Utilizing Emulators to Explore the Climate Model Parameter Space

Zoe Ludena
zludena@ucsd.edu

Eric Pham
erpham@ucsd.edu

Ylesia Wu
xw001@ucsd.edu

Duncan Watson-Parris
dwatsonparris@ucsd.edu

Abstract

Earth System Modeling is one of the primary tools we use to address uncertainties about the future of our climate and planet. But with it comes a daunting task—addressing the unpredictability in how we, as a planet or society, respond to the challenge of climate change. There are only a finite number of supercomputers on Earth capable of running simulations to address these questions, but an infinite combination of choices within the large n-dimensional parameter space of climate models. Additionally, running these simulations is expensive and time-consuming, which is why we turn to climate model emulators. These machine-learning models take typical Earth system model input data and learn the outputs based on data generated from previous climate model runs. These emulators are capable of predicting everything from spatial distributions of precipitation patterns to global average temperature over time, making them powerful tools that can help us answer questions about how particular scenarios of climate change may play out.

Code: https://github.com/zoeludena/ResearchOnClimate

1 Introduction . . . 2
2 Methods . . . 4
3 Results . . . 7
4 Discussion . . . 9
5 Conclusion . . . 11
References . . . 11
Appendices . . . A1

1 Introduction

1.1 First Look

There are a multitude of future climate scenarios that are possible over the next few decades, all of which come down to how we address our emissions. While it's important to understand what a realization of each of these pathways looks like in terms of climate patterns, it is infeasible to explore each of these scenarios because running climate models requires a lot of time and resources. As such, we turn to climate model emulators, which can help us explore the outcome of different emission scenarios at a much quicker rate.

1.2 Prior Work

"ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections" (Watson-Parris et al. 2022) is the paper we are trying to recreate in Fall 2024. This paper is the first benchmarking framework that uses a set of baseline machine learning models on an Earth System Model to emulate the response of different climate variables. This allows people to predict annual mean global distributions of temperature, diurnal temperature ranges, and precipitation given a wide range of emissions and concentrations of carbon dioxide, methane, sulfur dioxide, and black carbon. This allows the baseline models to explore the unexplored. The paper found the most accurate three baseline models were neural networks, gaussian processes, and random forests.

The latest UN Intergovernmental Panel on Climate Change (IPCC) Synthesis Report (Intergovernmental Panel on Climate Change IPCC) is a summary for policymakers that explains different Shared Socio-economic Pathways (SSPs). The summary also explains observed changes and impacts, current status and trends, future climate change, risks, and long-term responses, and responses in the near term. Each one explains who is being affected the most, the financial struggles, and the ecosystem strains. There are many insightful figures like Figure SPM.2 Projected changes of annual maximum daily maximum temperature, annual mean total column soil moisture, and annual maximum 1-day precipitation at global warming levels of 1.5 degrees Celsius, 2 degrees Celsius, 3 degrees Celsius, and 4 degrees Celsius relative to 1850 to 1900. This is directly related to the Climate Bench, as the paper explores different SSP scenarios (different warming levels).

"Predicting global patterns of long-term climate change from short-term simulations using machine learning" (Mansfield et al. 2020) is about using model simulations to learn relationships between short-term and long-term temperature responses while reducing the cost of scenario computations. The researchers analyzed scenario predictions across the world using ridge regression, gaussian process regression, and pattern scaling. In Fig 3 it appears that Gaussian Process regression had the smallest boxplot of error distribution around scenario predictions. This relates to the ClimateBench paper because they found Gaussian Processes were successful predictors (second to Neural Networks).

"Climate model genealogy: Generation CMIP5 and how we got there" (Knutti, Masson and Gettelman 2013) is about how models are strongly tied to each other because they share a common code. CMIP5 supports the idea that models improve over time because strengths and weaknesses of some models are passed onto newer model versions because of the exchange of code and ideas. This is important to the ClimateBench argument that the models are not independent, so a more complex one is not the answer. It will be better instead to use an emulator for NorESSM2 because they are simpler and will encompass what the mixed models are attempting to achieve.

1.3 Description of Data

There are two classes of data files that we need for the emulators. After preprocessing, both will be provided to the various models as input for training, validation, and ultimately testing. Both of these data files are binary .nc (NetCDF) files, which are essentially multidimensional data structures with "indexes". In this case, both of the data files are indexed along "lat" (latitude), "lon" (longitude), and "time".

