Exploring Vision Transformers for Early Detection of Climate Change Signals

Sungduk Yu[1]*, Brian L. White[2], Anahita Bhiwandiwalla[1], Yaniv Gurwicz[1], Musashi Hinck[1], Matthew Lyle Olson[1], Raanan Rohekar[1] Vasudev Lal[1]

[1]Intel Labs, [2]UNC Chapel Hill

Abstract

This study evaluates Vision Transformers (ViTs) for detecting anthropogenic climate change signals, crucial for effective policy planning and risk assessment. Compared to previously suggested models like CNN, MLP, and ridge regression, ViTs consistently detect forced climate signals earlier across three reanalysis datasets (ERA5, JRA-3Q, and MERRA-2). Interpretation with Integrated Gradients reveals consistent spatial patterns, suggesting ViTs utilize physically grounded signals. This work highlights ViTs' potential to advance climate change detection and attribution tasks.

1 Introduction

Understanding the drivers behind climate change is essential for developing accurate climate projections and shaping effective policies. At the heart of this endeavor is the Detection and Attribution (D&A) of climate change signals, a process focused on distinguishing the signals of human-induced warming from the inherently noisy patterns of natural climate variability. The Intergovernmental Panel on Climate Change (IPCC) has highlighted the critical importance of accurately identifying these anthropogenic influences to inform targeted mitigation and adaptation strategies effectively [1].

Objective. This paper explores the application of advanced artificial intelligence models, specifically Vision Transformers (ViTs), to extend the ongoing efforts of climate change Detection and Attribution (D&A) studies. Considering that climate fields are globally interconnected spatially, we hypothesize that the global attention mechanism of ViTs could significantly improve D&A models. This approach is particularly promising given the recent successes of MLPs and CNNs in similar tasks.

Related Work. Traditional approaches in climate Detection and Attribution (D&A) have largely relied on statistical methods such as Principal Component Analysis (PCA) to identify spatial fingerprints from *long-term* climate records, indicative of external forcings like greenhouse gas emissions [2–4]. Recent advancements have seen a shift towards incorporating machine learning techniques, including ridge regression, Multilayer Perceptrons (MLPs), and Convolutional Neural Networks (CNNs), adept at detecting climate change signals even in *daily* weather snapshots [5–8].

2 Dataset and Methods

2.1 Dataset

To train our climate change detection models, we used global climate model outputs from CMIP6 archive. For evaluation on observation records, we utilized three modern reanalysis products.

*sungduk.yu@intel.com

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2024.

CMIP6. CMIP6 is a globally coordinated climate modeling initiative, featuring over 100 models from more than 50 research groups [9]. It encompasses historical simulations covering 1850 to 2014, along with ScenarioMIP projections extending from 2015 to 2100 under diverse socioeconomic pathways, thereby offering a comprehensive dataset for climate research. For model training, we utilize the ClimDetect dataset [10], a dataset curated from CMIP6 specifically designed for climate change detection tasks, covering the period of 1950-2100. For the hypothesis testing, we prepare an independent (that is, not used in model training) CMIP6 dataset for 1850-1949 for estimating natural variability. It comprises the same 7 climate models included in the ClimDetect test set.

Reanalysis. Reanalysis datasets synthesize observations with model outputs to create continuous, globally comprehensive climate records, offering a more consistent alternative to sparse observational data alone. We use three modern reanalysis dataset that serves as the closest alternative to direct observations where continuous data coverage is required for robust model evaluation and hypothesis testing: ERA5 (data span: 1940-present) [11], JRA-3Q (1950-present) [12], and MERRA2 (1980-present) [13]. In the context of our study, 'observations' refers to reanalysis datasets.

2.2 Climate Change Detection

Our method builds on Sippel et al (2020) [7], but distinguishes itself by employing modern AI architectures instead of traditional ridge regression models. For a visual illustration, see Figure S1.

Step 1: Climate Change Detection Model Training. Detection models in most prior studies focus on extracting 'fingerprints'—spatial patterns anticipated to emerge due to external forcings such as greenhouse gas emissions. With the application of nonlinear machine learning models, these 'fingerprints' are reinterpreted as complex nonlinear functions. These models are trained to discern anthropogenic climate signals from the natural variability present in daily climate data. Specifically, these functions (F_θ) are trained on the CMIP6 dataset to map input *daily* climate fields ($\mathbf{X}$) to a *annual* scalar target variable (y), a key climate change indicator, establishing a model for climate change signal detection, i.e., $y = F_\theta(\mathbf{X})$[2].

Step 2: Hypothesis Testing. The null hypothesis posits that the predicted test statistic falls within the range expected under natural variability. We estimate the distribution of natural variability, $P(y_{\text{hist}}) = P(F_\theta(\mathbf{X}_{\text{hist}}))$, by predicting test statistics from the historical ("pre-warming") CMIP6 dataset for the period 1850–1949. Then, we apply the trained model to reanalysis datasets to obtain observed test statistics $y_{\text{obs}} = F_\theta(\mathbf{X}_{\text{obs}})$. Finally, we test the null hypothesis by assessing if y_{obs} is distinguishable from the estimated natural variability, e.g., 2.5th–97.5th percentile range of $P(y_{\text{hist}})$.

