Supplement

S1 False Discovery Rate

The false discovery rate (FDR) (Benjamini and Hochberg 1995) seeks to account for the multiple test comparison problem. To implement this control, ***n*** local hypothesis tests (equal to the number of reanalysis grid cells in our analysis) are conducted, with their p-values sorted in ascending order, $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$. Then a new threshold p-value (or critical value), p_{FDR} is generated using Eq. (1):

$$p_{FDR} = \max_{i=1,\ldots,n} [p_i : p_i \leq \left(\frac{i}{n}\right) \alpha_{FDR}] \quad (1)$$

where p_i is the ith p-value from the sorted list of p-values, and α_{FDR} sets the expected rate of Type 1 errors. After the p_{FDR} is calculated, local p-values below p_{FDR} result in a rejected local null hypothesis, and local p-values above p_{FDR} mean the local null hypothesis (in our case, H_0: no change in response time between periods) is not rejected. Testing with simulated climate datasets has indicated that FDR is robust to spatially dependent data (Wilks 2006; Ventura et al. 2004).

S2 Random forest algorithm

Random Forest (RF) regression uses an ensemble of regression trees to predict target variables from predictor variables. Here, our target variable is the response time during the time period 1951-2020 at each grid cell, while hydroclimatic variables serve as predictor variables (Tables 1 and 2). Our RF models each consist of 1000 regression trees, with each tree having a maximum depth of 10. To limit overfitting, each tree in an RF is trained on a bootstrapped resample of the original data (hydroclimatic data and response time at each grid cell), and only a random subset of predictor variables is considered at each split. Here, we consider a random one-third selection of predictor variables at each split. Once trained, each individual tree in the forest can make a prediction of response time at a grid cell given the hydroclimatic variables for that grid cell. The RF uses the mean response time prediction across all 1000 regression trees to make an overall prediction of response time at a grid cell from hydroclimatic variables for that grid cell. As we only use the RF for evaluating feature importance, predictions are done only to evaluate model performance on test data.

S3 Variable clustering method

To counter the issue of correlated features in a model sharing permutation importance, we follow a solution outlined by the creators of the Python predictive data science package, *scikit-learn* (Pedregosa et al. 2011). The variable clustering is described here. First, predictor variables are hierarchically clustered using an unweighted pair group with arithmetic mean (UPGMA), with distances corresponding to $1 - |corr|$, where $|corr|$ is the absolute value of the

spearman correlation between two features. UPGMA was chosen due to its simplicity and its ability to handle non-Euclidean distances such as $1 - |corr|$ (Datta and Datta 2006), and Spearman correlation is chosen as we do not expect the correlation between features to be linear. The result of the UPGMA is a dendrogram, a graphical representation of the hierarchical clustering of the features, showing how individual features are progressively clustered into larger groups. Clusters are selected manually from the resulting dendrogram and one feature from each cluster is retained reflecting the effect of that cluster.

S4 Partial dependence algorithm

PDPs are generated through the following method (Hastie et al. 2009). An RF is first fit on the full dataset of grid cell level pair hydroclimate and response time data, i.e., without a training/testing split. Then, for a given hydroclimate variable, $\boldsymbol{X}$, being analyzed, an interval of potential values is generated for that variable. Here, we use a standard practice of 100 equally spaced values between the 5th and 95th percentile values of the variable in the original data (Pedregosa et al. 2011). Then, for each of the equally spaced values ($\boldsymbol{x_i}$), every value of $\boldsymbol{X}$ is set to that value for every grid cell in the dataset while keeping all other hydroclimate variables at their original value. Model output (i.e., response time) is generated for every grid cell in the dataset, and an average of those predicted values is determined. Thus, for every value of $\boldsymbol{x_i}$, an average predicted value is generated while keeping every other non-$\boldsymbol{X}$ variable constant.

S5 Spatially-blocked 5-fold cross validation

First, the dataset of grid cell level hydroclimate predictors and response times is blocked into 2° x 2° blocks, after testing indicated that performance metrics did not noticeably change when block size increased past 2° x 2°. Those blocks are then split into 5 training/testing splits, with each block appearing in exactly one testing dataset. For each training/testing data split, model performance is quantified using R^2, and variable importance is measured through permutation importance.

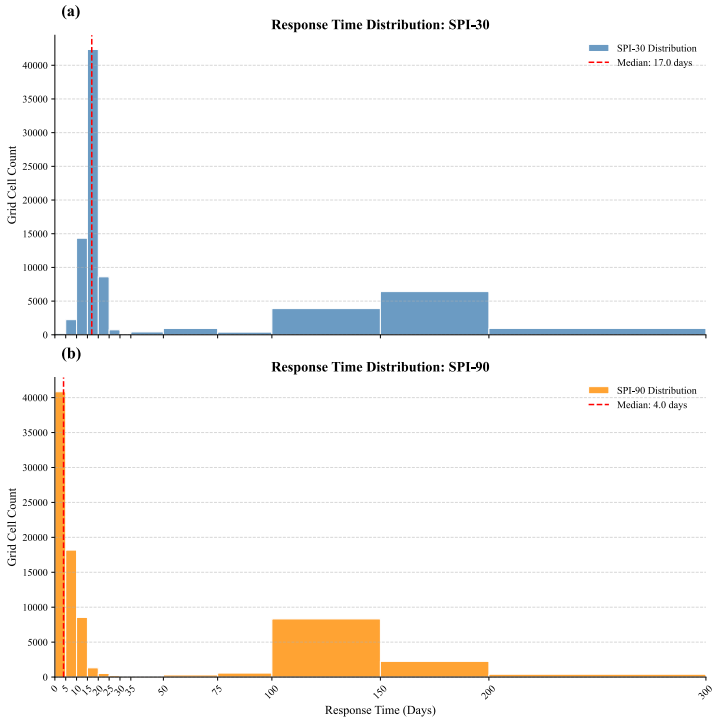

Figure S1: Histogram of response times of JJA SSI-30 across CONUS to (a) SPI-30 and (b) SSI-30.