Finding the Perfect Fit: Applying Regression Models to ClimateBench v1.0

Anmol Chaure
Research Scholar
Bhilai Institute of Technology
Durg, India

Ashok Kumar Behera
Associate Professor
Bhilai Institute of Technology
Durg, India

Sudip Bhattacharya
Associate Professor
Bhilai Institute of Technology Durg, India

ABSTRACT

Climate projections using data driven machine learning models acting as emulators, is one of the prevailing areas of research to enable policy makers make informed decisions. Use of machine learning emulators as surrogates for computationally heavy GCM simulators reduces time and carbon footprints. In this direction, ClimateBench [1] is a recently curated benchmarking dataset for evaluating the performance of machine learning emulators designed for climate data. Recent studies have reported that despite being considered fundamental, regression models offer several advantages pertaining to climate emulations. In particular, by leveraging the kernel trick, regression models can capture complex relationships and improve their predictive capabilities. This study focuses on evaluating non-linear regression models using the aforementioned dataset. Specifically, we compare the emulation capabilities of three non-linear regression models. Among them, Gaussian Process Regressor demonstrates the best-in-class performance against standard evaluation metrics used for climate field emulation studies. However, Gaussian Process Regression suffers from being computational resource hungry in terms of space and time complexity. Alternatively, Support Vector and Kernel Ridge models also deliver competitive results and but there are certain trade-offs to be addressed. Additionally, we are actively investigating the performance of composite kernels and techniques such as variational inference to further enhance the performance of the regression models and effectively model complex non-linear patterns, including phenomena like precipitation.

General Terms

Earth System Science, Machine Learning

Keywords

Gaussian Process Regression, Surrogate Model, Climate Modelling, Kernel Ridge Regression, Support Vector Regression

1. INTRODUCTION

The depletion of the ozone layer sparked an intense scientific investigation into viewing the Earth as a comprehensive system. The first extensive evaluation of global climate change was documented in the Charney Report [2], based on the circulation model developed by Syukuro Manabe [3]. This report concluded that the doubling of CO_2 concentration would lead to a 3°C rise in the global temperature. By the 1980s, it was evident from the Keeling curve [4] that the level of carbon dioxide (CO_2) was increasing, pointing directly to the impact of anthropogenic activities.

The fundamental physical equations that governed the Manabe circulation system have remained constant [5]. However, with the advancement of digital computers, significant progress has been made in terms of refining and enhancing climate predictions. General Circulation Models (GCMs) and Earth system models (ESMs) have been the forefront models for climate projection for several decades. These models have allowed decision makers to explore the Shared Socioeconomic Pathways(developed by IPCC in accordance with the Paris Climate Agreement), for the mitigation(Controlling emissions) and adaptations(Being equipped for the unavoidable effects) of Climate Change [5,6].

General Circulation Models (GCMs) have been criticized for their computational expensiveness and iterative calibrations. Furthermore, the models are unable to consider the climate fluctuations from the past data which may hold important lessons, leading to an enhanced role for data-driven(Machine Learning) emulators.

Machine learning emulators are able to overcome these challenges, by their ability to train and run faster, better sensor calibration, and increased accuracy and pattern recognition due to the availability of petabytes of data generated using the Coupled Model Intercomparison Project (CMIP6). Regression Models, although simple, are still very popular machine learning models and have demonstrated good performance across many climate datasets. In this work compares three non-linear regression models: Support Vector Regression, Kernel Ridge Regression and Gaussian Process Regression on ClimateBench v1.0. These models were selected for this study because of their ability to train efficiently on non-linear and sparse samples. Furthermore, Gaussian Process Regression [7] exhibits the ability to quantify uncertainty, which is an essential feature for climate datasets.

2. CLIMATE CHANGE

In 1986, Svante Arrhenius [8] developed a climate model revealing the relationship between atmospheric CO_2 and temperature change. His model demonstrated that atmospheric CO_2 increases geometrically, which in turn leads to a nearly arithmetic progression of temperature change. Arrhenius projected that the doubling of CO_2 solely from fossil fuel combustion would occur within a span of 500 years and result in a global warming of approximately 5°C. There has been an exponential rise in global warming since. The evidence of climate change has been increasing, storms, forest fires, and floods are becoming stronger and more frequent. Climate Action has been designated as the 13^{th} Sustainable Development Goal by the United Nations, recognizing the immense challenges faced by the most vulnerable populations due to the impacts of climate change. Governing bodies like IEA (International Energy Agency) and IPCC (Intergovernmental Panel on Climate Change) [9] call for urgent action. The IEA has recently devised an ambitious net-

zero emission pathway [10] that guides the transition toward achieving a carbon-neutral future [11].

3. GENERAL CIRCULATION MODEL(GCMS)

General Circulation models, three-dimensional models that simulate and analyse the complex interactions and dynamics of the Earth's atmosphere and oceans, have been the forefront of climate modelling since their initial introduction by Syukuro Manabe and Kirk Bryan [3]. GCMs is utilizes the equations of fluid dynamics and thermal exchange to describe atmosphere-ocean interaction.

$$\frac{\partial x}{\partial t} = R(x) + U(x) + P(x) + F \quad (1)$$

where x is the state vector, R(x) + U(x) represents the Navier Stokes equation, P represents the thermodynamic processes and F represents external forcings over the system [12]. GCMs are employed to examine the climate system's response to variations in the parameter F. However, they have high computational cost, requires continuous tuning of hundreds of parameters, and fails to recognize the overall uncertainty of the climate pattern [12–14].

