

Machine Learning Techniques for Data Reduction of Climate Applications

Xiao Li*, Qian Gong[†], Jaemoon Lee*, Scott Klasky[†], Anand Rangarajan*
Sanjay Ranka*

*University of Florida

[†]Oak Ridge National Laboratory

Abstract—Scientists conduct large-scale simulations to compute derived quantities-of-interest (QoI) from primary data. Often, QoI are linked to specific features, regions, or time intervals, such that data can be adaptively reduced without compromising the integrity of QoI. For many spatiotemporal applications, these QoI are binary in nature and represent presence or absence of a physical phenomenon.

We present a pipelined compression approach that first uses neural-network-based techniques to derive regions where QoI are highly likely to be present. Then, we employ a Guaranteed Autoencoder (GAE) to compress data with differential error bounds. GAE uses QoI information to apply low-error compression to only these regions. This results in overall high compression ratios while still achieving downstream goals of simulation or data collections. Experimental results are presented for climate data generated from the E3SM Simulation model for downstream quantities such as tropical cyclone and atmospheric river detection and tracking. These results show that our approach is superior to comparable methods in the literature.

I. INTRODUCTION

Scientific applications are producing data at an unprecedented rate, with both volume and velocity expanding rapidly, outpacing advancements in storage, computation, and network capacities [1], [2]. This growth is also seen in next-generation experimental and observational facilities, making data reduction an essential functionality because it's often impractical to retain data at the original scale. Data reduction necessitates the development of algorithms that can reconstruct data with rigorous quality assurance so that application scientists can trust the reduced data for use in their analyses.

Because the errors on derived quantities or quantities of interest (QoI) may deviate from these on lossy reduced primary data (PD), it is important that reduction techniques can maintain the accuracy of QoI and provide high compression ratios. Long-term and large-scale scientific simulations often model dynamic processes over an extended period and write out field data at time steps. In this paper, we target the spatiotemporal data analysis with climate simulation, as the presence or absence of a physical phenomenon (i.e., QoI) can be represented by binary masks. Throughout this paper, we refer to locations where climate features are presented as *critical regions* or *regions-of-interest* (ROIs). Our goal in this paper is to demonstrate a hybrid, feature-driven compression approach. The proposed pipeline consists of three stages (Figure 1):

- **ROI Detection:** We predict regions where these QoI are likely to be present. We employ historical data sets and a customized UNet model. The output is binary masks that indicate the presence of climate phenomena of interest. In this work, we propose to hybridize neural network (NN) prediction with differential compression. Our approach relies on a UNet-based neural network that can jointly predict ROIs for different climate events, such as AR and TC. Our neural network model generates probability maps for climate events where climate ROIs may be present. The allocation of ROIs can be fine-tuned by adjusting the acceptance threshold. This approach offers a convenient way to balance between false positive rates and compression ratios.
- **Guaranteed Autoencoder:** The original data are divided into smaller spatiotemporal blocks. a 3D convolutional autoencoder is utilized to capture the spatiotemporal correlations within each block. To guarantee the error bound of the reconstructed data, Principal Component Analysis (PCA) is applied to the residual between the original and reconstructed data. This yields a basis matrix, which is then used to project the residual of each instance. The resulting coefficients are retained to enable accurate recovery of the residual. The number of coefficients saved is incrementally increased until the error bound is satisfied. Additionally, quantization and entropy coding techniques are applied to both the latent data from the GAE and the PCA coefficients. This further improves the compression ratio of the overall process.
- **Differential compression:** We use lower error bounds on predicted critical regions and higher error bounds on the rest of space to ensure minimal false negatives on QoI and large compression ratios globally. Currently, most state-of-the-art lossy compressors [3]–[5] are mostly focused on controlling global error such as L^∞ and L^2 . Without local error control, the application of differential compression requires the partitioning of ROIs and the rest of the data and compression of each part separately. In this paper, we employ a multilevel-decomposition-based compression method discussed in [4] because it supports pixel-wise changed error bounds. The differentially compressed data can be reconstructed using the same algorithm and code as the uniformly compressed data.

We validate our hybrid method using the simulation data generated by Energy Exascale Earth System Model (E3SM) [6]

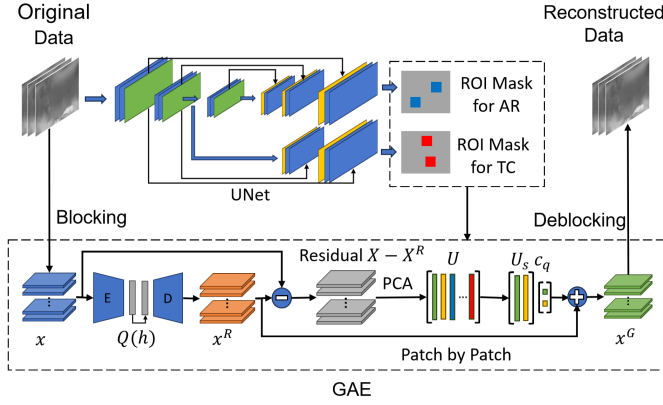


Fig. 1: Method Overview: x denotes AE the input block, while x^R denotes AE the output block. x^G denotes the error-bounded output block. $Q(h)$ represents latent space quantization. U denotes PCA basis matrix. U_s and C_q are selected basis vectors and corresponding quantized coefficients, respectively.

and assess the impact of compression errors on two climate QoI: tropical cyclone (TC) [7] and atmospheric river (AR) [8] tracking. In comparison to previous research [4], our method achieves a substantially higher compression ratio while also attaining a lower false negative (FN) ratio for both TC and AR detection, along with significantly improved quality of information preservation.

