Supporting Information for "Exploring a data-driven approach to identify regions of change associated with future climate scenarios"

Zachary M. Labe[1], Thomas L. Delworth[2], Nathaniel C. Johnson[2], and

William F Cooke[2]

[1]Atmospheric and Oceanic Sciences Program, Princeton University, NJ, USA

[2]NOAA/OAR/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

Contents of this file

Corresponding author: Zachary M. Labe (zachary.labe@noaa.gov)

Text S1: Artificial Neural Network Parameters

For each classification task (e.g., predicting 5 climate scenarios or 2 climate scenarios) and climate variable (temperature or precipitation) (Figure S1), we find a unique artificial neural network (ANN) which scores the highest in validation data accuracy. These final architecture details are listed in Table S2, and each one is selected by identifying the median accuracy of different ANN iterations for a range of network complexities. For networks with similar median skill, we select the higher ridge regularization parameter to help reduce overfitting and improve interpretability. The iterations are conducted by randomly selecting different SPEAR ensemble members used for training, testing, and validation data and alternating different random initialization seeds. This is conducted three times each for the 5-class ANN and five times each for the binary ANNs, and these results are shown in Figures S5-S6 and S7-S10, respectively. The relatively small number of random iterations for each network is due to the high computational cost of this machine learning task (i.e., slow training process for a comprehensive hyperparameter sweep), but overall we find that adding more iterations does not change our skill score results (not shown).

Each of the neural networks is fully-connected and receives vectorized maps of temperature or precipitation at the input layer that have a size equal to 207,360 units, which is comprised of 360 latitude points by 576 longitude points. No other information is provided at the input layer or during the training process, and therefore the ANN has no direct knowledge of which year is associated with each climate map. The output layer contains either two or five nodes depending on the classification network (e.g., number of predicted climate scenarios) (Figure S1). All classes are balanced with 86 years of annual mean maps input for each scenario (either 1929-2014 or

2015-2100). Before inputting any data into the ANN, all climate maps are standardized by subtracting the mean of the training data and dividing by the training standard deviation. This is conducted across all years, relevant climate scenarios, and training ensemble members for every grid point.

In short, a neural network training process consists of iteratively updating the model weights and biases until the loss function is minimized. For training each ANN, we use 24 ensemble members (80% of the data). There are 4 ensemble members then used for validation, and 2 ensemble members are used as testing data for independent classification evaluations. We consistently use one random initialization seed and the same subsets of individual ensemble members for training, testing, and validation for the main results of this study. Skill metrics for these specific ANNs, including testing accuracy, recall (proportion of classifications out of all possible samples in a given climate scenario class), precision (proportion of climate scenario classifications actually from that particular class), and the F1 score (harmonic mean of precision and recall) (Johnson & Khoshgoftaar, 2019), are shared in the main text and figures of the manuscript (e.g., Figure 2). Across all ANNs, we use a batch size of 128, learning rate of 0.0001, a stochastic gradient descent optimizer (Ruder, 2016) using Nesterov momentum (0.9) (Nesterov, 1983), a categorical cross-entropy loss function, the rectified linear unit (ReLu; Agarap, 2018) for nonlinear transformation in the hidden layers, and a softmax activation function applied to the output layer.

To help limit overfitting, we apply several different approaches to each classification network. First, we include a ridge regularization (L_2) parameter (L_2; Friedman, 2012), which acts to

penalize larger weights across the input data and subsequently reduces autocorrelation in the gridded fields of temperature and precipitation (Sippel et al., 2019; Barnes et al., 2020; Labe et al., 2024). We test a number of different combinations of regularization values and ANN architectures and then select the L_2 separately for each variable and classification network. These final values are given in Table S2. Interestingly, we find that ANN classification accuracy is more sensitive to the choice of L_2, rather than the complexity of the network itself (i.e., number of hidden layers and nodes). In general, our networks here are relatively shallow (one to three layers) and similar to recent studies applying feed-forward neural networks to climate science applications (e.g., Toms et al., 2021; Labe & Barnes, 2022; Martin et al., 2022; Rader et al., 2022). Although a slightly deeper ANN is sometimes selected for the binary classification prediction problem (Table S2), we acknowledge that this does not necessarily imply that a more complex network is necessarily needed given such similar skill is found between architectures and training iterations. We further apply early stopping to each training process, which stops model training if there is no improvement in validation accuracy (i.e., minimizing the loss function) after 10 epochs. The network with the best weights is then returned after this technique, and note that each ANN trains for no more than 1500 epochs. Lastly, we include a dropout layer after the first hidden layer (dropout rate = 0.4), which is another form of regularization that forces the ANN to learn more slowly and acts to lessen overfitting on new unseen data (Hinton et al., 2012; Srivastava et al., 2014).

To find a more comprehensive introduction to machine learning, we recommend resources provided by Goodfellow, Bengio, and Courville (2016) and Russell and Norvig (2021). In addition,

overviews specifically related to the atmospheric sciences can be found in Chase, Harrison, Lackmann, and McGovern (2022); Chase, Harrison, Burke, Lackmann, and McGovern (2022); de Burgh-Day and Leeuwenburg (2023), including for the use of explainability methods (Toms et al., 2020; Flora et al., 2023).

Text S2: Software Programs and Other Tools

As suggested by Irving (2016) on improving data and method standards in climate science, we provide references that document the important computational packages utilized in this work. Preprocessing of the large ensemble data was completed using CDO v1.9.10 (Schulzweida, 2019) and NCO v5.0.1 (Zender, 2008). Python code for the machine learning models and other statistical analysis is available from Labe, Delworth, Johnson, and Cooke (2023). The majority of this study uses Python v3.9.13 (Rossum & Drake, 2009) with the Conda v23.1.0 (Anaconda, 2023) environment and package management system. Specific Python packages that make up the majority of the analysis include Numpy v1.22.4 (Harris et al., 2020), SciPy v1.8.1 (Virtanen et al., 2020), Scikit-learn v1.1.1 (Pedregosa et al., 2011), TensorFlow/Keras v2.7.0 (Abadi et al., 2016; Chollet, 2015), iNNvestigate v2.0.2 (Alber et al., 2019), Matplotlib v3.5.2 (Hunter, 2007), Basemap v1.3.6, (*Basemap*, 2022), CMasher v1.6.3 (van der Velden, 2020), and cmocean v2.0 (Thyng et al., 2016).

Table S.1. List of the GFDL SPEAR Large Ensemble experiments (medium resolution configuration (MED)) evaluated using the neural network framework. More information on the model can be found at `https://www.gfdl.noaa.gov/spear_large_ensembles/`, and it is comprehensively documented in Delworth et al. (2020).

Experiment Name	Climate Scenario	Years	# Members
SPEAR_MED_SSP119	SSP1-1.9	2015-2100	30
SPEAR_MED_SSP245	SSP2-4.5	2015-2100	30
SPEAR_MED_SSP585	SSP5-8.5	2015-2100	30
SPEAR_MED_NATURAL	Only Natural Forcing	2015-2100	30
SPEAR_MED_HISTORICAL	CMIP6 Historial Forcing	1929-2014	30
SPEAR_MED_SSP534OS	SSP5-3.4OS	2015-2100	30
SPEAR_MED_SSP534OS_10ye	SSP5-3.4OS, but with CO_2/CH_4 mitigation starting 10 years earlier	2015-2100	30

Table S.2. Parameters for the artificial neural network (ANN) architecture that is ultimately selected for each classification network. These choices are determined by identifying the best performing network after a hyperparameter tuning process conducted for each separate variable (temperature and precipitation) and sequence of predicted climate scenarios, as shown in Figures S5-S10. This is done by identifying the combination of ridge regularization parameter and architecture (i.e., number of layers and nodes) with the highest median categorical accuracy after comparing several networks with random seeds. See Text S1 for more details.

Artificial Neural Network – Possible Classes	Variable	# Layers	# Nodes Per Layer	Ridge regularization (L_2)
Historical, Natural, SSP1-1.9, SSP2-4.5, SSP5-8.5	Temperature	1	100	0.1
Historical, Natural, SSP1-1.9, SSP2-4.5, SSP5-8.5	Precipitation	1	100	0.1
SSP2-4.5, SSP5-8.5	Temperature	1	20	0.2
SSP2-4.5, SSP5-8.5	Precipitation	3	100	0.05
SSP1-1.9, SSP2-4.5	Temperature	2	20	0.05
SSP1-1.9, SSP2-4.5	Precipitation	3	100	0.05