4. MACHINE LEARNING EMULATORS (SURROGATES)

With the rise in availability of climate observation data produced by numerous satellites and climate modelling projects, Data-driven emulators are becoming the forefront for projections. Machine learning emulators are used to model and approximate the behaviour of complex systems. They address the challenges prevalent in GCMs, by using automatic calibration techniques, being computationally efficient and using new observations to enhance the accuracy of the models. Regression is a widely recognized applied learning model. Any regressor aims to model the relationship between the regressor or independent variable(x) and the response or dependent variable(y). If there exists a relationship it is represented using the function f,

$$y = f(x, \beta) + \epsilon \quad (2)$$

Here, *represents* the error term or the residual (the differences between the predictions and real samples.), which accounts for the variability or randomness in the observed dependent variables that cannot be explained by the nonlinear function f and the parameters β [15].

In this paper, three nonlinear regression models, namely, Gaussian

Process Regression, Support Vector Regression and Kernel Ridge Regression, are used. Along with the ability to capture non linearity these models employ special mathematical function called" kernel". Kernels transform the original data into a new feature space where the complex and intricate patterns of data become more evident. This feature is especially useful in use case of climate data. Although the equation remains the same, but the kernel transformation helps us capture more nuanced and non-linear connections between the variables.

5. DATASET

Here, ClimateBench, a recently developed benchmark specifically curated for evaluation of machine learning model with the primary objective to predict air surface temperature, diurnal temperature, precipitation and the 90th percentile precipitation, has been used. This prediction is done from the four most common forcings in the Anthropocene era, Carbon dioxide, Methane, Sulphur dioxide and Black Carbon.

ClimateBench is devised using a selection of simulations like ScenarioMIP, CMIP6(Coupled Model Intercomparison Project, Phase-6) and DAMIP (Detection and Attribution Model Intercomparison Project). This dataset does not provide temporal predictions, but rather offers climate projections based on the scenarios designed by the IPCC.

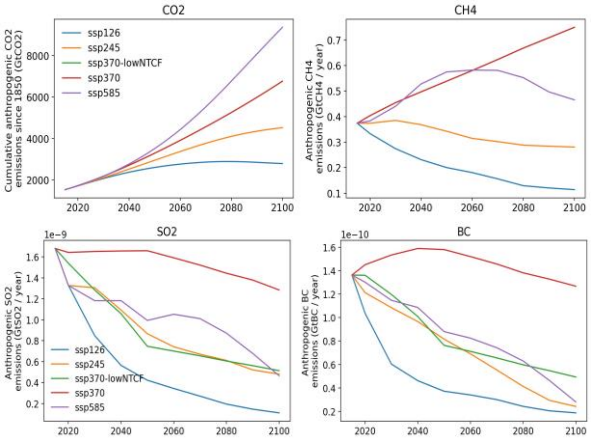

Fig. 1: IPCC Emission and Concentration Scenarios in ClimateBench for CO_2, CH_4, SO_2, and BC

6. GAUSSIAN PROCESS REGRESSION

Gaussian process regression (GPR) [16] is a supervised learning, probabilistic framework. Similar to regression modelling, it seeks to find the relationship between input and output variables. But, in addition to providing the predicted value GPR also provides a confidence interval.

6.1 Working of GPR

Gaussian processes provide an elegant solution to the problem of fitting data by assigning a probability to each function among the potentially infinite number of functions that can fit a given set of training points [16]. GPRs have the property of being closed under conditioning and marginalization, i.e., the output distributions will also be Gaussian distribution [17].

Firstly, it is necessary to concatenate the training data (X) and testing data (Y) in order to derive the joint probability distribution denoted as $P_{x,y}$. The objective is to acquire insight into the underlying distribution of Y in relation to X. Consequently, this joint distribution encompasses the complete set of potential function values targeted for prediction. In the context of regression, the approach adopted is akin to Bayesian inference, involving the adjustment of the initial prior knowledge to a posterior state upon the introduction of new observations (X). This leads to the requirement for the conditional probability $P_{Y|X}$, and due to the property of Gaussian Process Regressions (GPRs) to maintain closure under conditioning, the resultant distribution is also Gaussian in nature.

The subsequent step involves the computation of both the mean, denoted as μ, and the covariance matrix, represented by Σ. In the context of Gaussian distributions, the convention is that μ equals 0; however, if μ does not equal 0, it is considered as 0. The covariance matrix serves the purpose of conveying insights into both the data distribution and the inherent features of the predictive function. In order to determine this matrix, a specific mathematical function is required, known as the covariance function or kernel (k). This function operates with

two input values, x and x', generating an output that signifies their degree of similarity.

$$\Sigma = cov(X, X^0) = k(x, x^0) \quad (3)$$

This equation is calculated for each test point to create a covariance matrix. Since the kernel describes the similarity between the values of function, it controls the possible shape that a fitted function can adopt. Kernel essentially plays an important role in the accurate modelling of the relationship between X and Y. By means of a conditioning process, the distribution denoted as $P_{Y|X}$ is derived from the initial distribution $P_{X,Y}$. Notably, this resulting distribution aligns its characteristics with the distribution of the test points, where the count of test points is designated as N, equivalent to the size of the set Y. This procedure yields a multivariate Gaussian distribution encompassing the entirety of the test dataset.

