Do machine learning climate models work in changing climate dynamics?

Maria Conchita Agana Navarro
Centre for Artificial Intelligence, Department of Computer Science
University College London
m.navarro.23@ucl.ac.uk

Geng Li
The Hong Kong University of Science and Technology
glibd@connect.ust.hk

Theo Wolf
University of Oxford
theo@robots.ox.ac.uk

María Pérez-Ortiz
Centre for Artificial Intelligence, Department of Computer Science
University College London
maria.perez@ucl.ac.uk

Abstract

Climate change is accelerating the frequency and severity of unprecedented events, deviating from established patterns. Predicting these out-of-distribution (OOD) events is critical for assessing risks and guiding climate adaptation. While machine learning (ML) models have shown promise in providing precise, high-speed climate predictions, their ability to generalize under distribution shifts remains a significant limitation that has been underexplored in climate contexts. This research systematically evaluates state-of-the-art ML-based climate models in diverse OOD scenarios by adapting established OOD evaluation methodologies to climate data. Experiments on large-scale datasets reveal notable performance variability across scenarios, shedding light on the strengths and limitations of current models. These findings underscore the importance of robust evaluation frameworks and provide actionable insights to guide the reliable application of ML for climate risk forecasting.

1 Introduction

Climate change is driving more frequent and unprecedented extreme climate events (Arias et al., 2021; Kornhuber et al., 2024; Team et al., 2023). Climate models are vital tools for managing these risks, simulating interactions among the atmosphere, oceans, land, and ice (Kaltenborn et al., 2024). Machine learning (ML) has emerged as a transformative addition to climate modeling, offering computationally efficient approximations of climate processes (Kaltenborn et al., 2023). However, ML-based models have been shown to fail to generalize well under distribution shifts, when the real-world data deviates significantly from the ML training data (Gagnon-Audet et al., 2023). This raises concerns about the reliability of using ML for climate modelling in out-of-distribution (OOD) scenarios. For example, unprecedented combinations of greenhouse gas concentrations or tipping points such as ice sheet collapse could challenge ML models' predictive robustness. While benchmarks like ClimateBench (Watson-Parris et al., 2022) and ClimateSet (Kaltenborn et al., 2023) provide standardized ML-ready datasets, they currently evaluate models under limited conditions (e.g. only SSP-2.45). Even recent, state-of-the-art models like NeuralGCM (Kochkov et al., 2024) face challenges when extrapolating to extreme warming scenarios. Understanding the reliability of climate variable predictions, such as temperature and precipitation, is critical, with implications for early warning systems, agricultural planning, and water resource management. In high-vulnerability regions, where adaptive capacity is limited, unreliable predictions can exacerbate risks.

This paper introduces a novel evaluation methodology that integrates principles of OOD testing to assess the robustness of ML-based climate models under shifting climate dynamics, addressing a gap in more holistic assessments of model robustness. This study contributes to understanding the reliability of ML models in predicting climate risks under real-world conditions.

2 Data and Methodology

Climate data is time-series data, consisting of sequences of data points indexed over time; climate observations, such as temperature or precipitation, are collected at different time steps. Each data point X_t represents a specific observation at time $t \in T$ and the corresponding label Y_t could represent a future climate state, such as temperature or precipitation projections.

The WOODS framework (Gagnon-Audet et al., 2023) provides a systematic approach to evaluating OOD performance for time-series tasks across diverse domains, including neurophysiology, sign language, and energy consumption. OOD settings occur when ML models trained on one type of data distribution encounters data from a different, unseen distribution. The WOODS framework proposes different types of shifts in time-series data. Below we outline how these shifts can be applied to ML climate modeling tasks.

2.1 Adapting Existing OOD Evaluation Frameworks to ML Climate Models

Climate data is typically drawn from different geographical regions or climate regimes. We can define each domain d as a specific geographical area or climate regime, with associated distribution $P_d(X_t, Y_t)$. For ML climate models, the objective is to create a model f that generalizes well across different climate domains. To achieve this, we define the training data as:

$$E_{\text{train}} = \{d_1, d_2, ..., d_n\}$$

where E_{train} represents the set of geographical regions or climate regimes used for training. The model is then evaluated on unseen domains E_{all}, where $E_{\text{train}} \subseteq E_{\text{all}}$.

