ESM-VAE: a VAE Approach For Data Compression On Earth System Model Data

Nina Ervin
Department of Computer Science
Western Washington University
ervinn@wwu.edu

Cooper Cox
Department of Computer Science
Western Washington University
coxc25@wwu.edu

Seth Basseti
Department of Computer Science
Utah State University

Brian Hutchinson
Department of Computer Science
Western Washington University
Foundational Data Science
Pacific Northwest National Laboratory
hutchib2@wwu.edu

Claudia Tebaldi
Joint Global Change Research Institute
Pacific Northwest National Laboratory
claudia.tebaldi@pnnl.gov

Abstract

Earth System Models (ESMs) are a critical tool used by climate scientists to analyze future climate scenarios given different green house gas emissions. These ESMs are not deterministic, so many runs are needed to calculate the probability of extreme weather events. ESMs are computationally intensive to run, so daily data is stored for future analysis. The Sixth Coupled Model Inter-comparison Project (CMIP6) data, which is a collection of different ESM models, will take 28 petabytes of storage [1]. These storage demands increase the need for an effective compression algorithm, allowing for high compression without sacrificing important climate information, which would hinder accuracy on future climate predictions. In this paper, we utilize the strengths of Variational Autoencoders (VAEs) as a data compression and regeneration technique for ESM data. We propose using the VAE encoder as a compression method, the latent variable as a compressed representation, and the VAE decoder as the reconstruction model.

1 Introduction

ESMs are nondeterministic generative models that are useful for taking information about the current Earth system and applying it to predict past and future climate patterns [2]. They are composed of model components that simulate the individual parts of the climate system taking into account the physics, chemistry, and biology to simulate the transfer of energy and mass between land, ocean, atmosphere, and sea ice [2, 3]. This includes complete feedback cycles such as predicting extreme weather patterns like droughts or hurricanes and impacts on these fluctuations in how they affect other related variables [2]. Climate science is one of the most popular areas of research and ESMs are one of the main reasons why.

The Coupled Model Inter-comparison Project (CMIP) is an international climate modeling project with the aim of improving existing model simulations to better understand past, present, and future

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

climates [4]. CMIP6, the latest phase with the most up-to-date climate data, contains data from more than 100 ESMs that span over 50 modeling centers around the world [4]. Each new iteration of the CMIP phases brings improved protocols, standards, and data distribution mechanisms, but that does not mean it comes without cost [4]. CMIP6 alone has generated more than 28 PB of data due to growing global participation and complexity of model comparisons and experimental design [1]. This is a significant increase from previous phases (40 TB in CMIP3; 2 PB in CMIP5) with data compression techniques having grown in popularity to combat this rapidly expanding issue [1, 5]. The challenge then becomes, how much we can compress the data while keeping reconstruction error as low as possible? Since the problem involves time-series data, compressibility, and low reconstruction error, we naturally looked to VAEs.

Not only are VAEs strong in the aforementioned areas, but when compared to other compression model architectures, VAEs are simple and scalable in structure, easily allowing for adding features like attention and residual blocks to the model architecture [6]. For our specific problem, we focus our analysis on two separate CMIP6 ESMs, the sixth version of the Model for Interdisciplinary Research on Climate (MIROC6) and the Institute Pierre Simon Laplace (IPSL) ESMs [7, 8]. Additionally, to ensure our model is accustomed to the effects of climate change, we train on Representative Concentration Pathway (RCP) 8.5 and evaluate on RCP 4.5. These represent a high emission scenario where greenhouse gases continue to rise through 2100 and a moderate emission scenario where they begin to stabilize and decline from 2040 onward [9, 10]. Finally, our model performs compression and reconstruction on a daily scale. Future work could look to expand to monthly and yearly compression, respectively.

In this paper, we contribute the following: i) a VAE for large scale CMIP6 data compression and ii) an in-depth analysis into the reconstruction error for VAEs across differing variables, latent space dimensionality, seasons, and decades. The goal of the paper is to deepen our understanding over VAEs to determine their strengths and weaknesses with the CMIP6 data as well as their tradeoffs between data compression and reconstruction accuracy.

2 Related Work

2.1 VAEs

Autoencoders are a machine learning architecture that learns to reconstruct the input data by going through a neural network with a bottle neck [11]. This makes the model learn how to compress the given information in such a way where reconstruction loss is minimized. Autoencoders are made up of two models, the encoder, which learns how to map data x into latent variable z, and the decoder, which learns how to take the latent variable z and regenerate data x.