Year of Emergence. We quantify hypothesis testing outcomes with the Year of Emergence (YOE), an important metric for climate projections and policy planning. YOE is defined as the first year when climate change signals statistically surpass daily natural variability (Figure S2). An earlier YOE indicates a more sensitive detection model, implying better performance in extracting climate change signals. For robust detection, we establish an ad-hoc threshold for the emergence fraction (EF; defined as the ratio of days on which climate change is detected to the total days in a year) at 97.5%, equivalent to 356 days.

Table 1: RMSE across different models and experiments, calculated over the ClimDetect test set that spans 150 years (1950-2100) [Unit: °C]. RMSE values are underlined if their 95% confidence interval, determined by resampling the test set with replacement 10,000 times, overlap with that of the best RMSE.

	tas-huss-pr	tas	pr	huss	tas-huss-pr_mr	tas_mr
CLIP	**0.1411**	**0.1482**	0.8935	0.1801	0.1690	<u>0.2410</u>
DINOv2	0.1439	0.1645	0.7995	0.1942	0.1731	0.2552
MAE	0.1430	<u>0.1484</u>	0.6451	**0.1571**	**0.1672**	0.2531
ViT-b/16	0.1425	0.1610	0.7132	0.1604	0.1763	0.2562
ResNet-50	0.1471	0.1687	**0.6137**	0.1661	0.1835	0.2693
MLP	0.1488	0.1557	0.7502	0.1804	0.2192	<u>0.2409</u>
ridge	0.1508	0.1542	0.9708	0.2304	0.2156	**0.2404**

[2] $y \in \mathbb{R}$; $\mathbf{X} \in \mathbb{R}^{64\times128\times3}$, where 64 for latitude, 128 for longitude, and 3 for channels (climate variables).

Table 2: RMSE calculated over the most recent 24 years (1980-2023) of ERA-5 data [Unit: °C]. Corresponding RMSE tables for JRA-3Q and MERRA-2 are presented in Appendix A.3.

	tas-huss-pr	tas	pr	huss	tas-huss-pr_mr	tas_mr
CLIP	0.1069	0.1287	**0.5218**	0.1925	0.1797	0.1853
DINOv2	0.1123	0.1376	0.5876	0.1920	0.1596	0.1807
MAE	**0.0941**	0.1076	0.7649	**0.1319**	**0.1317**	0.1534
ViT-b/16	0.1039	**0.0878**	1.0109	0.1678	0.1343	**0.1480**
ResNet-50	0.0982	0.0912	0.6478	0.1764	0.1613	0.1885
MLP	0.1009	0.1091	0.6185	0.1626	0.1708	0.1744
ridge	0.0967	0.1026	0.5228	0.1861	0.1498	0.1792

2.3 Models

We finetuned four pretrained Vision Transformers (ViTs) with a regression head—CLIP [14], DINOv2 [15], MAE [16], and ViT-b/16 [17]. To benchmark these ViTs against models suggested by previous studies, we also trained additional models including CNN (ResNet-50 [18]), MLP, and ridge regression. Details on model setups and training dataset specifics are provided in Appendix A.2.

2.4 Evaluation Configurations

We evaluate our models using six configurations detailed in the `ClimDetect` [10] to predict annual global mean temperature (AGMT). The input variables—surface air temperature ("tas"), surface specific humidity ("huss"), and precipitation rate ("pr")—reflect those commonly used in climate change detection and attribution studies. Single-variable experiments are named after the input variable (e.g., "tas"). The combined variable setup is termed "tas-huss-pr," and mean-removed versions are indicated by "_mr" suffixes (e.g., "tas_mr"), presenting more challenging tasks where models must rely solely on spatial patterns without mean signals.

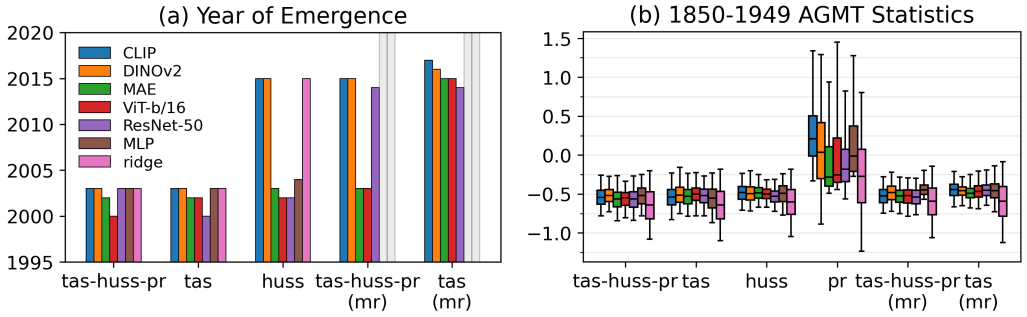

Figure 1: (a) Year of emergence (YOE), defined as the first year when the majority of daily climate fields show a distinguishable climate change signal from natural variability. Grey bars indicate instances where a model failed to capture YOE within the reanalysis period of 1980-2023. "pr" is omitted since no detection model can capture YOE. (b) Box plots showing the distribution of test statistics (AGMT) under natural variability, $P(y_{\text{hist}})$. The boxes represent the median and interquartile range; the whiskers the 2.5th–97.5th percentile range.