The rest of the paper is organized as follows. Section II describes our compression pipeline. Experimental results using our method are presented in Section III for different levels of compression and accuracy. Section IV presents related work. Conclusions are provided in Section V.

II. METHODOLOGY

The overview of our compression approach is given in Figure 1. Our method includes two parts: ROI detection and GAE. A UNet is employed for the detection of ROIs. We perform differential compression with GAE by applying lower error bound for ROIs and higher error bound for other regions. In the following, we briefly describe the ROI detection model and the GAE.

ROI Detection: We used a UNet [9] architecture with a U-shaped interconnection topology, which includes a contracting pathway to capture contextual information and an expanding pathway for accurate localization. Our network accepts input data with dimensions of $C \times H \times W$, where C denotes the number of climate variables, while H and W correspond to the height and width of the input images, respectively. To meet the requirement of jointly predicting the locations of TC and AR, we have made the following revisions to the original UNet architecture and extended the decoder into two branches (Figure 1) so that the two decoders share the same encoder. We use a shallow branch for TC detection, while using a deep branch for AR detection. The reason for this design choice is that we have found that using a deep neural network can yield better results for AR, while a shallow network is sufficient for TC detection. The output of each

decoder is passed through a sigmoid function to constrain the output within the range $[0, 1]$. The network’s output takes the form of a confidence map with dimensions $H \times W$, effectively represented as an image comprising probability values. These probabilities can then be suitably thresholded to derive ROIs.

Given that TCs are relatively rare and infrequent phenomena on Earth, it can be challenging for the model to learn robust features under the supervision of very sparse coordinates. Therefore, we convert the TC’s location coordinates into a Gaussian heatmap using the provided equation:

$$H_i(u, v) = A \cdot \exp\left(-\frac{(u - u_i)^2 + (v - v_i)^2}{2\sigma^2}\right), \quad (1)$$

where (u_i, v_i) denotes the location of the TC in the image coordinate system, σ represents the standard deviation controlling the spread of the TC representation, and A is a scaling factor that controls the overall peak intensity of the heatmap.

Because there could be more than one TC in a single data frame, we generate a heatmap for each cyclone within one frame and then combine them into a single heatmap using the following equation:

$$H_{\text{combined}}(x, y) = \max_{i=1}^N H_i(x, y), \quad (2)$$

where N is the number of TCs in one frame. The combined heatmap serves as the ground-truth label for model training.

The locations of ARs are represented using binary masks, where each pixel indicates the presence or absence of an AR at the corresponding location.

GAE: An autoencoder (AE) is a neural network designed to learn efficient representations of data by compressing the input into a lower-dimensional latent space and then reconstructing the original input from this compressed representation. It consists of two main parts: an encoder and a decoder. The encoder compresses the input data from high-dimensional to lower-dimensional latent space representation, while the decoder reconstructs the original input from this compressed representation. [10]

AEs can be trained by comparing the uncompressed and reconstructed data. The initial step towards achieving an effective codec involves minimizing the reconstruction error adequately. In this study, we employ the standard mean-squared error (MSE) loss function to measure compression-incurred errors in PD.

After training, the decoder and the latent representations $H = \{h_i\}_{i=1}^N$ need to be stored. We store the decoder without compression, given its small size. However, storing floating-point latent vectors is not an efficient approach. To overcome this challenge, we employ a compression technique that involves float-point quantization followed by entropy encoding to compress the latent space data. Our approach to enhancing compression efficiency employs a two-step strategy. Firstly, we uniformly quantize these coefficients into discrete bins, each with a bin size of d . This discretization process effectively represents all values within each bin by its central

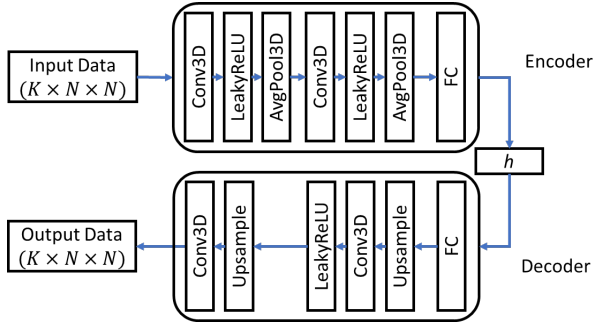


Fig. 2: The structure of the autoencoder: Conv3D denotes the 3D convolution layer, FC denotes the fully connected layer, and h denotes the latent space. Leaky ReLU is adopted as the activation function. We use a 3D average pooling layer to reduce the size of features and trilinear upsampling for interpolation.

value, transforming the originally continuous data into a discrete form. Subsequently, we utilize Huffman coding to compress these quantized coefficients. Huffman coding assigns shorter codes to frequently occurring quantized coefficients, optimizing the representation of the data and achieving higher compression efficiency. This combined technique significantly reduces the data's size while retaining crucial information.

