DOI: 10.1111/cgf.13707
Eurographics Conference on Visualization (EuroVis) 2019
M. Gleicher, H. Leitte, and I. Viola
(Guest Editors)

Volume 38 (2019), Number 3

Evaluating image quality measures to assess the impact of lossy data compression applied to climate simulation data

A. H. Baker[1], D. M. Hammerling[2] and T. L. Turton[3]

[1]National Center for Atmospheric Research, Boulder, Colorado, U. S. A
[2]Colorado School of Mines, Golden, Colorado, U. S. A
[3] Los Alamos National Laboratory, Los Alamos, New Mexico, U. S. A.

Abstract

Applying lossy data compression to climate model output is an attractive means of reducing the enormous volumes of data generated by climate models. However, because lossy data compression does not exactly preserve the original data, its application to scientific data must be done judiciously. To this end, a collection of measures is being developed to evaluate various aspects of lossy compression quality on climate model output. Given the importance of data visualization to climate scientists interacting with model output, any suite of measures must include a means of assessing whether images generated from the compressed model data are noticeably different from images based on the original model data. Therefore, in this work we conduct a forced-choice visual evaluation study with climate model data that surveyed more than one hundred participants with domain relevant expertise. In addition to the images created from unaltered climate model data, study images are generated from model data that is subjected to two different types of lossy compression approaches and multiple levels (amounts) of compression. Study participants indicate whether a visual difference can be seen, with respect to the reference image, due to lossy compression effects. We assess the relationship between the perceptual scores from the user study to a number of common (full reference) image quality assessment (IQA) measures, and use statistical models to suggest appropriate measures and thresholds for evaluating lossily compressed climate data. We find the structural similarity index (SSIM) to perform the best, and our findings indicate that the threshold required for climate model data is much higher than previous findings in the literature.

Categories and Subject Descriptors (according to ACM CCS): E.4 [Coding and Information Theory]: Data Compaction and Compression—I.5.2 [Design Methodology]: Feature evaluation—H.1.2 [User/Machine Systems]: Human factors—

1. Introduction

Climate model simulations greatly contribute to understanding and predicting the Earth's climate system. Recent advances in high-performance computing have enabled such simulations to run with higher resolutions and higher throughput, resulting in increasingly large data volumes that many climate research computing centers are struggling to store (e.g., [HWK*13, BXD*14, BHM*16, KKL16, Zen16]). For example, raw data requirements for climate models for the upcoming Coupled Model Comparison Project Phase 6 (CMIP6) [MMT*14] are likely to be nearly ten petabytes in size for a single model. In general, scientists must be increasingly cognizant of data volumes when designing experiments and make difficult decisions in terms of running fewer or shorter simulations, using lower resolutions, or outputting data less frequently. Our particular interest in this work is in applying data compression to reduce data volumes from the Community Earth System Model (CESM™) [HHG*13], a popular climate model whose development is led by the National Center for Atmospheric Research (NCAR).

We focus on lossy data compression schemes to reduce climate data storage requirements, as it is well known that lossless compression schemes (i.e., schemes that exactly preserve the data when decompressed) are relatively ineffective on floating-point simulation data (e.g., [LI06, LSE*11, LLW*13]). We note that there have been a number of studies advocating lossy data compression for climate data in particular, such as [WMB*11, HWK*13, KKL16, BXD*14]. The use of lossy data compression requires care: we must ensure that its effects on the original data do not affect scientific conclusions drawn from the data. To this end, striking a balance between effectively reducing data volume and preserving the quality of the climate simulation data is critical. Unfortunately, evaluating whether "data quality" has been preserved is an ill-defined and non-trivial task that requires determining how to quantify the loss of information due to compression. Simple measurements such as the mean squared error (MSE) appear to be insufficient for detecting lossy compression-induced artifacts of interest to climate scientists (e.g., [BXD*14], [BHM*16]). CESM, like most climate models, outputs a large number variables with a diversity of characteristics and varying importance to scientists, further complicat-

ing analysis. Efforts specific to evaluating the impact of lossy compression on CESM data thus far include developing measures that compare lossy compression-induced error to the internal variability of the climate model [BXD*14], engaging climate scientists in a blind study of a climate ensemble consisting of both reconstructed (data that has undergone compression and reconstruction) and original data [BHM*16], and comparing distinct types of lossy approaches on individual CESM variables via a suite of test measurements [BXH*17].

While post-processing analysis of CESM data takes a variety of forms, visual assessments are ubiquitous in post-processing workflows. In fact, visualizations as diagnostics are quite important to climate scientists and typically provide their first interaction with the simulation output data. Visualizations generated by the Atmosphere Working Group Diagnostics Package (AMWG-DP) and the Climate Variability Diagnostic Package (CVDP) [PDF14] are particularly popular with climate scientists and are typically included with the public releases of large CESM simulation data sets, such as the CESM Large Ensemble Community Project [KDP*15]. As an example, the AMWG-DP generates on the order of 1300 diagnostic images for a typical CESM simulation, and this number only represents images for the atmospheric model component (additional diagnostic packages exist for the other CESM components, such as the land, ocean, and sea-ice components). CESM scientists often view these diagnostic images from new simulations right away so as to verify that the data looks reasonable and/or meets their expectations in some sense before proceeding with further analysis. The work in [BHM*16] suggests that a goal in responsibly applying lossy compression to output data from CESM is that the reconstructed and original data be indistinguishable during analyses. Given the importance of diagnostic images to most climate scientists (particularly when initially engaging with a new dataset), and more generally, the importance of visualization in climate research, it seems reasonable to require that the diagnostic images generated not be noticeably different. Indeed, by providing assurance that the loss of information due to data compression does not negatively affect the diagnostic images, we aim to reduce the climate modeling community's hesitancy to fully adopting lossy data compression (e.g., see [BHM*16]).

In the recent work [BXH*17], the suite of tests used to compare lossy compression algorithms on CESM data includes a single image quality assessment (IQA) measure: the structural similarity index (SSIM) [WBSS04]. The authors in [BXH*17] indicate that the SSIM threshold chosen was largely inspired by suggestions from the medical imaging field, where lossy compression is typically used to reduce unmanageable data volumes (e.g., [CP16, KR12, Wan11, KS06]). Indeed, the loss of critical information (a concern shared by climate scientists) is understandably a concern with procedures such as computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound imaging (e.g., [KBMG13, KMLH10, GKS*12b]). However, an acceptable SSIM threshold varies by application, and the authors in [BXH*17] acknowledge that further research is required to confidently select one for CESM images. Beyond the selection of an application-appropriate threshold, we add that further research should also be undertaken to determine whether SSIM is really the most useful IQA measure in this context.

In this work, we address the selection of an appropriate IQA (and corresponding threshold) that can be used to indicate whether climate scientists would be able to detect a difference in CESM diagnostic images after lossy compression. We note that because the original data is used in other proposed quality measures, we limit our investigation to full-reference IQA measures. We design a visual evaluation study to determine when scientists start to perceive visual differences between the original and reconstructed images. We then evaluate a number of popular IQA algorithms in the context of the scores from the user study to determine which is most applicable to the CESM model data images. It is important to note that in choosing an IQA measure and threshold, we are not attempting to determine whether or not the difference in images matters in terms of drawing conclusions from the climate simulation data, but rather answering the easier – and more conservative – question as to whether *any difference* between images is noticeable.

This paper is organized as follows. In Section 2, we introduce the IQA measures that we evaluate. In Section 3, we describe the particular CESM data that we include in the study as well as the chosen lossy data compressors. In Section 4, we describe the setup of the user study. We present the results of our analysis in Section 5 and provide concluding remarks in Section 6.

2. Image quality measures

Our interest is in so-called full reference (FR) IQA measures, which require the original, or reference, image for comparison. In this study, the reference image is the CESM image generated with the original (unaltered) model output data and is readily available. The altered image is the CESM image generated from the reconstructed data (data that has undergone lossy compression, followed by decompression). While the mean-squared error (MSE), or the related peak signal-to-noise ratio (PSNR), have long been used to evaluate image quality, recent decades have seen the introduction of a number of arguably more comprehensive FR-IQA methods that enjoy wide popularity, such as the SSIM (e.g., [WBSS04, WB09, Wan11]).

