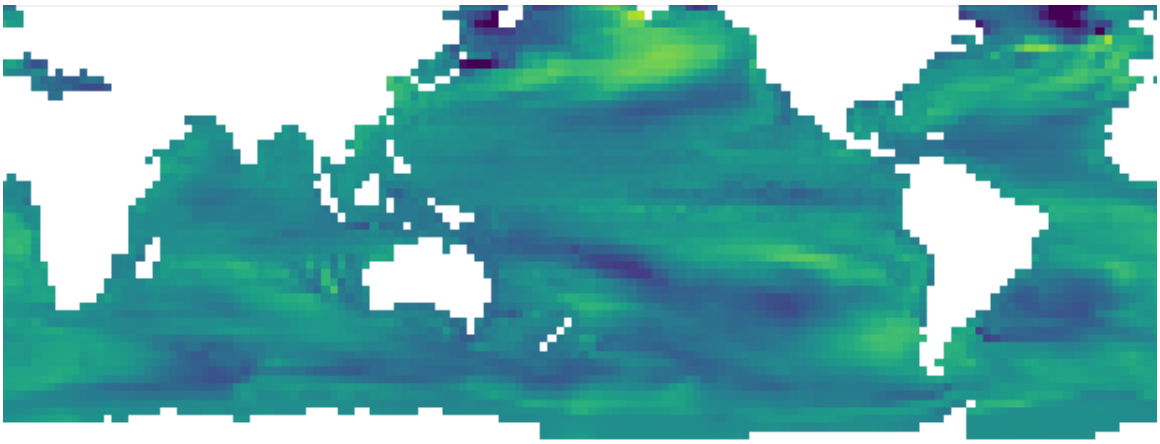
EPFL

Swiss Federal Institute of Technology

Generative models for robust global warming projections



Author:
Luc Harrison

Supervisor:
Victor Cohen

May 31, 2025

Abstract

This project tackles the challenge of estimating the forced responses of sea surface temperature (SST) anomalies from a single, noisy climate simulation. This is an important problem due to the lack of existing data and high computational cost of generating new samples. By framing the problem as a denoising and pattern extraction task, this project aims to apply both linear methods such as Reduced-Rank Regression (RRR) and a novel Variational Autoencoder architecture, TrendVAE, designed to capture the dominant spatial and temporal patterns in the data. The results of this project demonstrate that Reduced-Rank Regression effectively extracts the main forced response trends with strong robustness despite data noise and limited sample sizes, providing a reliable linear baseline. Meanwhile, TrendVAE successfully captures more complex, nonlinear spatial and temporal patterns by incorporating convolutional layers, however the predictions remain noisy due to the lack of data for training. Overall, both methods achieved significant reductions in noise, yielding more robust forced response predictions. Both these methods highlight the potential for improving climate model predictions as well as generating new, high quality data that can advance future climate research.

Contents

1 Introduction

1.1 Problem Setting

Projections of global warming over the 21st century remain highly uncertain, posing challenges for both scientific understanding and future planning around the globe. A central difficulty lies in separating the part of climate change driven by external factors (forced response) from the natural fluctuations within the climate system (internal variability).

Climate models are used to simulate the evolution of sea surface temperatures (SST) and other climate variables over time. Each model produces runs, which are individual simulations that include both the systemic warming pattern due to external factors (the forced response) and natural, chaotic fluctuations (internal variability). We assume that the forced response of a given model corresponds to the average across all its runs, effectively canceling out internal variability and yielding a robust estimate of the model's true response to external forcing. However, generating enough runs to perform this averaging is computationally expensive, and in many cases, only a few runs or even a single run may be available.

The objective of this project is to develop generative models that can accurately predict the forced response of an unknown model using only a single noisy simulation (i.e a single run). Formally, each run is discretized over d spatial grid cells and over T time steps. For each climate model $m \in \{1, \dots, M\}$, we have access to a certain number of runs (which varies with each model), which will be indexed as r.