- Climate Model Input Data
  - "CO2" (Carbon Dioxide) (NOAA Climate.gov 2024). One of Earth's most important greenhouse gases because it absorbs and radiates heat. It is a stable molecule and can remain in the atmosphere for several thousand years.
  - "CH4" (Methane) (NASA Climate Change 2024). Second largest contributor to climate warming after CO2. Methane is a much more potent greenhouse gas, but has a much shorter half-life of only 8-9 years.
  - "SO2" (Sulfur Dioxide) (NASA Earth Observatory 2017). Sulfur dioxide can react with the atmosphere to form aerosol particles which helps make clouds. It negatively affects air quality (a critical air pollutant) because it mainly comes from burning coal (coal-fired power plants). It can also react with water vapor to form acid rain.
  - "BC" (Black Carbon) (Office of Environmental Health Hazard Assessment OEHHA). Absorbs light and contributes to climate change by releasing heat energy into the atmosphere. It is considered a short-lived pollutant. They can cause greater warming effects than CO2 even with its short lifespan. Are causing snow, glaciers, and ice to darken and melt.
- Climate Model Output Data
  - "dtr" (Diurnal Temperature Range). Diurnal temperature range is calculated by diurnal temperature range $=$ tasmax $-$ tasmin (the difference between daily maximum and minimum temperature). Measured in Kelvin.
  - "tas" (Surface Air Temperature). Average monthly surface air temperature two meters above the ground. Measured in Kelvin.
  - "pr" (Precipitation). Average monthly precipitation. Measured in millimeters per day.
  - "pr90" (90th percentile of precipitation data). The threshold amount of monthly precipitation that is exceeded only 10% of the time over a given period. Measured in millimeters per day.

2 Methods

2.1 Normalizing the Data

“Normalizing data” means to adjust values from different scales into the same one. This makes it easier to compare and analyze the data. All of the models started by normalizing the emission data. More specifically we normalized the carbon dioxide and methane variables.

2.2 Data Used in Models

In this study, we employed historical data alongside various Shared Socioeconomic Pathway (SSP) scenarios. Specifically, we utilized SSP 126, SSP 370, and SSP 585 to train our models and evaluated their performance using SSP 245.

The chosen training scenarios encompass a broad spectrum of potential futures. SSP 126 represents a low-emissions pathway characterized by rapid decarbonization, reduced fossil fuel dependency, widespread renewable energy adoption, and sustainable land-use practices. SSP 370 reflects a medium-high emissions trajectory marked by regional self-reliance, limited climate policy, and moderate adaptation efforts. SSP 585 describes a high-emissions scenario driven by delayed or weak climate mitigation, high greenhouse gas emissions, rapid urbanization, and significant global population growth. This diverse set of training scenarios allows the models to learn from a wide range of socioeconomic and emissions conditions, spanning sustainable to high-emissions futures.

For testing, we selected SSP 245, a moderate-emissions scenario. This pathway combines elements of sustainability and fossil fuel usage with incremental climate policies and uneven global cooperation. As an intermediate scenario, SSP 245 was not included in the training set, making it an ideal candidate for evaluation. It provides a realistic and balanced context to assess model performance under conditions that are neither extreme nor strongly polarized.

2.3 Pattern Scaling

Pattern Scaling Overview The pattern scaling model is the simplest emulator at our disposal. The model consists of many linear regression models trained on global mean temperature in different emission scenarios. These models regress desired variables (precipitation, diurnal temperature range, etc.) on global mean temperature which is the “scaling” element of the model. Once trained, the model takes a vector of global mean temperatures from a particular emission scenario, and predicts the desired variables using the inputs. This model is powerful yet simple because it can predict local values of particular variables using only globally averaged inputs.

Though simple, the pattern scaling model is actually one of the most powerful in terms of predictive power. Additionally, it has no hyperparameters, which means no tuning is necessary for this model. It may be beneficial to try using other types of regression models (LASSO, Ridge, Elastic-Net), but the test errors are not vastly changed when using them. The only input needed are the global mean temperatures for the training and test scenarios, as well as the desired variables from the training scenarios.

However, because it is based on the result of linear regression models, this pattern scaling emulator can fail to capture non-linear processes, which arise often when dealing with climate change. Several aspects of the climate are governed by feedback loops, and these non-linear processes will not be captured well by the pattern-scaling model.

2.4 Gaussian Process

Gaussian Process Overview. A Gaussian Process (GP) model, used in the Climate Bench paper, is a probabilistic framework ideal for regression and classification tasks. GPs model functions by defining a prior characterized by a mean function, $m(x)$, representing the expected value at x, and a covariance function, $k(x, x)$, which measures similarity between inputs x and x. Using Bayesian inference, GPs update this prior with training data to produce a posterior distribution. For new inputs, predictions are made as a distribution with a mean (most likely value) and variance (uncertainty estimate).

GP models are well-suited for climate prediction. Climate systems are governed by complex, smooth, and often nonlinear relationships, which GPs can model through appropriately chosen kernels. Moreover, their ability to provide uncertainty estimates is invaluable when working with limited or noisy climate data, as these estimates can highlight regions where the model is less confident in its predictions. Finally, the interpretability of GP models aligns well with scientific practices, allowing researchers to explore the relationships captured by the covariance function and gain insights into the modeled climate dynamics.