VAEs are a type of autoencoder that introduces a probabilistic structure to the latent dimension [6]. The VAE objective is to model $p(x)$ using latent variables as defined as $p(x, z) = p(z)p(x|z)$ where $p(x, z)$ is the joint probability of the given data x and latent variable z, $p(z)$ is the prior distribution over the latent variable, and $p(x|z)$ is the likelihood function that generates data x from latent variable z. Since $p(x|z)$ is unknown, we will train a model to learn the approximate distribution which will be the VAE decoder. For the encoder we will approximate $p(z|x)$ by using a neural network $q(z|x)$. The marginal likelihood of $p(x) = \int p(z)p(x|z)\, dz$ is intractable, so VAEs optimize the lower bound $log(p(x)) \geq \mathbb{E}_{q(z|x)}[log(p(x|z))] - D_{KL}(q(z|x)||p(z))$.

2.2 Other Generative Climate Modeling

Other the years there have been many attempts at using generative machine learning on ESM data. One example that has worked well is using generative diffusion models to emulate ESM predictions. These models take in a conditioning map of yearly or monthly averages and are able to generate realistic daily samples. This approach takes significantly less computational time and energy as the models only take a few days to train and a few hours to generate new realizations after training where ESMs would take weeks to months for each realization. Being able to generate data faster is very important since ESMs and generative models are non-deterministic, so each run is averaged to find the likelihood of extreme weather. For more information on the diffusion model, check out this paper [12].

3 Methods

In this section, we train a VAE to perform daily compression and reconstruction of both MIROC6 and IPSL climate data. To achieve this, we utilize both attention and residual networks in the encoder and decoder as well as data preprocessing techniques for model stability and efficiency.
We separate our training, validation, and test splits uniquely to ensure clear separation between each set. The model is trained on RCP8.5 data only, while RCP4.5 is used as our test set. This maximizes the model's understanding of climate change behaviors while recognizing that current environmental awareness and policies indicate that emissions may stabilize before 2100. The model is trained on data from realizations 3, 4, and 6, while the model is always evaluated on realization 2 data. CMIP6 data is initially stored as .nc chunk files, which we load using xarray's. We pre-process the data into standardized precipitation and temperature metrics and then apply data normalization to transform the data to a predictable range for our model.

3.1 Model Architecture

Our VAE contains an encoder to encode the data into a compressed latent space, which we have set as a hyperparameter for tuning, and a decoder for data reconstruction. The encoder is composed of three distinct blocks: the input, middle, and output blocks. The input block contains a 2D convolutional layer to learn more complex feature representations of the data. Then, four layers of downsampling blocks each containing two residual network blocks and a single convolutional layer with a stride of 2 to downsample the data. The middle block is structured with an attention block and two residual network blocks. Performing attention allows our VAE to capture information from distant features that our convolutional layer may miss, improving data reconstruction. The residual network is crucial for gradient stability in our model and is also applied in the input block to down-sample in layers as opposed to all at once. The output block applies group normalization [13] which works better for our small batch size of 16, the Sigmoid Linear Unit (SiLU) activation function [14] and a final convolutional layer to create our latent representation of the data. The decoder is similar in structure, taking the latent representation and then applying a convolutional layer followed by a mid-block, upsampling blocks, and finally reconstructs the data through its output block. Since our task is data reconstruction rather than generation, rather than creating a complex decoder, we emphasize having a strong encoder that preserves key information and a decoder that recreates the original input in a symmetrical way.

3.2 Evaluation

For model evaluation we treat the decoder as our generator and implement a discriminator to create a Variational Autoencoder Generative Adversarial Network (VAE-GAN) [15, 16, 17]. While GAN style architectures often yield lower numerical precision [18, 19], we balance this by utilizing a running loss containing Kullback-Leibler divergence, Mean Absolute Error (MAE), and discriminator weights. By using VAE-GAN model evaluation, we can maximize the visual sharpness of our reconstructed images while minimizing the numerical error in the reconstruction. In addition to our running loss we evaluated our model's performance using the Kolmogorov-Smirnov Test (KL test) [20] to compare the likelihood of both original and reconstructed data coming from the same distribution.

4 Results

In this section we will discuss our findings from the VAE compression on ESM data. These results are all computed using RCP 4.5 and both IPSL and MIROC6 data were used. All results for Near-Surface Air Temperature (TAS) are reported in degrees Celsius and Precipitation (PR) reported in millimeters per day (mm/day).

Figure 1 and Figure 2 histograms show the variation between the difference of our compressed TAS samples to a TAS random realization using the same RCP, from IPSL or MIROC6 respectively, shown in orange, and the original data with difference to the same random realization shown in blue. These analyses are done for four different scenarios i) Average Monthly Hot Days, ii) Average Monthly Temperature, iii) Average 90th Quantile, and iv) Average Monthly Hot Streak.