Table 1 contains the IQA methods that we evaluate. While our list of IQA measures is not exhaustive, we chose a variety that includes several different types of approaches. MATLAB® implementations for all methods listed in Table 1 were either publicly available from the internet (e.g., the authors' website) or, in the case of SSIM, PSNR, and MSE, available from the MATLAB Image Processing Toolkit™ (IPT).

We include the simple mean-square error (MSE) and peak signal-to-noise (PSNR) measures, as these approaches are still popular in many application areas. Additionally, we calculate the normalized absolute error (NAE), which is simply the sum of the absolute errors at each location, normalized by the values in the original image. We note that PSNR, MSE, and NAE all use the scaled images resulting from MATLAB's *im2double()* function. Our measures also include the popular SSIM method [WBSS04], which evaluates the image structure, and its variant multi-scale SSIM (MS-SSIM) [WSB03], which is designed to evaluate structures at different scales. The visual information fidelity (VIF) [SB06] index represents a different type of approach that uses an information theory framework to quantify how much information is

Method	Description	Possible values	Identical value
MSE	mean-squared error (e.g., as in [WB09])	≥ 0	0
PSNR	peak signal-to-noise ratio (e.g., as in [WB09])	≥ 0	∞
NAE	normalized absolute error	≥ 0	0
SSIM	structural similarity index [WBSS04]	$[0, 1]$	1
MS-SSIM	multi-scale SSIM [WSB03]	$[0, 1]$	1
VIF	visual information fidelity measurement [SB06]	$[0, 1]$	1
MAD	most apparent distortion [LC10]	≥ 0	0
FSIM	feature similarity index [ZZMZ11]	$[0, 1]$	1
GMSD	gradient magnitude similarity deviation [XZMB14]	≥ 0	0
NLP-dist	normalized Laplacian pyramid distance [LBBS16]	$[0, 1]$	0

Table 1: *A list of the IQA measures evaluated in this study, followed by a description, the range of possible values, and the value that indicates that the images are identical.*

preserved between images. In addition, the most apparent distortion (MAD) [LC10] method is interesting as it uses a combination of two different model strategies to evaluate quality: detection-based and appearance-based. We also evaluate the feature similarity index (FSIM) [ZZMZ11], which uses gradient information to examine low-level features, as well as the more recent gradient magnitude similarity deviation (GMSD) [XZMB14] approach, which also focuses on local gradient similarities. Finally, we include the recently developed normalized Laplacian pyramid distance (NLP-dist), which is essentially a root mean-square error (RMSE) in a multi-scale decomposition or "normalized Laplacian" domain [LBBS16]. We note that our primary interest in this evaluation is to find an IQA that most agrees with the perceptual scores from our user study (described in the following section) on CESM data, and thus we ignore the cost of applying the IQA measures at this time.

3. Data and compression methods

To determine which of the objective IQA measures described in the previous section are most consistent with visual evaluations by climate scientists and other domain relevant experts, we designed a two-alternative forced choice experiment (Section 4). In that experiment, participants with domain relevant expertise compare reference and reconstructed CESM diagnostic images to determine when differences due to compression are visible. In this section, we explain our choices for which CESM data to evaluate as well as the lossy compression methods applied.

3.1. CESM version, setup, and variables

Data for this study was obtained from the CESM 1.1 series public release. We investigate output data from the atmospheric component (the Community Atmosphere Model, or CAM5) using a spectral-element dynamical core on a grid with a resolution of approximately 1-degree (i.e., $ne = 30$). The global grid is a cubed-sphere, with 48,602 horizonal grid points output as a 1D array and 30 vertical levels. Note that while CESM computations are performed in double-precision, data for the CESM history files (time-slices) are truncated to single-precision when the NetCDF format output file is written. As this work is motivated by the use of SSIM in [BXH*17] to evaluate compression artifacts in CESM images, we use the same CESM version and setup as in that study, the same compression methods (Section 3.2), as well as a subset of the so-called representative variables examined in detail in [BXH*17]. In particular, we chose the following four variables in our study:

- TS: surface temperature (2D)
- FSNTC: clear sky net solar flux at the top of model (2D)
- NUMLIQ: grid box averaged cloud liquid number (3D)
- PRECCDZM: convective precipitation rate (2D)

These particular four variables were chosen from the variables in [BXH*17] primarily because (1) they are quite different in their characteristics, and (2) their behavior varies under the two types of compression used in this study. TS is a variable of interest to nearly all climate study disciplines, in particular, atmospheric and ocean modellers. TS diagnostics are almost always examined in a first look at new simulation output. Further, because TS data are relatively smooth, with a modest dynamic range of values, and no zeros, TS can be compressed relatively easily and effectively by most compression approaches. In contrast, NUMLIQ data are nearly half zeros and have a very large dynamic range (20 orders of magnitude), which makes NUMLIQ difficult to compress for the two compression methods applied (only lossless compression passed the measures proposed in [BXH*17]). Note that because NUMLIQ is a 3D variable, we only include an image from a single level in the study. Precipitation data is also typically included in a first look at simulation data, and thus we included PRECCDZM as it is one of two precipitation-related variables studied in [BXH*17] (the other, PRECSC, contains more than 75% zeros and is challenging for one of the compression approaches in the study). Additionally, both PRECCDZM and NUMLIQ contain many zero or near zero values, which leads to more white space in the visualization (an attribute that potentially affects how easy it is to notice differences). Finally, we include the FSNTC variable as top of the model fluxes are of interest for initial energy balance concerns. FSNTC is more easily compressed than both NUMLIQ and PRECCDZM, but more difficult than TS.

The visualizations of each of the four uncompressed variables can be seen in the top panels of Figures 1, 2, 3, and 4, which also contain a comparison visualization(s) from compressed data (to be discussed in later sections). To ensure that the study images and colormaps felt familiar to those typically evaluated by CESM scientists, the images were created with the NCAR Command Lan-

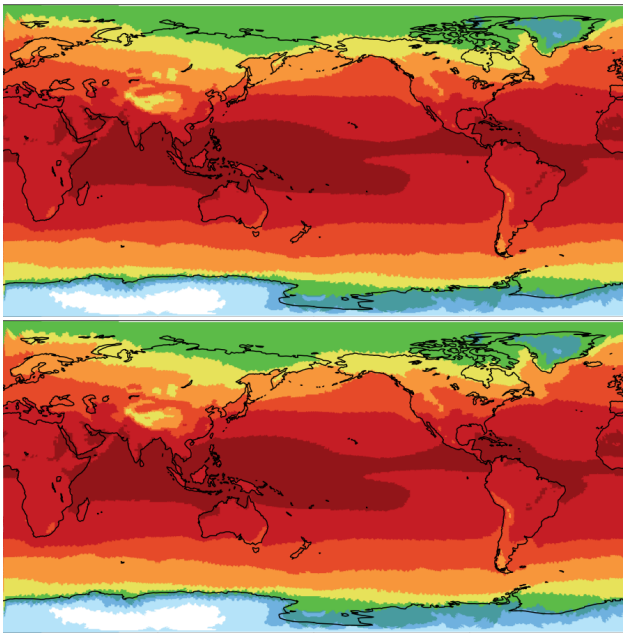

Figure 1: *The surface temperature variable, TS, is visualized for the uncompressed data (top) and for speck_4 compression (bottom). User study responses for this example were fairly evenly split with about 60% seeing no difference and 40% seeing a difference. Note that axes and legends have been removed to allow participants to focus on the visualizations.*

guage (NCL) [UCA17]. NCL is a popular post-processing utility for the Earth science community (and CESM scientists in particular) and is used to create the images in the aforementioned CVDP and AMWG-DP post-processing tools (see Section 1). As will be emphasized in the next section, our interest in this work is in detecting differences in the visualizations (and not in quantifying values), and for that reason, we did not include any axes, text, or colormaps in the figures that would impact the visual assessment. Also note that because the dynamic ranges of the four variables are different, the colormaps are different for each variable (but kept fixed for all images from the reconstructed data for each variable). For reference, all visualizations included in the user study can be found in the supplemental material.

