GEOPHYSICAL RESEARCH LETTERS

Supporting Information for "Quantifying the Effect of Climate Change on Midlatitude Subseasonal Prediction Skill Provided by the Tropics"

Kirsten J. Mayer [1]and Elizabeth A. Barnes [1]

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

Contents of this file

1. Text S1 to S6
2. Figures S1 to S9

Corresponding author: Kirsten J. Mayer, kirsten.j.mayer@gmail.com

March 10, 2022, 9:39pm

Overview In the supporting information, we provide details on the robustness of our results to changes in the number of training ensemble members and variations in the neural network architecture and hyperparameters. The sensitivity of the results to the choice of members used for validation and testing are examined as well. We also include information about the bootstrapping analysis, and provide additional information on the neural network explainability technique and the seasonal filtering results, along with the corresponding figures. Confidence versus accuracy diagrams for seasonal predictions are also included.

Text S1: Network Sensitivity to the Number of Training Members To test whether 8 training ensemble members (members #1-8) are sufficient for this analysis, 100 neural networks are trained with different sized training sets, starting with only 1 member and increasing to 8 members iteratively (moving left to right in Figure S1). Figure S1a,b includes the accuracies of the testing member (#10) for all predictions and Figure S1c,d includes the accuracies for the corresponding 20% most confident predictions. In the North Pacific (Figure S1a,c), the skill for the historical *and* future periods plateaus at about 5 training members for both all and the most confident predictions. The North Atlantic (Figure S1b,d) shows more skill variability with training size, but generally maintains the same range of skill for each period when 3 or more ensemble members are used. The skill variability in the North Atlantic may be related to multidecadal variability in the ensemble members (e.g. Simpson et al. 2018). While the Pacific is also impacted by longer timescales, it is more prominently impacted by decadal variability instead (e.g.

Cassou et al. 2018). Therefore, if a lower multidecadal predictability state in the North Atlantic is dominating the majority of the training years for a given ensemble member, this could ultimately impact how well the network can learn.

Text S2: Network Architecture and Sensitivity to Hyperparameters The neural network used in this study consists of two layers of 128 and 8 nodes, respectively. The rectified linear unit "relu" activation function is applied to the hidden layers. Categorical cross entropy is used for the loss function and the batch size is set to 256 samples. Adam (Kingma and Ba 2014) is used as an optimizer with a learning rate of 0.001. We reduce the learning rate exponentially by $e^{-0.1}$ for each epoch after 10 epochs to assist the network in minimizing the loss. To reduce overfitting on the training data, ridge regression ($L_2 = 1.0$; Friedman 2012) is applied to the first hidden layer and early stopping is implemented. Ridge regression is used to direct the network to account for spatial autocorrelation within the input field (tropical precipitation). Early stopping monitors the validation prediction accuracy, so when the validation prediction accuracy does not increase for more than 20 epochs, the network stops training and reverts back to the network weights from 20 epochs before. Otherwise, the network concludes training at 100 epochs. We find that a patience of 20 epochs is useful for this problem to reduce overfitting since the network never trains for the full 100 epochs when early stopping is implemented. The output layer consists of 2 nodes and uses the softmax activation function. The softmax activation function converts the output into two numbers which sum to 1 and can be interpreted as the likelihood of a given prediction, referred to as "model confidence". A more detailed description of network training for a similar artificial neural network is provided in the supporting information of Mayer and Barnes (2021) and additional information on artificial neural networks in general can be found in Nielsen (2015) and Goodfellow et al. (2016).

To test the sensitivity of our conclusions to the network architecture and hyperparameter choice, the learning rate, ridge regression parameter, nodes per layer and the number or layers were all varied and the validation accuracy compared (Figure S3-S4). Figure S3 (S4) shows results for the North Pacific (Atlantic), where the validation member #9 accuracy of 10 trained models with different initial weights are shown for each hyperparameter variation and time period. The network hyperparameters and architecture for this analysis were ultimately chosen because it has some of the highest validation skill for both the historical and the future time period in the North Atlantic, but also performs well in the North Pacific (Figure S3-S4). We initially focus on the skill of the network in the North Atlantic because it is more difficult for the network to predict than the North Pacific. We also see that slight variations of these hyperparameters show similar skill to the network chosen.

We note that for the North Pacific hyperparameter sweep (Figure S3), validation member #9 shows a decrease in skill for all predictions between the historical and the future period which is not seen with the testing member (Figure 2a). We believe that the decrease in skill in the validation data between the two time periods is likely a result of slight overfitting of the validation during the historical time period due to its use for early stopping (not shown).

Text S3: Accuracy Bootstrapping Analysis Due to the computational costs of training 100 networks for each grid point in the Northern Hemisphere, 10 neural networks are trained for each location instead. To check whether these changes identified in the North Pacific and North Atlantic with 100 networks can be seen using only 10 networks, and to provide a reference of the magnitude of significant skill changes for the other grid points in Figure 3, we used the 100 models trained for both the North Pacific and North Atlantic to conduct a bootstrapping analysis.

For each location, from the 100 models trained, 10 models are randomly selected and the top three networks are chosen, defined using the three highest 20% most confident validation accuracies. The mean of the 20% most confident testing accuracies is then calculated for these three models, identical to the method used to calculate the testing accuracy for each grid point in Figure 3. This is repeated 1000 times for each time period with the resulting distributions plotted in Figure S5. For each region, we find that the direction of change in skill for 10 networks is the same as that for 100 networks, and the future accuracy is statistically different than the historical time period using a one-sided Welch's t-test (Welch 1947) at a 95% confidence level (p-value < 0.0001). For the North Pacific (Atlantic), we test whether the future period is statistically less (greater) than the historical. Therefore, we find that 10 networks is sufficient for identifying these subseasonal prediction skill changes.