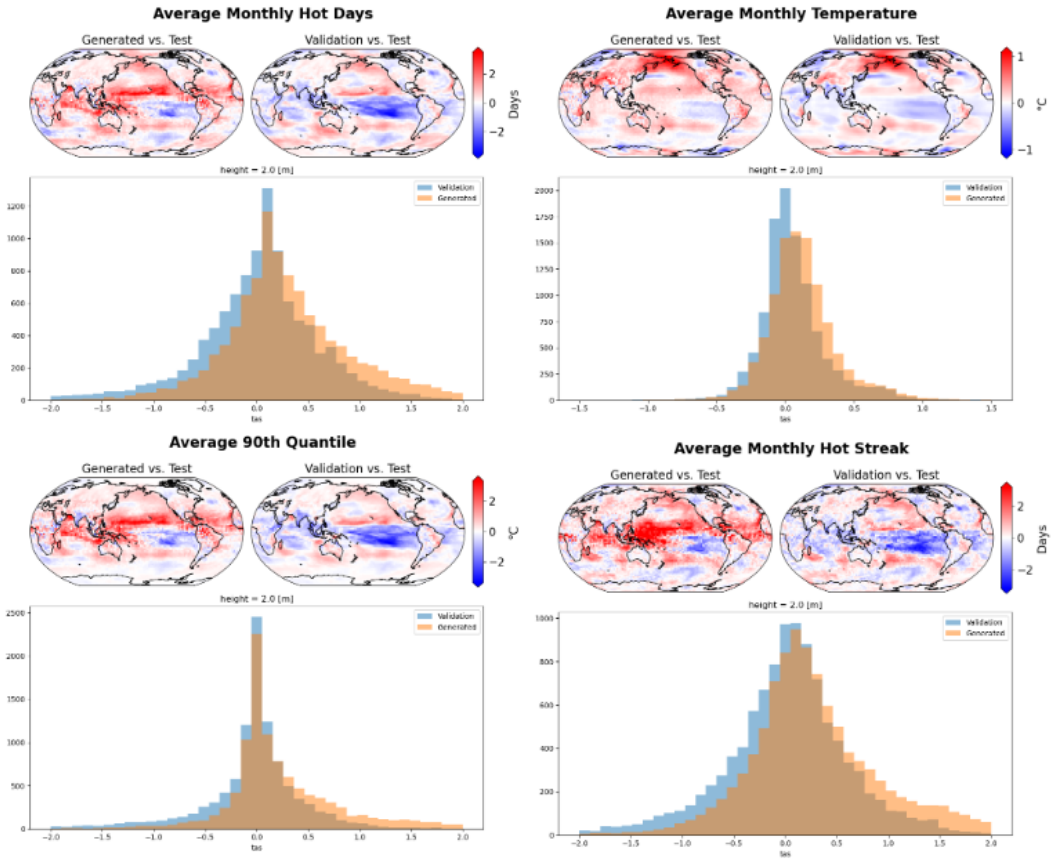

Figure 1: Difference histograms on the compression and reconstruction of realization 2 compared against realization 1 on MIROC6 RCP 4.5 generating TAS for the years 2080-2100.

Figure 3 and Figure 4 histograms show the variation between the difference of our compressed PR samples to a random PR realization using the same RCP, from IPSL or MIROC6 respectively, shown in orange, and the original data with difference to the same random realization shown in blue. These analyses are done for four different scenarios i) Average Monthly Precipitation, ii) Average Simple Daily Intensity Index (SDII), iii) Average Rainy Streak, and iv) Average Rainy Days.

Figure 5 and Figure 6 show the Kolgomorov-Smirnov (KS) Test for IPSL and MIROC6 in both TAS and PR. This test is used to show whether or not two data samples come from the same distribution by comparing the cumulative distribution functions. The score and image on the left shows the variation between the difference of our compressed samples to a random realization using the same RCP, from IPSL or MIROC6 and TAS and PR respectively, and the right image and score shows the original data with difference to the same random realization shown in blue.

Figure 7 shows our compressed and reconstructed predictions for TAS for both IPSL and MIROC6 data in 2080-2100 in blue, compared to the the predictions from the original data before being compressed.

Figure 8 shows our compressed and reconstructed predictions for pr for both IPSL and MIROC6 data in 2080-2100 in blue, compared to the predictions from the original data before being compressed.

