Supporting Information for "Detecting climate signals using explainable AI with single-forcing large ensembles"

Zachary M. Labe[1]and Elizabeth A. Barnes[1]

[1]Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA

Contents of this file

1. Table S.1 to S.2
2. Figures S.1 to S.14
3. References

Corresponding author: Zachary M. Labe (zmlabe@rams.colostate.edu)

Table S.1. Description of climate model data sets used for the primary analysis in this study.

Name	Forcing	Years	# Members	Reference
ALL	Historical (to 2005), RCP 8.5	1920–2080	20	CESM-LE - Kay et al. (2015)
AER+	ALL, but fixed greenhouse gases to 1920 levels	1920–2080	20	XGHG - Deser et al. (2020)
GHG+	ALL, but fixed industrial aerosols to 1920 levels	1920–2080	20	XAER - Deser et al. (2020)

Table S.2. Description of observational data sets used for the primary analysis in this study.

Name	Data Set	Years	Reference
20CRv3	NOAA-CIRES-DOE 20th Century Reanalysis V3	1920–2015	Slivinski et al. (2019)

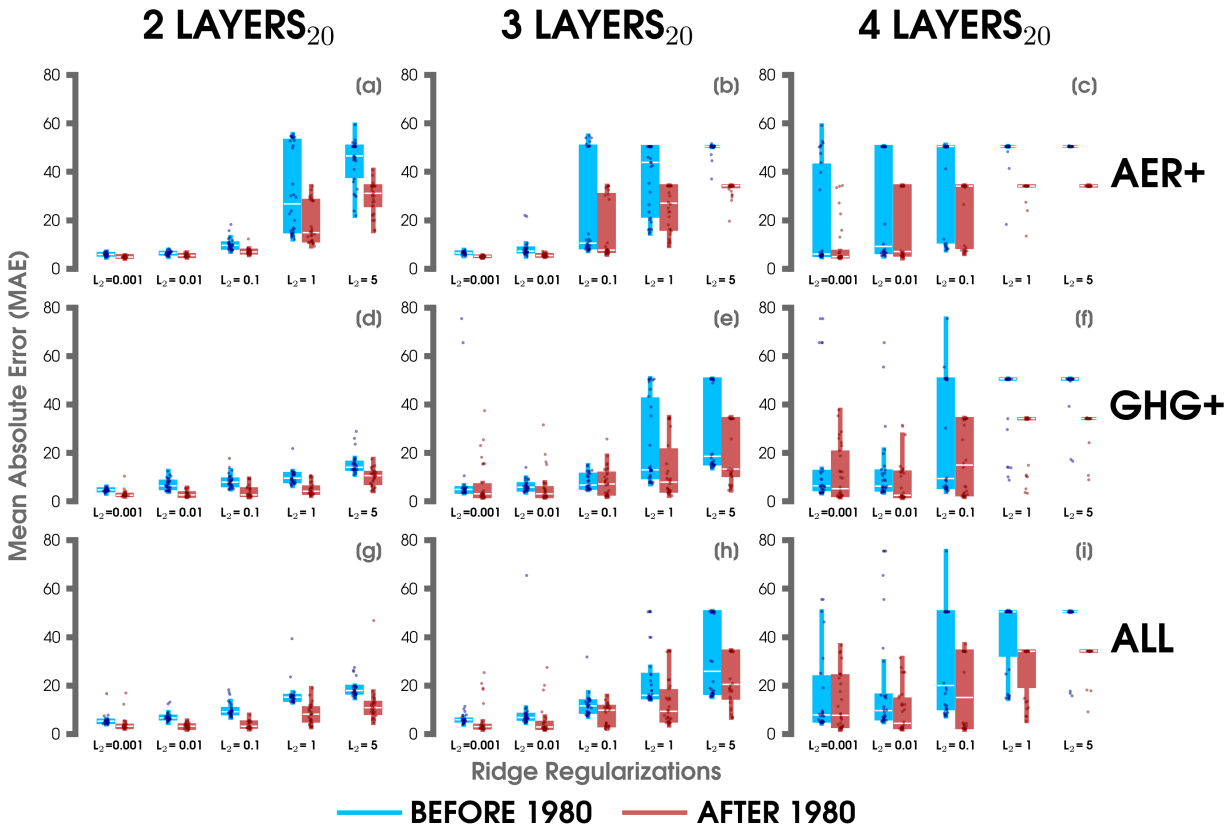

Figure S.1. Box-and-whisker plots showing the mean absolute error (MAE) of testing years before 1980 (blue) and after 1980 (red) for the ANNs trained separately on each large ensemble experiment (AER+; a-c, GHG+; d-f, ALL; g-i). Results are shown for ANN architectures using 2 hidden layers of 20 nodes each (a,d,g), 3 hidden layers of 20 nodes each (b,e,h), and 4 hidden layers of 20 nodes each (c,f,i) and different L_2 regularization values (0.001, 0.01, 0.1, 1, 5). Each box-and-whisker distribution of ANNs is comprised of 10 iterations (different combinations of training and testing data and random initialization seeds) for 3 separate epochs (100, 500, 1500).