Using the Model. The GP model was created using the `esem` library (Duncan Watson-Parris 2021a). The `esem` library has `gp_model` (Duncan Watson-Parris 2021b), which allows users to provide training data, specific kernel(s), and a way to combine multiple kernels together.

In the original baseline model there was a different GP model for each variable, so surface air temperature, precipitation, diurnal temperature range, and 90th quantile of precipitation. For training data X we gave each model `leading_historical_inputs`, which was a DataFrame containing normalized carbon dioxide, normalized methane, black carbon measurements, and sulfur dioxide measurements. The training data Y contained the historical measurements of the different variables and data from SSP 585. The code was simple, `gp_model(leading_historical_inputs, Y[“variable”])`, where “`variable`” is replaced with the acronym corresponding to the desired variable. This means the GPs were using no specific kernel!

Surprisingly, the root mean squared error (RMSE) when comparing the predictions from the GP models and the true values, from a super computer’s prediction, were low! When we ran the same experiment using `gp_model`’s default values we found very similar results. We found the GP model tended to underpredict the southern hemisphere. It also overcompensated at times making places that would have cooler surface air temperatures warmer.

Hyperparameters. To evaluate the performance of different kernel combinations in Gaussian models, we compared the RMSE values of individual kernels (`["Linear"]`, `["RBF"]`, and `["Polynomial"]`), kernel additions, and kernel multiplications.

The original RMSE values for the Gaussian model were:

- `tas`: 0.8067
- `dtr`: 0.1806
- `pr`: 0.6111
- `pr90`: 1.7520

Individual Kernels The kernel `["RBF"]` achieved the lowest RMSE for most metrics, with notable reductions in `dtr` (0.1779) and `pr` (0.6043) compared to the original values. The `["Polynomial"]` kernel performed the worst overall, particularly for `tas` and `pr90`.

Kernel Additions Looking at the three kernels and combining different permutations of them (`["Linear", "Polynomial"]`, `["Linear", "RBF"]`, and `["RBF", "Polynomial"]`) through addition we were able to see how making the model more complex would affect the RMSE. The addition of `["Linear", "Polynomial"]` provided the best overall performance closely matching the original model for most metrics:

- `tas`: 0.8076
- `dtr`: 0.1821
- `pr`: 0.6111
- `pr90`: 1.7520

Kernel Multiplications Looking at the same three models, but this time combined through multiplication we were able to find `["Linear", "Polynomial"]` performed the best again, reducing RMSE for `tas` (0.8303) and `pr` (0.6314). However, it was slightly less effective than additive combinations for `pr90`.

Best Performing Model The combination of `["Linear", "Polynomial"]` with additive operations showed the most consistent performance across all metrics, achieving values very close to the original Gaussian model while minimizing deviations.

2.5 Random Forest

Random Forest Overview. Random Forest is an ensemble method that aggregates the predictions of multiple decision trees to enhance predictive performance. Decision trees, as the base models, are particularly effective at capturing non-linear relationships and interactions between variables but are prone to overfitting. Random Forest addresses this limitation by averaging the predictions of all individual trees, which reduces variance and increases robustness. This makes it well-suited for climate model emulation, where separate models are often developed for multiple target variables.

One key advantage of Random Forest in climate model emulation is its interpretability, which aids in informing decision-making. While a common drawback of Random Forest is its inability to extrapolate beyond the range of training data, this is not a significant concern in this context. Relevant predictions in climate modeling typically lie within the range defined by historical climate data and plausible scenarios, such as the low-emissions SSP126 and high-emissions SSP585 pathways. This makes Random Forest an effective and practical choice for emulating climate models.

Features and Hyperparameters. The features for all four models consist of the first five principal components of SO2 and BC, CO2, and CH4. For hyperparameters, `max_features` was changed from `‘auto’` to `‘sqrt’` to accommodate a different version of `rf_model` while achieving a similar result. The rest of the hyperparameters are kept unchanged from the original paper, tuned using random search of the training data without replacement (Table 1).

Table 1: Hyperparameters for Random Forest Models

Model	**n_estimators**	**min_samples_split**	**min_samples_leaf**	**max_depth**
`rf_tas`	250	5	7	5
`rf_pr`	150	15	8	40
`rf_pr90`	250	15	12	25
`rf_dtr`	300	10	12	20

3 Results

To evaluate the emulator’s performance, we calculated the root mean square error between the emulator projection and the climate model projection. Subsequently, we plot the true and the emulated projections side-by-side for each target variable. The results are overall promising with the RMSE scores being relatively low (Table 2) and the patterns in the figures matching (Figure 3).

Table 2: Root Mean Square Error (RMSE) for Emulated Results Compared to Climate Model Projections. Let **RF** be Random Forest, **GP** be Gaussian Process, and **PS** be Pattern Scaling.