$x_{i,t}^{m,r}$ is the SST anomaly at time t, grid cell i, for model m and run r. Therefore, since the runs generated by the models are just the forced response with some unknown additive noise, the run data can be written as:

$$x_{i,t}^{m,r} = y_{i,t}^{m} + v_{i,t}^{m,r}$$

Where:

- $y_{i,t}^{m}$ is the forced response of model m at time t and grid cell i
- $v_{i,t}^{m,r}$ is the internal variability (the noise) specific to run r of model m

The goal is to learn a warming pattern function $f_\theta(\cdot)$ that, for an unknown time series $x \in \mathbb{R}^{d \times T}$, that predicts a robust estimate of the underlying forced response $y \in \mathbb{R}^{d \times T} : \hat{y} = f_\theta(x)$

To summarize, we aim to solve the following optimization problem:

$$\min_{\theta} \; L(Y, f_\theta(X)) + \Omega(\theta)$$

Where L is a suitable loss function measuring the discrepancy between the predicted and true forced responses, and $\Omega(\theta)$ is a regularization term to ensure model robustness. The key challenge is that the model must recover the forced response from only a single realization, effectively de-noising the input data.

1.2 Data and Preprocessing

The dataset used in this project comes from CMIP6 [2], which provides sea surface temperature (SST) anomaly simulations from a range of global climate models under historical and future forcing scenarios. For each model m, there is a set of simulation runs (the number of which depends on the model used), where each of these runs is represented by a spatio-temporal matrix over d grid cells and T time steps.

To ensure that training and evaluation are numerically stable, grid cells located above 60°N latitude are removed. These polar regions are characterized by persistent low SST temperature variability, often to temperatures being near the freezing point. As noted in [6], this low variability can lead to numerical instability when computing metrics such as the normalized mean squared error, where very small standard deviations can amplify errors disproportionately.

Additionally, in this project only models with at least 4 runs are retained. This is because the forced response of a model is assumed to be the average over the runs for that given model, and a sufficiently large number of runs is necessary to reduce the influence of internal variability through averaging (under the assumption that the internal variability is zero mean).

In total, the data from 34 separate climate models was used to train the models.

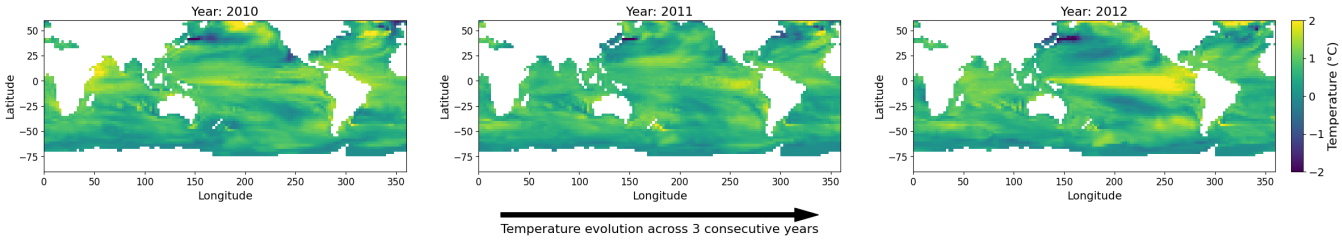


Figure 1: Example of the forced response shown as consecutive spatial maps over time. The arrow indicates the direction of temporal evolution over 3 years.

2 Literature

Accurately estimating the forced component of climate change is a difficult task, especially due to the inherent internal variability present in each model, which itself is often unknown. In [5], the authors introduced a robust statistical learning approach utilizing anchor regression to extract a climate change fingerprint that is robust to variations in internal climate variability. Their findings indicate that it is highly likely that at least 85% of the observed warming over the past 40 years is attributable to external forcing.

In the field of generative modeling, diffusion models have emerged as powerful tools for data generation tasks, particularly for tasks requiring the synthesis of complex, often high-dimensional data. In the context of climate science, acquiring such data is very costly and time consuming. To combat this, the authors in [4] introduce a diffusion-based model that emulates weather forecasts by leveraging historical data, with the objective to generate large ensembles of high quality data.

Extending generative modeling to time-series data, TimeVAE, presented in [1], introduces a Variational Auto-Encoder based architecture that incorporates multiple, interpretable temporal structures - such as trend and seasonality - into the generation process. This design enables the generation of realistic time-series data while providing insights into the temporal patterns.

3 Models

3.1 Reduced-Rank Regression

Reduced-Rank Regression, as introduced in [3], is a method for estimating the regression coefficient matrix under a rank constraint, making it well-suited for multivariate linear models with high-dimensional data and potential multicollinearity. In the context of climate datasets—where outputs are spatial temperature fields with thousands of grid cells—RRR reduces model complexity and enhances interpretability by projecting the response onto a lower-dimensional subspace that captures the dominant patterns of the signal.

