Correction to "Graph-Guided Regularized Regression of Pacific Ocean Climate Variables to Increase Predictive Skill of Southwestern U.S. Winter Precipitation" by Stevens et al., J. of Climate 34(2), 2021

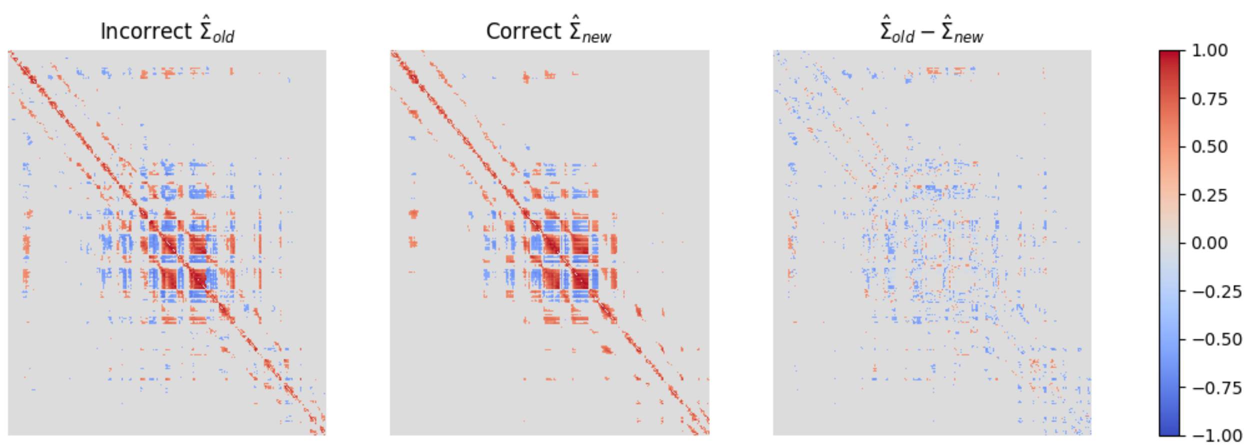
Our team discovered a small bug in the code used to preprocess the CESM-LENS data to be used to define the GTV regularization term. The regression model we proposed imposes a graph total variation constraint to align the estimated coefficients with the correlation structure of Pacific SSTs. The theoretical guarantees of our estimator rely on the existence of side information with which to estimate the covariance matrix, and in our work, we use the trajectories from CESM-LENS to do this. In our work, we proposed using the trajectories from CESM-LENS to robustly estimate the covariance matrix.

While exploring next steps in our work, we discovered that in an early preprocessing step, there was an error in how we constructed the CESM-LENS covariance matrix. In section 3.3 of our paper, we state the following:

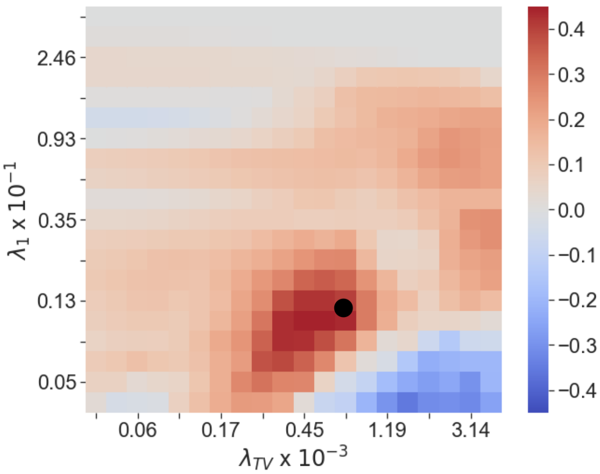
> Letting $\mathbf{X_{CL}} \in \mathbb{R}^{\mathbf{n}\times\mathbf{p}}$ be the detrended and standardized (zero mean and unit variance) matrix of stacked SST variables from all the CESM-LENS members, we define $\hat{\mathbf{\Sigma}}_{\mathbf{CL}}$ as the sample covariance of $\mathbf{X_{CL}}$.

We discovered that, while forming the $\mathbf{X_{CL}}$ matrix to compute its sample covariance, rather than stacking the 40 trajectories, we incorrectly indexed into them to create a matrix of overlapping trajectories, essentially not using all ensemble members but mostly overlapping copies of the first and second trajectories. Having corrected the data matrix $\mathbf{X_{CL}}$ and reconstructed the covariance matrix using all 40 ensemble members, the results do not differ significantly but we report the updated main figures herein and as part of our posted and corrected code.

