Modeling Optimal Compression Settings on Spatiotemporal Climate Datasets

Alex Pinard

April 2022

TODOS: Add images of datasets in training, testing, validation. Fix or redo the classification report csv so accuracy isn't repeated several times.

1 Introduction

Lossy compression is a vital tool for reducing the enourmous datasizes produced by climate simulation codes. In previous work [1], [2], [3], we have determined certain metrics that relate the similarity of a compressed dataset to the original dataset. If a compressed dataset passes these metrics the dataset is considered statistically indistinguishable from the original dataset. If, in addition, the compressed dataset is smallest in storage size among all datasets that pass these metrics, we consider the dataset to be optimally compressed. Here, we shift focus to prediction of compression algorithm settings that will result in an optimally compressed dataset.

2 Data

We restrict our focus to the 47 spatiotemporal variables with daily output frequency from the CESM Large Ensemble Project (LENS) dataset. These variables include: bc_a1_SRF, dst_a1_SRF, dst_a3_SRF, FLNS, FLNSC, FLUT, FSNS, FSNSC, FSNTOA, ICEFRAC, LHFLX, pom_a1_SRF, PRECL, PRECSC, PRECSL, PRECT, PRECTMX, PSL, Q200, Q500, Q850, QBOT, SHFLX, so4_a1_SRF, so4_a2_SRF, so4_a3_SRF, soa_a1_SRF, soa_a2_SRF, T010, T200, T500, T850, TAUX, TAUY, TMQ, TREFHT, TREFHTMN, TREFHTMX, TS, U010, U200, U500, U850, VBOT, WSPDSRFAV, Z050, and Z500. More information about these variables, including their units and long form names, can be found at https://www.cesm.ucar.edu/projects/community-projects/LENS/data-sets.html. From these datasets, we extract the first 730 timesteps of each variable to use for prediction modeling.

3 Modeling

3.1 Features

Using the LDCPY software package [4], we compute single-value features on each of the uncompressed datasets, including mean, variance, north-south contrast variance, first differences (west-east), probability of a point being positive, number of zeroes, range, and median. Each of these values are aggregated over the whole dataset and averaged together, if applicable. In addition, other single-value features including autocorrelation and entropy, percent of unique values, and mode, which will be added when time permits. At this time, feature selection and standardization are not being performed, we are simply using all of the available features in the model. This will be addressed at a later time.

The next step is to introduce spatial features into the fitting and prediction algorithm. The implementation details are not yet clear, one option is to train a convolutional neural network on the entire 192x288 input datasets, or subsample before feeding the images to a neural network.

3.2 Models

We are primarily interested in the ability of our algorithm to predict the optimal compression settings of a previously unseen variable, which may have similar behavior to an existing variable or may exhibit completely different characteristics. For this reason, we split each variables into one of the training, testing, and validation datasets, as seen in table 1. This partitioning of the variable prevents data leakage, and makes prediction a more difficult problem (note that there is intra-variable variability in the optimal compression settings, so even if we train, validate and test on the same variable we would not reach 100% accuracy. Predicting the optimal compression of a subsequent timestep for a previously-seen variable is a related but slightly different problem).

TRAIN	VALIDATE	TEST
bc_a1_SRF	ICEFRAC	dst_a3_SRF
dst_a1_SRF	LHFLX	FSNS
FLNS	PRECT	FSNTOA
FLNSC	Q500	Q850
FLUT	TREFHTMN	TREFHTMX
pom_a1_SRF	TS	Z050
PRECL	U850	U010
PRECSC	WSPDSRFAV	PRECTMX
PRECSL	Z500	
PSL	FSNSC	
Q200		
QBOT		
SHFLX		
so4_a1_SRF		
soa_a1_SRF		
soa_a1_SRF		
T010		
T200		
T500		
T5850		
TAUX		
TAUY		
TMQ		
TREFHT		
U200		
U500		
VBOT		

Table 1: Assignment of each daily CESM-LENS2 variable to training, validation, and testing groups. The validation variables are selected so that there is similarity between what some of the variables in the training dataset represent, and some which represent entirely different types of variable. This is likewise done for the testing dataset, with some of the variables conceptually similar to those in the training dataset and some which are unlike anything in the training dataset.

Once the data is partitioned, we evaluate the performance of standard statistical models on the training, validation and testing data. Table 2 Lists each of the models that we used. The first column lists the type of model, the second lists the parameters attempted for each model, and the final column lists the best-performing parameters for the model. Table 3 lists the accuracy of each model. Currently, we restrict the modeling to only focus on predicting the optimal compression level or the varaibles for which the compression algorithm ZFP is ideal. Ultimately, we will also need to be able to discriminate between datasets where BG and ZFP are optimal, but at this initial stage we only test the model on the simpler problem of just predicting the optimal level for one algorithm.

The majority of these models are widely used and discussion of them can be found in a standard statistical text, for example, [5]. Here I will note relevant attributes of each model that explain their performance on the data, but some details will be glossed over.

For each statistical model, we include a classification matrix indicating the true optimal compression level (”target class”) and the prediction from the model (”output class”). Datasets which are categorized correctly will fall on the diagonal of this matrix. We also include a classification report. This report lists measures of performance within each class including precision (of the datasets identified as a certain class, the proportion that were correct), recall (of the datasets that are truly a certain class, how many were identified as such), and f1-score (the harmonic mean of the precision and recall scores). Also included are the number of datasets that truly fall in the class under

Model Name	Parameter Sweep	Best Parameters
Random Forest	max_depth = [2 5] random state = 0	max_depth = , random state = 0
AdaBoost	n_estimators = 50, learning rate = 0.1	n_estimators = 50, learning rate = 0.1
Neural Network	2 dense 10-node ReLu layers, adam optimizer	2 dense 10-node ReLu layers, adam optimizer
k-Nearest Neighbors	n_neighbors = [1 2 3 4 5 10 20 50]	
SVM	C: [1e-5 1e-4 1e-3 1e-2 1e-1 1, 1e2, 1e3, 1e4	
LDA	n_components = 1	n_components = 1
QDA		
Aggregate		

Table 2: List of statistical models applied to the data, the list of parameters tried and the parameters for the best performing model.

Model Name	Accuracy
Random Forest	0.446
AdaBoost	0.108
Neural Network	0.277
k-Nearest Neighbors	0.0120
SVM	0.446
LDA	0.337
QDA	0.265
Aggregate	0.446

Table 3: Statistical models with their corresponding accuracy using the best model parameters.

the "support" column. Below the list of classes, the accuracy of the model is listed, and two types of averages are presented. The first is the macro average, which weights each class equally, and the second is an average that weights the performance measures according to the support size of each class.

3.2.1 Random Forest

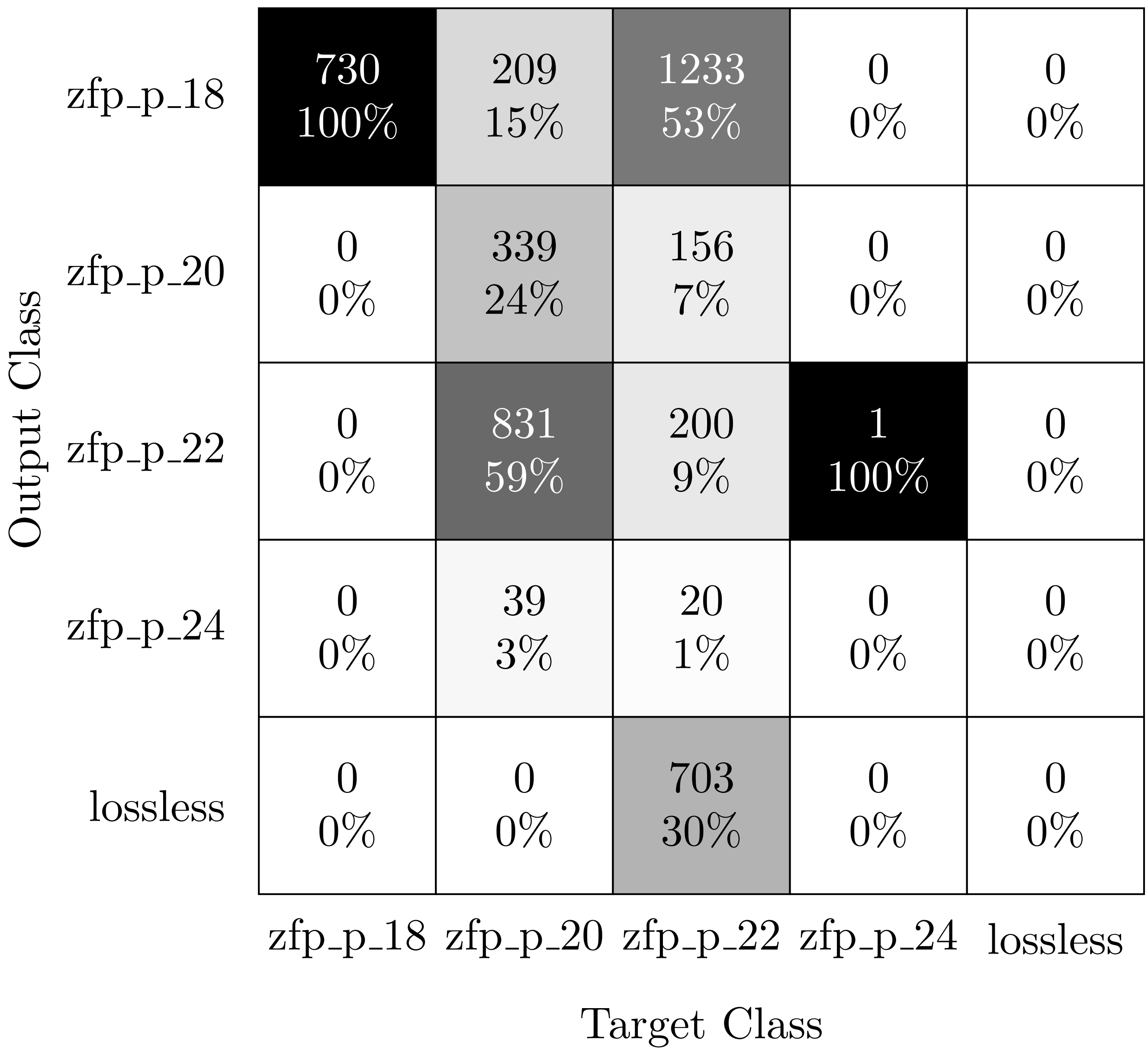


Figure 1: Classification Matrix for the Random tree statistical model.

class	precision	recall	f1-score	support
zfp_p_18	1.000	0.336	0.503	2172.000
zfp_p_20	0.239	0.685	0.354	495.000
zfp_p_22	0.087	0.194	0.120	1031.000
zfp_p_24	0.000	0.000	0.000	59.000
lossless	0.000	0.000	0.000	703.000
accuracy	0.285	0.285	0.285	0.285
macro avg	0.265	0.243	0.195	4460.000
weighted avg	0.534	0.285	0.312	4460.000

Table 4: Evaluation of Random Forest performance. Each class has an associated precision, recall, and f1-score associated with it. Below the classes the accuracy is listed, along with averages of the precision, recall, and f1-score over the entire dataset.

3.2.2 AdaBoost

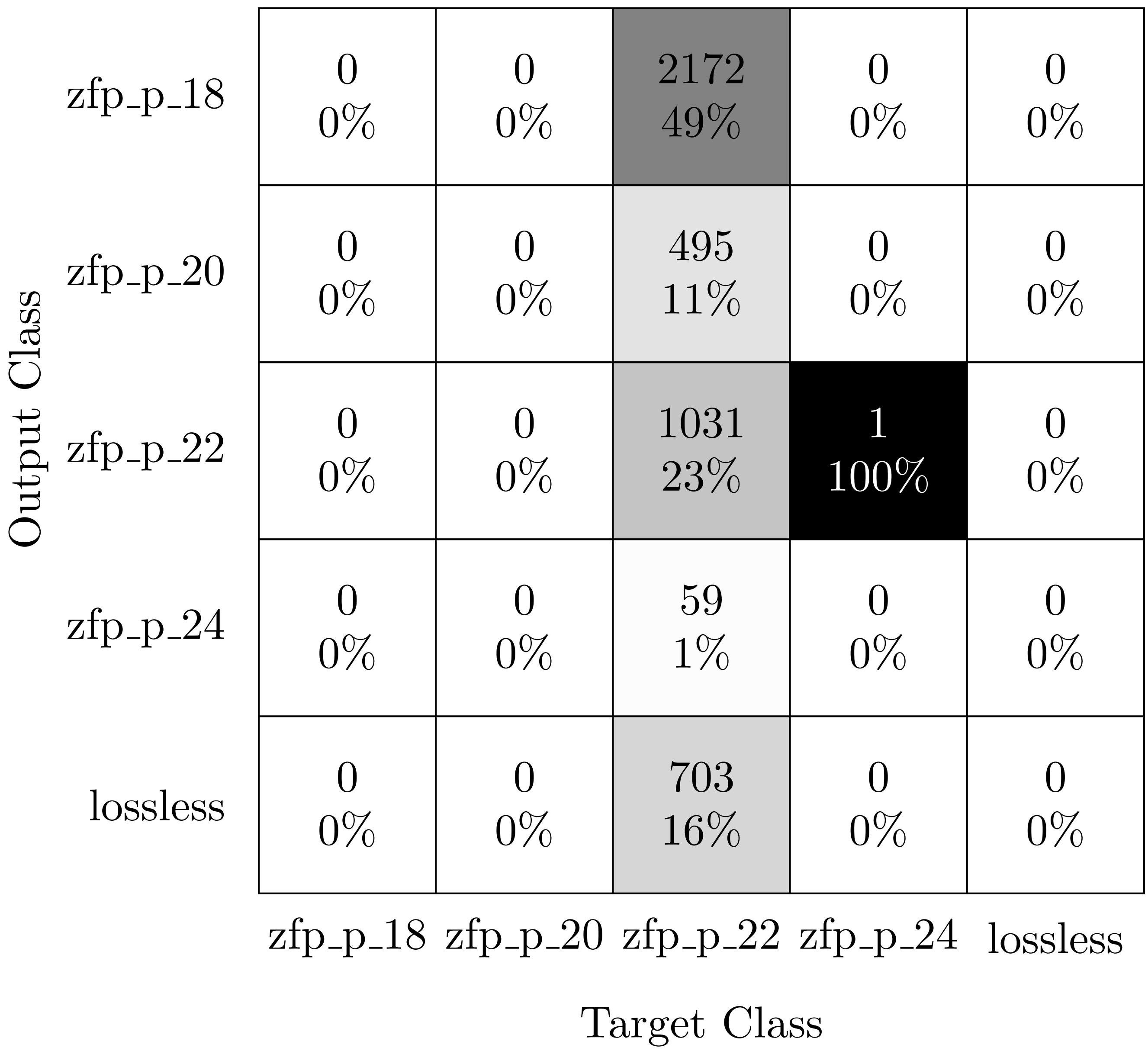


Figure 2: Classification Matrix for the AdaBoost tree statistical model.

class	precision	recall	f1-score	support
zfp_p_18	0.000	0.000	0.000	2172.000
zfp_p_20	0.000	0.000	0.000	495.000
zfp_p_22	0.231	1.000	0.376	1031.000
zfp_p_24	0.000	0.000	0.000	59.000
lossless	0.000	0.000	0.000	703.000
accuracy	0.231	0.231	0.231	0.231
macro avg	0.046	0.200	0.075	4460.000
weighted avg	0.053	0.231	0.087	4460.000

Table 5: Evaluation of AdaBoost performance. Each class has an associated precision, recall, and f1-score associated with it. Below the classes the accuracy is listed, along with averages of the precision, recall, and f1-score over the entire dataset.

3.2.3 Neural Network

Output Class \ Target Class	zfp_p_18	zfp_p_20	zfp_p_22	zfp_p_24	lossless
zfp_p_18	0 0%	938 33%	1233 75%	0 0%	0 0%
zfp_p_20	0 0%	339 12%	156 10%	0 0%	0 0%
zfp_p_22	0 0%	767 27%	243 15%	1 100%	0 0%
zfp_p_24	0 0%	59 2%	0 0%	0 0%	0 0%
lossless	0 0%	700 25%	3 0%	0 0%	0 0%

Figure 3: Classification Matrix for the Neural Network statistical model.

class	precision	recall	f1-score	support
zfp_p_18	0.948	0.623	0.752	2172.000
zfp_p_20	0.157	0.685	0.255	495.000
zfp_p_22	0.948	0.176	0.296	1031.000
zfp_p_24	0.000	0.000	0.000	59.000
lossless	0.000	0.000	0.000	703.000
accuracy	0.420	0.420	0.420	0.420
macro avg	0.410	0.297	0.261	4460.000
weighted avg	0.698	0.420	0.463	4460.000

Table 6: Evaluation of Neural Network performance. Each class has an associated precision, recall, and f1-score associated with it. Below the classes the accuracy is listed, along with averages of the precision, recall, and f1-score over the entire dataset.

3.2.4 k-Nearest Neighbors

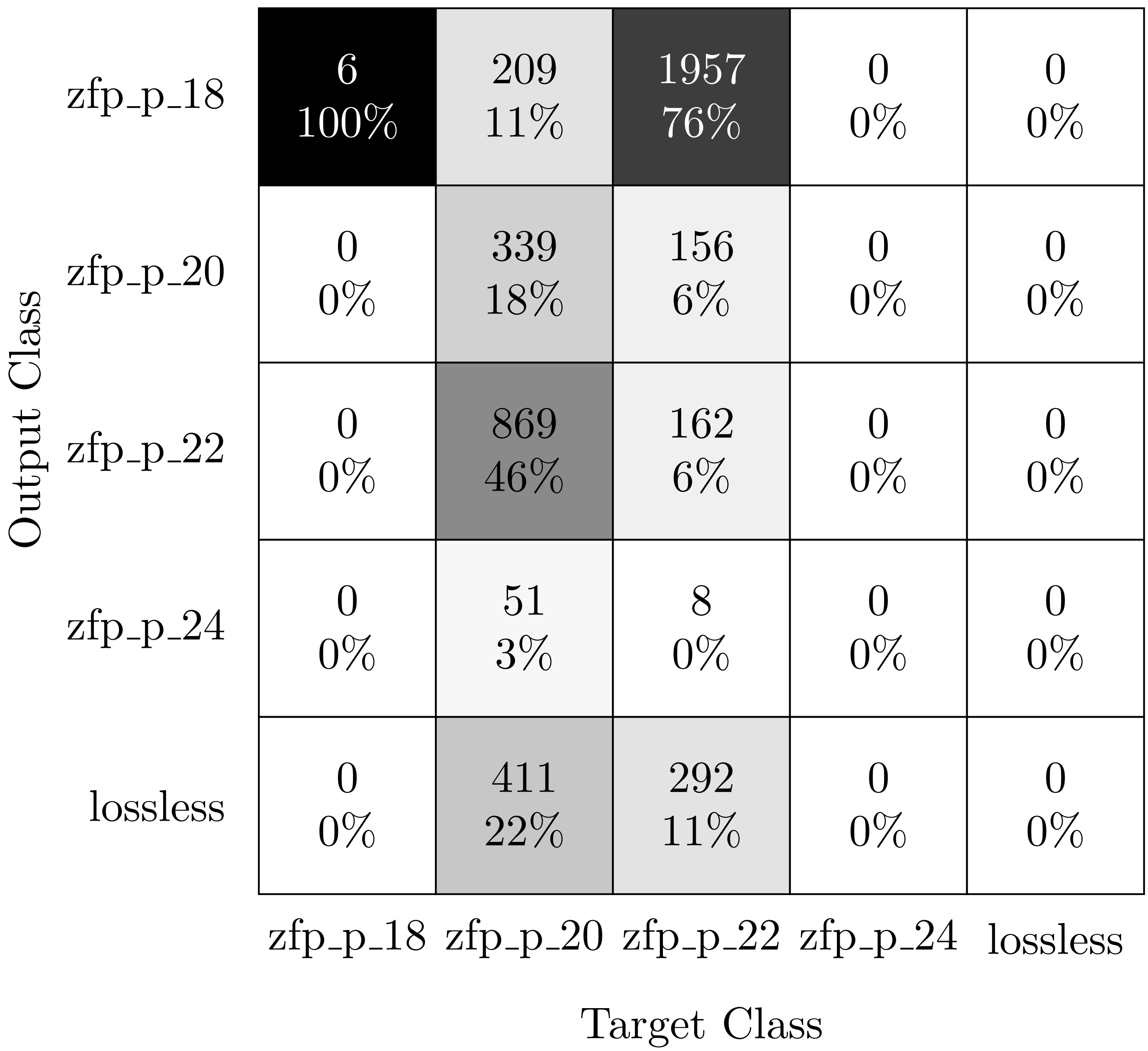


Figure 4: Classification Matrix for the k-Nearest Neighbors statistical model.

	precision	recall	f1-score	support
zfp_p_18	1.000	0.003	0.006	2172.000
zfp_p_20	0.180	0.685	0.286	495.000
zfp_p_22	0.063	0.157	0.090	1031.000
zfp_p_24	0.000	0.000	0.000	59.000
lossless	0.000	0.000	0.000	703.000
accuracy	0.114	0.114	0.114	0.114
macro avg	0.249	0.169	0.076	4460.000
weighted avg	0.522	0.114	0.055	4460.000

Table 7: Evaluation of KNN performance. Each class has an associated precision, recall, and f1-score associated with it. Below the classes the accuracy is listed, along with averages of the precision, recall, and f1-score over the entire dataset.

3.2.5 SVM

The SVM runtime is prohibitively long on these large datasets, for now the SVM results are omitted and not included as part of the aggregate model.

3.2.6 LDA

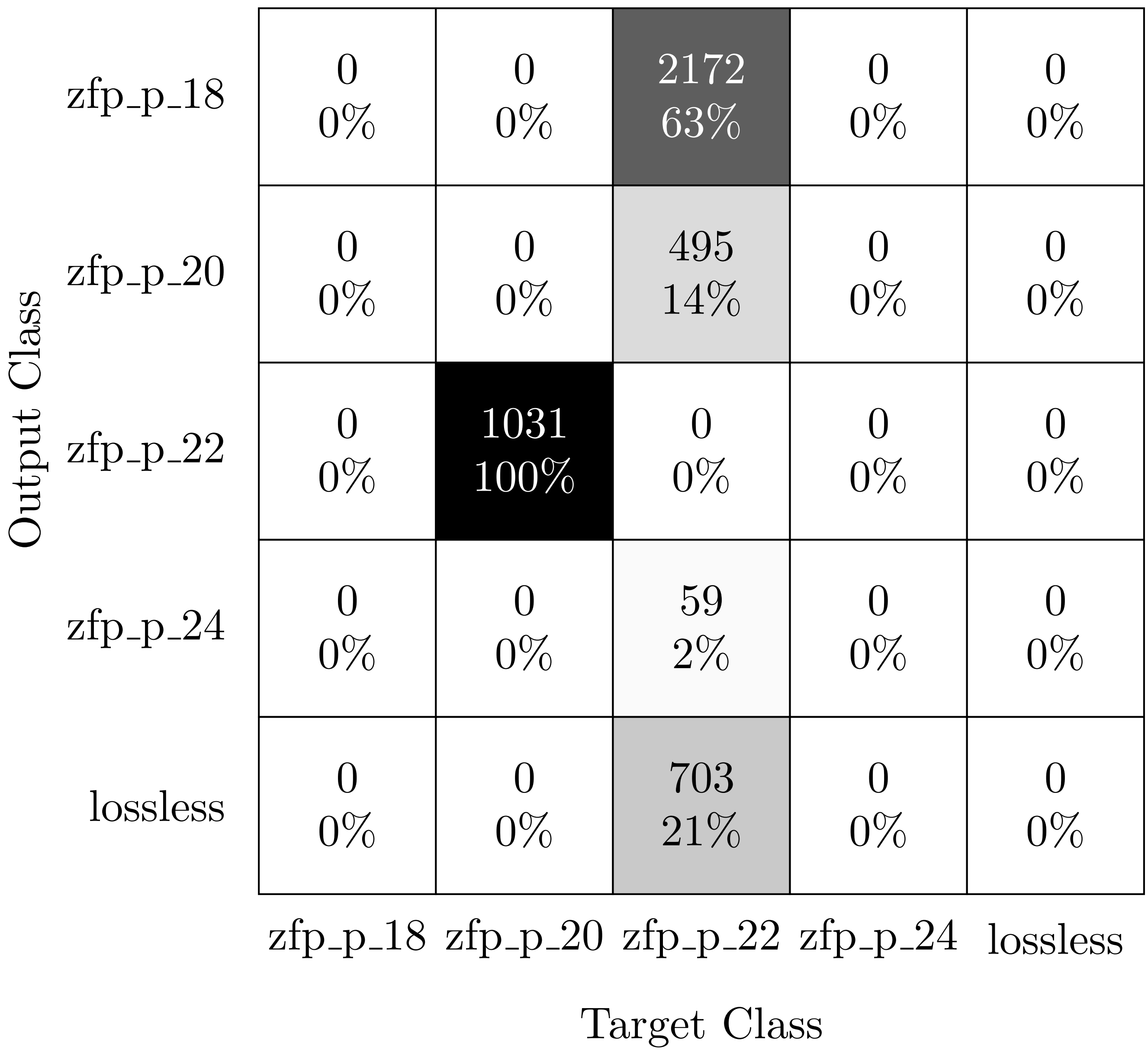


Figure 5: Classification Matrix for the LDA statistical model.

class	precision	recall	f1-score	support
zfp_p_18	0.000	0.000	0.000	2172.000
zfp_p_20	0.000	0.000	0.000	495.000
zfp_p_22	0.231	1.000	0.376	1031.000
zfp_p_24	0.000	0.000	0.000	59.000
lossless	0.000	0.000	0.000	703.000
accuracy	0.231	0.231	0.231	0.231
macro avg	0.046	0.200	0.075	4460.000
weighted avg	0.053	0.231	0.087	4460.000

Table 8: Evaluation of LDA performance. Each class has an associated precision, recall, and f1-score associated with it. Below the classes the accuracy is listed, along with averages of the precision, recall, and f1-score over the entire dataset.