Furthermore, through a technique called marginalization, the aim is to establish a probability distribution wherein the influence of multiple variables is contingent solely upon a single variable, represented as x. This is accomplished by integrating over all conceivable values associated with the remaining variables while keeping the variable x constant. The ensuing outcomes of this process involve the computation of both the mean and standard deviation for the distribution at each i-th test datapoint.

Utilizing this approach, the mean and standard deviation for each point within the set Y are attainable, and subsequently, the determination of a confidence interval becomes feasible by leveraging the standard deviation up to a specified threshold.

As one distances themselves from the training data points, a discernible trend emerges wherein the level of prediction uncertainty experiences a gradual augmentation. Conversely, regions proximate to the training data exhibit a comparatively diminished level of uncertainty in predictions [17].

6.2 GPR Kernels

The accuracy and efficiency of the Gaussian Process predominately depend upon the kernel functions. In addition to the wide variety of available kernels, they can also be combined to fit the data better. The most used kernel combination is addition and multiplication. There exist numerous types of kernels, the most commonly used include Linear, Exponential, Matern (1/2,3/2,5/2), Radial basis function or Squared Exponential [18]. In this project, the Matern 3/2 kernel was chosen for its ability to yield the most accurate results.

7. SUPPORT VECTOR REGRESSION(SVR)

Support Vector Regression [19] is considered a non-parametric technique due to its use of kernel functions. The model produced by support vector classification depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Analogously, the model produced by Support Vector Regression depends only on a subset of the training data, because the cost function ignores samples whose prediction is close to their target [20]. As a regression model, SVR tries to find a function f(x) = wx +b that best fits the model. In regression the goal is to quantify prediction, therefore, we aim to have the data points (observations) as close as possible to the hyperplane, in contrast to SVM for classification.

$$|y(x) - f(x)| \leq \epsilon + C \quad (4)$$

Here, f(x) is the predicted function, y(x) is the actual function, *is* the width of of marginal plane, C is the distance between data points outside marginal plane.

SVR is similar to simple regression techniques such as Ordinary Least Square, but with one key difference. It introduces an epsilon range on both sides of the hyperplane, which makes the regression function less sensitive to errors. Ultimately, SVR in regression has a boundary, similar to SVM in classification, but the purpose of the boundary in regression is to reduce the sensitivity of the regression function to errors. In contrast, the boundary in classification is designed to create a clear distinction between classes in the future (which is why it's referred to as the safety margin).

The efficiency of SVR depends heavily on the selection of appropriate kernel, because kernel selection helps in mapping the data into higher dimensional without actually calculating the coordinates for that space. Therefore, reducing the computational need and increasing the accuracy of mapping non linear data points.

8. KERNEL RIDGE REGRESSION(KRR)

Kernel ridge regression (KRR) combines ridge regression (linear least squares with l2-norm regularization) with the kernel trick. KRRs are regression models that can capture both linear and nonlinear relationships between predictor variables and outcomes. These models are considered non-parametric, meaning they do not rely on strict assumptions about the underlying data distribution. However, the performance of kernel ridge regression heavily depends on the selection of hyperparameters. The objective of kernel ridge regression is to minimize the sum of squared errors between the predicted values and the actual outcomes, while also minimizing the complexity of the model. The model can be represented by the equation:

$$f(x) = \Sigma(\alpha_i * k(x_i, x)) + b \quad (5)$$

where f(x) is the predicted outcome for a new input x, α_i are the learned coefficients, x_i represents the training inputs, $K(x_i, x)$ is the kernel function that measures the similarity between x_i and x in the feature space, and b is a bias term. To address this sensitivity, Kernel Ridge Regression offers a solution by employing k-fold cross-validation on predefined grids of hyperparameter values. This technique allows for a systematic evaluation of various hyperparameter combinations. By performing cross-validation, the optimal hyperparameter values can be determined, leading to more reliable and accurate predictions in the regression model. It thus learns a linear function in the space induced by the respective kernel and the data. For non-linear kernels, this corresponds to a non-linear function in the original space.

9. RELATED WORK

Recently, a generalized machine learning model ClimaX [21] was introduced for weather and climate. It presents flexible and generalized deep learning model for weather and climate science can be trained using heterogeneous datasets spanning different variables, spatio-temporal coverage, and physical groundings.

Another paper by L.A Mansfield et al. [22], involves training of ridge and Gaussian regression using short-term simulations and historical climate data. By integrating relevant environmental variables and extracting meaningful features from the short-term simulations, the models are designed to capture the underlying patterns and relationships that contribute to long-term climate change. The authors carefully curate and

preprocess the input features to ensure the models' robustness and generalizability.

ESEm v1.0.0 [23], an openly accessible and scalable Earth System Emulator library developed by Duncan Watson-Parris, incorporates GP fitting using the open-source GPFlow, a machine learning library that leverages Graphical Processing Units (GPUs) to expedite GP training, as mentioned earlier. The paper demonstrates the utilization of ESEm to generate emulated responses by employing GP emulation of aerosol optical depth (AAOD) with a 'Bias + linear' kernel and showcases its remarkable capability in accurately reproducing the spatial characteristics of AAOD while minimizing errors.

Additionally, the use of GPR in climate emulation is discussed [24], the paper also explores the utilization of Gaussian Process Regression (GPR) in climate emulation, specifically focusing on its application in modelling and predicting precipitation patterns in the Hindu Kush Karakoram Himalayan region.

GPR has been utilized as a baseline emulator in ClimateBench [1]. Gaussian process regression is used to develop an emulator for a computationally expensive climate model. The emulator is then used to generate a large ensemble of simulations, which enables the exploration of the model's parameter space in a more efficient and systematic manner.

