Geophysical Research Letters

Supporting Information for

Sea Surface Salinity Provides Subseasonal Predictability for Forecasts of Opportunity of U.S. Summertime Precipitation

Marybeth C. Arcodia[1], Elizabeth A. Barnes[1], Paul J. Durack[2], Patrick W. Keys[1], Juliette Rocha[3]

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

[2]Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory, Livermore, CA, USA.

[3] Department of Atmospheric Science, Texas A&M University, College Station, TX, USA.

Contents of this file

Text S1. Data Preprocessing

Daily data from 10 CESM2 ensemble members (Table S1) are used from 1850-1949 for May-August. All CESM2 data are interpolated from a 1x1 degree resolution to 2.5 x 2.5 degree resolution via bilinear interpolation for computational efficiency.

Sea surface salinity anomalies are computed via subtraction of the linear trend at each grid point of the ensemble mean for each day of the year to remove the forced response and retain only internal variability. A 3-day running average is applied to smooth the data while retaining high frequency fluctuations. Similar analyses with a 1-day and 5-day running averages of the precipitation data yielded similar results. The data are normalized by subtracting the mean and dividing by the standard deviation at each grid point.

The precipitation data are raw CESM2 data (e.g. not anomalies) averaged over the Midwest region (Fig. 1a). The daily precipitation in this region is summed cumulatively for 3 days.

Our goal is to evaluate the predictability of precipitation events across lead times spanning from the weather to subseasonal range. Therefore, we apply a Poisson weighting (Fig. S1) to the data to smooth the timeseries as lead time increases. Large weights are applied to the day being predicted for short-term forecasts (e.g. 7-day lead predictions; orange line in Fig. S1). Weights are distributed more widely as lead time increases, eventually widening into a centered nearly-Gaussian average as the upper limit of a Poisson distribution is the Gaussian distribution (e.g. 56-day lead predictions, yellow line in Fig. S1). After the precipitation time series are smoothed with the Poisson weighting (Fig. S1), the 3-day periods of precipitation are then ranked by magnitude. Periods above the 80th percentile of precipitation are classified as heavy events,

designated as a 1, and the remaining 80% of the time period are classified as light events, designated as a 0. Predictions are made using the 3-day trailing average sea surface salinity map to make the prediction of the 3-day forward-cumulative sum beginning with the day of each respective lead time (0-day, 7-day, 14-day, 21-day, 28-day, 35-day, 42-day, 49-day, 56-day). For example, a 0-day lead prediction made on May 4, 1850 uses the averaged salinity input from May 1-3 to classify the precipitation event as light or heavy for May 4-7, 1850. The same input map would be used for a 7-day lead example, but to classify the precipitation event for May 11-14, 1850, and so on for all lead times. Classification is performed individually within each ensemble member for each smoothed time series based on prediction lead time.

Training of the neural networks is performed using seven ensembles with each network initialized with 5 random seeds for robustness of the results. Two members are used for validation, and one member is used for testing. The training, validation, and testing ensemble members are then randomly reselected to train another set of neural networks with 5 random seed initializations. This strategy ensures that training of networks is performed individually so that no knowledge of the test data is used in the training of the networks. This process is repeated 5 times, for a total of 25 trained neural networks (5 networks with 5 random initializations each) per lead time.

Based on the nature of the classification of the output by percentile, the training, validation, and testing data are heavily imbalanced. For effective training of the networks to learn both the light and heavy precipitation event classes, we undersample our data via randomly selecting light precipitation events to remove from the training set to balance the classes for an even 50-50 split (e.g. Prusa et al., 2015). Although 60% of the data is discarded in this process, the benefit of large ensemble climate model data used here ensures that we still have enough data

Text S2. Neural Network Architecture

The neural network architecture is depicted in the schematic in Fig. 1a. The architecture is identical for networks trained for predictions from leads of 0-35 days and then a slightly different architecture was used for leads of 42-56 days. Hyperparameter tuning was performed using the KerasTuner (O'Malley et al., 2019) to find the optimal set of parameters determined via validation accuracy. For the shorter lead forecasts, the network architecture consists of 1 hidden layer with 128 nodes with a rectified linear activation function applied (ReLU), a dropout rate of 50% and ridge regression coefficient of 0.1 to reduce overfitting, batch size of 32 samples, and a learning rate of 1.618e-5. For the longer lead forecasts beyond 35 days, the network architecture consists of 2 hidden layers with 160 and 192 nodes with a rectified linear activation function applied (ReLU) to each, a dropout rate of 80% and ridge regression coefficient of 0.01 to reduce overfitting, batch size of 32 samples, and a learning rate of 2.886e-6. All networks have a set global seed of 147483648 and are initialized with the following random seeds: 6, 26, 19, 54, 68. Networks are trained using the categorical cross-entropy loss function. Networks are trained with early stopping when the validation loss does not decrease after 25 epochs.