3.2. Lossy compression approaches

As volumes of floating-point scientific data have exploded across scientific modeling disciplines in recent years, lossy compression methods are increasingly receiving more attention, and a number of state-of-the art methods have been developed. Transform methods are a popular type of lossy approach, modeling the data with wavelets or discrete cosine transforms, for example, and reducing the data size by retaining only a subset of the transform coefficients. Such approaches include the well-known JPEG2000, the more recent ZFP [Lin14] compressor, and SPECK (a discrete wavelet tranform with the set partitioned embedded block coder algorithm) [IP98]. The so-called predictive lossy approaches are widely used

Method	TS	FSNTC	NUMLIQ	PRECCDZM
fpzip_12	.01	.03	.11	.11
fpzip_16	.04	.10	.15	.21
fpzip_20	.15	.21	.21	.32
fpzip_24	.28	.32	.28	.42
fpzip_28	.35	.43	.34	.53
fpzip_32	.46	.54	.40	.64
speck_1	.03	.03	.03	.03
speck_2	.06	.06	.06	.06
speck_4	.11	.11	.11	.11
speck_8	.22	.22	.22	.22
speck_12	.34	.34	.34	.34
speck_24	.67	.67	.67	.67
speck_32	.90	.90	.90	.90

Table 2: *A list of the compression ratios corresponding to the four variables used in the study images: TS (surface temperature), FSNTC (clear sky net solar flux), NUMLIQ (averaged cloud liquid number), and PRECCDZM (convective precipitation rate). Note that because* SPECK *is a fixed-rate method, the compression ratios (CRs) are equivalent across variable type.*

as well. As the name implies, these approaches traverse the data and predict upcoming data values based on previously visited values, typically retaining the residual between the predicted and actual data value. For example, the SZ compressor ([DC16, TDCC17]) is a predictive method that uses adaptive error-controlled quantization, and the well-used FPZIP compressor (which can be lossy or lossless, depending upon whether any least significant bits are discarded) encodes residual values with a fast entropy encoder.

Again, given that our motivation for this study comes primarily from the work in [BXH*17], we use the same compression methods in this study: FPZIP [LI06] and SPECK [IP98]. To indicate the amount of compression applied by each approach, we follow the naming convention defined in [BXH*17]. In particular, the SPECK compressor is a fixed-rate method, and we use the notation speck_B to indicate the bit target rate by B. We use $B \in \{1, 2, 4, 8, 12, 24, 32\}$, noting that speck_1 is the most aggressive and speck_32 is the least. As in the work in [BXH*17], we map the 1D array of grid points output by CESM (in space-filling curve ordering) onto the six 2D faces of the cubed-sphere and apply compression to each face separately. This mapping step improves the spatial coherence of the data (and increases its dimension to 2D to match a natural latitude-longitude ordering), which greatly improves the transform method's compression ratios. For the FPZIP compressor, we use fpzip_N to denote FPZIP, where N is the number of bits retained before quantization. In this study, we use $N \in \{12, 16, 20, 24, 28, 32\}$, noting that fpzip_32 is lossless as we are compressing single-precision data.

Table 2 contains the compression ratio (CR) for each of the four CESM variables for each compressor variant. We define the CR as the ratio of the size of the compressed file to that of the original file, meaning that a smaller number indicates greater compression.

4. User study

In image quality measures, subjective evaluation – in the form of mean opinion scores (MOS) – is often considered the ground truth [SWH16, SSB06] to measure the effect of a "deformation" or change to an image. In this work, any changes to the images are due to increasing levels of compression, and the classic psychophysical two-alternative forced-choice approach (2AFC) [CW12] can be used to provide an objective evaluation of the impact for each level of compression.

In this difference-threshold experiment, the research question asks which standard image quality measure most closely models the ability of domain relevant experts to see a difference in climate data visualization due to lossy compression. Participants were shown two images simultaneously – the uncompressed baseline image as the standard stimulus and a test stimulus chosen from the set of images derived from the independent study variables, the two different types of compression (SPECK and FPZIP) and the multiple levels of compression for each compression type. The dependent variable was the *two alternative choices* in this experimental design: to see a difference or to not see a difference. The experiment used a between-subject approach where each participant saw a limited number of trials to minimize any learning effect and to ensure the study was short enough to encourage participation.

4.1. Method

The experimental stimuli images mirrored the images used to calculate the image quality measures. These images were cropped to remove text, axes, and color map. Because we are only interested in whether participants can detect a difference, not in quantifying variables, removing the axes and color map allowed participants to focus on the visual differences. The study was implemented using Qualtrics survey software, and the two stimuli images were presented vertically using a modified version of the 2AFC module from the Evaluation Toolkit [TBRA17]. Each image was 890 px in width and 450 px in height. The study included a screen size check to ensure participants could see both images on the screen at the same time. A hidden check also prevented participants from using a mobile device.

A brief introduction included a description of the study and a series of example images highlighting subtle differences due to compression. The example images were FSNTC images that had been further cropped (to minimize potential learning effects). Each participant saw a randomized subset of 16 comparisons drawn from the possible compression types and variables. In order to avoid participant learning effects, each participant saw only four images from each variable type. To ensure that each participant would see significant compression effects for each variable, one of these four images (for each variable) was from speck_1 compression, which is very aggressive. The other three images were then randomly selected from the remaining compression levels, which were divided into *high* and *low* compression categories as shown in Table 3, with two images randomly selected from the low compression category and one randomly chosen image from the high compression category. This randomization approach increased the number of trials for the low compression levels, where differences were more difficult to detect, while still showing participants at least a few comparisons with obvious differences.

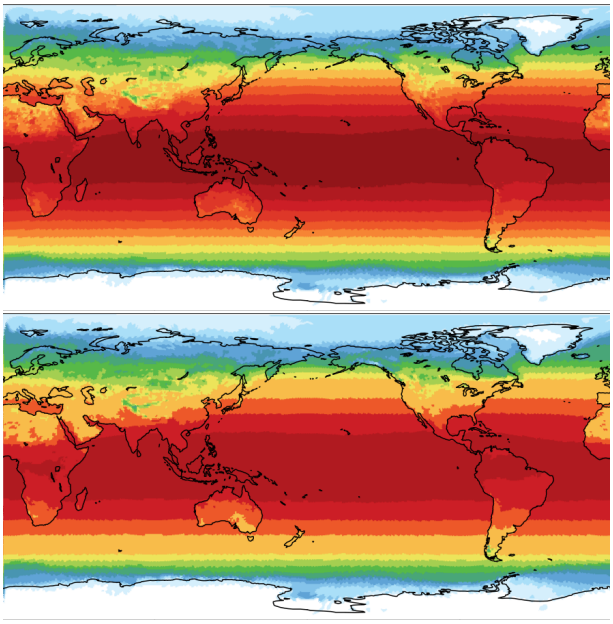

Figure 2: *The net solar flux variable, FSNTC, is visualized for the uncompressed data (top) and for fpzip_12 compression (bottom). User study responses for this example were fairly uniform with about 96% seeing a difference. Note that axes and legends have been removed to allow participants to focus on the visualizations.*

High Compression		Low Compression	
speck_2	fpzip_12	speck_8	fpzip_16
speck_4		speck_12	fpzip_20
		speck_24	fpzip_24
		speck_32	fpzip_28
			fpzip_32

Table 3: *The high and low compression categories used to determine the randomized set of image comparisons seen by each participant. One image was chosen from the high compression set and two from the low compression set. Note that speck_1 (which would be considered high compression) is not listed as it was seen by all participants for each variable.*

The study question was given as: *You will see a series of comparisons of two images, one of which is uncompressed. Your task will be to decide if they appear IDENTICAL or DIFFERENT.* Since the images were arranged vertically, the two button options were arranged horizontally to avoid biasing the viewer. IDENTICAL was the left button; DIFFERENT was the right button. In addition to the button choice, the amount of time spent on each stimuli set was saved.