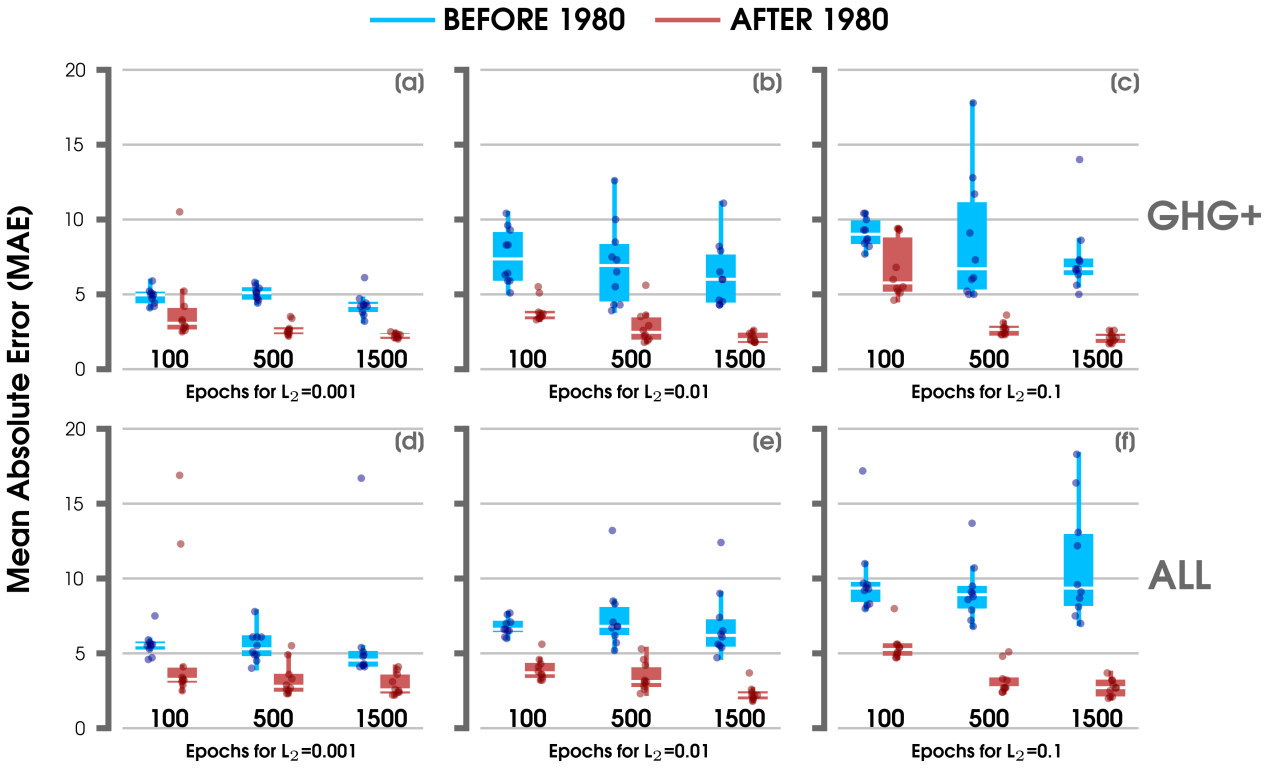

Figure S.2. Box-and-whisker plots showing the mean absolute error (MAE) of testing years before 1980 (blue) and after 1980 (red) for the ANNs trained separately on two large ensemble experiments (GHG+; a-c, ALL; d-f) using architectures with 2 hidden layers of 20 nodes each, three different epochs (100, 500, 1500), and L_2 regularization values of 0.001 (a,d), 0.01 (b,e), and 0.1 (c,f). Each box-and-whisker distribution of ANNs is comprised of 10 iterations using different combinations of training and testing data and random initialization seeds.

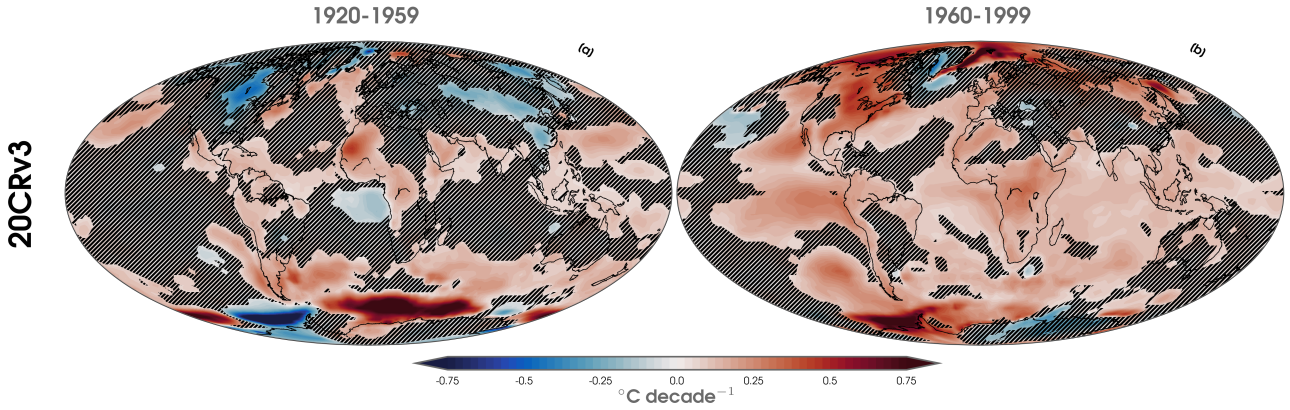

Figure S.3. Annual linear least squares trends of 2-m temperature (°C per decade) over 1920 to 1959 (a) and 1960 to 1999 (b) using 20CRv3 reanalysis (observations). Statistically significant trends are shown with shaded contours at the 95% confidence level following the Mann-Kendall (MK) test (Mann, 1945; Bevan & Kendall, 1971). Insignificant trends are masked out using black hatch marks.

