

# Forecasting the Future with Yesterday's Climate: Temperature Bias in AI Weather and Climate Models

Jacob B. Landsberg<sup>1,2</sup>, Elizabeth A. Barnes<sup>1,2,3</sup>

<sup>1</sup>Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA

<sup>2</sup>Faculty of Computing and Data Sciences, Boston University, Boston, MA, USA

<sup>3</sup>Department of Earth and Environment, Boston University, Boston, MA, USA

## Key Points:

- AI weather and AI climate models produce cold-biased boreal winter land temperatures that resemble those from 15-20 years earlier.
- The weather model cold bias is strongest for the hottest temperatures, suggestive of limited training exposure to modern extreme heat.
- The climate model cold bias is largest in regions, seasons, and parts of the temperature distribution where climate change has been largest.

---

Corresponding author: Jacob B. Landsberg, [jlandsbe@bu.edu](mailto:jlandsbe@bu.edu)

**Abstract**

AI-based climate and weather models provide fast, skillful forecasts yet face a key challenge: predicting future climates while being trained with historical data. We investigate this issue by analyzing boreal winter land temperature biases in AI weather (FourCastNet V2 Small and Pangu Weather) and climate (Ai2 Climate Emulator version 2) models. We evaluate these models during time periods that are significantly more recent than the bulk of their training data, allowing us to assess how well they generalize to more modern, conditions. We find that all models produce cold-biased mean temperatures, resembling climates from 15–20 years earlier than their prediction period. Furthermore, FourCastNet’s and Pangu’s cold bias is strongest for the hottest predicted temperatures, indicating limited training exposure to modern extreme heat events. In contrast, ACE2’s bias is more evenly distributed but largest in regions, seasons, and parts of the temperature distribution where historic global warming is most pronounced.

**Plain Language Summary**

AI-based climate and weather models, which learn from historical data, can struggle to accurately predict future conditions, especially as the climate changes. We probe this issue by analyzing boreal winter land temperature biases in AI weather models (FourCastNet and Pangu) and an AI climate model (ACE2). We find that all models produce temperatures that better resemble climates from 15–20 years earlier than the period they are predicting. In some regions, like the Eastern U.S., the predictions resemble climates from as much as 20–30 years earlier. Further analysis shows that FourCastNet’s cold bias is strongest in the hottest predicted temperatures, indicating that these models may not have seen enough examples of modern extreme heat events in the past data. In contrast, ACE2’s bias is more evenly distributed but largest in regions, seasons, and parts of the temperature distribution where climate change has been most pronounced. These findings underscore the challenge of training AI models exclusively on historical data and highlight the need to account for such biases when applying them to future climate prediction.

## 1 Introduction

Over the last five years, a new generation of fully data-driven AI models has emerged, reimagining weather forecasts and exploring early applications to climate prediction (H. Zhang et al., 2025). Unlike traditional dynamical models, which are governed by physical equations, these AI models learn relationships between variables directly from large datasets (e.g., Ebert-Uphoff and Hilburn (2023); Rasp et al. (2020); Bonev et al. (2023)). This approach has largely been successful, with many recent AI models achieving state-of-the-art performance (e.g., Pathak et al. (2022); Bonev et al. (2023, 2025); Bi et al. (2023); Lam et al. (2023); Lang et al. (2024)). Furthermore, these models are much less computationally expensive than dynamical models, allowing for faster predictions and more extensive ensemble simulations (Liu et al., 2024).

One of the key challenges with fully data-driven AI models is that they are most often trained on historical data, which may not accurately represent future conditions. This can lead to biases in the model’s predictions, particularly in the context of a changing climate (Lindsey & Dahlman, 2020). For example, these models may be tasked with predicting temperatures that largely lie outside the bulk of their training distribution (Beucler et al., 2024). Rackow et al. (2024) examined this phenomenon by assessing the performance of three prominent AI weather models: Pangu Weather (Pangu) (Bi et al., 2023), Graphcast (Lam et al., 2023), and the AIFS (Lang et al., 2024) under different climate regimes. The former two models were trained on reanalysis data from 1979–2017 (Bi et al., 2023; Lam et al., 2023), while the later was trained on data from 1979–2020 (Lang et al., 2024). Rackow et al. (2024) confronted these models with a preindustrial climate (1955), a modern climate (2023), and a future, warmer climate (2049). They found that for these three different years, while the models’ biases varied, in general, they exhibited warmer biases in the preindustrial climate, slight cold biases in the modern climate, and significant cold biases in the future climate. Furthermore, Z. Zhang et al. (2025); Kent et al. (2025) both find that AI weather and AI climate models perform worse than traditional models when predicting record-breaking events. These works, therefore, underscore the challenges of training AI models on historical data.

In this study, we analyze boreal winter land temperature predictions of two AI weather models and an AI climate model with explicit CO<sub>2</sub> forcing more broadly. Specifically, we quantify the extent to which FourCastNet V2 small (FourCastNet) (Pathak et al., 2022), Pangu (Bi et al., 2023), and Ai2 Climate Emulator version 2 (ACE2) (Watt-Meyer et al., 2025) reflect their evaluation-period climate as opposed the mean climate of their training data. For FourCastNet and Pangu, we focus on predictions from 2020–2025, while for ACE2 we analyze those from 1996–2010. These periods are outside of their respective training sets and are also warmer than their training climatologies. In doing so, we assess the persistence of training-climate influence on the models’ temperature distributions. Finally, we probe when and where these biases are most pronounced across the different types of models.

## 2 Data and Models

### 2.1 FourCastNet and Pangu

FourCastNet (Bonev et al., 2023) and Pangu (Bi et al., 2023) are both fully-data-driven AI weather models designed by NVIDIA and Huawei, respectively. Pangu is trained with ECMWF Reanalysis v5 (ERA5) data (Hersbach et al., 2020) from 1979–2017 and a transformer architecture, while FourCastNet utilizes ERA5 data from 1979–2015 and a Spherical Fourier Neural Operator (SFNO) architecture. Thus, the training data for FourCastNet and Pangu are centered around 1997 and 1998, respectively. We use both models’ 2m temperature (2mT) outputs generated by Radford et al. (2025) for 2-day and 9-day leads. Each forecast is initialized with 0000 UTC NOAA Global Forecasting Sys-

tem (GFS) data (NOAA Office of Satellite and Product Operations, 2020) from 2 or 9 days prior to the forecast date. Predictions are rolled out in 6-hour timesteps during the December-January-February (DJF) period between December 2020 and February 2025. As the training sets for FourCastNet and Pangu are centered around the turn of the century, this timeframe is not only more modern than any training year, it is also  $\sim 25$  years more modern than the average training year. We compute the daily average of these 6-hour forecasts to obtain daily mean temperature forecasts. Both models are missing a small number of initialization dates, although this represents less than 0.7% of the total data for FourCastNet and less than 0.9% of the data for Pangu. For more details see Appendix A.

## 2.2 ACE2

ACE2 is an atmosphere-only AI climate model designed to produce stable  $\sim 100$ -year simulations of Earth’s climate, capturing atmospheric variability, global temperature trends, and tropical phenomena like the Madden Julian Oscillation and El Niño Southern Oscillation (Watt-Meyer et al., 2025). ACE2 similarly uses an SFNO architecture, a 6-hour autoregressive structure, and is trained on ERA5 data from 1940–1995, 2011–2019, and 2021–2022. Unlike FourCastNet, ACE2 includes specific CO<sub>2</sub> forcing from the Climate Model Intercomparison Project – Phase 6 (Meinshausen et al., 2017) and NOAA Global Monitoring Laboratory (Conway et al., 1994). We generate a 5-member ensemble of daily surface temperatures from 1940–2020. The first ensemble member is generated by running the model with January 1940 initial conditions from ERA5. For each subsequent ensemble member, we initialize the model using the day-1 forecast from the previous ensemble member as its initial condition. All ensemble members use the same sea surface temperature and CO<sub>2</sub> forcings and differ only in their initializations. We then compute the daily mean temperature for each ensemble member, and extract data between 1996–2010. While none of these years are within ACE2’s training data, we do include 1996–2000, which are in the validation set. We find, including or excluding these years has minimal impact on the global bias ( $-.33$  K compared to  $-.35$  K). We chose this earlier period, rather than the 2020–2025 time range we use for the weather models, as ACE2 includes training data up through 2022. Nonetheless, as with the weather models, 1996–2010 is still 25–30 years more modern than the average training year ( $\sim 1975$ ).