We note that for the lead of 7 days, the network architecture with the highest validation accuracy was slightly different than the one used here. However, the same architecture which resulted in the highest validation accuracy for leads 0, 14, 21, 28, and 35 resulted in a validation accuracy on the order of 0.001 less than the highest performing architecture. Therefore, we used the same architecture for all leads 0-35 days for simplicity.

Text S3. Water Accounting Model

The WAM2layers uses ERA5 climate reanalysis data, including hourly, 2-dimensional surface pressure, evaporation and precipitation, and hourly 3-dimensional specific humidity, and zonal and meridional winds. We use data from 2008-2021 for this analysis. We use the backtracking function of the WAM2layers, which permits the tracing of precipitated water back through the atmosphere to its origins as evaporation. In this study, we spin-up the model for six months prior to the event of interest, to ensure full saturation of the atmospheric column.

Text S4. Skill Scores

The Threat Score is a biased verification metric for categorical forecasts in which the score is based on the frequency of the event. It is defined as *hits/(hits+false alarms+misses)* in which a hit is a correctly forecasted heavy precipitation event, a false alarm is the prediction of a heavy precipitation event but it does not occur, and a miss is a prediction of a light precipitation event but a heavy event occurred. It does not account for correct rejections, e.g. correctly forecasted light events. The Gilbert Skill Score is an unbiased verification metric which accounts for the number of hits due to random chance, i.e. *chance hits*. It is defined as (*hits-chance hits)/(hits+false alarms+misses-chance hits)* where *chance hits = (hits+false alarms)*(hits+misses)/total number of forecasts*. For both skill scores, a score of zero denotes no skill, or random chance, and a skill of one is a perfect score.

Text S5. Case Studies with the WAM2layers Model

We compute the percentage of moisture that originated from a certain location for the two case studies. Specifically, we compute the amount of moisture that originated over the Caribbean Sea and Gulf of Mexico region (262-320E; 11-30N) which eventually fell in the Midwest during the event. This value is divided by the total moisture that precipitated in that region during the event. The local moisture recycling percentage is computed by the amount of moisture that originated in the analyzed region (e.g. red boxes in Fig. 4c and S5) divided by the total moisture that precipitated in that region during the event.

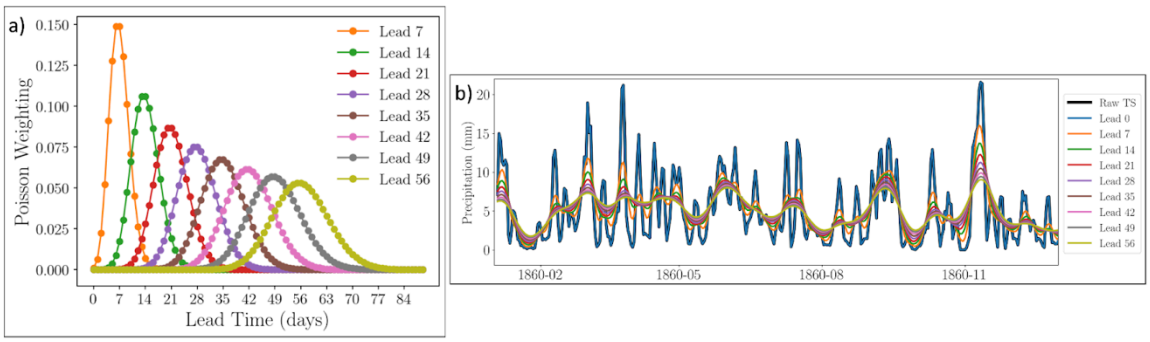

Figure S1. a) The Poisson distribution of the weights applied to the forecast period as a function of lead time. b) An example time series of the 3-day cumulative sum of precipitation in the U.S. Midwest for 1860 from ensemble member #0 showing the smoothed time series based on the Poisson weighting in (a). No weights are applied to lead of 0 days, so the raw time series (Raw TS; black line) and the time series for a lead of 0 days (Lead 0; blue line) are the same.