Mathematically, given paired data matrices $X \in \mathbb{R}^{T\times d}$ (inputs) and $Y \in \mathbb{R}^{T\times d}$ (outputs), and some **unknown** noise ε:

$$Y = XB + \varepsilon$$

Where $X \in \mathbb{R}^{T\times d}$ represents the input features (i.e the climate simulation realizations), $Y \in \mathbb{R}^{T\times d}$ the forced response of the model and $\hat{B}_{\text{OLS}} \in \mathbb{R}^{d\times d}$, the Ordinary Least Squares solution to the above problem, which is given by:

$$\hat{B}_{\text{OLS}} = (X^\top X)^{-1} X^\top Y$$

While OLS estimates the full-rank coefficient matrix, it can lead to overfitting and poor generalization in higher dimensions (which is our case due to the high number of grid cells). To address this, Reduced-Rank Regression (RRR) constrains the rank of the regression by solving the following problem:

$$\min_B \|Y - XB\|_F^2,$$

Where $||\cdot||_F$ is the Frobenius norm. In this problem setting we assume that B has a rank of $r \leq d$ Then by involving the OLS solution and performing SVD on the term $X\hat{B}_{\text{OLS}}$ and finally truncating the first r singular vectors, the following closed form solution for RRR is derived:

$$\hat{B}_{\text{RRR}} = \hat{B}_{\text{OLS}} V_r V_r^\top$$

Where $V_r \in \mathbb{R}^{d\times r}$ contains the top r right singular vectors. This gives the rank-r matrix that best approximates the OLS solution. The idea of using RRR comes from the fact that we aim to model the trend of the SST and capture the essential patterns in the data by constraining the rank of the matrix.

3.2 Variational Autoencoders

To complement the linear approach described above, we also consider Variational Autoencoders (VAEs) [1], which can essentially be considered as a non linear version of the Reduced-rank Regression. This class of model can learn non-linear representations of high-dimensional data, while being able to generate new, realistic samples from randomly sampled inputs. In the context of climate science, VAEs are particularly pertinent when it comes to learning compact latent representations of data, such as temperature fields or other geological patterns, enabling uncertainty aware predictions. The forced responses are given by a multivariate time series $Y \in \mathbb{R}^{T\times d}$, where T denotes the number of time steps and d the number of grid cells.

The VAE aims to approximate the true data distribution $p(Y)$, however this distribution is unknown. To tackle this problem, the VAE introduces latent variables $z \in \mathbb{R}^k$(with k $\ll d$), which serves as a lower-dimensional representation of the data. The model operates in two steps:

- **Encoding:** the input X is passed through a probabilistic encoder $q_\phi(z|X)$ which maps the high dimensional time series to a distribution over latent variables. Rather than returning a single point, the encoder outputs the parameters of a Gaussian distribution (mean and variance).
- **Decoding:** a decoder $p_\theta(Y|z)$ then attempts to reconstruct the forced response from the sampled latent variable. This allows the model not only to compress the data into a meaningful latent space but also to generate new data samples by sampling $z \sim \mathcal{N}(0, I)$ and decoding them.

The training of a VAE is performed by maximizing the Evidence Lower Bound (ELBO) on the log-likelihood of the data:

$$\mathcal{L}_{\theta,\phi} = -\mathbb{E}q_\phi(z|Y)[\log p_\theta(Y|z)] + \mathrm{D}_{\mathrm{KL}}(q_\phi(z|Y)\|p(z))$$

Where $\mathrm{D}_{\mathrm{KL}}(q_\phi(z|Y)\|p(z))$ is the KL-divergence between 2 distributions. The first term encourages the decoder to accurately reconstruct the input time series from the latent variable z, and the second term regularizes the encoder to remain close to the prior distribution (typically a unit Gaussian).

Building on the TimeVAE architecture proposed in [1], which introduces interpretability into the latent space and aims to capture components such as trends and seasonality, we developed a specialized model named TrendVAE. This variant is designed to capture and extract long-temporal trends in the climate data by using convolutions as well as a dedicated "Trend block".