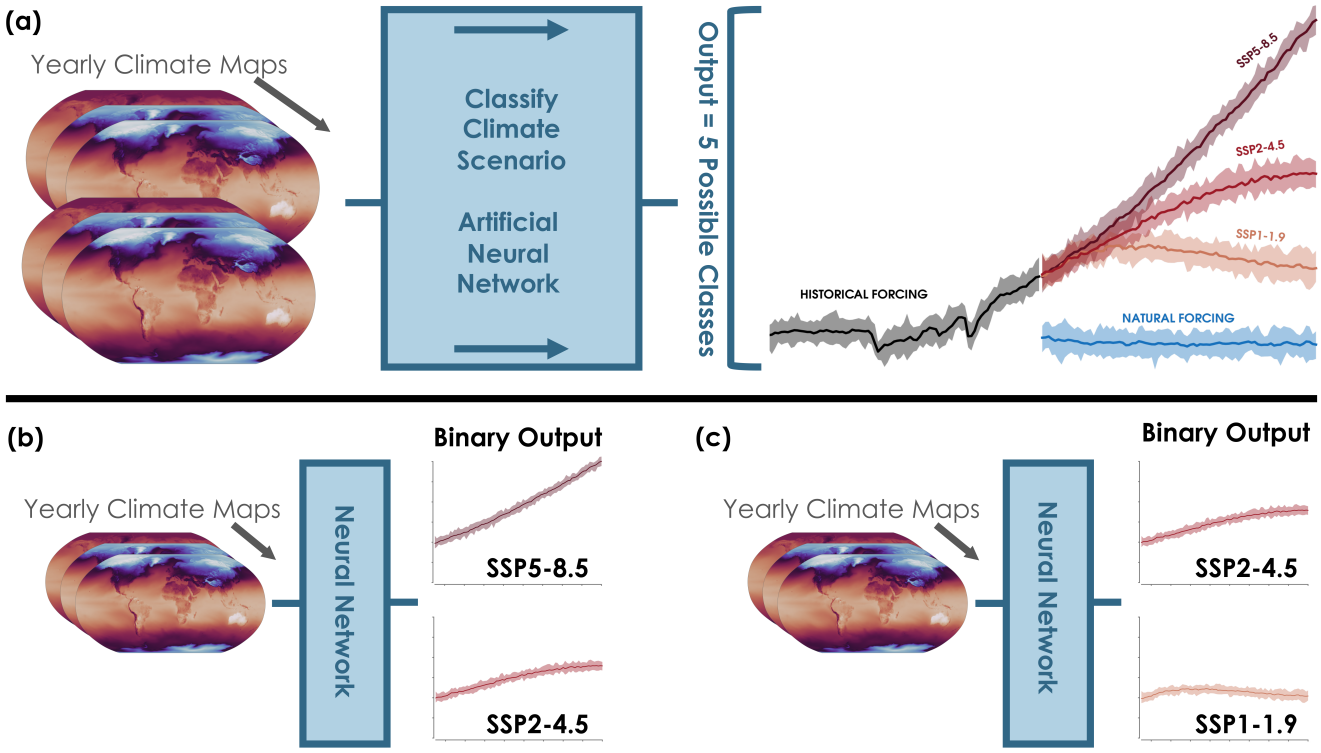

Figure S1. Outline of our approach for classifying maps of climate variables to individual climate scenarios. (a) A classification ANN that takes inputs of global maps of annual mean near-surface temperature or total precipitation and then outputs whether each map is from a historical forcing scenario, a natural forcing scenario, Shared Socioeconomic Pathway (SSP) 1-1.9 (SSP1-1.9), SSP2-4.5, or SSP5-8.5. See Text S1 and Table S2 for the architecture specifications and hyperparameter choices. (b) As in (a), but for an ANN that only predicts two classes (SSP2-4.5 or SSP5-8.5). (c) As in (b), but instead predicts either SSP1-1.9 or SSP2-4.5.

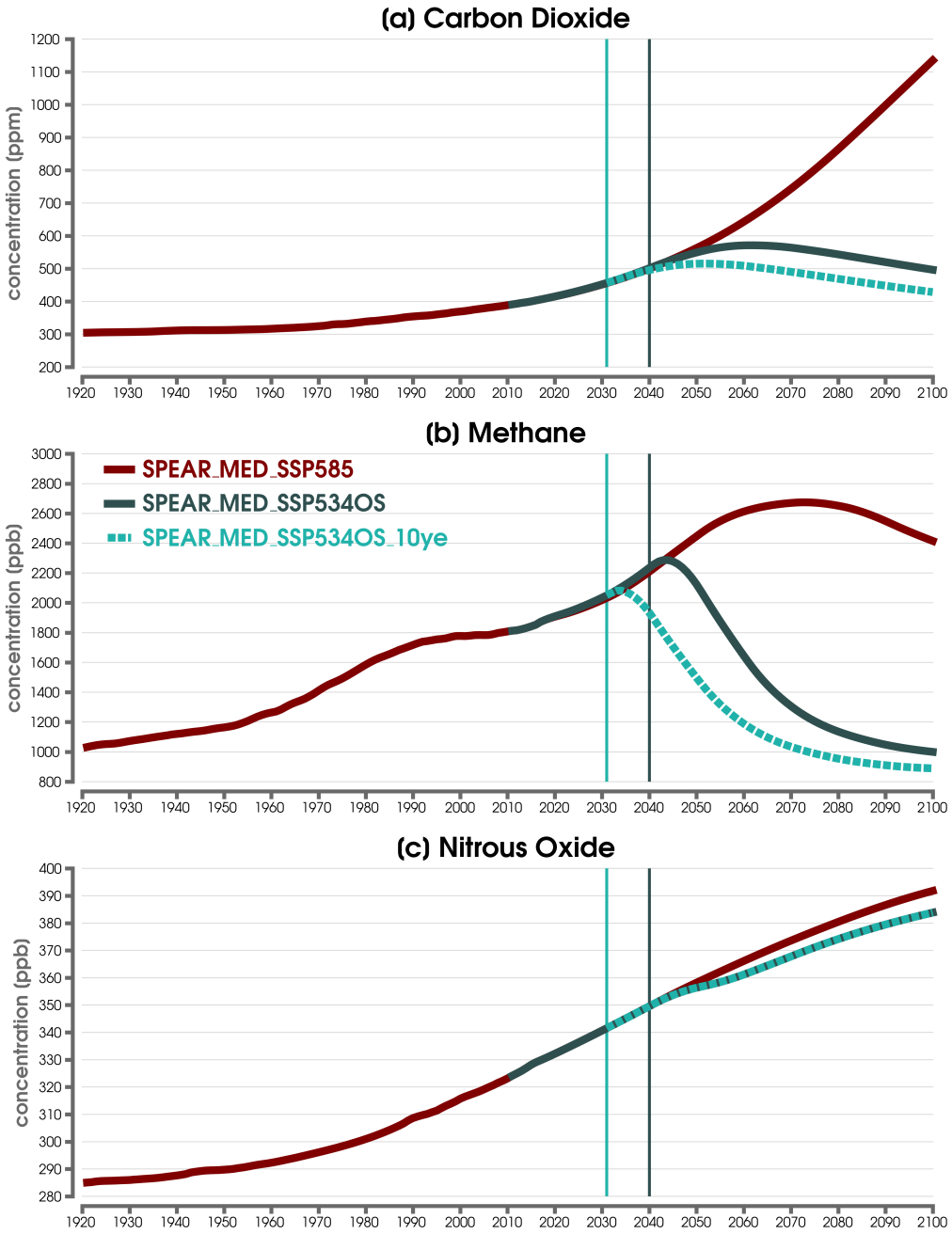

Figure S2. (a) Time series of annual mean carbon dioxide (CO_2; parts per million (ppm)) for the concatenated historical scenario and SSP5-8.5 scenario of SPEAR from 1921 to 2100 (solid red line; SPEAR_MED_SSP585), the SSP5-3.4OS scenario from 2015 to 2100 (solid dark green line; SPEAR_MED_SSP534OS), and the SSP5-3.4OS_10ye scenario from 2031 to 2100 (dashed bright green line; SPEAR_MED_SSP534OS_10ye). The vertical dark green line indicates the start of mitigation in 2040, and the bright vertical green line indicates the start of mitigation in 2031. (b) As in (a), but for methane (CH_4; parts per billion (ppb)). (c) As in (a), but for nitrous oxide (N_2O; parts per billion (ppb)).

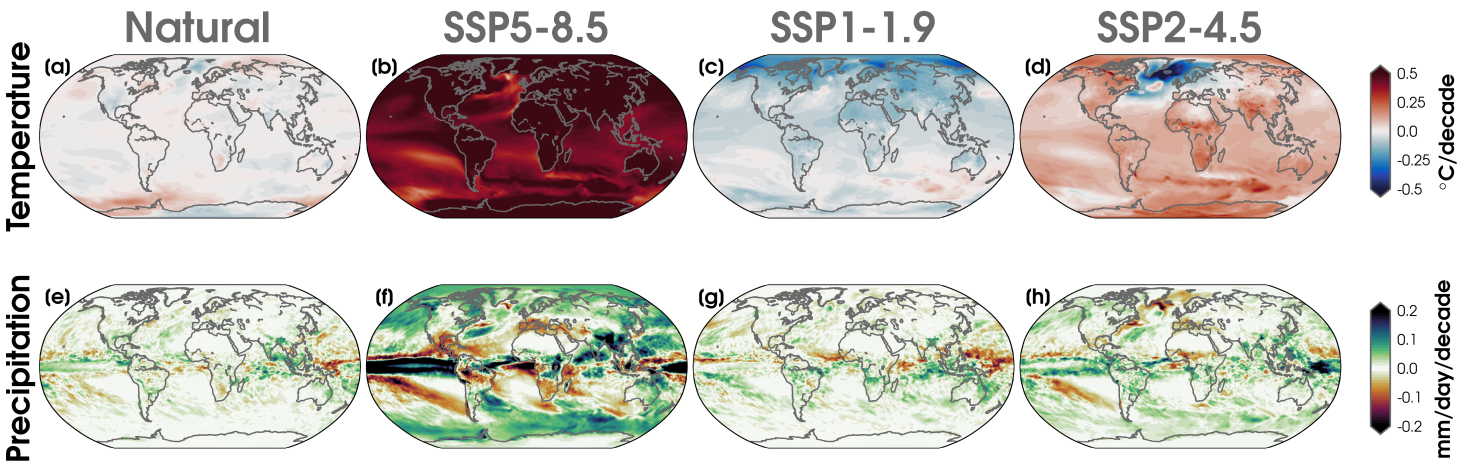

Figure S3. (a) Decadal trends of annual mean temperature (°C) from 2071 to 2100 for the ensemble mean of the natural forcing run of SPEAR. The map is calculated by considering the linear least-squares regression at every grid point in single ensemble members before averaging all members for the ensemble mean. (b) As in (a), but for the SSP5-8.5 future scenario. (c) As in (a), but for the SSP1-1.9 future scenario. (d) As in (a), but for the SSP2-4.5 future scenario. (e-h) As in (a-d), but calculated for fields of precipitation (mm/day).

