Supplementary Information for

Skillful Long-lead Prediction of Summertime Heavy Rainfall in the US Midwest from Sea Surface Salinity

Laifang Li[1,3,4*], Raymond W. Schmitt[2], Caroline C. Ummenhofer[2]

[1] Department of Meteorology and Atmospheric Science, the Pennsylvania State University, University Park, PA, 16802

[2] Department of Physical Oceanography, Woods Hole Oceanographic Institution, Woods Hole, MA, 02543

[3] Earth and Environmental Science Institute, the Pennsylvania State University, University Park, PA, 16802

[4] Institute of Computational and Data Science, the Pennsylvania State University, University Park, PA, 16802

*Correspondence to: Laifang Li (lfl5240@psu.edu)
ORCID: https://orcid.org/0000-0002-6721-9002

Contents of this file

1. Deriving the posterior distribution of the parameters in the Gaussian mixture model

Prior to implementing the Bayesian Gaussian mixture model (GMM), we applied a natural log transformation to daily rainfall so that the data samples are distributed in the range of $(-\infty, +\infty)$. In this model (Eq. 1), the three clusters, respectively, represent light, moderate, and heavy precipitation events according to the AMS Glossary of Meteorology (2013). It is noteworthy that the choice of the number of clusters is subjective when constructing finite GMM (e.g. Richardson and Green 1997; McLachlan and Peel 2000; Melnykov and Maitra 2010). Generally, adding more clusters to the mixture model usually better approximates the true distribution, but at the same time increases the risk of over-fitting (Lin *et al* 2007), blurs the physical meanings of each cluster, and hampers the interpretation of mechanisms that control the rainfall probability distribution. We adopt the 3-cluster model based on the conventionally practiced rainfall categorization based on the AMS criteria.

The distribution parameters (π, μ, and ϕ) in the likelihood model (Eq. 1) were obtained using Bayesian inference, and their prior distributions are as follows:

$$\pi_{\mathrm{h}}|a_{\mathrm{h}} \sim \mathrm{Dirichlet}(a_{\mathrm{h}}), \quad \text{(S1)}$$

$$(\mu_{\mathrm{h}}, \phi_{\mathrm{h}}) \sim \mathrm{Gauss}\left(\mu_{\mathrm{h}}\middle|\mu_{0\mathrm{h}}, \kappa_{\mathrm{h}}\phi_{\mathrm{h}}^{-1}\right)\mathrm{Gamma}(\phi_{\mathrm{h}}|\alpha_{\mathrm{h}}, \beta_{\mathrm{h}}). \quad \text{(S2)}$$

In Eq. (S2), κ is the degree of freedom, and the $\mathrm{Gamma}(\phi_{\mathrm{h}}|\alpha_{\mathrm{h}}, \beta_{\mathrm{h}})$ is parameterized to have mean $\alpha_{\mathrm{h}}/\beta_{\mathrm{h}}$ and variance $\alpha_{\mathrm{h}}/\beta_{\mathrm{h}}^2$. The parameters in the prior distributions (Eqs. S1-S2) were assigned according to the AMS definitions of light (0-6 $mm\ day^{-1}$), moderate (6-18 $mm\ day^{-1}$), and heavy (>18 $mm\ day^{-1}$) rainfall (AMS Glossary of Meteorology 2009), but were kept weakly informative to incorporate more data information into the posterior distributions: $a_{\mathrm{h}} = (0.5, 0.35, 0.15)$, $\mu_{0\mathrm{h}} = (0.0, 2.0, 3.0)$[1], $\kappa_{\mathrm{h}} = (1,1,1)$, $\alpha_{\mathrm{h}} = (1.0,1.0,0.4)$, and $\beta_{\mathrm{h}} = (1.0,1.0,1.0)$.