To quantify generalization, we minimize the model's worst-case risk across all possible domains:

$$\min_f \max_{d \in E_{\text{all}}} R_d(f)$$

where $R_d(f)$ is the risk or loss on domain d, defined as:

$$R_d(f) = \mathbb{E}_{(X_t,Y_t)\sim P_d(X,Y)}[\ell(f(X_t), Y_t)]$$

where ℓ is the loss function, and $f(X_t)$ is the predicted climate state.

WOODS highlights two key scenarios for distribution shifts in time-series data:

1. **Time-Domain Shifts:** Occur when the data distribution changes over time, which can be due to long-term trends, seasonal variations, or unexpected events.
2. **Source-Domain Shifts:** Occur when training data comes from a domain that differs in underlying factors or conditions, such as different data sources.

ML climate models are susceptible to both time and source-domain shifts. Climate data distributions can change over time, due to seasonal patterns, long-term climate trends like global warming, or events such as El Niño. Human-driven changes, such as land-use changes and urbanization, may also contribute to temporal shifts in climate data. Climate models can also encounter domain-shifts with training data originating from regions or climate regimes that differ from data encountered in test settings or real-world use. Examples include shifts based on differences in the datasets' geographical regions (e.g., tropical vs. temperate climates) or data sources (e.g., climate change scenarios defined by Shared Socioeconomic Pathways vs ground-based observations).

Building on time-domain and source-domain shifts, we propose two distinct methodologies to evaluate the ability of ML climate models to generalize under such conditions.

Method 1 - Split Based on Time Period (Time-Domain Shift): This method evaluates how well ML climate models generalize across data from different time periods. The models are trained on data from 1850–2014 and tested on data from 2015–2023, treating the recent period as an OOD scenario. We hypothesize that temporal shifts in climate patterns over time, such as increased global warming, may challenge the models' ability to generalize. This period is also chosen for testing, as ClimaX's pretraining data is limited to up to 2015 (Nguyen et al., 2023).

Method 2 - Evaluate Across Multiple SSP Scenarios (Source-Domain Shift): This method assesses model performance when trained and tested on different Shared Socioeconomic Pathways (SSPs). Each SSP represents a distinct trajectory of global development, characterized by differences in socioeconomic factors such as population growth or energy use, which leads to varying climate forcing trajectories and temperature projections (Kaltenborn et al., 2024). We hypothesize that training on one set of SSPs and testing on another can simulate OOD conditions because the scenarios embody shifts in the distributions of climate drivers and their downstream effects. For example, SSP scenarios with high emissions (e.g. SSP5) differ in climate forcings compared to low-emission pathways (e.g., SSP1). While existing benchmarks often evaluate models against a single SSP scenario, this method expands on that by evaluating performance across multiple SSPs.

3 Experiments and Results

3.1 Experiment Setup: Dataset, models, task, and evaluation metrics

We use the ClimateSet dataset, which integrates inputs and outputs of temperature and precipitation from Input4MIPs and CMIP6 across 36 traditional climate models (Kaltenborn et al., 2023). The dataset focuses on four major climate forcing agents (CO_2, CH_4, BC, and SO_2) and four SSPs. It provides a diverse set of temporal and spatial features across multiple geographic regions, enabling us to assess model performance under various distribution shift conditions. The ClimateSet authors have also documented benchmark performance for state-of-the-art (SoTA) ML climate models, serving as a comparison point for the performance of these models in out-of-distribution settings. We utilize the implementation of these ML models within ClimateSet for consistency.

Models: We employ four advanced machine learning models (U-Net, ConvLSTM, ClimaX, and ClimaX Frozen) which have demonstrated strong performance on ClimateSet's benchmark datasets.