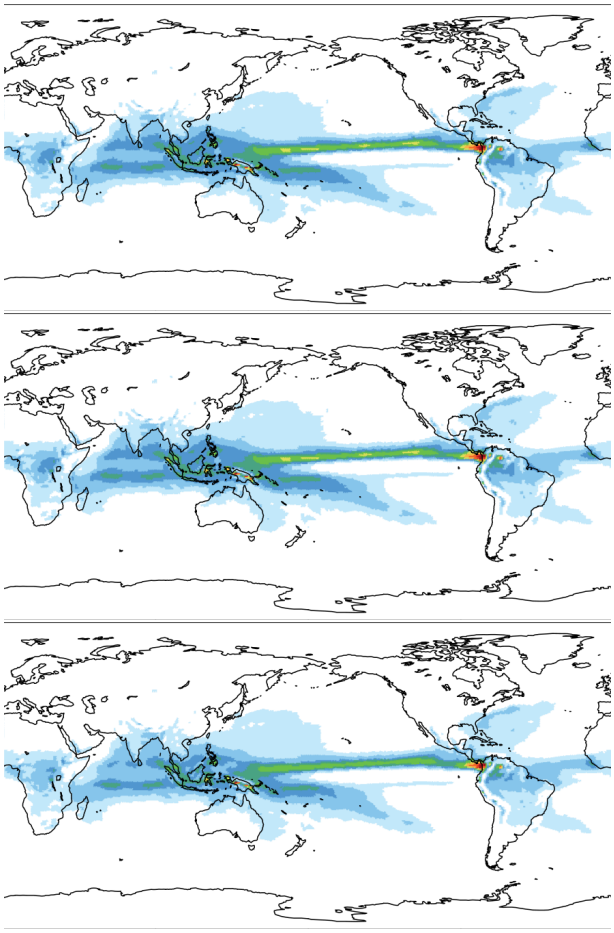

Figure 3: *Three images are shown for PRECCDZM (convective precipitation rate). The top is the visualization from the uncompressed data. The middle image has the speck_4 compression applied and the bottom has fpzip_12 applied. Despite the two "high compression" approaches (speck_4 and fpzip_12) resulting in nearly identical compression ratios, there are a number of obvious differences particularly near the equator. In this example, the speck_4 image is closer to the original (e.g., see Table 5), and 42% of the study participants noticed a difference. For fpzip_12, 100% of the study participants noted a difference.*

4.2. Participants

In order to ensure maximum engagement on the part of participants, the participant pool was solicited (through email) from the ranks of people familiar with and comfortable viewing scientific data. Participants included people with backgrounds in climate science, computer science, biological and physical sciences, statistics, and mathematics. A plurality (45%) had backgrounds in climate-relevant sciences. Participants were assigned a random identification number. In all, there were 113 participants with 85 male (72.2%), 27 female (23.9%), and one declining to answer. The participants ranged in age from 19 to 67 with a median age of 41 for those choosing to answer the age question. There were three participants (2.7%) without a college degree (students). The majority (59.3%) of participants had a doctoral or other professional degree and another 26.5% had a master's level degree.

4.3. Visual Acuity Checks

When doing online presentation of studies where color is critical, it is important to minimize the potential for contamination of the subject pool by participants with visual acuity issues or color vision deficiencies (CVD). In this study, this was done both via self-identification and through online testing for CVD. Each participant was asked if they had any visual acuity issues or color vision deficiencies and, if so, was asked to describe those issues. Any participant self-identifying with CVD or visual acuity issues was removed from the participant pool, regardless of their result on the CVD test. Each subject completed an online presentation of the Farnsworth D-15 (FD-15) color cap arrangement test for CVD [CJJ93]. In the FD-15, a subject is required to order a standard set of 15 color patches. While each type of CVD has somewhat typical presentations of incorrectly ordered results, the spectrum of CVD can result in a wide range of errors. Taking a conservative approach, any participant whose FD-15 ordering included more than one set of flipped patches was excluded from the participant pool. The set of people self-identifying with CVD or a visual acuity issue or failing the CVD test was 17 people (15.0%). While this is larger than might be expected from the general populace, it does reflect a quite conservative approach to minimizing CVD effects. The median time spent on the full study was 16 minutes.

4.4. Study data

Figures 1 and 2 show example visualizations for TS and FSNTC, respectively. For TS in Figure 1, study participants were fairly evenly split as to whether the images were identical or not. For FSNTC in Figure 2, nearly all study participants noticed a difference. For reference, we include Table 5, which lists the IQA measure values (in the same order as in Table 1) for each image and compression level, followed by the user study responses for each image. Before discussing our modeling and analysis in the next section, we note that many interesting observations can be made by visual inspection of the raw data in Table 5 together with the *CR* values from Table 2 (reproduced in Table 5), and we discuss a few here.

First, we observe that speck_1 is quite aggressive; its $CR = .03$ represents a roughly 33x reduction in the size of the data. (Recall that we specifically chose to show speck_1 to each participant for each variable.) The fpzip_12 approach is similarly aggressive (in terms of *CR*) on variables TS and FNSTC (which are the two "easier-to-compress" variables), and the user study results indicate that the resulting images were obviously different as all but a single participant noted the differences for these two variables (with speck_1 and fpzip_12). On the other hand, variables PRECCDZM and NUMLIQ are compressed by speck_1 about 4x more aggressively than by fpzip_12. But despite fpzip_12's more conservative rate, all study participants correctly identified differences in the resulting images. However, with speck_1, differences in the images were not uniformly identified by participants, particularly for NUMLIQ. This result for the images is somewhat in conflict

with the assertion made in [BXH*17] that transform methods, such as SPECK, do not work very well on hard-to-compress variables such as NUMLIQ. Certainly for the visual analysis, SPECK leaves fewer visible artifacts in this case. In particular, for NUMLIQ, both speck_4 and fpzip_12 have the same $CR = .11$, but all study participants noted a visual difference with fpzip_12 and no participants noted a difference with speck_4. We note that the work in [BXH*17] discusses the strengths and weaknesses of these two distinct lossy compression approaches (SPECK and FPZIP) on CESM variables with differing characteristics.

The PRECCDZM variable additionally highlights how differently the two types of lossy approaches can affect the visualization. Consider the images for speck_4 and fpzip_12 in Figure 3, both of which are categorized as "high compression" in Table 3 and have the same $CR = .11$. While the speck_4 image (middle) has few minor features that differ from the original image (top), the fpzip_12 image (bottom) has quite obvious differences.

Another point to note is the difference in how quickly FPZIP reaches the "ideal" IQA measure value relative to SPECK for the two easy-to-compress variables, TS and FNSTC – not the case for the remaining "difficult" variables. The reason for this is that FPZIP performs lossy compression by discarding least significant bits. For variables TS and FNSTC, which are smooth and have modest dynamic ranges, the least significant bits are truly unimportant (and are likely small-scale noise). Even though transform methods are known to work well on smooth data, they achieve lossy compression by eliminating coefficients of the least important basis functions. It follows then that this strategy affects more than just the most aggressive levels of SPECK compression. We do not mean to imply by this observation that the transform approach is inferior, but simply that it has different implications for the visualization. In contrast, the transform approach appears superior in the context of the previous discussion of the PRECCDZM and NUMLIQ variables.