3 Results

Evaluation on CMIP6 (1950-2100) from ClimDetect Test Set. In experiments like "tas" and "tas_mr", simple models such as ridge regression and MLP perform on par with ViTs and CNNs (Table 1). However, ViTs demonstrate enhanced capabilities in more complex configurations, such as with multi-variable inputs ("tas-huss-pr") and particularly when the mean signal is removed ("tas-huss-pr_mr"), suggesting their potential uses in more sophisticated climate detection tasks. Challenges remain in detecting climate signals using only precipitation rate, likely due to its sparsity and the significant uncertainties associated with precipitation responses under climate change [19].

Evaluation on Observation (1980-2023). Despite subtle differences, RMSE on ERA5 broadly aligns RMSE on CMIP6 test set, showing the ViTs performs better in most experiments (Table 2). While MAE and ViT-b/16 consistently show low RMSE for most variables except "pr", CLIP and DINOv2 do not uniformly outperform simpler models like MLP and Ridge Regression, particularly in configurations such as "tas-huss-pr", "huss", and "tas_mr". Notable discrepancies across other reanalysis, like JRA-3Q and MERRA2 (Tables S1, S2), suggest variability in model performance, likely due to the differing assimilation models and observational inputs used in these reanalysis systems. Additionally, RMSE values are generally lower on observation data than on the CMIP6 data, likely due to varying evaluation periods rather than model generalization. For example, the uncertainties in CMIP6 output increase over the projection period.

Year of Emergence. In contrast to RMSE, YOE distinctly highlights the effectiveness of sophisticated models like ViTs and CNNs (Figure 1a). Across all experiments, MAE, ViT-b/16, and ResNet-50 consistently show the earliest YOE. Conversely, ridge regression and MLP perform comparably to less effective ViTs such as CLIP and DINOv2, and fail to detect an emergence in the mean-removed experiments (tas_mr and tas-huss-pr_mr) at the 97.5% EF threshold. This finding is consistent across various EF thresholds (Figure S4), showing that MAE, ViT-b/16, and ResNet-50 either outperform or match the performance of simpler models across different experimental settings.

Interpretability. Physical interpretability remains crucial for establishing data-driven models as a new tool in climate science. We show preliminary model interpretations using Integrated Gradients [20] for the "tas-huss-pr_mr" experiment, where a wide performance gap between simple vs. advanced models are seen, revealing distinct differences between nonlinear ML models and ridge regression (Figure 2). Unlike ridge, ViTs (along with CNN and MLP) exhibit a diminished focus on land-sea contrasts and a greater positive dependence on the Antarctic Ocean. These consistent patterns across different architectures suggest that ViTs may be underpinned by physical processes.

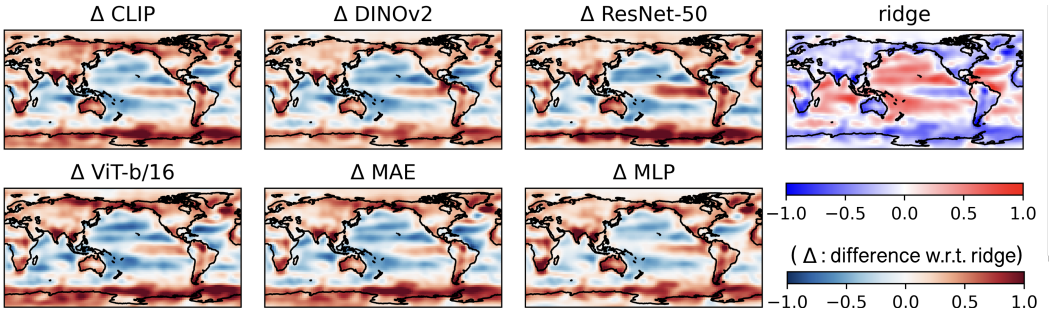

Figure 2: Visualization of Integrated Gradients (IG) times Input for the tas-huss-pr_mr experiment, highlighting regions influencing the prediction of AGMT. IG×Input values were calculated for 26k samples from the ClimDetect test set, where AGMT falls within [1.5, 2.5). These values were averaged, smoothed using a Gaussian filter, and then normalized by the maximum IG×Input value. Differences (Δ) with respect to the ridge regression model are displayed for all models except ridge. Appendix A.4 includes IG×Input visualizations for other experiments.

4 Summary and Future Work

We demonstrated the potential of ViTs for climate change D&A tasks, noting their ability to detect climate change signals earlier than simpler models like MLP or ridge regression. While MLP and ridge regression can match the accuracy of ViTs in certain setups (Table 1), ViTs (and CNNs) appear to excel in filtering out natural variability during the pre-warming period, thereby increasing detection sensitivity (Figure 1b). Future work will further probe the physical basis of ViTs, assess consistency across various explainable AI (XAI) methods, and validate our findings with additional datasets to ensure our models' robustness is not merely an artifact of their training and evaluation data.

References

[1] IPCC. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, volume In Press. Cambridge

University Press, Cambridge, United Kingdom and New York, NY, USA, 2021.

[2] Benjamin D Santer, Wolfgang Brüggemann, Ulrich Cubasch, Klaus Hasselmann, Heinke Höck, Ernst Maier-Reimer, and Uwe Mikolajewica. Signal-to-noise analysis of time-dependent greenhouse warming experiments. *Climate Dynamics*, 9(6):267–285, 1994.