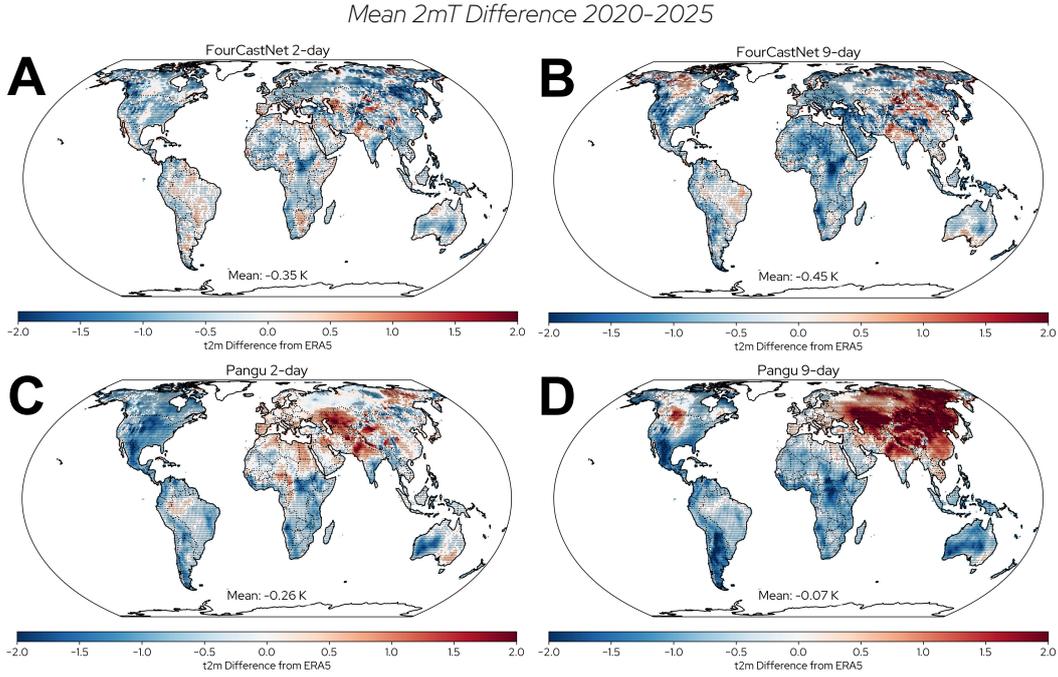
## 2.3 Temperature Comparison

To evaluate the daily temperature biases of the models, we subset the data to boreal winter land temperatures, which we define to be data in the DJF period and all land gridpoints except those in Antarctica and Greenland. We focus on boreal winter, as both cold extremes in the Northern Hemisphere and land temperatures in particular have been shown to be warming more rapidly than the global average (Gross et al., 2020; Crimmins et al., 2023). We analyze the biases of these models by comparing their predictions to daily ERA5 formed by sampling at 6-hour frequency—the dataset with which all three models were trained. We use ERA5 data at  $0.25^\circ$  resolution, which we then coarsen to  $1^\circ$ , when comparing with FourCastNet and Pangu and directly at  $1^\circ$  resolution when comparing to ACE2, as those are the native resolutions of the respective AI models. To compute bias, we take the time-mean difference at every grid point between ERA5 and the model predictions. Global mean biases are then reported as the cosine-latitude-weighted average of these gridpoint biases. To compute the spatial significance of these biases, we use a two-sided Z-test. We compute the population variance by subtracting the seasonal mean temperature from ERA5 and computing variance from 1980–2025 for the weather models and from 1940–2022 for ACE2. We perform this test with the null hypothesis that the mean bias is equal to zero at an  $\alpha = 0.1$ . We use the lag-1 autocorrelation to estimate the effective sample size when computing the Z-statistic for all models (Equation 6 in Santer et al. (2000)).

### 3 Results

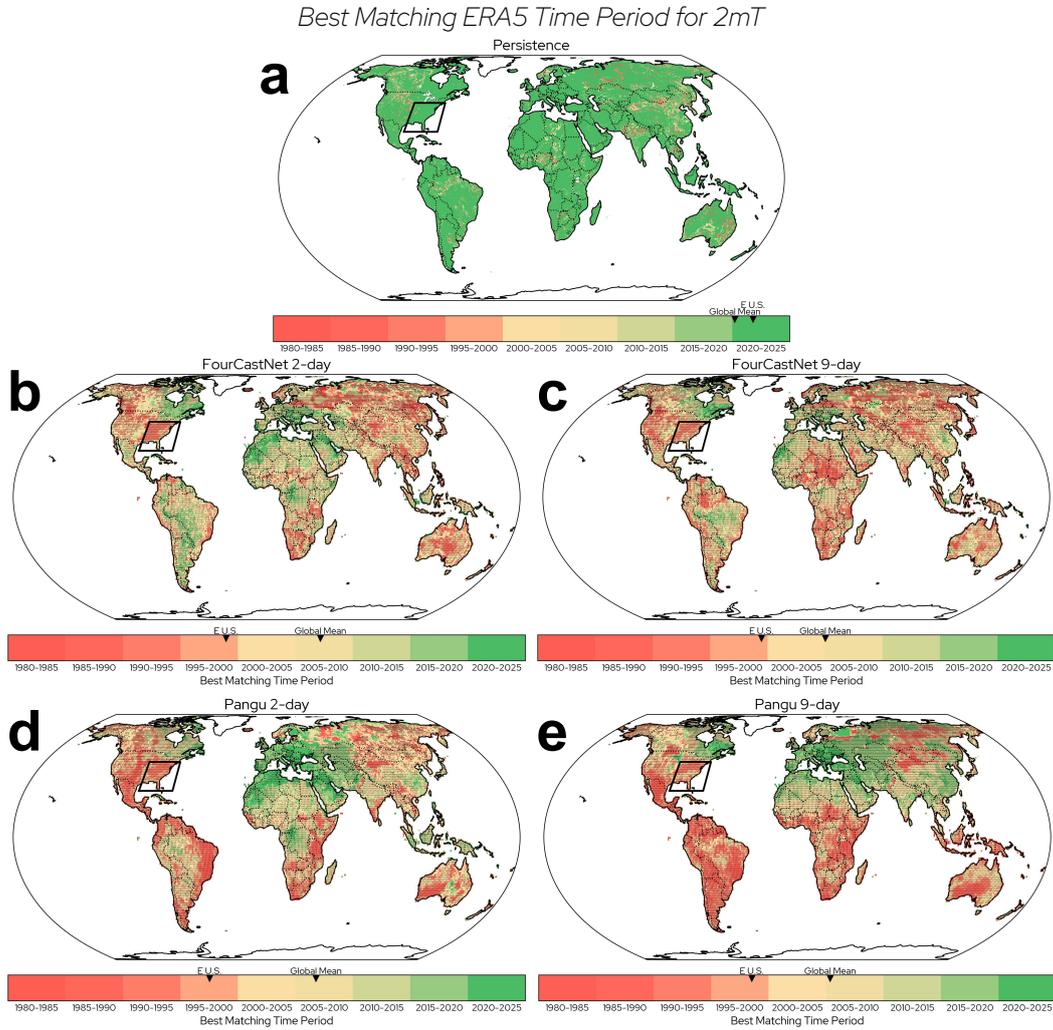
#### 3.1 Weather Models

We find that at 2-day and 9-day lead times, FourCastNet and Pangu both produce forecasts of 2020-2025 boreal winter land temperatures that are too cold relative to ERA5 (Figure 1), a pattern that holds across lead times (Figure S1). While both models are cold, FourCastNet is colder than Pangu with global mean differences of  $-0.35$  K and  $-0.45$  K compared to  $-0.26$  K and  $-0.07$  K for 2-day and 9-day leads respectively. Moreover, this cold bias is distributed nearly globally, with the exception of Asia in Pangu’s 9-day lead forecasts.