10. RESULTS

Among the three regression models, Gaussian Process Regression

(GPR) demonstrated the best performance, achieving the lowest RMSE compared to SVR and KRR. This suggests that GPR was the most accurate and effective model in capturing the complex relationships between the predictor variables and air surface temperature. The lower RMSE indicates that the predictions made by GPR were closer to the actual observed values, signifying its superior predictive capabilities in this specific regression task. These findings highlight the suitability of GPR for the given regression task, as it outperformed the other two models in terms of prediction accuracy. The probabilistic nature of GPR, which models the uncertainty in the predictions, may have contributed to its superior performance when dealing with the uncertainties and nonlinear relationships present in the dataset.

Table 1. : RMSE values for different Regressors against different lead times

RMSE (Root Mean Squared Error)						
Year	**2050**	**2100**	**2045 - 2055**	**2090 - 2100**	**2050 - 2100**	**20Y**
GPR(ClimateBench)	0.33	0.393	0.349	0.421	0.373	**0.246**
GPR	0.308	0.350	0.353	0.372	0.38	**0.184**
SVR	0.299	0.487	0.366	0.529	0.462	**0.402**
KRR	0.337	0.38	0.39	0.48	0.379	**0.227**

The study's results support the use of Gaussian Process Regression as the preferred model for predicting air surface temperature based on the chosen predictors, methane (CH4), carbon dioxide (CO2), sulphur dioxide (SO2), and black carbon (BC). It underscores the importance of choosing an appropriate regression model for specific datasets and demonstrates the value of GPR in climate modelling and environmental research where accurate predictions of air surface temperature are crucial for understanding climate patterns and their potential impacts.

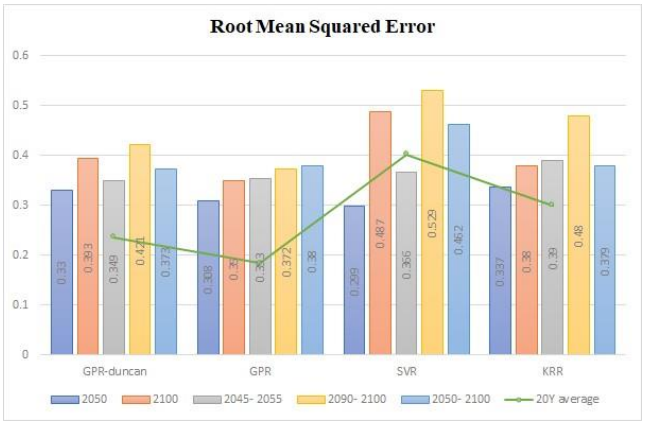

Fig. 2: Comparing Root Mean Square Error over Different Lead Time for all models

*** For the visualization of results generated using Cartopy and panoply, please refer to the Appendix.**

11. CONCLUSION

Machine learning emulators have demonstrated the capability to achieve comparable accuracy to large-scale Earth System models while requiring significantly less computational resources. However, each model has its own strengths and weaknesses, although they generally excel in evaluation metrics and effectively reproduce the temperature response of NorESEM2 in a realistic future scenario. Among these models, GPR stands out as the top performer in evaluation metrics and benefits from its ability to quantify uncertainty. Nevertheless, training the model to understand non-linear responses like precipitation has posed challenges.

Despite the notable performance of regression models, they are often considered fundamental. Researchers are more leveraging more advanced models such as neural networks, transformers, and generative models for climate projections. These advanced models come with a cost of increased carbon footprints. The objective of this work was to disseminate the effectiveness of simple regression models for climate projections with the purview of carbon footprint and model complexity. Results clearly denote the effectiveness of these models and the identifies Gaussian Process Regression as the best-in-class model for the task over the considered dataset.

12. REFERENCES

[1] Journal Article Author, D Watson-Parris, Y Rao, D Olivie,´ Ø Seland, P Nowack, G Camps-Valls, P Stier, S Bouabid, M Dewey, E Fons, J Gonzalez, P Harder, K Jeggle, J Lenhardt, P Manshausen, M Novitasari, L Ricard, and C Roesch.

[2] ETH Library ClimateBench v1.0: A Benchmark for DataDriven Climate Projections Rights / license: Creative Commons Attribution 4.0 International Funding acknowledgement: 860100-innovative MachIne learning to constrain Aerosol-cloud climate Impacts (EC) ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections Citation. *Journal of Advances in Modeling Earth Systems*, 14(10), 2022.

[3] J. G. Charney et al. Carbon Dioxide and Climate: A Scientific Assessment . *National Academies, 1979*, 96(8), 1979.

[4] Syukuro Manabe and Richard T. Wetherald. The Effects of Doubling the CO ¡sub¿2¡/sub¿ Concentration on the

climate of a General Circulation Model. *Journal of the Atmospheric Sciences*, 32(1):3–15, 1 1975.

[5] Kirk W. Thoning, Pieter P. Tans, and Walter D. Komhyr. Atmospheric carbon dioxide at Mauna Loa Observatory: 2. Analysis of the NOAA GMCC data, 1974-1985. *Journal of Geophysical Research: Atmospheres*, 94(D6):8549–8565, 6 1989.

[6] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P.

[7] Kording, Carla Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling Climate Change with Machine Learning. 6 2019.

[8] Tim Lenton. *Earth System Science: A Very Short Introduction*. Oxford University Press, 2 2016.