Variable	**PS**	**GP**	**RF**
tas (Near-Surface Air Temperature)	0.3648	0.8076	0.6823
dtr (Diurnal Temperature Range)	0.1503	0.1821	0.1654
pr (Precipitation)	0.5275	0.6905	0.6111
pr90 (90th Percentile of Precipitation)	1.5322	1.7520	1.5880

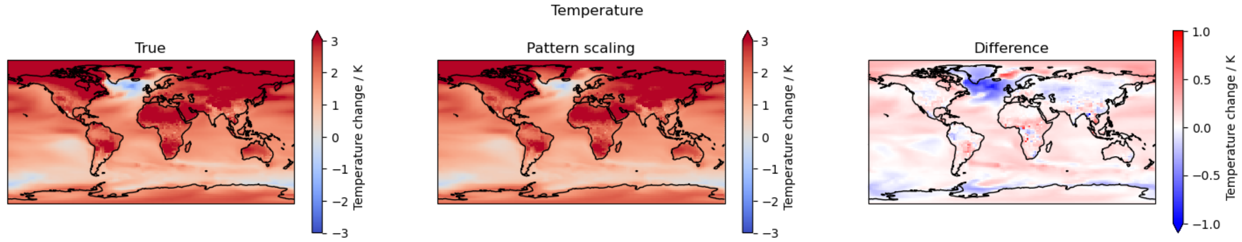
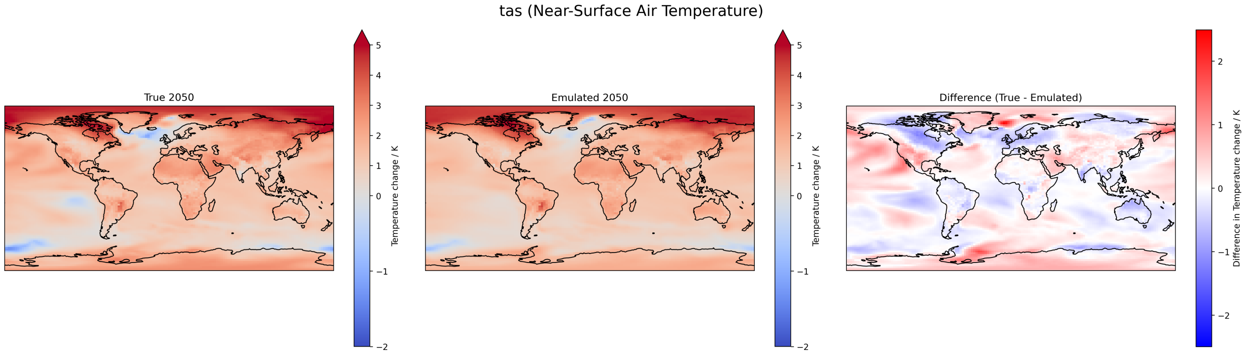


Figure 1: Climate model projection vs. Pattern Scaling model projection for Near-Surface Air Temperature in 2050.



tas (Near-Surface Air Temperature)
True 2050
Emulated 2050
Difference (True - Emulated)
Temperature change / K
Difference in Temperature change / K
5
4
3
2
1
0
−1
−2

Figure 2: Climate model projection vs. Gaussian Process model projection for Near-Surface Air Temperature in 2050.

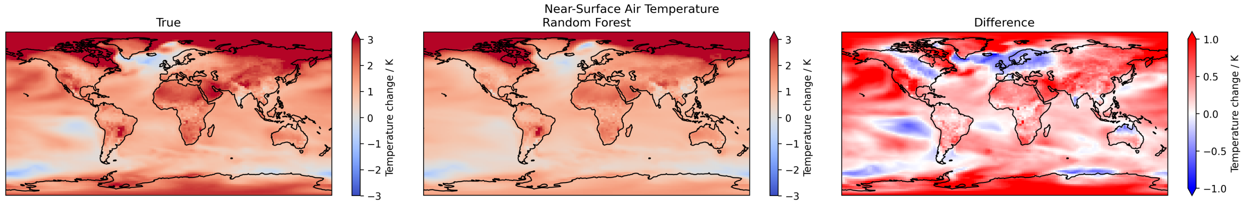


Figure 3: Climate model projection vs. Random Forest projection for Near-Surface Air Temperature in 2050.

4 Discussion

4.1 Pattern Scaling

Base-lining. The pattern scaling model's performance is among the best relative to the other emulators, even though it is only based on a series of linear regressions. This is an interesting observation because there are a lot of non-linear processes that go into future climate predictions, but this may be more indicative that the climate models used for training do not predict or capture many non-linear relationships.

Advantages and Drawbacks. Its simple structure and moderately good performance mean that the pattern-scaling model is a reliable baseline to which we can compare the other emulators. We can use this information to make decisions about whether or not more tuning is required of the hyper-parameters for a particular emulator. Still, this model is limited by its inability to capture nonlinear relationships. If nonlinear relationships are present in different climate model runs, then we can expect the error for pattern scaling models to be a bit worse than what we are observing right now.