Task: The core task being evaluated is climate emulation, where the ML models aim to replicate the outputs of traditional climate models. We use 5 traditional climate models (AW1-CM-1-1-MR, EC-Earth3, FGOALS-f3-L, BCC-CSM2-MR, MPI-ESM1-2-HR) used in ClimateSet's baseline experiments, allowing for comparison with their experiments.

Evaluation: We evaluate the model's performance using the latitude-longitude weighted root mean squared error (RMSE), the metric reported from ClimateSet (Kaltenborn et al., 2023). Predictions of monthly surface air temperature and precipitation are assessed against outputs from traditional climate models.

3.2 Experiment Steps

1) Run ClimateSet baselines: We first run the baseline; ClimateSet single emulator specifications (Kaltenborn et al., 2023), for each ML model, establishing a clear baseline for comparison. Each emulator receives as input the climate forcing emission fields of CO_2, CH_4, BC, and SO_2, with the output variables being climate model predictions of temperature and precipitation. The baseline training dataset includes 165 years of historical data (1850-2014) as well as 86-years of climate predictions for 3 SSP scenarios (2015-2100). For training, the historical data, SSP1-2.6, SSP3-7.0, and SSP5-8.5 are utilized. A random 10% of this data is withheld for validation, and the SSP2-4.5 scenario is used for testing.

Table 1: Percent change (%) in RMSE values between baseline performance and performance under time-domain and source-domain shifts for each ML model single-emulating surface air temperature (TAS) and precipitation (PR) across five climate models.

			U-Net		Conv-LSTM		ClimaX		ClimaX $_{\text{Frozen}}$	
			TAS	PR	TAS	PR	TAS	PR	TAS	PR
Time-Domain Shift		AWI-CM-1-1-MR	-19.94	-18.40	-4.49	-5.71	-7.98	-7.49	-22.14	-21.48
		EC-Earth3	-22.16	-21.83	1.93	1.94	-7.09	-6.06	-21.38	-21.38
		FGOALS-f3-L	-9.07	-12.97	1.66	-1.87	-3.72	-7.09	-13.16	-17.06
		BCC-CSM2-MR	-15.70	-15.94	-3.73	-4.86	-8.47	-9.80	-17.93	-18.51
		MPI-ESM1-2-HR	-25.14	-25.47	-3.78	-3.24	-1.49	-1.97	-14.43	-16.05
Source-Domain Shift	**SSP1-2.6**	AWI-CM-1-1-MR	66.98	56.13	-2.70	-3.20	60.84	56.93	85.67	83.75
		EC-Earth3	52.85	49.09	-1.72	-2.20	68.66	67.28	75.03	74.52
		FGOALS-f3-L	48.28	45.14	0.00	-0.07	30.20	29.20	56.25	56.48
		BCC-CSM2-MR	44.82	46.56	2.25	1.28	76.37	68.54	72.64	68.78
		MPI-ESM1-2-HR	58.33	57.09	-4.01	-4.09	134.87	137.02	104.12	105.59
	SSP3-7.0	AWI-CM-1-1-MR	-14.33	-14.11	1.35	1.60	0.38	0.75	-6.42	-6.48
		EC-Earth3	-11.84	-11.90	0.36	0.26	4.58	5.00	-4.85	-5.45
		FGOALS-f3-L	-10.84	-10.65	-0.75	0.28	-0.86	0.57	-4.02	-3.31
		BCC-CSM2-MR	-11.46	-10.74	-0.50	-1.06	3.65	3.34	-0.74	0.12
		MPI-ESM1-2-HR	-24.95	-24.85	1.10	2.16	14.98	16.28	-1.39	-1.03
	SSP5-8.5	AWI-CM-1-1-MR	69.47	69.02	1.57	2.53	10.00	10.86	17.06	19.30
		EC-Earth3	52.61	57.30	1.73	1.47	19.90	22.10	26.39	24.31
		FGOALS-f3-L	41.64	46.45	1.60	2.82	12.05	14.44	2.83	2.55
		BCC-CSM2-MR	27.78	27.35	-4.03	-3.07	-1.45	0.24	12.79	11.36
		MPI-ESM1-2-HR	50.58	55.90	-1.36	0.10	36.58	42.04	12.55	11.05