This Trend block models the underlying trend in the data by assuming that it is a combination of polynomial basis functions whose coefficients are estimated by the latent representation. This approach allows the model to learn interpretable trends, for example depending on which coefficient is larger, we could assume that the model is learning that is a more linear or quadratic increase/decrease in temperature.

By structuring the latent space and using 1D convolutions, TrendVAE can learn global spatial and temporal patterns and represent the noisy input data in a more robust way, representing clear trends over time.

4 Results and discussion

The performance of three models is evaluated: the Reduced Rank Regression (RRR), a baseline Variational Autoencoder (which doesn't take into account the time dependency of the problem) and finally the custom-made TrendVAE. Each model is evaluated based on its ability to reconstruct forced response trajectories Y from the simulation runs X, as well as its capacity to generate new runs (in the case of the VAEs).

4.1 Reduced Rank Regression (RRR)

RRR was evaluated using a leave-one-out strategy: the model was trained on $m-1$ climate models and tested on the held-out one. To select the best hyperparameters - the rank and the regularization parameter λ, a grid search was performed and the best parameters were the ones that minimized the following objective:

$$\min_{\lambda,\text{rank}} \left(\Lambda_{MSE} \cdot \mathbb{E}[\text{MSE}] + \Lambda_{std} \cdot \text{std}(\text{MSE})\right)$$

Where Λ_{MSE} and Λ_{std} were the values used to balance the trade-off between average performance and robustness across models. In our case, we set $\Lambda_{MSE} = 0.7$ and $\Lambda_{std} = 0.3$. to prioritize minimizing the average reconstruction error while still accounting for variability across the m training and testing splits. This selection strategy encourages both accuracy and generalizability by avoiding parameter sets that would for example perform extremely well on certain models and very poorly on others (which could have a very low average MSE but generalize poorly). The best overall combination of rank and λ was then trained on all m splits of the dataset (trained on m-1 models and tested on the one that was left out), the results can be seen in Figure 2.

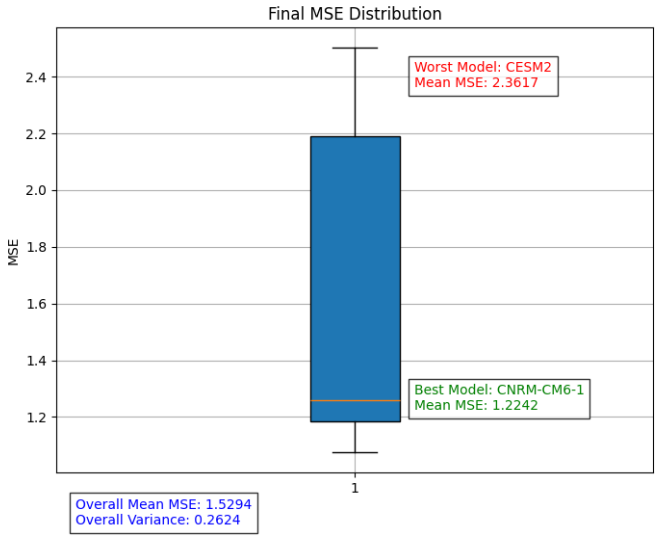


Figure 2: MSE distribution of the model trained on m-1 models and evaluated on the m^{th} model (using the best combination of λ and rank ($rank = 10$, $\lambda = 100$).

4.1.1 Normalization and evaluation metric

Initially, the training data was standardized by removing the mean and dividing by the standard deviation per model. For the test data, global standardization was used, applying the mean and standard deviation computed over the entire training set (i.e., all $m-1$ models combined).

This approach posed issues in training: some training models had very few samples, and certain grid cells (especially in polar regions) exhibited near-zero variance, making standardization not statistically relevant and leading to stability issues in training.

To mitigate this, we excluded all grid cells with latitudes above 60°. This filtered out regions with unreliable standard deviation estimates and improved numerical stability.

After removing these regions, we only centered the training data (i.e., subtracted the mean) without dividing by the standard deviation during training. For evaluation, we used a normalized mean squared error (NMSE), defined as:

$$\text{NMSE} = \frac{1}{T \cdot d} \sum_{t=1}^{T} \sum_{i=1}^{d} \left(\frac{\hat{Y}_{t,i} - Y_{t,i}}{\sigma_i} \right)^2$$

Where σ_i is standard deviation for the i^{th} grid cell. Given the very high variability present in the inputs (i.e present in each run), the RRR, despite being a simple linear model, was able to capture consistent spatial and temporal patterns. As illustrated in Figure 3, which shows the time series prediction at a relatively central grid cell, the RRR model effectively reduces the variability of the observed runs. This smoothing shows the model's ability to extract the forced response from noisy data even when constrained with a low rank (in our case $rank = 10$).