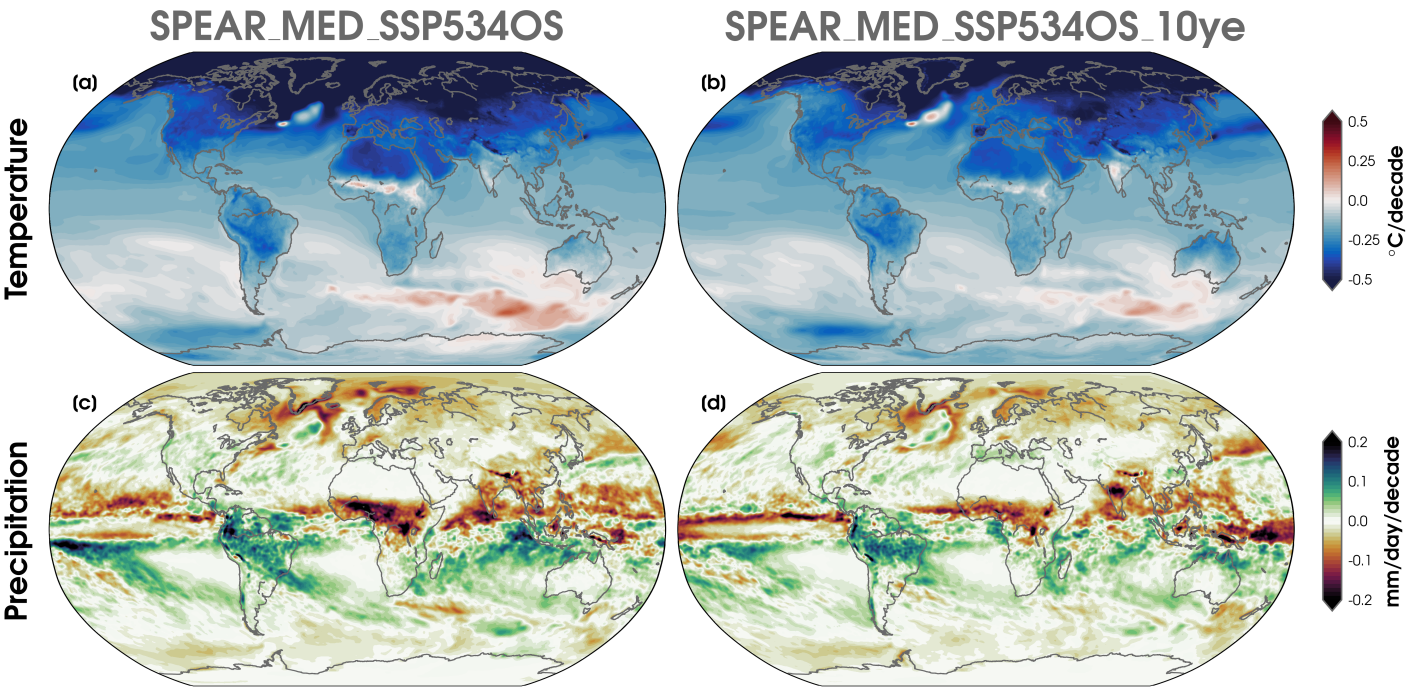

Figure S4. As in Figure S3, but for the SSP5-3.4OS future scenario (a,c) and the SSP5-3.4OS_10ye future scenario (b,d).

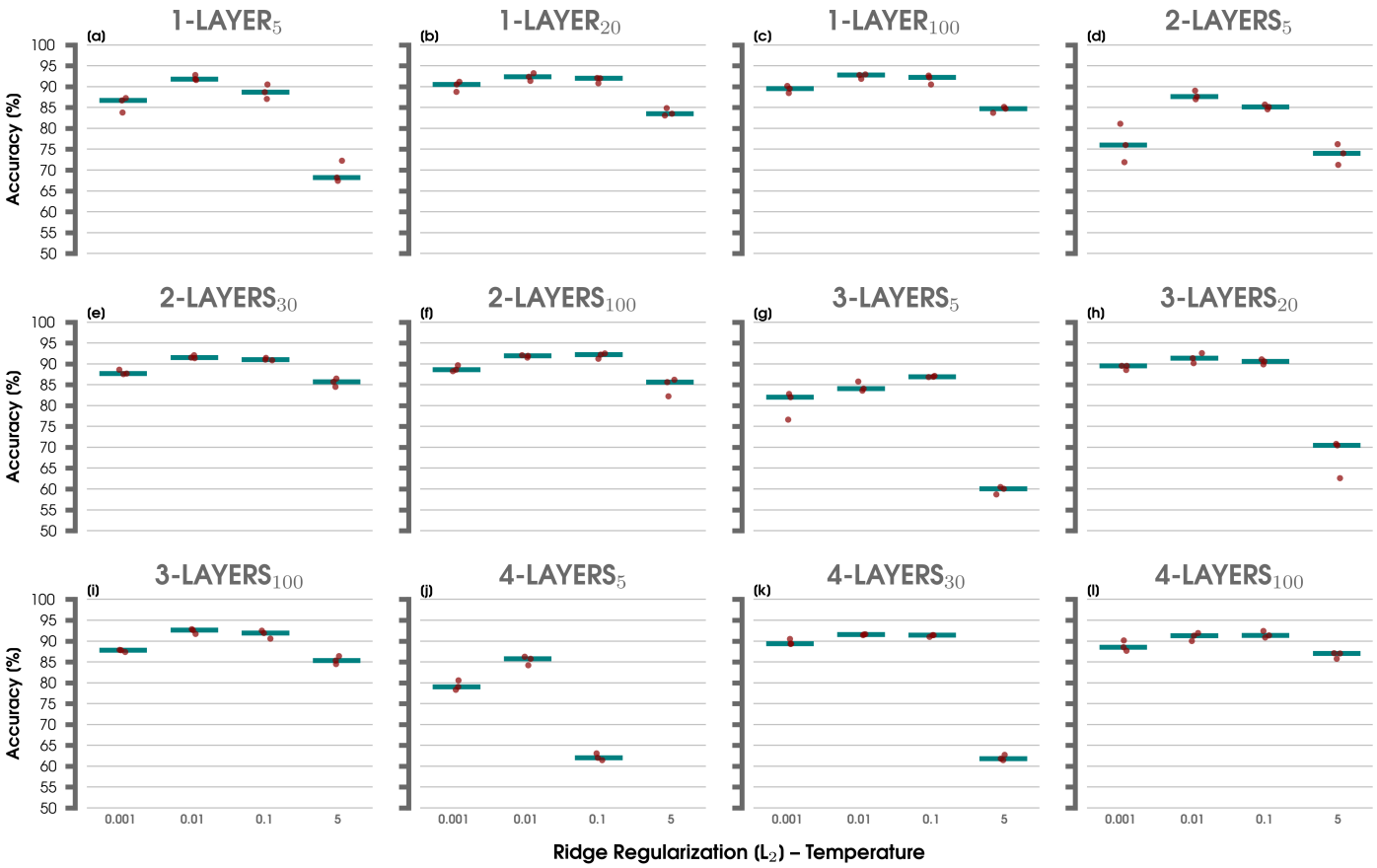

Figure S5. Scores for the total class accuracy of validation data using the 5-class artificial neural network (ANN) and inputs of global maps of annual mean temperature. (a) The ANN architecture consists of 1 hidden layer and 5 nodes. Four different L_2 regularization values (0.001, 0.01, 0.1, 5) are compared using this same ANN architecture. Each set of red points is the distribution of accuracies from 3 ANN iterations (randomized combinations of ensemble members used for training, validation, and testing and selection of random initialization seeds). The median accuracy is shown with a blue horizontal line and organized by L_2 parameter. (b-l) As in (a), but for ANN architectures of 1 hidden layer and 20 nodes, 1 hidden layer and 100 nodes, 2 hidden layers of 5 nodes each, 2 hidden layers of 30 nodes each, 2 hidden layers of 100 nodes each, 3 hidden layers of 5 nodes each, 3 hidden layers of 20 nodes each, 3 hidden layers of 100 nodes each, 4 hidden layers of 5 nodes each, 4 hidden layers of 30 nodes each, and 4 hidden layers of 100 nodes each.