2) Evaluate model performance under distribution shifts: We then run experiments assessing model performance under distribution shifts. For Method 1 (time-domain shift), the training data is restricted to data from the period 1850-2014, while the test data covers 2015-2023. For Method 2 (source-domain shift), the train-test split is varied across the different Shared Socioeconomic Pathways (SSPs). Three scenarios are tested, each SSP scenario being used as the test set in turn: SSP1-2.6, SSP3-7.0, and SSP5-8.5.

3.3 Results

The results reveal that ML models exhibit different strengths and weaknesses depending on the nature of the OOD data. Under a time-domain shift, most models demonstrated improved or comparable performance to the baseline for both temperature and precipitation predictions (Table 1). Improvements were indicated by negative percent changes in RMSE scores, with ClimaX consistently achieving the lowest RMSE scores across all climate models for both variables (Appendix C Figure 1). This aligns with ClimaX's strong performance in ClimateSet's benchmark experiment, showcasing its robustness in generalization. The enhanced performance under time-domain shifts suggests that ClimaX's architecture may better capture temporal dependencies in climate data. Simpler models like U-Net also exhibited notable improvements, indicating that even less complex architectures can generalize well in this setting. The overall trend suggests the models may be biased towards patterns resembling historical data, with stronger generalization to near-future conditions.

Under source-domain shifts, RMSE values generally increased compared to the baseline, indicating reduced performance (Table 1, Appendix C Figure 2). However, certain scenarios presented exceptions where models performed comparably or better. Among the models, ConvLSTM emerged as the most consistent performer across scenarios, exhibiting minimal performance variation. It achieved the best results for SSP1-2.6, indicating strong adaptability to lower-emission scenarios. In contrast, U-Net struggled significantly with SSP1-2.6 and SSP5-8.5 but excelled in SSP3-7.0. ClimaX and ClimaX Frozen also encountered challenges with SSP1-2.6 and SSP5-8.5, suggesting potential limitations in capturing the variability introduced by extreme SSP scenarios.

Variations in performance carry relevance for policymakers and planners relying on climate forecasts; an observed 20–30% increase in RMSE under specific OOD scenarios could, for example, indicate underestimations of the frequency and severity of temperature and precipitation events by ML models. Recognizing these inaccuracies during evaluation can inform decision-makers about models' predictive reliability in evolving climate conditions and guide any actions, such as further review, before decisions are made.

4 Conclusion and Future Work

This study evaluated the performance of SoTA ML climate models under OOD scenarios, developing a novel evaluation framework tailored to performance assessment under evolving climate conditions. Key recommendations include rotating time and source-domain scenarios between training and testing to ensure a comprehensive evaluation of model robustness across diverse climate domains. The framework is adaptable and can be used to evaluate other climate models. The exclusion of more recent SoTA ML models such as NeuralGCM (Kochkov et al., 2024) limits the scope of comparison. Future research should aim to evaluate these SoTA models within the framework. Other promising avenues include expanding the framework to address other climate tasks beyond global climate model emulation, such as downscaling and regionalized predictions, as well as consider multiple random seeds and re-analysis datasets for more generalized insights. As ML's role advances to support our understanding, proactive evaluation will be crucial for reliable decision-making and ensuring these tools effectively contribute to addressing our changing climate.