Text S4: Layer-wise Relevance Propagation Layer-wise Relevance Propagation (LRP; Bach et al. 2015; Montavon et al. 2019) is a neural network (attribution) explainability technique that creates a heatmap of the estimated “relevance” of the input for a given prediction. Here, we use the LRP_z rule, which has been shown to perform well for specific geoscience applications (Mamalakis et al. 2021). For an individual prediction, LRP_z backpropagates relevance information from an output node through the network to create a heatmap of the estimated regions of the input that the network found most relevant for its prediction, where positive (negative) relevance denotes positive (negative) contributions to the final output. The softmax activation function is removed before back propagation and the heatmap for each prediction is normalized by dividing by the absolute maximum relevance value in that map. Figure S6 shows the average of the correct and confident predictions' heatmaps for an example neural network. We find that other networks produce similar LRP maps to this example.

In both the North Pacific and North Atlantic, the differences in relevance between the two time periods are most evident in the equatorial Pacific. In the North Pacific, the relevance of this region is generally reduced and the focus shifts westward in the future period. In the North Atlantic, the relevance of this region increases in the future, mainly over the western equatorial Pacific. The change in relevance of the equatorial Pacific corresponds with the change in prediction skill for both regions and suggests that the network's changing focus in the equatorial Pacific is related to the changes in subseasonal prediction skill.

Text S5: Seasonal Filtering Analysis To further examine the possible role of seasonal variability influencing future subseasonal prediction skill, we task the neural network to predict the sign of z500 anomalies using only z500 variability on *shorter* than 60 day (subseasonal) timescales. The z500 anomalies are filtered by removing the forward 60 day running mean. This filtering is used to direct the network to focus on tropical precipitation specifically related to midlatitude subseasonal variability in the z500 anomalies. Thus, changes to prediction skill between the two time periods are a result of changes in the ability to specifically predict midlatitude variability with shorter than 60 day periods. We use this approach to identify if the changes in skill are mainly related to changes in midlatitude subseasonal variability or if the changes are related purely to changes in seasonal variability, or a combination of the two. We note that this filtering analysis could have been conducted for the main paper, however, we wanted to retain seasonal variability information to identify possible skill changes that could be seen in a typical subseasonal forecast. For each time period and region, 100 networks are again trained and their accuracies across model confidence thresholds are computed (Figure S7).

Overall, the removal of seasonal variability reduces the information the network can use for its predictions, so the filtering leads to a decrease in skill for both time periods compared to the unfiltered predictand. In the North Pacific (Figure S7a-b), there is virtually no difference between the historical and future period when seasonal variability is removed because the historical skill decreases more than the future skill, resulting in similar accuracies across model confidence thresholds. This implies that the historical period relies more on seasonal variability for subseasonal prediction than the future period, consistent with the LRP analysis. The lack of skill change between the two time periods also implies that the *change* in subseasonal prediction skill seen in the unfiltered analysis is related to midlatitude seasonal variability instead of subseasonal. In the North Atlantic, we see that the future period still has higher prediction skill compared to the historical, although, the overall skill for both time periods is reduced (Figure S7c-d). A reduction in skill for both time periods is expected because the LRP maps suggest that both time periods rely, at least partially, on the ENSO regions for the predictions. However, even with midlatitude seasonal variability removed from z500, there is still an increase in skill from the historical to the future time period over the North Atlantic, suggesting there are other shorter timescale variability contributors to the increase in midlatitude subseasonal prediction skill in the future.

Text S6: Seasonal Predictions To check whether the neural networks are using more than seasonal information for their predictions, we train 100 neural networks for East Asia, the North Pacific and the North Atlantic for leads of 60 and 90 days (Figure S8-S9). East Asia is also analyzed here because of the unexpected increase in subseasonal prediction skill in the future (Figure 3). By training the networks at seasonal lead times, we can assess whether the prediction skill in each region *only* comes from seasonal variability. In other words, if only seasonal variability is contributing to the prediction skill, there should be no difference in skill between a lead of 21 days and a lead of 60 or 90 days.

We see that in East Asia (Figure S8a-b, S9a-b) the neural networks have similar skill whether trained at a lead of 21, 60 or 90 days. This suggests that the skill seen at 21 days is likely skill from seasonal variability alone. On the other hand, the North Pacific (Figure S8c-d, S9c-d) and the North Atlantic (Figure S8e-f, S9e-f) both show higher skill at a lead of 21 days, suggesting that in these regions the neural network is using more than seasonal variability for its predictions. Lastly, Figures S8 and S9 demonstrate that the *changes* in skill between the historical and future periods at a lead of 21 days (Figure 2), are similar to those for seasonal lead predictions, particularly in the North Pacific. In the North Atlantic, the change in skill is larger for the seasonal lead predictions than the 21 day lead. This implies that the networks for the 21 day lead prediction use sources of predictability other than seasonal variability to make predictions, ultimately impacting how much the skill changes between time periods. Overall, this analysis again suggests that seasonal variability is playing a role in the changes to subseasonal prediction skill, but the magnitude of the seasonal influence varies by region.