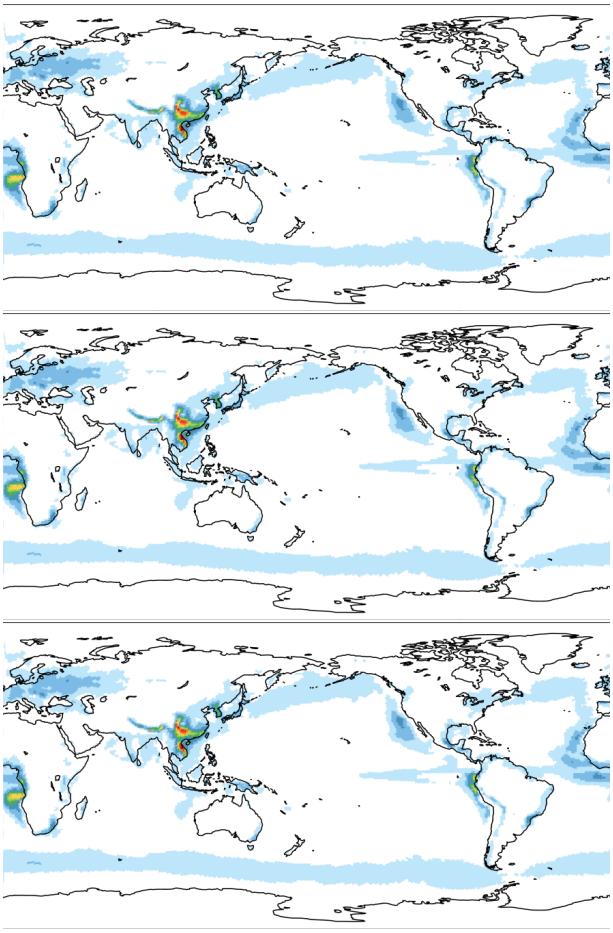

Figure 4: *Example visualizations are shown for NUMLIQ, the grid box averaged cloud liquid number. The top image is from the uncompressed data. The middle image has the fpzip_20 compression applied to the data, and no study participants noticed the difference. The bottom image has speck_8 applied to the data, and while the resulting image is equivalent to the original image, one participant noted a difference.*

5. Analysis of survey data

Our goal is to identify which of the IQA measures best describes the data from the user evaluation study. For practical considerations in an operational implementation, we are looking to use a single measure, rather than a combination. Once identified, we can then determine a threshold best-suited for this measure to ensure that visual renderings of the compressed climate model data are indistinguishable from renderings of the original data. The selected measure (and its threshold) can then be incorporated in the suite of measures used to safeguard against compression negatively affecting the scientific integrity of the climate model output.

5.1. Methods

To model the relationship between the responses by the participants and each of the candidate measures, we employ generalized linear model regression (see [MN89] for an overview). We model the responses shown in the last two column of Table 5 as coming from a binomial distribution, which provides the flexibility to model proportions rather than direct counts. This way we can incorporate the varying sample sizes for each image resulting from our randomized setup. When modeling proportions, the response should be in the range from 0 to 1, which is achieved by using specific link functions in the regression setup. The canonical link function for the binomial distributions is the logit function, which we test in addition to the probit and complementary log-log link functions. A detailed discussion of link functions for the binomial distribution can be found in [Col02].

Given the size and range of the measures, we model each predictor using its original values as well as a log-transformed version. (Note that because MSE, PSNR, and NAE are highly correlated, we limit our evaluation to NAE.) For consistency between measures, we subtract the measures having 1 as their identical value (see Table 1) from 1. For the log-transformed version, we also add a small number, namely the closest power of 10 below the lowest value occurring in the data, to avoid numerical issues with the log

	logit	probit	comp log-log
NAE	57.13	61.24	111.84
SSIM	34.61	37.60	62.28
MS-SSIM	39.99	42.07	112.15
VIF	38.59	44.72	39.23
MAD	63.31	67.93	85.27
FSIM	42.61	45.57	76.88
GMSD	37.78	46.66	55.78
NLP-dist	37.68	44.31	76.11

Table 4: *Deviance as a function of predictor and link function.*

being undefined at zero. Deviance is used as the main criterion for model fit. Alternatively, we also evaluate the Anscombe residuals (discussed in Chapter 5 in [Col02]) as a secondary means of model diagnostics. The only form of quality control we perform on the participant data is imposing a forced zero for any "DIFFERENT" values for those images where the sum of the absolute differences between the original and the compressed images is zero. Clearly, if there is truly no difference and all measures are accordingly, and correctly, at their identical value, it is not sensible to try to model those images as anything other than identical.

5.2. Results

The models based on the log-transformed measures perform universally better, and we will from here onwards limit our discussion to the model results based on the log-transformed predictors. Deviance and Anscombe residuals led to the exact same ranking between measures and within link function choices. For succinctness, only the deviance results are shown in Table 4. The best fitting model, corresponding to the lowest deviance value, is using the log-transformed SSIM as a predictor and the logit link function. A visualization of the fit for the model using SSIM for all three link functions along with the observed responses are shown in Figure 5. For comparison, the visualization of the fit for the model using MAD, which corresponds to the highest deviance value (for the logit link function), is shown in Figure 6. (Images of the fits for all the IQA measures can be found in the supplemental material.) We note that neither the MAD measure nor the others are unreasonable, but the SSIM performs the best with regard to our study data. While it would be interesting (and useful) to understand the underlying reason as to why SSIM performs better than the others, we do not have a hypothesis at this time.

Given these results, it is interesting to note that the SSIM threshold used in [BXH*17] was chosen as 0.98 based on suggestions from the medical literature (e.g, [GKS*12a], [Weg10]). Using a SSIM threshold of 0.98, the estimated proportion different for this threshold value is 0.9588 (0.9740, 0.9354), where the values in parenthesis provide the 95% confidence bounds for the estimate. Given that our model indicates that almost every scientist viewing data compressed at this threshold would perceive a difference, the threshold used by [BXH*17] appears too lenient to be of practical use. While the exact determination of a threshold warrants further study, a SSIM threshold value on the order of 0.99995 corresponding to an estimate of proportion different of 0.0100 (0.0237, 0.0042) seems a more appropriate choice for climate model data.

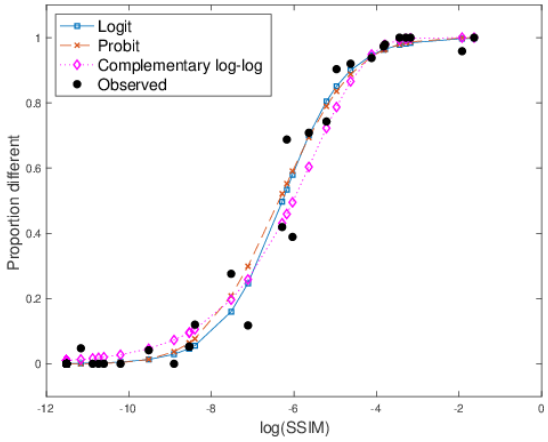

Figure 5: *Best fitting model using log-transformed SSIM as a predictor and the logit link function.*

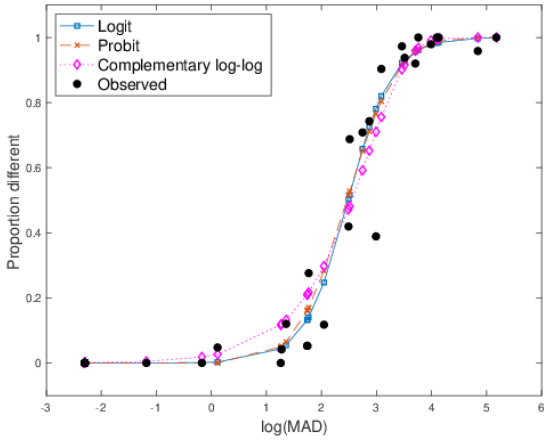

Figure 6: *Worst fitting model using log-transformed MAD as a predictor and the logit link function.*

6. Concluding remarks and future work

In this manuscript, we describe the results of a large-scale user study conducted with domain relevant participants to evaluate a number of well-known image quality measures in the context of lossy compressed climate model data. While all the measures that we evaluate show reasonable predictive ability to gauge when expert users perceive differences, the SSIM IQA measure performs best. It is interesting to note that the SSIM values at which scientific users perceive differences are much higher than what was previously found in the literature. Consider an SSIM value of 0.98, which appears to be considered a normal to conservative threshold in the medical literature. However, virtually all scientists in our study perceive a difference in climate model output. Our findings indicate that using an SSIM threshold on the order of 0.99995

might be required for climate model output to ensure visually identical data.