Given the likelihood model (Eq. 1) and the prior distributions (Eqs. S1-S2), the full conditional posterior distributions can be derived analytically (Gelfand 2000). The Gibbs sampler for posterior computation using Markov Chain Monte Carlo (MCMC) algorithm is as follows:

$$(\pi_{\mathrm{h}}|-) \sim \mathrm{Dirichlet}(a_{\mathrm{h}} + n_{\mathrm{h}}), \quad \text{(S3)}$$

$$P(z_{\mathrm{i}} = \mathrm{h}|-) = \frac{\pi_{\mathrm{h}}\mathrm{Gauss}(y_i|\mu_{\mathrm{h}}, \phi_{\mathrm{h}}^{-1})}{\sum_{h=1}^{3} \pi_{\mathrm{h}}\mathrm{Gauss}(y_{\mathrm{i}}|\mu_{\mathrm{h}}, \phi_{\mathrm{h}}^{-1})}, \quad \text{(S4)}$$

$$(\mu_{\mathrm{h}}, \phi_{\mathrm{h}}|-) \sim \mathrm{Gauss}\left(\mu_{\mathrm{h}}\middle|\hat{\mu}_{0\mathrm{h}}, \hat{\kappa}\phi_{\mathrm{h}}^{-1}\right)\mathrm{Gamma}\left(\phi_{\mathrm{h}}\middle|\hat{\alpha}_{\mathrm{h}}, \hat{\beta}_{\mathrm{h}}\right), \quad \text{(S5)}$$

where $n_{\mathrm{h}} = \sum_{i=1}^{n} 1(z_i = \mathrm{h})$ denotes the number of samples in cluster h, $\hat{\kappa}_{\mathrm{h}} = \left(\kappa_{\mathrm{h}}^{-1} + n_{\mathrm{h}}\right)^{-1}$, and $\hat{\mu}_{0\mathrm{h}} = \hat{\kappa}_{\mathrm{h}}\left(\kappa_{\mathrm{h}}^{-1}\mu_0 + n_{\mathrm{h}}\bar{y}_{\mathrm{h}}\right)$ ($\bar{y}_{\mathrm{h}} = n_{\mathrm{h}}^{-1}\sum_{i:z_i=\mathrm{h}} y_i$ is the sample mean of cluster h). $\hat{\alpha}_{\mathrm{h}} = \alpha_{\mathrm{h}} + \frac{n_h}{2}$, and $\hat{\beta}_{\mathrm{h}} = \beta_{\mathrm{h}} + \frac{1}{2}\left\{\sum_{i=1}^{n_h}(y_i - \bar{y}_{\mathrm{h}})^2 + \left(\frac{n_{\mathrm{h}}}{1+\kappa_{\mathrm{h}}n_{\mathrm{h}}}\right)(\bar{y}_{\mathrm{h}} - \mu_{0\mathrm{h}})^2\right\}$.

In this study, the MCMC algorithm (Eqs. S3-S5) was applied to US Midwest daily precipitation during the 1948-2019, with physical constraint placed upon μ_{h} ($\mu_1 \leq \mu_2 \leq \mu_3$, meaning that heavy rainfall intensity is weighted more heavily for stronger than moderate and light rainfall) to deal with label switching issues (Stephens 2000). The MCMC algorithm consists of 1000-iteration runs, which stabilizes after 100 (200) iterations for μ_h (π_h) (Fig. S1). Thus, the first 200 burn-in samples were discarded, and the remaining 800 post burn-in samples are used in the analysis.

[1] Here μ_{0h} corresponds to the mean of $\ln Pr$, which is equivalent to precipitation of 1.0, 8.0, and 20.0 for light, moderate and heavy rainfall.

2. Relationship between heavy rainfall frequency and SSTA and SSSA predictors: subsample the Gaussian mixture model

We define SSTA and SSSA-based indices over the corresponding regions in the North Pacific and North Atlantic (Fig. 3a-b), in order to further quantify the response of Midwest heavy rainfall frequency to the abovementioned preseason SSTA and SSSA patterns. These indices include Niño 3.4 SSTA, Gulf of Mexico (GOM) SSTA, tropical western Pacific SSSA dipole index, and North Atlantic SSSA (Fig. S4). Here, Niño 3.4 SSTA is calculated as monthly SSTA averaged over the 170°W-120°W, 5°S-5°N (Fig. S4a). GOM SSTA is domain averaged SSTA over the area of 100°W-80°W, 20°N-30°N (Fig. S4b). The tropical western Pacific SSSA dipole index is defined as the SSSA difference between the southern domain (165E-170W, 20S-10S) and equator domain (150E-180, 5S-5N) based on SSSA distribution in Fig. 3a (Fig. S4c). In addition, the North Atlantic SSSA is the SSSA averaged over the northwestern portion of the subtropical North Atlantic (Fig. S4d), the same as the definition by Li *et al*. (2016, 2018).