Bounding the Reconstruction Error: We aim to limit reconstruction errors for all instances in an AE. Although any appropriate error bound can be applied within our framework, our emphasis lies in constraining the ℓ_2 norm of each instance residual, denoted as $\|x - x^R\|_2$. To optimize compression ratios, we only bound the AE reconstruction error for instances whose residual ℓ_2 norm exceeds the specified threshold τ . After obtaining the reconstructed data from the autoencoder, we apply Principal Component Analysis (PCA) to the residual of the entire dataset to extract the principal components or basis matrix, denoted as U . These basis vectors are sorted in descending order according to their corresponding eigenvalues. These principal components capture the directions of maximum variance in the residual data. Although we compress the data block by block, we treat each patch of data as a single instance and compute the basis matrix at the patch level. To guarantee the error bound for each patch of data, we project the residual of each patch of data onto the space spanned by U and select the leading coefficients such that the ℓ_2 norm of the corrected residual falls below the specified threshold τ . These coefficients, representing the residual, are derived from the equation:

$$c = U^T (x - x^R). \quad (3)$$

It's important to note that the complete recovery of the residual $x - x^R$ can be achieved by computing Uc , yielding the coefficient vector $c \equiv [c_1, \dots, c_D]$. Given that the error bound criterion is based on ℓ_2 , we compute $\{c_k^2\}_{k=1}^D$ and sort the positive values. The coefficients are therefore sorted in the order of largest contribution to the error. The top M coefficients and corresponding basis vectors are selected to satisfy the target error bound τ . To minimize the storage cost of these

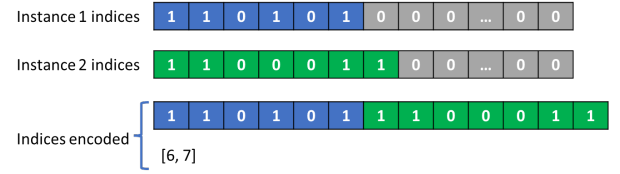


Fig. 3: Indices encoding

coefficients, we compress the selected coefficients c_s using a method similar to that employed for compressing AE latent coefficients. These coefficients are quantized before being used for reconstructing the residual. The corrected reconstruction x^G is

$$x^G = x^R + U_s c_q, \quad (4)$$

where c_q is the set of selected coefficients c_s after quantization and U_s is the set of selected basis vectors. We increase the number of coefficients until we achieve $\|x - x^G\|_2 \leq \tau$. Therefore, in order to guarantee the error bound, we need to store the basis matrix, the selected coefficients c_q for each patch data and their basis vector indices.

To explore the spatiotemporal correlation within scientific data, we integrate 3D convolution into our AE architecture that incorporates both spatial and temporal correlations. By harnessing the capabilities of 3D convolutional operations, our autoencoder excels in capturing intricate spatial patterns and temporal dynamics simultaneously. We partition the original data into non-overlapping $N \times N$ patches at each data frame. Then, we group K consecutive patches from the same location across time into a single block of data as an input to the AE. Each block of data is processed independently by the AE. The structure of the AE is shown in Figure 2.

To store the basis vector indices for each patch of data effectively, we start by encoding the indices using a binary sequence. Remarkably, we notice that the initial basis vectors (associated with larger eigenvalues) are selected more frequently. As a result, these binary sequences frequently end with a sequence of zeros. We decide to store only the shortest prefix subsequence that contains all ones. Furthermore, we introduce an additional value to indicate the length of this subsequence. An illustrative example is provided in Figure 3.

To guarantee QoI while maintaining a high compression ratio, we categorize all patches of data into three classes, each with a distinct error bound: ROI patches, buffer zone patches, and non-ROI patches. These are based on thresholding the output of the previously described UNet model that produces binary masks indicating the presence or absence of TC/AR. To determine if a patch belongs to the ROI category, we evaluate the proportion of pixels predicted as ROI within each patch data. If this percentage surpasses a specific threshold, the patch is designated as an ROI patch. Furthermore, within the non-ROI patches, we identify neighboring ROI patches as a Buffer Zone patch. This procedure generates a set of error bounds for each patch. Details of the overall approach are described in Algorithm 1.

MGARD: We compare our method with a state-of-the-art lossy compressor – MGARD [11], as it can take in customized

Algorithm 1 The GAE Algorithm

Input: Input data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, reconstructed data $\mathbf{X}^R = \{\mathbf{x}_i^R\}_{i=1}^N$, target error bounds $\{\tau_i\}_{i=1}^N$.
Output: Corrected reconstruction $\mathbf{X}^G = \{\mathbf{x}_i^G\}_{i=1}^N$, coefficients $\mathbf{C} = \{\mathbf{c}_i\}_{i=1}^N$, indices $\mathbf{I} = \{I_i\}_{i=1}^N$ where I_i is an index set, basis matrix $\mathbf{U}$.