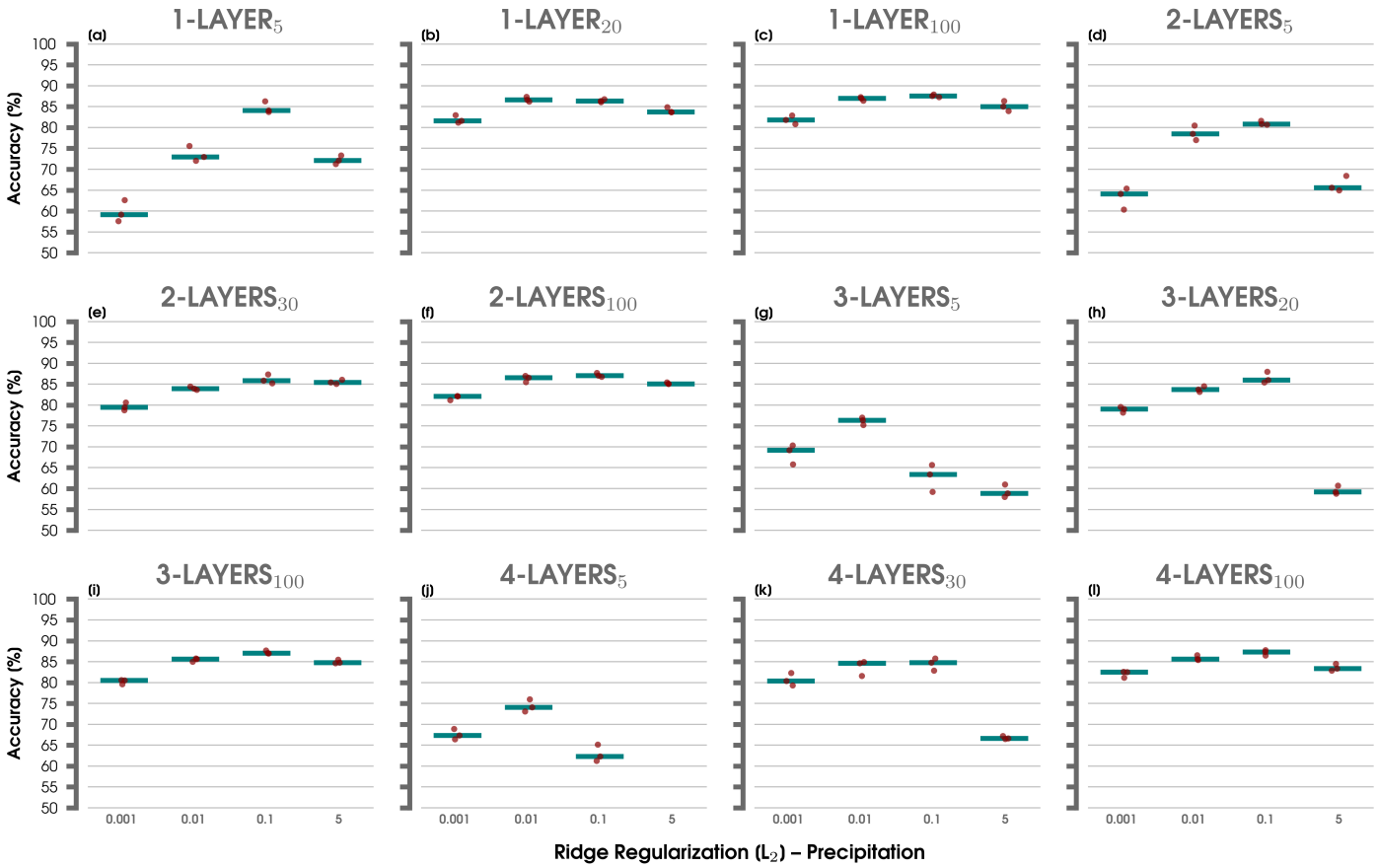

Figure S6. As in Figure S5, but for global maps of annual mean precipitation.

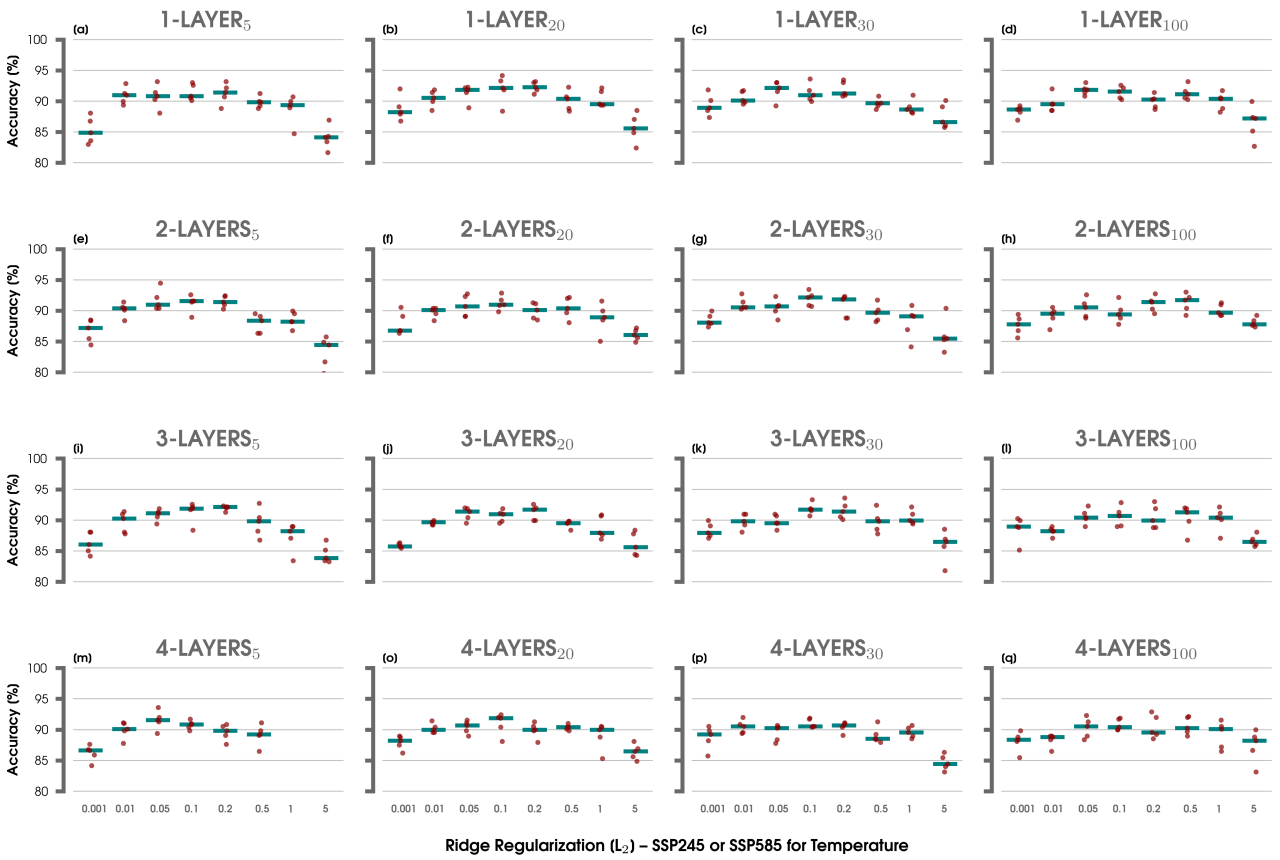

Figure S7. Scores for the total class accuracy of validation data using the binary ANN framework (either SSP2-4.5 or SSP5-8.5) and inputs of global maps of annual mean temperature. (a) The ANN architecture consists of 1 hidden layer and 5 nodes. Eight different L_2 regularization values (0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 1, 5) are compared using this same ANN architecture. Each set of red points is the distribution of accuracies from 5 ANN iterations (randomized combinations of ensemble members used for training, validation, and testing and selection of random initialization seeds). The median accuracy is shown with a blue horizontal line and organized by L_2 parameter. (b-q) As in (a), but for ANN architectures of 1 hidden layer and 20 nodes, 1 hidden layer of 30 nodes, 1 hidden layer of 100 nodes, 2 hidden layers of 5 nodes each, 2 hidden layers of 20 nodes each, 2 hidden layers of 30 nodes each, 2 hidden layers of 100 nodes each, 3 hidden layers of 5 nodes each, 3 hidden layers of 20 nodes each, 3 hidden layers of 30 nodes each, 3 hidden layers of 100 nodes each, 4 hidden layers of 5 nodes each, 4 hidden layers of 20 nodes each, 4 hidden layers of 30 nodes each, and 4 hidden layers of 100 nodes each.