[3] B. D. Santer, C. Mears, F. J. Wentz, K. E. Taylor, P. J. Gleckler, T. M. L. Wigley, T. P. Barnett, J. S. Boyle, W. Brüggemann, N. P. Gillett, S. A. Klein, G. A. Meehl, T. Nozawa, D. W. Pierce, P. A. Stott, W. M. Washington, and M. F. Wehner. Identification of human-induced changes in atmospheric moisture content. *Proceedings of the National Academy of Sciences*, 104(39):15248–15253, 2007.

[4] Benjamin D. Santer, Stephen Po-Chedley, Mark D. Zelinka, Ivana Cvijanovic, Céline Bonfils, Paul J. Durack, Qiang Fu, Jeffrey Kiehl, Carl Mears, Jeffrey Painter, Giuliana Pallotta, Susan Solomon, Frank J. Wentz, and Cheng-Zhi Zou. Human influence on the seasonal cycle of tropospheric temperature. *Science*, 361(6399), 2018.

[5] Elizabeth A. Barnes, James W. Hurrell, Imme Ebert-Uphoff, Chuck Anderson, and David Anderson. Viewing Forced Climate Patterns Through an AI Lens. *Geophysical Research Letters*, 46(22):13389–13398, 2019.

[6] Elizabeth A. Barnes, Benjamin Toms, James W. Hurrell, Imme Ebert-Uphoff, Chuck Anderson, and David Anderson. Indicator Patterns of Forced Change Learned by an Artificial Neural Network. *Journal of Advances in Modeling Earth Systems*, 12(9), 2020.

[7] Sebastian Sippel, Nicolai Meinshausen, Erich M. Fischer, Enikő Székely, and Reto Knutti. Climate change now detectable from any single day of weather at global scale. *Nature Climate Change*, 10(1):35–41, 2020.

[8] Yoo-Geun Ham, Jeong-Hwan Kim, Seung-Ki Min, Daehyun Kim, Tim Li, Axel Timmermann, and Malte F. Stuecker. Anthropogenic fingerprints in daily precipitation revealed by deep learning. *Nature*, 622(7982):301–307, 2023.

[9] Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

[10] Sungduk Yu, Brian L. White, Anahita Bhiwandiwalla, Musashi Hinck, Matthew Lyle Olson, Tung Nguyen, and Vasudev Lal. Climdetect: A benchmark dataset for climate change detection and attribution, 2024.

[11] Cornel Soci, Hans Hersbach, Adrian Simmons, Paul Poli, Bill Bell, Paul Berrisford, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Raluca Radu, Dinand Schepers, Sebastien Villaume, Leopold Haimberger, Jack Woollen, Carlo Buontempo, and Jean-Noël Thépaut. The era5 global reanalysis from 1940 to 2022. *Quarterly Journal of the Royal Meteorological Society*, 2024.

[12] Yuki Kosaka, Shinya Kobayashi, Yayoi Harada, Chiaki Kobayashi, Hiroaki Naoe, Koichi Yoshimoto, Masashi Harada, Naochika Goto, Jotaro Chiba, Kengo Miyaoka, et al. The jra-3q reanalysis. *Journal of the Meteorological Society of Japan. Ser. II*, 102(1):49–109, 2024.

[13] Ronald Gelaro, Will McCarty, Max J Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A Randles, Anton Darmenov, Michael G Bosilovich, Rolf Reichle, et al. The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of climate*, 30(14):5419–5454, 2017.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[19] C. Tebaldi, K. Debeire, V. Eyring, E. Fischer, J. Fyfe, P. Friedlingstein, R. Knutti, J. Lowe, B. O'Neill, B. Sanderson, D. van Vuuren, K. Riahi, M. Meinshausen, Z. Nicholls, K. B. Tokarska, G. Hurtt, E. Kriegler, J.-F. Lamarque, G. Meehl, R. Moss, S. E. Bauer, O. Boucher, V. Brovkin, Y.-H. Byun, M. Dix, S. Gualdi, H. Guo, J. G. John, S. Kharin, Y. Kim, T. Koshiro, L. Ma, D. Olivié, S. Panickal, F. Qiao, X. Rong, N. Rosenbloom, M. Schupfner, R. Séférian, A. Sellar, T. Semmler, X. Shi, Z. Song, C. Steger, R. Stouffer, N. Swart, K. Tachiiri, Q. Tang, H. Tatebe, A. Voldoire, E. Volodin, K. Wyser, X. Xin, S. Yang, Y. Yu, and T. Ziehn. Climate model projections from the scenario model intercomparison project (scenariomip) of cmip6. *Earth System Dynamics*, 12(1):253–293, 2021.