- 1: Run PCA on the residual $\mathbf{X} - \mathbf{X}^R$, obtaining basis matrix $\mathbf{U}$
- 2: **for** $i = 1$ **to** N **do**
- 3: $\mathbf{x} \leftarrow \mathbf{x}_i$, $\mathbf{x}^R \leftarrow \mathbf{x}_i^R$, $\tau \leftarrow \tau_i$.
- 4: Compute ℓ_2 norm $\delta = \|\mathbf{x} - \mathbf{x}^R\|_2$.
- 5: **if** $\delta > \tau$ **then**
- 6: Project residual $\mathbf{c} = \mathbf{U}^T(\mathbf{x} - \mathbf{x}^R)$ and sort c_k^2 , $\forall k$.
- 7: $M \leftarrow 1$
- 8: **while** $\delta > \tau$ **do**
- 9: $\mathbf{c}_s, U_s \leftarrow$ Top M coefficients in $\mathbf{c}$ and corresponding basis vectors in $\mathbf{U}$.
- 10: $\mathbf{c}_q \leftarrow \text{Quantize}(\mathbf{c}_s)$
- 11: $\mathbf{x}^G \leftarrow \mathbf{x}^R + U_s \mathbf{c}_q$.
- 12: $\delta \leftarrow \|\mathbf{x} - \mathbf{x}^G\|_2$
- 13: $M \leftarrow M + 1$
- 14: **end while**
- 15: $\mathbf{c}_i \leftarrow \mathbf{c}_q$
- 16: $I_i \leftarrow$ Index set for $\mathbf{c}_q$
- 17: $\mathbf{x}_i^G \leftarrow \mathbf{x}^G$
- 18: **end if**
- 19: **end for**

ROI masks then conduct regionally-varied, error-bounded compression. MGARD utilizes a multilevel decomposition-based compression algorithm and mathematically ensures the errors induced by lossy compression to stay below user-prescribed error tolerance. MGARD treats input data u in d dimensions as the values taken by a continuous function on a grid N_L . It decomposes u into a set of multi-level coefficients $\mathbf{u}_{\text{mc}}$, residing on a hierarchy of subgrids $\mathcal{N}_{L-1}, \dots, \mathcal{N}_l, \dots, \mathcal{N}_0$ downsampled from N_L through L^2 projection and multilinear interpolation operations. MGARD then leverages linear-scaling quantization, Huffman encoding [12], and GZIP/ZSTD compression [13], [14] to convert $\mathbf{u}_{\text{mc}}$ into a reduced representation subject to user-prescribed error bounds.

III. EXPERIMENTAL RESULTS

This section presents the experimental results of compressing the E3SM through our pipeline. We provide a concise description of the data to be compressed and evaluation metrics, followed by a comparison of accuracy and compression ratio achieved by our method as compared to previous work.

Dataset: The testing data come from a HR configuration atmosphere E3SM simulation spanning one year. In this simulation, E3SM utilizes a grid resolution of 25 km, resulting in a total of 350,000 float32 data points per variable per

time-slice. The grid spacing used in the data simulation is 0.25° . The TC and AR detection are treated as blackboxes in our approach. We used TempestExtremes version 2.1 [15], a widely recognized software package designed for feature tracking and scientific analysis of global earth-system data. The detected TCs are nodal points in latitude and longitude coordinates, and the atmospheric rivers (AR) are areas formed by connected grid points. Eligible candidate across consecutive timesteps can be stitched to form TC and AR trajectories. The outcomes of TempestExtremes for the AR and TC locations at individual snapshot (i.e., timestep output) are used as binary masks and serve as the ground-truth labels for both model training and testing purposes.

UNet Configure: To achieve better compression ratios, for each snapshot, we transform variables from their original 1D grids to three 2D snapshots correlated in longitude and latitude on the cubed sphere. We split the dataset into two subsets: a training set and a testing set, which include 34,560 images and 21,600 images, respectively. We use the following TC tracking variables as inputs: pressure at sea level (PSL), temperature at 500 hPa and 200 hPa (T500, T200), meridional and zonal wind speed at surface (VBOT, UBOT), and 1 AR tracking variable –integrated vapor transport ($\text{IVT} = \sqrt{\text{TVQ}^2 + \text{TUQ}^2}$, where TVQ and TUQ denote meridional and zonal water flux, respectively) into the developed model and simultaneously make predictions on the above two QoI.

AE Configure: To train the autoencoder, we partition the dataset into uniform blocks. Experimentally, each block is defined by a size of $8 \times 16 \times 16$, where the spatial dimensions are 16×16 , and data from 8 consecutive frames at the same location are combined into a single block. Using the mean squared error (MSE) loss function to train the network, our method is assessed across all the climate variables: PSL, T200, T500, VBOT, UBOT, TVQ, and TUQ. The first five variables serve for TC detection, while the latter two are utilized for AR detection. Separate autoencoders are trained for each variable.