Different Regressions. As a good extension of this linear model, we can explore the value of using different regression models to train the pattern scaling model in more depth. In particular, tuning the parameters for a different linear model may be interesting to see if we can get even more significant reductions in the error.

4.2 Gaussian Process

Impact and Applicability. Our Gaussian Process (GP) model is pretty close to the more intensive truth. This precision enables researchers and policymakers to simulate diverse ”what-if” scenarios, reflecting a range of possible futures. For instance, GP models can be instrumental in studying Shared Socioeconomic Pathways (SSPs), facilitating robust predictions and analyses of climate outcomes under varying socioeconomic and environmental trajectories.

Limitations. As shown in Figure 2, the GP model tends to underpredict the truth. This limitation is probably due to the limited size of the training dataset, which constrains the model's ability to fully capture the complexities of the true system. Additionally, while our model incorporates hyperparameter tuning, the exploration was restricted to three kernels and their basic combinations. A more extensive search over hyperparameter spaces and kernel combinations could improve the model's predictive accuracy. Providing larger and more diverse datasets would further enhance its performance.

Future Work. Future research should explore a broader range of kernels, more sophisticated kernel combinations, and advanced hyperparameter optimization techniques, such as Bayesian optimization or grid search. Expanding the training dataset and incorporating additional features could also improve the model's generalizability and accuracy in capturing complex climate dynamics.

4.3 Random Forest

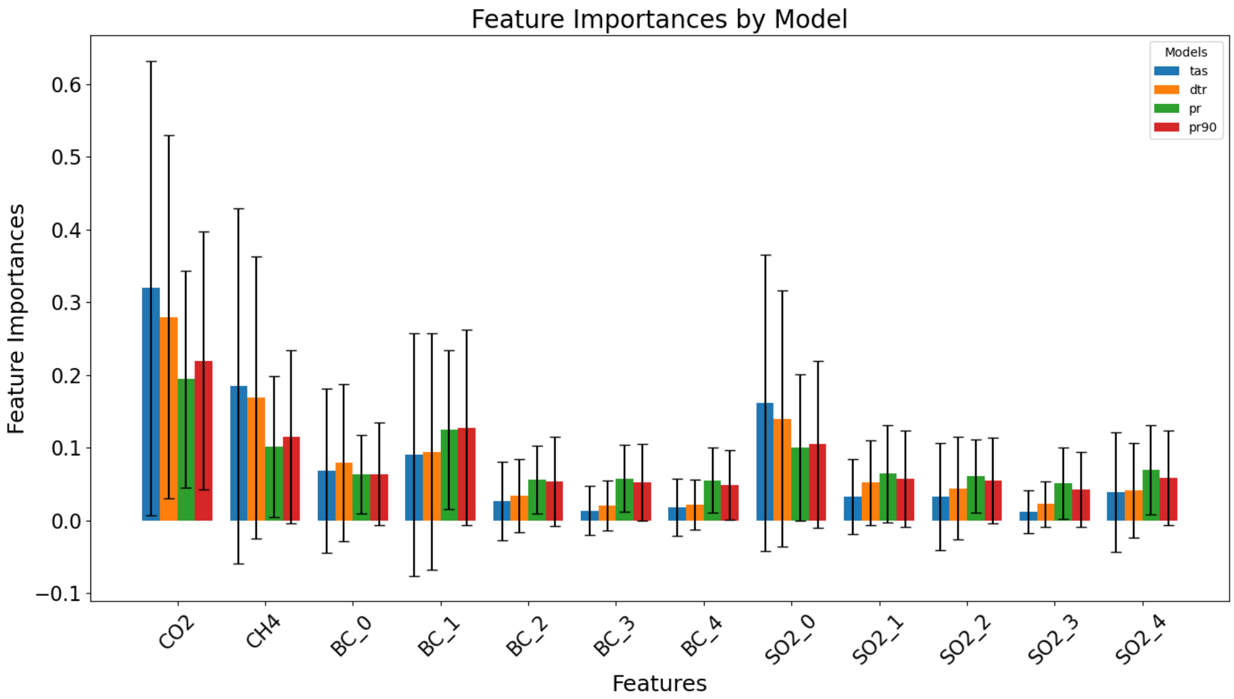


Figure 4: Feature Importance for Random Forest models.

Feature Importance. As we can observe in Figure 4, CO2 is the most important feature in the four models, followed by CH4, the first principal component of SO2, and the second principal component of BC. This aligns with our expectations. CO2 is the most prevalent greenhouse gas and plays an important role in climate change. CH4, although more potent, is slightly less prevalent in the atmosphere (U.S. Environmental Protection Agency 2024).

Impact and Applicability. One main advantage of using the Random Forest emulator is the interpretability of the results, in contrast to some ”black-box” models. A “black-box” model is an input-output model based on data that hides its internal workings. Being able to interpret the Random Forest model allows researchers to better inform policy makers when addressing climate change.