[20] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

A Supporting Information for "Exploring Vision Transformers for Early Detection of Climate Change Signals"

A.1 Visual Illustration of Climate Change Detection Method

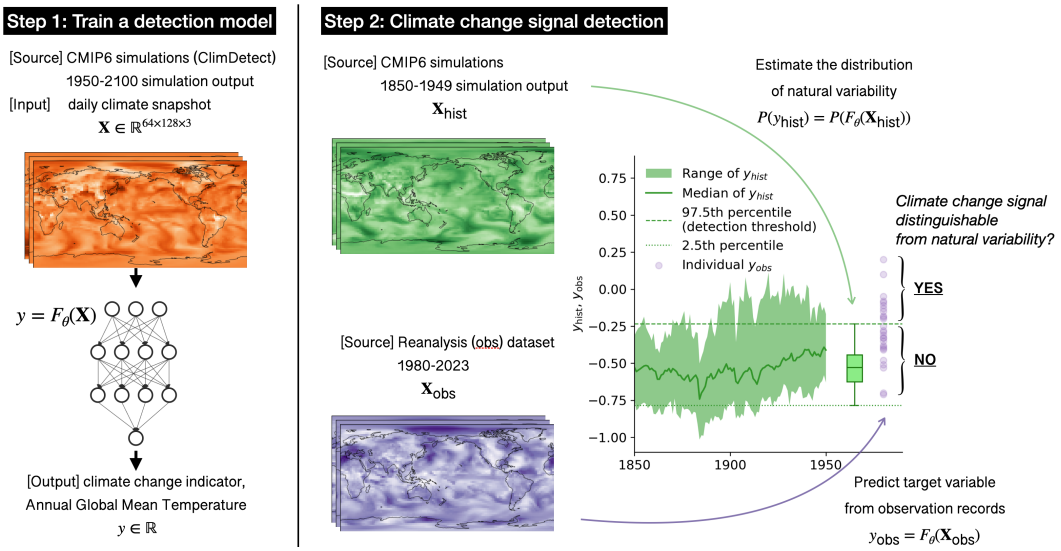

Figure S1: Illustration of our climate change signal detection methodology. The diagram features climate field maps distinguished by color to denote independent datasets: the training dataset in orange, the historical (i.e., pre-warming) dataset in green, and the observation dataset in purple. F_θ denotes a detection model (e.g., vision transformer, CNN, etc.), where θ represents the parameters of the model. One purple dot represent an individual estimates from a single observation sample. For detailed information, see Section 2.2.

A.2 Training details

Vision Transformers. We adopted four Vision Transformer (ViT) models—ViT-b/16, CLIP, MAE, and DINOv2—as described in the ClimDetect baseline models, adhering to their specified configurations and training settings. Their pretrained checkpoints were sourced from Hugging Face (checkpoint name `google/vit-base-patch16-224`, `openai/clip-vit-large-patch14-336`, `facebook/vit-mae-base`, and `facebook/dinov2-large`). Each model was finetuned with a regression head using a batch size of 512. The learning rate was set at 5e-4, with a warm-up period during the first half of an epoch followed by a fixed linear decay at 5% for the remainder of the training. The models were trained over 10 epochs using the AdamW optimizer, with all parameters being updated during training. We used the best checkpoints based on the lowest validation loss.

CNN. We chose the ResNet-50 architecutre for our CNN model. ResNet-50 was trained from a Hugging Face checkpoint (`microsoft/resnet-50`) with regression head (that is, num_labels=1). The effective batch size was 64. The learning rate was set at 1e-4 with a warm-up period over the first epoch followed by a 5% linear decay for the remaining epochs. The training was conducted over 10 epochs, and then the best checkpoints were selected based on validation loss.

MLP and Ridge Regression. A ridge regression model was fit with $\alpha = 10^6$, and a multilayer perceptron (MLP) featured five hidden layers, each with 100 units. The MLP's learning rate was set at 5e-5 with cyclic adjustments and included L2 regularization set at $\alpha = 0.01$.

Training Dataset Size and Split. We utilized the ClimDetect dataset [10] for detection model training, adhering to its default data split. The dataset comprises 627,581 training samples, 80,227 validation samples, and 108,405 test samples, corresponding to a split ratio of approximately 77:10:13. ClimDetect provides prenormalized, training ready inputs (e.g., daily climatology removal and z-score standardization) and outputs AGMT values relative to the 1980-2014 mean. For additional details, refer to the ClimDetect documentation.

A.3 RMSE calculated on JRA-3Q and MERRA-2

	tas-huss-pr	tas	pr	huss	tas-huss-pr_mr	tas_mr
CLIP	0.1169	0.1265	**0.4453**	0.1775	0.2010	0.2276
DINOv2	0.1291	0.1235	0.4684	0.1727	0.1748	0.2424
MAE	0.1086	0.1039	0.4586	**0.1319**	**0.1579**	0.2034
ViT-b/16	0.1318	**0.0994**	0.5299	0.1634	0.1667	0.1851
ResNet-50	0.1215	0.1208	0.5829	0.1564	0.1944	**0.1706**
MLP	0.1142	0.1259	0.7178	0.1547	0.2167	0.2258
ridge	**0.1065**	0.1156	0.5237	0.1675	0.1748	0.2193

Table S1: Similar to Table 2 in the main text, but with RMSE calculated over the 1980-2023 period using JRA-3Q data.

	tas-huss-pr	tas	pr	huss	tas-huss-pr_mr	tas_mr
CLIP	0.1284	0.1576	**0.5223**	0.2019	0.1685	0.2364
DINOv2	0.1327	0.1733	0.5689	0.2009	0.1709	0.2420
MAE	**0.1136**	0.1357	0.6032	0.1747	0.1453	0.1897
ViT-b/16	0.1153	**0.1159**	0.7986	0.1769	**0.1411**	**0.1783**
ResNet-50	0.1211	0.1223	0.5270	**0.1448**	0.1784	0.2384
MLP	0.1311	0.1341	0.6757	0.1757	0.2924	0.2723
ridge	0.1256	0.1262	0.6170	0.1782	0.2546	0.2552

Table S2: Similar to Table 2 in the main text, but with RMSE calculated over the 1980-2023 period using MERRA-2 data.