Evaluation Metrics: We evaluate our method from three aspects: ROI detection, QoI preservation, and primary data error. The overall compression ratio that can be achieved is directly proportion to the percentage of total area or volume covered by ROIs as well as the false negatives on the QoI. For a given QoI, the FN percentage reflects the proportion of incorrectly classifying genuine ROIs as negative. The ROI ratio indicates the portion of the region designated as ROIs based on prediction results. This assessment helps us gauge the model’s effectiveness in accurately detecting and encompassing all ROIs. The resulting reconstructed climate data are then inputted into TempestExtremes to generate AR and TC predictions. Simultaneously, we also run TempestExtremes on the uncompressed climate data. This is done by a comparison between the detection results obtained using reconstructed data and original uncompressed data.

For TC evaluation, we compare the locations of cyclones obtained from the compressed and uncompressed data. we use TC error rate to measure the performance of our method on TC preservation. The TC error rate is defined as:

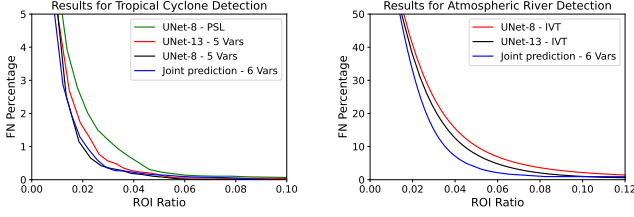


Fig. 4: ROI ratio versus false negative ratio curve for TC and AR. These results show that a NN-based prediction approach for both TC and AR can achieve close to zero false negatives while keeping the ROIs to be less than 12 percent.

$$\text{TC Error Rate} = \frac{N_{\text{error}}}{N_{\text{total}}} * 100\% \quad (5)$$

where N_{total} represents the total count of ground truth TCs, and N_{error} denotes the count of TCs from the compressed data that deviate from their ground truth locations.

For AR evaluation, we adopt the Jaccard index or intersection-over-union (IoU) to evaluate the performance of our method. IoU can be expressed as

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \quad (6)$$

where TP, FP, FN denote the number of true positive, false positive, and false negative predictions for AR.

For primary data, we utilize the relative l_2 norm as the evaluation metric, defined as:

$$\text{Relative } l_2 = \frac{\|\mathbf{x} - \hat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2},$$

where $\mathbf{x}$ denotes the original uncompressed data and $\hat{\mathbf{x}}$ denotes the reconstructed data.

ROI Evaluation: We conducted multiple experiments to assess the effectiveness of our approach. Our network simultaneously predicts TC and AR. For each of these cases, we tried two UNet structures with eight layers and 13 layers, respectively. We also varied the the input variables that were used for the prediction:

- ‘PSL’ and ‘IVT’ are the most important variables for TC detection and AR detection, respectively. For experiments marked with ‘PSL’ or ‘IVT’, we utilized only one variable as input.
- ‘5 Vars’ corresponds to the use of five variables: PSL, T500, T200, UBOT, and VBOT.
- All variables correspond to using all the six variables as input to jointly detect ARs and TCs.

A comparison of ROI ratio and FN for different approaches is shown in Figure 4. For TC detection, our method achieves a false negative ratio of 0, coupled with an ROI ratio of approximately 0.1. For AR detection, our approach attains an FN of less than 0.9% for a corresponding ROI ratio of 0.08. In comparison to Qian et al. [4], which achieves a 0.7% false negative ratio with a 0.22 ROI ratio for TC and a 5.8% false negative ratio with a 0.36 ROI ratio for AR,

our method exhibits substantial improvements in both TC and AR detection accuracy with lower ROI ratio. Thus, by using our approach, where the QoI are provided by existing application, we are able to more precisely determine the regions where they are present. At this juncture, any approach that uses differential compression should generally use higher compression in areas outside ROIs and should result in an overall better compression. We want to clarify again that the purpose of using the prediction model (that uses historical QoI) is only for determining ROIs. In the following sections, we show that both PD error and QoI errors are bounded. All the results are on unseen data used by the prediction mode.

QoI Evaluation: We applied GAE to perform differential compression and decompression on climate data with the ROIs identified by UNet model. Subsequently, we utilized Tempes-tExtremes on both the reconstructed and original climate data, followed by a comparison of the results for TC and AR. TC preservation was assessed using the TC error rate, while AR preservation was evaluated using the Intersection over Union (IoU), as discussed in the previous section.

In our TC comparison, we conducted a series of experiments detailed in Table Ia, comparing GAE with MGARD across various experimental setups. Experiments labeled ‘Uniform’ employed a consistent error bound across all regions, while those marked ‘UNet’ utilized region-adaptive error bounds. Qian et al. [4] utilized region-adaptive compression using MGARD with ROIs detected by MGARD coefficients. When comparing MGARD-Uniform with GAE-Uniform at similar compression ratios, GAE achieved a lower TC error rate. Similarly, comparing MGARD-UNet with the approach by Qian et al. [4], UNet notably improved compression ratio by over 2 times while achieving a lower TC error rate. Moreover, in the comparison between MGARD-UNet and GAE-UNet, GAE further enhanced the compression ratio by approximately 50% while maintaining the same TC error rate. Additionally, we present the compression ratio versus TC error rate curve, as depicted in Figure 5 (a-c). We specifically choose PSL, T200, and T500 as examples, given their significant influence on TC detection. Region-adaptive compression with UNet substantially outperformed uniform compression in terms of TC preservation. When comparing GAE with MGARD, GAE exhibited comparable or superior performance in both uniform and region-adaptive compression scenarios.