Limitations. An overall pattern of under-prediction can be observed for all four target variables, as shown in Figure 3. The random forest models likely underpredict due to their bias toward the mean, under representation of high target values in the training data, or insufficient modeling of spatial and temporal dependencies. Adjusting hyperparameters, improving features or using methods better suited to extremes could help.

Future Work. Future attempts of using random forest models to emulate climate models should try to address the issue of underprediction. Future attempts can also try more combinations of features, following the insights on feature importance presented above.

5 Conclusion

This study demonstrates the potential of machine learning emulators in addressing the computational challenges of exploring large climate model parameter spaces. By employing Pattern Scaling, Gaussian Processes, and Random Forest models, we effectively replicated key climate variables, including surface air temperature, precipitation, diurnal temperature range, and extreme precipitation percentiles. Among the tested methods, the Pattern Scaling model worked the best (as it had the lowest RMSE)!

Despite promising results, limitations in model accuracy, particularly in underpredicting extreme values and southern hemisphere temperatures, highlight the need for larger datasets and advanced tuning methods. Future work should focus on exploring richer datasets and robust hyperparameter optimization techniques. These enhancements could significantly improve emulator accuracy and adaptability, providing more precise insights into Shared Socioeconomic Pathways and their implications for future climate scenarios.

The findings highlight the transformative role of emulators in climate science, enabling broader access to predictive tools, and facilitating informed policymaking. As climate challenges intensify, such computational advancements will be indispensable for timely and actionable responses to global climate uncertainties.

References

Duncan Watson-Parris. 2021a. "Emulating with ESEm." [Link]

Duncan Watson-Parris. 2021b. "esem.gp_model." [Link]

Intergovernmental Panel on Climate Change (IPCC). 2023. "Climate Change 2023: Synthesis Report. Summary for Policymakers." https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_SPM.pdf

Knutti, Reto, David Masson, and Andrew Gettelman. 2013. "Climate model genealogy: Generation CMIP5 and how we got there." *Geophysical Research Letters* 40(6): 1194–1199. [Link]

Mansfield, Louise A., Peer J. Nowack, Mark Kasoar, and et al. 2020. "Predicting global patterns of long-term climate change from short-term simulations using machine learning." *npj Climate and Atmospheric Science* 3, p. 44. [Link]

NASA Climate Change. 2024. "Vital Signs of the Planet: Methane." [Link]

NASA Earth Observatory. 2017. "The Ups and Downs of Sulfur Dioxide in North America." [Link]

NOAA Climate.gov. 2024. "Climate Change: Atmospheric Carbon Dioxide." [Link]

Office of Environmental Health Hazard Assessment (OEHHA). 2022. "Atmospheric Black Carbon Concentrations." [Link]

U.S. Environmental Protection Agency. 2024. "Importance of Methane." https://www.epa.gov/gmi/importance-methane

Watson-Parris, Duncan, Yash Rao, Dirk Olivié, Christopher Kadow, Nicholas Leach, Jessica Vial, and Veronika Eyring. 2022. “ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections.” *Journal of Advances in Modeling Earth Systems* 14(10), p. e2021MS002954. [Link]

Appendices

A.1 Additional Figures A1

A.1 Additional Figures

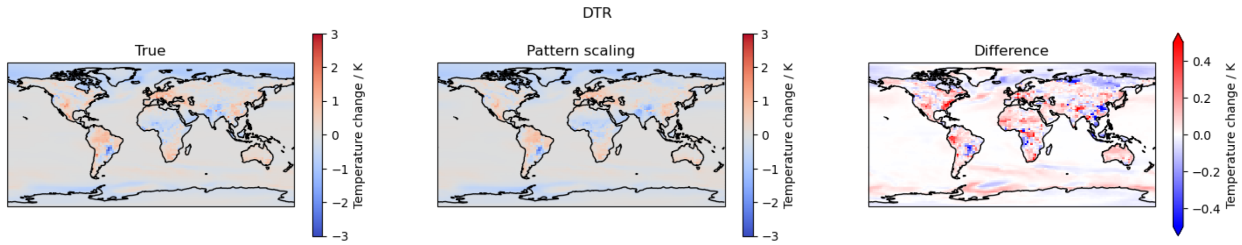


Figure A 1: Climate model projection vs. Pattern Scaling model projection for Diurnal Temperature Range.

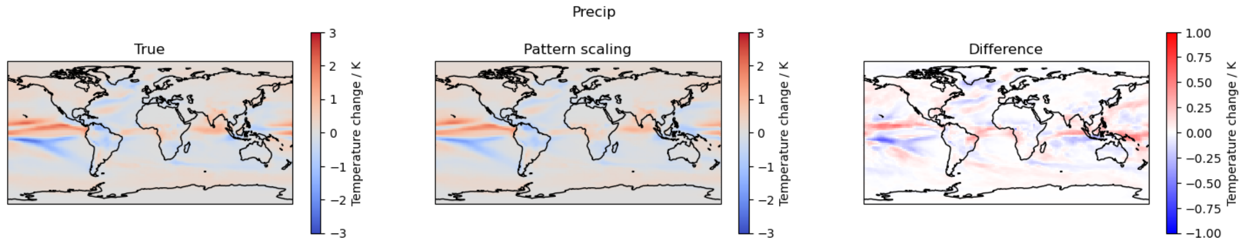


Figure A 2: Climate model projection vs. Pattern Scaling model projection for Precipitation.