3.2.7 QDA

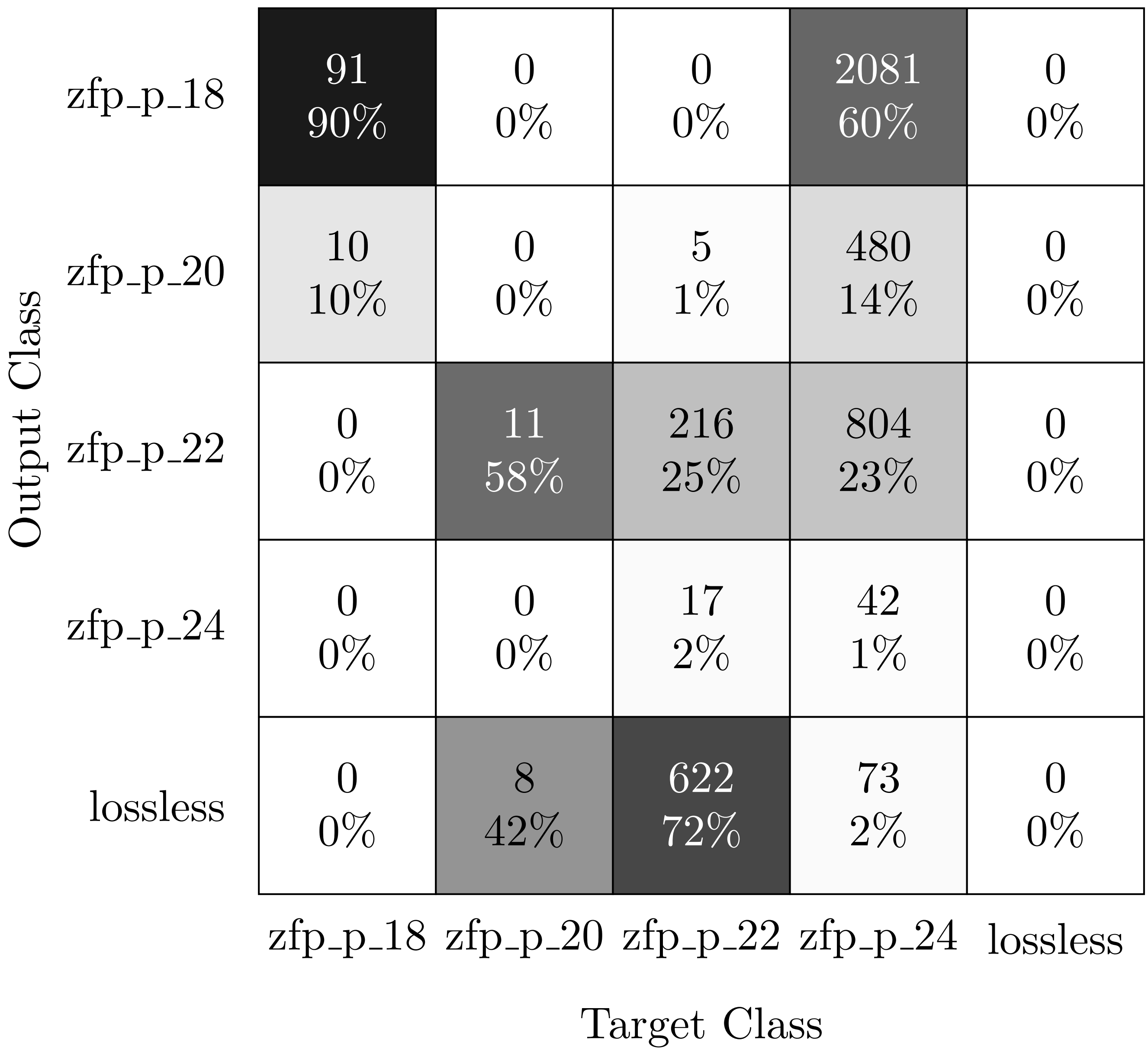


Figure 6: Classification Matrix for the QDA statistical model.

class	precision	recall	f1-score	support
zfp_p_18	0.901	0.042	0.080	2172.000
zfp_p_20	0.000	0.000	0.000	495.000
zfp_p_22	0.251	0.210	0.228	1031.000
zfp_p_24	0.012	0.712	0.024	59.000
lossless	0.000	0.000	0.000	703.000
accuracy	0.078	0.078	0.078	0.078
macro avg	0.233	0.193	0.066	4460.000
weighted avg	0.497	0.078	0.092	4460.000

Table 9: Evaluation of QDA performance. Each class has an associated precision, recall, and f1-score associated with it. Below the classes the accuracy is listed, along with averages of the precision, recall, and f1-score over the entire dataset.

3.2.8 Aggregate Model

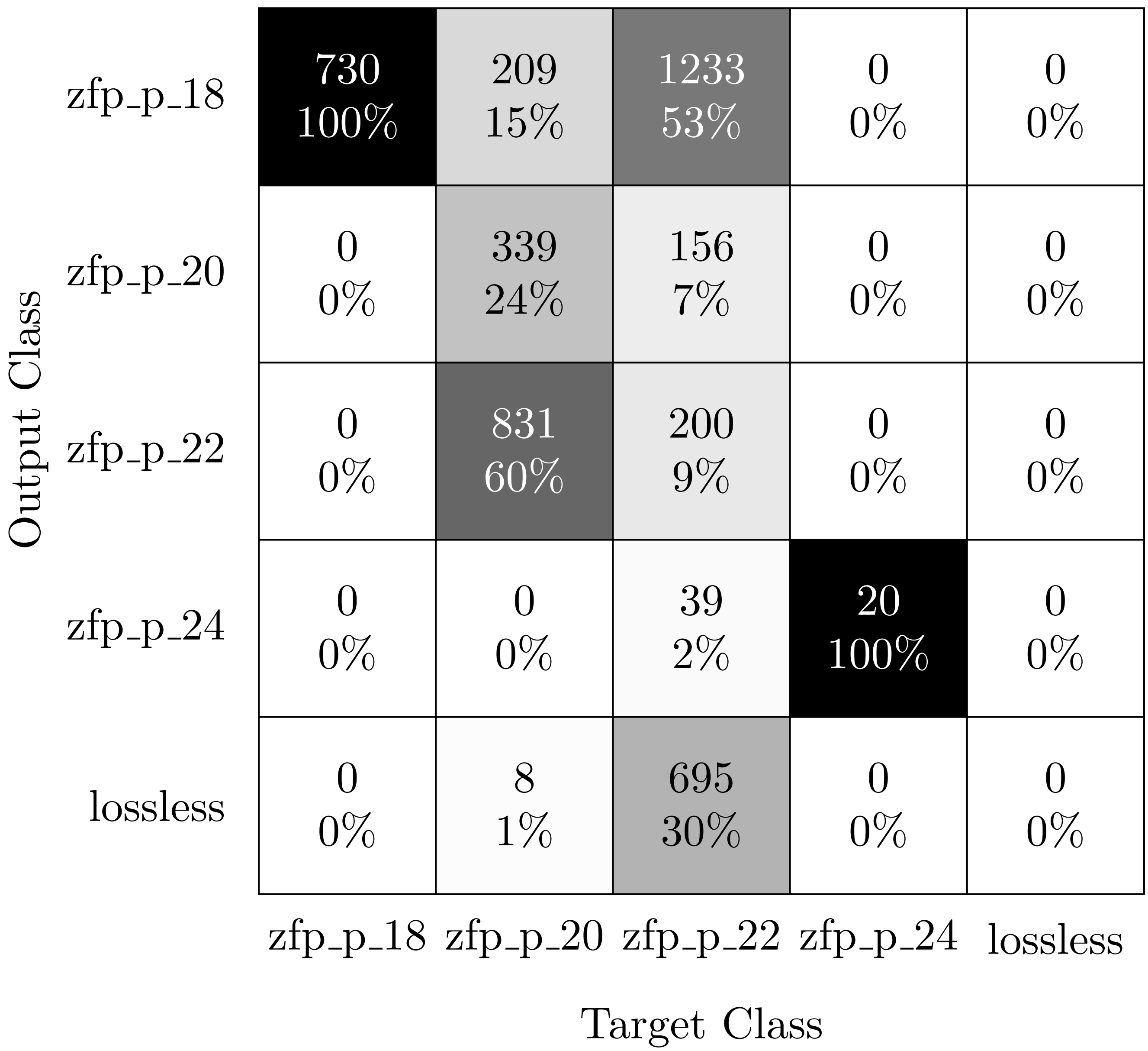


Figure 7: Classification Matrix for the aggregate model.

class	precision	recall	f1-score	support
zfp_p_18	1.000	0.336	0.503	2172.000
zfp_p_20	0.238	0.685	0.353	495.000
zfp_p_22	0.087	0.194	0.120	1031.000
zfp_p_24	0.000	0.000	0.000	59.000
lossless	0.000	0.000	0.000	703.000
accuracy	0.285	0.285	0.285	0.285
macro avg	0.265	0.243	0.195	4460.000
weighted avg	0.533	0.285	0.312	4460.000

Table 10: Evaluation of aggregate model performance. Each class has an associated precision, recall, and f1-score associated with it. Below the classes the accuracy is listed, along with averages of the precision, recall, and f1-score over the entire dataset.

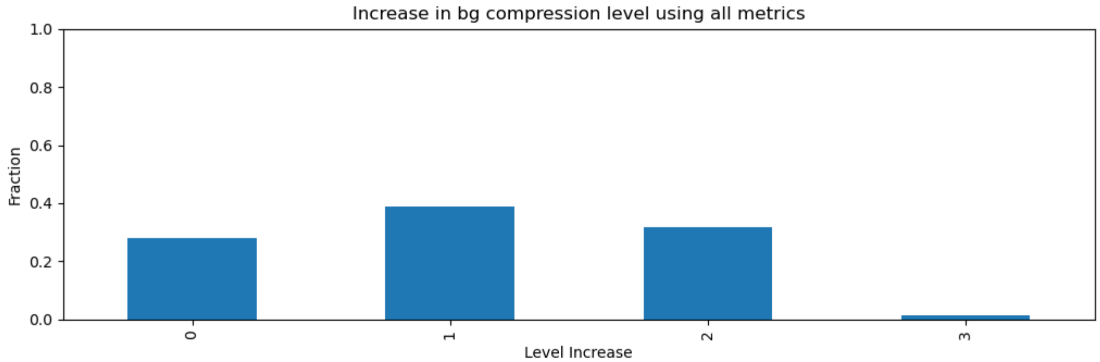


Figure 8: Fraction of all time slices where the compression level (e.g. bg_2 to bg_3 would count as a one-level increase) increased by the given amount when using all metrics instead of just the DSSIM. Note that bg_7 to lossless also counts as a single level increase.

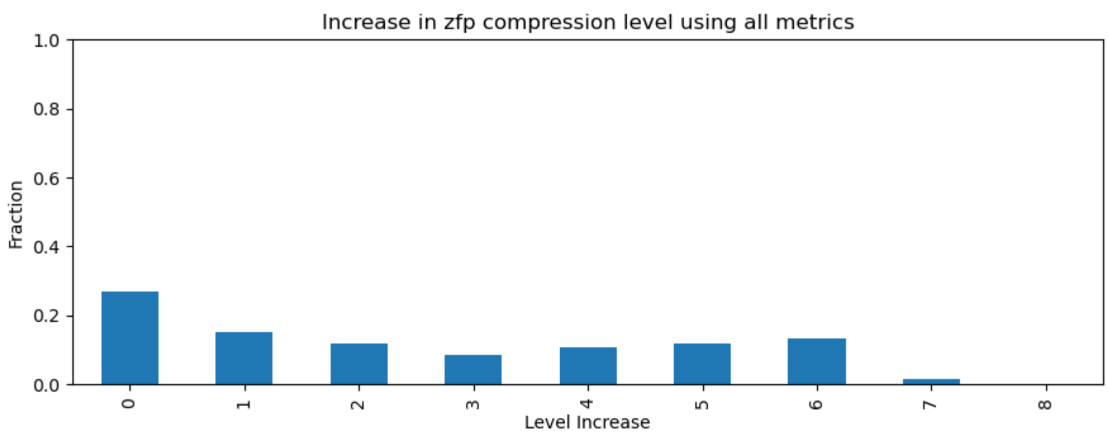


Figure 9: Fraction of all time slices where the compression level (e.g. zfp_14 to zfp_16 would count as a one-level increase) increased by the given amount when using all metrics instead of just the DSSIM. Note that zfp_24 to lossless also counts as a single level increase.

4 Using multiple metrics vs. just DSSIM

The initial metric used for comparison is the DSSIM. We introduce other metrics to capture artifacts we have found that the DSSIM is not always able to capture. The below plots show how much the optimal compression level increases when including the other metrics. The thresholds used here are less than 5% of points showing a spatial relative error of at least 0.0001, a Kolmogorov-Smirnov p-value of at least 0.05, a pearson correlation coefficient of at least 0.99999, and a DSSIM of at least 0.9995, and a maximum spatial relative error of less than 0.1.

Shown in Figures 4, 4 are the increase in compression level compared to just using the DSSIM metric. In Figures 4, 4 the fraction of time each metric is considered the "hardest" or "easiest" metric to pass is displayed, and in FIgures 4, 4 are the fraction of time slices where the optimal compression level increased by at least two levels.

5 Alternate Thresholds

This analysis is repeated with alternate thresholds: For the spatial relative error, less than 1% of points showing an error of at least 0.0001, a Kolmogorov-Smirnov p-value of at least 0.01, a pearson correlation coefficient of at least 0.9999, and a DSSIM of at least 0.99995, and a maximum spatial relative error of less than 0.05.

Shown in Figures 5, 5 are the increase in compression level compared to just using the DSSIM metric. In Figures 5, 5 the fraction of time each metric is considered the "hardest" or "easiest" metric to pass is displayed, and in

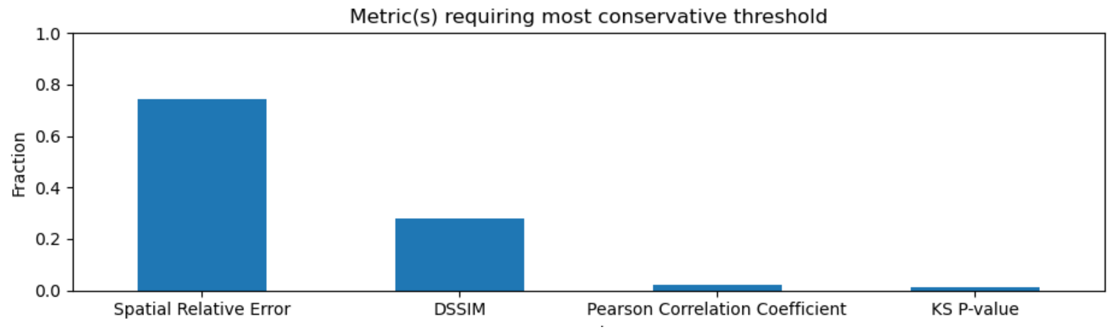


Figure 10: Fraction of all time slices where each metric is the "most selective", requiring the least-compressed compression level. In the case of ties, both metrics are counted (summing the fractions will add to over 100 percent).

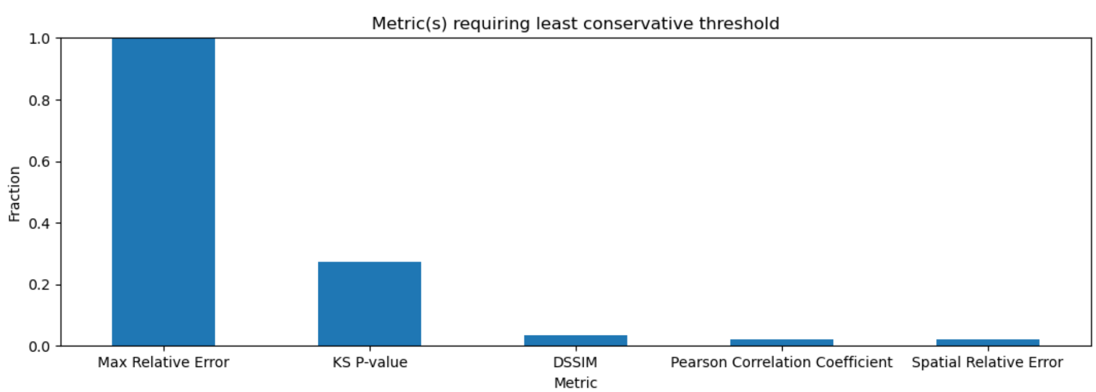


Figure 11: Fraction of all time slices where each metric is the "least selective". In the case of ties, both metrics are countd (summing the fractions will add to over 100 percent).

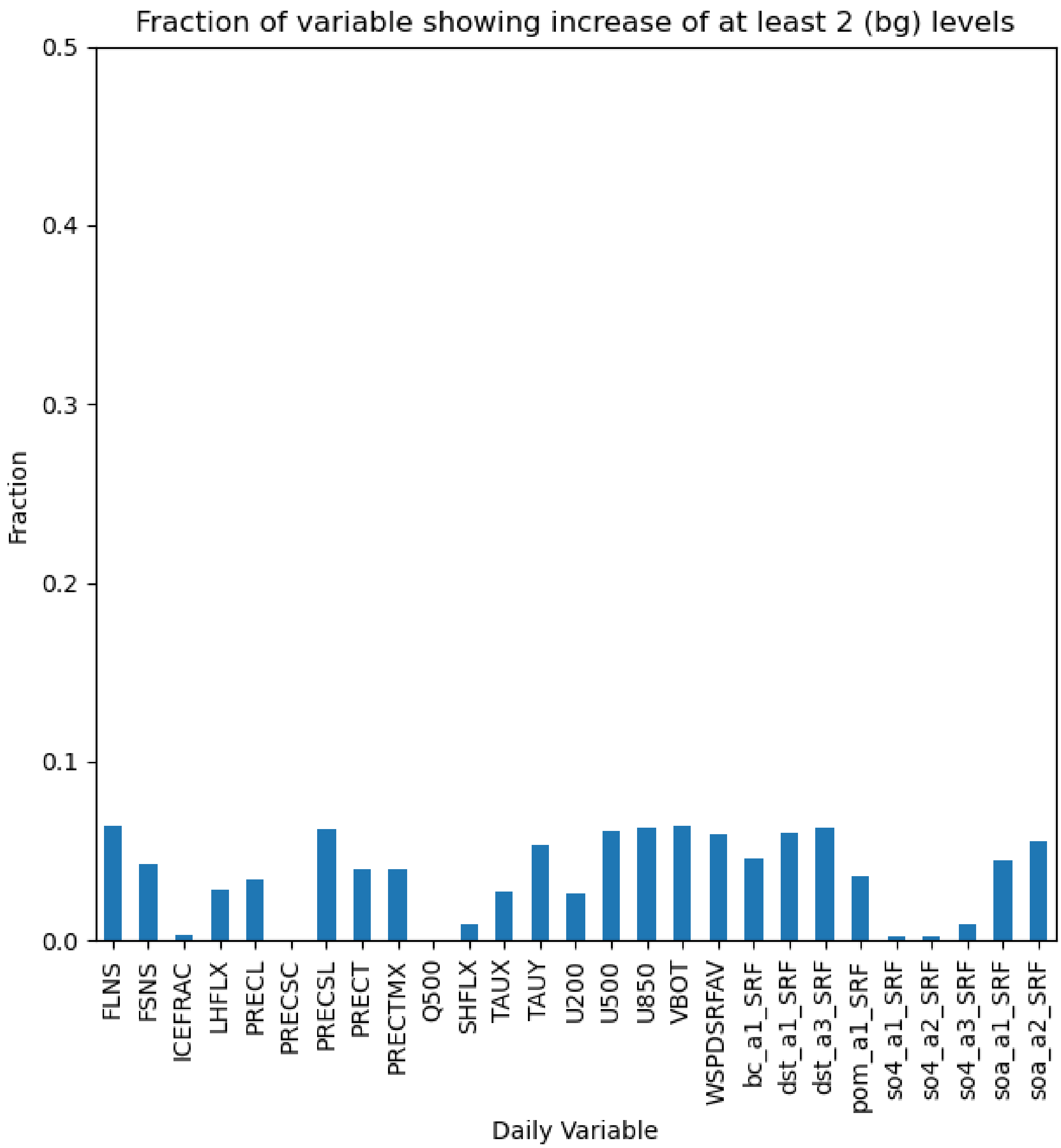


Figure 12: Counts of the number of increases for each daily variable showing an increase of over 2 (bg) levels when using all metrics, compared to just the DSSIM.