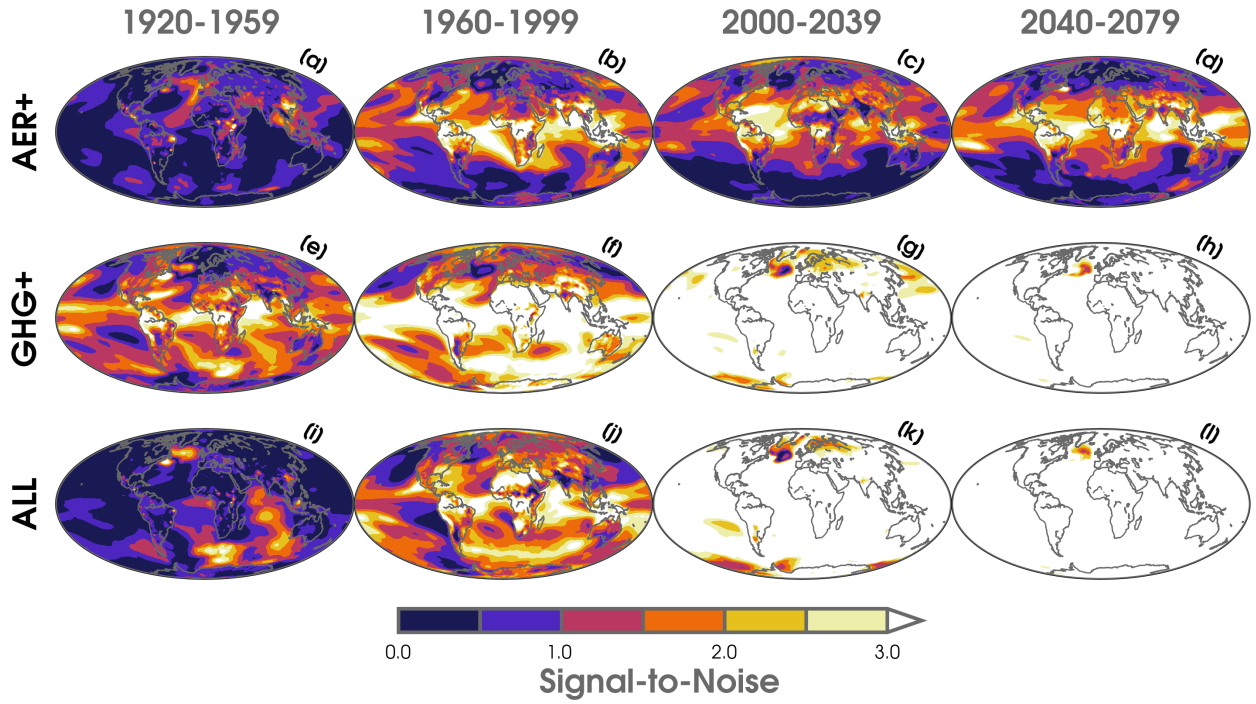

Figure S.4. Signal-to-noise ratio (SNR) maps of annual mean 2-m temperature over 1920 to 1959 (a,e,i), 1960 to 1999 (b,f,j), 2000 to 2039 (c,g,k), and 2040 to 2079 (d,h,l) for three large ensemble simulations (AER+; a-d, GHG+; e-h, ALL; i-l). SNR is calculated here as the absolute value of the ensemble mean trend (forced response) normalized by the standard deviation of trends across individual ensemble members (internal variability) for each 40 year period.

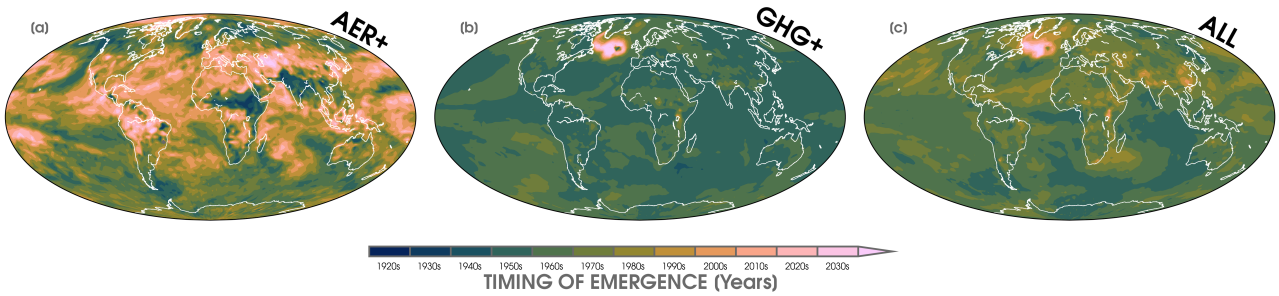

Figure S.5. Average timing of emergence (ToE) maps defined as the first year the 10-year running-mean 2-m (annual mean) temperature exceeds and stays above the mean 1920-1949 period by more than two standard deviations (e.g., Lehner et al., 2017) for each ensemble member in the three large ensemble simulations (AER+; a, GHG+; b, ALL; c).