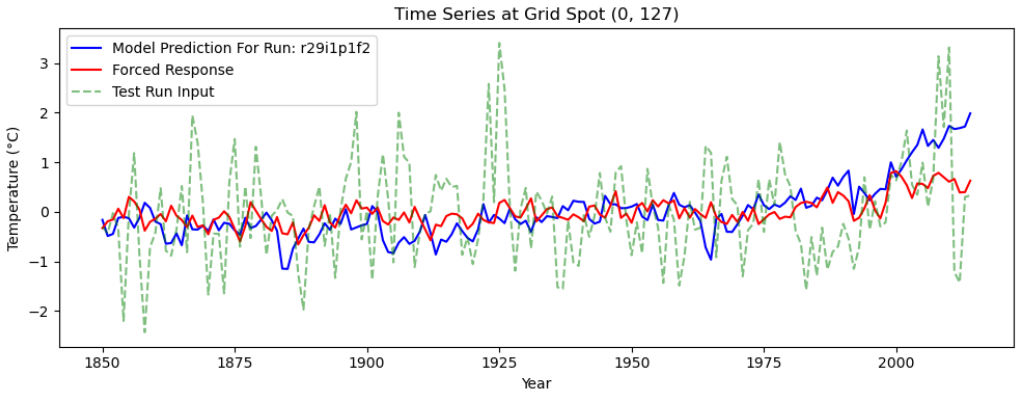


Figure 3: RRR model trained on $m-1$ models and tested on a run selected from model MIROC-ES2L (using the best combination of λ and rank ($rank = 10$, $\lambda = 100$).

Furthermore, in Figure 4, we display the predictions of the RRR model for all individual runs along with the ground truth forced response of the model. While the individual runs retain a certain level of variation, they consistently follow the overall trend of the forced response, particularly showing a significant warming pattern towards the end of the time series This emerging pattern is consistent with our current observations of global warming due to anthropogenic factors, suggesting that the RRR model is successful in isolating this long-term warming pattern given a single noisy input of a model.

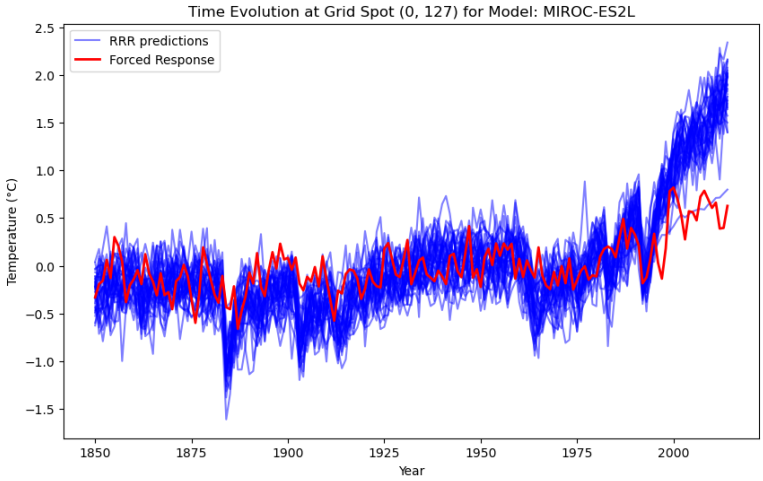


Figure 4: RRR model trained on $m-1$ models and tested on all the runs from model MIROC-ES2L (using the best combination of λ and rank ($rank = 10$, $\lambda = 100$).

Finally, Figure 5, highlight the spatial predictions from the RRR model, showing that even with a reduced rank ($rank = 10$), the model is able to capture and retain characteristic spatial structures in its predicted forced response. For instance, we observe a horseshoe-like warming pattern in the Atlantic Ocean, which corresponds to oceanic circulation, demonstrating RRR's ability to extract meaningful spatial features despite its linear nature.