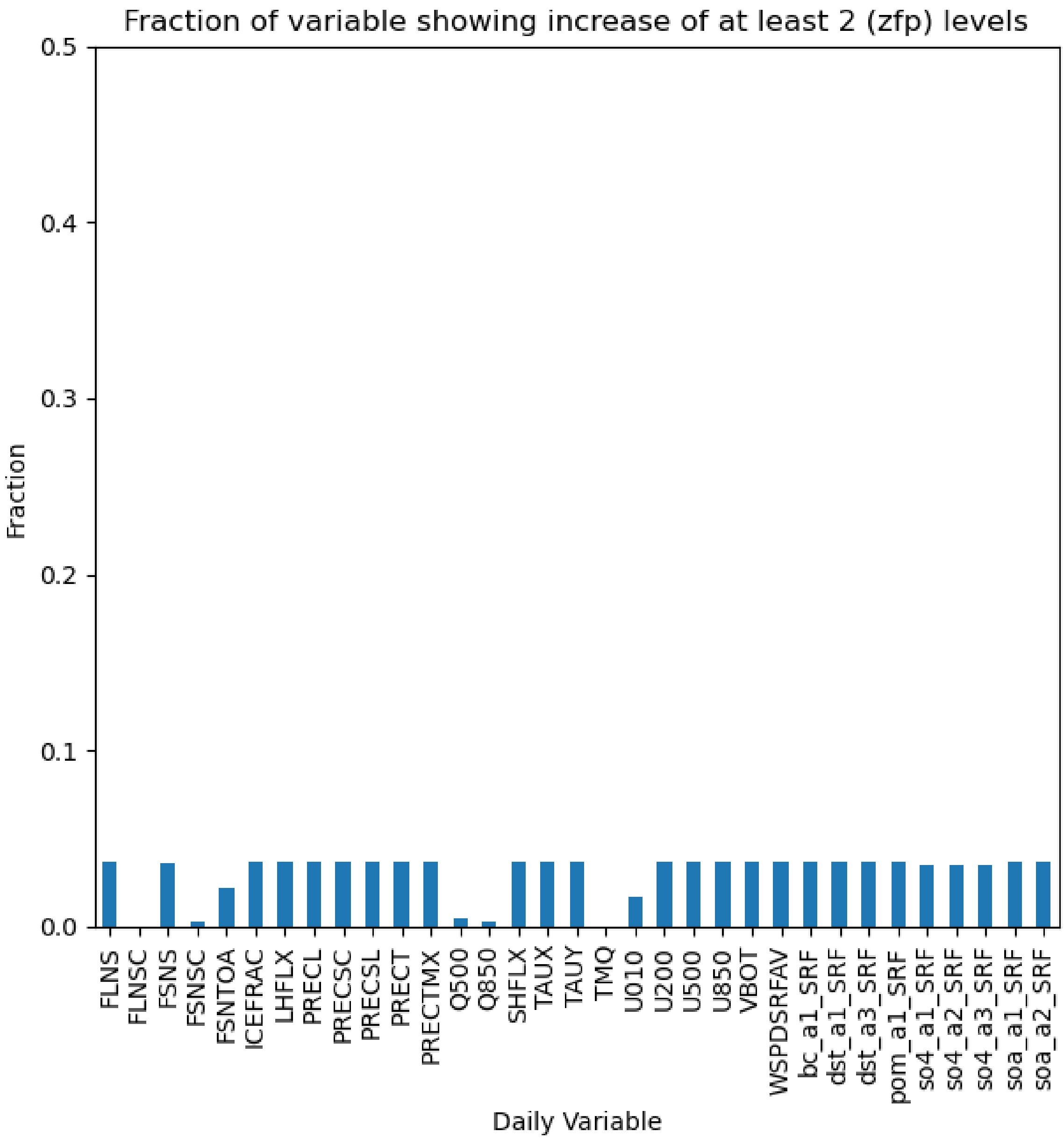


Figure 13: Counts of the number of increases for each daily variable showing an increase of over 2 (zfp) levels when using all metrics, compared to just the DSSIM.

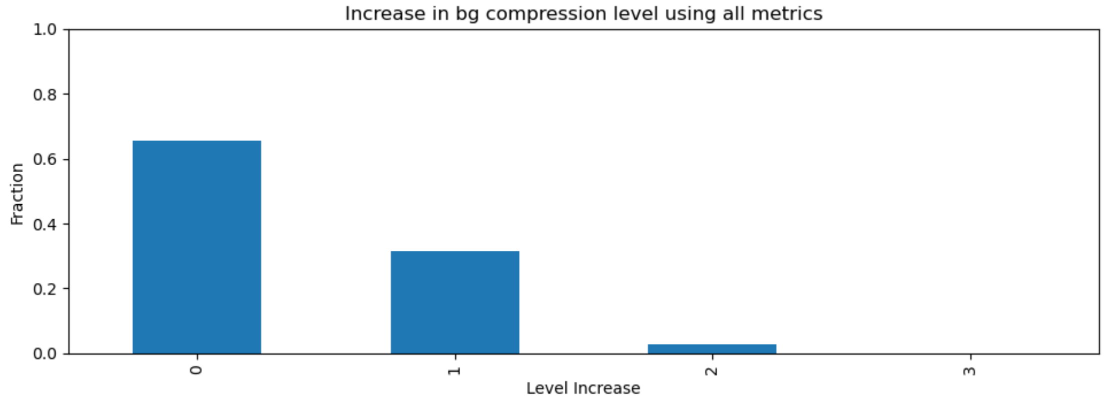


Figure 14: Fraction of all time slices where the compression level (e.g. bg_2 to bg_3 would count as a one-level increase) increased by the given amount when using all metrics instead of just the DSSIM. Note that bg_7 to lossless also counts as a single level increase.

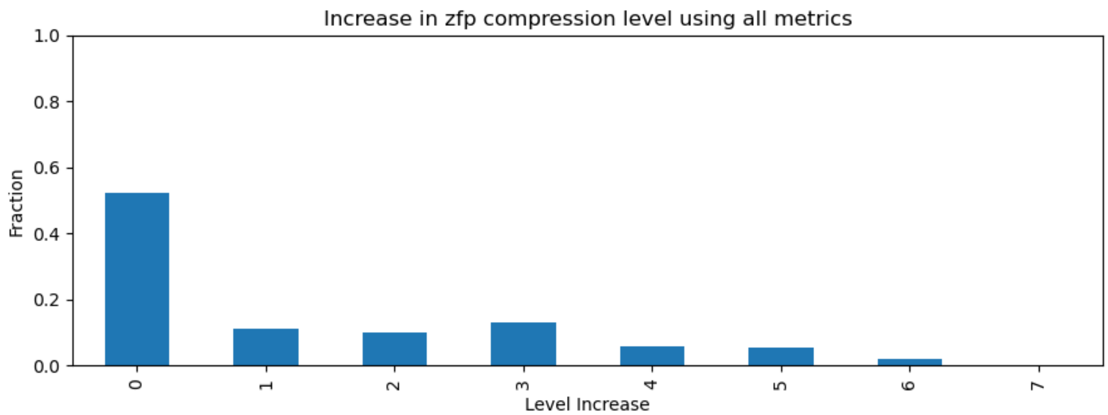


Figure 15: Fraction of all time slices where the compression level (e.g. zfp_14 to zfp_16 would count as a one-level increase) increased by the given amount when using all metrics instead of just the DSSIM. Note that zfp_24 to lossless also counts as a single level increase.

Figures 5, 5 are the fraction of time slices where the optimal compression level increased by at least two levels.

6 Histograms of Optimal Slices

References

[1] A. H. Baker, H. Xu, D. M. Hammerling, S. Li, and J. P. Clyne, "Toward a multi-method approach: Lossy data compression for climate simulation data," in *International Conference on High Performance Computing*. Springer, 2017, pp. 30–42.

[2] A. Pinard, A. H. Baker, and D. M. Hammerling, "A statistical approach to obtaining a data structural similarity index cutoff threshold," National Center for Atmospheric Research, Tech. Rep. NCAR/TN-568+STR, 2021.

[3] ——, "Examining variations in the optimal compression level of spatiotemporal datasets determined using the data structural similarity index measure (dssim)," National Center for Atmospheric Research, Tech. Rep. NCAR/TN-570+STR, 2021.

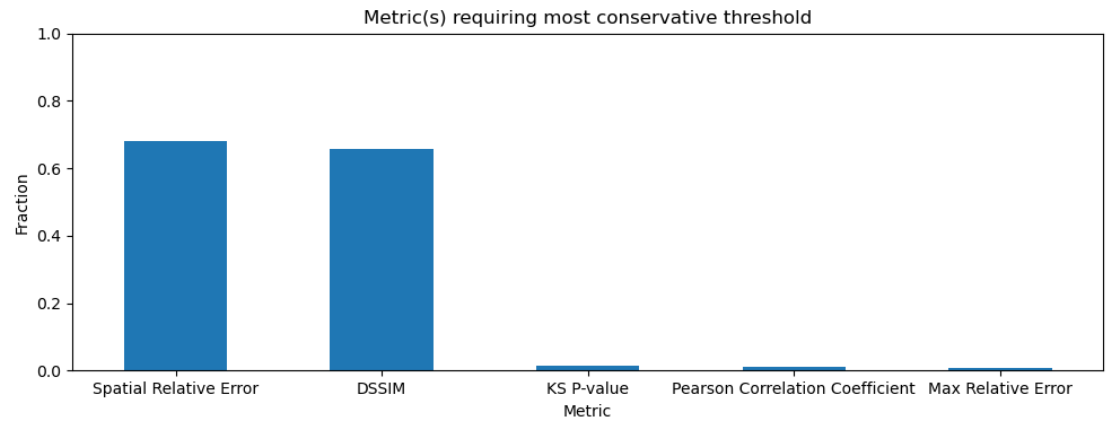


Figure 16: Fraction of all time slices where each metric is the "most selective", requiring the least-compressed compression level. In the case of ties, both metrics are counted (summing the fractions will add to over 100 percent).

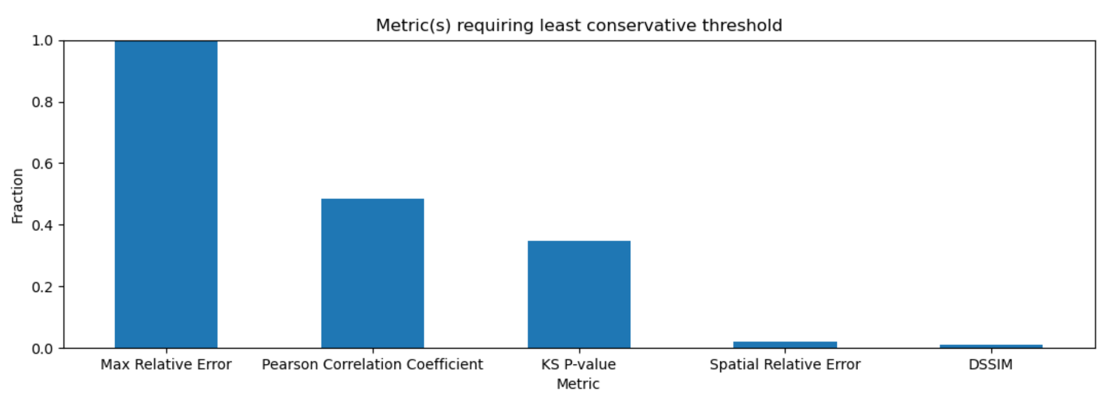


Figure 17: Fraction of all time slices where each metric is the "least selective". In the case of ties, both metrics are countd (summing the fractions will add to over 100 percent).

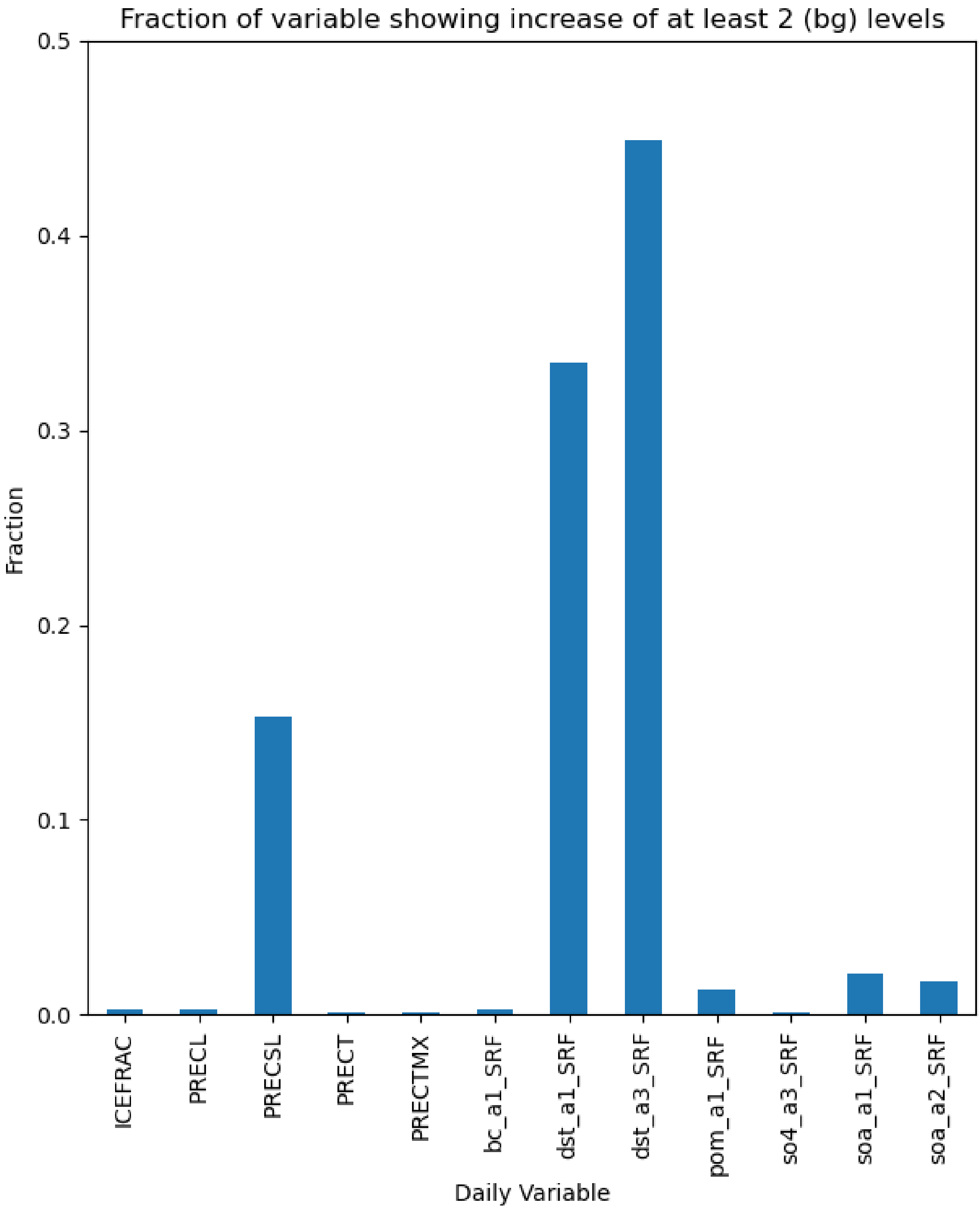


Figure 18: Counts of the number of increases for each daily variable showing an increase of over 2 (bg) levels when using all metrics, compared to just the DSSIM.

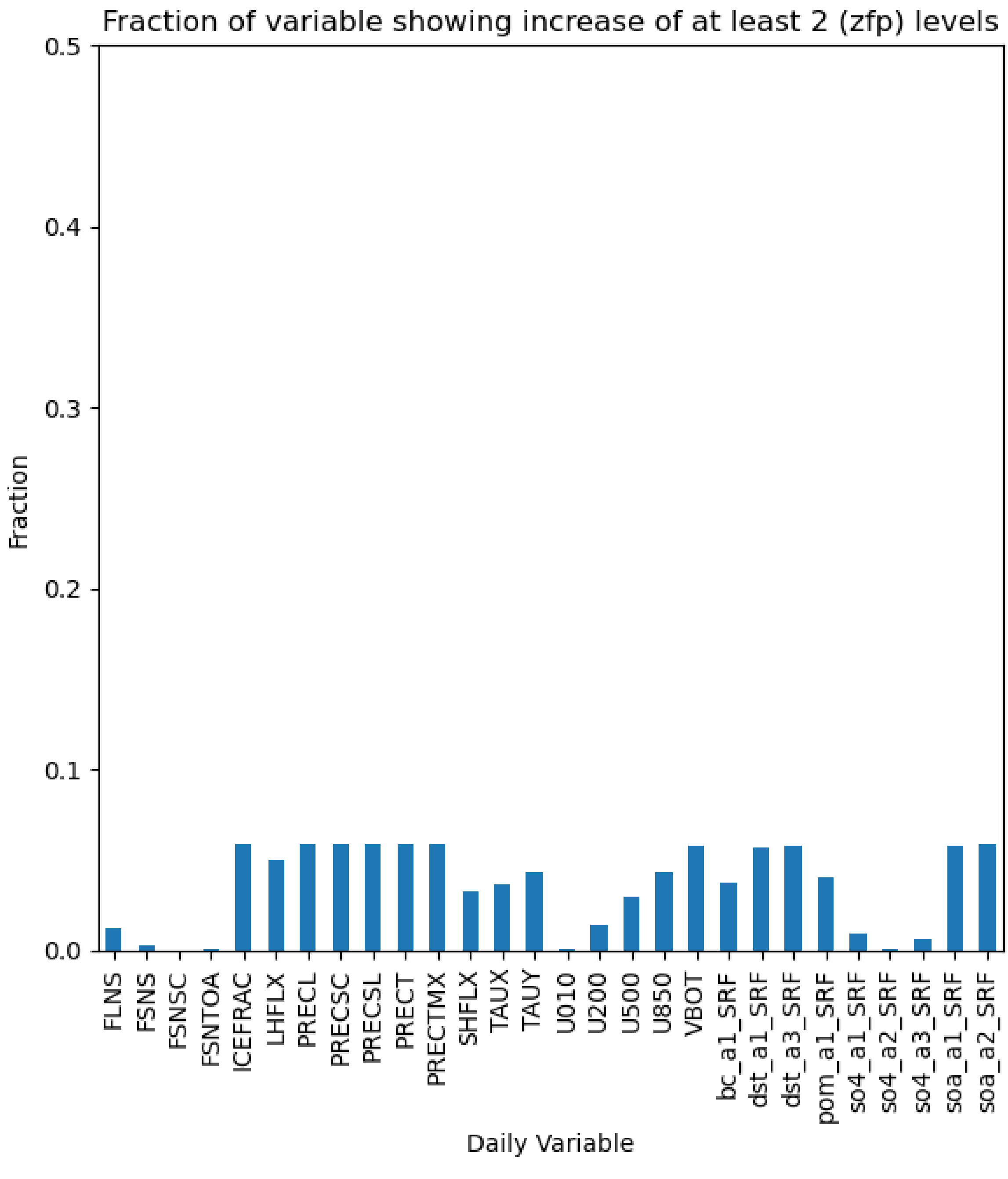
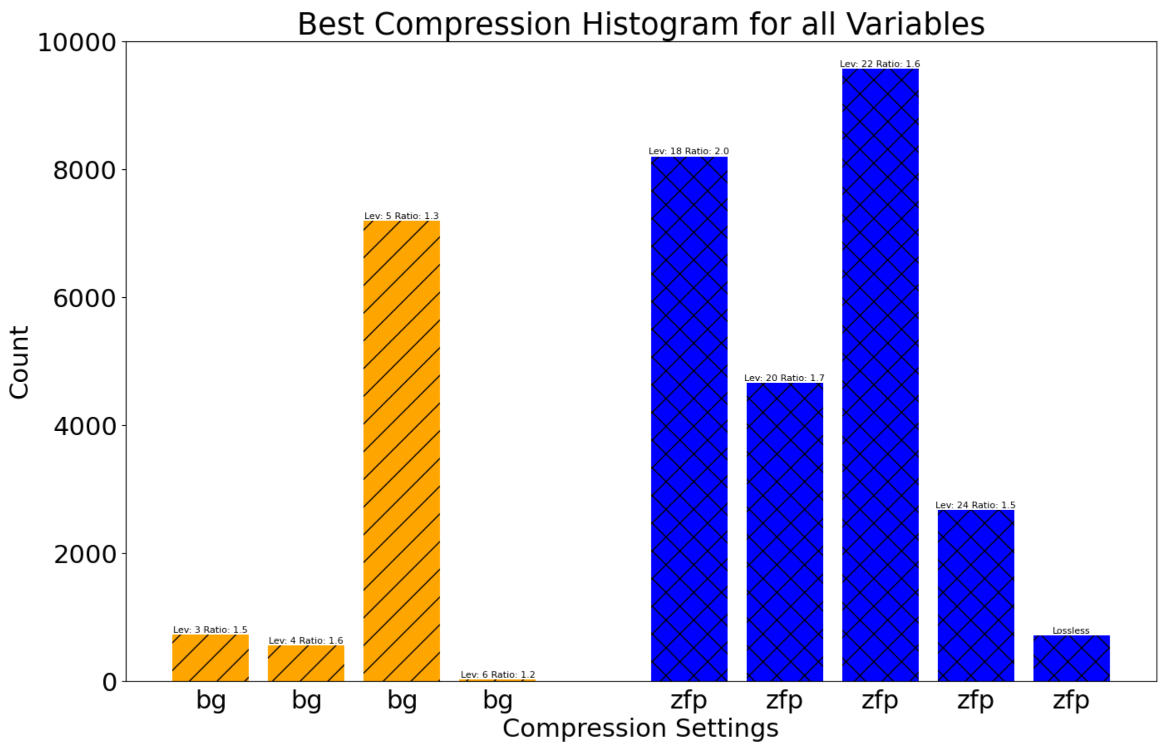
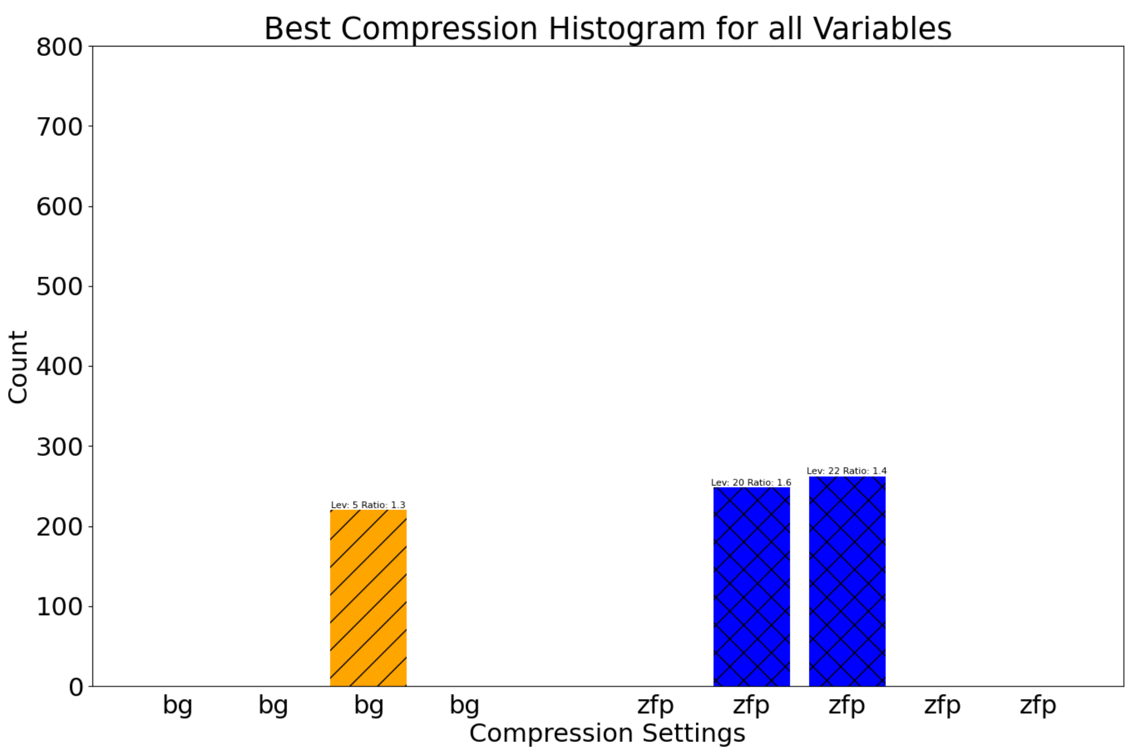


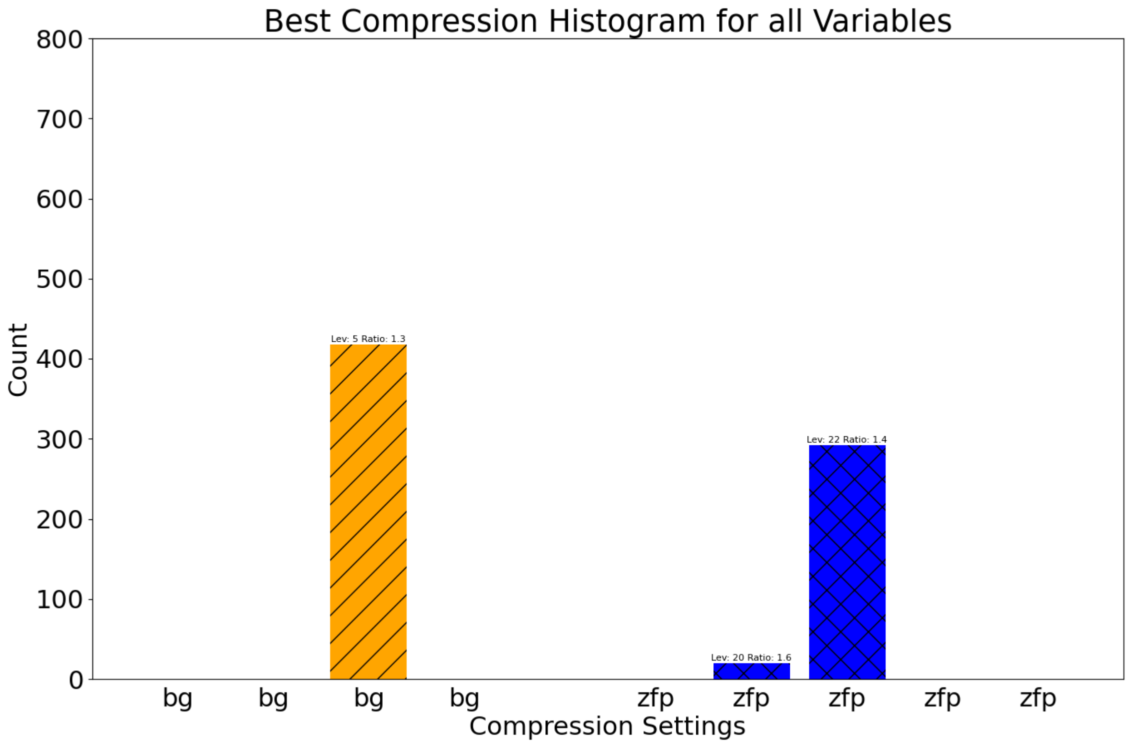
Figure 19: Counts of the number of increases for each daily variable showing an increase of over 2 (zfp) levels when using all metrics, compared to just the DSSIM.