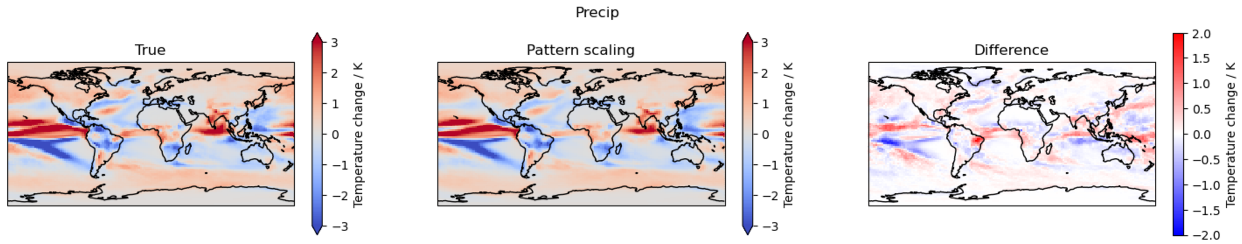


Figure A 3: Climate model projection vs. Pattern Scaling model projection for Precipitation 90th percentile.

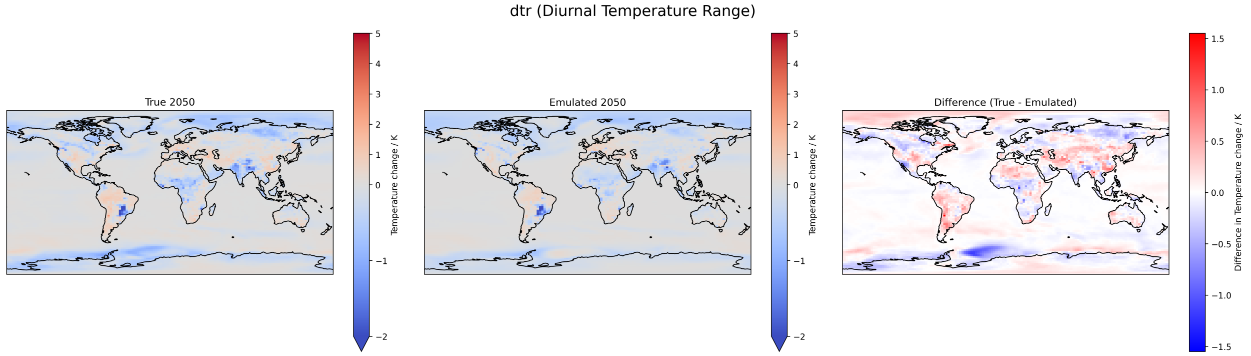


Figure A 4: Climate model projection vs. Gaussian Process model projection for Diurnal Temperature Range.

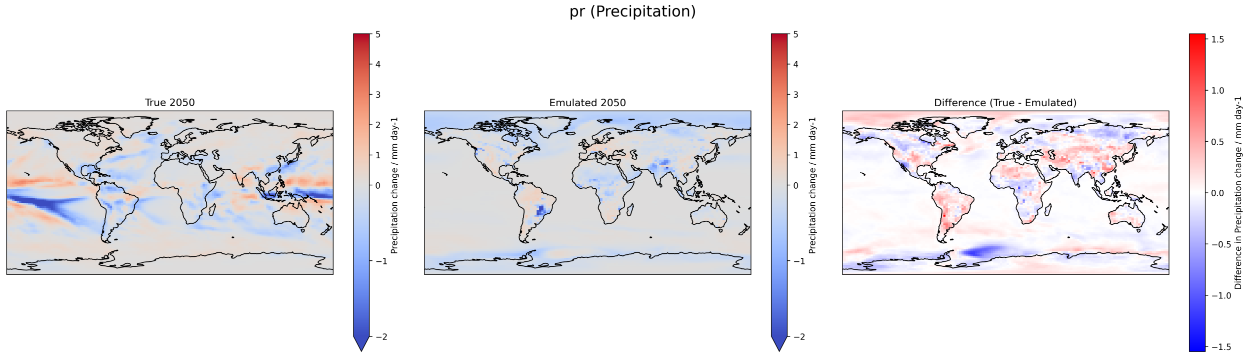

pr (Precipitation)
True 2050
Emulated 2050
Difference (True - Emulated)
Precipitation change / mm day-1
Precipitation change / mm day-1
Difference in Precipitation change / mm day-1
5
4
3
2
1
0
−1
−2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

Figure A 5: Climate model projection vs. Gaussian Process model projection for Precipitation.