Similar trends are evident in the AR comparison, as summarized in Table Ib and illustrated in Figure 5 (d). When comparing Qian et al.’s approach [4] with MGARD-UNet at a close IoU, leveraging ROIs identified by UNet leads to a compression ratio improvement of over 50%. Further comparison between MGARD-UNet and GAE-UNet demonstrates an additional enhancement of approximately 60% in compression ratio with the integration of GAE. By combining the advantages of GAE and UNet, GAE-UNet surpasses the performance of Qian et al. [4] by 1.5 times in terms of compression ratio.

PD Error: To evaluate the performance of GAE on primary data, we conducted several comparisons between dif-

Method	ROI Ratio	Compression Ratio						TC Error Rate
		PSL	T200	T500	VBOT	UBOT	Overall	
Qian et al. [4]	0.22	10.8	14.6	17.1	18.3	18.0	15.2	0.38%
MGARD-Uniform	-	18.3	21.6	22.6	26.1	24.7	22.1	1.7%
MGARD-UNet	0.07	20.9	48.7	51.3	51.5	50.1	39.4	0.29%
GAE-Uniform	-	20.7	22.4	20.3	21.5	20.3	21.0	1.2%
GAE-UNet	0.07	54.3	63.0	66.8	66.3	53.0	60.1	0.29%

(a) QoI preservation results for TC.

Method	ROI Ratio	CR	IoU
Qian et al. [4]	0.37	9.1	99.3%
MGARD-Uniform	-	12.1	99.1%
MGARD-UNet	0.22	14.4	99.2%
GAE-Uniform	-	17.4	99.2%
GAE-UNet	0.22	23.4	99.4%

(b) QoI preservation results for AR.

TABLE I: QoI preservation results for TC and AR. The ‘Uniform’ corresponds to compressing the entire region with the same error bound. The ‘UNet’ corresponds to differential compression where ROIs identified by UNet are compressed with lower error bound than other regions. These results show that our method has a significantly lower ROI and a higher compression ratio, but achieves lower QoI error as compared to previous work.

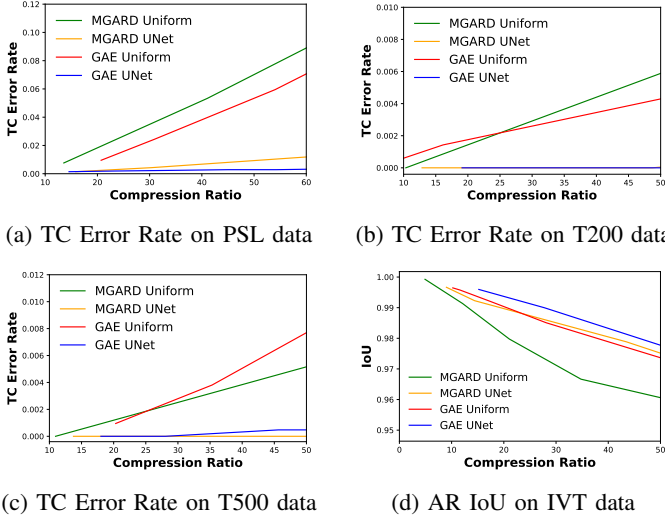


Fig. 5: QoI error for TC and AR. These results show that GAE with UNet is able to achieve higher compression ratio while achieving comparable or better accuracy to other methods.

ferent experimental setups. The comparisons were conducted on all seven climate variables, but we only show four here due to the page constraints. We compared MGARD with GAE and also compared uniform compression with region-adaptive compression. The experimental results are shown in Figure 6. When measuring the error globally, uniform compression achieves lower PD error than region-adaptive compression. This finding holds true for both MGARD and GAE. When comparing MGARD with GAE, it’s notable that GAE consistently achieves a similar or lower PD error than MGARD at the same compression ratio, irrespective of whether uniform or region-adaptive compression is employed.

IV. RELATED WORK

Error-bounded lossy compression has drawn increasing attention as it allows more significant data reduction and ensures the quality of the reconstructed data to meet the user’s request. In general, lossy compressors reduce data through multiple steps including de-correlation, quantization/thresholding, and lossless compression. Based on the de-correlation algorithms employed by lossy compressors, they can be categorized into prediction-based [3], [16], [17], transformation-based [5], [18], [19], and decomposition-based methods [11], [20]–[22]. Most lossy compressors [3]–[5] only support applying uniform error

bounds (e.g., L^2 or L^∞) across the entire data space. Previous research on varying the compression quality for dynamically changing data space typically shows a need to include region segmentation types of methods [23]–[25] and compressing foreground and background data separately. For ROIs with irregular shapes, partitioning them from background can result in oversized bounding boxes or fragmented data pieces, which leads to suboptimal compression ratios. Recently, Liang et al. [26], [27] derived local error bounds which can be used to preserve critical point features for a prediction-based compressor, SZ. However, their method demands a mask of ROIs and different reconstruction algorithms to assemble pixels compressed with varied error bounds. In previous work [4], [28], Gong et al. employ the multilevel coefficients from MGARD decomposition to assess local data turbulence. QoI are treated as variational features, and detected using a hierarchical mesh refinement approach which tracks the clusters of large-magnitude coefficients. The authors only categorize the QoI into nodal-based and areal-based features when making different mesh refinement parameters choices.