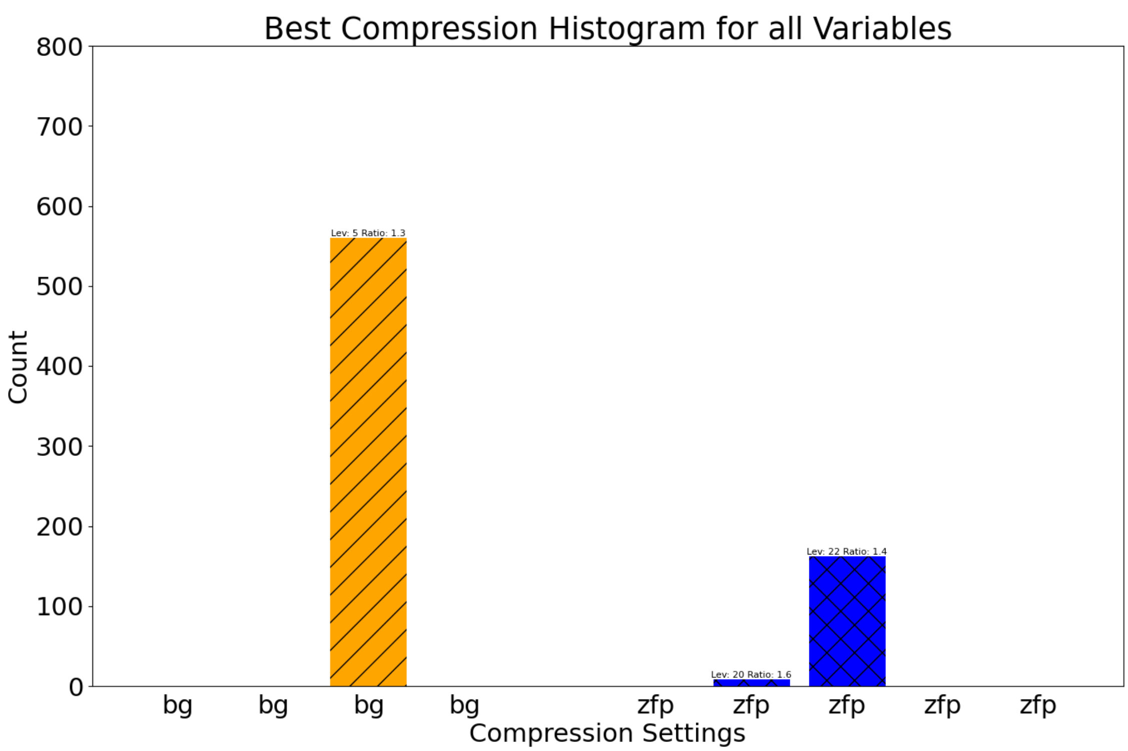


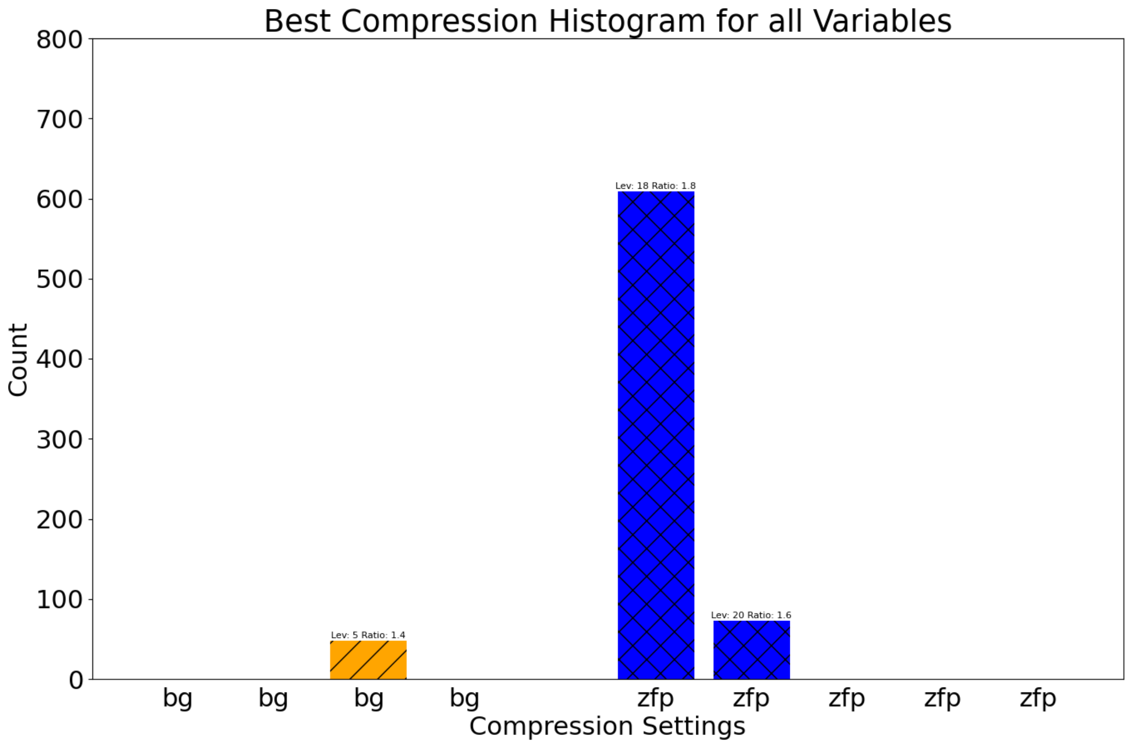
Best Compression Histogram for all Variables
Count
Compression Settings
bg
zfp
Lev: 3 Ratio: 1.5
Lev: 4 Ratio: 1.6
Lev: 5 Ratio: 1.3
Lev: 6 Ratio: 1.2
Lev: 18 Ratio: 2.0
Lev: 20 Ratio: 1.7
Lev: 22 Ratio: 1.6
Lev: 24 Ratio: 1.5
Lossless

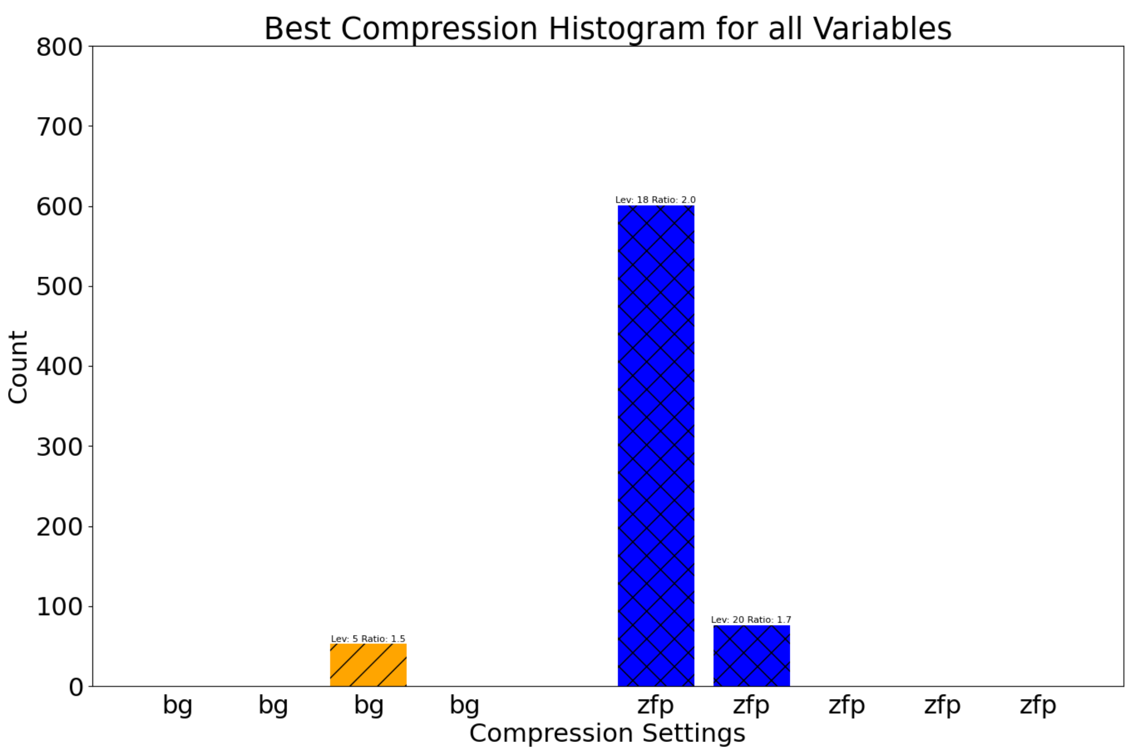


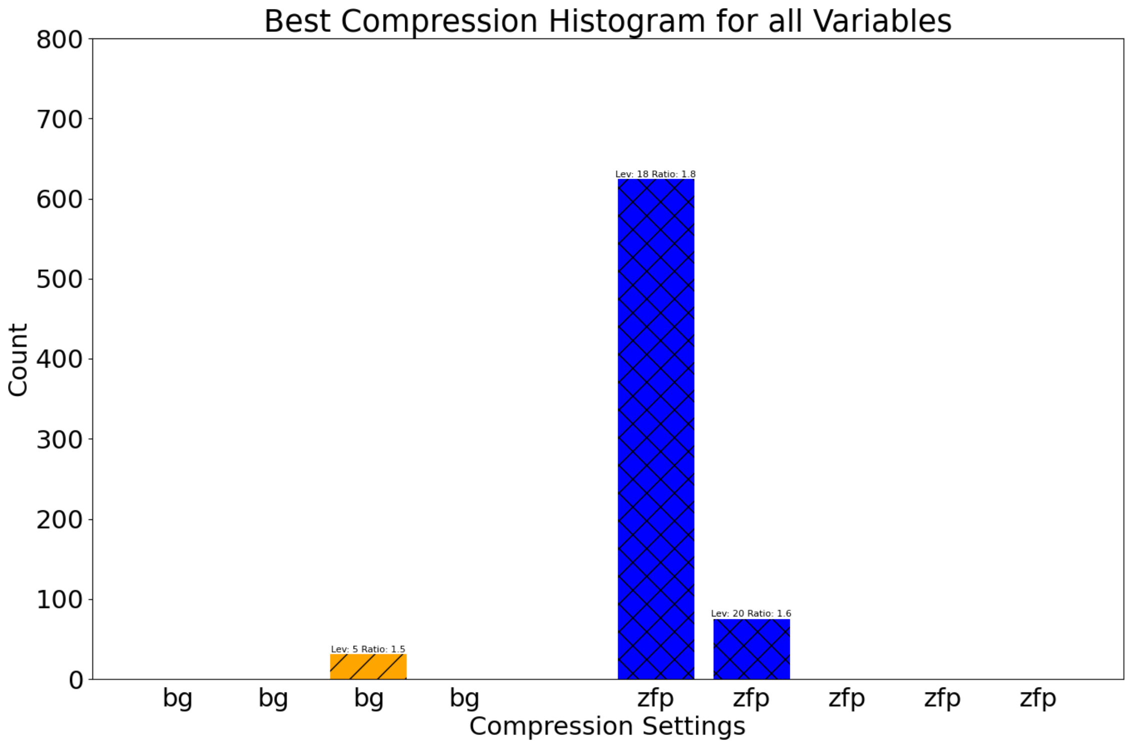
Best Compression Histogram for all Variables
Count
Compression Settings
bg
zfp
Lev: 5 Ratio: 1.3
Lev: 20 Ratio: 1.6
Lev: 22 Ratio: 1.4

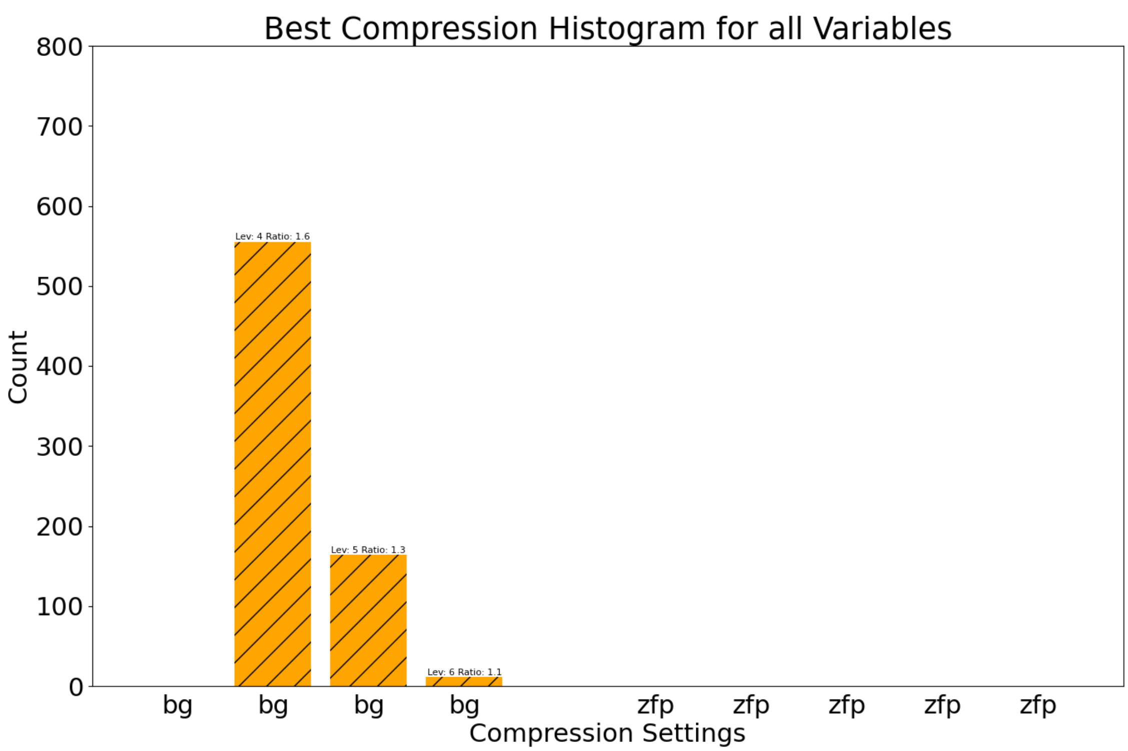

Best Compression Histogram for all Variables
Count
0
100
200
300
400
500
600
700
800
Lev: 5 Ratio: 1.3
Lev: 20 Ratio: 1.6
Lev: 22 Ratio: 1.4
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

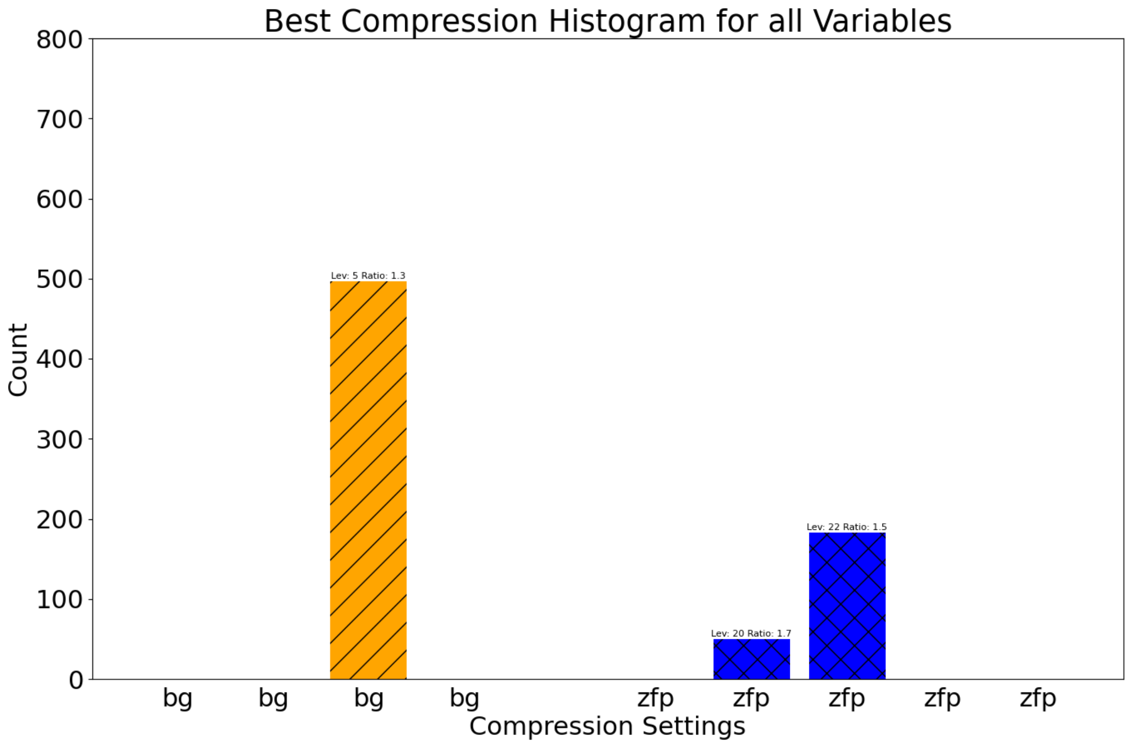

Best Compression Histogram for all Variables
Count
0
100
200
300
400
500
600
700
800
Lev: 5 Ratio: 1.3
Lev: 20 Ratio: 1.6
Lev: 22 Ratio: 1.4
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

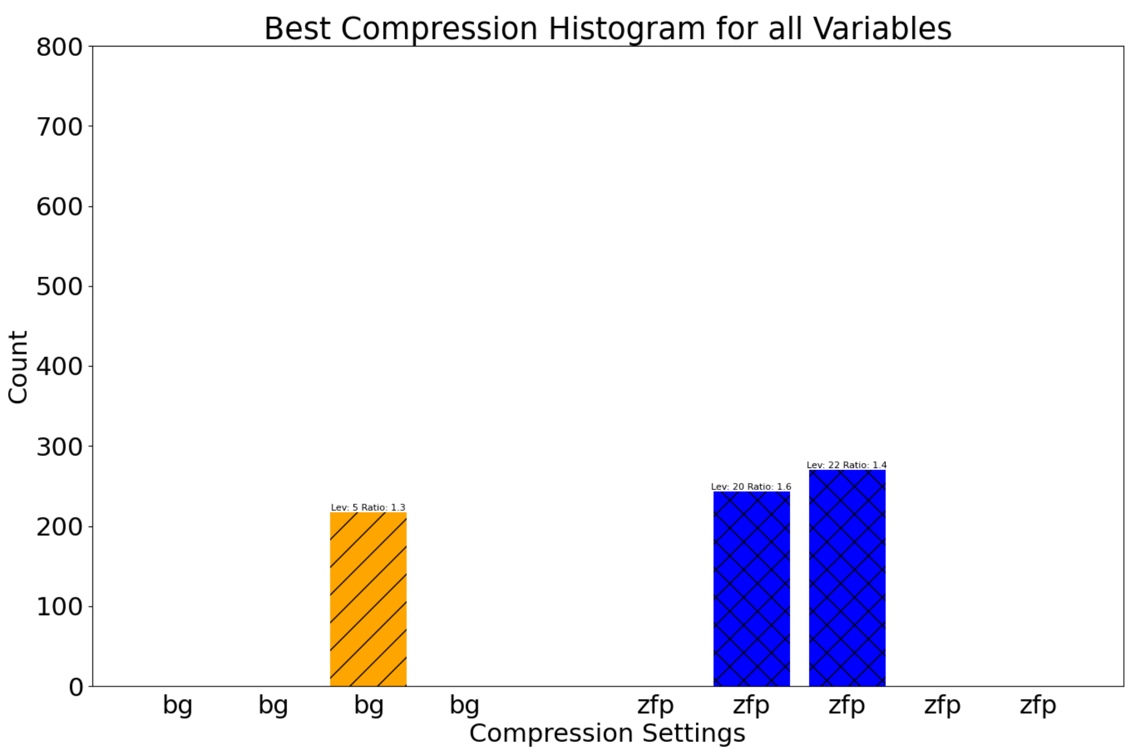

Best Compression Histogram for all Variables
Count
Compression Settings
Lev: 5 Ratio: 1.4
Lev: 18 Ratio: 1.8
Lev: 20 Ratio: 1.6
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp

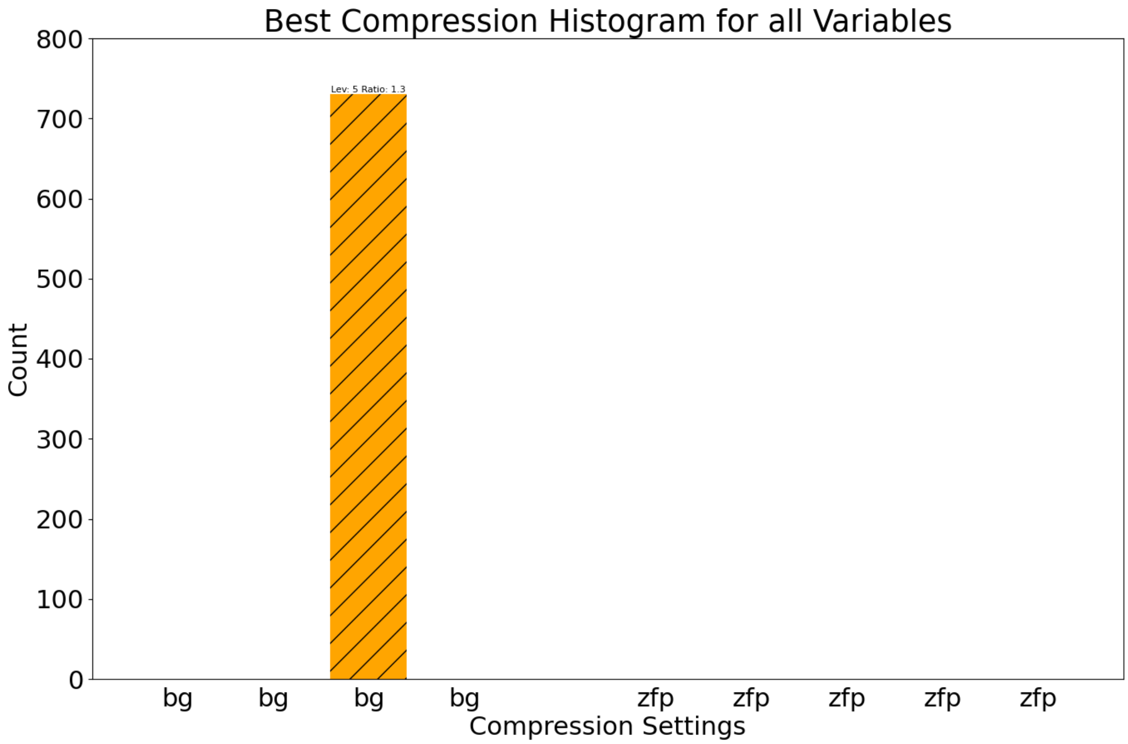

Best Compression Histogram for all Variables
Count
Compression Settings
Lev: 5 Ratio: 1.5
Lev: 18 Ratio: 2.0
Lev: 20 Ratio: 1.7
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp

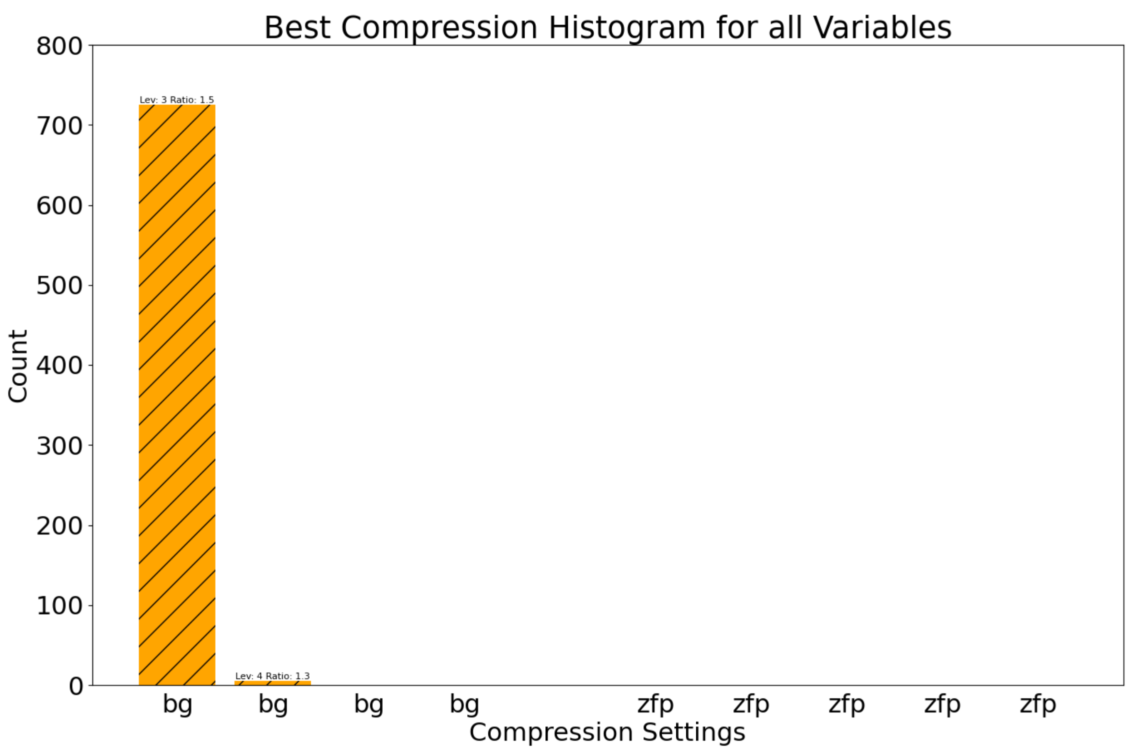

Best Compression Histogram for all Variables
Count
800
700
600
500
400
300
200
100
0
Lev: 5 Ratio: 1.5
Lev: 18 Ratio: 1.8
Lev: 20 Ratio: 1.6
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

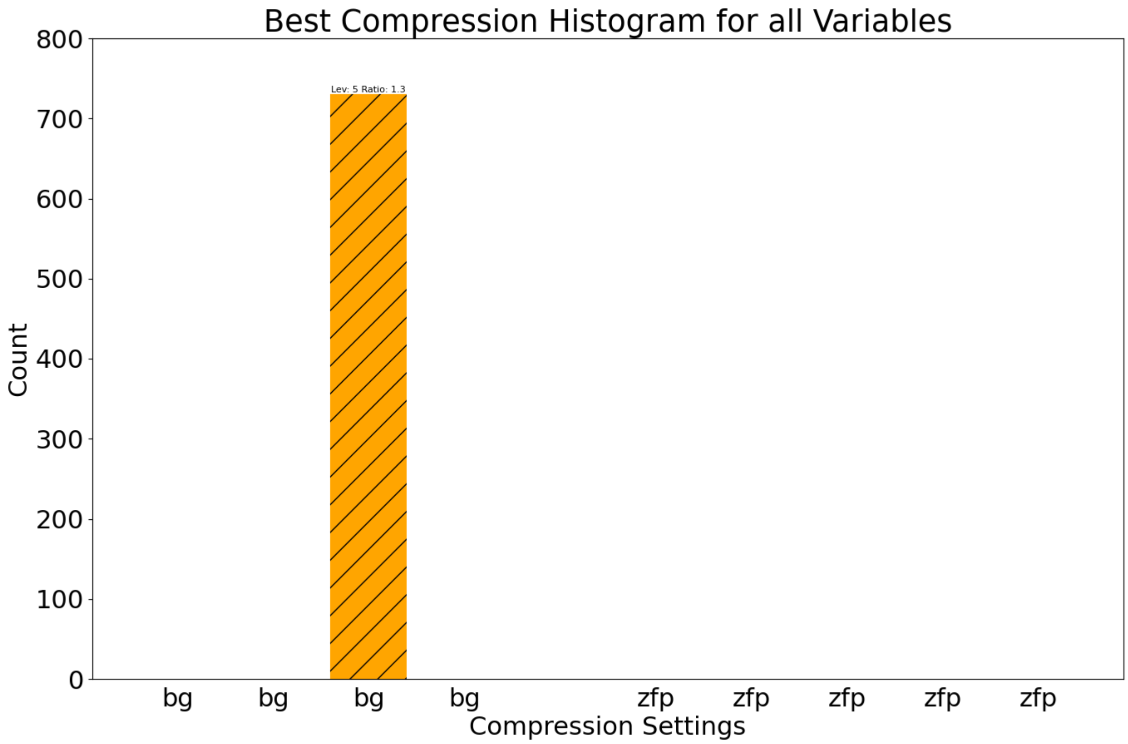

Best Compression Histogram for all Variables
Count
800
700
600
500
400
300
200
100
0
Lev: 4 Ratio: 1.6
Lev: 5 Ratio: 1.3
Lev: 6 Ratio: 1.1
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

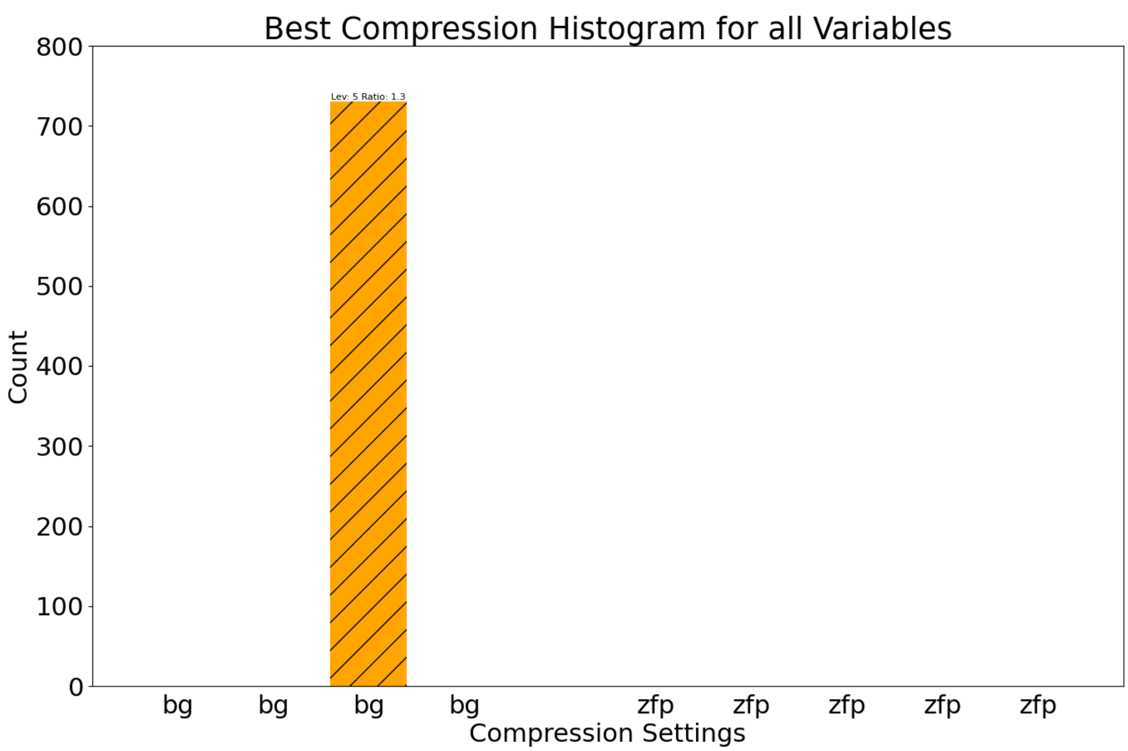

Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 5 Ratio: 1.3
Lev: 20 Ratio: 1.7
Lev: 22 Ratio: 1.5
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

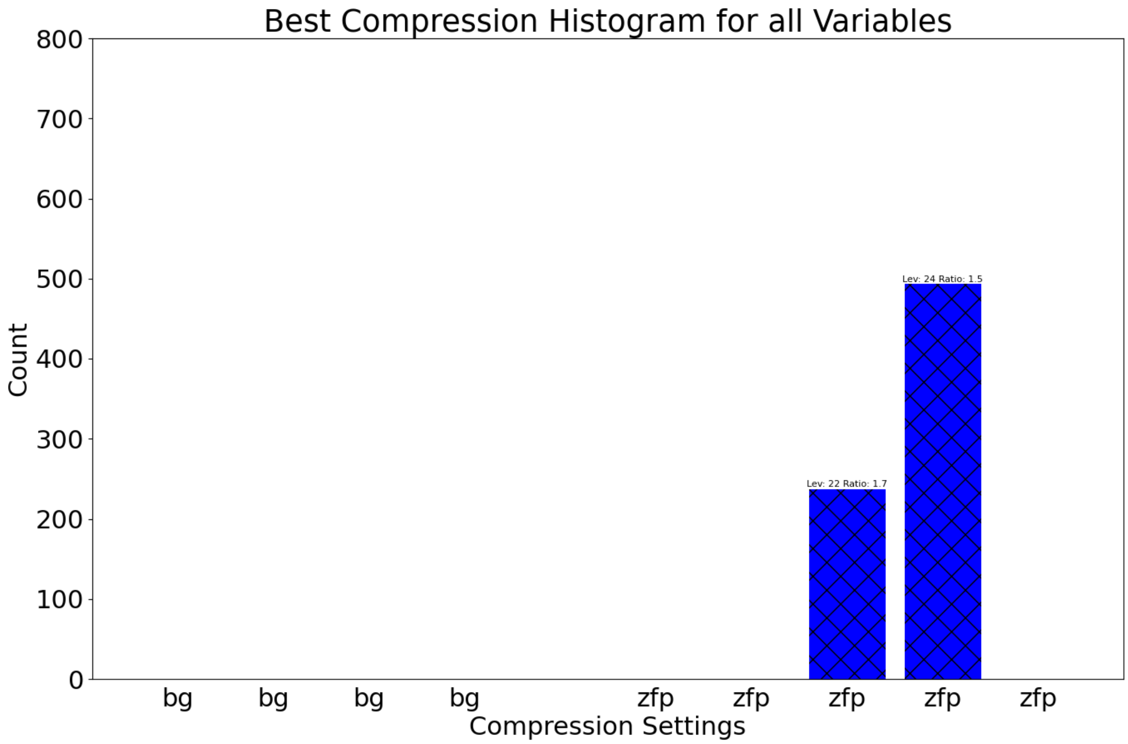

Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 5 Ratio: 1.3
Lev: 20 Ratio: 1.6
Lev: 22 Ratio: 1.4
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

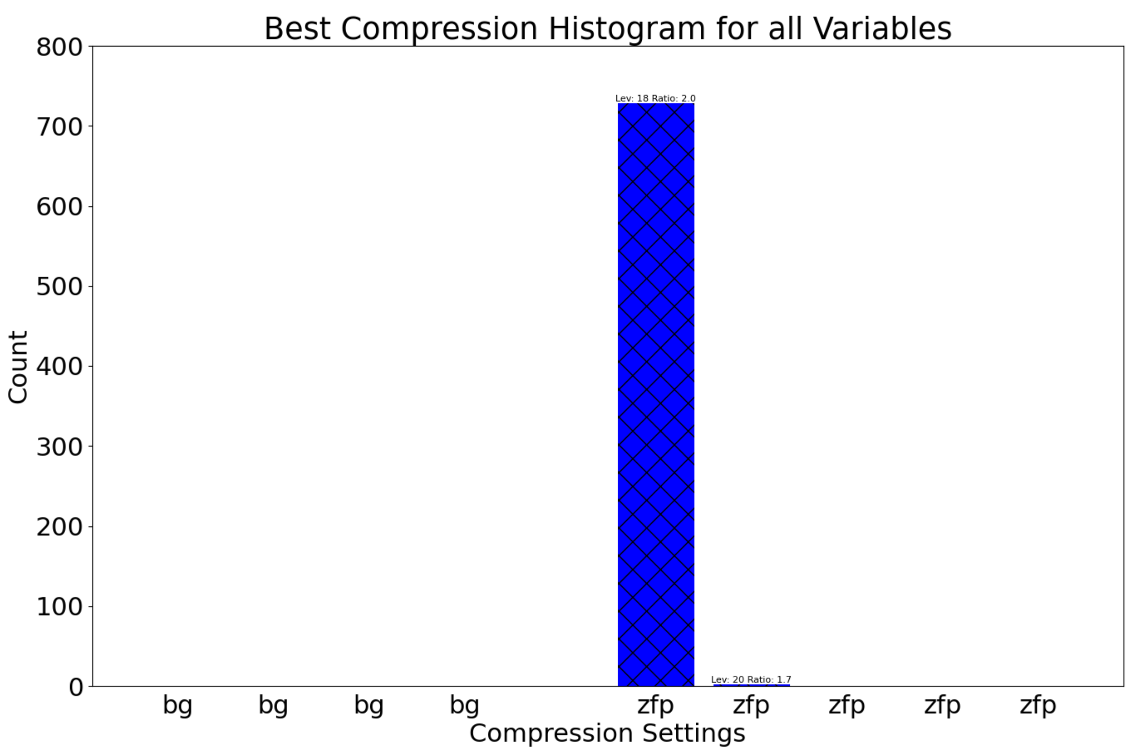

Best Compression Histogram for all Variables
Lev: 5 Ratio: 1.3
Count
800
700
600
500
400
300
200
100
0
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

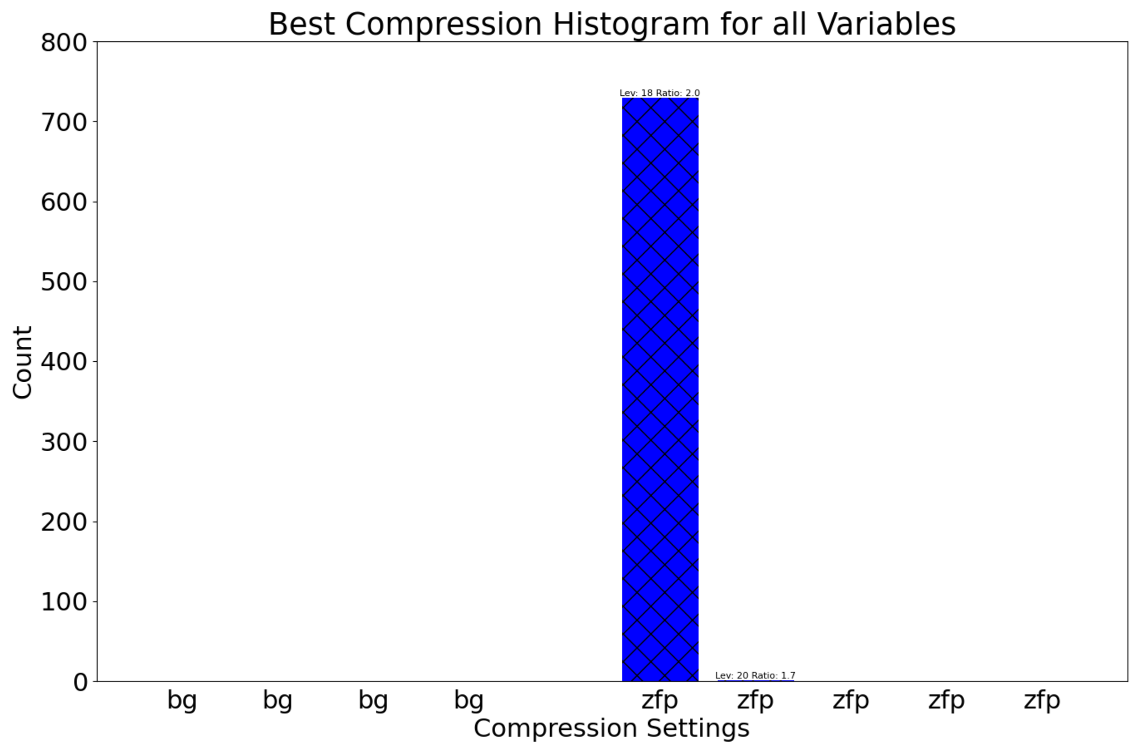

Best Compression Histogram for all Variables
Lev: 3 Ratio: 1.5
Lev: 4 Ratio: 1.3
Count
800
700
600
500
400
300
200
100
0
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

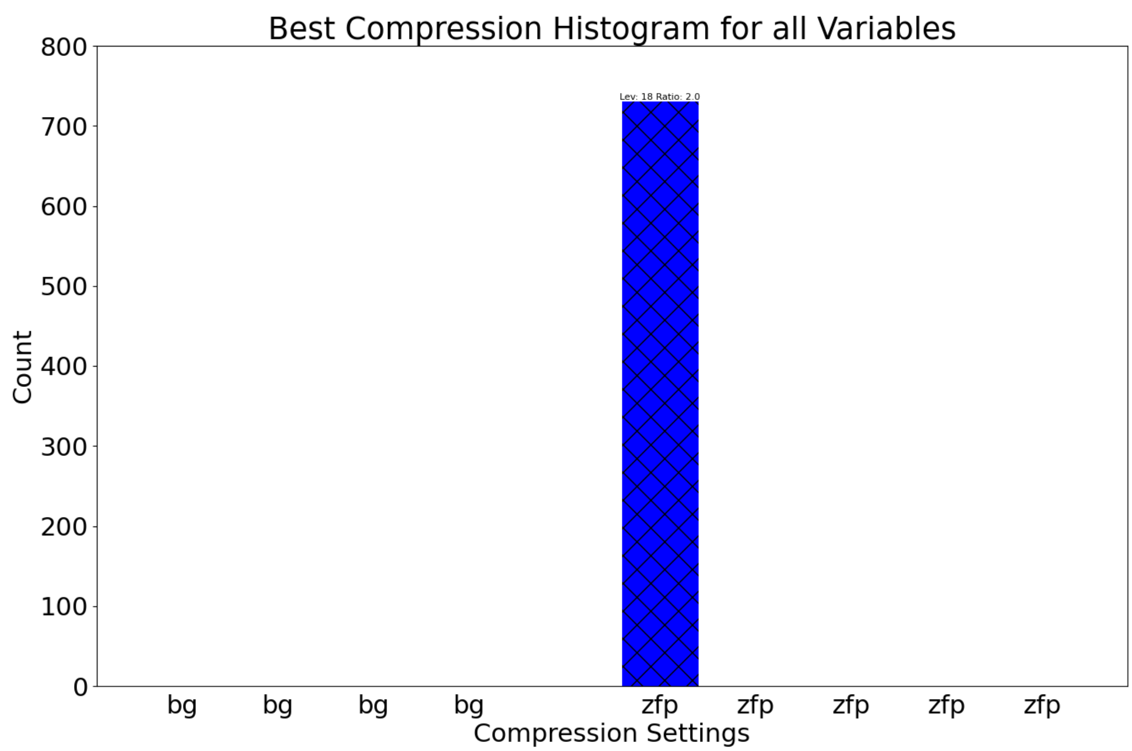

Best Compression Histogram for all Variables
Count
Compression Settings
Lev: 5 Ratio: 1.3
800
700
600
500
400
300
200
100
0
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp

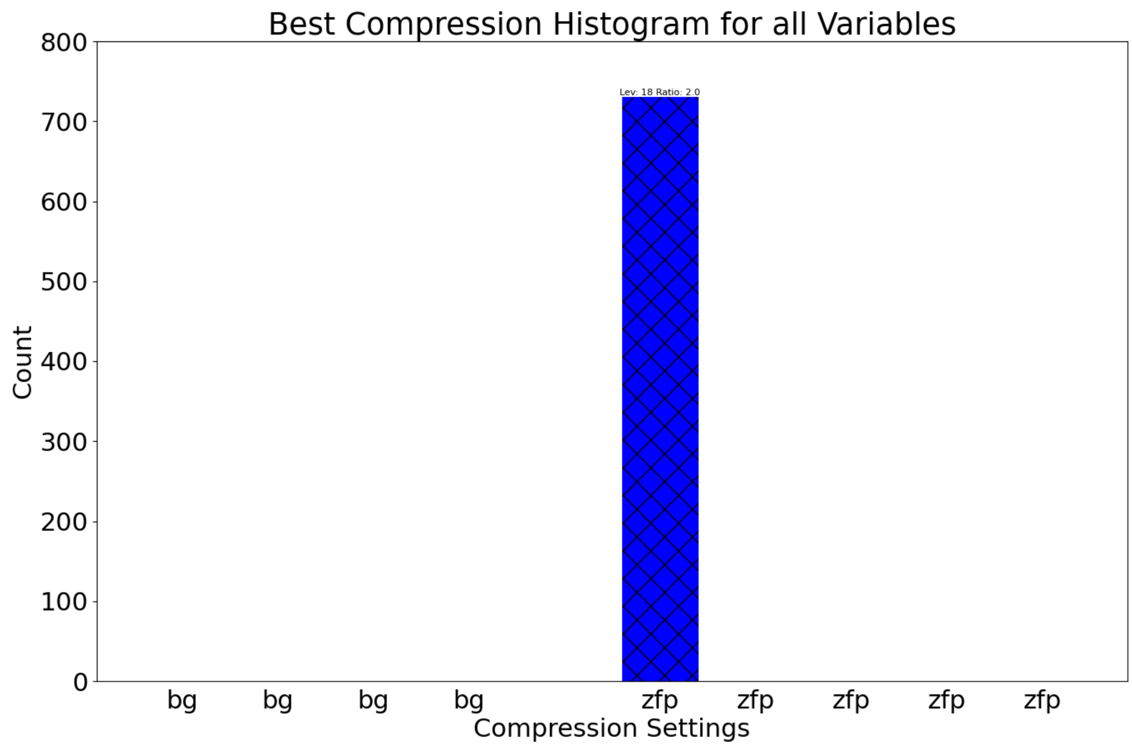

Best Compression Histogram for all Variables
Count
Compression Settings
Lev: 5 Ratio: 1.3
800
700
600
500
400
300
200
100
0
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp

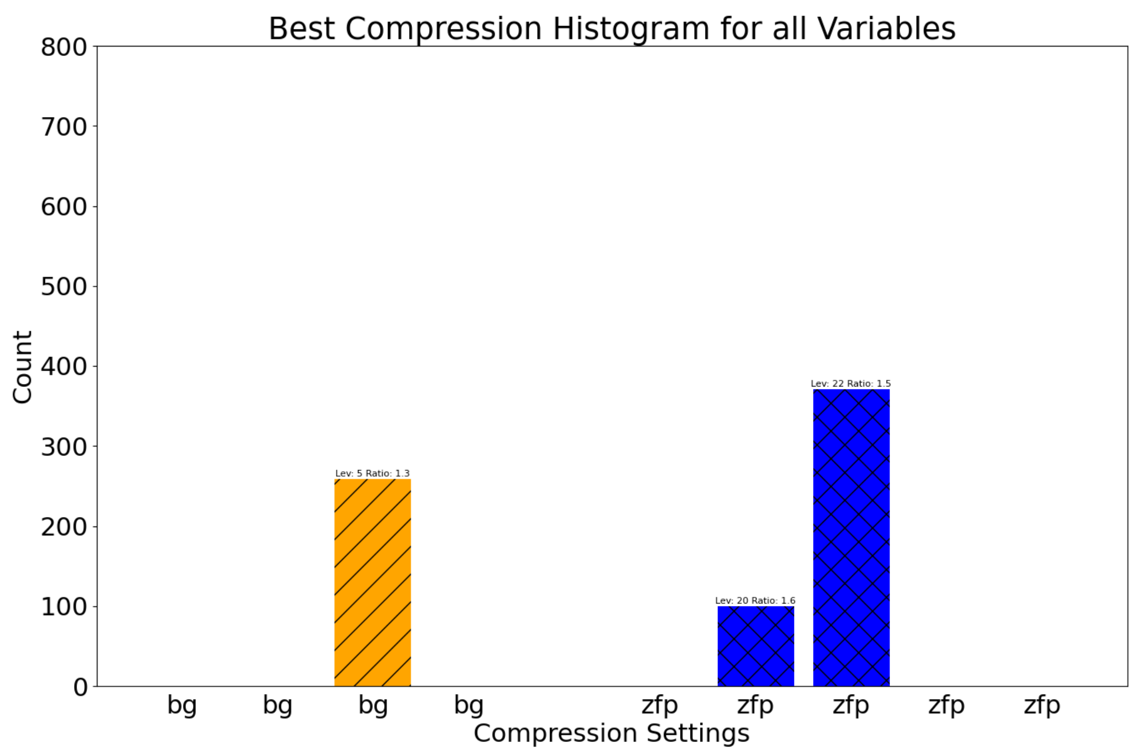

Best Compression Histogram for all Variables
Count
Compression Settings
0
100
200
300
400
500
600
700
800
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Lev: 22 Ratio: 1.7
Lev: 24 Ratio: 1.5

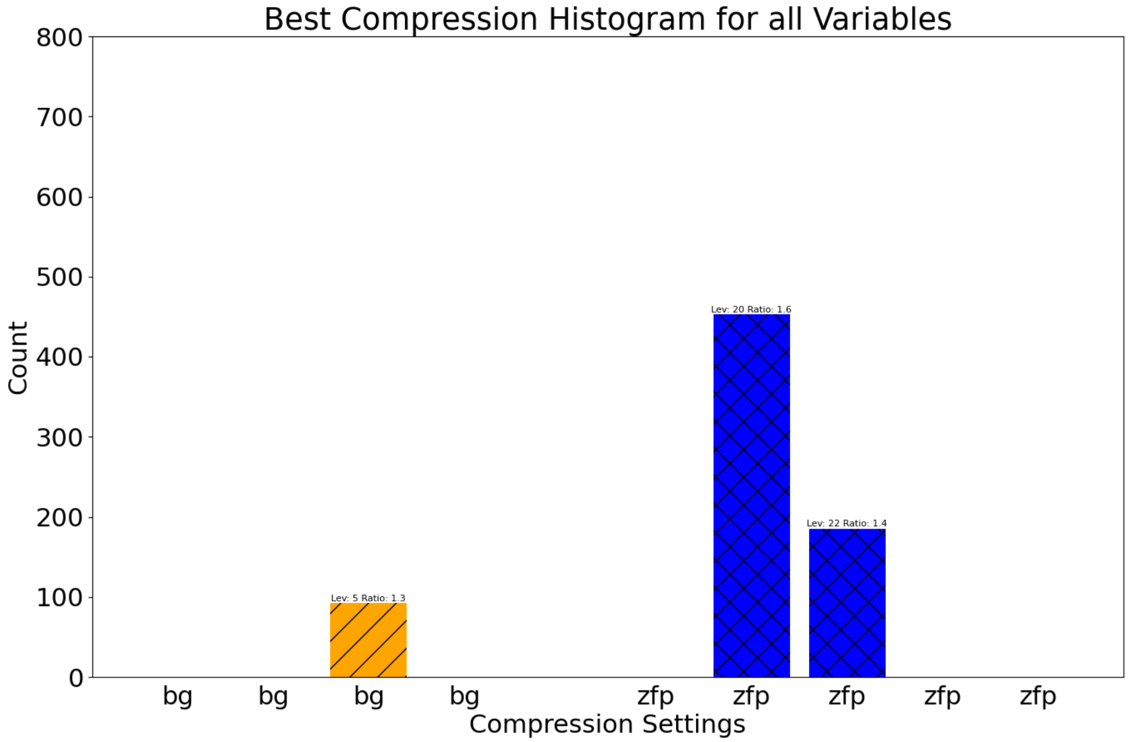

Best Compression Histogram for all Variables
Count
Compression Settings
0
100
200
300
400
500
600
700
800
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Lev: 18 Ratio: 2.0
Lev: 20 Ratio: 1.7

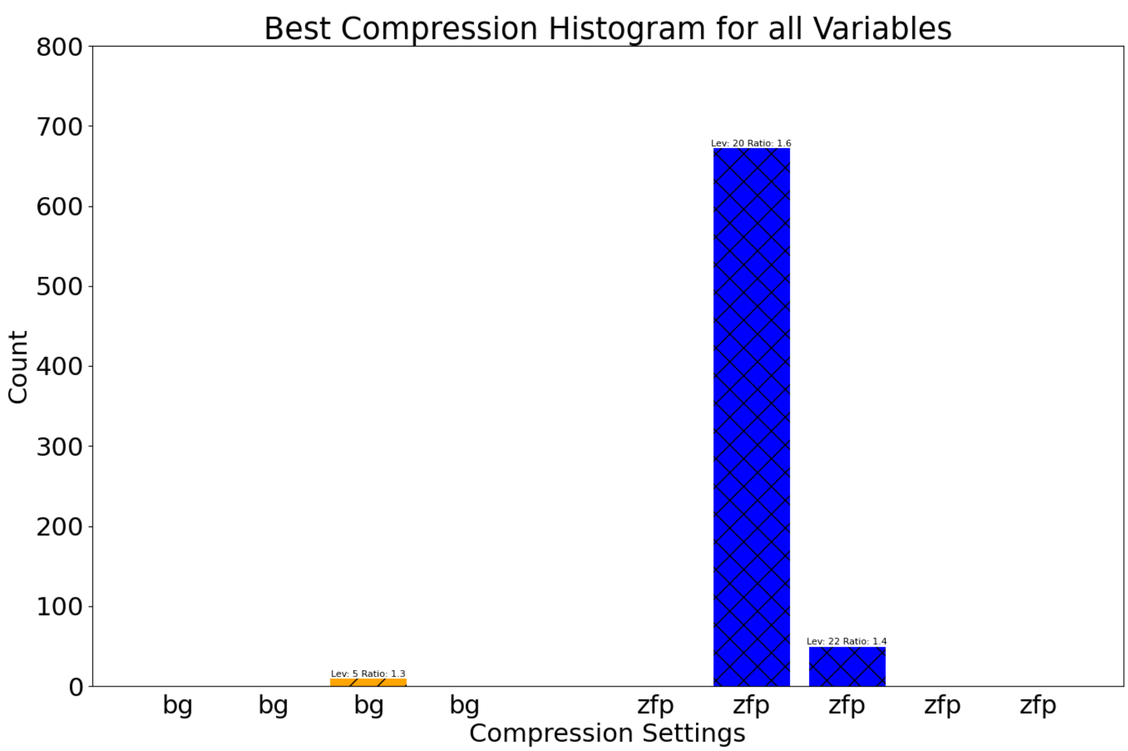

Best Compression Histogram for all Variables
Lev: 18 Ratio: 2.0
Lev: 20 Ratio: 1.7
Count
Compression Settings
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp

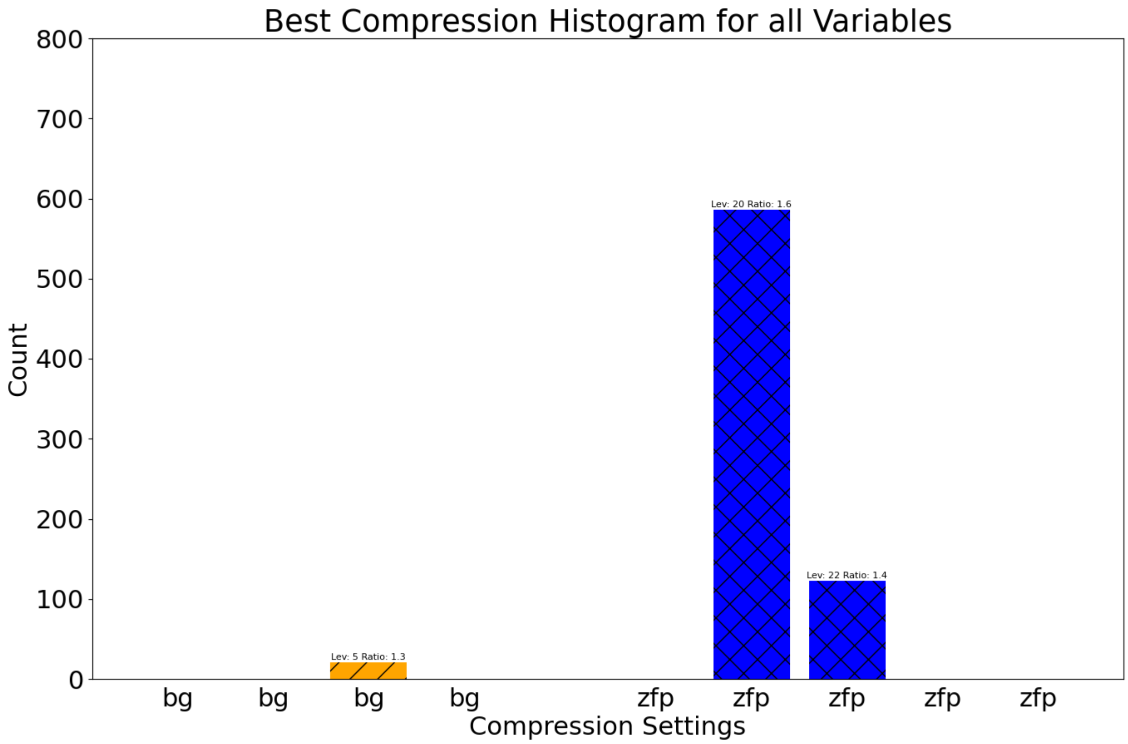

Best Compression Histogram for all Variables
Lev: 18 Ratio: 2.0
Count
Compression Settings
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp

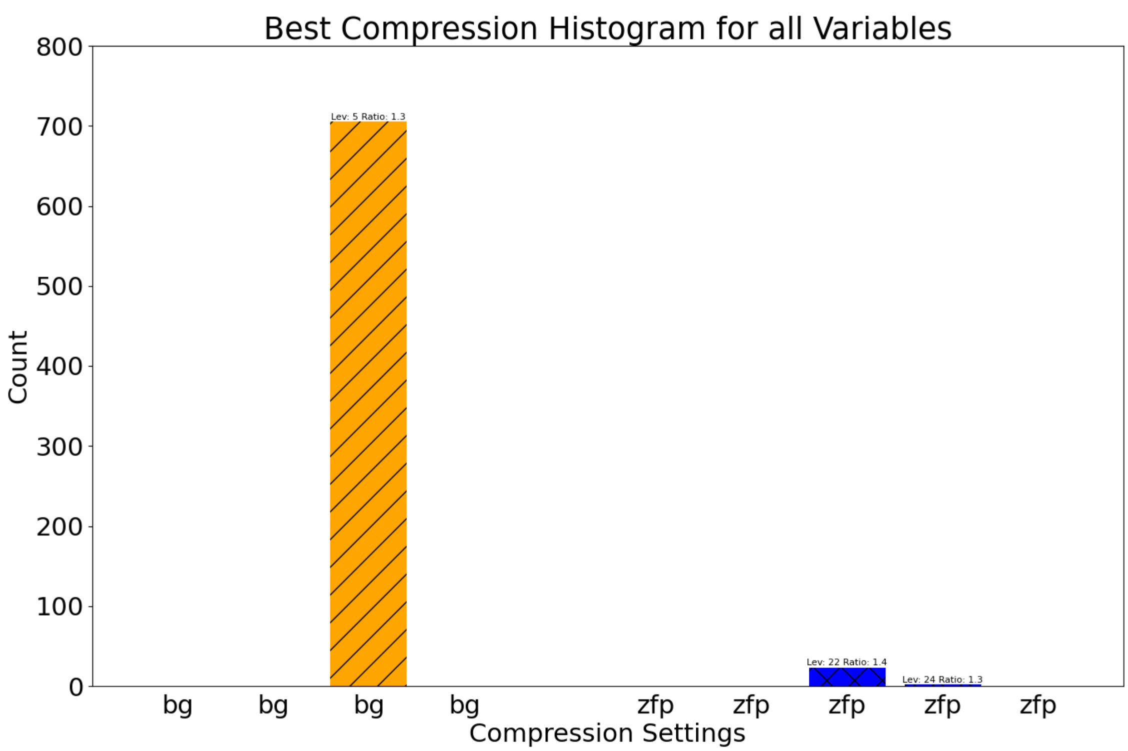

Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 18 Ratio: 2.0
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

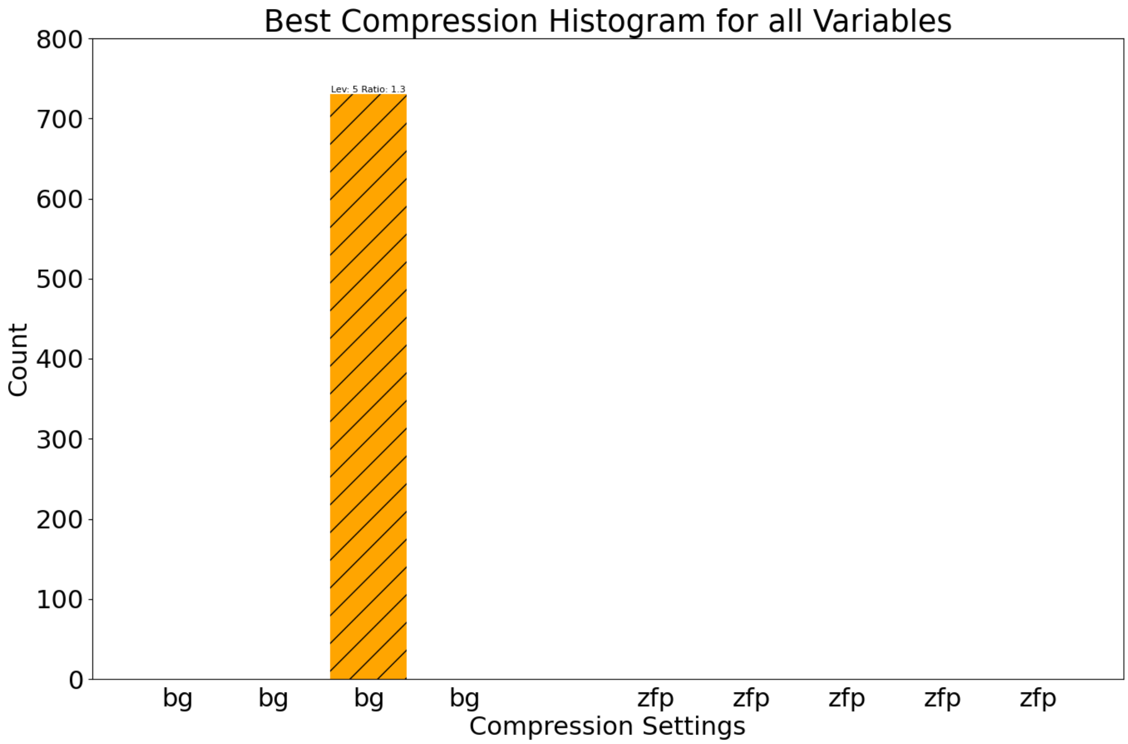

Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 5 Ratio: 1.3
Lev: 20 Ratio: 1.6
Lev: 22 Ratio: 1.5
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

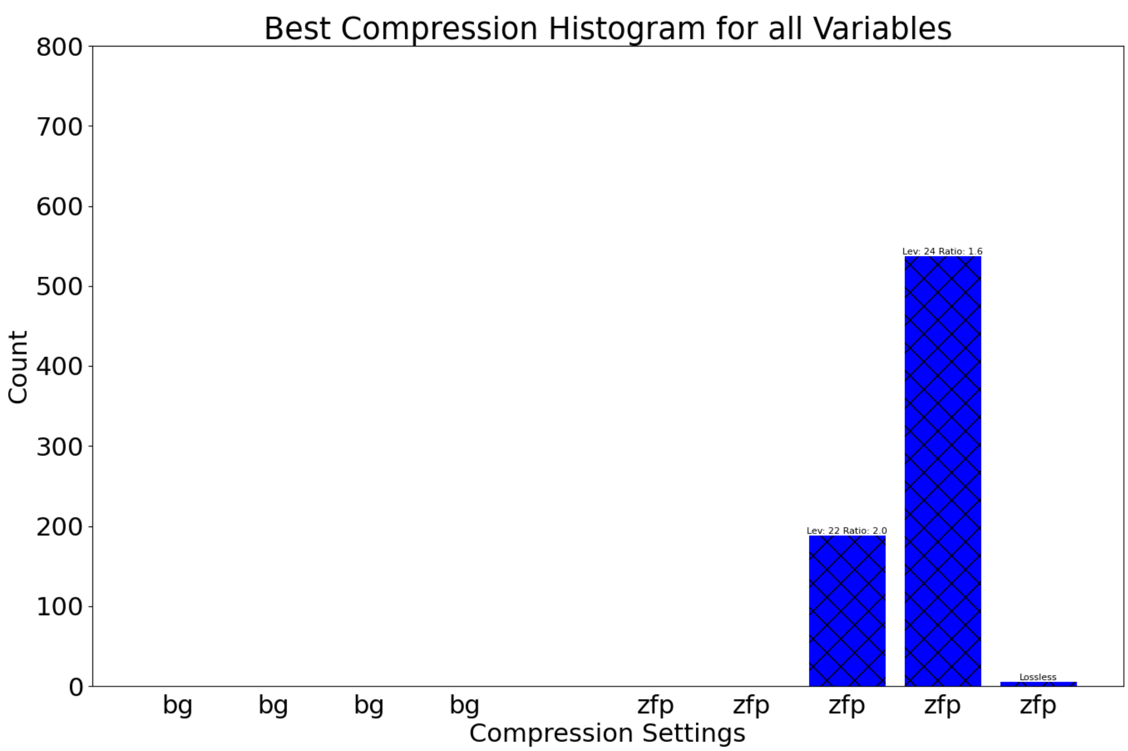

Best Compression Histogram for all Variables
Count
800
700
600
500
400
300
200
100
0
Lev: 5 Ratio: 1.3
Lev: 20 Ratio: 1.6
Lev: 22 Ratio: 1.4
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

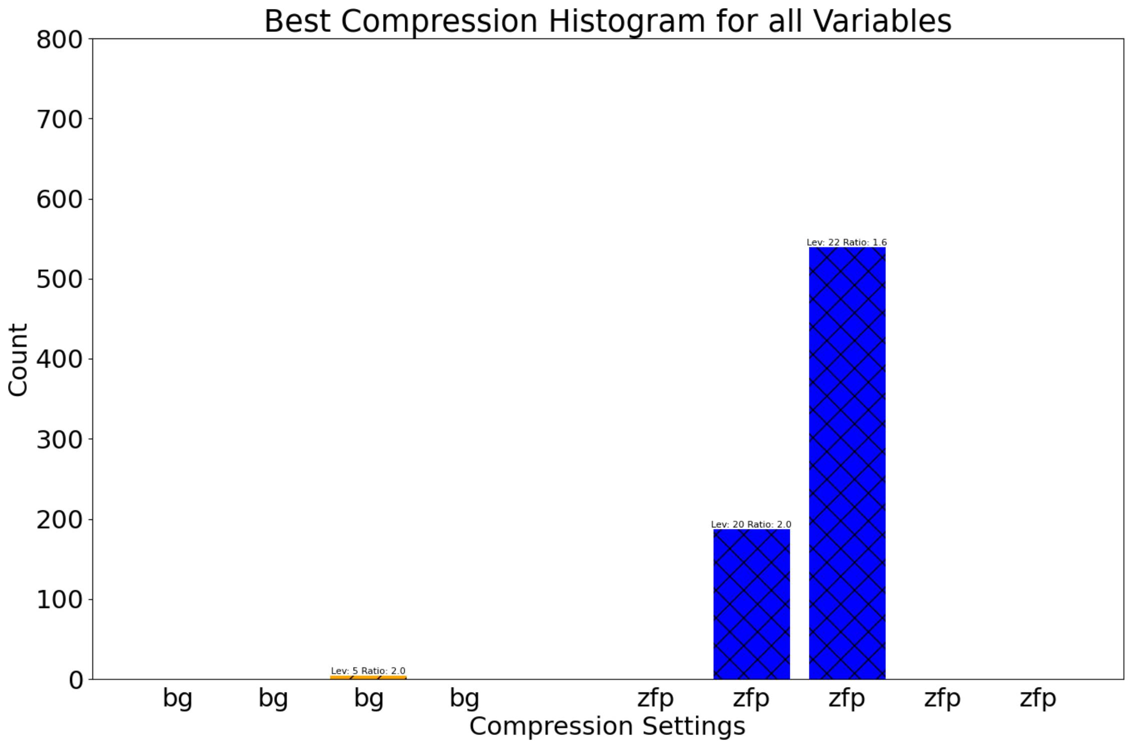

Best Compression Histogram for all Variables
Count
800
700
600
500
400
300
200
100
0
Lev: 20 Ratio: 1.6
Lev: 22 Ratio: 1.4
Lev: 5 Ratio: 1.3
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

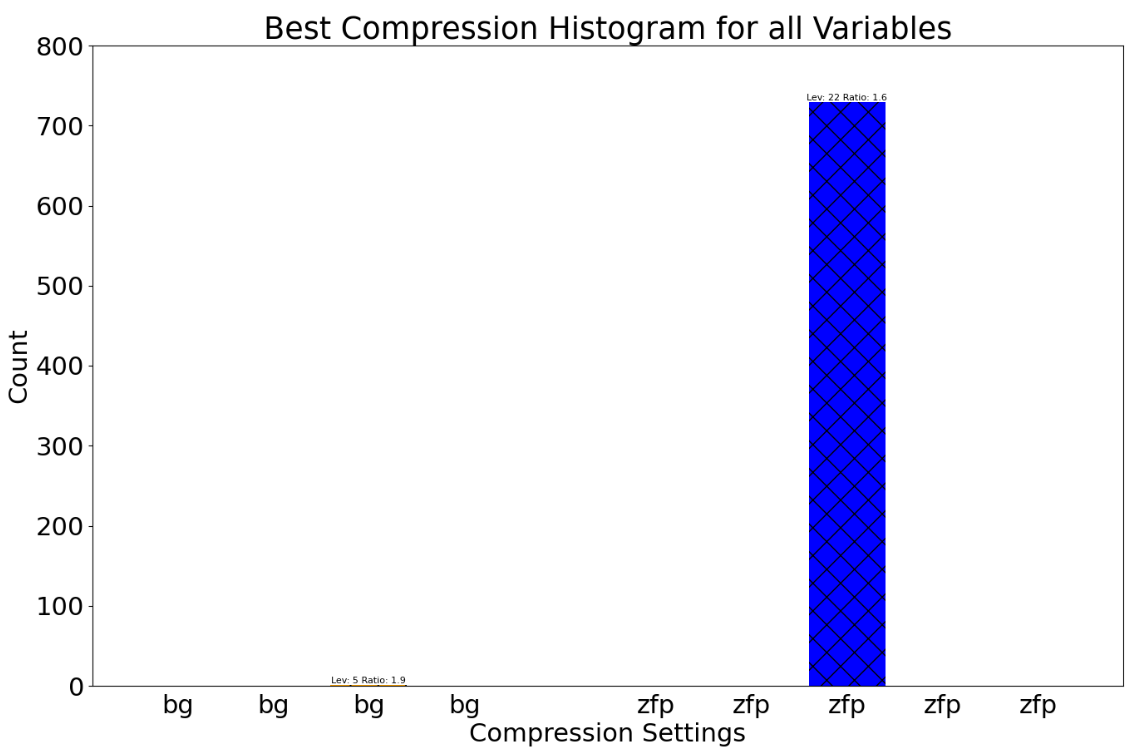

Best Compression Histogram for all Variables
Count
800
700
600
500
400
300
200
100
0
Lev: 5 Ratio: 1.3
Lev: 20 Ratio: 1.6
Lev: 22 Ratio: 1.4
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