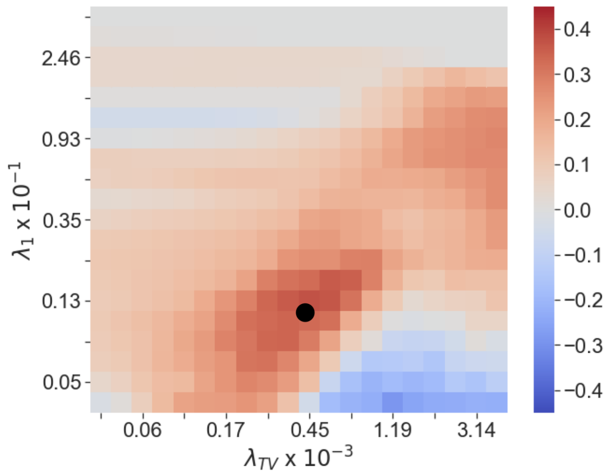
The incorrect and corrected covariance matrices are shown below (after the threshold of 0.5 has been applied), corresponding to Figure 4 in the paper. There are minor discrepancies between the covariance matrices, but no significant and systematic biases are observed.



We reproduced the contour plots of test R^2 values from Figure 4 using the incorrect and correct covariance estimates. We see that, with the corrected CESM-LENS, the patterns of predictability remain the same, and the test R^2 value at the optimal parameter values went down only slightly, from $R^2 \approx 0.42$ to $R^2 \approx 0.35$.

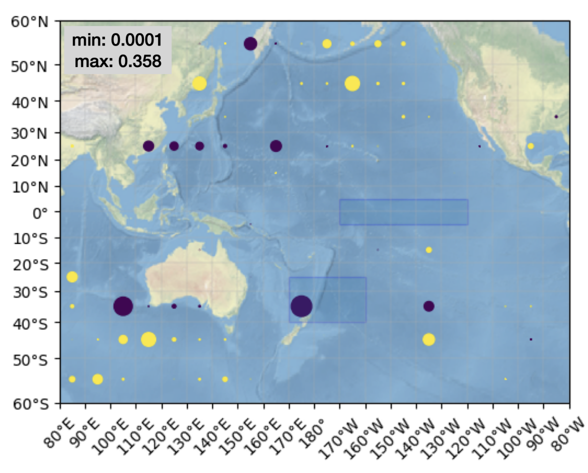


(a) Test R^2 using incorrect CESM-LENS

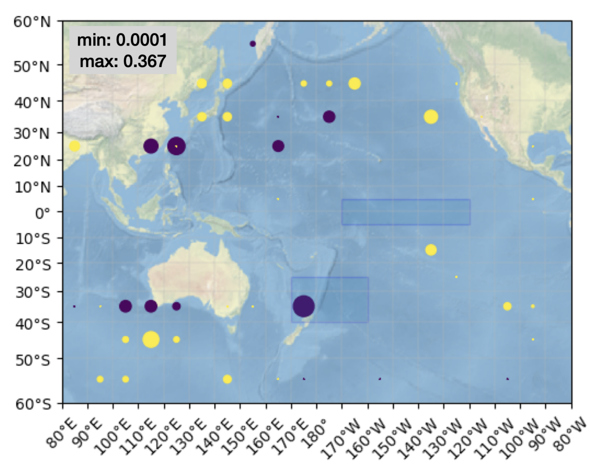


(b) Test R^2 using correct CESM-LENS

We then investigated the spatial patterns of the estimated coefficients at the optimal parameter values (Figure 7d). Again, we see minor discrepancies between them but overall spatial coherence in the emergent patterns.