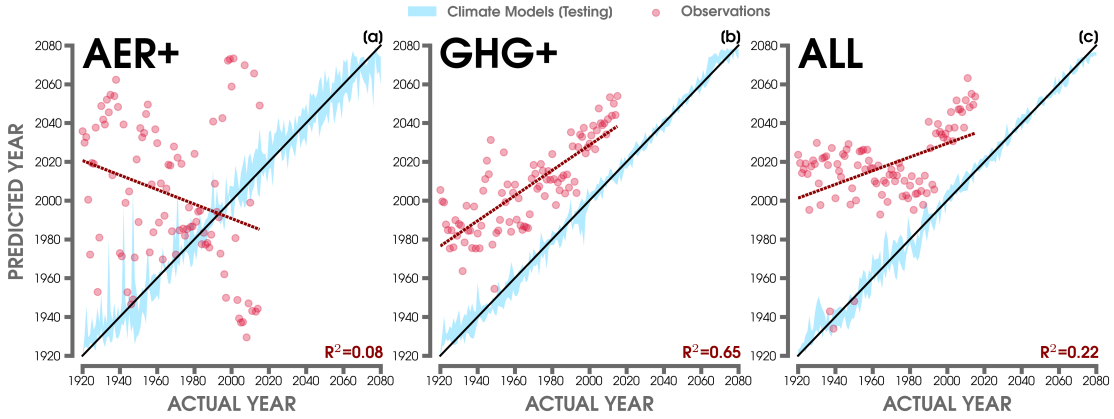

Figure S.6. (a) Predictions of the year by the ANN (y-axis) compared to the actual year (x-axis) from maps of annual 2-m temperature, but with the global mean temperature removed in AER+. (b) Same as (a) but for GHG+. (c) Same as (a) but for ALL. The blue shading highlights the 5th-95th percentiles of predictions from the large ensemble testing data. The red points show the ANN predictions using 20CRv3 observations. The red dashed line shows the linear least squares fit through the predicted observations in each model, and the associated R^2 is shown in the lower right-hand corner. The 1:1 line (or perfect prediction) is overlaid in black.

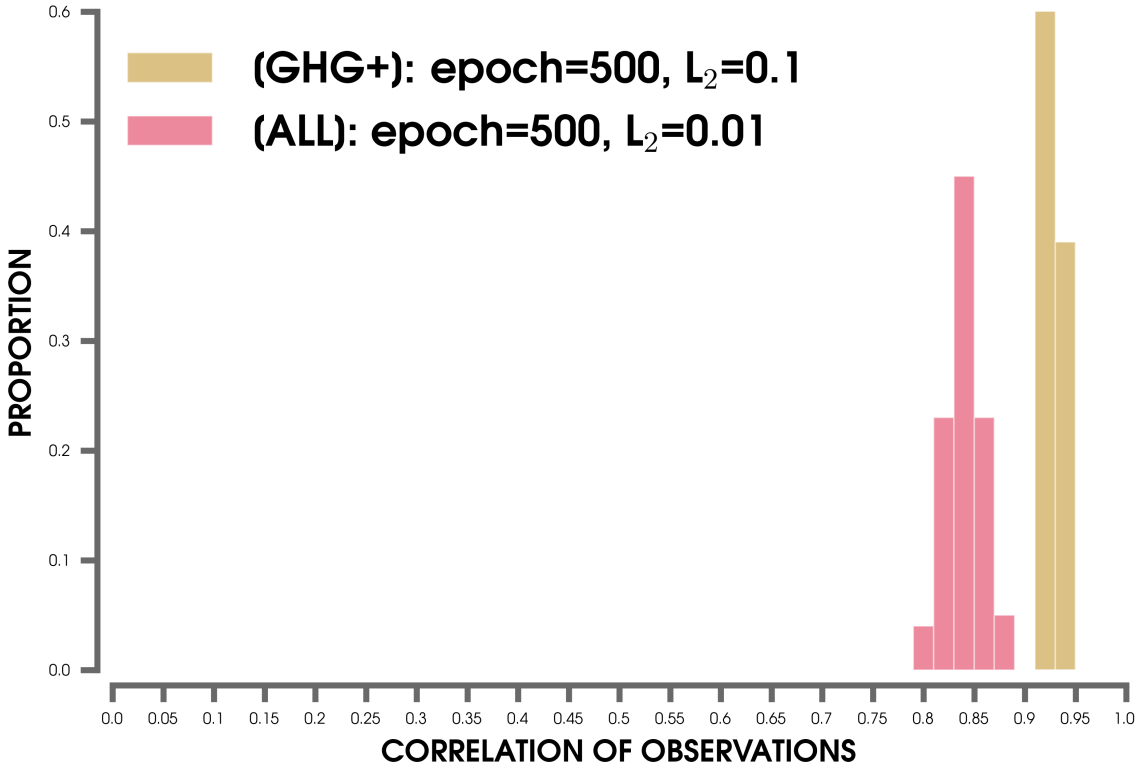

Figure S.7. Histogram of correlations between the actual years and the ANN-predicted years based on maps from 20CRv3 observations for the GHG+ (brown) and ALL (red) ANNs (created using different combinations of training and testing data). The distributions are selected by their respective ANN architecture with the highest median correlation over the six hyperparameter combinations (epochs and L_2 regularization) shown in Figure 4.