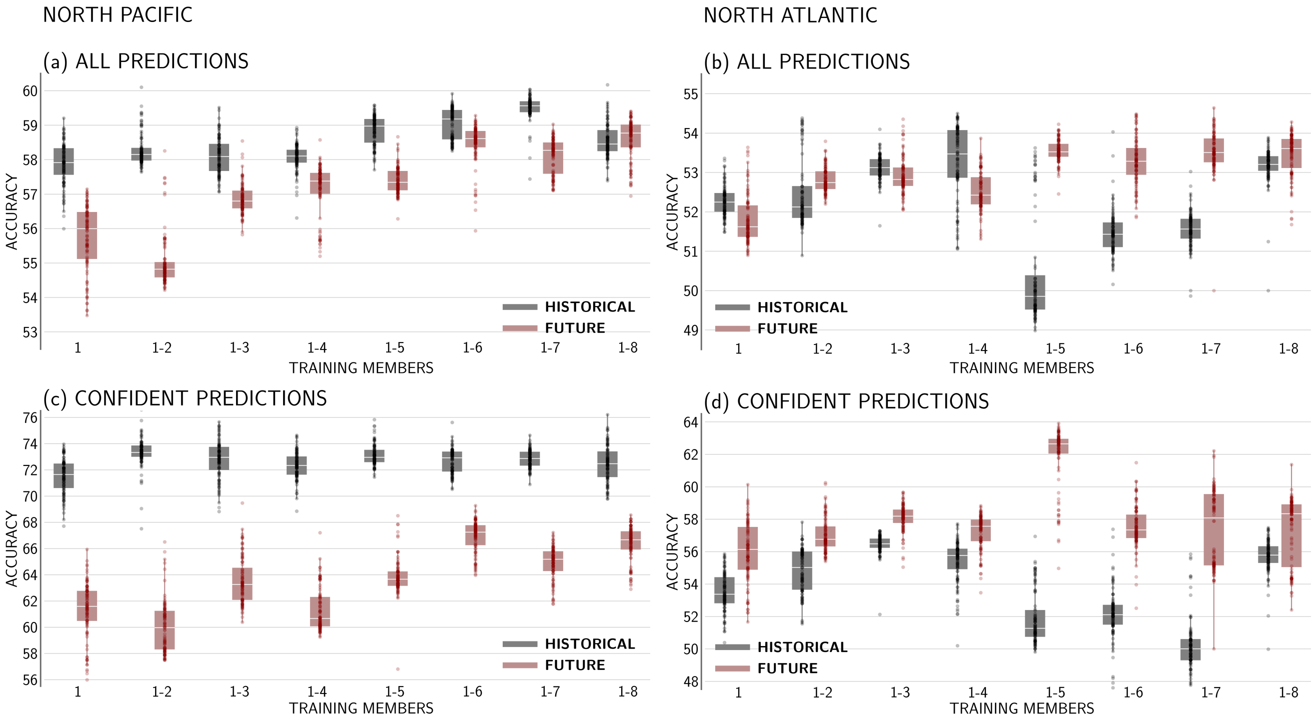

Figure S1. Box and whisker plots of (a,b) all prediction and (c,d) the 20% most confident prediction accuracies for testing ensemble member #10 for the (a,c) North Pacific and (b,d) North Atlantic using increasing numbers of ensemble members for training. Training members #1-8 are used for the main analysis. The black (red) denotes the historical (future) period and the x-axis are the members used to train. The dots indicate individual accuracy for each of the 100 models trained. The white line across each box is the median of the models and the edges of the boxes are the 25th and 75th percentiles.

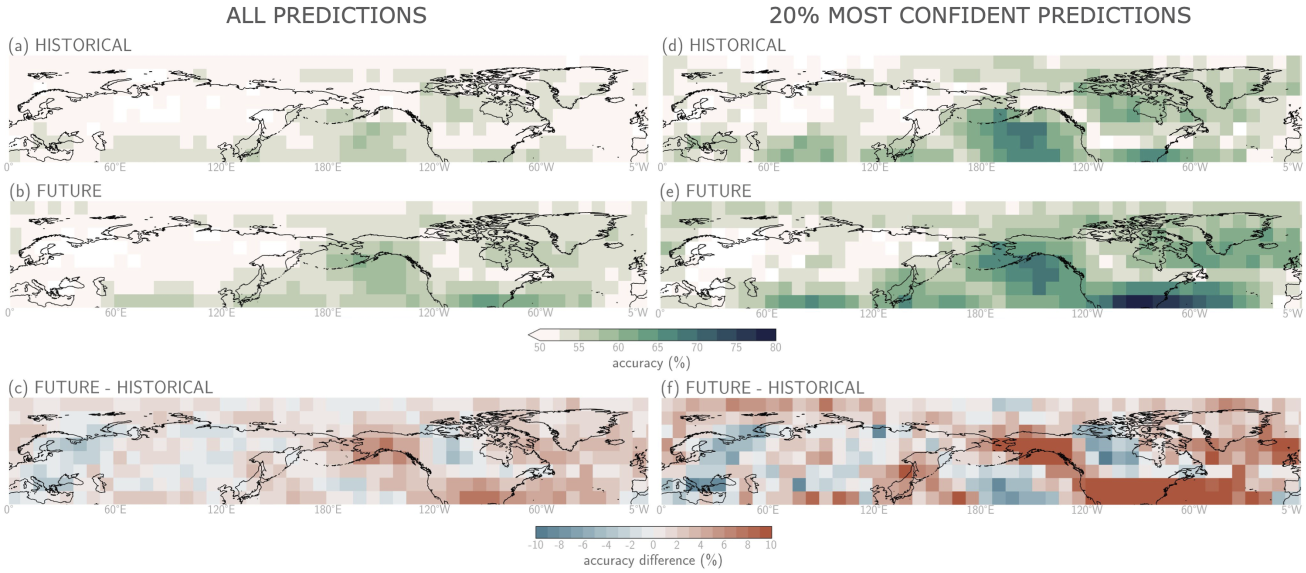

Figure S2. As in main text Figure 3, but with ensemble members #3-10 for training, member #2 for validation and member #1 for testing.