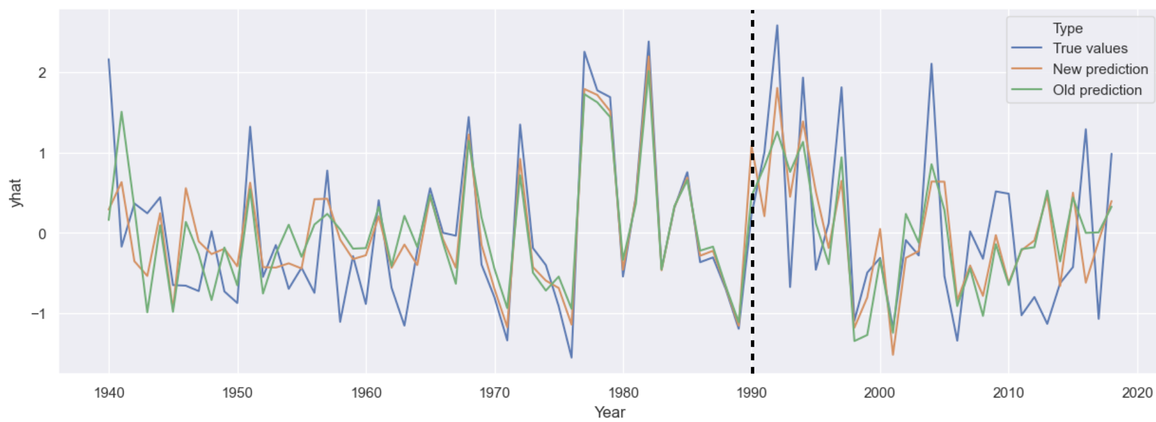


(a) Emergent predictors using incorrect CESM-LENS



(b) Emergent predictors using correct CESM-LENS

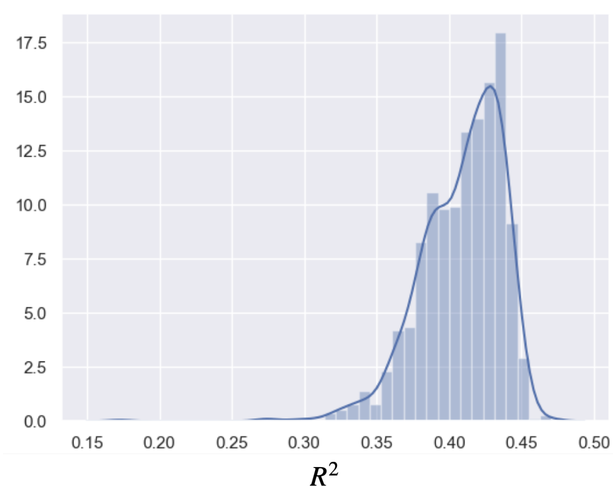
The actual series of predicted SWUS precipitation values also shows minimal difference (Figure 5a):



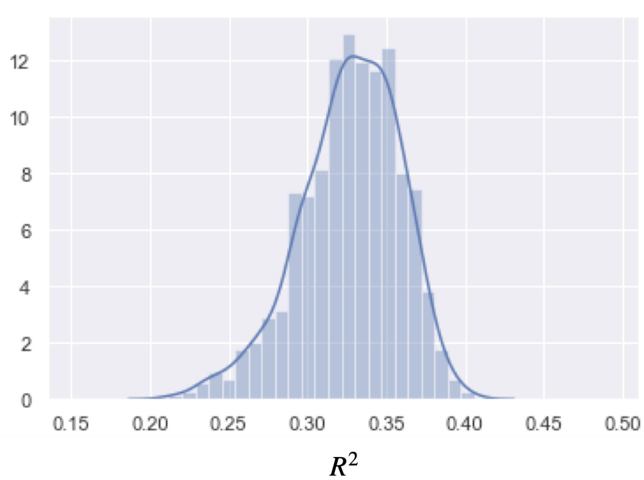
Finally, we repeated the bootstrap procedure from Section 4d, which is the piece of our analysis that it potentially most impacted by the mistake. As described in the paper,

> Namely, rather than stacking all 40 CESM-LENS trajectories to form the covariance matrix, we resample the 40 trajectories (with replacement) and compute the sample covariance of the new sample. Next, we form our GTV regularization term using this resampled covariance matrix, then fit the GTV scheme in the training period, and finally calculate the coefficient of determination (R^2) in the test period. By repeating this procedure 1000 times, we can quantify how the uncertainty in the covariance matrix propagates to uncertainty in the regression coefficients $\hat{\beta}$ and model performance.

With our incorrect construction of $\mathbf{X_{CL}}$, instead of resampling entire trajectories, we sampled the incorrect overlapping trajectories. Our bootstrapped results thus did not reflect the variability across CESM-LENS trajectories. In repeating this analysis with the correct construction of $\mathbf{X_{CL}}$, we find higher variance in the resulting distribution, which reflects the additional variability introduced from correctly using all 40 trajectories rather than just the first few.



(a) Bootstrapped test R^2 using incorrect LENS



(b) Bootstrapped test R^2 using correct LENS