V. CONCLUSIONS

Scientists conduct large-scale simulations to compute derived quantities from primary data, and it is crucial for them to find compression techniques that maintain low or bounded errors on these derived quantities called quantities of interest (QoI). Additionally, for many spatiotemporal applications these QoI are binary in nature (presence or absence of a physical phenomenon). We demonstrated a multilevel compression approach that first uses neural-network-based techniques to derive regions where QoI are highly likely to be present. Then, an autoencoder with error bound guarantee is proposed to perform differential compression based on these regions. This results in overall high compression ratios while still achieving downstream goals of simulation or data collections.

Experimental results are presented for climate data generated from the E3SM simulation model for downstream QoI such as tropical cyclone and atmospheric river detection and tracking. Compared to previous research [4], our method achieves a significantly higher compression ratio while simultaneously achieving a lower false negative (FN) ratio for both TC and AR detection.

REFERENCES

- [1] H. Zou, Y. Yu, W. Tang, and H. M. Chen, “Improving i/o performance with adaptive data compression for big data applications,” in *2014 IEEE*

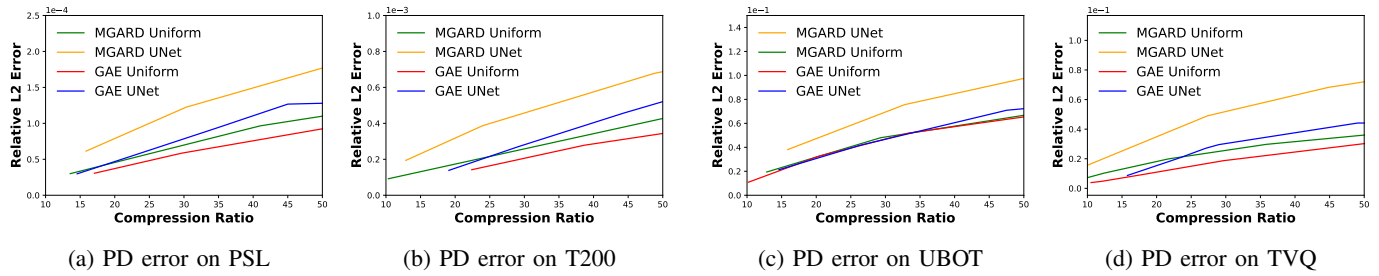


Fig. 6: Comparison of PD error for four variables using the four methods described in the paper. The other three variables have similar relative behavior for the four methods. Overall, GAE-based approaches provide higher compression than other methods for the same error levels and comparable QoI. The QoI are for a subset of these variables, presented in Figure 5.