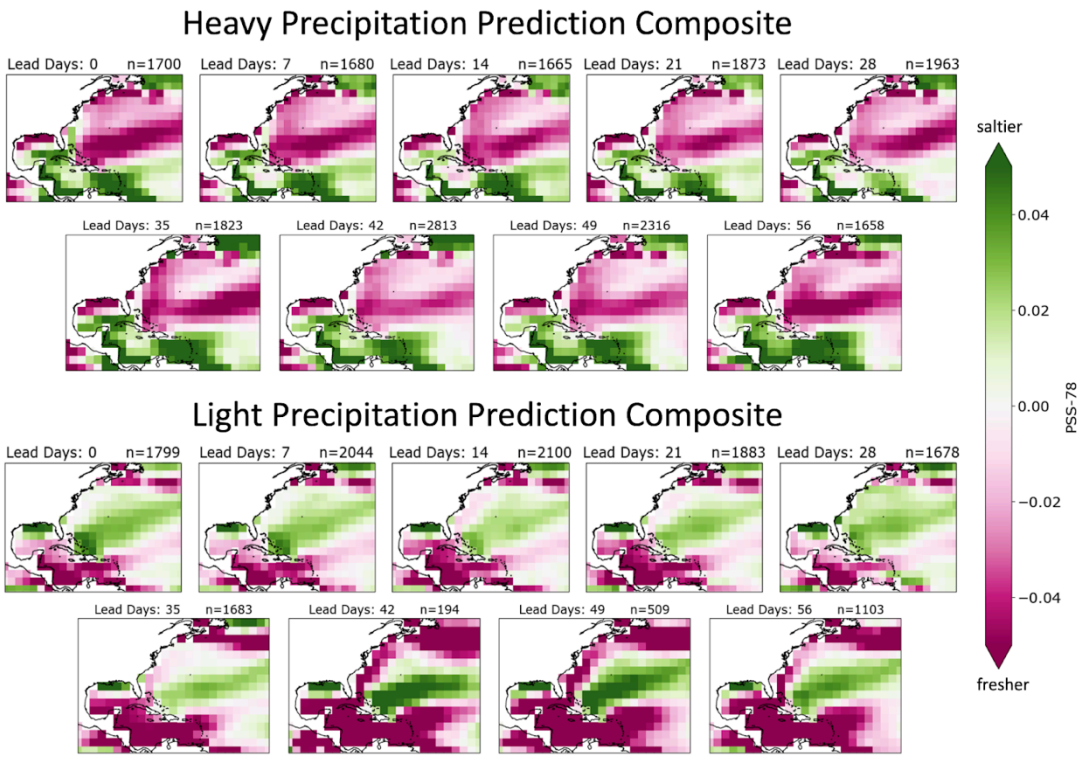

Figure S2. Composite of the sea surface salinity anomalies in PSS-78 for input maps of the 20% most confident, correct predictions for all leads for heavy predictions (top) and light predictions (bottom). Green colors represent positive sea surface salinity anomalies, or saltier waters, while pink colors represent negative sea surface salinity anomalies, or fresher waters. The number *n* represents the number of samples per composite.