Figures 9 and 10 are our difference histograms and provide a reference into our reconstruction performance over decades and seasons. We do observe any particular trends across decades in our analysis, highlighting our model's robustness in reconstructing data long-term. However, we notice that our model performs worse in the Winter and especially in the Summer months. This is more evident in our temperature figures, where it often reconstructs hotter temperatures. This is even more apparent once the latent dimension shrinks or expands. This is one of many figures that allowed us to

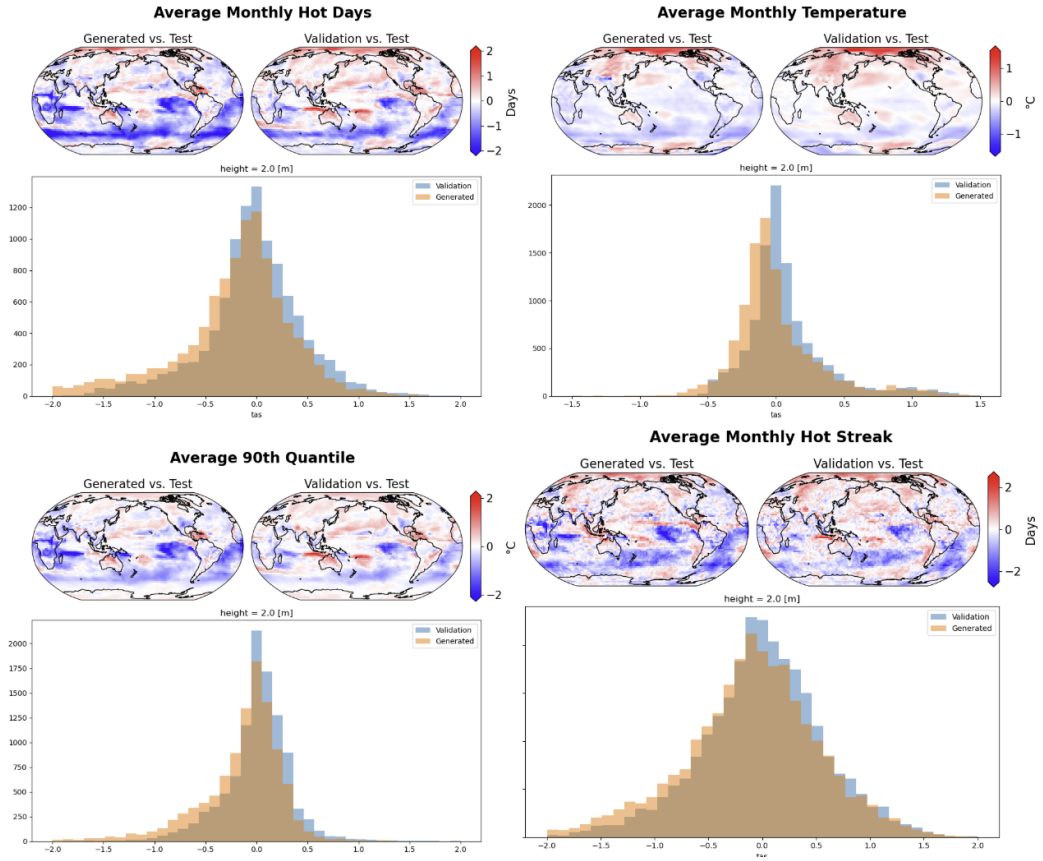

Figure 2: Difference histograms on the compression and reconstruction of realization 2 compared against realization 1 on IPSL RCP 4.5 generating TAS for the years 2080-2100.

identify how much we can compress our data before months with extreme climates began to have worse reconstructions.

5 Conclusion & Future Work

In this paper, we have shown that VAEs are a better solution to the data storage problem for ESM CMIP6 data than other lossy data compression like JPEG. By using that latent variable as a compressed learned representation of the daily climate predictions, we are able to compress and regenerate ESM temperature and precipitation daily data with minimal reconstruction error. We have also shown that our model generalizes to unseen realizations, making our modeled more storage efficient, since less model weights have to be stored.

Future work that could be done with VAEs in ESM storage would be to leverage the dependencies of similar days. By compressing 28 day sequences or months together, we could take advantage of time temporal dependencies possibly allowing for further compression results. Another way to utilizes dependencies within the ESM data is to compress multiple variables together. Past work in jointly generating precipitation and temperature for ESM emulators has shown to improve results [21]. That conclusion could be applied to data compression allowing for smaller overall latent spaces for the combined variables as information would not have to be encoded and stored twice.

Another avenue to explore is latent diffusion [22]. As diffusion models are currently the state of the art in image modeling, there has been extensive work in applying diffusion to ESM emulation. With our new exploration of data compression using VAEs, these diffusion models have the ability to accelerate computational time and energy even more.