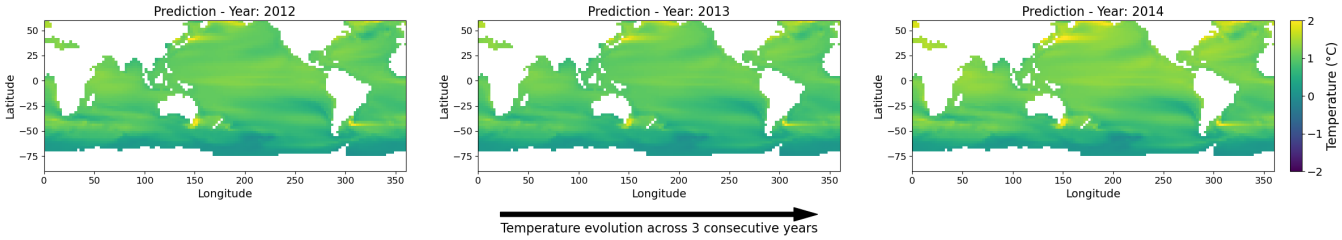


Figure 5: RRR model trained on $m-1$ models and tested on all the runs from model MIROC-ES2L (using the best combination of λ and rank (rank $= 10$, $\lambda = 100$)).

Despite its simplicity, RRR fails to capture a meaningful temporal variability (as can be observed in the videos in the notebook RRR.ipynb), the patterns stay relatively consistent throughout time but do not move around the globe (even if they do in the input data).

Overall, RRR successfully captures consistent spatial patterns and the long-term forced response despite high input variability. While it struggles with temporal dynamics, it serves as a strong, interpretable baseline for modeling large-scale climate trends.

4.2 TrendVAE

As described in 3.2, Variational autoencoders can essentially be viewed as a non-linear extension of RRR, by constraining the distribution of the latent space to staying near a $\mathcal{N}(0, 1)$

Normal distribution through the KL-divergence term. This makes them particularly well conditioned for modeling non-linear patterns, between the noisy input runs X and the target forced response Y. The objective being, to test, what advantage a non-linear model has over less computationally expensive methods such as the RRR described earlier.

The first VAE tested was a standard 1-layer VAE that treated the input as a 1-D vector, by flattening the temporal dimension. While this simple model was able to somewhat lead to the inputs becoming more robust, it lacked the ability to reconstruct smooth predictions and extract long-term trends.

Recognizing the importance of temporal structures, an upgraded version of the VAE was tested that simply preserved the temporal structure throughout the pipeline and latent space. While this upgraded version of the initial model yielded decent results, it still failed in effectively capturing the slow evolving forced signal and spatial patterns.

To address these limitations, TrendVAE, as described in 3.2 was created. This model puts special emphasis on the temporal structure of the data by incorporating a Trend block and keeping it throughout the latent space. This allowed the model to treat each time step separately, helping it learn temporal evolution in the latent space, leading to more time-consistent reconstruction. These improvements enabled the model to learn long-term temporal evolutions and spatial patterns present in the data (see Figure 6).

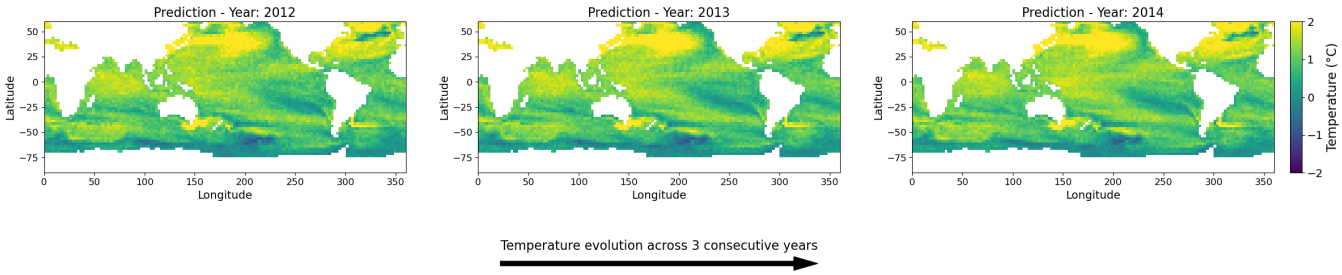


Figure 6: TrendVAE predictions for 3 consecutive years.