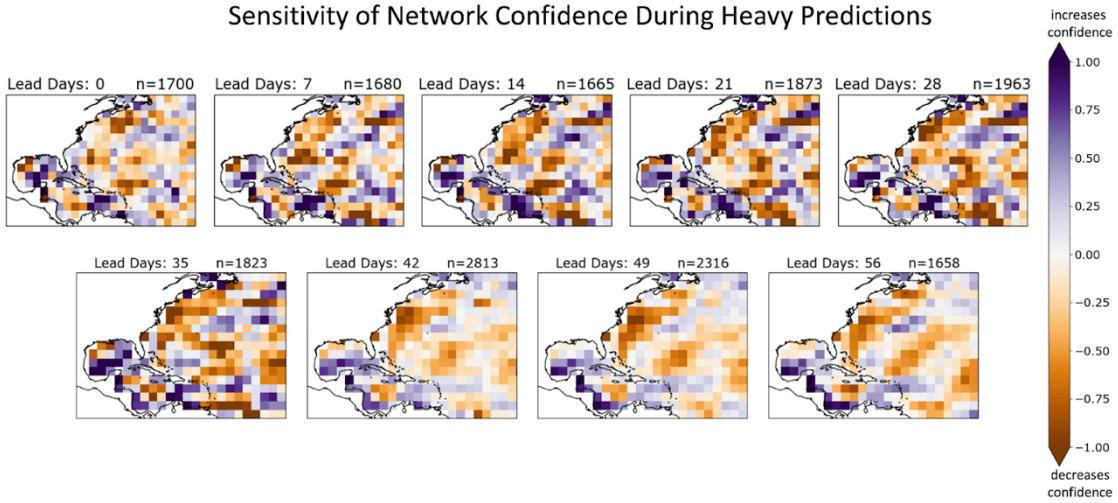

Figure S3. Saliency XAI composited heatmaps for the same days as the input maps of the 20% most confident, correct predictions for all leads for heavy predictions. Darker purple colors designate increased network confidence for positive salinity anomalies, and vice versa for orange colors. The colorbar is a unitless measure of sensitivity. The colorbar is a unitless measure of sensitivity. The number *n* represents the number of samples per composite.

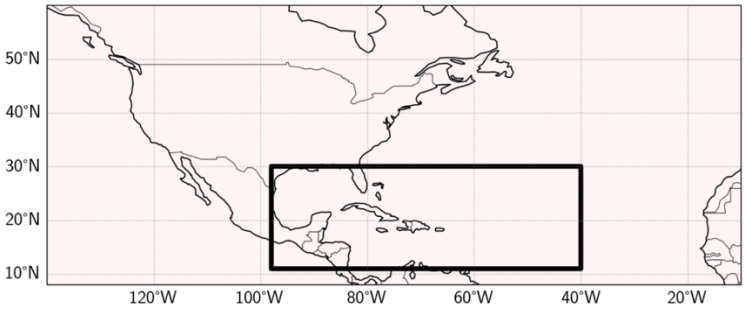

Figure S4. Region over which the moisture origin source is computed. The black box outlines 262-320E; 11-30N.

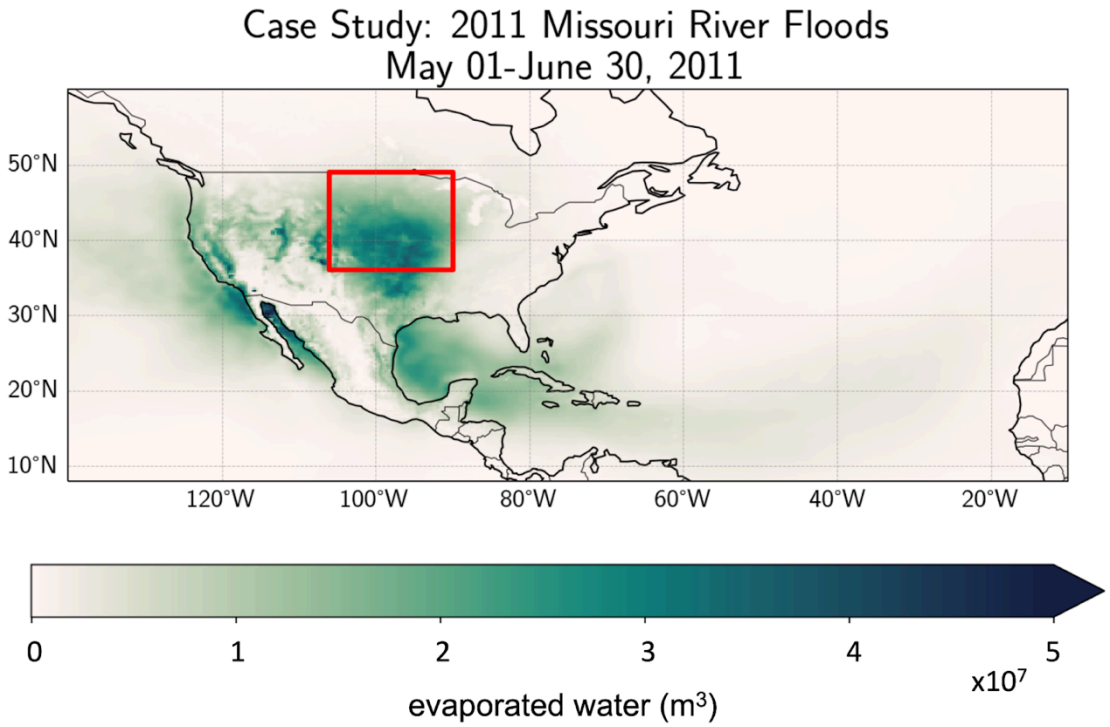

Figure S5. The sum of the evaporated water (in cubic meters) which fell as precipitation in the red boxed region computed using the WAM2layers backtracking algorithm for May 1 through June 30, 2011.