A.4 Year of Emergence

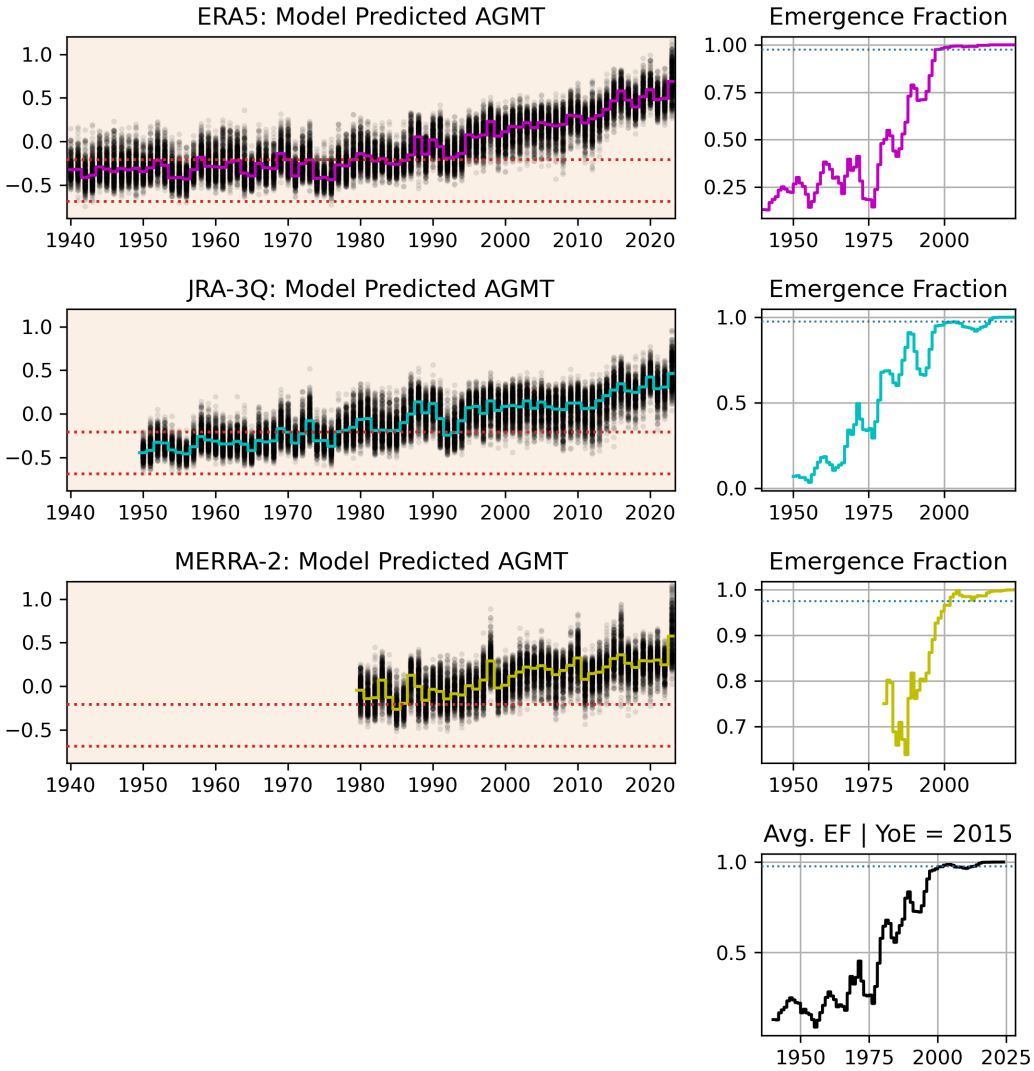

Figure S2: Detection model: **ViT-b/16**; Experiment: "tas_mr". (Left) Model-predicted test statistic, AGMT, from three different reanalysis datasets, displayed as 365 black dots per year with their mean represented by the colored line. The red lines indicate the 2.5th to 97.5th percentile range of natural variability for the test statistics, which was estimated from the 1850-1949 CMIP6 model simulation output. (Right) Emergence fraction (EF) per year, defined as the fraction of days where predicted AGMT exceeds the upper bound (the 97.5th percentile of natural variability) within one year. Centered 5-year window moving averaging is applied to EF time series. (Bottom Right) The black line represents the average of the three colored lines shown in the upper panels. The Year of Emergence (YOE) is calculated from this average, defined as the first year where the averaged EF surpasses the 97.5% threshold (blue line), corresponding to 356 days of the year.