[9] Drew Bagnell, Venkatraman Narayanan, M Koval, and P Parashar. Gaussian Processes. Technical report.

[10] Svante Arrhenius. On the Influence of Carbonic Acid in the Air upon the Temperature of the Ground. Technical report, 1896.

[11] Intergovernmental Panel on Climate Change (IPCC). *Climate Change 2022 – Impacts, Adaptation and Vulnerability*. Cambridge University Press, 6 2023.

[12] International Energy Agency. Net Zero by 2050 - A Roadmap for the Global Energy Sector. Technical report, 2050.

[13] Mark Maslin. *Climate : a very short introduction*.

[14] V. Balaji, Fleur Couvreux, Julie Deshayes, Jacques Gautrais, Fred´ eric Hourdin, and Catherine Rio. Are general circulation´ models obsolete? *Proceedings of the National Academy of Sciences of the United States of America*, 119(47), 11 2022.

[15] J. E. Kay, C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, J. M. Arblaster, S. C. Bates, G. Danabasoglu, J. Edwards, M. Holland, P. Kushner, J.-F. Lamarque, D. Lawrence,

[16] K. Lindsay, A. Middleton, E. Munoz, R. Neale, K. Oleson, L. Polvani, and M. Vertenstein. The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability. *Bulletin of the American Meteorological Society*, 96(8):1333–1349, 8 2015.

[17] Jessie Carman, Thomas Clune, Francis Giraldo, Mark Govett, Brian Gross, Anke Kamrath, Tsengdar Lee, David Mccarren, John Michalakes, Scott Sandgathe, and Tim Whitcomb. Position paper on high performance computing needs in Earth system prediction. National Earth System Prediction Capability. 2017.

[18] G. A. F. Seber and C. J. Wild. *Nonlinear Regression*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2 1989.

[19] Carl Edward. Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[20] Jochen Gortler, Rebecca Kehlbeck, and Oliver Deussen. A Vi-¨ sual Exploration of Gaussian Processes. *Distill*, 4(4), 4 2019.

[21] David Duvenaud. The Kernel Cookbook: Advice on Covariance functions. Technical report.

[22] Harris Drucker, Chris J C Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support Vector Regression Machines. Technical report.

[23] Fan Zhang and Lauren J. O'Donnell. Support vector regression. In *Machine Learning*, pages 123–140. Elsevier, 2020.

[24] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. 1 2023.

[25] L. A. Mansfield, P. J. Nowack, M. Kasoar, R. G. Everitt, W. J. Collins, and A. Voulgarakis. Predicting global patterns of long-term climate change from short-term simulations using machine learning. *npj Climate and Atmospheric Science*, 3(1):44, 11 2020.

[26] Duncan Watson-Parris, Andrew Williams, Lucia Deaconu, and Philip Stier. Model calibration using ESEm v1.0.0-an open, scalable Earth System Emulator. Technical report.

[27] Vidhi Lalchand, Kenza Tazi, Talay M. Cheema, Richard E. Turner, and Scott Hosking. Kernel Learning for Explainable Climate Science. 9 2022.