However, despite these improvements, VAE-based models still require large amounts of data and many training iterations, and often produce noisy outputs. This is partly due to the randomness from sampling in the latent space. One solution was to add a smoothness penalty between timestamps, but this overly constrained the model and reduced performance. Another simple and effective solution was simply to train the model for longer.

To evaluate the model's performance, the distribution of the mean squared error was examined. As with the RRR model, Leave-One-Out (LOO) cross-validation was employed to ensure robustness when testing on each of the models in the dataset (this was used when testing multiple architectures and learning rates for instance), see Figure 7.

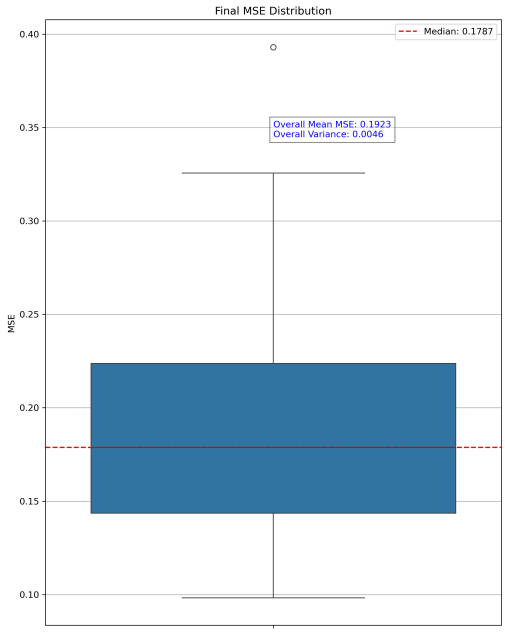


Figure 7: MSE distribution of the model trained on m-1 models and evaluated on the m^{th} model.

In terms of inference results, TrendVAE tends to overestimate the target response compared to the RRR, particularly in recent years (as observed in 8, when observing TrendVAE's prediction for a specific grid location). This likely occurs due to the model overfitting to certain patterns becoming more spurious in recent times.

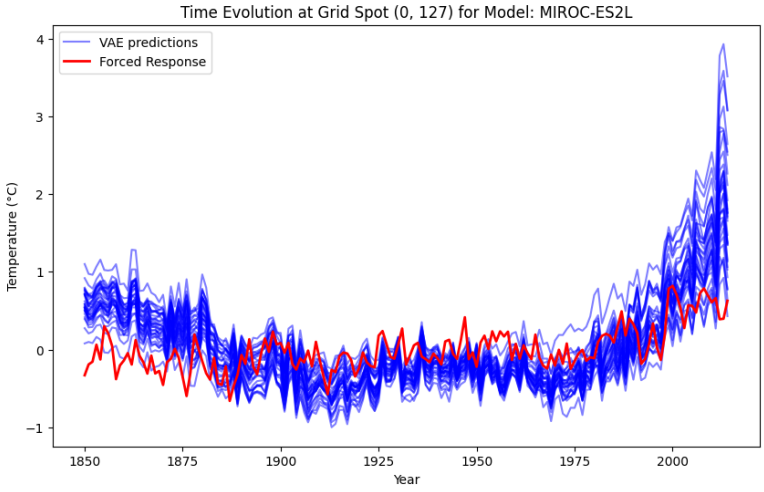


Figure 8: TrendVAE model trained on $m-1$ models and tested on all the runs from model MIROC-ES2L.

Nonetheless, TrendVAE can accurately capture the temporal and spatial dynamics (for instance, the same sort of horseshoe pattern in 6 can be observed as with the RRR predictions in 5). Beyond its predictive capability, TrendVAE can be used as a generative model by sampling directly from the latent space due to the reparameterization trick. By drawing random samples from $\mathcal{N}(0, 1)$ and decoding them to generate plausible realizations of forced responses.

As observed in 9, the generated forced responses present certain spatial patterns that we see when passing an input through the model, indicating that the model was able to learn similar patterns to those observed in the training data.