The association between US Midwest heavy rainfall and these identified indices is manifested in the case of 1993, a record-breaking wet year in the region's history. In this year, the March-April May (MAM) SSTA shows a positive phase of El Nino Southern Oscillation (ENSO) in the central and eastern tropical Pacific (Fig. S5a), and the El Niño SSTA decays in the summer (Fig. S5b). At the same time, the SSTA in the GOM is significantly cooler than normal in the spring months (Fig. S5a). Thus, the SSTA pattern is consistent with the regression analysis as shown in Fig. 3a. Similarly, the SSSA also shows consistency with Fig. 3b. The salinity along the western part of the subtropical North Atlantic is saltier than normal by 0.6 PSU, which exceeds 3 standard deviation of the interannual SSSA variability in the region (Fig. S5c). In addition, the salinity dipole is also significant in the tropical western Pacific, where a significant freshening is observed along the equator with a salty patch to the south (Fig. S5c-d). Thus, the SSTA and SSSA in the year 1993 verifies the findings in Fig. 3a-b. The El Niño SSTA, cold GOM, saltier than normal subtropical North Atlantic, and positive salinity dipole in the western tropical Pacific are all favorable for an increase in heavy rainfall events in the US Midwest, which might have collectively contributed to the wettest summer in US Midwest instrumental records.

According to the SSTA and SSSA indices (Fig. S4), we subsample the precipitation datasets and re-run the GMM (Eqs. 1 and S1-S5) to evaluate the changes in the probability distribution of heavy rainfall frequency (π_3) in response to the preseason SSTA and SSSA. The daily rainfall samples included in each simulation by the GMM is summarized in Table S1.

The Bayesian inference suggests that the knowledge of pre-season Nino3.4 and GOM SSTA could moderately improve the forecasting the summertime heavy rainfall frequency over the US Midwest. Specifically, heavy rainfall frequency increases (decreases) from 15% to 16.8% (12.9%) when the Nino3.4 index is 1.5STD above (below) normal (Fig. S6a), a 3.9% difference between the opposite phase of ENSO. The opposite GOM SSTAs would lead to a similar 4% differences in the heavy rainfall frequency over the US Midwest (Fig. S6b). In addition, the combination of warm tropical Pacific and cold GOM would result in a 5.2% increase in heavy rainfall frequency as compared to the scenarios with cold Pacific and warm GOM (Fig. S6c), which is less than the summation of the two individual cases (7.9%). Compared to SSTA, the pre-season SSSA indices are more skillful in predicting the heavy rainfall frequency especially when both indices are taken into account. Specifically, the Pacific salinity dipole (North Atlantic SSSA) alone corresponds to 5.1% (5.7%) changes in heavy rainfall frequency (Fig. S6d-e). In combination, a positive Pacific salinity dipole and a salty subtropical North Atlantic could increase the heavy rainfall frequency to 22.6%, which almost doubles during the years with a negative dipole and fresh North Atlantic (12.1%; Fig. S6f).