13. APPENDIX

Results with Gaussian Process Regression

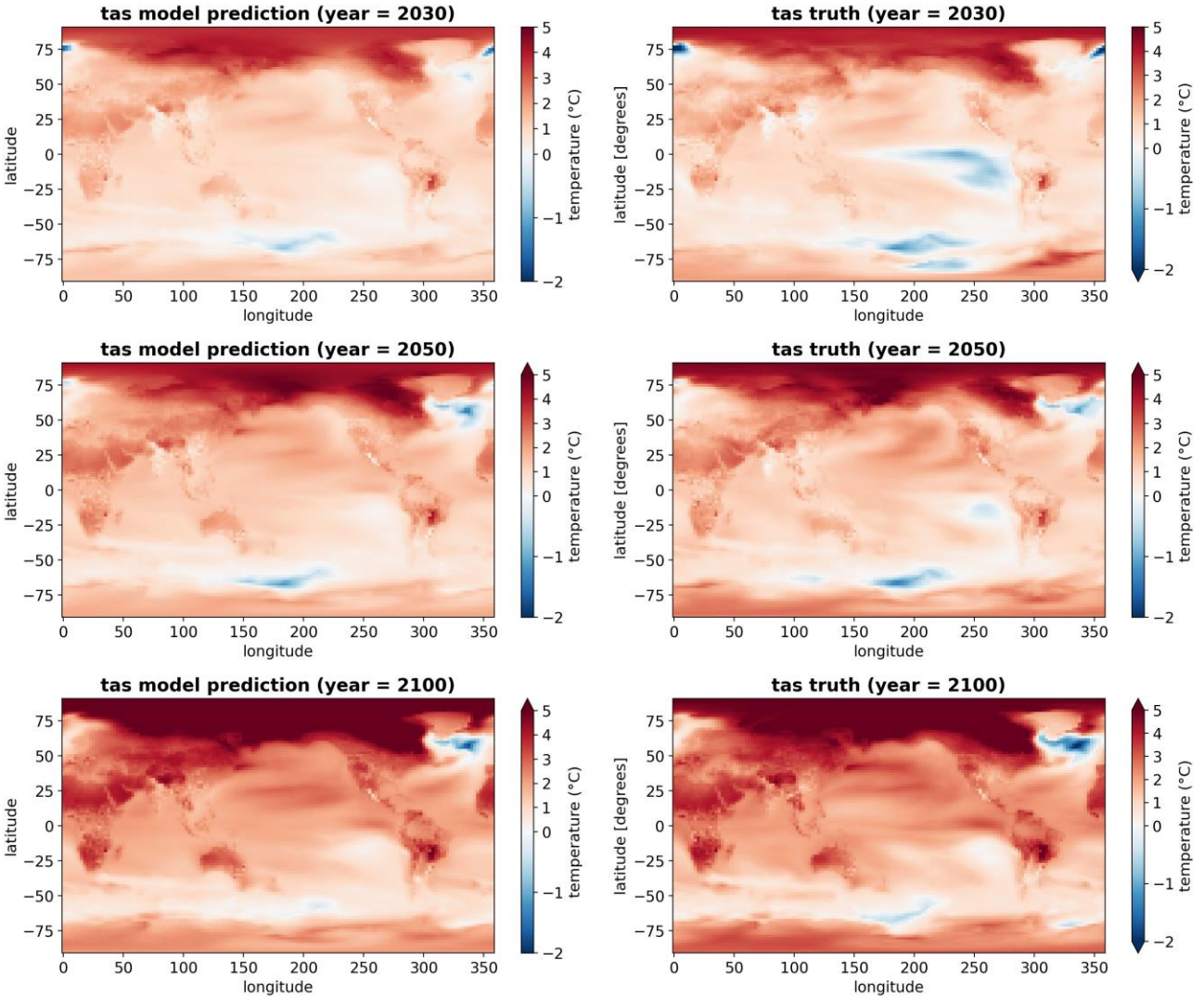

Fig. 3: Results of Year Wise Prediction of Air Surface temperature using GPR for different lead times