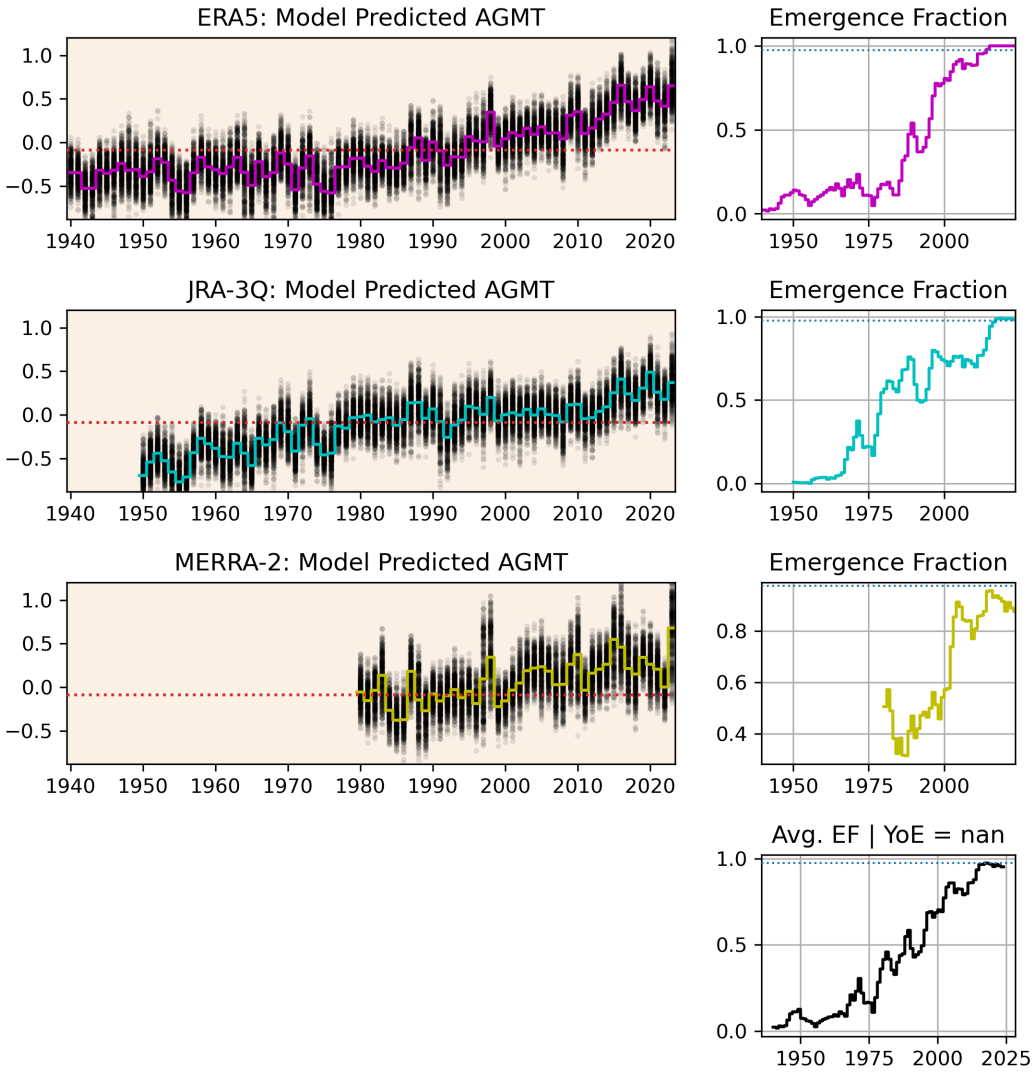

Figure S3: Similar to Figure S2 but **Ridge regression** is used as a climate change detection model.

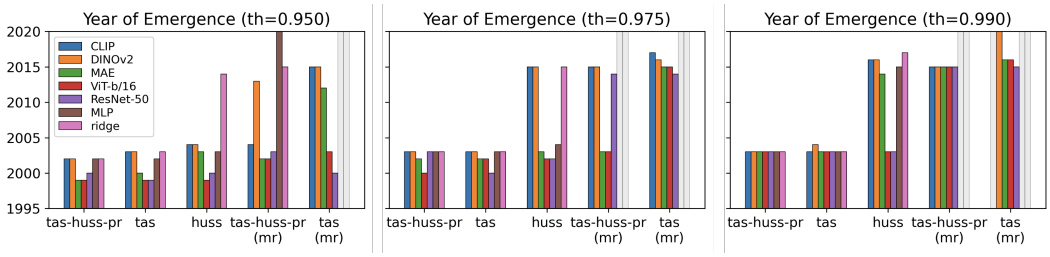

Figure S4: Similar to Figure 1a, but with three different emergence fraction threshold: (left) 0.95, (middle) 0.975, and (right) 0.99.