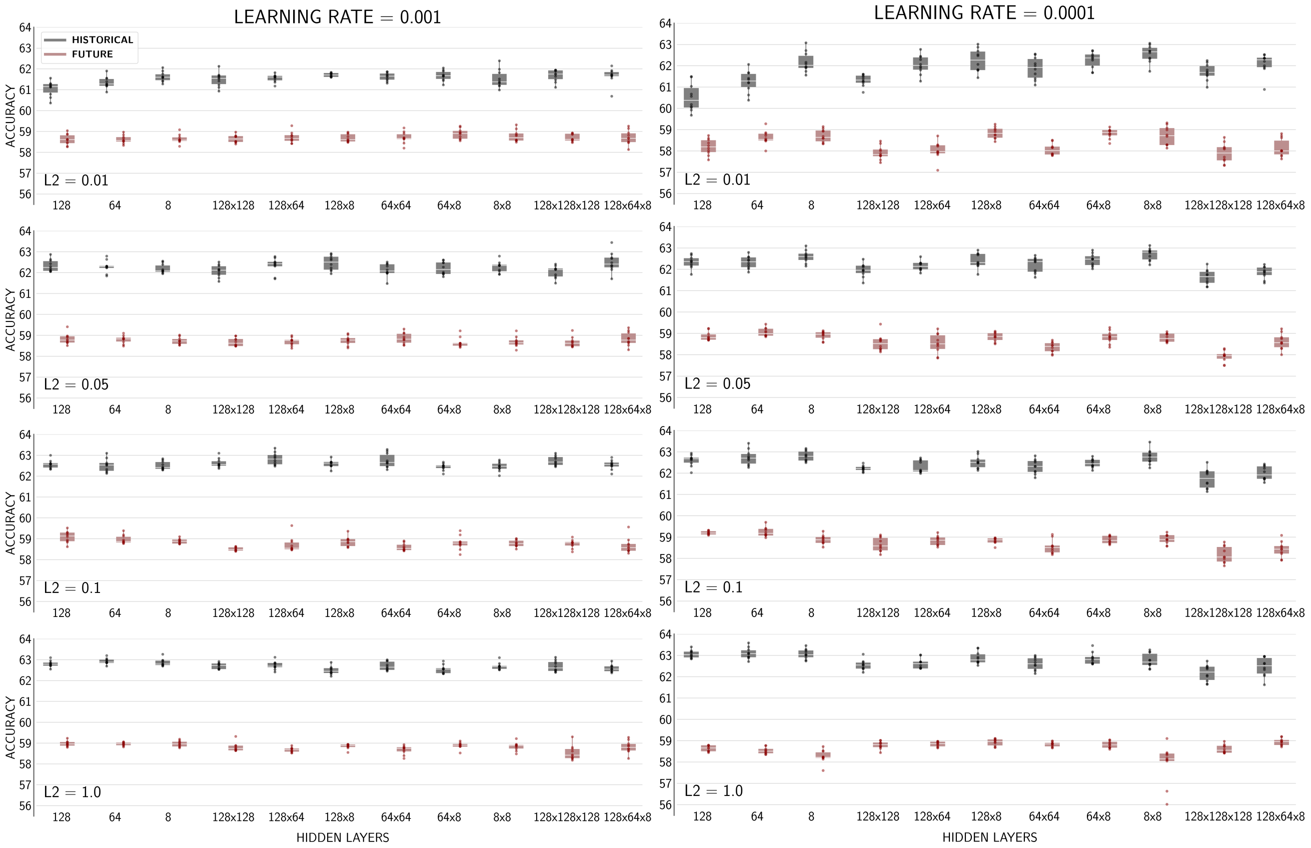

Figure S3. Validation (member #9) box and whisker plots of accuracies for 10 trained models in the North Pacific for variations combinations of the learning rate, ridge regression (L2), nodes per layer, and number of layers. Networks accuracies for a learning rate of 0.001 (0.0001) are in the left (right) column. Ridge regression values (denoted in the bottom left of each figure) increase from top to bottom and the network depth increases from left to right, where the number(s) represent the number of nodes per layer.