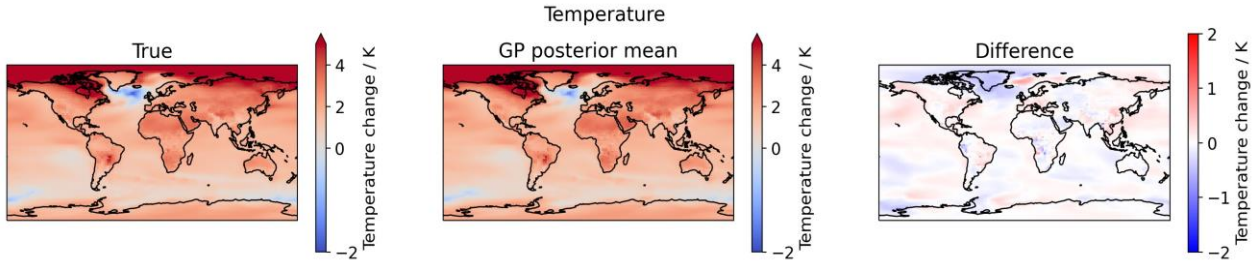

Fig. 4: Difference of Air Surface temperature using GPR

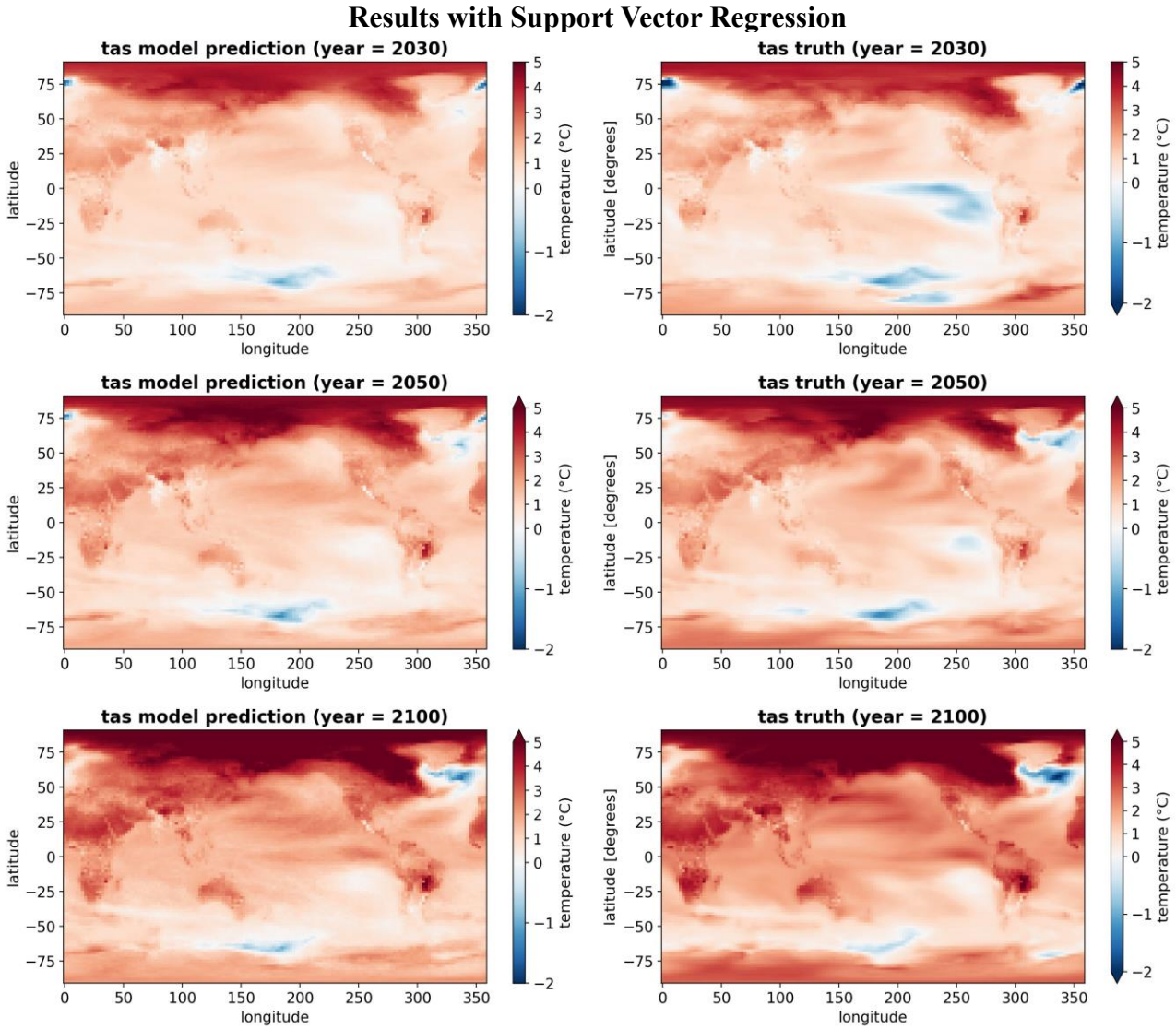

Fig. 5: Results of Year Wise Prediction of Air Surface temperature using SVR for different lead times