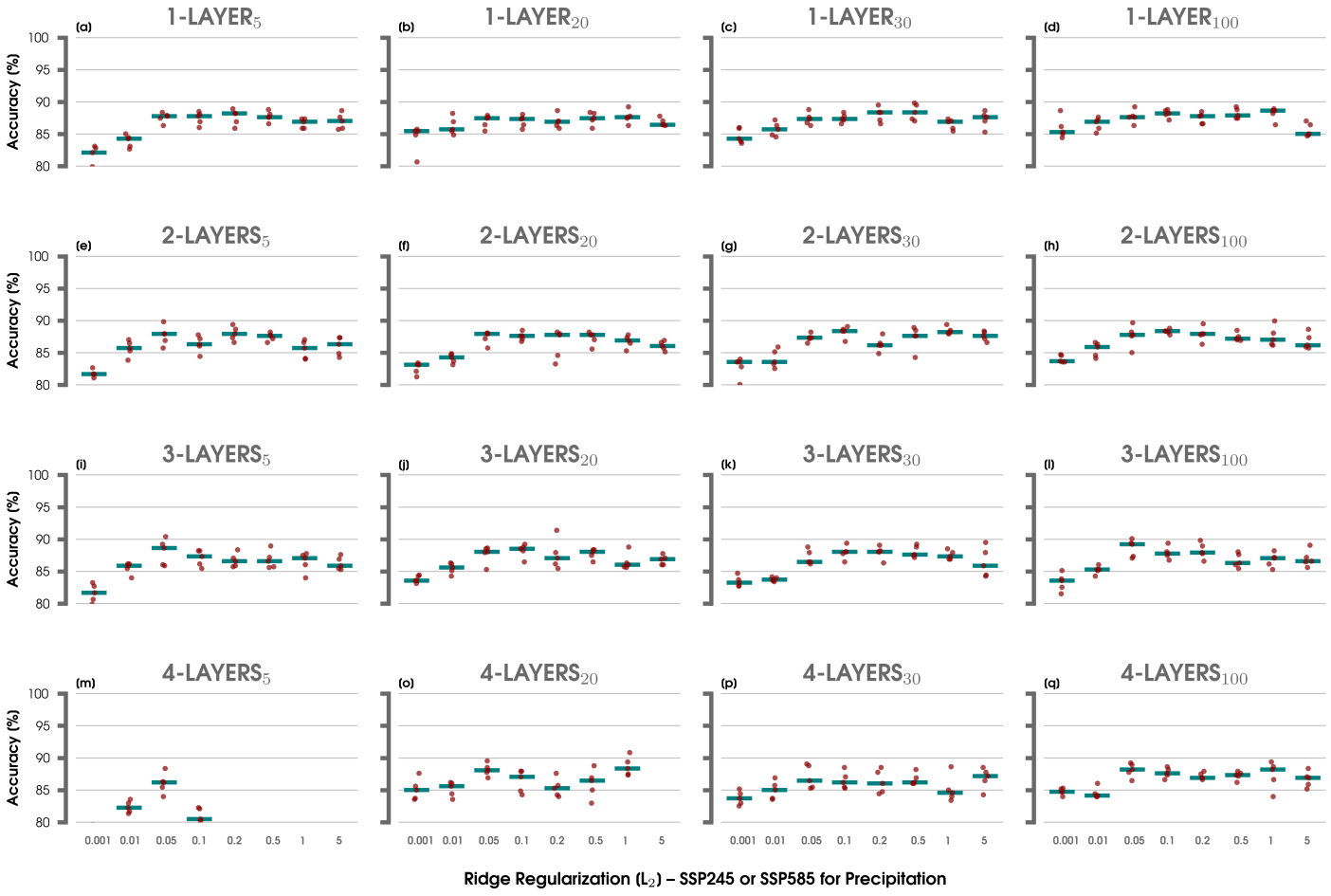

Figure S8. As in Figure S7, but for global maps of annual mean precipitation.

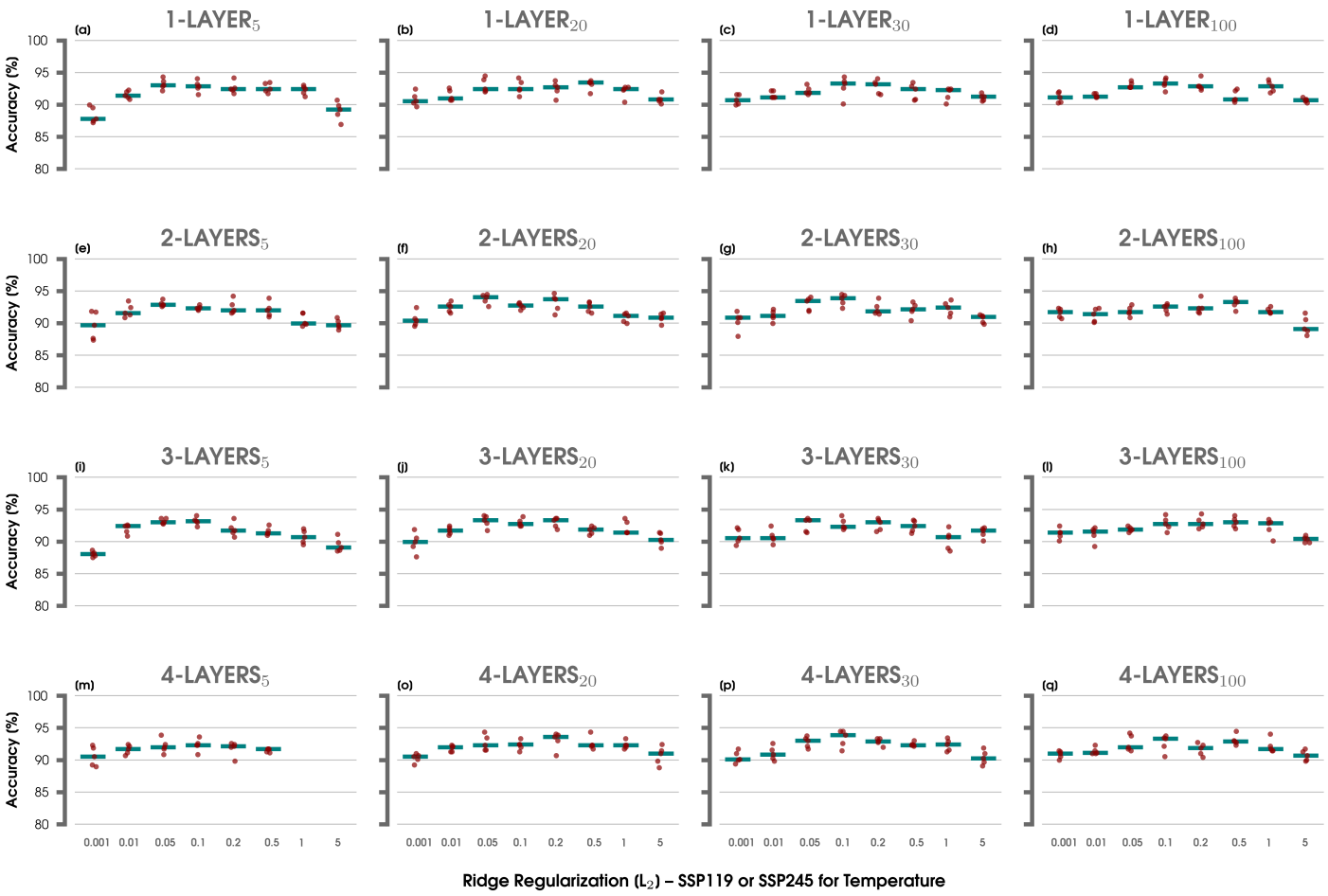

Figure S9. As in Figure S7, but for the binary ANN framework that predicts either SSP1-1.9 or SSP2-4.5 climate scenarios.