The Bayesian GMM subsampling results are consistent with the logistic regression model, in which SSSA-based prediction explains 92% more heavy rainfall variance compared to SSTA-based predictors (Fig. 3c-e). The combination of the two SSSA-based predictors produces a prediction that explains 0.27 of the observed heavy rainfall variance. Here, the R^2 between observations and the prediction by the North Pacific SSSA dipole is 0.15, and is 0.13 between the North Atlantic SSSA. Thus, the two SSSA-based predictors contribute equally to the prediction of heavy rainfall, and their predictive skill is overall linearly additive (0.28 compared to 0.27 from the prediction using both predictors). It is noteworthy that the prediction skill of the Logistic regression model (Eq. 2 and Fig. 3c-e) is largely leveraged by the extremely wet year of 1993 when 30% of the daily precipitation is categorized as heavy rainfall. According to our assessment, this case along accounts for about 14% of the total model skill in predicting the heavy rainfall frequency. By excluding the 1993 case, the R^2 between observation and the logistic regression model (with all four predictors) drops from 0.46 to 0.32. The SSTA-based predictors only generate an R^2 of 0.07, while the SSSA-based prediction maintains an R^2 of 0.21.

3. North Atlantic pre-season SSSA and US Midwest summer season precipitation relationship

The linkage between springtime SSSA in the subtropical North Atlantic and summer precipitation in the US Midwest is through soil moisture feedback (Li *et al*. 2016; Carter *et al*. 2021). Specifically, positive springtime SSSA in the subtropical North Atlantic is associated with increased moisture export from the ocean, which converges over the Southern Great Plains by prevailing southeasterly wind along the western edge of the Azores high (Fig. 4a; Li *et al*. 2016; 2018). The increased oceanic moisture input elevates local soil moisture content, and the soil moisture anomaly is preserved in the subsequent season as the typical time scale and land surface memory is are 3-6 months.

In the subsequent season, the soil moisture anomaly interacts with the overlying atmosphere and promotes moisture flux convergence into the US Midwest through both thermodynamic and dynamic processes (Fig. 4b; see detailed moisture budget analysis and mechanism in Li *et al*. 2016). Thermodynamically, the increased soil moisture content moistens the boundary layer in the Southern Great Plains, enhancing moisture supply into the US Midwest carried by the GPLLJ (Meehl and Washington 1988; Delworth and Manabe 1989; Ek and Holtslag 2004; Aligo *et al*. 2007). Dynamically, the soil moisture anomaly in the southern US created a zonal soil moisture contrast along the eastern slope of the Rocky Mountains, which by modulating pressure distribution intensifies the GPLLJ (Fast and McCorcle 1990, 1991). The dual effects of soil moisture on the regional water cycle favors heavy precipitation in the US Midwest both thermodynamically and dynamically (the detailed moisture budget analysis is documented in Li *et al*. 2016). This mechanism is supported by the extreme summer precipitation in the summer seasons of the 2008 and 2015 (Li *et al*. 2018).

It is noteworthy that, as the water cycle intensify and the subtropical North Atlantic gets saltier (e.g., Durack *et al*. 2012; Cheng *et al*. 2019; Skliris *et al*. 2020), the North Atlantic SSSA is expected to generate higher predictive skills for US Midwest heavy rain (Fig. S9). From Fig. S9, there is a clear breakdown in the regression slope between SSSA and heavy rainfall frequency at neutral SSSA over the subtropical North Atlantic. For SSSA<0 portion of the data, one standard deviation increases in springtime SSSA corresponds to 0.7% increase in summertime heavy rainfall frequency in the Midwest (Fig. S9). In contrast, for SSSA>0 portion of the data, the same one standard deviation increases in MAM SSSA is followed by 6% increase in summertime heavy rainfall frequency (Fig. S9).

References

Aligo, E. A., Gallus, W. A., & Segal, M. (2007). Summer rainfall forecast spread in an ensemble initialized with different soil moisture analyses. *Weather Forecast*., **22**, 299–314.

Carter, E., Herrera, D. A., & Steinschneider, S. (2021) Feature engineering for subseasonal to seasonal warm-season precipitation forecasts in the Midwestern US: towards a unifying hypothesis of anomalous warm-season hydroclimatic circulation. *J. Climate*, **34**, 8291-8318.

Delworth, T., & Manabe, S. (1989). The influence of soil wetness on near-surface atmospheric variability. *J. Climate*, **2**, 1447–1462.

Ek, M. B., & Holtslag, A. A. M. (2004). Influence of soil moisture on boundary layer cloud development. *J. Hydrometeor*., **5**, 86–99.