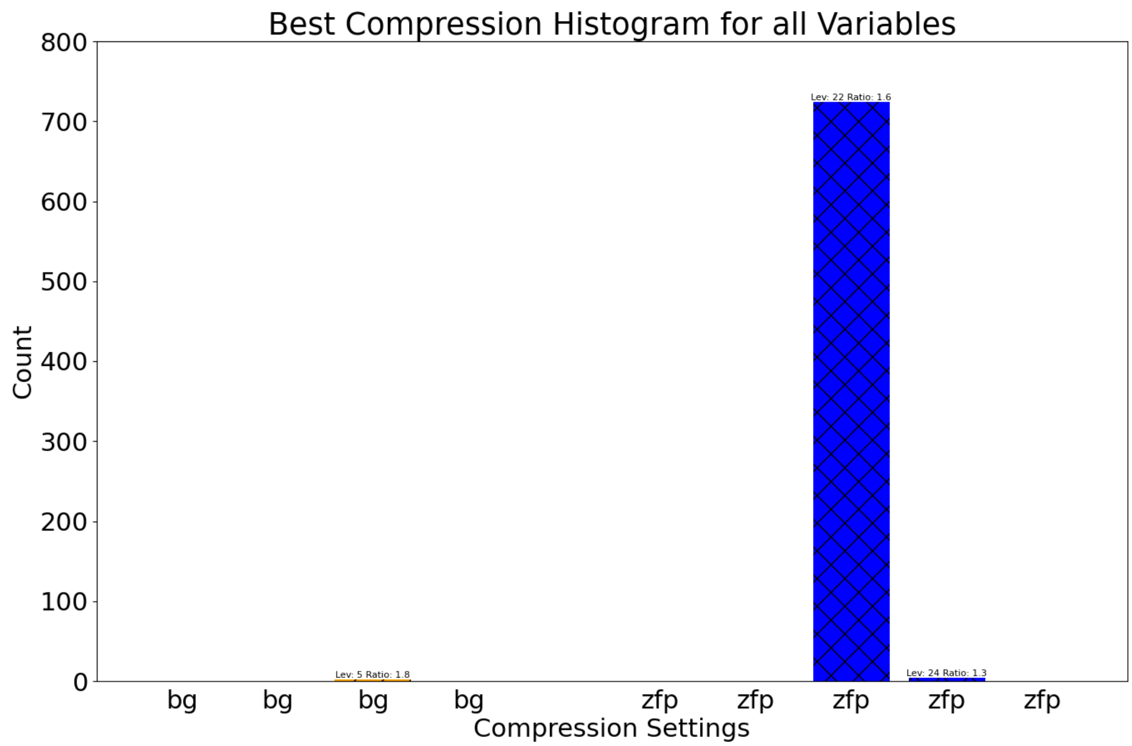

Best Compression Histogram for all Variables
Count
800
700
600
500
400
300
200
100
0
Lev: 5 Ratio: 1.3
Lev: 22 Ratio: 1.4
Lev: 24 Ratio: 1.3
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

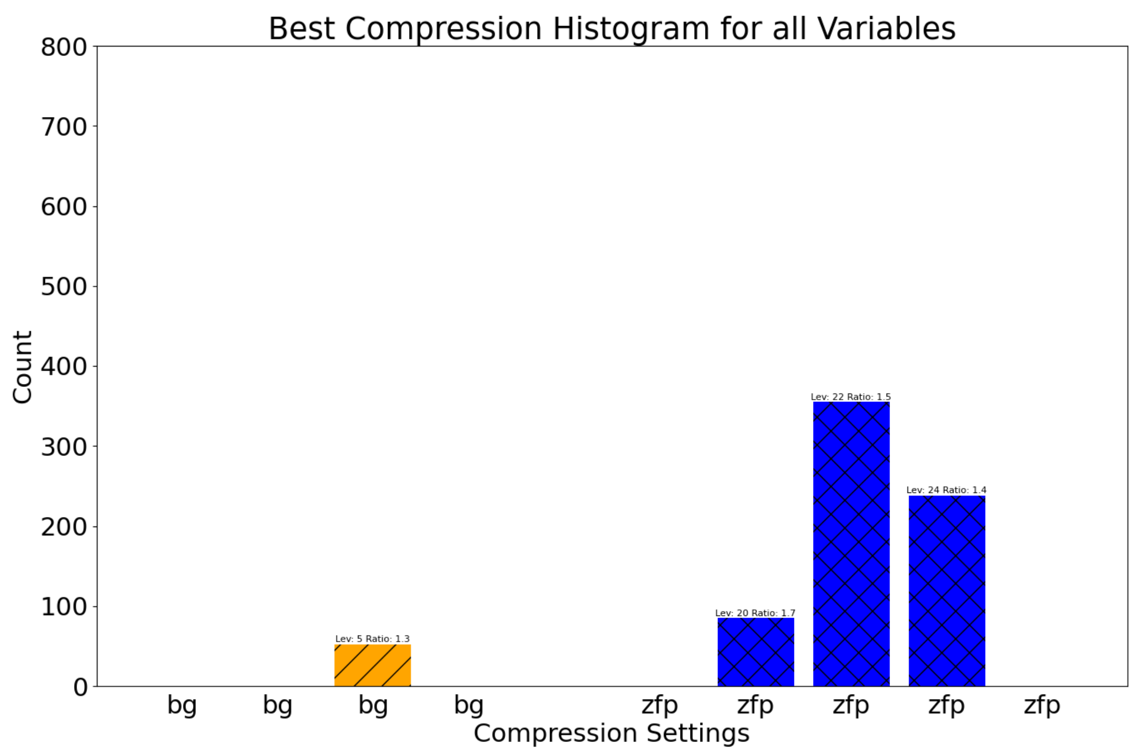

Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 5 Ratio: 1.3
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

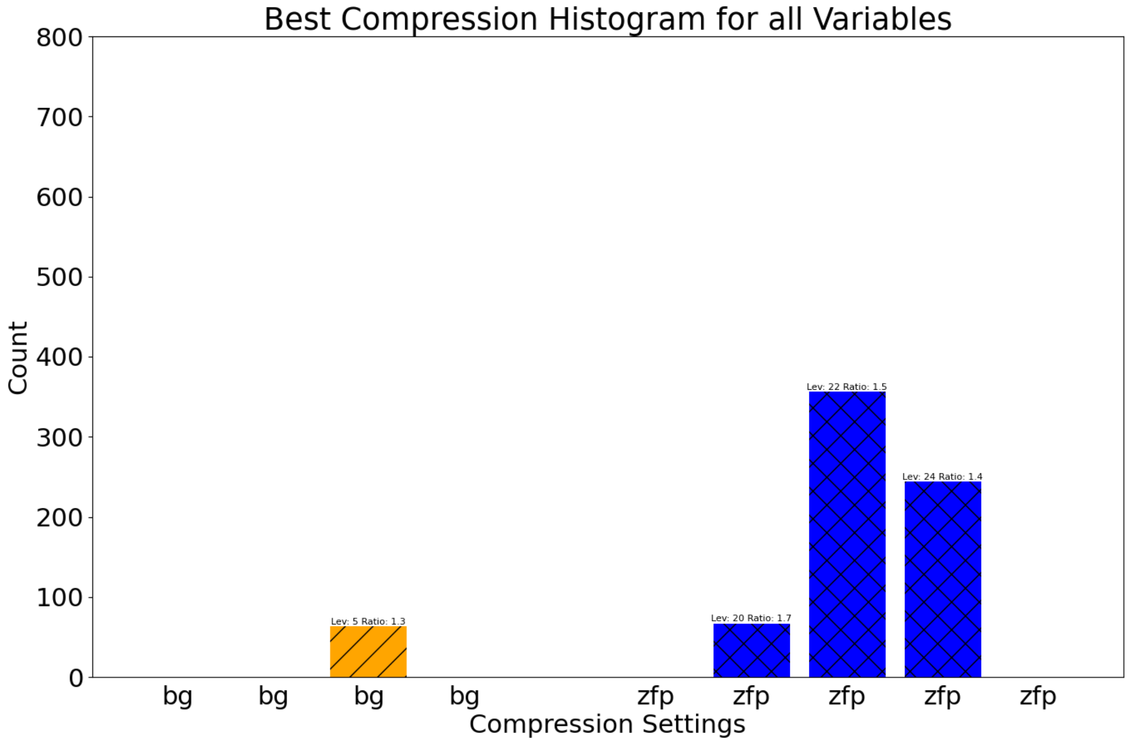

Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 22 Ratio: 2.0
Lev: 24 Ratio: 1.6
Lossless
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

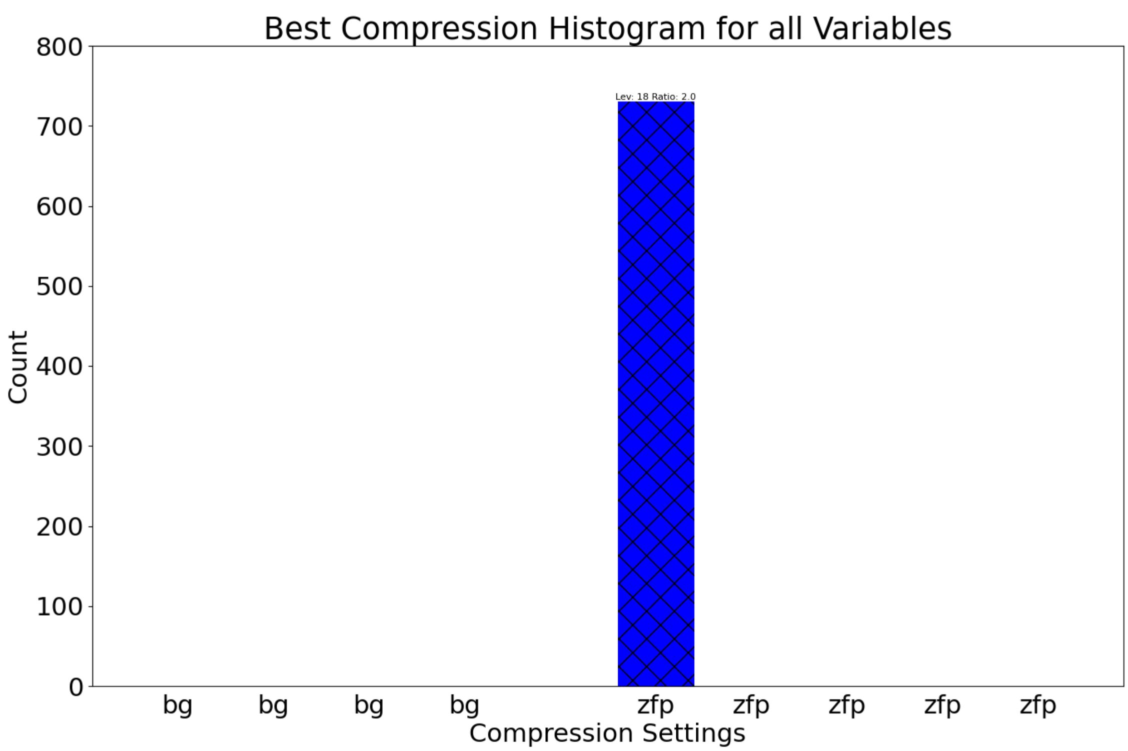

Best Compression Histogram for all Variables
Count
800
700
600
500
400
300
200
100
0
Lev: 5 Ratio: 2.0
Lev: 20 Ratio: 2.0
Lev: 22 Ratio: 1.6
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

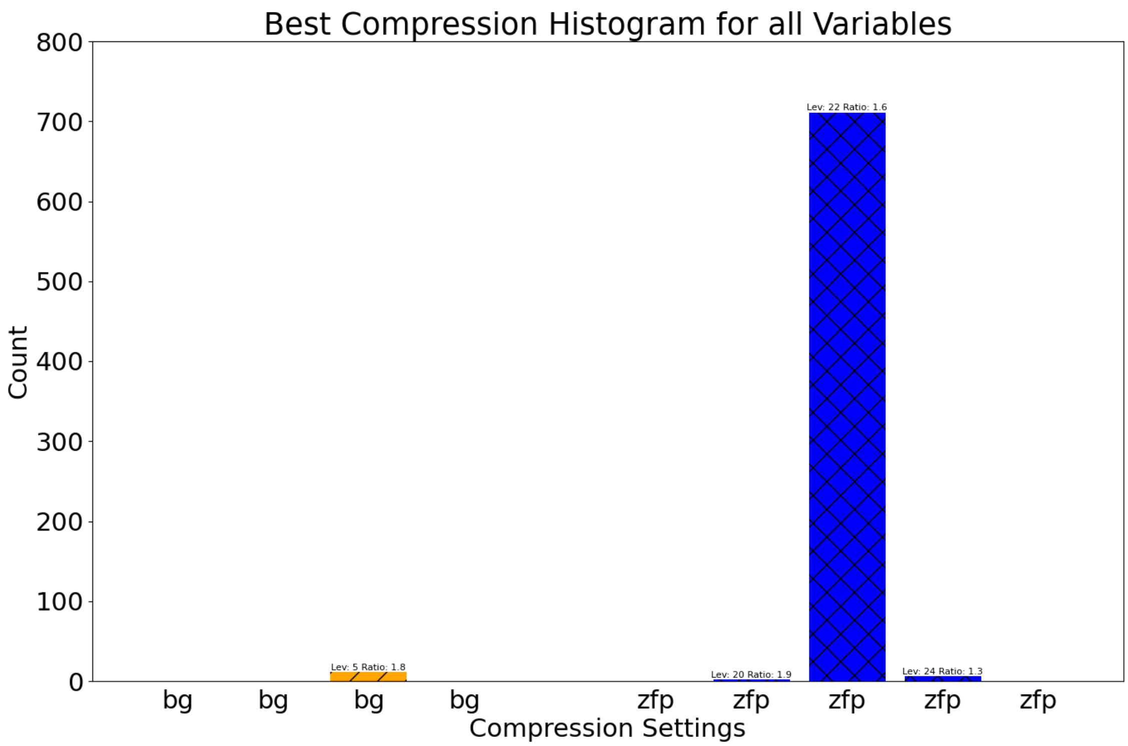

Best Compression Histogram for all Variables
Count
800
700
600
500
400
300
200
100
0
Lev: 5 Ratio: 1.9
Lev: 22 Ratio: 1.6
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

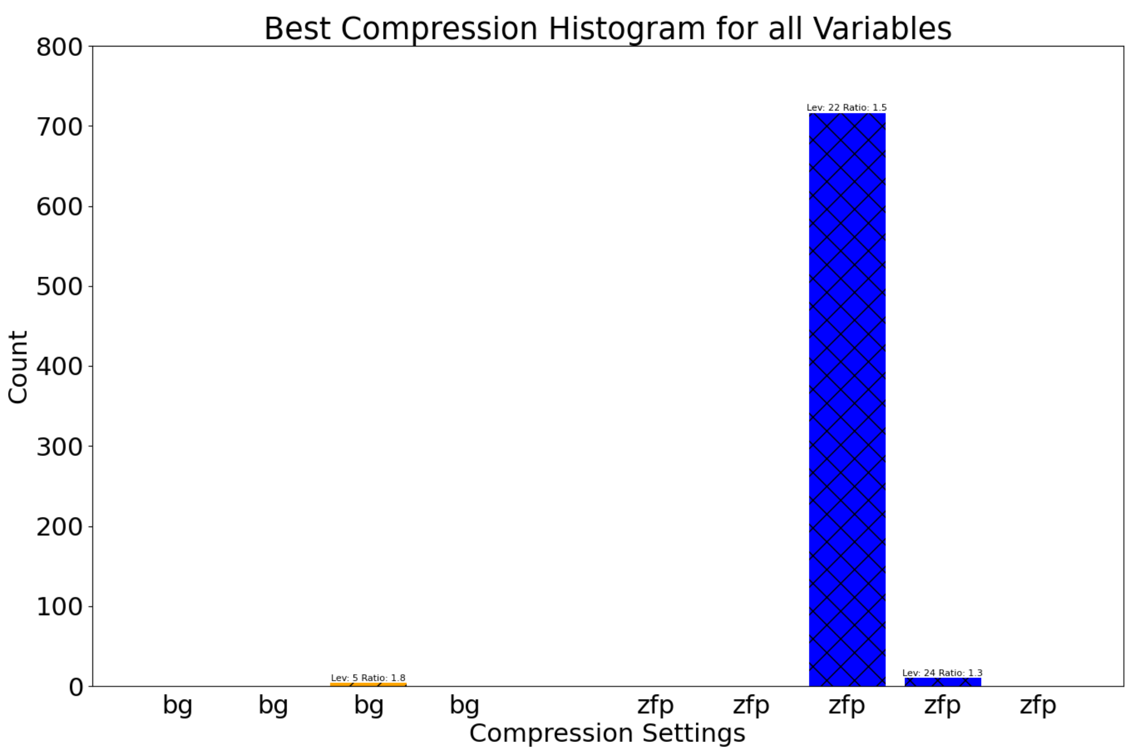

Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 22 Ratio: 1.6
Lev: 5 Ratio: 1.8
Lev: 24 Ratio: 1.3
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

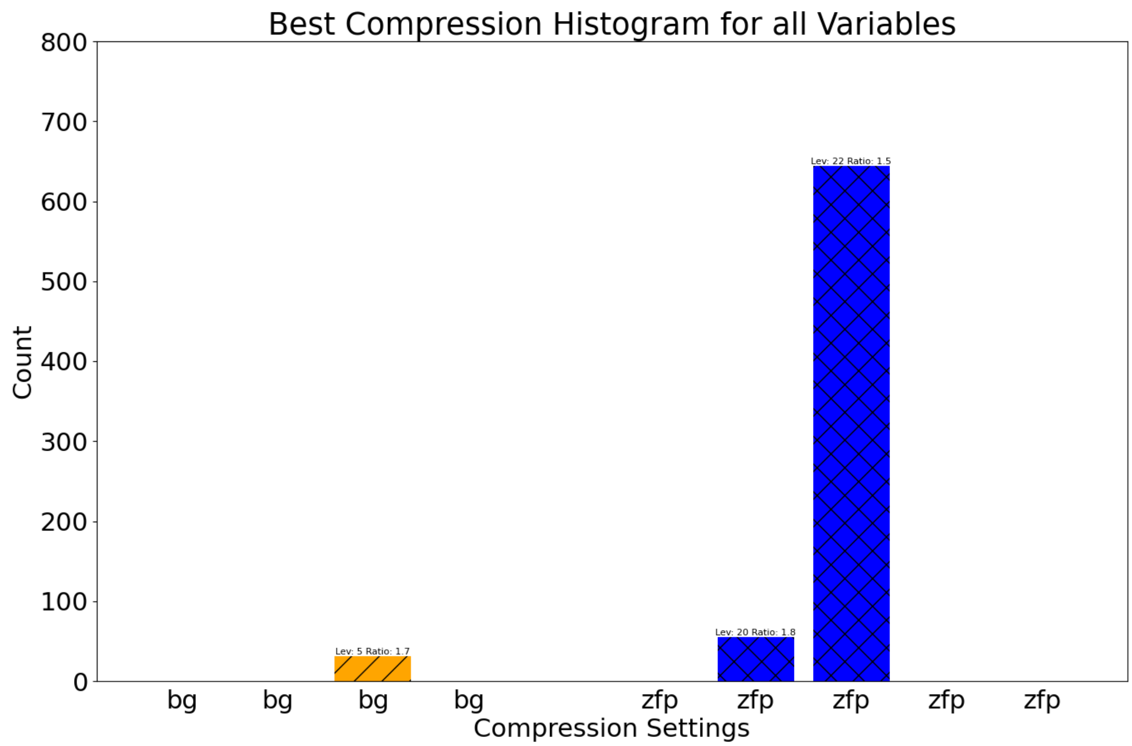

Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 22 Ratio: 1.5
Lev: 24 Ratio: 1.4
Lev: 20 Ratio: 1.7
Lev: 5 Ratio: 1.3
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

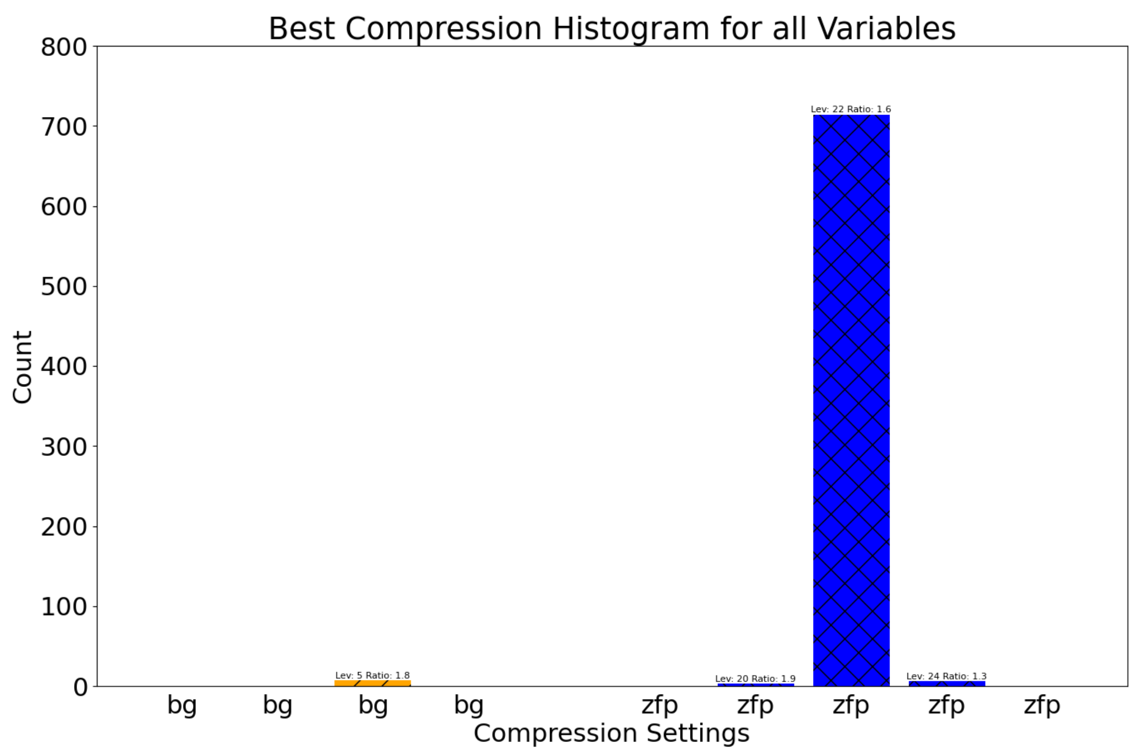

Best Compression Histogram for all Variables
Count
Compression Settings
Lev: 5 Ratio: 1.3
Lev: 20 Ratio: 1.7
Lev: 22 Ratio: 1.5
Lev: 24 Ratio: 1.4
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp