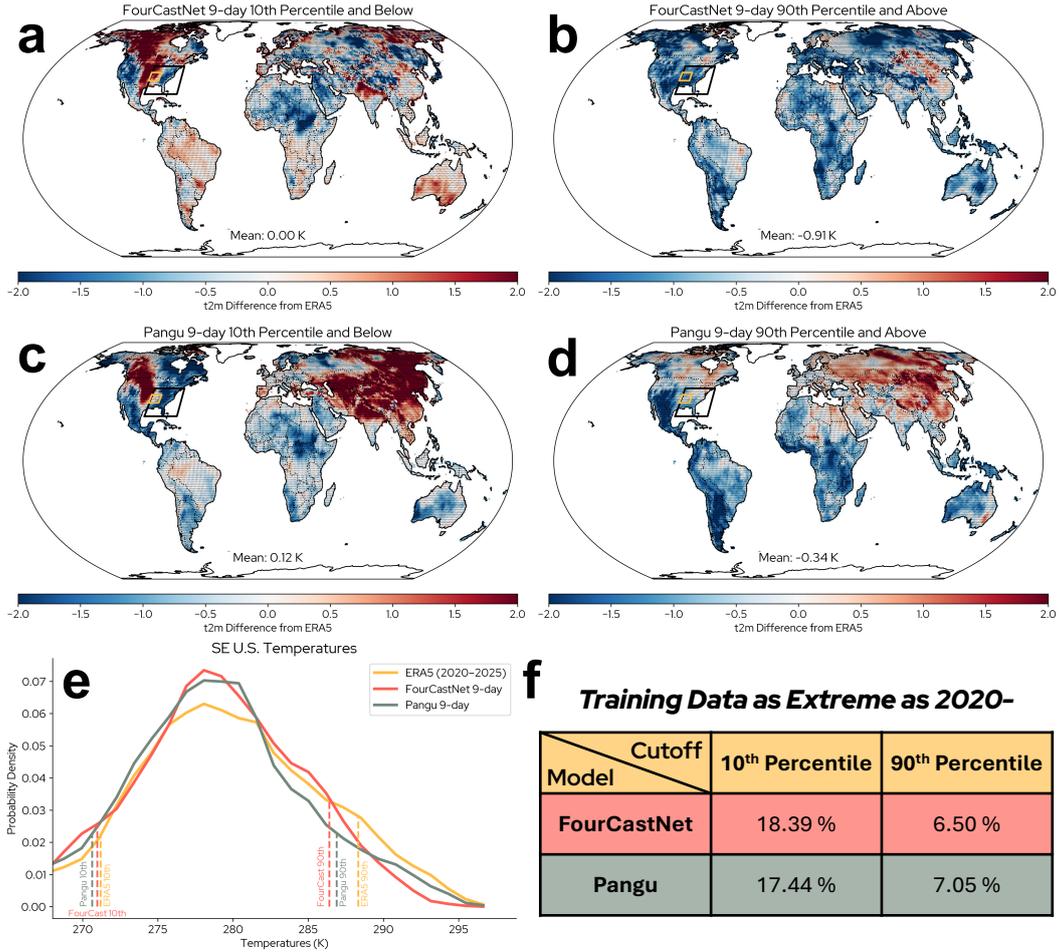


**Figure 1.** Mean 2mT differences for 2020-2025 boreal winter land temperatures compared to ERA5 for (a) FourCastNet 2-day lead, (b) FourCastNet 9-day lead, (c) Pangu 2-day lead, and (d) Pangu 9-day lead. Global means are shown at the bottom of each panel, with stippling indicating statistically significant non-zero bias (see Methods).

In fact, the temperatures generated by FourCastNet and Pangu for 2020-2025 more closely resemble temperatures from 15-20 years earlier (Figure 2b-e). For some regions, like the Eastern U.S. ( $25^{\circ}\text{N}$  to  $42^{\circ}\text{N}$  and  $70^{\circ}\text{W}$  to  $95^{\circ}\text{W}$ ), this bias is even more pronounced, with the model’s forecasts most closely resembling ERA5 temperatures from 20-25 years earlier. This suggests these models may be struggling to fully generalize to 2020-2025 which lies beyond their training data’s climate, which is centered about 25 years prior. We compare these models’ lagging climate to that generated by a 9-day persistence forecast (Figure 2a). We find that a persistence forecast shows essentially matching mean temperatures to the prediction period of 2020-2025. While this is perhaps unsurprising, as most of the ERA5 persistence forecasts are the same data as the ground truth ERA5 data, it highlights that a much simpler prediction model can offer a more temporally consistent mean climate than FourCastNet and Pangu.



**Figure 2.** The closest matching 5-year span of ERA5 land temperatures to FourCastNet and Pangu’s 9-day lead forecasts of 2020-2025 boreal winter land temperatures for a) a 9-day persistence forecast, b) FourCastNet 2-day prediction, c) FourCastNet 9-day prediction, d) Pangu 2-day prediction, and e) Pangu 9-day prediction. The Eastern U.S. (highlighted by the black box) and global mean time period are shown in the legend. Stippling indicates grid points that have statistically significant non-zero bias.

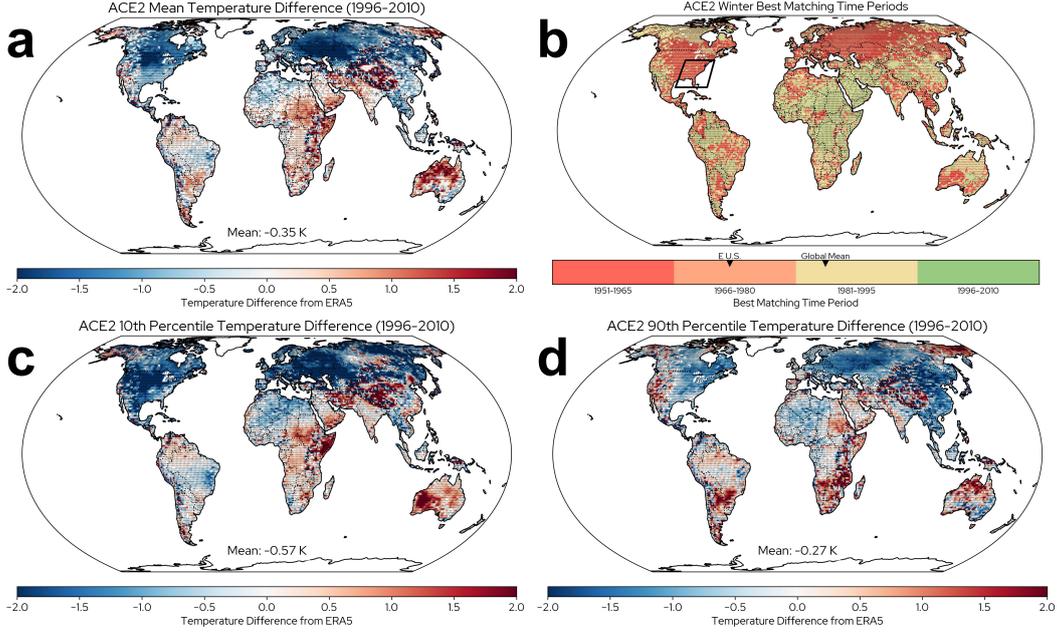


**Figure 3.** Mean 2mT differences as in Figure 1 but for the 10th and 90th percentiles of FourCastNet’s (a, b) and Pangu’s (c,d) 9-day lead forecasts, with stippling indicating statistically significant non-zero bias. An example of the tail behavior for the SE U.S. (bounded by the yellow box in a-d) is shown in e. The global mean percent of training data as or more extreme than the 10th and 90th percentiles of 2020–2025 ERA5 temperatures is displayed in f.