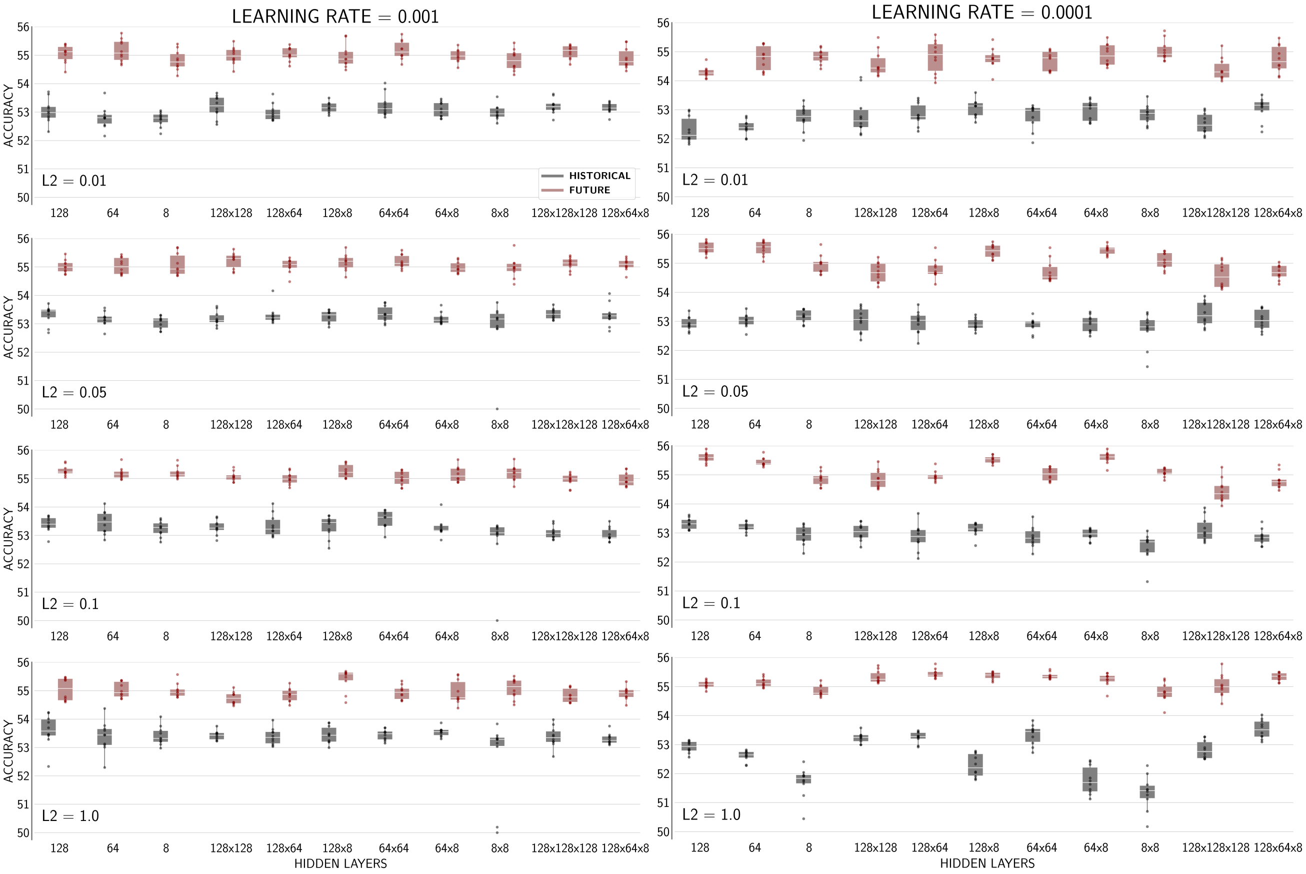

Figure S4. As in Figure S2, but for the North Atlantic.

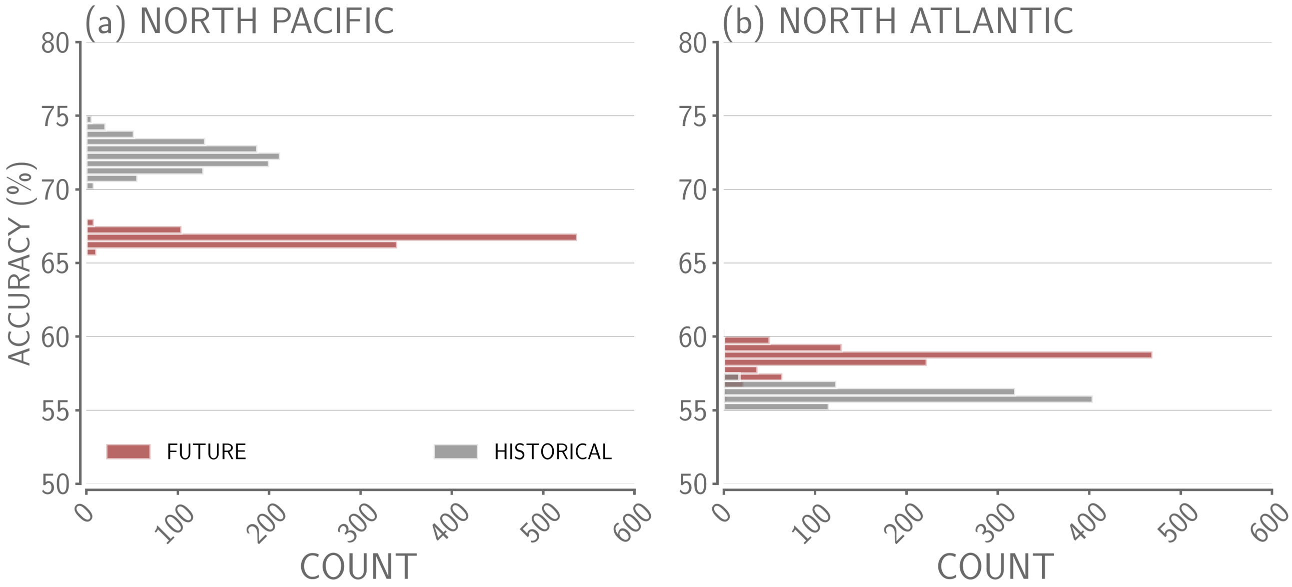

Figure S5. Histograms of bootstrapped top 3 models' mean 20% most confident testing accuracies with a bin size of 0.5% for (a) the North Pacific and (b) the North Atlantic, where grey and red refer to the historical and future, respectively.