Fast, J. D., & McCorcle, M. D. (1990). A two-dimensional numerical sensitivity study of the Great Plains low-level jet. *Mon. Wea. Rev.*, **118**, 151–163.

Fast, J. D., & McCorcle, M. D. (1991). The effect of heterogeneous soil moisture on a summer baroclinic circulation in the central United States. *Mon. Wea. Rev*., **119**, 2140–2167.

Gelfand, A. E. (2000). Gibbs sampling. *J. Am. Stat. Assoc*. **95**, 1300-1304.

Li, L., Schmitt, R. W., Ummenhofer, C. C. & Karnauskas, K. B. (2016b). Implications of North Atlantic sea surface salinity for summer precipitation over the US Midwest: Mechanisms and predictive values. *J. Climate*, **29**, 3143-3159.

Li, L., Schmitt, R. W. & Ummenhofer, C. C. (2018). The role of the subtropical North Atlantic water cycle in recent US extreme precipitation events. *Clim. Dyn*. **50**, 1291-1305.

Lin, T. I., Lee, J. C. & Yen, S. Y. (2007). Finite mixture modeling using the skew normal distribution. *Stat. Sinica*, **17**, 909–927.

McLachlan, G. J., & Peel, D. (2000). Finite Mixture Models (New York: Wiley)

Meehl, G. A., & Washington, W. M. (1988). A comparison of soil moisture sensitivity in two global climate models. *J. Atmos. Sci*., **45**, 1476–1492.

Melnykov, V., & Maitra, R. (2010). Finite mixture models and model-based clustering. *Stat. Surv*., **4**, 80–116.

Richardson, S., & Green P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. R. Stat. Soc. B*, **59**, 731–792.

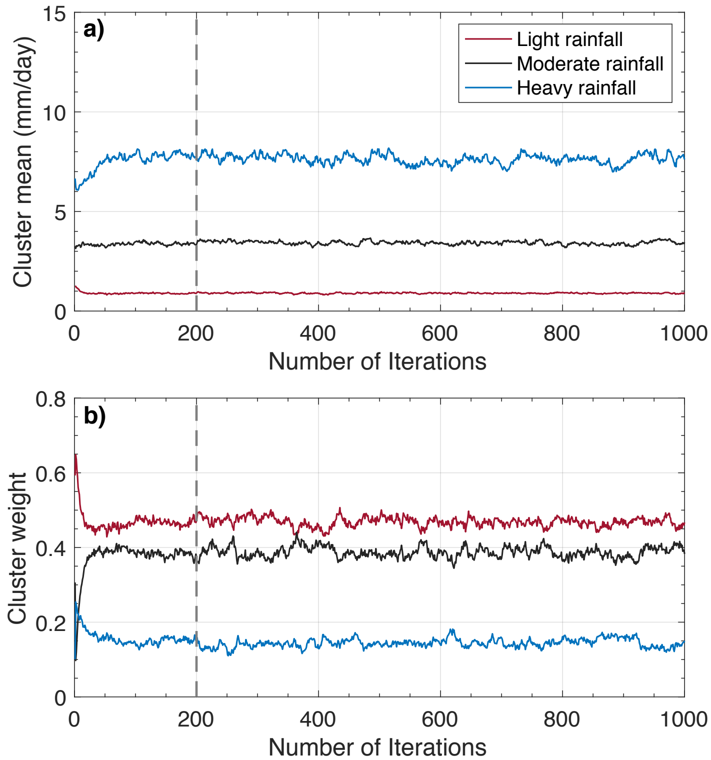

Figure S1 a) Cluster mean (μ_h) and b) cluster weight (π_h) calculated from each iteration of the Gibbs sampler using MCMC algorithm. The red, black, and blue lines corresponding to the light, moderate, and heavy rainfall clusters. The dashed gray lines mark the 200th iteration before which the samples are discarded.