For lossy compression to be accepted and adopted by the CESM modeling community, users must have confidence in the data. While ensuring that visualizations in the post-processing workflow are not noticeably altered may seem quite conservative, this step should positively contribute to user confidence. However, it is important to note that "passing" a visual check alone, even with a full reference IQA, is not sufficient to determine that a particular compression type and level is acceptable for a certain climate model variable. For example, consider the NUMLIQ variable images, shown in Figure 4. Both speck_8 and fpzip_20 had a similar *CR* value (.22 and .21, respectively), and their corresponding images are nearly identical in terms of the IQA measures (e.g., Table 5) – in fact, the speck_8 image is identical to the original. However, the authors in [BXH*17] classify NUMLIQ as a variable that requires lossless compression due to its failure on other (non-visual) measures. Indeed, we advocate that a visual measure be part of a comprehensive suite of well-designed measures that can be shown to detect problematic data loss due to lossy compression, particularly given that visualizations are ubiquitous in post-processing workflows of climate simulation data. Because climate scientists have shown reluctance to embrace lossy compression (e.g., [BHM*16]), such a suite should bolster CESM users' confidence and willingness to use lossy compression.

The methodology used in this paper can be applied across a wide range of data types, compression approaches, and potentially be improved by considering other color maps. However, this work did require certain design choices. We chose a single climate data set and four variables. The variables were chosen to represent common variables as well as variables that are challenging to compress (e.g., cloud liquid and precipitation), and to explore both fields with many zeros (so less happening visually) and fields with no zeros (surface temperature). Therefore, we expect that the threshold and measure recommended would be appropriate for most of the CESM atmospheric model output variables (as with the other non-visual measures recommended in [BXH*17]). Further, the two lossy compression algorithms used in this study (SPECK and FPZIP) are quite different, but represent the two most commonly used categories of lossy compression approaches: transform and predictive methods. Because our IQA measure (and threshold) recommendations are applicable to both of these disparate approaches, we have no reason to believe that they would not be appropriate for any lossy method belonging to these two popular categories.

We made the choice to use the rainbow color map because of its familiarity to climate and computational scientists (domain relevance). We acknowledge the perceptual limitations of the rainbow color map. However, a cursory count of the data sets in the NCAR Climate Data Guide [NCA] still has a majority of thumbnails using a rainbow-based color map, and the color map used in the study images represents typical output for diagnostic images produced by CESM tools. That said, the question of color map choice is certainly an avenue for future work. Using a more perceptually uniform color map or one with higher discriminative power may well push the threshold for visually identical data to a higher level.

Another design choice was to use participants with domain relevant expertise. Although we were not attempting to have the participants complete a domain relevant task (such as in [DPR*18]), we did want to ensure engagement on the part of participants. Participants were not asked to rate their level of comfort with scientific visualizations or specific level of expertise, but the high percentage of participants with advanced degrees or who cited climate science in their background gives us confidence in a sufficient level of engagement with the comparison task that was presented. Note that this relatively narrow study (one color map, four variables, two compression types, 6 or 7 levels of compression) required over 30 hours of participant time. Using only domain relevant experts may be a prohibitive requirement to extending this work to a wider range of color maps, data domains, compression approaches, or variable types. However, a widening body of literature using crowdsourced participant pools for visualization tasks (e.g., [HB10] or [WTB*18]) indicates that a simple task, such as the identical/different comparison in this paper, may be appropriate for crowdsourcing, allowing interested researchers to expand and generalize this work to their particular areas of interest.

Our work here indicates that the SSIM metric is well-suited for the type of diagnostic images commonly evaluated by climate scientists. As far as applicability beyond climate data, the higher threshold recommendation for SSIM may be desirable for other types of model simulation data where visualization is important in post-processing workflows and any detectable artifacts from compression are undesirable. We note that this criteria is more stringent than, for example, the "diagnostically lossless" requirement (see, for example, [SoRE11, KUBW*14]) often desired in the medical literature, where the goal is to reach the same diagnosis/conclusion from the data – not necessarily avoid any visual differences. Further research is needed to determine whether different types of simulation data would require different IQA measure thresholds for evaluating compression and if generalizations can be drawn.

Acknowledgements

We thank Haiying Xu for generating the compressed data and plots. We thank Kevin Raedar for valuable feedback on experimental design and express our gratitude to everyone who took the time to take this study. This research used computing resources provided by the Climate Simulation Laboratory at NCAR's Computational and Information Systems Laboratory (CISL), sponsored by the National Science Foundation and other agencies. The user evaluation study was conducted under a subcontract from the National Center for Atmospheric Research to the University of Texas at Austin. This research is released under LA-UR-19-22420.

Variable and Compression	CR	IQA metrics MSE	PSNR	NAE	SSIM	MS-SSIM	VIF	MAD	FSIM	GMSD	NLP-dist	User study responses Identical	Different
TS													
fpzip_12	0.01	5.92E-02	12.2796	1.48E-03	0.8045411	0.4470091	0.4081	177.492	0.7299792	0.4959	0.3188	0	35
fpzip_16	0.04	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	14	0
fpzip_20	0.15	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	22	0
fpzip_24	0.28	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	17	0
fpzip_28	0.35	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	27	0
fpzip_32	0.46	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	29	0
speck_1	0.03	3.71E-4	35.3051	1.70E-5	0.9774348	0.9892419	0.7502	53.961	0.9736993	0.0771	0.1201	2	94
speck_2	0.06	1.56E-4	38.0646	6.93E-6	0.9902174	0.9954218	0.8776	40.630	0.9875604	0.0512	0.0787	2	23
speck_4	0.11	2.76E-5	45.5900	1.51E-6	0.9976131	0.9989622	0.9649	19.818	0.9976458	0.0205	0.0343	22	14
speck_8	0.22	3.65E-6	54.3712	1.55E-7	0.9998129	0.9999103	0.9959	5.594	0.9998766	0.0033	0.0093	18	1
speck_12	0.34	3.65E-8	72.1084	6.05E-9	0.9999958	0.9999974	0.9999	1.019	0.9999990	0.0004	0.0012	20	1
speck_24	0.67	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	20	0
speck_32	0.90	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	19	0
FNSTC													
fpzip_12	0.03	4.46E-3	23.5102	2.27E-4	0.8541187	0.8254689	0.5066	126.646	0.8064006	0.3666	0.2293	1	23
fpzip_16	0.10	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	17	0
fpzip_20	0.21	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	21	0
fpzip_24	0.32	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	17	0
fpzip_28	0.43	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	22	0
fpzip_32	0.54	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	24	0
speck_1	0.03	4.00E-4	33.9773	2.64E-5	0.9582047	0.9847898	0.6581	42.900	0.964863	0.0896	0.1337	0	96
speck_2	0.06	2.13E-4	36.7068	1.40E-5	0.9782505	0.9923561	0.7791	31.843	0.9815798	0.0662	0.0992	1	36
speck_4	0.11	4.97E-5	43.0371	3.32E-6	0.9945575	0.9981294	0.9338	17.556	0.9949551	0.0341	0.0469	9	26
speck_8	0.22	1.20E-6	59.1964	9.98E-8	0.9997848	0.9999387	0.9967	3.792	0.9998806	0.0052	0.0085	22	3
speck_12	0.34	1.74E-7	67.5952	1.49E-8	0.9999882	0.9999982	0.9994	0	0.9999852	0.0011	0.0019	17	0
speck_24	0.67	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	22	0
speck_32	0.90	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	25	0
NUMLIQ													
fpzip_12	0.11	4.62E-4	33.3518	1.27E-5	0.9626904	0.9782720	0.7345	60.020	0.9572827	0.1066	0.1444	0	36
fpzip_16	0.15	2.25E-6	46.4727	6.57E-7	0.9979233	0.9992209	0.9640	12.249	0.9980807	0.0195	0.0316	5	11
fpzip_20	0.21	1.91E-6	57.1824	5.32E-8	0.9998742	0.9999561	0.9960	3.430	0.9998585	0.0037	0.0086	24	0
fpzip_24	0.28	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	26	0
fpzip_28	0.34	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	25	0
fpzip_32	0.40	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	19	0
speck_1	0.03	3.46E-5	44.6078	1.03E-6	0.9964475	0.9986568	0.9466	15.504	0.9970395	0.0250	0.0386	28	68
speck_2	0.06	5.64E-6	52.4886	1.74E-7	0.9994682	0.9998267	0.9883	5.763	0.9996166	0.0092	0.0151	21	8
speck_4	0.11	3.01E-7	65.2190	1.06E-8	0.9999911	0.9999974	0.9991	0.742	0.9999880	0.0013	0.0028	31	0
speck_8	0.22	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	22	1
speck_12	0.34	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	11	0
speck_24	0.67	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	17	0
speck_32	0.90	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	26	0
PRECCDZM													
fpzip_12	0.11	5.07E-4	32.9426	1.46E-5	0.9679866	0.9810663	0.7952	62.252	0.9559590	0.1063	0.1413	0	34
fpzip_16	0.21	8.78E-6	50.5638	2.96E-7	0.9991952	0.9997103	0.9847	7.643	0.9992886	0.0094	0.0192	15	2
fpzip_20	0.32	6.85E-7	61.6417	2.28E-8	0.9999365	0.9999771	0.9986	3.484	0.9999743	0.0023	0.0062	23	1
fpzip_24	0.42	3.14E-7	65.0272	1.01E-8	0.9999852	0.9999972	0.9992	0	0.9999996	0.0005	0.0029	16	0
fpzip_28	0.53	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	16	0
fpzip_32	0.64	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	21	0
speck_1	0.03	1.61E-4	37.9426	5.25E-6	0.9837388	0.9937612	0.8403	33.438	0.9847874	0.0518	0.0845	6	90
speck_2	0.06	6.22E-5	42.0592	2.07E-6	0.9930488	0.9972905	0.9151	21.838	0.9939472	0.0314	0.0536	3	28
speck_4	0.11	1.67E-5	47.7701	5.50E-7	0.9981441	0.9993690	0.9723	11.984	0.9983560	0.0157	0.0274	18	13
speck_8	0.22	1.52E-6	58.1735	5.11E-8	0.9998141	0.9999457	0.9969	5.367	0.9997921	0.0058	0.0091	18	1
speck_12	0.34	4.96E-7	63.0424	1.64E-8	0.9999729	0.9999956	0.9987	0.207	0.9999859	0.0006	0.0037	29	0
speck_24	0.67	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	24	0
speck_32	0.90	0	∞	0	1.0	1.0	1.0	0	1.0	0	0	24	0