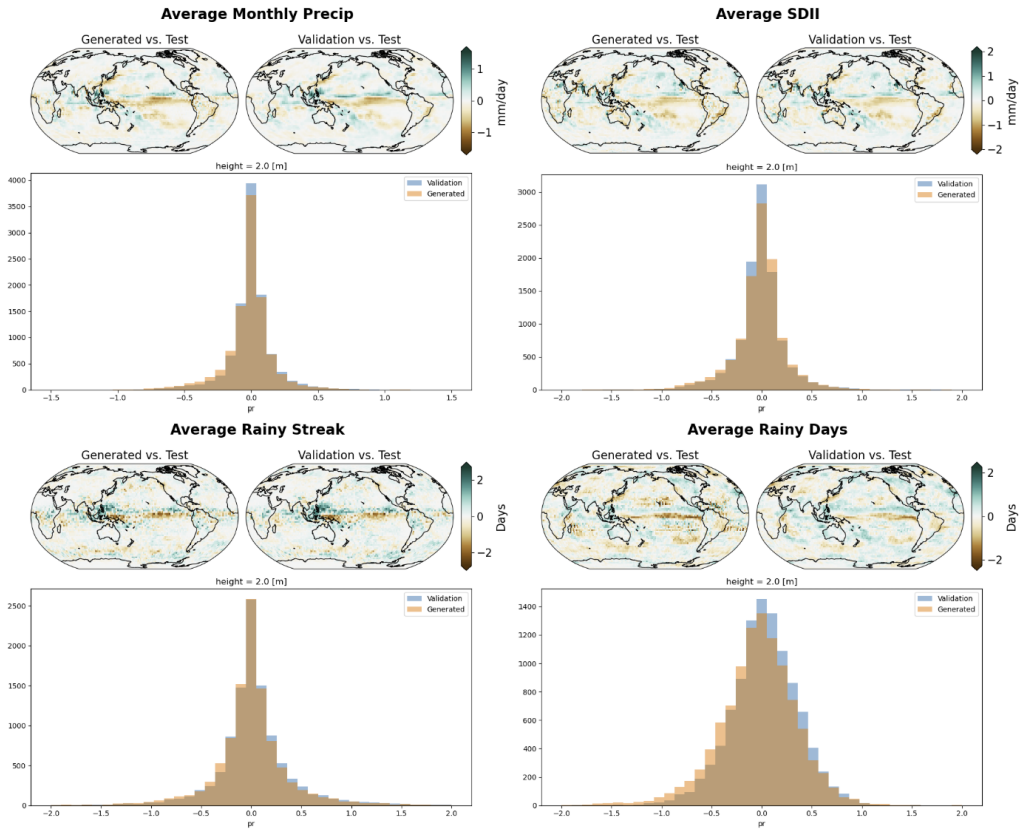

Figure 3: Difference histograms on the compression and reconstruction of realization 2 compared against realization 1 on MIROC6 RCP 4.5 generating PR for the years 2080-2100.

References

[1] Sameh Abdulah, Allison H. Baker, George Bosilca, Qinglei Cao, Stefano Castruccio, Marc G. Genton, David E. Keyes, Zubair Khalid, Hatem Ltaief, Yan Song, Georgiy L. Stenchikov, and Ying Sun. Boosting Earth System Model Outputs And Saving PetaBytes in their Storage Using Exascale Climate Emulators, August 2024. arXiv:2408.04440 [stat].

[2] Studying and Projecting Climate Change with Earth System Models | Learn Science at Scitable. Cg_cat: Studying and Projecting Climate Change with Earth System Models Cg_level: MED Cg_topic: Studying and Projecting Climate Change with Earth System Models.

[3] thomas harrisson. CMIP6: the next generation of climate models explained, December 2019.

[4] CMIP Overview - Coupled Model Intercomparison Project, April 2023.

[5] Mrunmayee Dhapre. Using Variational AutoEncoders (VAE) for Time-Series Data Reduction, September 2024.

[6] Max Welling Diederik Kingma. Auto-encoding variational bayes, 2022.

[7] Hiroaki Tatebe, Tomoo Ogura, Tomoko Nitta, Yoshiki Komuro, Koji Ogochi, Toshihiko Takemura, Kengo Sudo, Miho Sekiguchi, Manabu Abe, Fuyuki Saito, Minoru Chikira, Shingo Watanabe, Masato Mori, Nagio Hirota, Yoshio Kawatani, Takashi Mochizuki, Kei Yoshimura, Kumiko Takata, Ryouta O'ishi, Dai Yamazaki, Tatsuo Suzuki, Masao Kurogi, Takahito Kataoka, Masahiro Watanabe, and Masahide Kimoto. Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, 12(7):2727–2765, July 2019. Publisher: Copernicus GmbH.