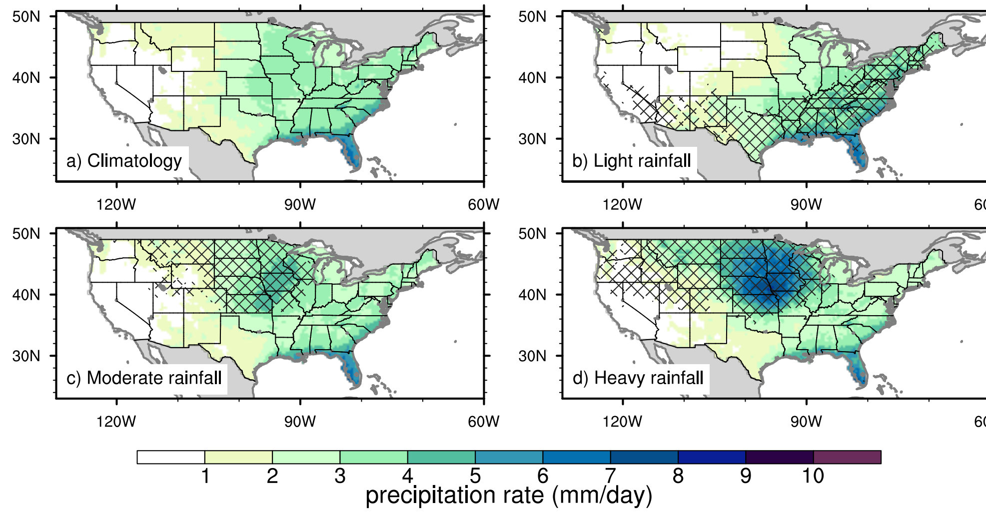

Figure S2 Spatial distribution of US June-July-August (JJA) precipitation corresponding to different categories of US Midwest precipitation: a) 1948-2019 climatology; b) light rainfall composite; c) moderate rainfall composite; and d) heavy rainfall composite. The hatched regions are where precipitation is significantly higher than climatology at $\alpha = 0.05$ level based on 1000-trial Monte Carlo simulations.

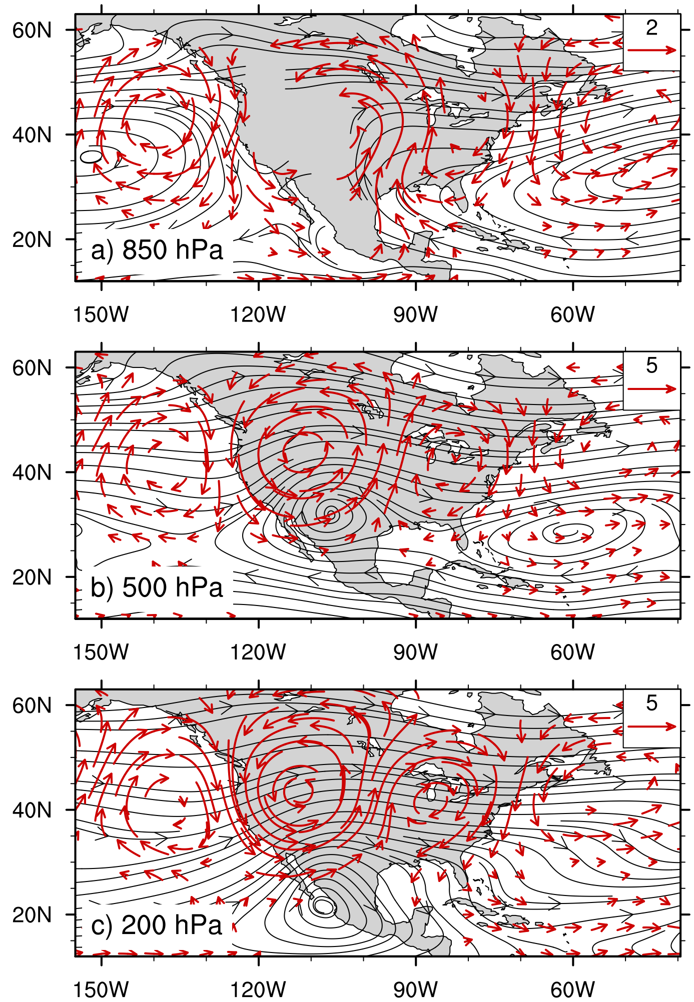

Figure S3 Atmospheric circulation associated with heavy precipitation events: a) 850hPa; b) 500hPa; and c) 200hPa. The streamlines are wind pattern associated with normal cases, and the red arrows are the wind anomalies composite upon heavy precipitation events in the US Midwest. Only wind anomalies statistically significant at $\alpha = 0.05$ level based on 1000-iteration of Monte-Carlo simulations are shown.