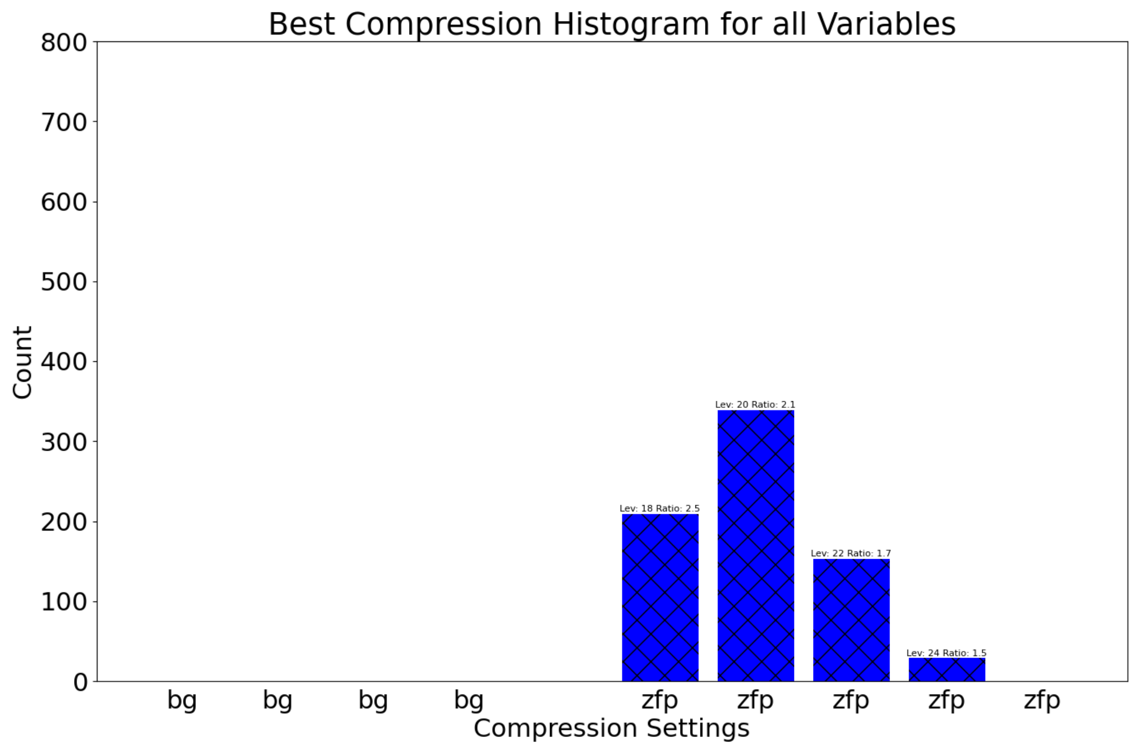

Best Compression Histogram for all Variables
Count
Compression Settings
Lev: 18 Ratio: 2.0
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp

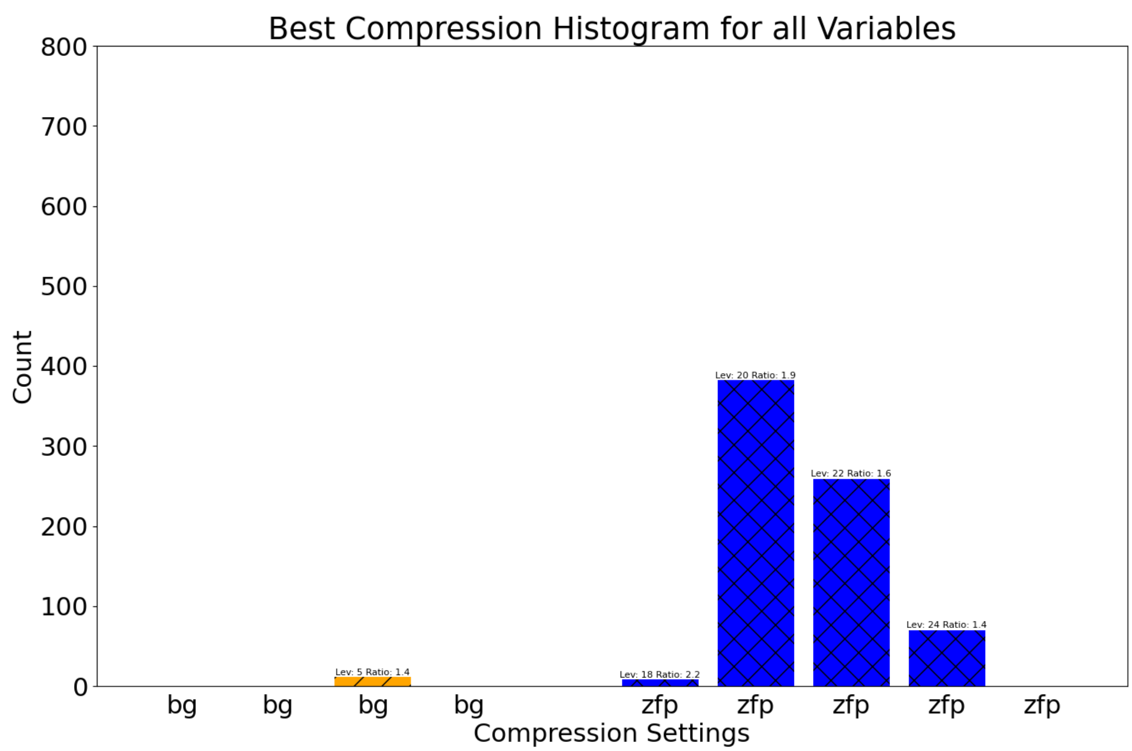

Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 22 Ratio: 1.6
Lev: 5 Ratio: 1.8
Lev: 20 Ratio: 1.9
Lev: 24 Ratio: 1.3
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

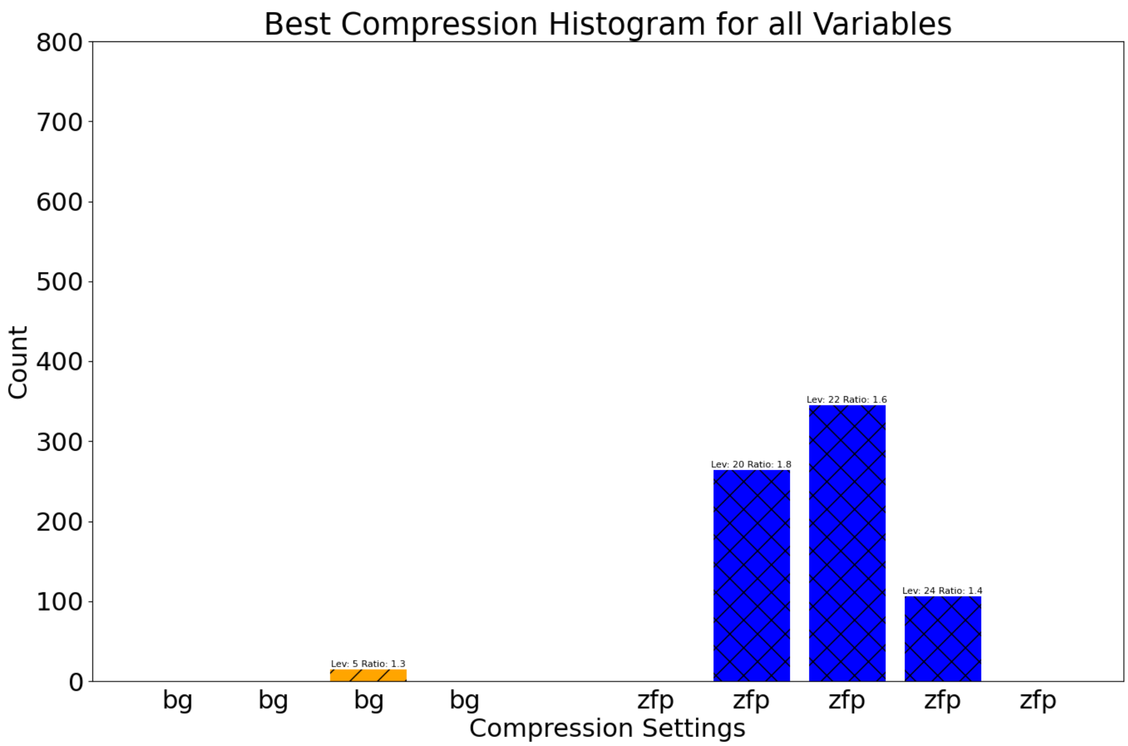

Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 22 Ratio: 1.5
Lev: 5 Ratio: 1.8
Lev: 24 Ratio: 1.3
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

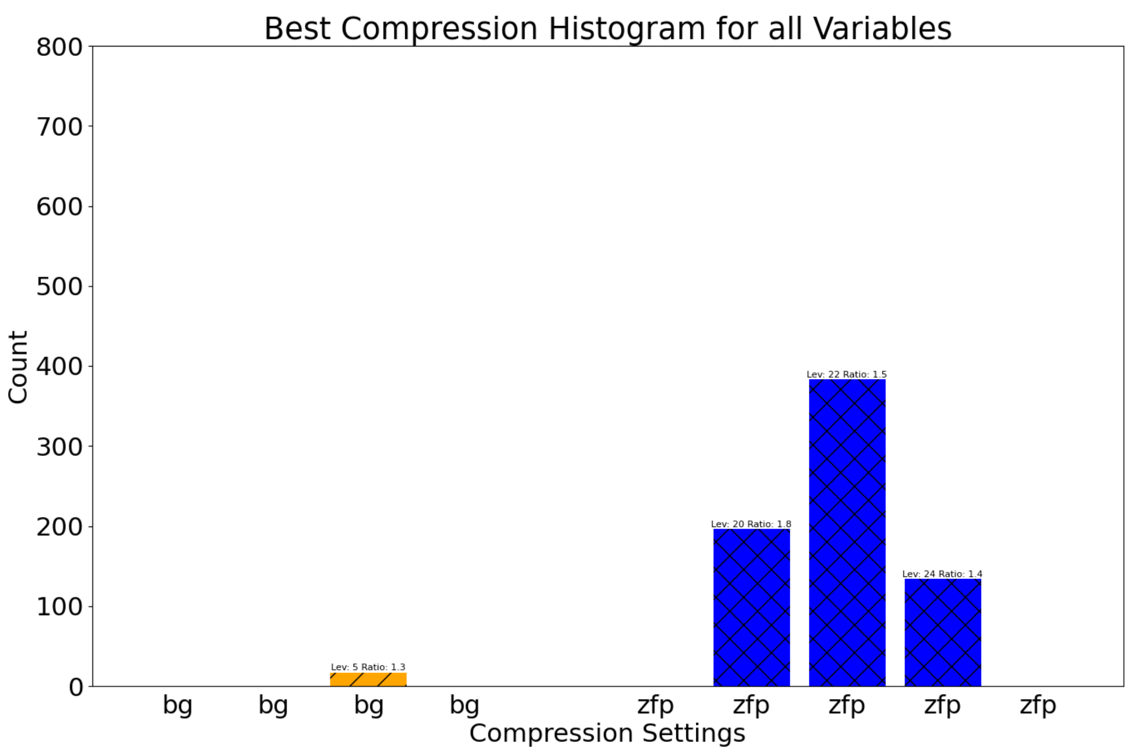

Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 5 Ratio: 1.7
Lev: 20 Ratio: 1.8
Lev: 22 Ratio: 1.5
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

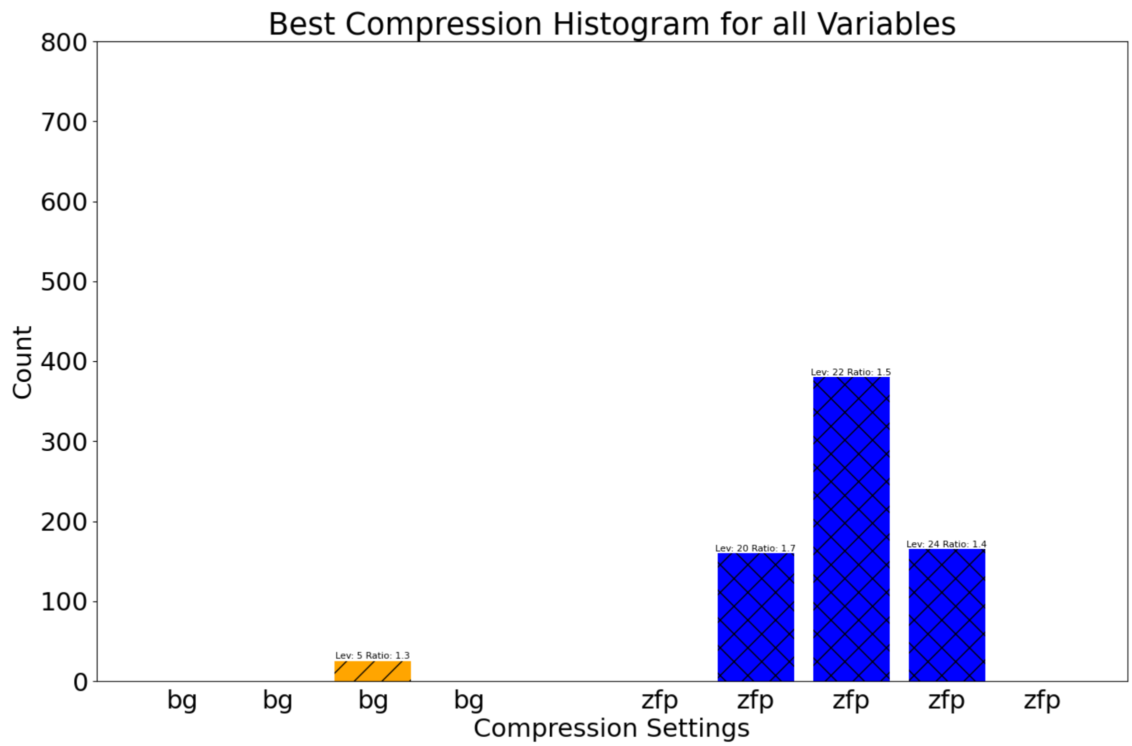

Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 5 Ratio: 1.8
Lev: 20 Ratio: 1.9
Lev: 22 Ratio: 1.6
Lev: 24 Ratio: 1.3
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

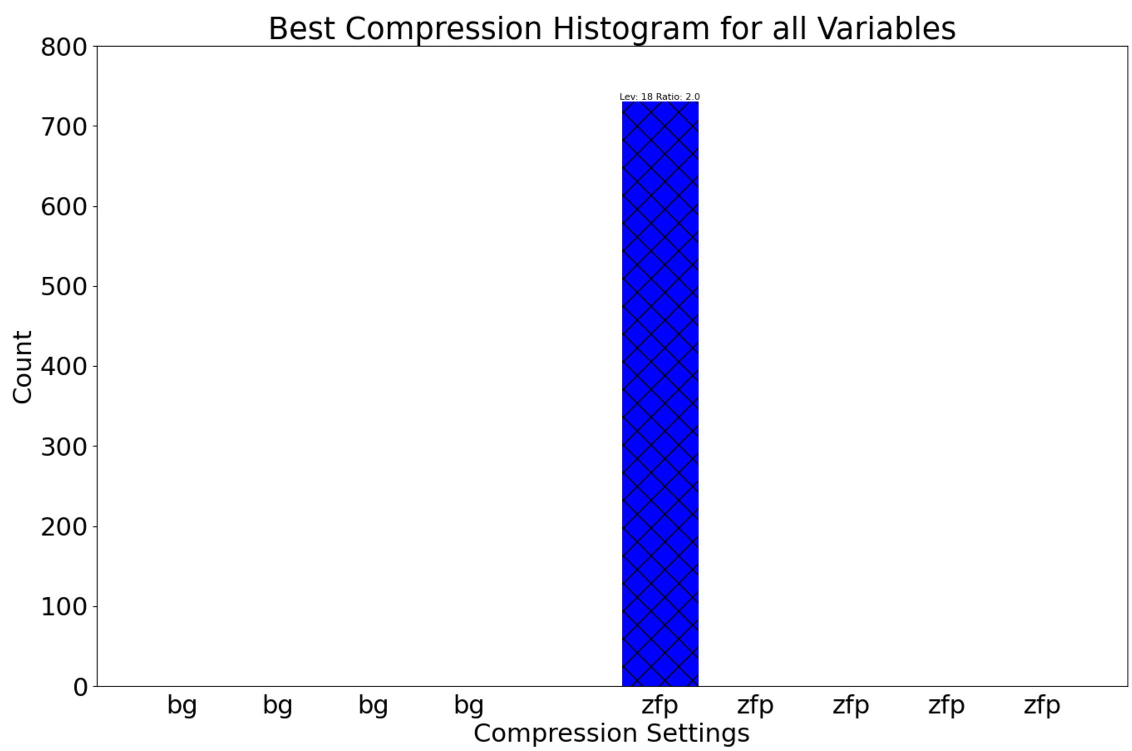

Best Compression Histogram for all Variables
Count
Compression Settings
Lev: 18 Ratio: 2.5
Lev: 20 Ratio: 2.1
Lev: 22 Ratio: 1.7
Lev: 24 Ratio: 1.5
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp

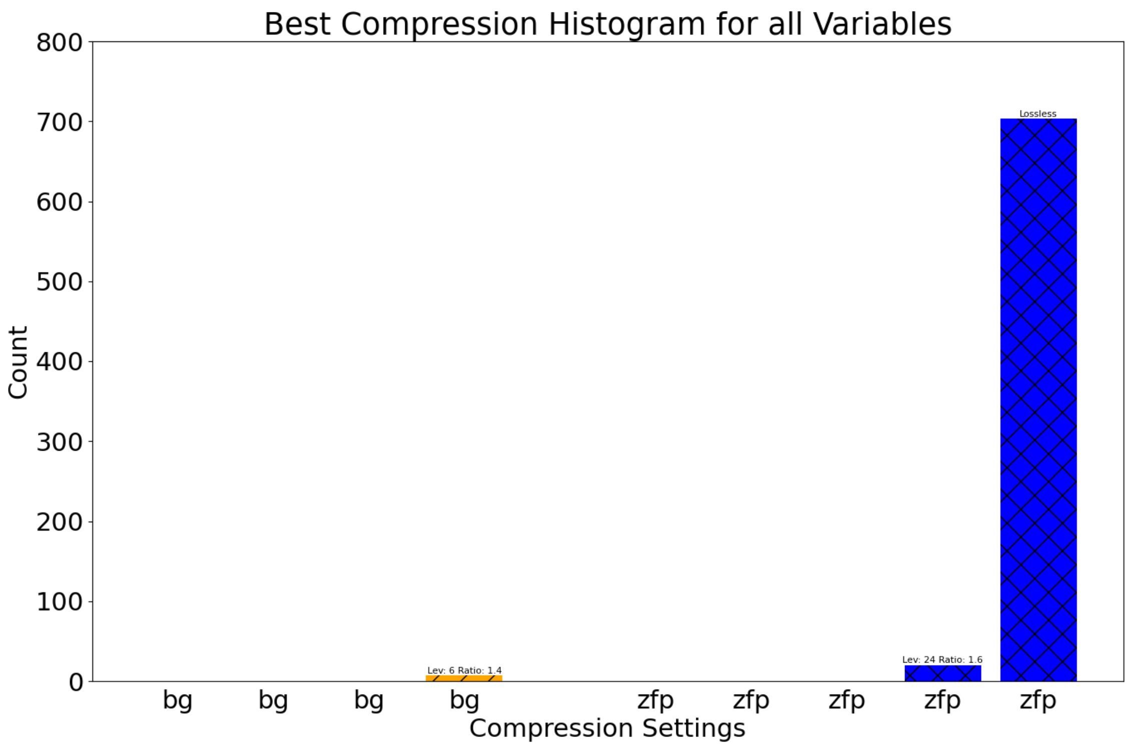

Best Compression Histogram for all Variables
Count
Compression Settings
Lev: 5 Ratio: 1.4
Lev: 18 Ratio: 2.2
Lev: 20 Ratio: 1.9
Lev: 22 Ratio: 1.6
Lev: 24 Ratio: 1.4
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp

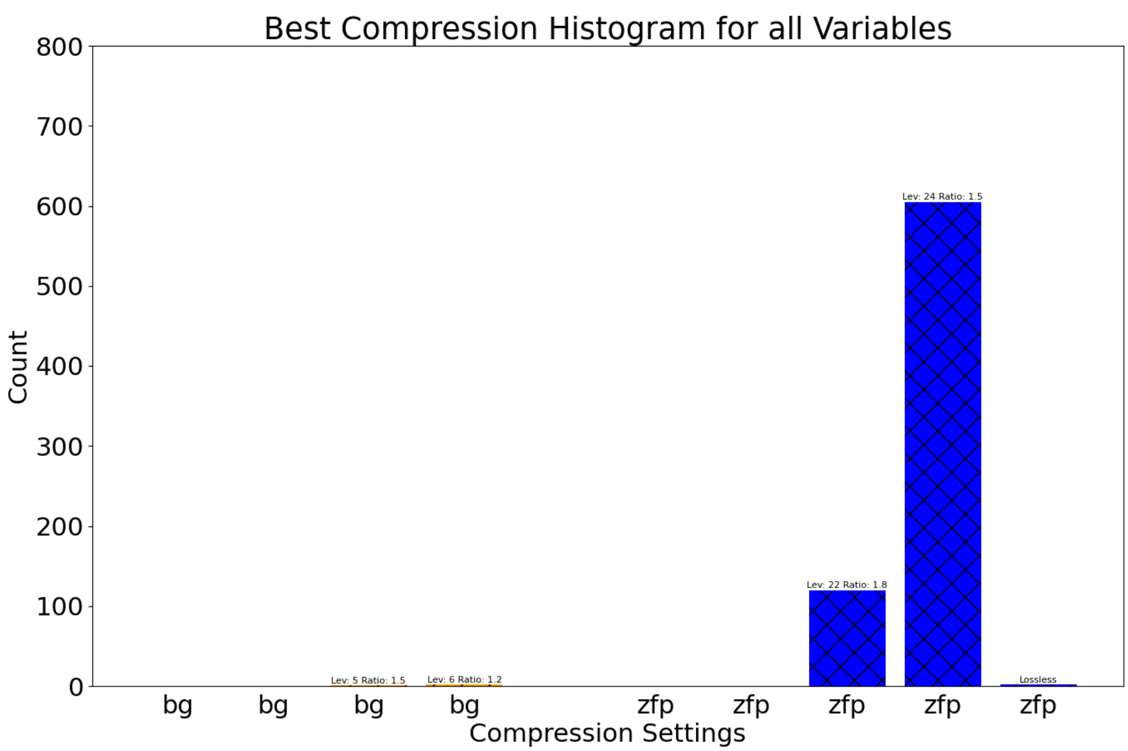

Best Compression Histogram for all Variables
Count
Compression Settings
800
700
600
500
400
300
200
100
0
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Lev: 5 Ratio: 1.3
Lev: 20 Ratio: 1.8
Lev: 22 Ratio: 1.6
Lev: 24 Ratio: 1.4


Best Compression Histogram for all Variables
Count
Compression Settings
800
700
600
500
400
300
200
100
0
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Lev: 5 Ratio: 1.3
Lev: 20 Ratio: 1.8
Lev: 22 Ratio: 1.5
Lev: 24 Ratio: 1.4


Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 5 Ratio: 1.3
Lev: 20 Ratio: 1.7
Lev: 22 Ratio: 1.5
Lev: 24 Ratio: 1.4
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings


Best Compression Histogram for all Variables
800
700
600
500
400
300
200
100
0
Count
Lev: 18 Ratio: 2.0
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings


Best Compression Histogram for all Variables
Count
800
700
600
500
400
300
200
100
0
Lossless
Lev: 24 Ratio: 1.6
Lev: 6 Ratio: 1.4
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings


Best Compression Histogram for all Variables
Count
800
700
600
500
400
300
200
100
0
Lev: 24 Ratio: 1.5
Lev: 22 Ratio: 1.8
Lev: 5 Ratio: 1.5
Lev: 6 Ratio: 1.2
Lossless
bg
bg
bg
bg
zfp
zfp
zfp
zfp
zfp
Compression Settings

[4] A. Pinard, D. M. Hammerling, and A. H. Baker, "Assessing differences in large spatio-temporal climate datasets with a new python package," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 2699–2707.

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, New York, 2013.