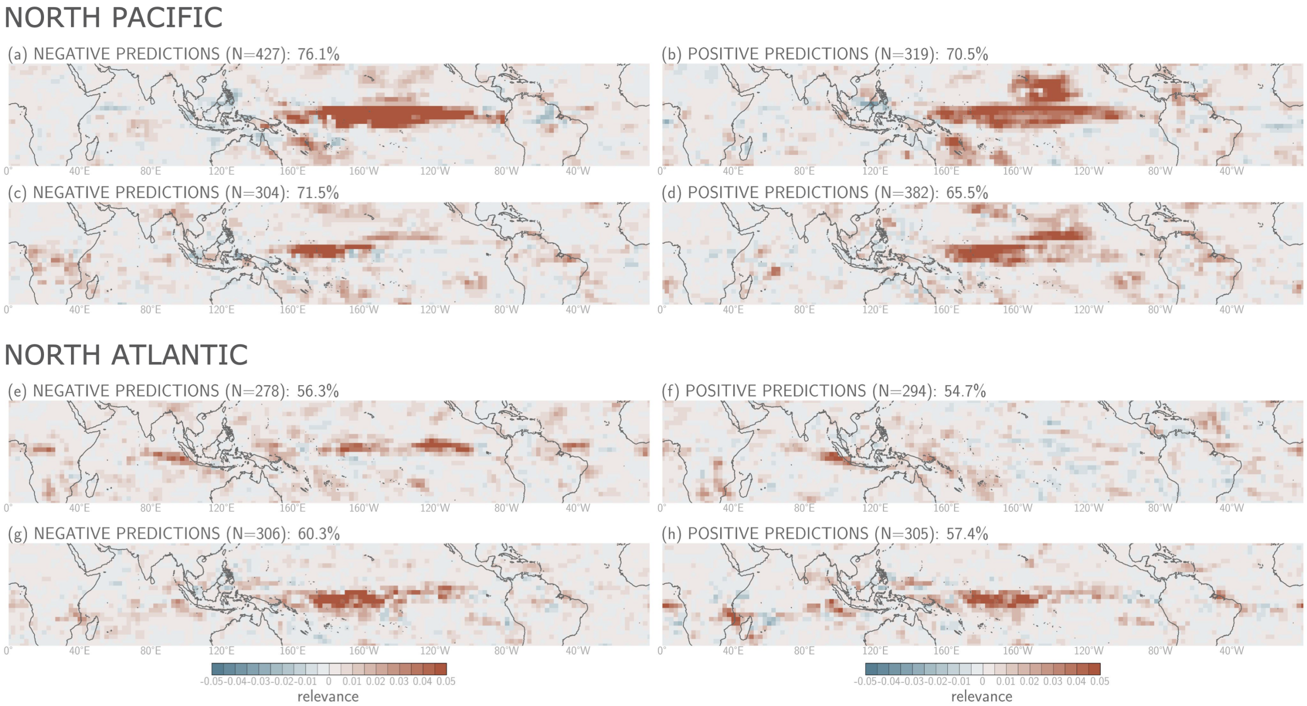
NORTH PACIFIC
(a) NEGATIVE PREDICTIONS (N=427): 76.1%
(b) POSITIVE PREDICTIONS (N=319): 70.5%
(c) NEGATIVE PREDICTIONS (N=304): 71.5%
(d) POSITIVE PREDICTIONS (N=382): 65.5%
NORTH ATLANTIC
(e) NEGATIVE PREDICTIONS (N=278): 56.3%
(f) POSITIVE PREDICTIONS (N=294): 54.7%
(g) NEGATIVE PREDICTIONS (N=306): 60.3%
(h) POSITIVE PREDICTIONS (N=305): 57.4%
0°
40°E
80°E
120°E
160°E
160°W
120°W
80°W
40°W
-0.05 -0.04 -0.03 -0.02 -0.01 0 0.01 0.02 0.03 0.04 0.05
relevance

Figure S6. Example average layer-wise relevance plots for the 20% most confident and correct predictions in the North Pacific (a-d) and the North Atlantic (e-h). The top two panels for each locations (a-b, e-f) are the historical period and the bottom two panels for each location (c-d, g-h) are the future period. The left column includes heatmaps for the negative predictions and the right column includes heatmaps for the positive predictions. Red (blue) colors indicate the location had a positive (negative) contribution to the correct prediction. The percentage at the top of each panel is the conditional accuracy for each sign prediction and 'N' is the number of samples in each average.

