

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. *Journal of graduate medical education , 4* (3), 279-282.

United States Geological Survey. (2016, 12 2). *How much water is there on, in, and above the Earth?* Retrieved 7 27, 2017, from https://water.usgs.gov/edu/earthhowmuch.html

Woodring, J., Ahrens, J., Figg, J., Wendelberger, J., Habib, S., & Heitmann, K. (2011). In-situ Sampling of a Large-Scale Particle Simulation for Interactive Visualization and Analysis. In *Computer Graphics Forum* (Vol. 30, pp. 1151-1160). Wiley Online Library.

Woodring, J., Mniszewski, S., Brislawn, C., DeMarle, D., & Ahrens, J. (2011). Revisiting wavelet compression for large-scale climate data using JPEG 2000 and ensuring data precision. *2011 IEEE Symposium on Large Data Analysis and Visualization (LDAV)* (pp. 31-38). IEEE.