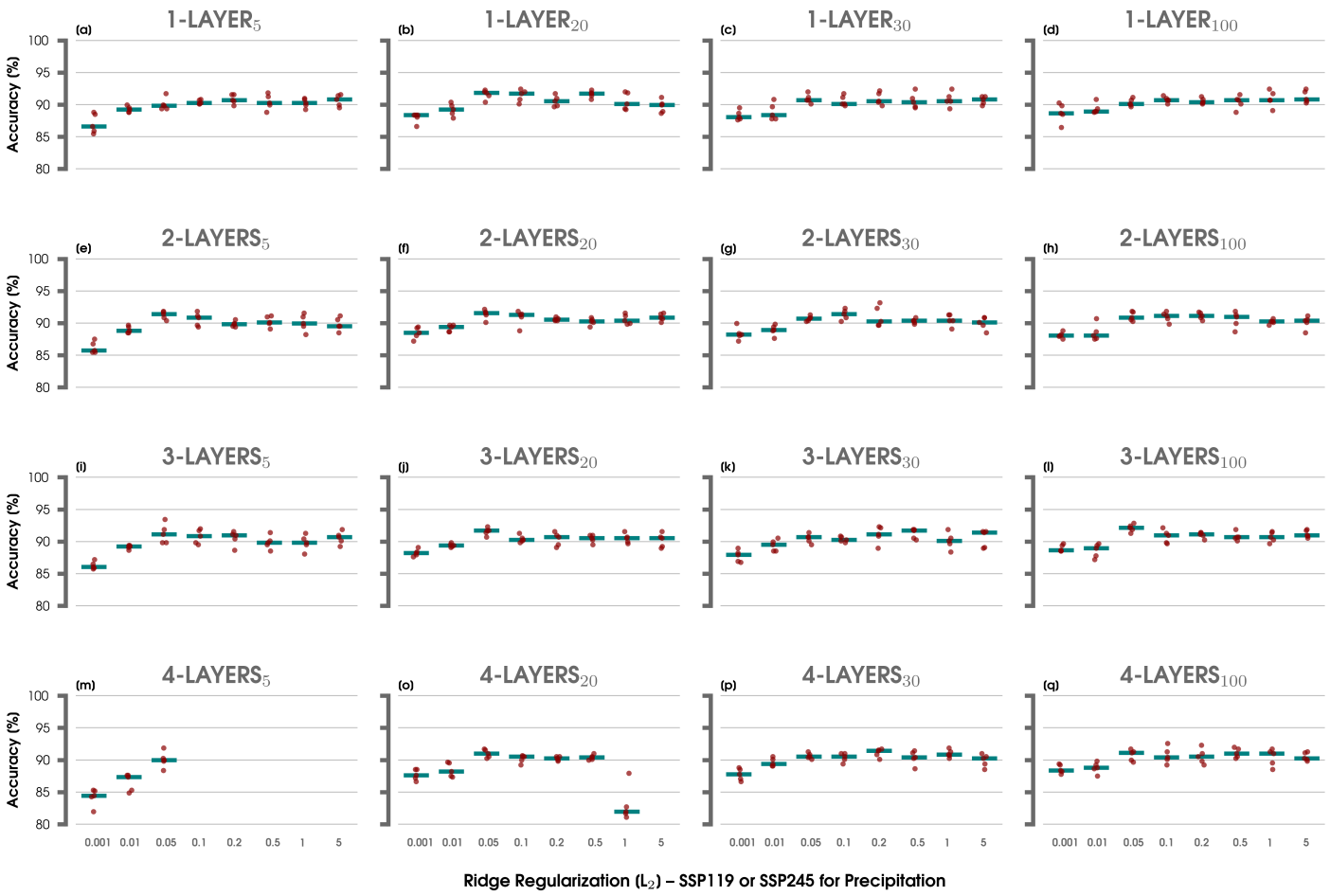

Figure S10. As in Figure S8, but for the binary ANN framework that predicts either SSP1-1.9 or SSP2-4.5 climate scenarios.

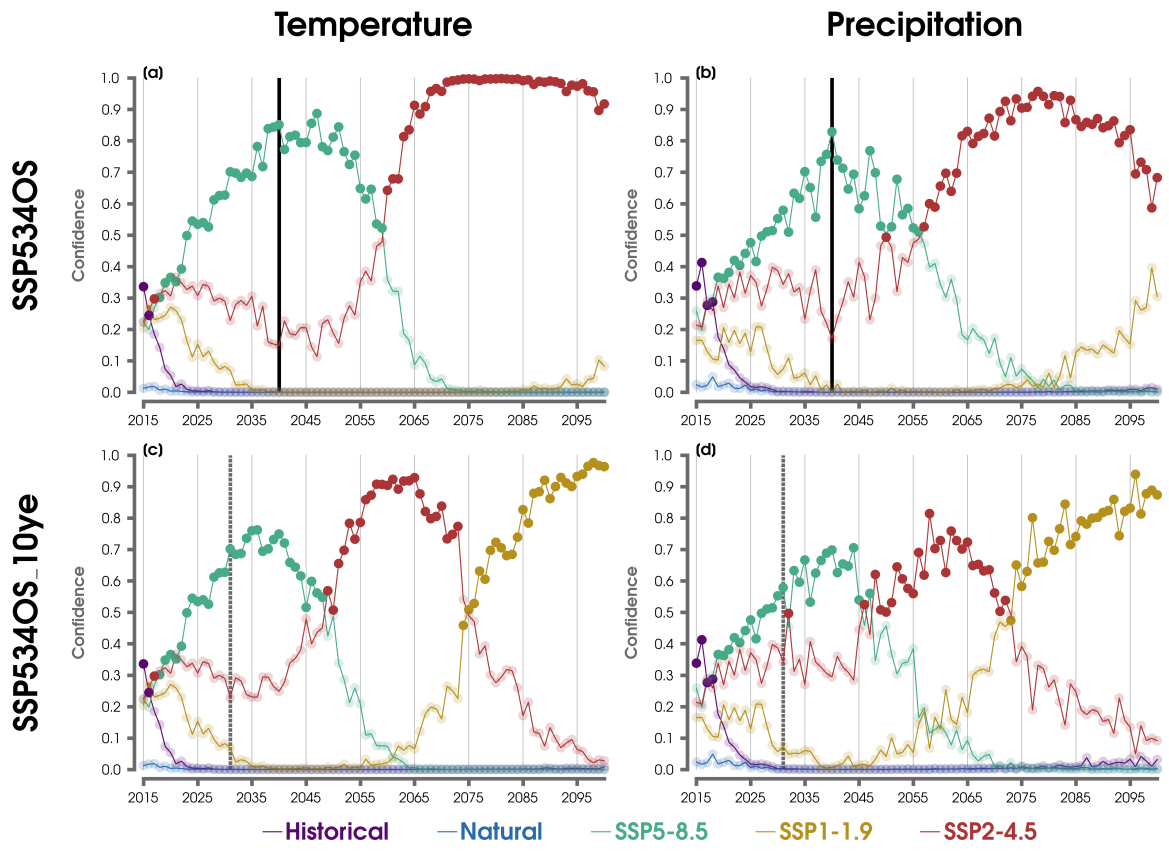

Figure S11. (a) The ensemble mean of the confidence values (after the softmax operator) for the ANN with 5 climate scenario classes (historical scenario (purple line), natural forcing scenario (blue line), SSP5-8.5 (green line), SSP1-1.9 (yellow line), or SSP2-4.5 (red line)) after making inferences on maps of temperature from the SSP5-3.4OS experiment for 2015 to 2100. The vertical black line indicates the start of climate mitigation for this experiment (year 2040). The darker colored lined are denoted for the climate scenario with the highest mean confidence value in each year, and the remaining classes subsequently have a lighter transparency shading. (b) As in (a), but for inputting maps of precipitation. (c) As in (a), but for the SSP5-3.4OS_10ye experiment. The vertical dashed gray line shows the start of mitigation in 2031 for this scenario. Note that the predictions from 2015 to 2030 are the same as the SSP5-3.4OS experiment in panel (a) (see Section 2.2). (d) As in (c), but for precipitation.