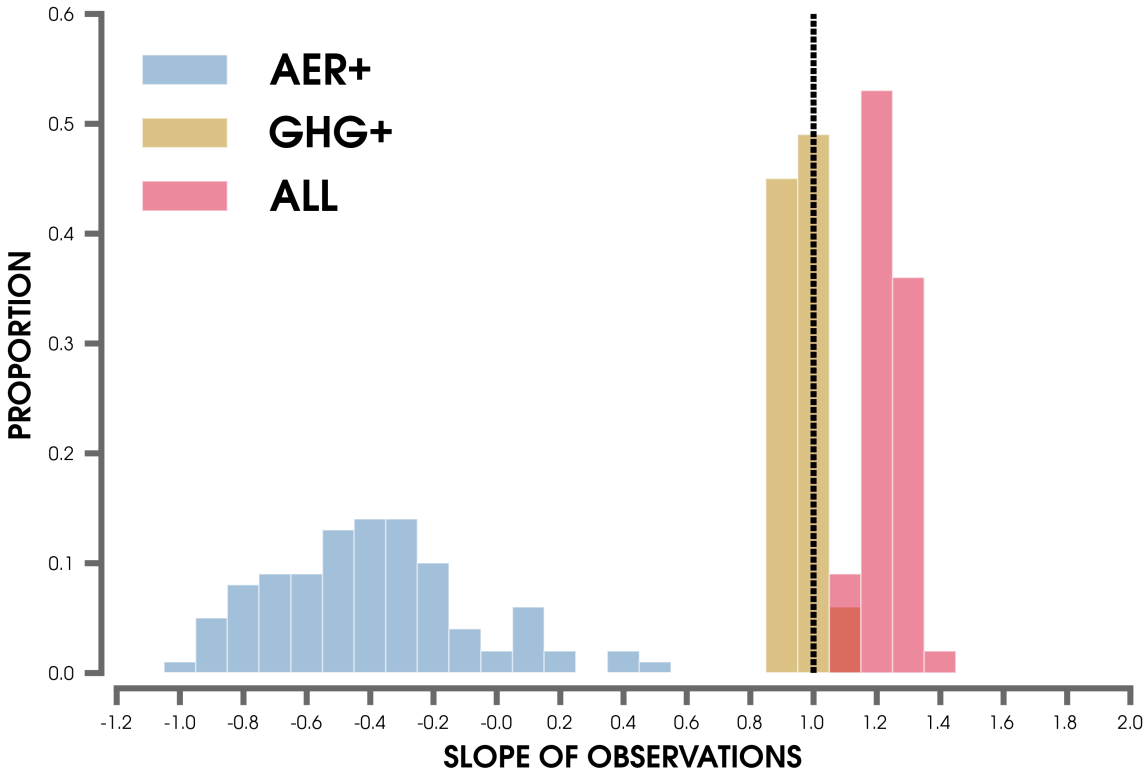

Figure S.8. Histogram of the possible slopes of predicted 20CRv3 observations after considering different combinations of training and testing data for each of the AER+ (blue), GHG+ (brown), and ALL (red) ANNs. The 1:1 is highlighted by the dashed gray line.