### 3.2 Extreme Modern Temperatures

We further investigate where this difference in modern temperatures is most pronounced by looking at the tails of the temperature distribution for the 9-day lead time forecasts. We find that the hottest temperature forecasts for both Pangu and FourCastNet exhibit a much stronger cold bias than those for the coldest temperatures. For instance, the hottest 10% of 2020–2025 temperature forecasts are on average 0.91 K colder than ERA5 for FourCastNet and 0.34 K colder for Pangu (Figures 3b and 3d), while the coldest 10% of FourCastNet’s temperatures exhibit minimal bias compared to ERA5 and Pangu’s are even 0.12 K warmer (Figures 3a and 3c). An example of the temperature distributions’ tail behavior for the SE U.S. (30°–35° N, 90°–100° W) is shown in 3e, with Pangu and FourCastNet both matching much more closely with ERA5 for the cold tail than the hot tail.

This stark difference in bias between the coldest and hottest temperatures may be a reflection of the models’ training data, which is primarily from a colder climate. For

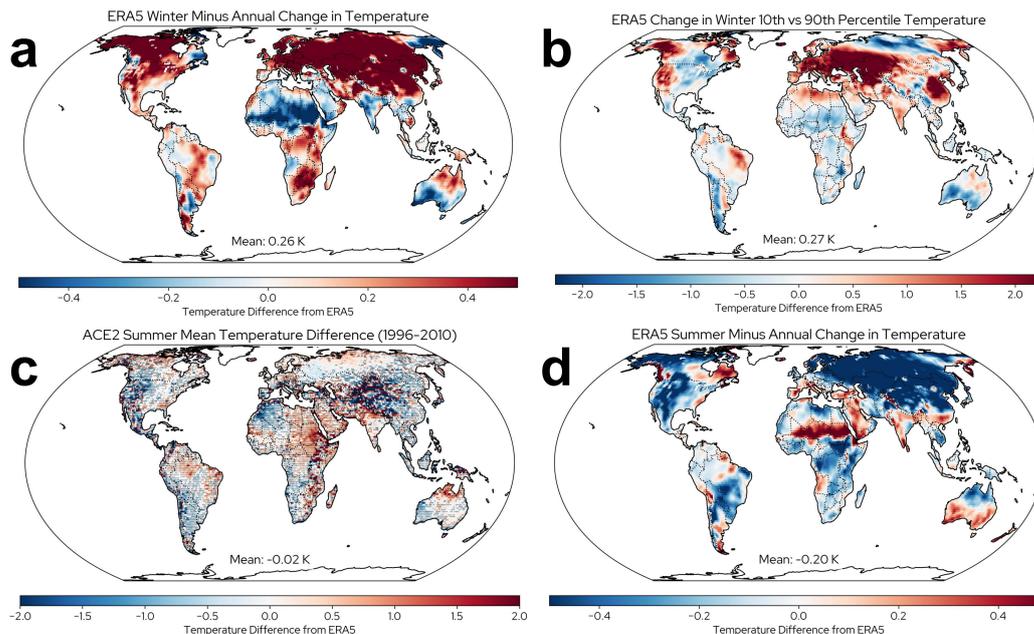


**Figure 4.** a) Mean surface temperature differences for 1996-2010 boreal winter land temperatures compared to ERA5 for ACE2. b) The closest matching 15-year span of ERA5 land temperatures to ACE2’s 1996-2010 boreal winter land temperatures. The Eastern U.S. (highlighted by the black box) and global mean time period are shown in the legend. c) Mean surface temperature differences as in (A) but for the 10th percentile of ACE2’s 1996-2010 predictions. d) Mean surface temperature differences as in (a) but for the 90th percentile of ACE2’s 1996-2010 predictions. Global means are shown at the bottom of (a), (c), and (d), while all figures show stippling at grid points of statistically significant non-zero bias.

instance, globally there is  $\sim 2\text{-}3\times$  as much training data that is as cold or colder than the 10th percentile of 2020-2025 ERA5 temperatures than there is training data that is as hot or hotter than the 90th percentile of ERA5 temperatures (Figures 3f and S2). Hence, this cold bias during the hotter forecasts is likely a pull toward the mean of the training dataset. These findings hold for various percentile thresholds used to define cold and hot tails (Figures S3 and S4).

### 3.3 Climate Models

We similarly analyze ACE2 temperatures, investigating how a climate model with an SFNO architecture, like FourCastNet, but adapted for long-term climate prediction through the inclusion of  $\text{CO}_2$  forcing, performs. We find that ACE2 is also too cold, with a global mean of  $-.35\text{ K}$  relative to ERA5, and is particularly cold over North America, Europe, and Russia (Figure 4a). This cold bias represents 35% of the average root mean square error compared to ERA5 over the five ensemble members. As with the weather models, this connotes a temperature pattern more similar to that of 15-20 years prior, with some regions, like the Eastern U.S. lagging  $\sim 30$  years behind (Figure 4b). This again is consistent with a pull towards the mean climate of the training data, which is centered around 1975. However, unlike FourCastNet and Pangu, which showed a stronger cold bias for their hottest forecasts, ACE2 exhibits a larger cold bias for its coldest predictions (Figures 4c and 4d).



**Figure 5.** a) Change boreal winter land surface temperatures between 1940-1979 and 1980-2022 relative to the annual mean change. b) Change in the 10th vs. 90th percentile of boreal winter land surface temperatures between 1940-1979 and 1980-2022. c) Mean surface temperature differences for 1996-2010 boreal summer land temperatures compared to ERA5 for ACE2. Stippling indicates statistically significant non-zero bias. d) Change in boreal summer land surface temperatures between 1940-1979 and 1980-2022. Global means are shown at the bottom of each panel.

We further analyze this asymmetry by situating ACE2’s biases in the context of climate change. To estimate the effect of climatic warming, we compute the difference between temperatures from 1980–2022 and those from 1940–1979. In line with previous work, (e.g. Gross et al. (2020)), we find winter temperatures have warmed more rapidly than the annual mean, particularly over North America, Europe, and Russia (Figure 5a)—the same regions where ACE2 is most cold-biased (Figure 4a). Similarly, when we look at the change in the 90th percentile of winter temperatures compared to the 10th percentile, we find that the coldest winter temperatures have warmed more rapidly than the hottest temperatures over much of the Northern Hemisphere (Figure 5b). Again, this mirrors the stronger bias ACE2 shows in the cold tail (Figures 4c-d). Notably, the weather models, which show the opposite tail behavior to ACE2, have training periods that lack a similar asymmetric climatic warming trend (Figure S5). This pattern is consistent across seasons as well; for example, in boreal summer, ACE2 exhibits lower bias (Figure 5c), consistent with the fact that summer temperatures have warmed less rapidly than both winter and the annual mean (Figure 5d). This shows that ACE2’s bias is largest in regions, seasons, and parts of the temperature distribution where climate change has been most pronounced.