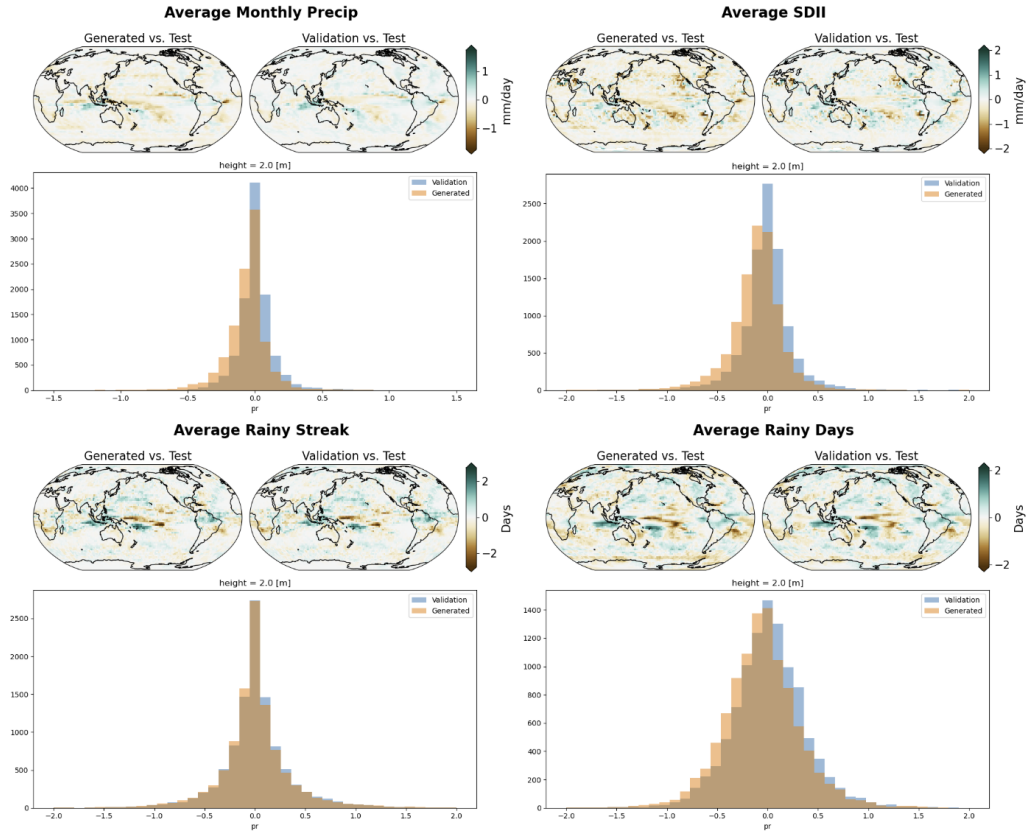

Figure 4: Difference histograms on the compression and reconstruction of realization 2 compared against realization 1 on IPSL RCP 4.5 generating PR for the years 2080-2100.

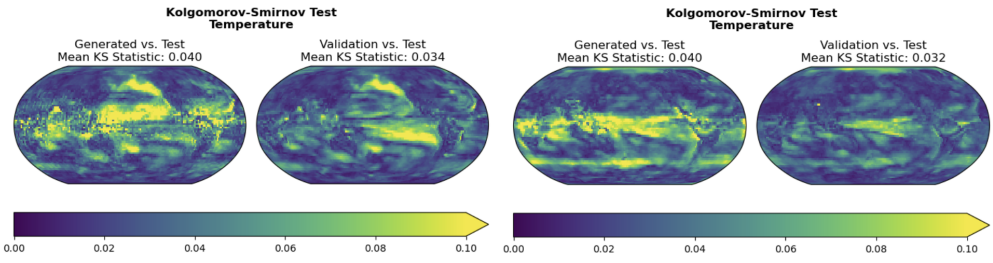

Figure 5: KS tests for MIROC6 (left two figures) and IPSL (right two figures) on RCP 4.5 for TAS for the years 2080-2100. Left figure shows the difference between realization 1 after compression and regeneration compared to realization 2 and the right figure is the original realization 1 compared to realization 2.

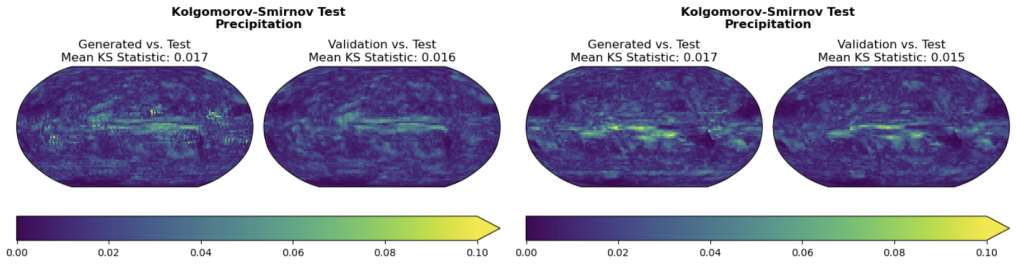

Figure 6: KS tests for MIROC6 (left two figures) and IPSL (right two figures) on RCP 4.5 for PR for the years 2080-2100. Left figure shows the difference between realization 1 after compression and regeneration compared to realization 2 and the right figure is the original realization 1 compared to realization 2.