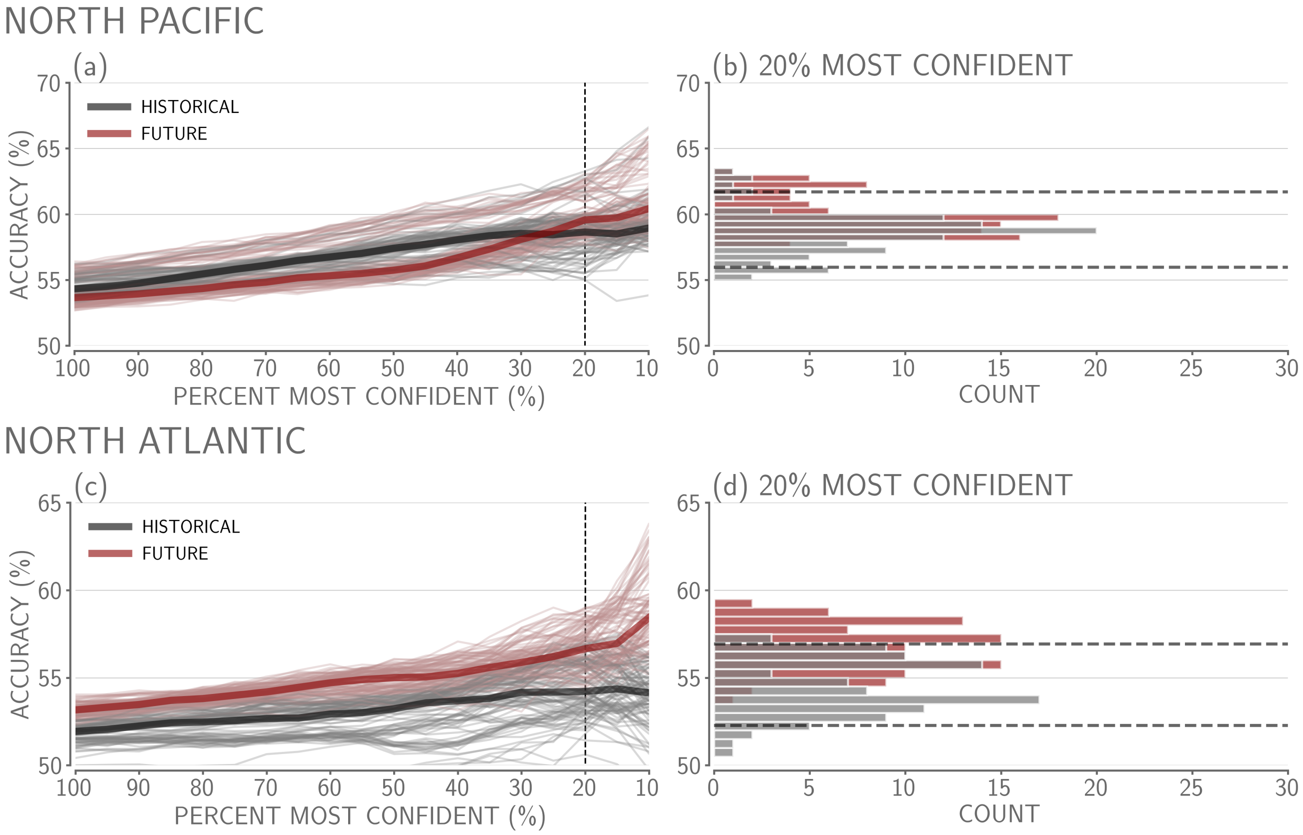

Figure S7. As in Figure 2 in the main text, but with 60+ day z500 anomaly variability removed from the predictand.