#### 4 Discussion and Conclusions

In this work we have shown that both AI weather and climate models exhibit cold biases when predicting modern climates that lie outside of the bulk of their training data (Figures 1 and 4). Instead, their boreal winter land temperatures better resemble those

of 15-20 years prior (Figures 2 and 4b). This is consistent with a pull toward the mean of their training data, as all models have training data centered  $\sim 25 - 30$  years prior to their prediction time period.

We did, however, find that the tails of the AI weather and AI climate temperature distributions displayed different behavior. FourCastNet and Pangu exhibited a cold bias almost exclusively for the hottest temperature predictions (Figure 1), which may be due to a lack of training data for modern extreme heat events (Figures 3f and S2). This finding is aligned with Z. Zhang et al. (2025), who found that AI weather models performed poorly when predicting record-breaking (i.e., outside of the training set) extremes. ACE2, on the other hand, exhibited a more pronounced cold bias for the coldest temperature predictions (Figures 4c-d). We attribute this asymmetric bias to the pattern of warming temperatures under climate change in ACE2’s training set. Moreover, we found a much more pronounced winter bias than summer bias, a seasonal pattern that may be more difficult to see when looking at annual means (Watt-Meyer et al., 2025). We show that these spatial, seasonal, and distributional patterns of ACE2’s bias align well with regions of rapid historical warming (Figure 5). Thus, although AI weather and AI climate models have different bias patterns in the tails of their temperature distributions, both are consistent with an anchoring to their training sets. While continued work is needed to fully understand the training mechanisms behind these biases, our findings highlight that simply including CO<sub>2</sub> forcing in an AI model (e.g., as in ACE2) is not sufficient to fully eliminate training-set artifacts.

Our work contributes to the growing body of literature documenting the limitations of AI models in extrapolating to climates outside their training domain (Rackow et al., 2024; Z. Zhang et al., 2025; Kent et al., 2025; Hernanz et al., 2022). We show that biases in both AI weather and climate models are already evident in present-day predictions, not only in future climates, and that these biases vary across space, season, and the temperature distribution. Several strategies have been proposed to mitigate such biases, including augmenting training data with climate model simulations that extend into the future (Clark et al., 2022) or transforming inputs to be “climate invariant” (Beucler et al., 2024). Advancing this focus on developing more climate-robust AI models is critical. Since many AI models already achieve skill comparable to traditional approaches (e.g., Pathak et al. (2022); Bonev et al. (2023, 2025); Bi et al. (2023); Lam et al. (2023); Lang et al. (2024)), addressing these biases will further strengthen their value for predicting both present and future climate.

## Appendix A Missing Data

Three initialization dates are missing from Radford et al. (2025) FourCastNet’s run: December 4th, 2021, December 1st, 2024, and January 22nd, 2025. Hence, for 2-day lead forecasts, December 6th, 2021; December 3rd, 2024; January 24th, 2025 and for 9-day lead forecasts, December 13th, 2021, December 10th, 2024, and January 31st, 2025 are excluded. These three missing dates represents only 0.67% of the total daily data we utilize from ERA5. Similarly, Pangu is missing: December 4th, 2021, December 1st, 2024, December 11th, 2024, and January 2nd, 2025. These four missing dates represent 0.89% of the total daily data we utilize from ERA5.

## **Open Research Section**

ERA5 data are available from (Hersbach et al., 2020) at <https://doi.org/10.24381/cds.4991cf48>. FourCastNet and Pangu data are available from Radford et al. (2025). The ACE2 model checkpoint and forcings are available from Watt-Meyer et al. (2025) at <https://doi.org/10.57967/hf/5377>. All code and initialization files for ACE are available at [https://github.com/jlandsbe/AI\\_Bias](https://github.com/jlandsbe/AI_Bias), which will be given a permanent DOI via Zenodo upon publication.

## **Conflict of Interest declaration**

The authors declare there are no conflicts of interest for this manuscript.

## **Acknowledgments**

This work is funded, in part, by NOAA grant #NA22OAR4310621 and Heising-Simons Foundation grant #2023-4720. The authors would also like to thank the two anonymous reviewers for their helpful comments and suggestions.

## References

- Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., . . . Pritchard, M. (2024). Climate-invariant machine learning. *Science Advances*, *10*(6), eadj7250. Retrieved from <https://www.science.org/doi/abs/10.1126/sciadv.adj7250> doi: 10.1126/sciadv.adj7250
- Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023, Jul 01). Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, *619*(7970), 533-538. Retrieved from <https://doi.org/10.1038/s41586-023-06185-3> doi: 10.1038/s41586-023-06185-3
- Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., Kashinath, K., & Anandkumar, A. (2023). *Spherical fourier neural operators: Learning stable dynamics on the sphere*. Retrieved from <https://arxiv.org/abs/2306.03838>
- Bonev, B., Kurth, T., Mahesh, A., Bisson, M., Kossaifi, J., Kashinath, K., . . . Keller, A. (2025). *Fourcastnet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale*. Retrieved from <https://arxiv.org/abs/2507.12144>
- Clark, S. K., Brenowitz, N. D., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., . . . Harris, L. M. (2022). Correcting a 200 km resolution climate model in multiple climates by machine learning from 25 km resolution simulations. *Journal of Advances in Modeling Earth Systems*, *14*(9), e2022MS003219. Retrieved from <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003219> (e2022MS003219 2022MS003219) doi: <https://doi.org/10.1029/2022MS003219>
- Conway, T. J., Tans, P. P., Waterman, L. S., Thoning, K. W., Kitzis, D. R., Masarie, K. A., & Zhang, N. (1994). Evidence for interannual variability of the carbon cycle from the national oceanic and atmospheric administration/climate monitoring and diagnostics laboratory global air sampling network. *Journal of Geophysical Research: Atmospheres*, *99*(D11), 22831-22855. Retrieved from <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JD01951> doi: <https://doi.org/10.1029/94JD01951>
- Crimmins, A. R., Avery, C. W., Easterling, D. R., Kunkel, K. E., Stewart, B. C., & Maycock, T. K. (2023). Fifth national climate assessment.
- Ebert-Uphoff, I., & Hilburn, K. (2023, July). The outlook for ai weather prediction. *Nature*, *619*(7970), 473-474. doi: 10.1038/d41586-023-02084-9
- Gross, M. H., Donat, M. G., Alexander, L. V., & Sherwood, S. C. (2020). Amplified warming of seasonal cold extremes relative to the mean in the northern hemisphere extratropics. *Earth System Dynamics*, *11*(1), 97-111. Retrieved from <https://esd.copernicus.org/articles/11/97/2020/> doi: 10.5194/esd-11-97-2020
- Hernanz, A., García-Valero, J. A., Domínguez, M., & Rodríguez-Camino, E. (2022). A critical view on the suitability of machine learning techniques to down-scale climate change projections: Illustration for temperature with a toy experiment. *Atmospheric Science Letters*, *23*(6), e1087. Retrieved from <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/asl.1087> doi: <https://doi.org/10.1002/asl.1087>
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., . . . Thépaut, J.-N. (2020, 2025/01/20). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049. Retrieved from <https://doi.org/10.1002/qj.3803> doi: <https://doi.org/10.1002/qj.3803>
- Kent, C., Scaife, A. A., Dunstone, N. J., Smith, D., Hardiman, S. C., Dunstan, T., & Watt-Meyer, O. (2025, Aug 25). Skilful global seasonal predictions from a machine learning weather model trained on reanalysis data. *npj Climate and Atmospheric Science*, *8*(1), 314. Retrieved from <https://doi.org/10.1038/s41612-025-01198-3> doi: 10.1038/s41612-025-01198-3

- Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416-1421. Retrieved from <https://www.science.org/doi/abs/10.1126/science.adi2336> doi: 10.1126/science.adi2336
- Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., ... Rabier, F. (2024). *Aifs - ecmwf's data-driven forecasting system*. Retrieved from <https://arxiv.org/abs/2406.01465>
- Lindsey, R., & Dahlman, L. (2020). Climate change: Global temperature. *Climate.gov*, *16*, 1-5.
- Liu, C.-C., Hsu, K., Peng, M. S., Chen, D.-S., Chang, P.-L., Hsiao, L.-F., ... Kuo, H.-C. (2024, Sep 28). Evaluation of five global ai models for predicting weather in eastern asia and western pacific. *npj Climate and Atmospheric Science*, *7*(1), 221. Retrieved from <https://doi.org/10.1038/s41612-024-00769-0> doi: 10.1038/s41612-024-00769-0
- Meinshausen, M., Vogel, E., Nauels, A., Lorbacher, K., Meinshausen, N., Etheridge, D. M., ... Weiss, R. (2017). Historical greenhouse gas concentrations for climate modelling (cmip6). *Geoscientific Model Development*, *10*(5), 2057-2116. Retrieved from <https://gmd.copernicus.org/articles/10/2057/2017/> doi: 10.5194/gmd-10-2057-2017
- NOAA Office of Satellite and Product Operations. (2020). *NOAA Global Forecast System (GFS) on AWS*. <https://registry.opendata.aws/noaa-gfs-bdp-pds/>. (Accessed: 2025-08-05)
- Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., ... Anandkumar, A. (2022). *Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators*. Retrieved from <https://arxiv.org/abs/2202.11214>
- Rackow, T., Koldunov, N., Lessig, C., Sandu, I., Alexe, M., Chantry, M., ... Jung, T. (2024). *Robustness of ai-based weather forecasts in a changing climate*. Retrieved from <https://arxiv.org/abs/2409.18529>
- Radford, J. T., Ebert-Uphoff, I., Stewart, J. Q., Musgrave, K. D., DeMaria, R., Tourville, N., & Hilburn, K. (2025). Accelerating community-wide evaluation of ai models for global weather prediction by facilitating access to model output. *Bulletin of the American Meteorological Society*, *106*(1), E68 - E76. Retrieved from <https://journals.ametsoc.org/view/journals/bams/106/1/BAMS-D-24-0057.1.xml> doi: 10.1175/BAMS-D-24-0057.1
- Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002203. Retrieved from <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002203> (e2020MS002203 10.1029/2020MS002203) doi: <https://doi.org/10.1029/2020MS002203>
- Santer, B. D., Wigley, T. M. L., Boyle, J. S., Gaffen, D. J., Hnilo, J. J., Nychka, D., ... Taylor, K. E. (2000). Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. *Journal of Geophysical Research: Atmospheres*, *105*(D6), 7337-7356. Retrieved from <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1999JD901105> doi: <https://doi.org/10.1029/1999JD901105>
- Watt-Meyer, O., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., ... Bretherton, C. S. (2025). ACE2: accurately learning subseasonal to decadal atmospheric variability and forced responses. *npj Climate and Atmospheric Science*, *8*(1), 205. Retrieved from <https://doi.org/10.1038/s41612-025-01090-0> doi: 10.1038/s41612-025-01090-0
- Zhang, H., Liu, Y., Zhang, C., & Li, N. (2025). Machine learning methods for weather forecasting: A survey. *Atmosphere*, *16*(1). Retrieved from

<https://www.mdpi.com/2073-4433/16/1/82> doi: 10.3390/atmos16010082  
Zhang, Z., Fischer, E., Zscheischler, J., & Engelke, S. (2025). *Numerical models outperform ai weather forecasts of record-breaking extremes.* Retrieved from <https://arxiv.org/abs/2508.15724>

# Geophysical Research Letters

Supporting Information for

## *Forecasting the Future with Yesterday's Climate: Temperature Bias in AI Weather and Climate Models*

Jacob B. Landsberg<sup>1,2</sup>, Elizabeth A. Barnes<sup>1,2,3</sup>

<sup>1</sup>Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA

<sup>2</sup>Faculty of Computing and Data Sciences, Boston University, Boston, MA, USA

<sup>3</sup>Department of Earth and Environment, Boston University, Boston, MA, USA

### **Contents of this file**

Figures S1 to S5

### **Introduction**

Here we include 5 additional figures referenced in the main text. All figures are generated using the same data and methods as described in the main text.

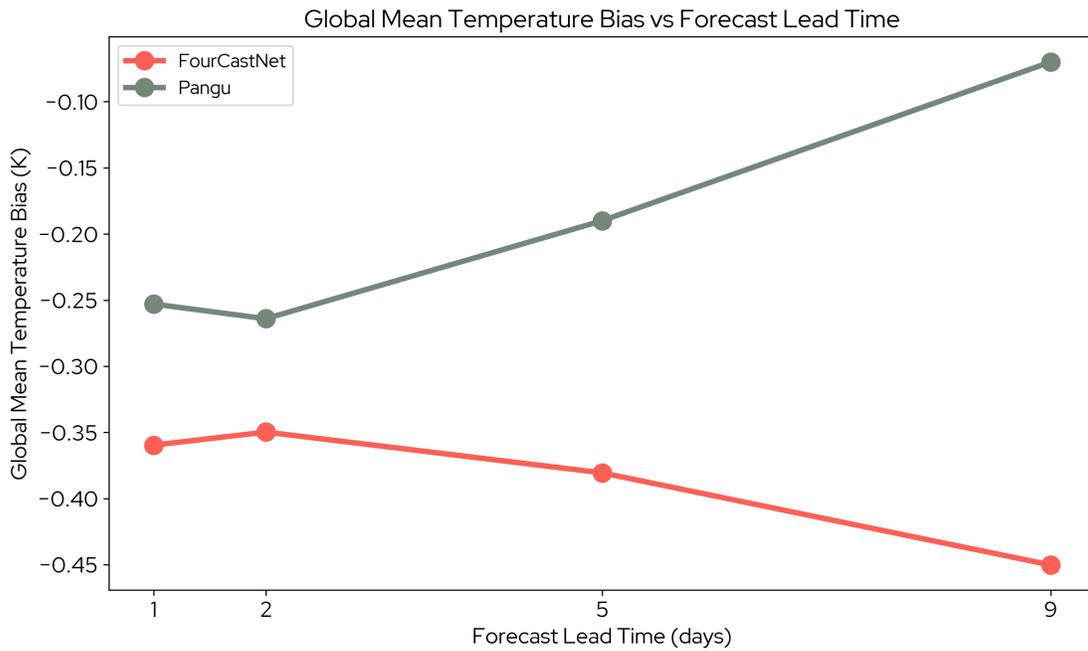


Figure S1: Global temperature bias at 1-day, 2-day, 5-day, and 9-day lead times for FourCastNet and Pangu. Even at 1-day lead, both models exhibit a cold bias, with the bias generally increasing with lead time for FourCastNet and decreasing with lead time for Pangu (although as shown in Figure 1, there is still significant, offsetting regional bias at longer leads).