Table 5: *For each variable and lossy compression method/level, we list the calculated IQA metric values (in the same order as in Table 1) for the resulting diagnostic image. In the rightmost two columns, we list the user study responses ("Identical" or "Different") for each image. The compression levels are listed from highest compression to lowest compression ratio.*

References

[BHM*16] BAKER A. H., HAMMERLING D., MICHELSON S. A., XU H., ET AL.: Evaluating lossy data compression on climate simulation data within a large ensemble. *Geoscientific Model Development 9*, 12 (2016). 1, 2, 9

[BXD*14] BAKER A., XU H., DENNIS J., LEVY M., NYCHKA D., MICKELSON S., EDWARDS J., VERTENSTEIN M., WEGENER A.: A methodology for evaluating the impact of data compression on climate simulation data. In *Proceedings of the 23rd International Symposium on High-performance Parallel and Distributed Computing* (2014), HPDC '14, pp. 203–214. 1, 2

[BXH*17] BAKER A. H., XU H., HAMMERLING D. M., LI S., CLYNE J. P.: Toward a multi-method approach: Lossy data compression for climate simulation data. In *International Conference on High Performance Computing* (2017), Springer, pp. 30–42. 2, 3, 4, 7, 8, 9

[CJJ93] CJ B., JC G., J H.: Comparison of the Farnsworth-Munsell 100-hue, the Farnsworth D-15, and the Anthony D-15 desaturated color tests. *Archives of Ophthalmology 111*, 5 (1993), 639–641. doi:10.1001/archopht.1993.01090050073032. 6

[Col02] COLLETT D.: *Modelling Binary Data, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2002. URL: https://books.google.com/books?id=3Zx_AwAAQBAJ. 7, 8

[CP16] CHOW L. S., PARAMESRAN R.: Review of medical image quality assessment. *Biomedical signal processing and control 27* (2016), 145–154. 2

[CW12] CUMMINGHAM D. W., WALLRAVEN C.: *Experimental Design – From User Studies to Psychophysics*. CRC Press (Taylor and Francis Group), Boca Raton, 2012, ch. 6. 5

[DC16] DI S., CAPPELLO F.: Fast error-bounded lossy HPC data compression with sz. In *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (2016), pp. 730–739. doi:10.1109/IPDPS.2016.11. 4

[DPR*18] DASGUPTA A., POCO J., ROGOWITZ B., HAN K., BERTINI E., SILVA C. T.: The effect of color scales on climate scientists' objective and subjective performance in spatial data analysis tasks. *IEEE Transactions on Visualization and Computer Graphics* (2018), 1–1. doi:10.1109/TVCG.2018.2876539. 9

[GKS*12a] GEORGIEV V. T., KARAHALIOU A. N., SKIADOPOULOS, S. G. A., S. N., KAZANTZI A. D.AND PANAYIOTAKIS G. S. C. L. I.: Quantitative visually lossless compression ratio determination of JPEG2000 in digitized mammograms. *Journal of digital imaging 26*, 3 (2012), 427–439. 8

[GKS*12b] GEORGIEV V. T., KARAHALIOU A. N., SKIADOPOULOS S. G., ARIKIDIS N. S., KAZANTZI A. D.AND PANAYIOTAKIS G. S., COSTARIDOU L. I.: Quantitative visually lossless compression ratio determination of JPEG2000 in digitized mammograms. *Journal of digital imaging 26*, 3 (2012), 427–39. 2

[HB10] HEER J., BOSTOCK M.: Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2010), CHI '10, ACM, pp. 203–212. URL: http://doi.acm.org/10.1145/1753326.1753357, doi:10.1145/1753326.1753357. 9

[HHG*13] HURRELL J., HOLLAND M., GENT P., GHAN S., KAY J., KUSHNER P., LAMARQUE J.-F., LARGE W., LAWRENCE D., LINDSAY K., LIPSCOMB W., LONG M., MAHOWALD N., MARSH D., NEALE R., RASCH P., VAVRUS S., VERTENSTEIN M., BADER D., COLLINS W., HACK J., KIEHL J., MARSHALL S.: The Community Earth System Model: a framework for collaborative research. *Bulletin of the American Meteorological Society 94* (2013), 1339–1360. doi:10.1175/BAMS-D-12-00121.1. 1

[HWK*13] HÜBBE N., WEGENER A., KUNKEL J. M., LING Y., LUDWIG T.: Evaluating lossy compression on climate data. In *Proceedings of the International Supercomputing Conference (ISC '13)* (2013), pp. 343–356. 1

[IP98] ISLAM A., PEARLMAN W. A.: Embedded and efficient low-complexity hierarchical image coder. In *Electronic Imaging'99* (1998), International Society for Optics and Photonics, pp. 294–305. 4

[KBMG13] KOFF D., BAK P., MATOS A., G. N.: Evaluation of irreversible compression ratios for medical images thin slice CT and update of Canadian Association of Radiologists (CAR) guidelines. *Journal of Digital Imaging 26* (2013), 440–446. doi:10.1007/s10278-012-9542-y. 2