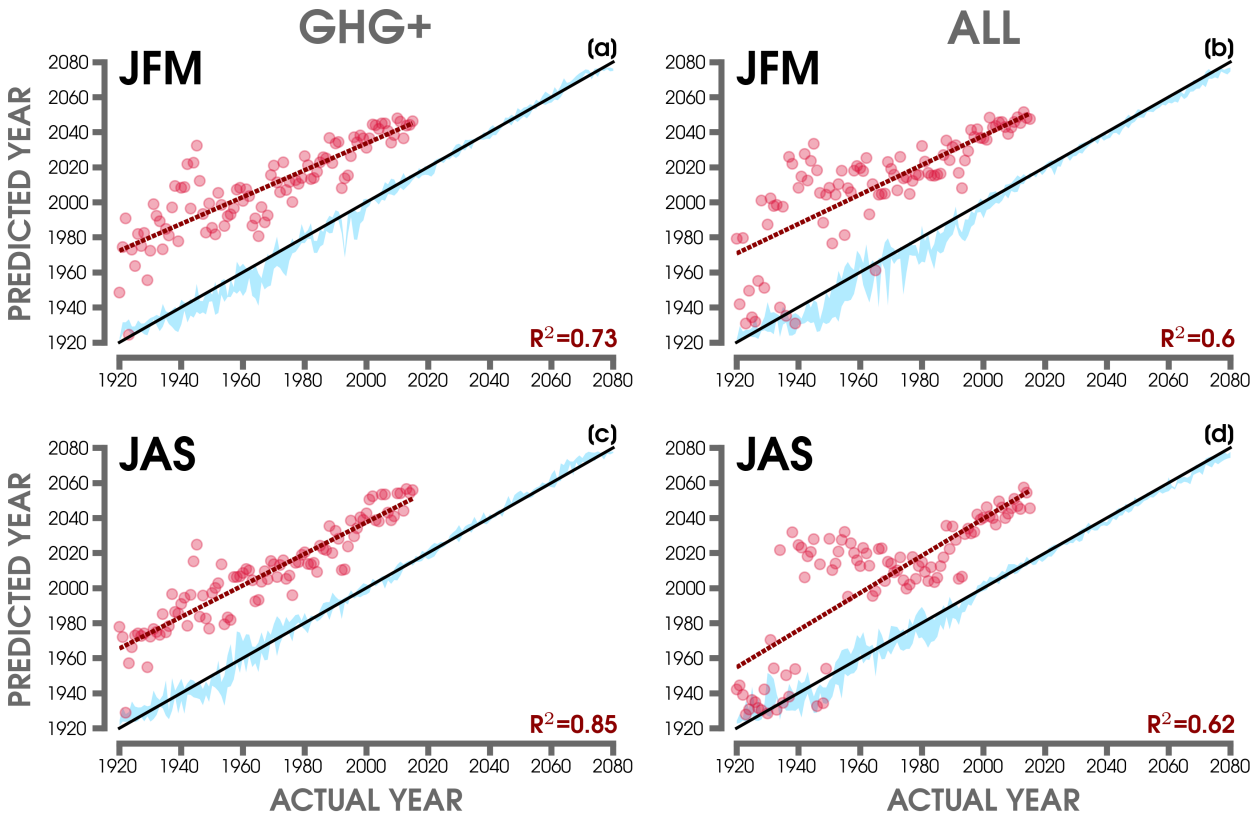

Figure S.9. (a) ANN predictions of the year (y-axis) compared to the actual year (x-axis) by models trained on global maps of 2-m temperature in GHG+ (a,c) and ALL (b,d) during January-February-March (JFM; a,b) and July-August-September (JAS; c,d). The blue shading highlights the 5th-95th percentiles of predictions from the large ensemble testing data. The red points show the predictions using 20CRv3 observations. The red dashed line shows the linear least squares fit through the predicted observations in each model, and its R^2 is shown in the lower right-hand corner. The 1:1 line is overlaid in black.

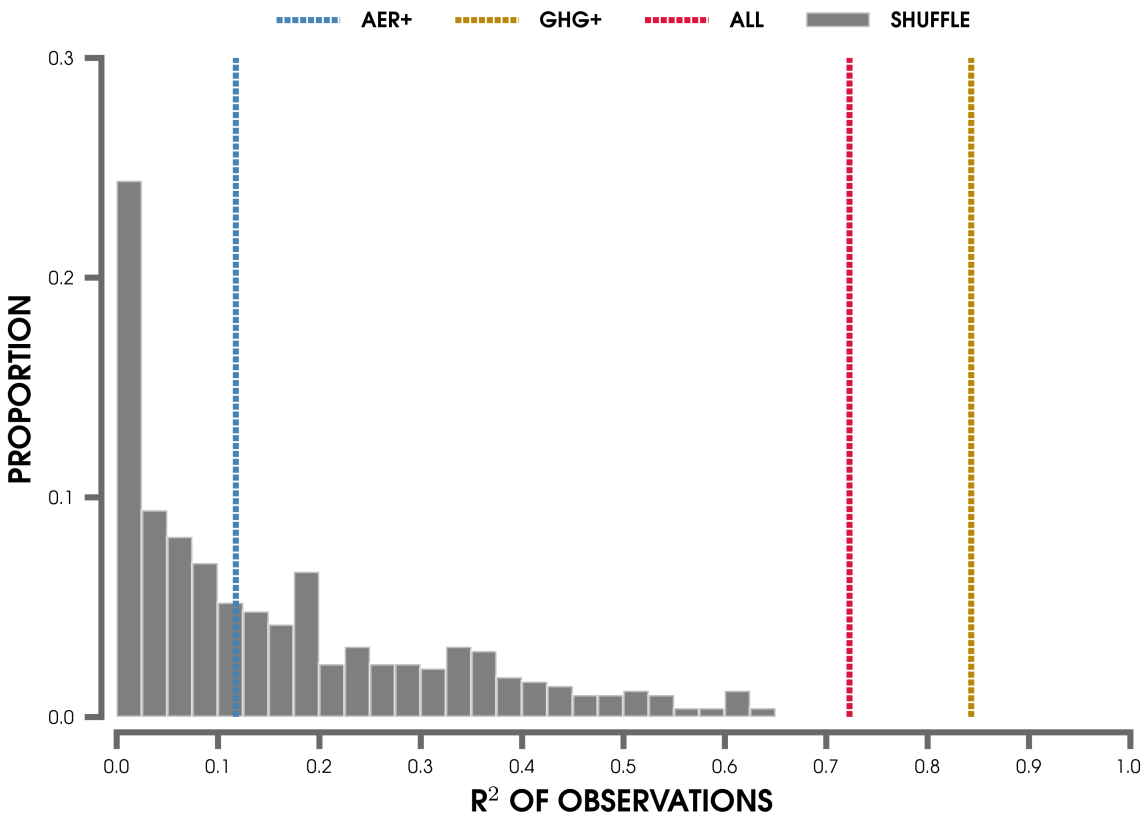

Figure S.10. Histogram of the possible R^2 values from the linear fit of predicted 20CRv3 observations after randomly shuffling the individual ensemble members and years of ALL input data to obtain 500 iterations of the ANN. The median R^2 values of the possible predictions of observations using the ANNs for AER+ (blue), GHG+ (brown), and ALL (red) are overlaid by the dashed vertical lines.