Ensemble	CESM2 Ensemble Member Name for Precipitation Variable	CESM2 Ensemble Member Name for Sea Surface Salinity Variable
0	b.e21.BHISTsmbb.f09_g17.LE2-1231.011.cam.h1.PRECT.*1850-1949*.nc	b.e21.BHISTsmbb.f09_g17.LE2-1231.011.pop.h.nday1.SSS.*1850-1949*.nc
1	b.e21.BHISTsmbb.f09_g17.LE2-1231.012.cam.h1.PRECT.*1850-1949*.nc	b.e21.BHISTsmbb.f09_g17.LE2-1231.012.pop.h.nday1.SSS.*1850-1949*.nc
2	b.e21.BHISTsmbb.f09_g17.LE2-1251.013.cam.h1.PRECT.*1850-1949*.nc	b.e21.BHISTsmbb.f09_g17.LE2-1251.013.pop.h.nday1.SSS.*1850-1949*.nc
3	b.e21.BHISTsmbb.f09_g17.LE2-1251.014.cam.h1.PRECT.*1850-1949*.nc	b.e21.BHISTsmbb.f09_g17.LE2-1251.014.pop.h.nday1.SSS.*1850-1949*.nc
4	b.e21.BHISTsmbb.f09_g17.LE2-1281.015.cam.h1.PRECT.*1850-1949*.nc	b.e21.BHISTsmbb.f09_g17.LE2-1281.015.pop.h.nday1.SSS.*1850-1949*.nc
5	b.e21.BHISTsmbb.f09_g17.LE2-1281.016.cam.h1.PRECT.*1850-1949*.nc	b.e21.BHISTsmbb.f09_g17.LE2-1281.016.pop.h.nday1.SSS.*1850-1949*.nc
6	b.e21.BHISTsmbb.f09_g17.LE2-1301.017.cam.h1.PRECT.*1850-1949*.nc	b.e21.BHISTsmbb.f09_g17.LE2-1301.017.pop.h.nday1.SSS.*1850-1949*.nc
7	b.e21.BHISTsmbb.f09_g17.LE2-1301.018.cam.h1.PRECT.*1850-1949*.nc	b.e21.BHISTsmbb.f09_g17.LE2-1301.018.pop.h.nday1.SSS.*1850-1949*.nc
8	b.e21.BHISTsmbb.f09_g17.LE2-1281.017.cam.h1.PRECT.*1850-1949*.nc	b.e21.BHISTsmbb.f09_g17.LE2-1281.017.pop.h.nday1.SSS.*1850-1949*.nc
9	b.e21.BHISTsmbb.f09_g17.LE2-1251.019.cam.h1.PRECT.*1850-1949*.nc	b.e21.BHISTsmbb.f09_g17.LE2-1251.019.pop.h.nday1.SSS.*1850-1949*.nc

Table S1. The naming convention of the ensemble members used in this study and the corresponding CESM2 ensemble member for precipitation and sea surface salinity. We used the smoothed biomass burning ensemble members (denoted by smbb). We italicize *1850-1949* because the data available are in 10-year increments. The data are downloaded then concatenated for the full time series.

Tunable Parameter	Search Space
Dropout	[0.0, 0.1, 0.2, 0.5, 0.8]
Ridge Regression (L2)	[0.01, 0.1,0.5,1.0,2.0,5.0,10.0]
# of Hidden Layers	[1,2]
# of Nodes per Layer	[32, 64, 96, 128, 160, 192, 224, 256]
Batch Size	[32, 64, 128, 256, 512]
Learning Rate	[1e-7, 1e-6, 1e-5, 1e-4, 1e-3]

Table S2. The hyperparameter search space evaluated using the KerasTuner to select the neural network architecture with the highest validation accuracy. The learning rate parameter space follows a logarithmic scale. 25 trails were performed using a random combination of the above parameters. Each network was trained for 5000 epochs with early stopping applied if validation loss increased after 25 epochs (e.g. the patience was 25).