References

Arias, P., Bellouin, N., Coppola, E., Jones, R., Krinner, G., Marotzke, J., Naik, V., Palmer, M., Plattner, G.-K., Rogelj, J., Rojas, M., Sillmann, J., Storelvmo, T., Thorne, P., Trewin, B., Achuta Rao, K., Adhikary, B., Allan, R., Armour, K., … Zickfeld, K. (2021). Technical summary. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (Eds.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (pp. 33–144). Cambridge University Press. https://doi.org/10.1017/9781009157896.002

Gagnon-Audet, J.-C., Ahuja, K., Darvishi-Bayazi, M.-J., Mousavi, P., Dumas, G., & Rish, I. (2023). Woods: Benchmarks for out-of-distribution generalization in time series. https://arxiv.org/abs/2203.09978

Kaltenborn, J., Lange, C. E. E., Ramesh, V., Brouillard, P., Gurwicz, Y., Nagda, C., Runge, J., Nowack, P., & Rolnick, D. (2023). Climateset: A large-scale climate model dataset for machine learning. https://arxiv.org/abs/2311.03721

Kaltenborn, J., Lange, C. E. E., Ramesh, V., Brouillard, P., Gurwicz, Y., Nagda, C., Runge, J., Nowack, P., & Rolnick, D. (2024). Welcome to climateset's documentation! - climateset 0.1 documentation [[Accessed: Aug. 05, 2024]].

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., & Hoyer, S. (2024). Neural general circulation models for weather and climate. *Nature*, *632*(8027), 1060–1066. https://doi.org/10.1038/s41586-024-07744-y

Kornhuber, K., Bartusek, S., Seager, R., Schellnhuber, H. J., & Ting, M. (2024). Global emergence of regional heatwave hotspots outpaces climate model simulations [_eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2411258121]. *Proceedings of the National Academy of Sciences*, *121*(49), e2411258121. https://doi.org/10.1073/pnas.2411258121

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). Climax: A foundation model for weather and climate. https://arxiv.org/abs/2301.10343

Team, C. W., Lee, H., & Romero, J. (Eds.). (2023). *Climate change 2023: Synthesis report. contribution of working groups i, ii and iii to the sixth assessment report of the intergovernmental panel on climate change*. IPCC. https://doi.org/10.59327/IPCC/AR6-9789291691647

Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., Stier, P., Bouabid, S., Dewey, M., Fons, E., Gonzalez, J., Harder, P., Jeggle, K., Lenhardt, J., Manshausen, P., Novitasari, M., Ricard, L., & Roesch, C. (2022). Climatebench v1.0: A benchmark for data-driven climate projections [e2021MS002954 2021MS002954]. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2021MS002954. https://doi.org/https://doi.org/10.1029/2021MS002954

A Training Details

Our baseline runs followed the ClimateSet single emulator specifications (Kaltenborn et al., 2023):

- **Training Process:** Each emulator is trained on data from a single climate model, predicting outputs for an entire sequence of monthly data for each year.
- **Pre-Processing:** The data has been pre-processed by ClimateSet to have a spatial resolution of approximately 250 km (144 x 96 longitude-latitude cells) and a temporal resolution of monthly data. The time series is divided into 1-year chunks, resulting in data with a shape of $\langle$ *scenarios, years * months, variables, longitude, latitude* $\rangle$.
- **Input and Output Shapes:** The input data has the shape $\langle$ *batch, sequence length, num vars, lon, lat* $\rangle$, where the sequence length is 12 (monthly data). The output has the shape $\langle$ *batch, sequence length, 2, lon, lat* $\rangle$, where the '2' corresponds to temperature (TAS) and precipitation (PR).
- **Training Parameters:** The models are trained for 50 epochs with an initial learning rate of 2e-4, using an exponential decay scheduler. For the non-frozen ClimaX models, training begins with a 5-epoch warm-up phase at 1e-8, followed by training at 5e-4.
- **Loss:** The latitude-longitude weighted mean squared error (LLMSE) as implemented in (Nguyen et al., 2023) is used.

B Evaluation Metrics Details

The metric reported in our experiments is the Latitude-Longitude Weighted Root Mean Squared Error (RMSE), implemented in (Nguyen et al., 2023) and used in (Kaltenborn et al., 2023) ClimateSet experiments. It is a modified version of the Normalized Root Mean Squared Error (NRMSE) that accounts for varying grid sizes at different latitudes (Nguyen et al., 2023).

C Detailed Results

Figures 1 and 2 summarize the performance of all ML models for surface air temperature (TAS) and precipitation (PR). These comparisons evaluate baseline performance against performance under time-domain and source-domain distribution shifts.