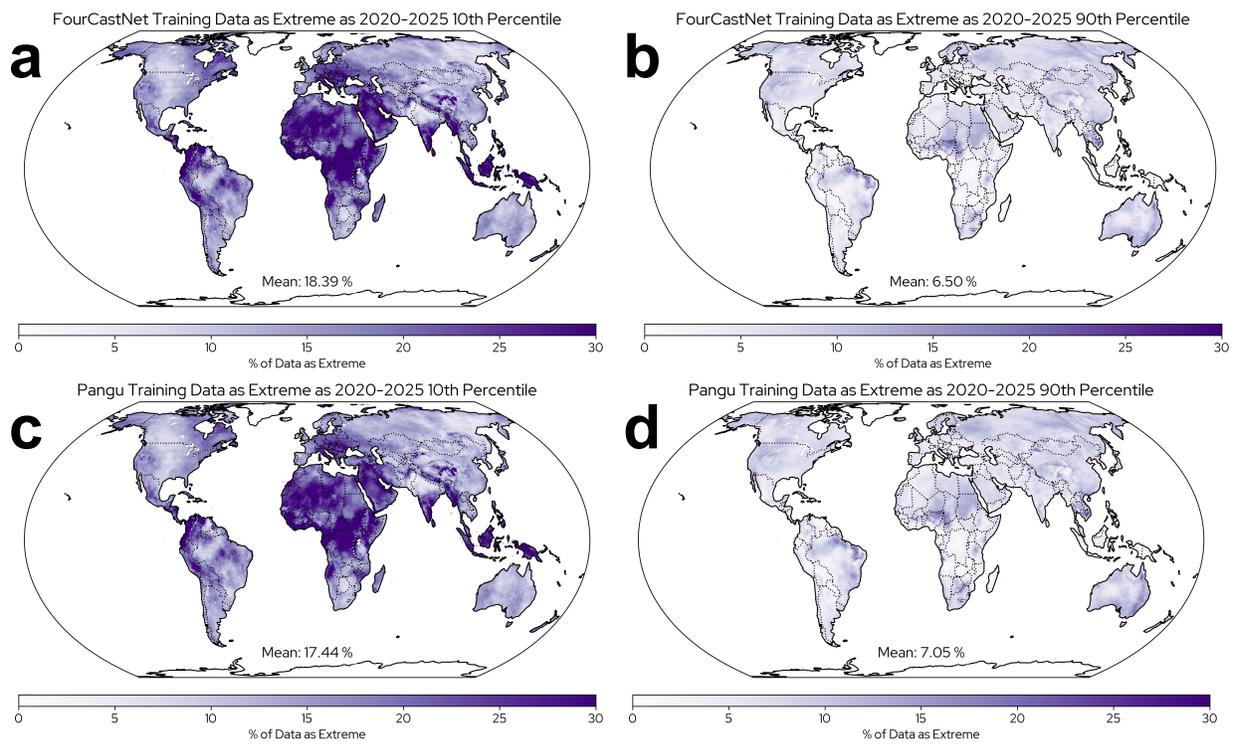


Figure S2: The percent of training data as or more extreme than the 10th and 90th percentiles of 2020-2025 ERA5 temperatures for FourCastNet (a, b) and Pangu (c, d). Global means are shown at the bottom of each panel and are similarly displayed in Figure 3f.

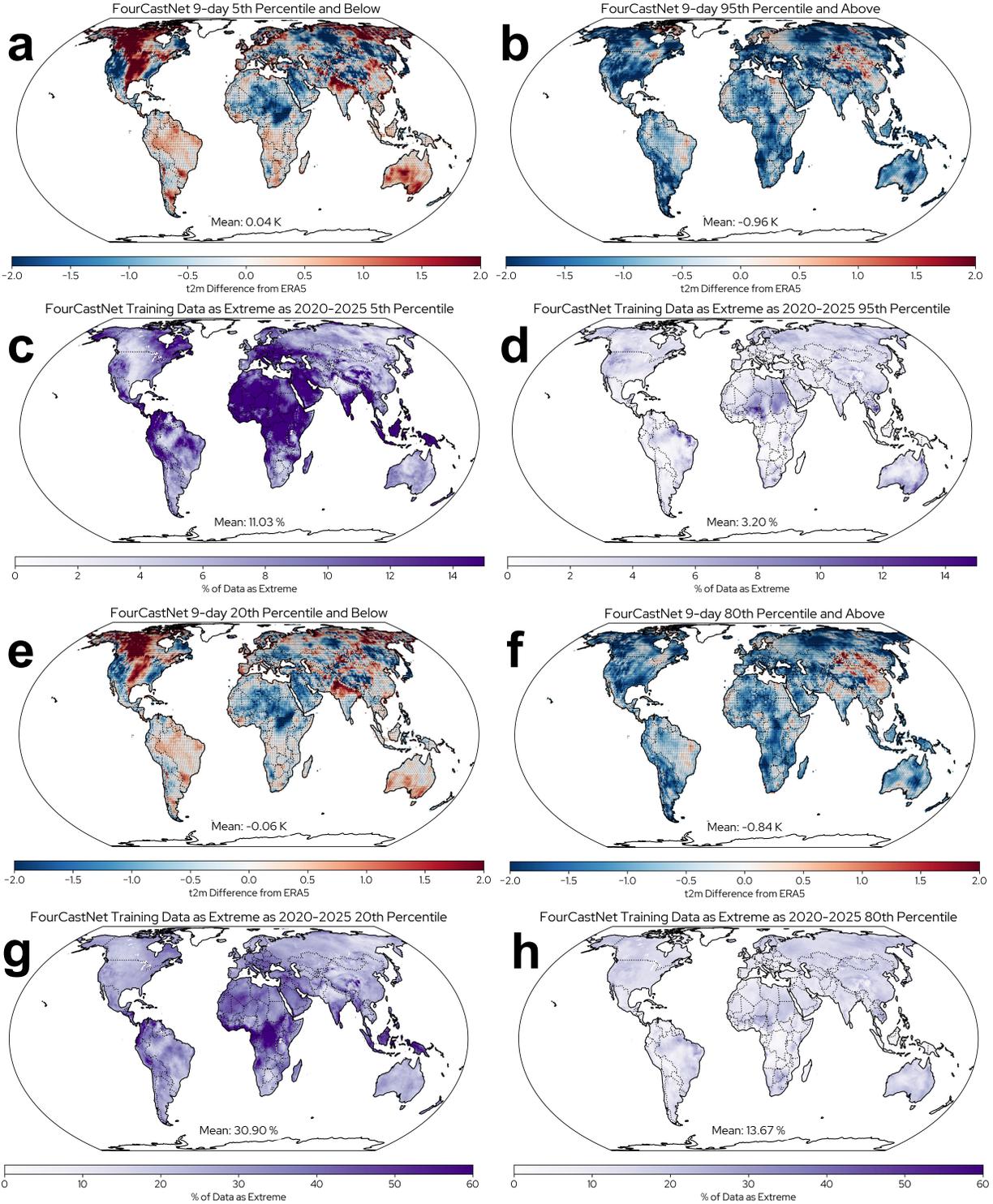


Figure S3: As in Figures 3 (a-d) and S2, but for the 5th and 95th (a, c and b, d) and 20th and 80th (e, g and f, h) percentiles of FourCastNet’s 9-day lead predictions. Global means are shown at the bottom of each panel. We see similar behavior as in Figure 3, with the hottest percentiles exhibiting a stronger cold bias than the coldest percentiles, in line with their being less training data for hot extremes. Stippling in a, b, e, and f indicates grid points where the bias is significantly non-zero.