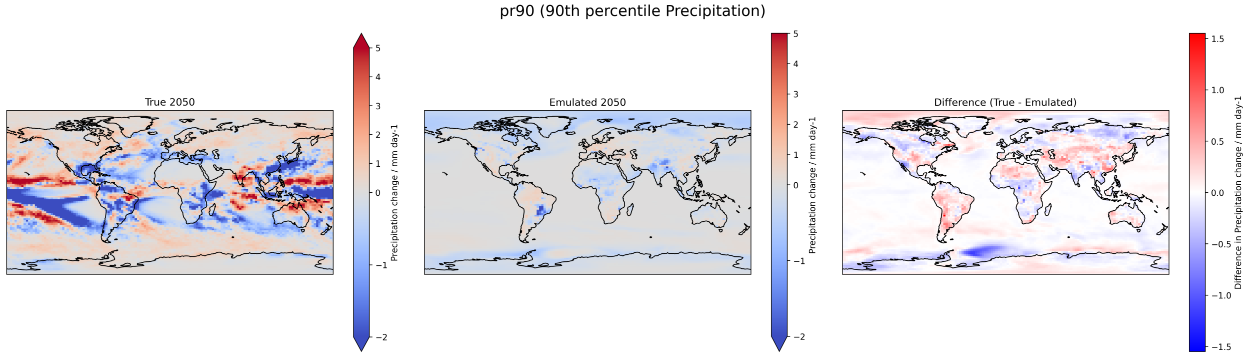

pr90 (90th percentile Precipitation)
True 2050
Emulated 2050
Difference (True - Emulated)
Precipitation change / mm day-1
Precipitation change / mm day-1
Difference in Precipitation change / mm day-1
5
4
3
2
1
0
−1
−2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

Figure A 6: Climate model projection vs. Gaussian Process model projection for Precipitation 90th percentile.

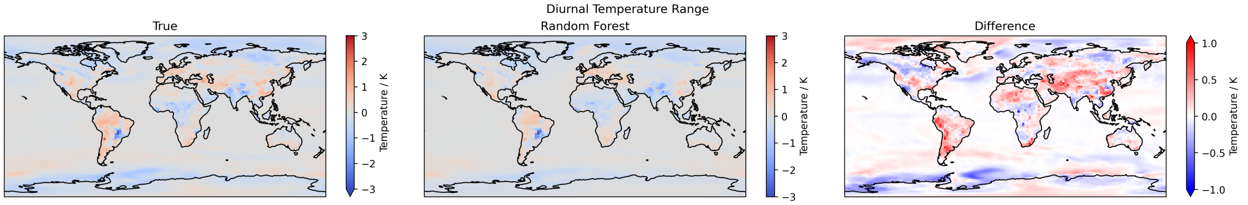


Figure A 7: Climate model projection vs. Random Forest projection for Diurnal Temperature Range.

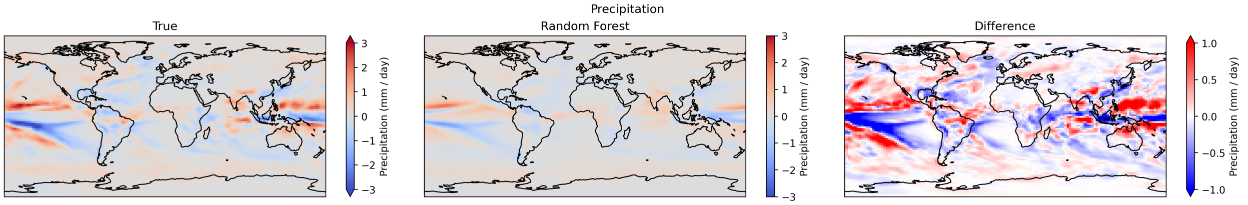


Figure A 8: Climate model projection vs. Random Forest projection for Precipitation.

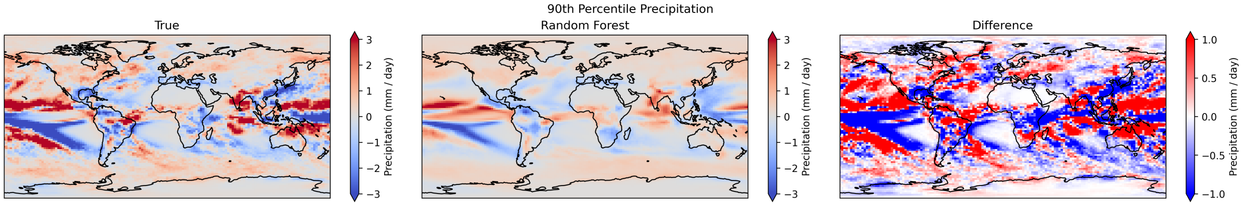


Figure A 9: Climate model projection vs. Random Forest projection for 90th Percentile Precipitation.