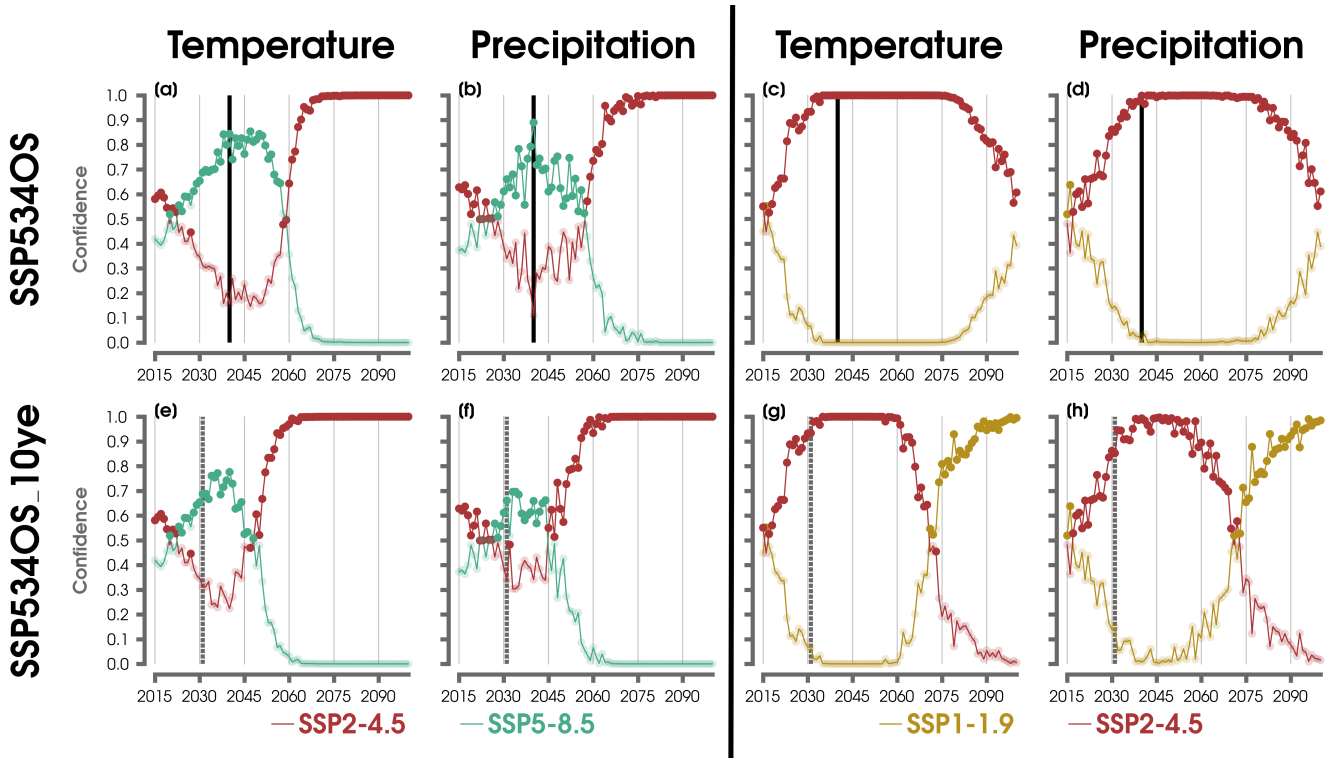

Figure S12. (a) The ensemble mean of network confidence values (after the softmax function) for the ANN with two climate scenario classes (SSP2-4.5 (red line) or SSP5-8.5 (green line)) after making inferences on maps of temperature from the SSP5-3.4OS experiment for 215 to 2100. The vertical black line indicates the start of climate mitigation for this experiment (year 2040). The darker colored lined are denoted for the climate scenario with the highest mean confidence value in each year, and the remain classes subsequently have a lighter transparency shading. (b) As in (a), but for inputting maps of precipitation. (e) As in (a), but for the SSP5-3.4OS_10ye experiment. The vertical dashed gray line shows the start of mitigation in 2031 for this scenario. Note that the predictions from 2015 to 2030 are the same as the SSP5-3.4OS experiment in panel (a). See methods in Section 2. (f) As in (a), but for precipitation. (c,d,g,h) As in (a,b,e,f), but for the binary ANN predicting either SSP1-1.9 (yellow line), or SSP2-4.5 (red line).

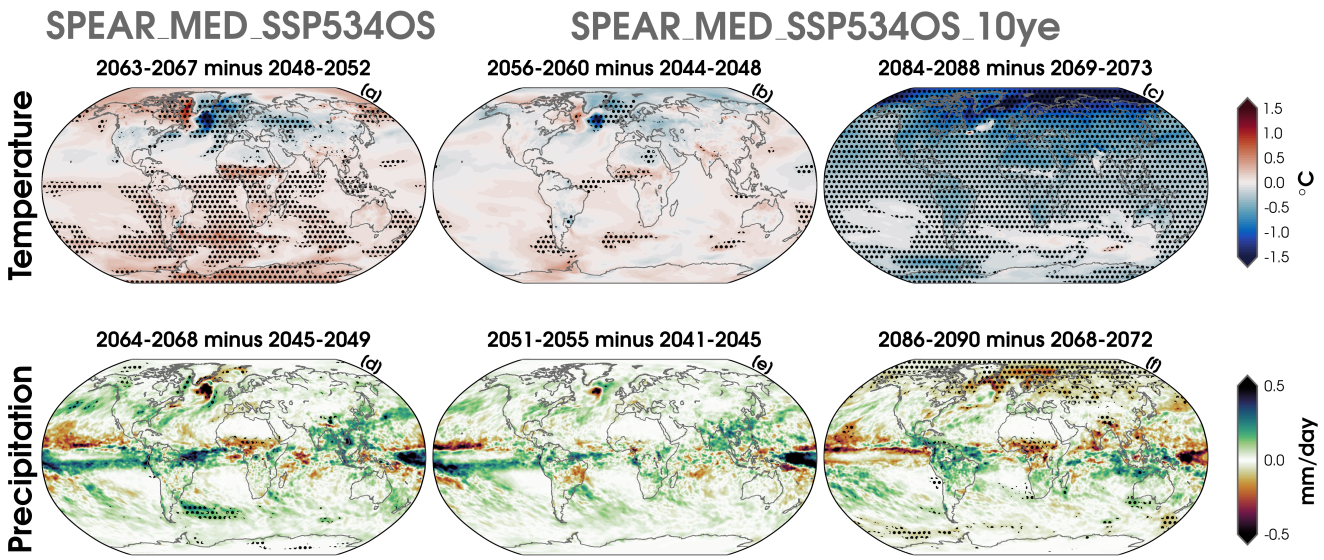

Figure S13. Difference in temperature (°C) for the ensemble mean of SSP5-3.4OS temperature predictions for the five years after the transition period in classifications from SSP8-8.5 to SSP2-4.5 minus the five years before the transition period (i.e., mean of 2063 to 2067 minus the mean of 2048 to 2052). See also Figure 4a. Statistically significant differences are overlaid with black stippling after using a two-sided Student's t test and adjusting for field significance using the false discovery rate (FDR; Benjamini & Hochberg, 1995; Wilks, 2006, 2016) with an FDR-adjusted p value less than 0.05. (b) As in (a), but for the ensemble mean of predictions using SSP5-3.4OS_10ye (i.e., years of 2056 to 2060 minus the mean of 2044 to 2048). See also Figure 4b. (c) As in (b), but for the five years after the transition period in classifications from SSP2-4.5 to SSP1-1.9 subtracted by the five years before this transition period (i.e., mean of 2084 to 2088 minus the mean of 2069 to 2073). (d-f) As in (a-c), but for maps of precipitation (mm/day) using transition periods around the years (a) 2064 to 2068 minus 2045 to 2049, (b) 2051 to 2055 minus 2041 to 2045, and (c) 2086 to 2090 minus 2068 to 2072. See also Figure 4c,d.

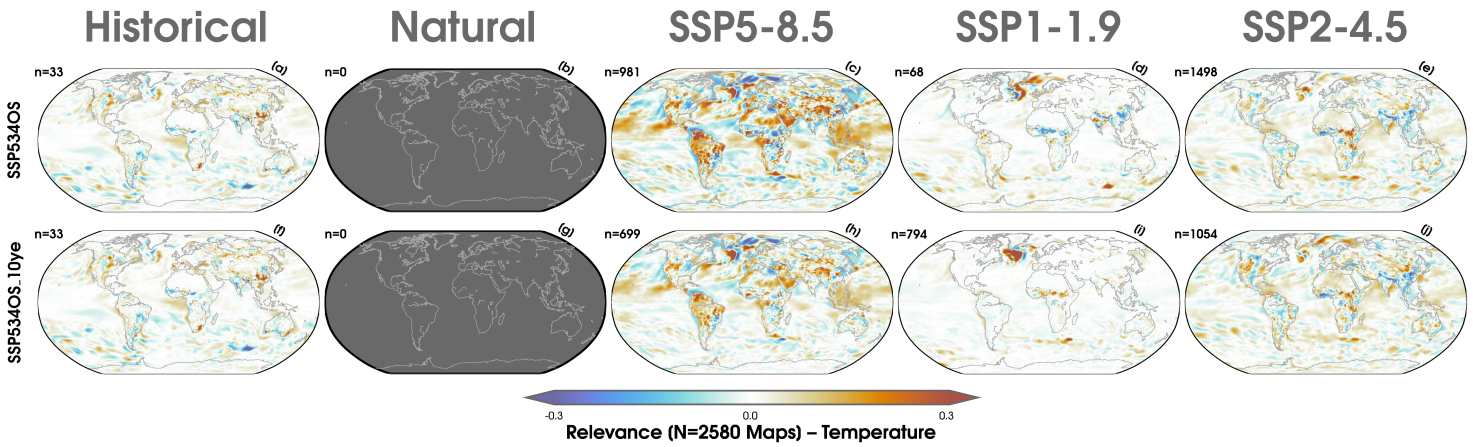

Figure S14. (a-e) Explainability composites using the Integrated Gradients method averaged for each climate scenario prediction using the 5-class ANN after inputting yearly maps of temperature from the SSP5-3.4OS experiment for 2015 to 2100. Thus, there are a total of 2580 possible predictions (N) in the top row (86 years times 30 ensemble members). The number of times each class was predicted (n) is denoted in the upper-left corner of every map composite. Gray shaded maps indicate that this climate scenario was never predicted. Positive areas of relevance (red shading) indicate that the region had a positive contribution to the ANN's prediction (i.e., pushed the network toward the ultimately predicted climate scenario). Negative areas of relevance (blue shading) indicate that the region had a negative contribution to the ANN's prediction (i.e., pushed the network toward predicting one of the other climate scenario classes). (f-j) As in (a-e), but for 30 ensemble members of the SSP5-3.4OS_10ye experiment. Note that the composites for years from 2015 to 2030 are the same as the SP5-3.4OS experiment in (a-e).