- International Parallel & Distributed Processing Symposium Workshops*. IEEE, 2014, pp. 1228–1237.
- [2] L. Wen, K. Zhou, S. Yang, and L. Li, “Compression of smart meter big data: A survey,” *Renewable and Sustainable Energy Reviews*, vol. 91, pp. 59–69, 2018.
 - [3] S. Di and F. Cappello, “Fast error-bounded lossy hpc data compression with sz,” in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2016, pp. 730–739.
 - [4] Q. Gong, B. Whitney, C. Zhang, X. Liang, A. Rangarajan, J. Chen, L. Wan, P. Ullrich, Q. Liu, R. Jacob *et al.*, “Region-adaptive, error-controlled scientific data compression using multilevel decomposition,” in *Proceedings of the 34th International Conference on Scientific and Statistical Database Management*, 2022, pp. 1–12.
 - [5] P. Lindstrom, “Fixed-rate compressed floating-point arrays,” *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 2674–2683, 2014.
 - [6] J.-C. Golaz, P. M. Caldwell, L. P. Van Roekel, M. R. Petersen, Q. Tang, J. D. Wolfe, G. Abeshu, V. Anantharaj, X. S. Asay-Davis, D. C. Bader *et al.*, “The doe e3sm coupled model version 1: Overview and evaluation at standard resolution,” *Journal of Advances in Modeling Earth Systems*, vol. 11, no. 7, pp. 2089–2129, 2019.
 - [7] K. Balaguru, L. R. Leung, L. P. Van Roekel, J.-C. Golaz, P. A. Ullrich, P. M. Caldwell, S. M. Hagos, B. E. Harrop, and A. Mametjanov, “Characterizing tropical cyclones in the energy exascale earth system model version 1,” *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 8, p. e2019MS002024, 2020.
 - [8] S. Kim, L. R. Leung, B. Guan, and J. C. Chiang, “Atmospheric river representation in the energy exascale earth system model (e3sm) version 1.0,” *Geoscientific Model Development*, vol. 15, no. 14, pp. 5461–5480, 2022.
 - [9] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
 - [10] J. Lee, A. Rangarajan, and S. Ranka, “Nonlinear-by-linear: Guaranteeing error bounds in compressive autoencoders,” ser. IC3-2023. New York, NY, USA: Association for Computing Machinery, 2023, p. 552–561. [Online]. Available: <https://doi.org/10.1145/3607947.3609702>
 - [11] M. Ainsworth, O. Tugluk, B. Whitney, and S. Klasky, “Multilevel techniques for compression and reduction of scientific data—the multivariate case,” *SIAM Journal on Scientific Computing*, vol. 41, no. 2, pp. A1278–A1303, 2019.
 - [12] A. Moffat, “Huffman coding,” *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–35, 2019.
 - [13] M. S. Abdelfattah, A. Hagiescu, and D. Singh, “Gzip on a chip: High performance lossless data compression on fpgas using opencl,” in *Proceedings of the international workshop on openCL 2013 & 2014*, 2014, pp. 1–9.
 - [14] Y. Collet and C. Turner, “Smaller and faster data compression with zstandard,” *Facebook Code [online]*, vol. 1, 2016.
 - [15] P. A. Ullrich, C. M. Zarzycki, E. E. McClenny, M. C. Pinheiro, A. M. Stansfield, and K. A. Reed, “Tempestextremes v2. 1: A community framework for feature detection, tracking, and analysis in large datasets,” *Geoscientific Model Development*, vol. 14, no. 8, pp. 5023–5048, 2021.
 - [16] X. Liang, S. Di, D. Tao, S. Li, S. Li, H. Guo, Z. Chen, and F. Cappello, “Error-controlled lossy compression optimized for high compression ratios of scientific datasets,” in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 438–447.
 - [17] J. Tian, S. Di, K. Zhao, C. Rivera, M. H. Fulp, R. Underwood, S. Jin, X. Liang, J. Calhoun, D. Tao *et al.*, “Cusz: An efficient gpu-based error-bounded lossy compression framework for scientific data,” in *Proceedings of the ACM International Conference on Parallel Architectures and Compilation Techniques*, 2020, pp. 3–15.
 - [18] J. Diffenderfer, A. L. Fox, J. A. Hittinger, G. Sanders, and P. G. Lindstrom, “Error analysis of zfp compression for floating-point data,” *SIAM Journal on Scientific Computing*, vol. 41, no. 3, pp. A1867–A1898, 2019.
 - [19] A. Fox, J. Diffenderfer, J. Hittinger, G. Sanders, and P. Lindstrom, “Stability analysis of inline zfp compression for floating-point data in iterative methods,” *SIAM Journal on Scientific Computing*, vol. 42, no. 5, pp. A2701–A2730, 2020.
 - [20] X. Liang, B. Whitney, J. Chen, L. Wan, Q. Liu, D. Tao, J. Kress, D. Pugmire, M. Wolf, N. Podhorszki *et al.*, “Mgard+: Optimizing multilevel methods for error-bounded scientific data reduction,” *IEEE Transactions on Computers*, vol. 71, no. 7, pp. 1522–1536, 2021.
 - [21] M. Ainsworth, O. Tugluk, B. Whitney, and S. Klasky, “Multilevel techniques for compression and reduction of scientific data—quantitative control of accuracy in derived quantities,” *SIAM Journal on Scientific Computing*, vol. 41, no. 4, pp. A2146–A2171, 2019.
 - [22] J. Chen, L. Wan, X. Liang, B. Whitney, Q. Liu, Q. Gong, D. Pugmire, N. Thompson, J. Y. Choi, M. Wolf *et al.*, “Scalable multigrid-based hierarchical scientific data refactoring on gpus,” *arXiv preprint arXiv:2105.12764*, 2021.
 - [23] M. Rahman, M. Islam, C. Holt, J. Calhoun, and M. Chowdhury, “Dynamic error-bounded lossy compression to reduce the bandwidth requirement for real-time vision-based pedestrian safety applications,” *Journal of Real-Time Image Processing*, pp. 1–15, 2022.
 - [24] S. R. Uddehal, T. Strutz, H. Och, and A. Kaup, “Image segmentation for improved lossless screen content compression,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
 - [25] H. Akutsu and T. Naruko, “End-to-end deep roi image compression,” *IEICE TRANSACTIONS on Information and Systems*, vol. 103, no. 5, pp. 1031–1038, 2020.
 - [26] P. Jiao, S. Di, H. Guo, K. Zhao, J. Tian, D. Tao, X. Liang, and F. Cappello, “Toward quantity-of-interest preserving lossy compression for scientific data,” *Proceedings of the VLDB Endowment*, vol. 16, no. 4, pp. 697–710, 2022.
 - [27] X. Liang, S. Di, F. Cappello, M. Raj, C. Liu, K. Ono, Z. Chen, T. Peterka, and H. Guo, “Toward feature-preserving vector field compression,” *IEEE Transactions on Visualization and Computer Graphics*, 2022.
 - [28] Q. Gong, C. Zhang, X. Liang, V. Reshniak, J. Chen, A. Rangarajan, S. Ranka, N. Vidal, L. Wan, P. Ullrich *et al.*, “Spatiotemporally adaptive compression for scientific dataset with feature preservation—a case study on simulation data with extreme climate events analysis,” in *2023 IEEE 19th International Conference on e-Science (e-Science)*. IEEE, 2023, pp. 1–10.