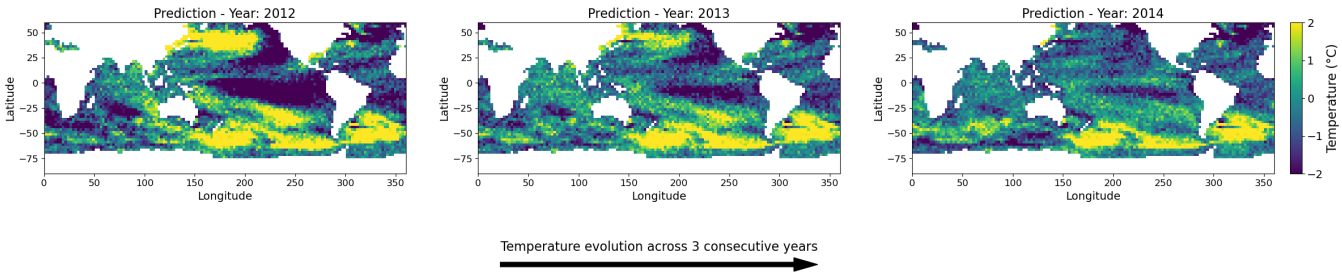


Figure 9: Generated predictions of TrendVAE by sampling the latent space.

This generative ability opens the door to new applications such as data augmentation, scenario exploration, or probabilistic forecasting—key advantages that linear methods like RRR cannot provide.

5 Conclusion and future work

This project explored the challenge of estimating the forced response of sea surface temperature (SST) anomalies from a single noisy climate simulation. By framing this as a denoising problem, both linear and non-linear generative models were tested to recover the robust warming patterns using limited observational data from 34 climate models.

Reduced-Rank Regression (RRR) proved to be a reliable method for extracting the dominant trend and despite it being a relatively simple method on paper, it yielded the most robust results overall, achieved through low-rank constraints. However in order to capture non-linear relationships, Variational Autoencoder based approaches were used. Building on the TimeVAE architecture (see [5]), the proposed TrendVAE incorporated convolutional layers and a trend block to force the model to learn long-term temporal patterns while enhancing both interpretability and performance.

Together, these methods demonstrate the potential of data driven techniques to generate robust predictions of forced responses for climate models, using only a single prediction. The overall results highlight the importance of domain knowledge as well as the different approaches to dealing with structured problems (for instance the issue of dealing with the temporal patterns and how to preserve them). One major obstacle encountered was the standardization of data, to finally resolve the issue, the data was centered and the standard deviation was used when evaluating model performance in the (Normalised Mean Square Error). Overall, the models developed during this project offer a promising direction for generating new and reliable datasets that can be used to further our current understanding of climate change, tackling a key issue which is the limited ensemble data currently at our disposal.

With access to larger datasets, diffusion models could be explored to better capture complex non-linear relationships as well as further improving the generative capabilities, potentially leading to high quality forced response data being created. Another direction could involve combining data-driven methods with physics-informed constraints to enforce known climate dynamics and improve interpretability.

6 Acknowledgements

I would like to sincerely thank Victor Cohen for his invaluable supervision throughout this project, his guidance greatly assisted me throughout the whole project pipeline. My gratitude also goes to Mathieu Salzmann for providing the opportunity to undertake this work. Finally, I would like to acknowledge the Swiss Data Science Center (SDSC) for their support and resources throughout the project timeline.

Glossary

forced response The systematic change in climate due to external forcing factors, such as greenhouse gas emissions or aerosols. 3, 4

internal variability Natural, chaotic fluctuations in the climate system that occur independently of external forcing, this includes phenomena such as El Niño events. 3, 4

SST Sea Surface Temperature, the temperature of the ocean's surface layer. 3, 4

References

[1] Abhyuday Desai, Cynthia Freeman, Zuhui Wang, and Ian Beaver. Timevae: A variational auto-encoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095*, 2021.

[2] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

[3] Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

[4] Lizao Li, Robert Carver, Ignacio Lopez-Gomez, Fei Sha, and John Anderson. Generative emulation of weather forecast ensembles with diffusion models. *Science Advances*, 10(13):eadk4489, 2024.

[5] Sebastian Sippel, Nicolai Meinshausen, Enikő Székely, Erich Fischer, Angeline G. Pendergrass, Flavio Lehner, and Reto Knutti. Robust detection of forced warming in the presence of potentially large climate variability. *Science Advances*, 7(43):eabh4429, 2021.

[6] Y. Wang, K. J. Heywood, D. P. Stevens, and G. M. Damerell. Seasonal extrema of sea surface temperature in cmip6 models. *Ocean Science*, 18(3):839–855, 2022.