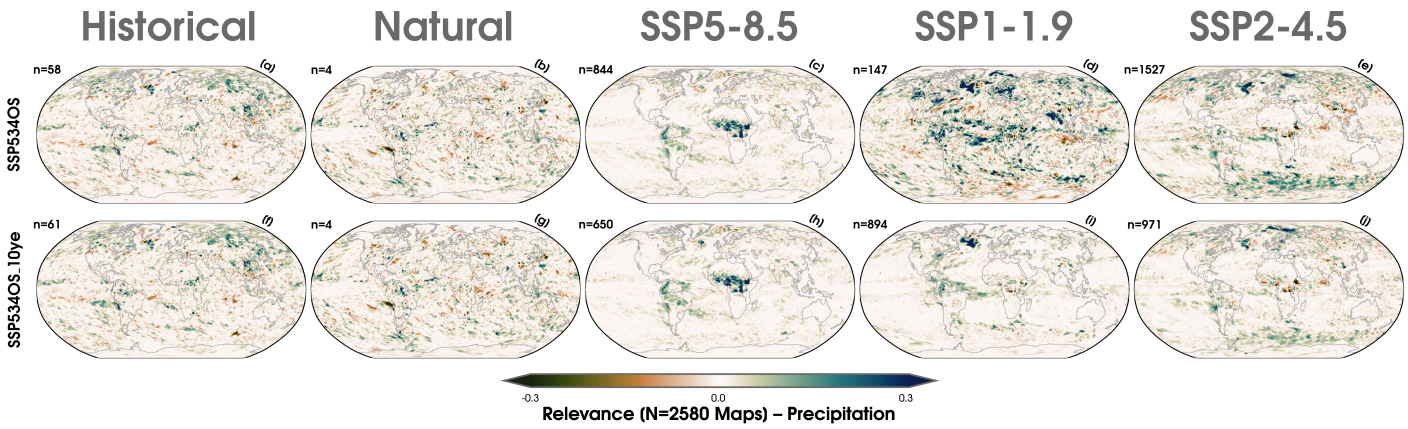

Figure S15. As in Figure S14, but for fields of precipitation.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... Zheng, X. (2016). Tensorflow: A system for large-scale machine learning..

Agarap, A. F. (2018, 3). Deep learning using rectified linear units (relu). *arXiv*. Retrieved from http://arxiv.org/abs/1803.08375

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., ... Kindermans, P. J. (2019). Innvestigate neural networks! *Journal of Machine Learning Research*, *20*.

Anaconda. (2023). *Conda (version 23.1.0) [software].* Retrieved from https://github.com/conda/conda/tree/mainhttps://docs.conda.io/projects/conda/en/23.1.x/index.html

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020, 9). Indicator patterns of forced change learned by an artificial neural network. *Journal of Advances in Modeling Earth Systems*, *12*. Retrieved from https://onlinelibrary.wiley.com/doi/10.1029/2020MS002195 doi: 10.1029/2020MS002195

Basemap. (2022, 10). Retrieved from https://github.com/matplotlib/basemap

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*. doi: 10.1111/j.2517-6161.1995.tb02031.x

Chase, R. J., Harrison, D. R., Burke, A., Lackmann, G. M., & McGovern, A. (2022, 8). A machine learning tutorial for operational meteorology. part i: Traditional machine learning. *Weather and Forecasting*, *37*, 1509-1529. Retrieved from https://journals.ametsoc.org/

view/journals/wefo/37/8/WAF-D-22-0070.1.xml doi: 10.1175/WAF-D-22-0070.1

Chase, R. J., Harrison, D. R., Lackmann, G., & McGovern, A. (2022, 10). A machine learning tutorial for operational meteorology, part ii: Neural networks and deep learning. *arXiv*. Retrieved from https://arxiv.org/abs/2211.00147v1 doi: 10.48550/arxiv.2211.00147

Chollet, F. (2015). Keras: The python deep learning library [software]. *Keras.Io*. Retrieved from https://github.com/keras-team/keras

de Burgh-Day, C. O., & Leeuwenburg, T. (2023, 11). Machine learning for numerical weather and climate modelling: a review. *Geoscientific Model Development*, *16*, 6433-6477. Retrieved from https://gmd.copernicus.org/articles/16/6433/2023/ doi: 10.5194/GMD-16-6433-2023

Delworth, T. L., Cooke, W. F., Adcroft, A., Bushuk, M., Chen, J.-H., Dunne, K. A., ... Zhao, M. (2020, 3). Spear: The next generation gfdl modeling system for seasonal to multidecadal prediction and projection. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001895. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2019MS001895 doi: 10.1029/2019MS001895

Flora, M. L., Potvin, C. K., McGovern, A., & Handler, S. (2023, 11). A machine learning explainability tutorial for atmospheric sciences. *Artificial Intelligence for the Earth Systems*. Retrieved from https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-23-0018.1/AIES-D-23-0018.1.xml doi: 10.1175/AIES-D-23-0018.1

Friedman, J. H. (2012, 7). Fast sparse regression and classification. *International Journal of Forecasting*, *28*, 722-738. doi: 10.1016/j.ijforecast.2012.05.001

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020, 9). Array programming with numpy. *Nature*, *585*, 357. Retrieved from https://doi.org/10.1038/s41586-020-2649-2 doi: 10.1038/s41586-020-2649-2

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012, 7). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv*, 1-18. Retrieved from https://arxiv.org/abs/1207.0580v1

Hunter, J. D. (2007, 5). Matplotlib: A 2d graphics environment. *Computing in Science and Engineering*, *9*, 99-104. doi: 10.1109/MCSE.2007.55

Irving, D. (2016, 7). A minimum standard for publishing computational results in the weather and climate sciences. *Bulletin of the American Meteorological Society*, *97*, 1149-1158. Retrieved from https://journals.ametsoc.org/view/journals/bams/97/7/bams-d-15-00010.1.xml doi: 10.1175/BAMS-D-15-00010.1

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*. doi: 10.1186/s40537-019-0192-5

Labe, Z. M., & Barnes, E. A. (2022, 7). Comparison of climate model large ensembles with observations in the arctic using simple neural networks. *Earth and Space Science*, *9*, e2022EA002348. Retrieved from https://doi.org/10.1029/2022EA002348 doi: 10.1029/2022EA002348

Labe, Z. M., Delworth, T. L., Johnson, N. C., & Cooke, W. F. (2023). *zmlabe detectmitigate [software].* Retrieved from https://zenodo.org/records/10083257

Labe, Z. M., Johnson, N. C., & Delworth, T. L. (2024, 2). Changes in united states summer temperatures revealed by explainable neural networks. *Earth's Future*, *12*, e2023EF003981. Re-

trieved from https://onlinelibrary.wiley.com/doi/abs/10.1029/2023EF003981 doi: 10.1029/2023EF003981

Martin, Z. K., Barnes, E. A., & Maloney, E. (2022). Using simple, explainable neural networks to predict the madden-julian oscillation. *Journal of Advances in Modeling Earth Systems*, *14*. doi: 10.1029/2021MS002774

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence o(1/k2). *Doklady AN USSR*, *269*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*.

Rader, J. K., Barnes, E. A., Ebert-Uphoff, I., & Anderson, C. (2022, 7). Detection of forced change within combined climate fields using explainable neural networks. *Journal of Advances in Modeling Earth Systems*, *14*, e2021MS002941. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1029/2021MS002941 doi: 10.1029/2021MS002941

Rossum, G. V., & Drake, F. L. (2009). *Python 3 reference manual [software]*. CreateSpace.

Ruder, S. (2016, 9). An overview of gradient descent optimization algorithms. *arXiv*. Retrieved from http://arxiv.org/abs/1609.04747

Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (global edition). *Artificial Intelligence: A Modern Approach*. Retrieved from https://aima.cs.berkeley.edu/

Schulzweida, U. (2019, 2). Cdo user guide [software]. *Zenodo*. Retrieved from https://zenodo.org/record/2558193 doi: 10.5281/ZENODO.2558193

Sippel, S., Meinshausen, N., Merrifield, A., Lehner, F., Pendergrass, A. G., Fischer, E., & Knutti, R. (2019). Uncovering the forced climate response from a single ensemble member using statistical learning. *Journal of Climate*, *32*. doi: 10.1175/JCLI-D-18-0882.1

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*.

Thyng, K., Greene, C., Hetland, R., Zimmerle, H., & DiMarco, S. (2016, 9). True colors of oceanography: Guidelines for effective and accurate colormap selection. *Oceanography*, *29*, 9-13. Retrieved from https://tos.org/oceanography/article/true-colors-of-oceanography-guidelines-for-effective-and-accurate-colormap doi: 10.5670/oceanog.2016.66

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020, 9). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*. Retrieved from https://onlinelibrary.wiley.com/doi/10.1029/2019MS002002 doi: 10.1029/2019MS002002

Toms, B. A., Barnes, E. A., & Hurrell, J. W. (2021, 6). Assessing decadal predictability in an earth-system model using explainable neural networks. *Geophysical Research Letters*, *48*, e2021GL093842. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021GL093842 doi: 10.1029/2021GL093842

van der Velden, E. (2020, 2). Cmasher: Scientific colormaps for making accessible, informative and 'cmashing' plots. *Journal of Open Source Software*, *5*, 2004. Retrieved from `https://joss.theoj.org/papers/10.21105/joss.02004` doi: 10.21105/JOSS.02004

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... Vázquez-Baeza, Y. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, *17*. doi: 10.1038/s41592-019-0686-2

Wilks, D. S. (2006, 9). On "field significance" and the false discovery rate. *Journal of Applied Meteorology and Climatology*, *45*, 1181-1189. Retrieved from `http://journals.ametsoc.org/doi/abs/10.1175/JAM2404.1` doi: 10.1175/JAM2404.1

Wilks, D. S. (2016, 12). "the stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bulletin of the American Meteorological Society*, *97*, 2263-2273. Retrieved from `http://journals.ametsoc.org/doi/10.1175/BAMS-D-15-00267.1` doi: 10.1175/BAMS-D-15-00267.1

Zender, C. S. (2008). Analysis of self-describing gridded geoscience data with netcdf operators (nco). *Environmental Modelling and Software*, *23*. doi: 10.1016/j.envsoft.2008.03.004