A.5 Integrated Gradients maps

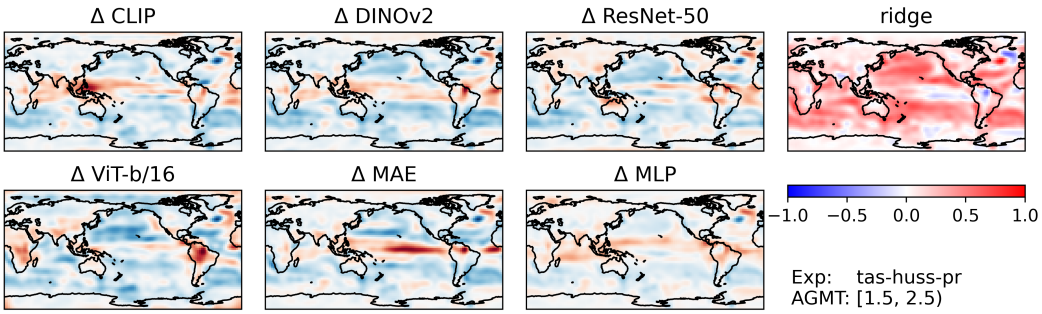

Figure S5: Similar to Figure 2, except for the **tas-huss-pr** experiment

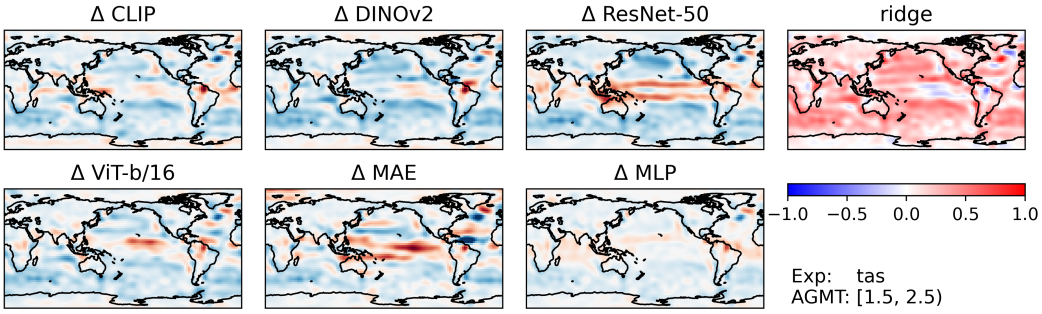

Figure S6: Similar to Figure 2, except for the **tas** experiment

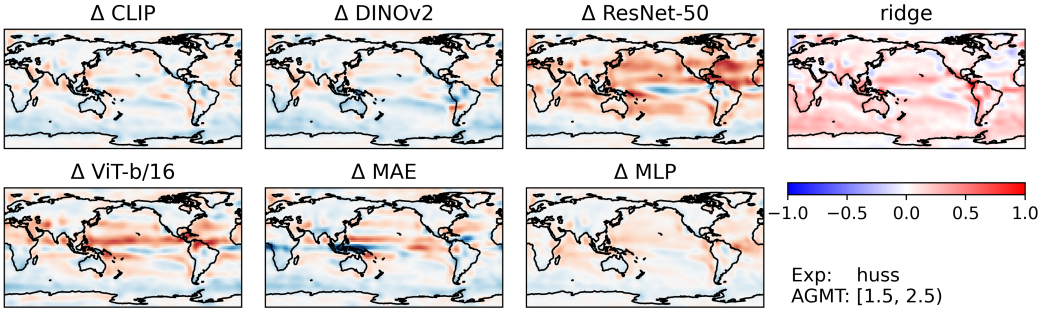

Figure S7: Similar to Figure 2, except for the **huss** experiment

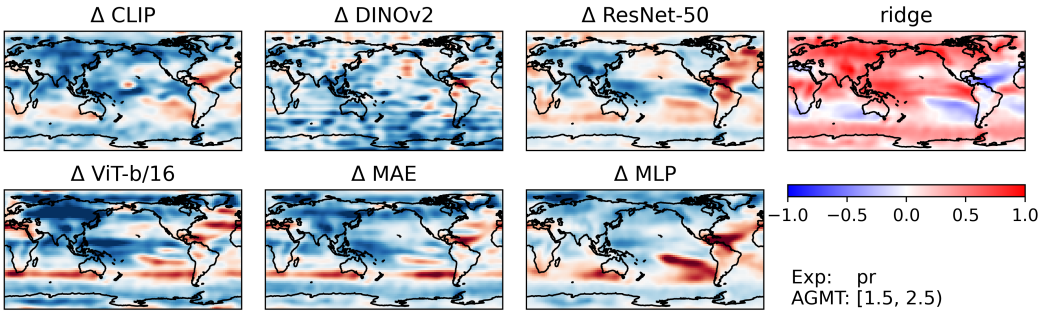

Figure S8: Similar to Figure 2, except for the **pr** experiment

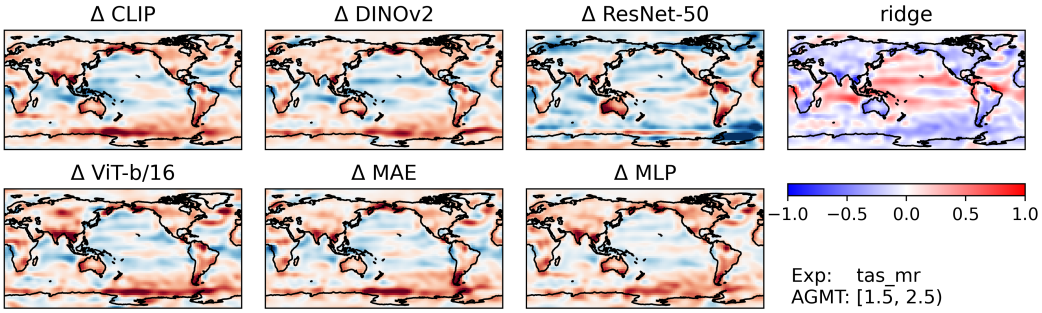

Figure S9: Similar to Figure 2, except for the **tas_mr** experiment