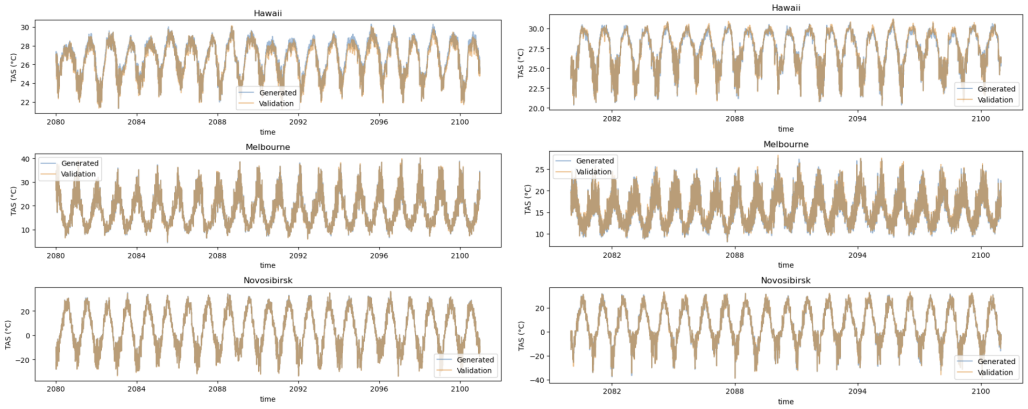

Figure 7: Daily predictions in Hawaii USA, Melbourne Australia, and Novosibirsk Russia using RCP 4.5 on MIROC6 (right) and ISPL (left) TAS for the years 2080-2100. Compressed and regenerated from realization 1 data in blue and original realization 1 data in orange.

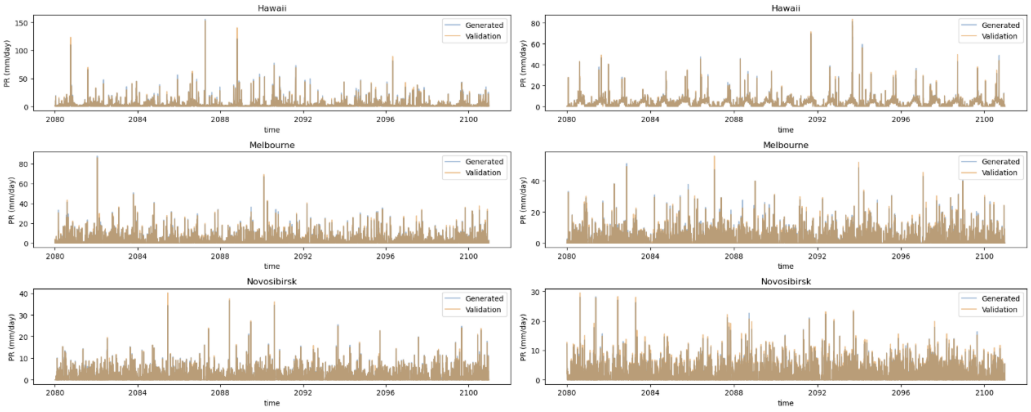

Figure 8: Daily predictions in Hawaii USA, Melbourne Australia, and Novosibirsk Russia using RCP 4.5 on MIROC6 (right) and ISPL (left) PR for the years 2080-2100. Compressed and regenerated from realization 1 data in blue and original realization 1 data in orange.

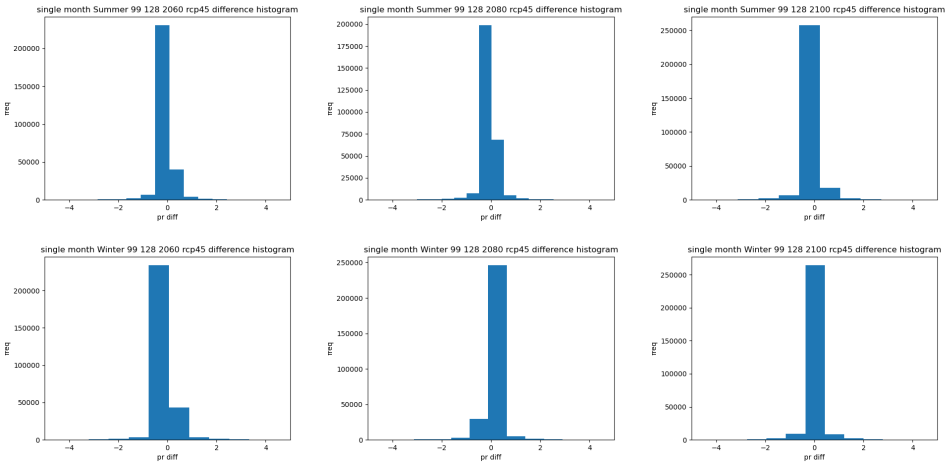

Figure 9: Precipitation geospatial difference histograms in mm. Trained on RCP8.5 data and evaluated on RCP4.5 data. Columns represent decades 2060, 2080, 2100, and rows represent season with the top row being Summer and bottom Winter.