[KDP*15] KAY J., DESER C., PHILLIPS A., MAI A., HANNAY C., STRAND G., ARBLASTER J., BATES S., DANABASOGLU G., EDWARDS J., HOLLAND M., KUSHNER P., LAMARQUE J.-F., LAWRENCE D., LINDSAY K., MIDDLETON A., MUNOZ E., NEALE R., OLESON K., POLVANI L., VERTENSTEIN M.: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society 96* (2015). 2

[KKL16] KUHN M., KUNKEL J., LUDWIG T.: Data compression for climate data. *Supercomputing frontiers and innovations 3*, 1 (2016), 75–94. 1

[KMLH10] KIM K., MANTIUK R., LEE K. H., HEIDRICH W.: Calibration of the visual difference predictor for estimating visibility of JPEG2000 compression artifacts in CT images. *Proceedings of SPIE - The International Society for Optical Engineering 7527* (02 2010), 75270. doi:10.1117/12.845292. 2

[KR12] KUMAR R., RATTAN M.: Analysis of various quality metrics for medical image processing. *International Journal of Advanced Research in Computer Science and Software Engineering 2*, 11 (2012), 137–144. 2

[KS06] KOFF D., SHULMAN H.: An overview of digital compression of medical images: can we use lossy image compression in radiology? *Canadian Association of Radiologists Journal 57* (2006), 211 –217. 2

[KUBW*14] KOWALIK-URBANIAK I., BRUNET D., WANG J., KOFF D., SMOLARSKI-KOFF N., R. VRSCAY E., WALLACE B., WANG Z.: The quest for "diagnostically lossless" medical image compression: A comparative study of objective quality metrics for compressed medical images. In *Progress in Biomedical Optics and Imaging - Proceedings of SPIE* (02 2014), vol. 9037. doi:10.1117/12.2043196. 9

[LBBS16] LAPARRA V., BALLÉ J., BERARDINO A., SIMONCELLI E. P.: Perceptual image quality assessment using a normalized laplacian pyramid. In *Proc. IS&T Int'l Symposium on Electronic Imaging, Conf. on Human Vision and Electronic Imaging* (2016), Rogowitz B., Pappas T. N., de Ridder H., (Eds.), no. 16 in 2016, Society for Imaging Science and Technology, pp. 1–6. doi:10.2352/ISSN.2470-1173.2016.16.HVEI-103. 3

[LC10] LARSON E., CHANDLER D.: Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging 19* (2010). doi:10.1117/1.3267105. 3

[LI06] LINDSTROM P., ISENBURG M.: Fast and efficient compression of floating-point data. *IEEE Transactions on Visualization and Computer Graphics 12* (2006), 1245–1250. 1, 4

[Lin14] LINDSTROM P.: Fixed-rate compressed floating-point arrays. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2674–2683. doi:10.1109/TVCG.2014.2346458. 4

[LLW*13] LANEY D., LANGER S., WEBER C., LINDSTROM P., WEGENER A.: Assessing the effects of data compression in simulations using physically motivated metrics. In *Supercomputing 2013 (SC'13)* (2013). 1

[LSE*11] LAKSHMINARASIMHAN S., SHAH N., ETHIER S., KLASKY S., LATHAM R., ROSS R., SAMATOVA N. F.: Compressing the incompressible with ISABELA: In-situ reduction of spatio-temporal data. In *Proceedings of the 17th International Conference on Parallel Processing* (Bordeaux, France, Aug 29 - Sep 2 2011), Euro-Par'11. 1

[MMT*14] MEEHL G., MOSS R., TAYLOR K., EYRING V., STOUFFER R., BONY S., STEVENS B.: Climate model intercomparisons: Preparing for the next phase. *Eos, Transactions American Geophysical Union 95*, 9 (2014), 77–78. doi:10.1002/2014EO090001. 1

[MN89] MCCULLAGH P., NELDER J.: *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989. URL: http://books.google.com/books?id=h9kFH2_FfBkC. 7

[NCA] NCAR Climate Data Guide. https://climatedataguide.ucar.edu/climate-data. 9

[PDF14] PHILLIPS A., DESER C., FASULLO J.: Evaluating modes of variability in climate models. *Eos, Transactions American Geophysical Union* (2014), 453–455. 2

[SB06] SHEIKH H. R., BOVIK A. C.: Image information and visual quality. *IEEE Transactions on Image Processing 15*, 2 (2006), 430–444. doi:10.1109/TIP.2005.859378. 2, 3

[SoRE11] SOCIETY OF RADIOLOGY (ESR E.: Usability of irreversible image compression in radiological imaging. a position paper by the european society of radiology (esr). *Insights into Imaging 2* (04 2011), 103–115. doi:10.1007/s13244-011-0071-x. 9

[SSB06] SHEIKH H. R., SABIR M. F., BOVIK A. C.: A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing 15* (2006), 3440–3451. 5

[SWH16] STREIJL R. C., WINKLER S., HANDS D. S.: Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems 22*, 2 (Mar 2016), 213–227. URL: https://doi.org/10.1007/s00530-014-0446-1, doi:10.1007/s00530-014-0446-1. 5

[TBRA17] TURTON T. L., BERRES A. S., ROGERS D. H., AHRENS J.: ETK: An Evaluation Toolkit for Visualization User Studies. In *EuroVis 2017 - Short Papers* (2017), Kozlikova B., Schreck T., Wischgoll T., (Eds.), The Eurographics Association. doi:10.2312/eurovisshort.20171131. 5

[TDCC17] TAO D., DI S., CHEN Z., CAPPELLO F.: Significantly improving lossy compression for scientific data sets based on multidimensional prediction and error-controlled quantization. In *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (2017), pp. 1129–1139. doi:10.1109/IPDPS.2017.115. 4

[UCA17] UCAR/NCAR/CISL/TDD: The NCAR Command Language [Software], 2017. Version 6.4.0. URL: http://dx.doi.org/10.5065/D6WD3XH5. 4

[Wan11] WANG Z.: Applications of objective image quality assessment methods [applications corner]. *IEEE Signal Processing Magazine 28*, 6 (2011), 137–142. doi:10.1109/MSP.2011.942295. 2

[WB09] WANG Z., BOVIK A. C.: Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine 26*, 1 (2009), 98–117. 2, 3

[WBSS04] WANG Z., BOVIK A. C., SHEIKH H. R., SIMONCELLI E. P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (2004), 600–612. 2, 3

[Weg10] WEGENER A.: Compression of medical sensor data. *IEEE Signal Processing Magazine 27*, 4 (July 2010), 125–130. 8

[WMB*11] WOODRING J., MNISZEWSKI S. M., BRISLAWN C. M., DEMARLE D. E., AHRENS J. P.: Revisiting wavelet compression for large-scale climate data using JPEG2000 and ensuring data precision. In *IEEE Symposium on Large Data Analysis and Visualization (LDAV)* (2011), Rogers D., Silva C. T., (Eds.), IEEE, pp. 31–38. 1

[WSB03] WANG Z., SIMONCELLI E. P., BOVIK A. C.: Multi-scale structural similarity for image quality assessment. In *in Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers, (Asilomar* (2003), pp. 1398–1402. 2, 3

[WTB*18] WARE C., TURTON T. L., BUJACK R., SAMSEL F., SHRIVASTAVA P., ROGERS D. H.: Measuring and modeling the feature detection threshold functions of colormaps. *IEEE Transactions on Visualization and Computer Graphics* (2018), 1–1. doi:10.1109/TVCG.2018.2855742. 9

[XZMB14] XUE W., ZHANG L., MOU X., BOVIK A. C.: Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing 23*, 2 (2014), 684–695. doi:10.1109/TIP.2013.2293423. 3

[Zen16] ZENDER C. S.: Bit grooming: statistically accurate precision-preserving quantization with compression, evaluated in the netcdf operators (nco, v4.4.8+). *Geoscientific Model Development 9*, 9 (2016), 3199–3211. URL: http://www.geosci-model-dev.net/9/3199/2016/, doi:10.5194/gmd-9-3199-2016. 1

[ZZMZ11] ZHANG L., ZHANG L., MOU X., ZHANG D.: FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing 20*, 8 (2011), 2378–2386. 3