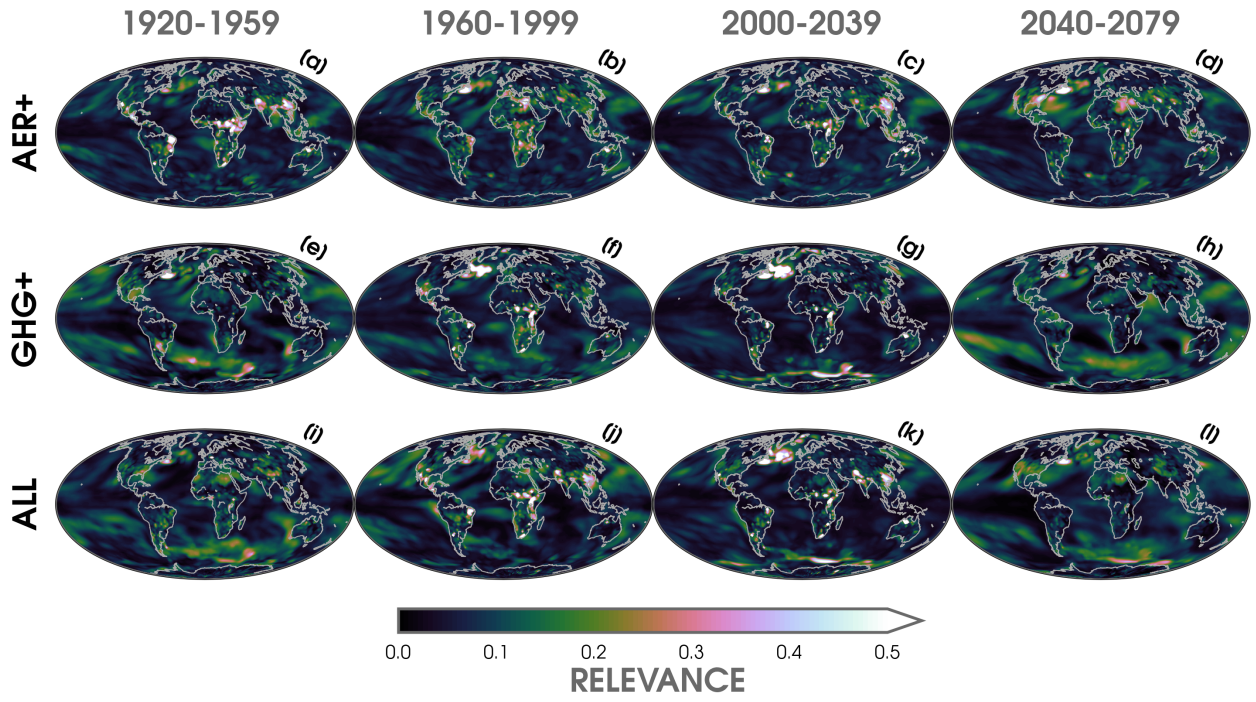

Figure S.11. LRP composite heatmaps averaged over 1920 to 1959 (a,e,i), 1960 to 1999 (b,f,j), 2000 to 2039 (c,g,k), and 2040 to 2079 (d,h,l) for the three large ensemble experiments (AER+; a-d, GHG+; e-h, ALL; i-l). Higher LRP values indicate greater relevance for the ANN's prediction.

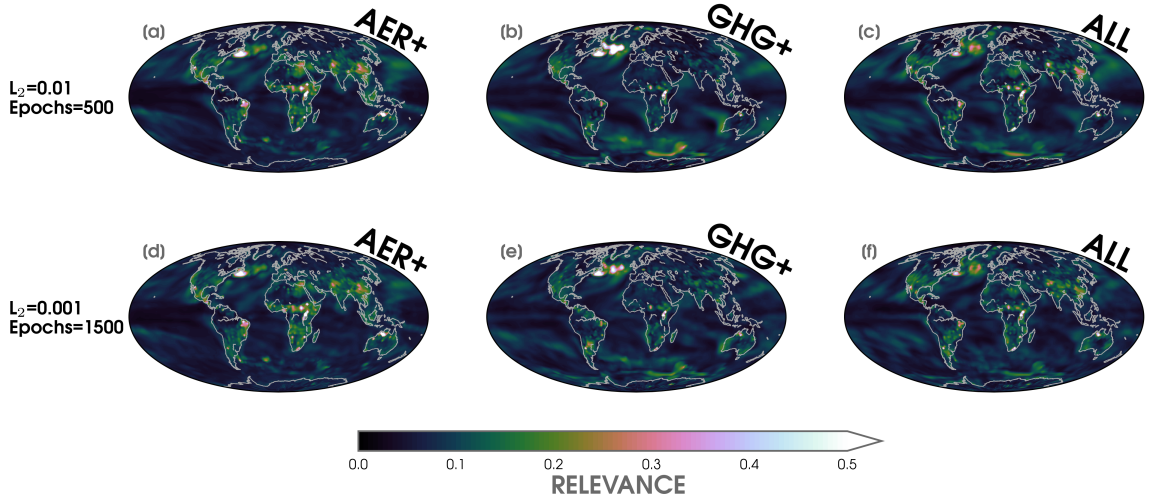

Figure S.12. (a) Composite of LRP heatmaps over 1920 to 2080 for inputs of annual 2-m temperature (global) maps from AER+ using the ANN architecture with 500 epochs and L_2 regularization set to 0.01. (b) Same as (a) but for GHG+. (c) Same as (a) but for ALL. (d-f) Same as top row, but for LRP heatmaps from an ANN architecture with 1500 epochs and L_2 regularization set to 0.001. LRP composites are generated by averaging across 100 possible ANN iterations by using different combinations of training and testing data for each large ensemble. Higher LRP values indicate greater relevance for the ANN's prediction.