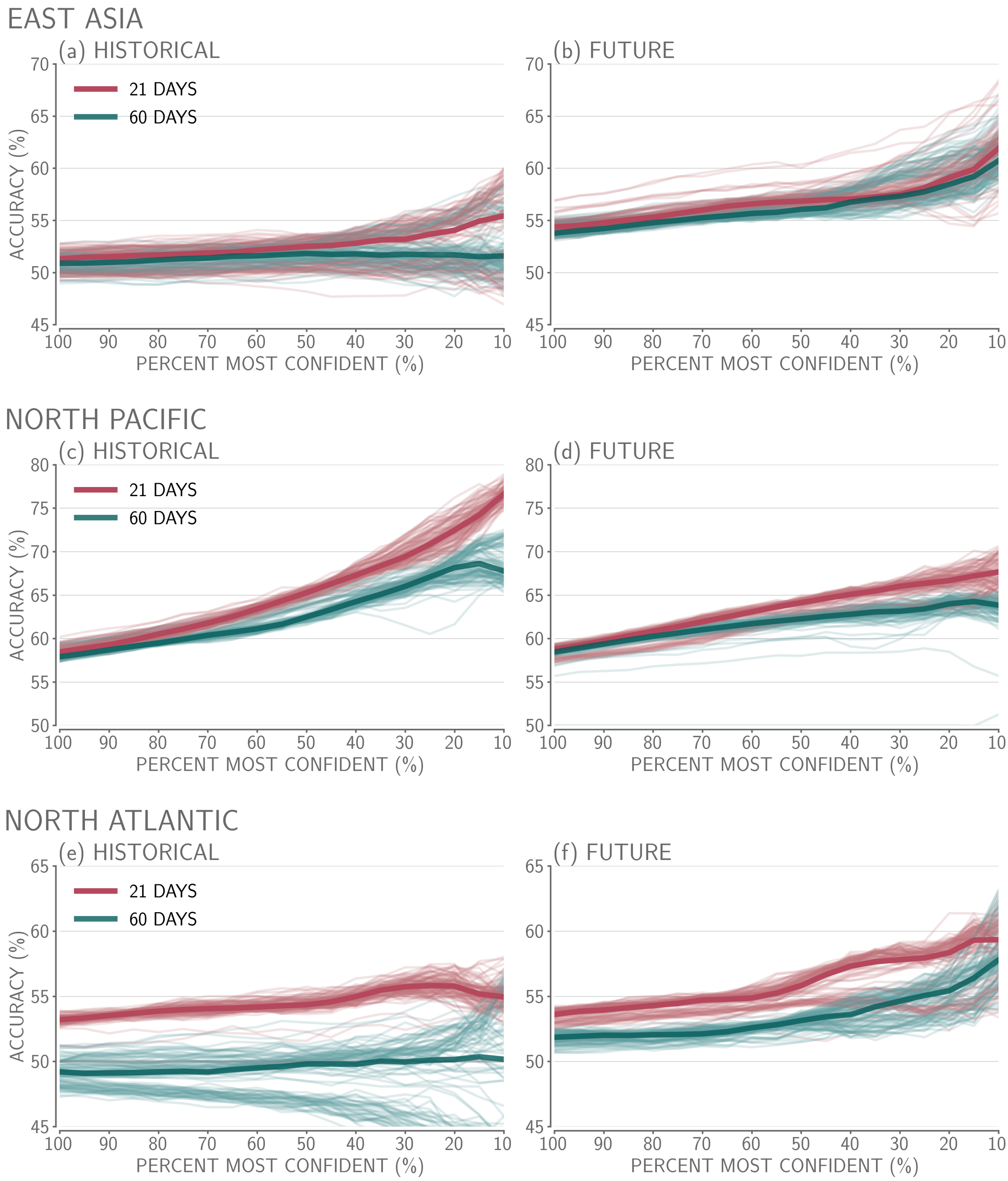

Figure S8. Accuracy versus confidence for 100 trained networks for the (left) historical and (right) future time period at leads of 21 (pink) and 60 (teal) days in (a,b) East Asia, (c,d) the North Pacific and (e,f) the North Atlantic. Accuracies are calculated using the testing member #10 and the thicker lines denote the median accuracy across the 100 networks at each confidence threshold. The pink lines are the same as the red/grey lines included in Figure 2 for the respective location and time period.

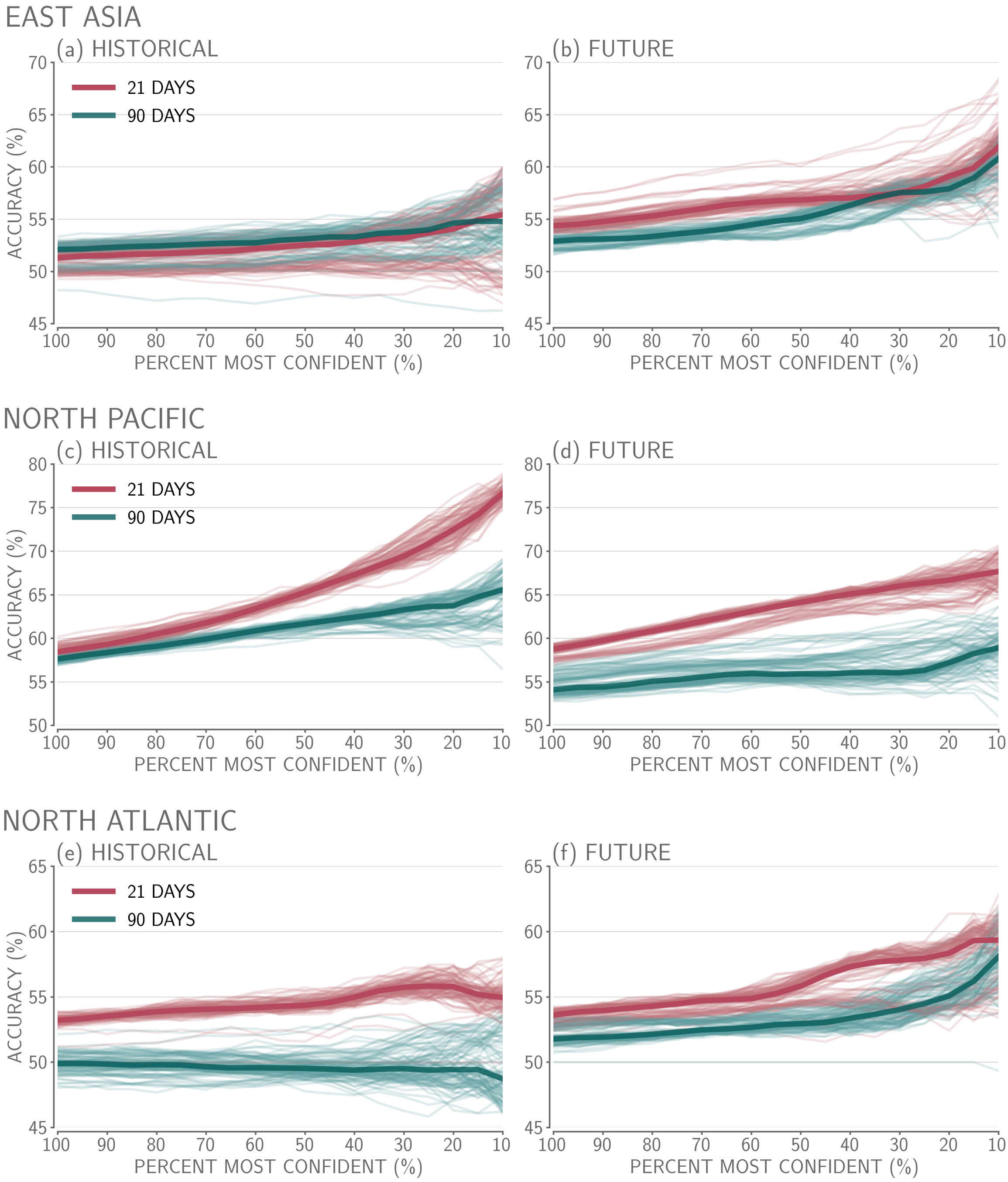

Figure S9. As in Figure S8, but for a lead of 90 days.