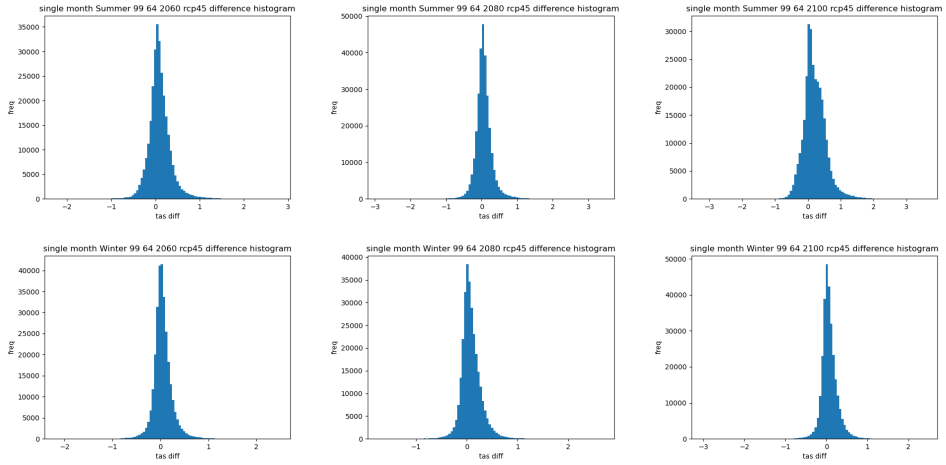

Figure 10: Temperature geospatial difference histograms in degrees celsius. Trained on RCP8.5 data and evaluated on RCP4.5 data. Columns represent decades 2060, 2080, 2100, and rows represent season with the top row being Summer and bottom Winter.

[8] IPSL-CM6A-LR – My CMS.

[9] Allison M. Thomson, Katherine V. Calvin, Steven J. Smith, G. Page Kyle, April Volke, Pralit Patel, Sabrina Delgado-Arias, Ben Bond-Lamberty, Marshall A. Wise, Leon E. Clarke, and James A. Edmonds. RCP4.5: a pathway for stabilization of radiative forcing by 2100. *Climatic Change*, 109(1-2):77–94, November 2011.

[10] thomas harrisson. Explainer: The high-emissions 'RCP8.5' global warming scenario, August 2019.

[11] R.J. Williams D.E. Rumelhart, G. E. Hinton. Learning internal representations by error propagation, 1986.

[12] Seth Bassetti, Brian Hutchinson, Claudia Tebaldi, and Ben Kravitz. DiffESM: Conditional Emulation of Temperature and Precipitation in Earth System Models with 3D Diffusion Models, September 2024. arXiv:2409.11601 [physics].

[13] Gousia Habib, Ishfaq Ahmed Malik, Jameel Ahmad, Imtiaz Ahmed, and Shaima Qureshi. Exploring the Efficacy of Group-Normalization in Deep Learning Models for Alzheimer's Disease Classification, April 2024. arXiv:2404.00946 [cs].

[14] SiLU — PyTorch 2.7 documentation.

[15] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric, February 2016. arXiv:1512.09300 [cs].

[16] Yuchao Liao, Tosiron Adegbija, Roman Lysecky, and Ravi Tandon. Skip the Benchmark: Generating System-Level High-Level Synthesis Data using Generative Machine Learning. In *Proceedings of the Great Lakes Symposium on VLSI 2024*, pages 170–176, June 2024. arXiv:2404.14754 [cs].

[17] Yichi Zhang, Paul Seibert, Alexandra Otto, Alexander Raßloff, Marreddy Ambati, and Markus Kästner. DA-VEGAN: Differentiably Augmenting VAE-GAN for microstructure reconstruction from extremely small data sets, February 2023. arXiv:2303.03403 [cs].

[18] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network, May 2017. arXiv:1609.04802 [cs].

[19] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch. EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis, July 2017. arXiv:1612.07919 [cs].

[20] Kolmogorov–Smirnov test, May 2025. Page Version ID: 1289549258.

[21] Katie Christensen, Lyric Otto, Seth Bassetti, Claudia Tebaldi, and Brian Hutchinson. Diffusion-Based Joint Temperature and Precipitation Emulation of Earth System Models, April 2024. arXiv:2404.08797 [physics].

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, April 2022. arXiv:2112.10752 [cs].