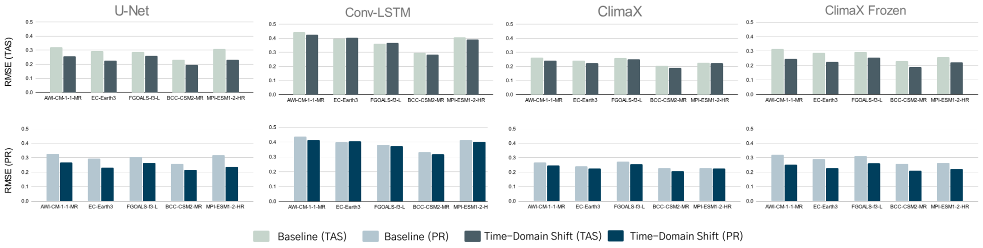

Figure 1: Comparison of RMSE between baseline performance and performance under a time-domain shift for each ML model emulating surface air temperature (TAS) and precipitation (PR) across five climate models.

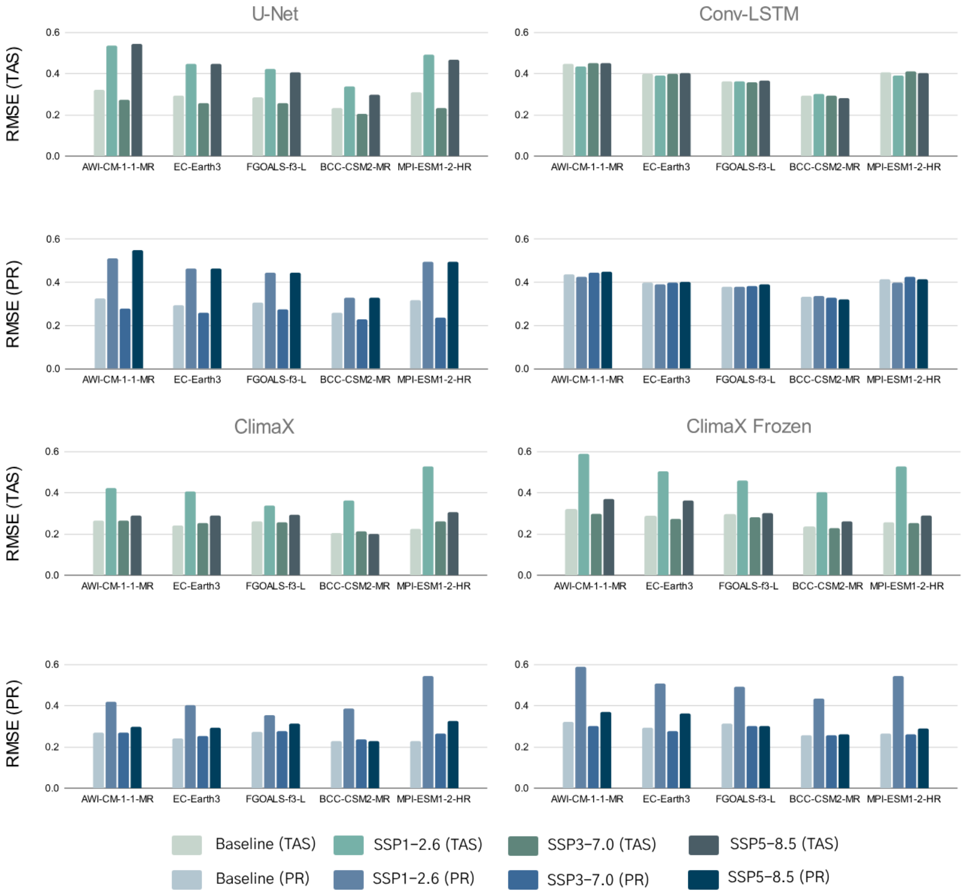

Figure 2: Comparison of RMSE between baseline performance and performance under source-domain shifts for each ML model emulating surface air temperature (TAS) and precipitation (PR) across five climate models.