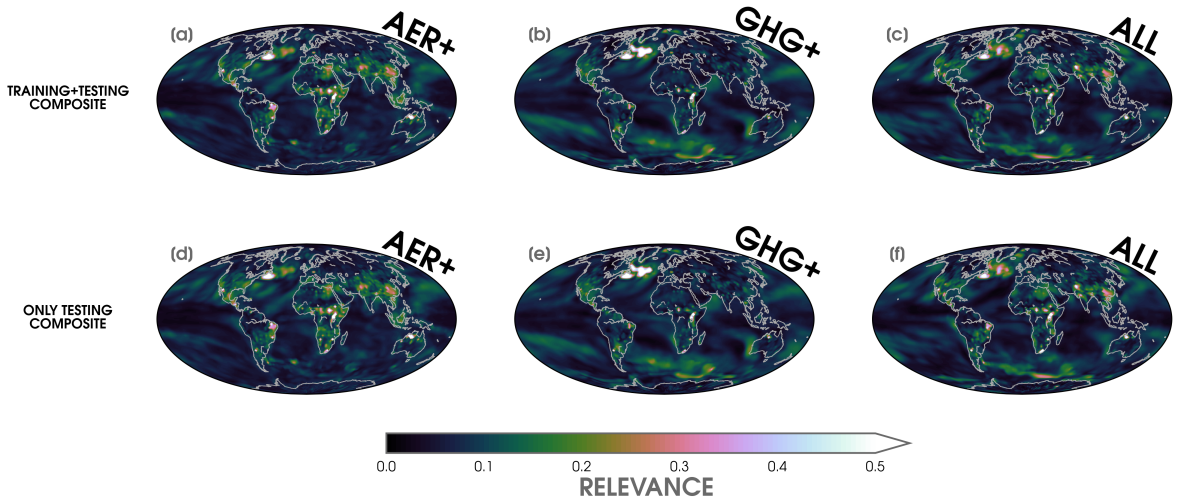

Figure S.13. (a) Composite of LRP heatmaps using both training and testing data over 1920 to 2080 for inputs of annual 2-m temperature (global) maps from AER+. (b) Same as (a) but for GHG+. (c) Same as (a) but for ALL. (d-f) Same as top row, but for LRP heatmaps using only testing data. Higher LRP values indicate greater relevance for the ANN's prediction.

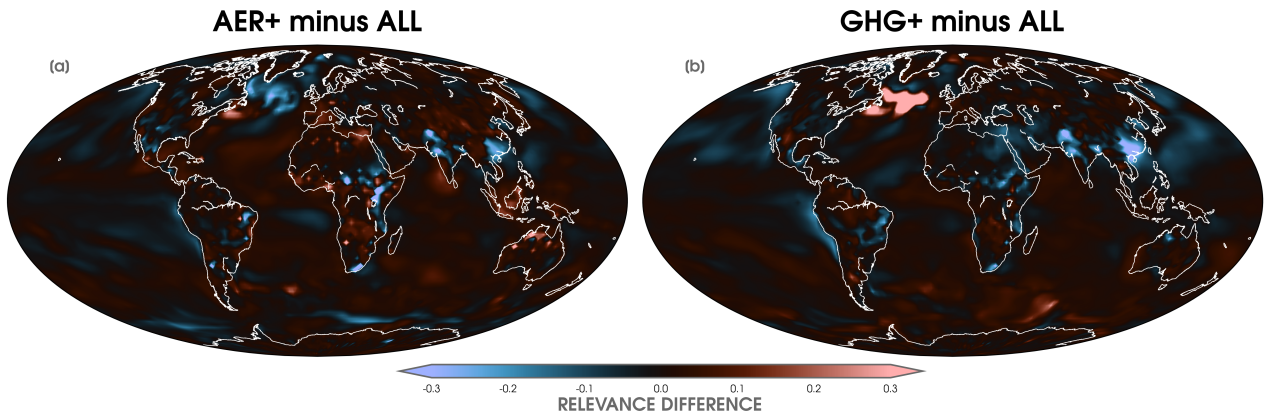

Figure S.14. Difference in composites of LRP heatmaps for AER+ minus ALL (a) and GHG+ minus ALL (b) over 1960 to 2039 for inputs of annual 2-m temperature (global) maps. LRP composites are first generated by averaging across 100 possible ANN iterations by using different combinations of training and testing data for each large ensemble.

References

Bevan, J. M., & Kendall, M. G. (1971). Rank Correlation Methods. *The Statistician*. doi: 10.2307/2986801

Deser, C., Phillips, A. S., Simpson, I. R., Rosenbloom, N., Coleman, D., Lehner, F., ... Stevenson, S. (2020, sep). Isolating the Evolving Contributions of Anthropogenic Aerosols and Greenhouse Gases: A New CESM1 Large Ensemble Community Resource. *Journal of Climate*, *33*(18), 7835–7858. Retrieved from `https://doi.org/10.1175/JCLI-D-20-` doi: 10.1175/JCLI-D-20

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., ... Vertenstein, M. (2015, aug). The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349. Retrieved from `http://journals.ametsoc.org/doi/10.1175/BAMS-D-13-00255.1` doi: 10.1175/BAMS-D-13-00255.1

Lehner, F., Deser, C., & Terray, L. (2017). Toward a new estimate of "time of emergence" of anthropogenic warming: Insights from dynamical adjustment and a large initial-condition model ensemble. *Journal of Climate*, *30*(19). doi: 10.1175/JCLI-D-16-0792.1

Mann, H. B. (1945). Nonparametric Tests Against Trend. *Econometrica*. doi: 10.2307/1907187

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., ... Wyszyński, P. (2019, oct). Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, *145*(724), 2876–2908. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3598` doi: 10.1002/qj.3598