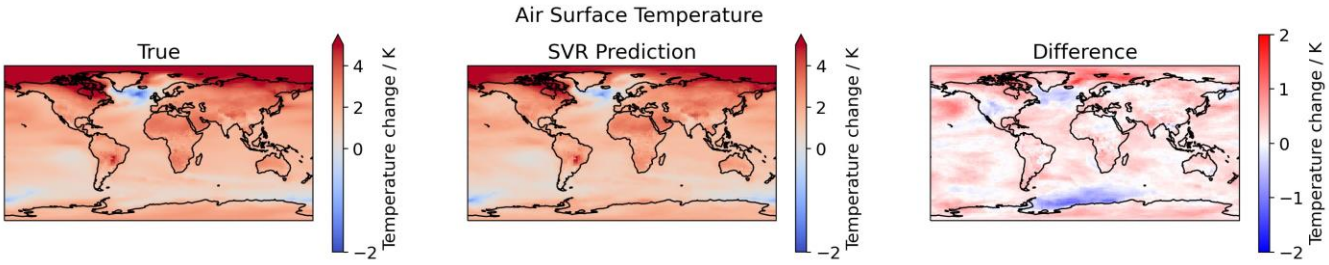

Fig. 6: Difference of Air Surface temperature using SVR

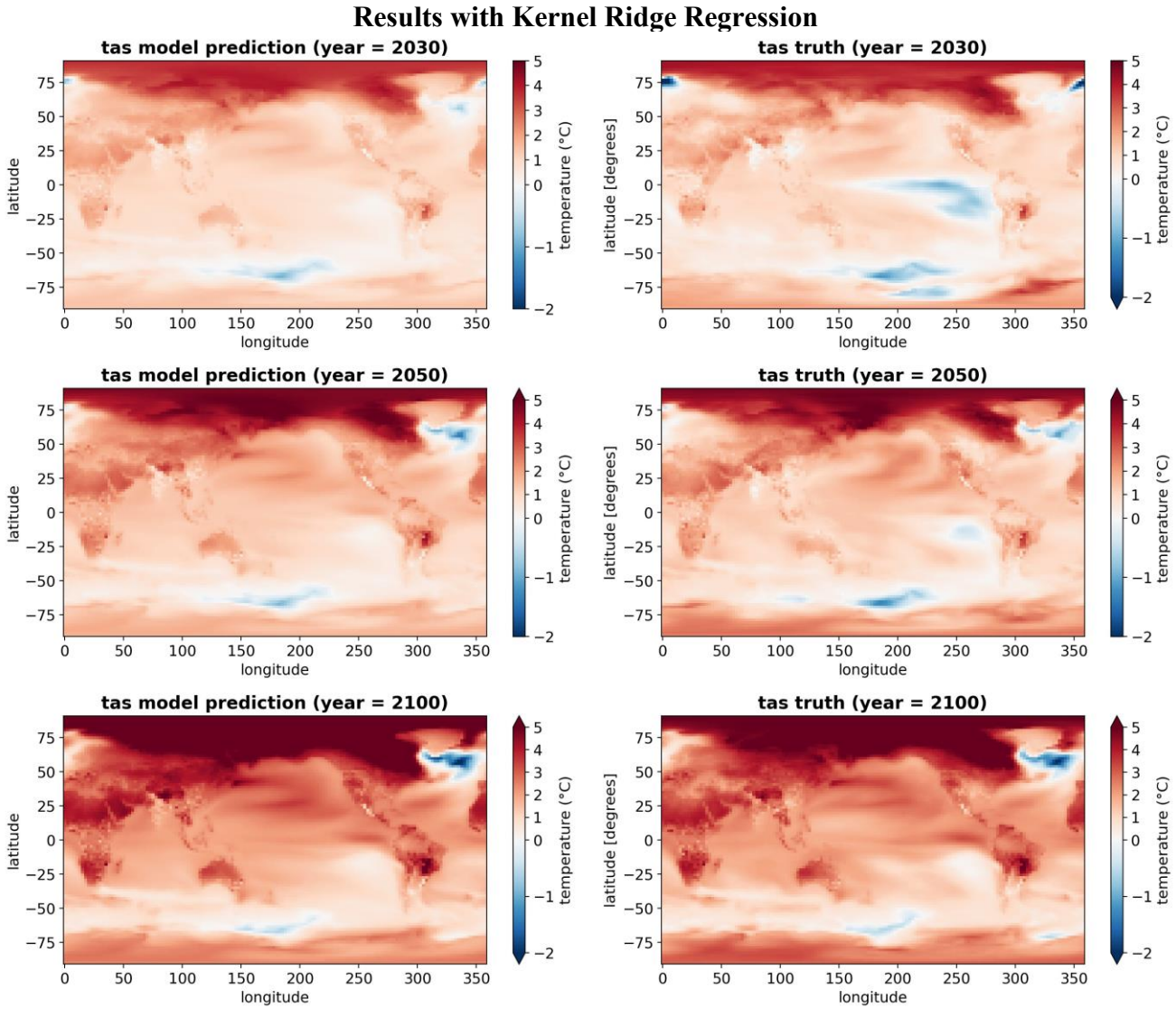

Fig. 7: Results of Year Wise Prediction of Air Surface temperature using SVR for different lead times

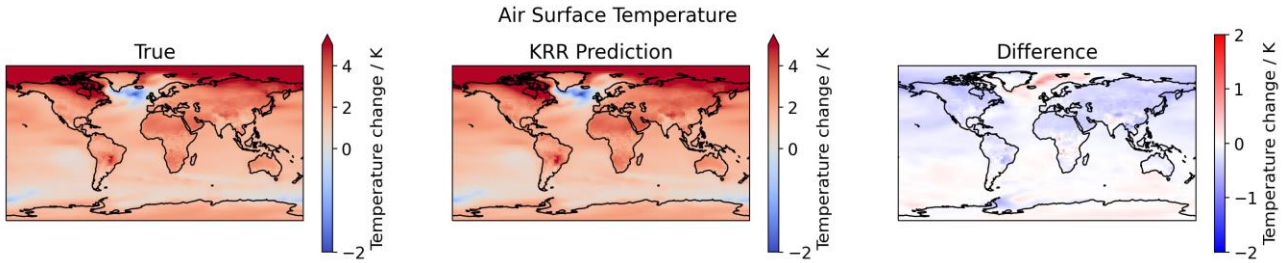

Fig. 8: Difference of Air Surface temperature using KRR

Visualization using Panoply

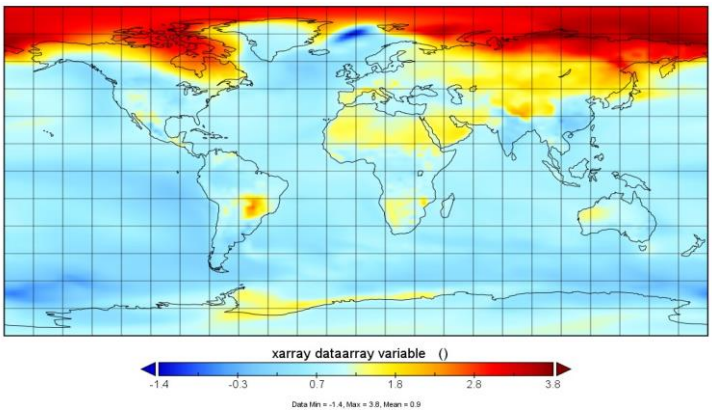

Fig. 9: Projection of GPR

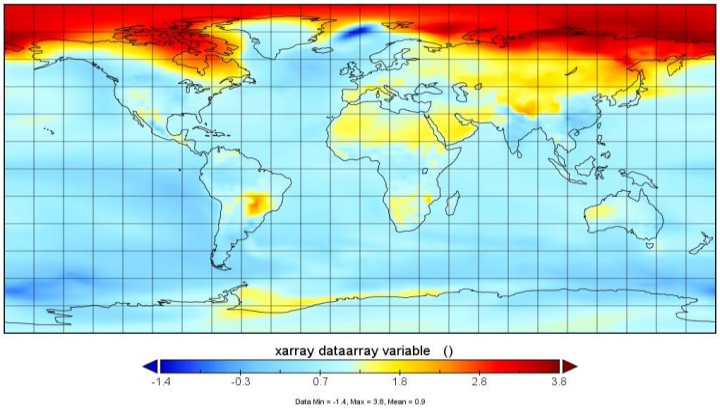

Fig. 10: Projection of SVR

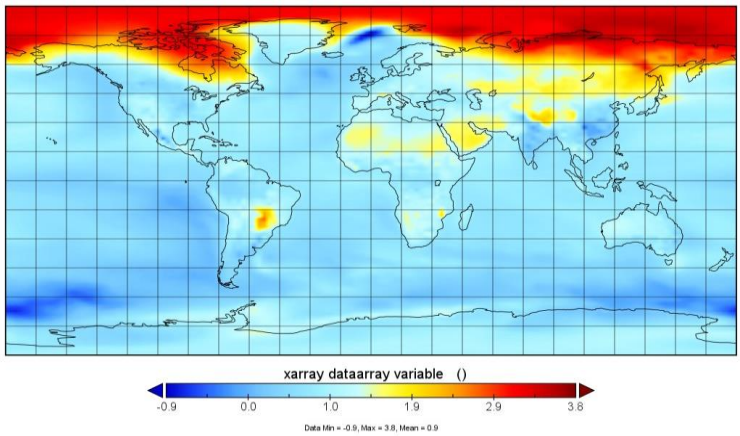

Fig. 11: Projection of KRR