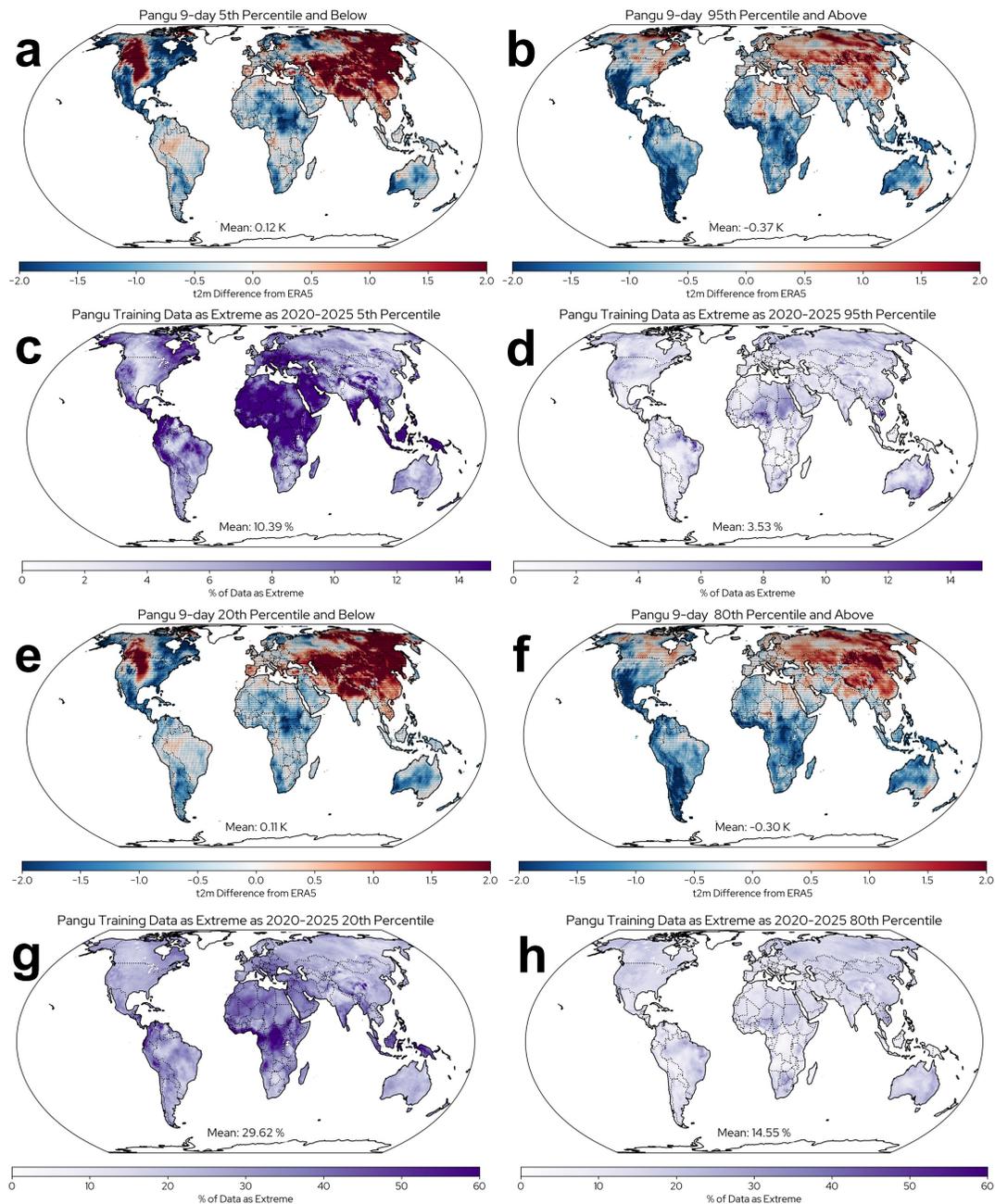


Figure S4: As in Figures 3 (a-d) and S2, but for the 5th and 95th (a, c and b, d) and 20th and 80th (e, g and f, h) percentiles of Pangu’s 9-day lead predictions. Global means are shown at the bottom of each panel. We see similar behavior as in Figure 3, with the hottest percentiles exhibiting a stronger cold bias than the coldest percentiles, in line with their being less training data for hot extremes. Stippling in a, b, e, and f indicates grid points where the bias is significantly non-zero.

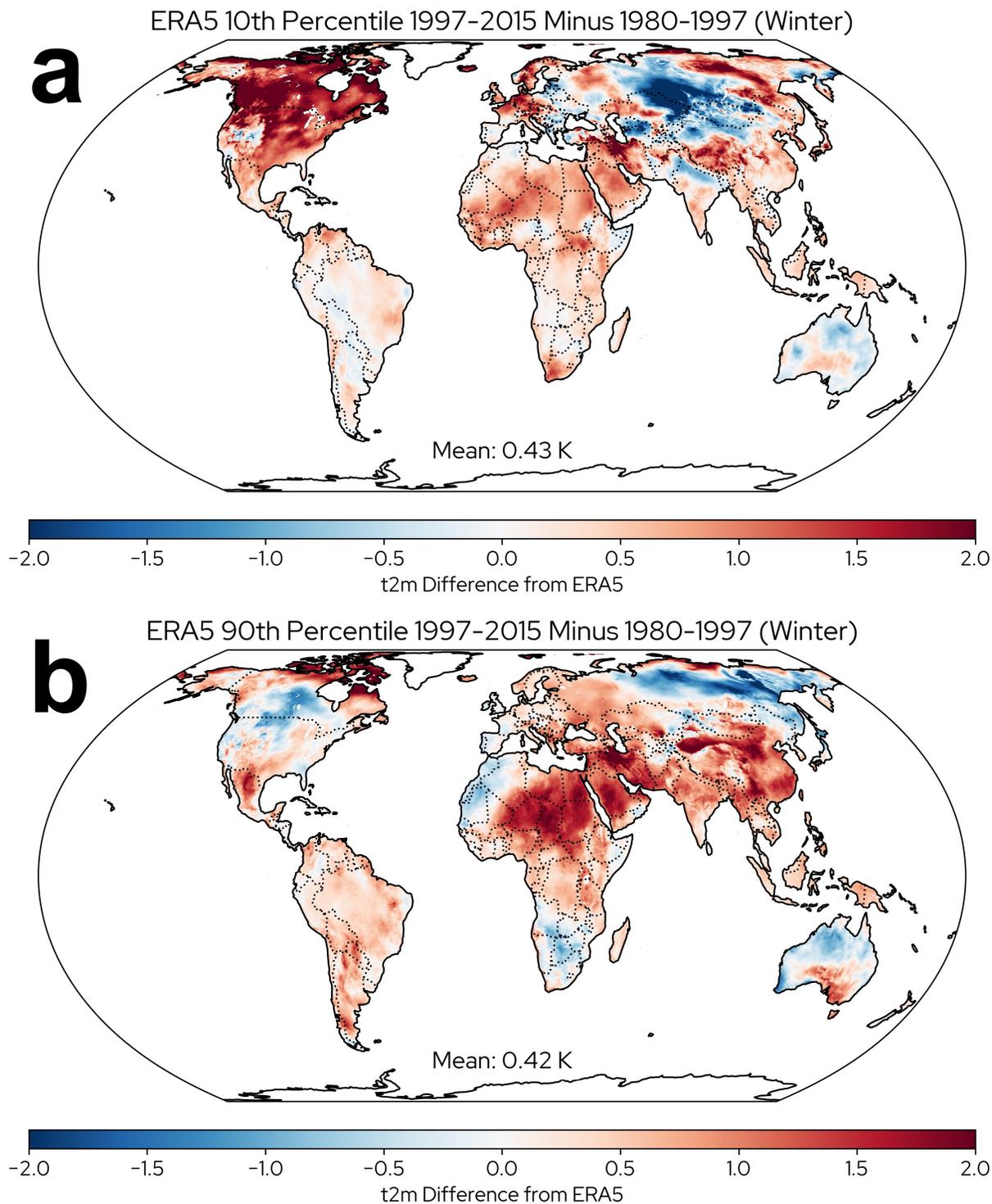


Figure S5: Difference in 10th (a) and 90th (b) percentile ERA5 winter temperatures between 1980-1997 and 1997-2015. We choose these dates as they approximately split the weather model's training sets in half. We see little global mean difference in warming between the 10th and 90th percentiles, unlike what we saw when assessing ACE2's training set (Figure 5b), which showed the change between 1940-1979 and 1980-2022.