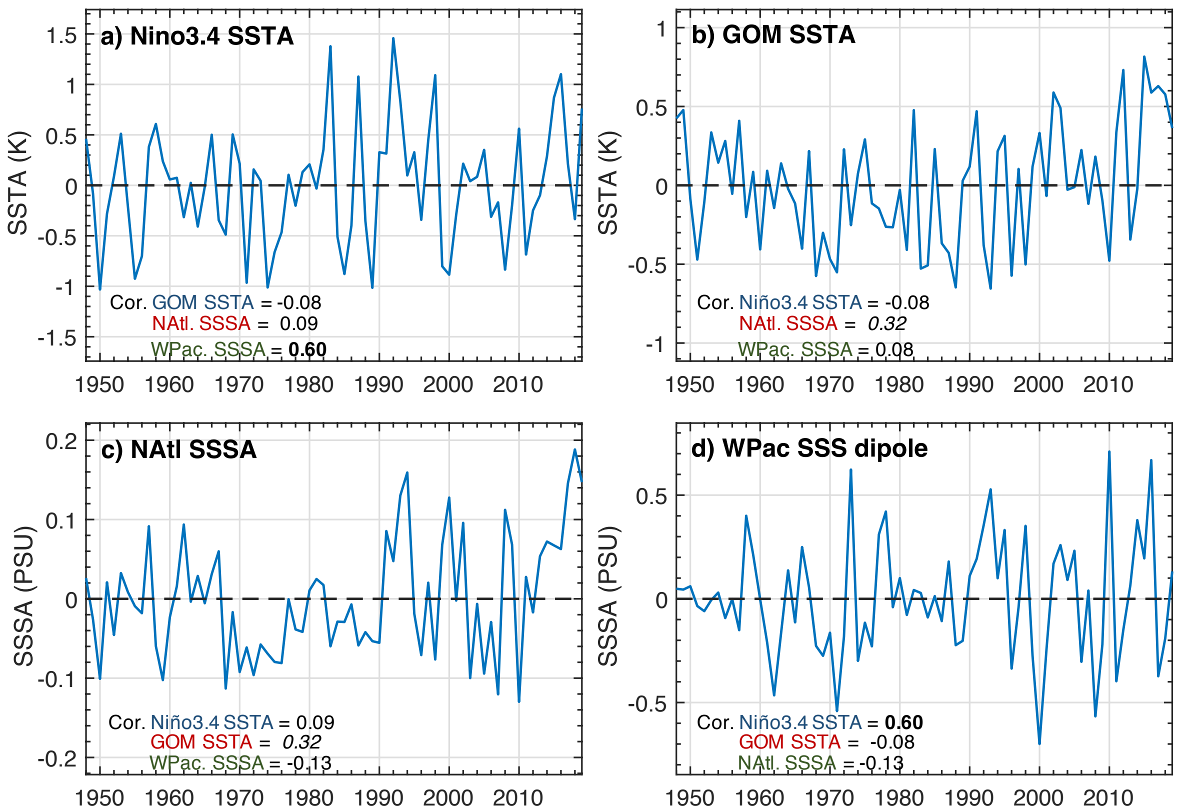

Figure S4 Time series (1948-2019) of the MAM SSTA and SSSA indices that are significantly correlated with the frequency of heavy rainfall events in the US Midwest: a) Niño3.4 SSTA, b) GOM SSTA, c) North Atlantic SSSA, and d) North Pacific SSSA dipole. Cross-correlation coefficients between predictors are reported. The bold font (italic) denotes correlation significant at $\alpha = 0.01(0.05)$ level.

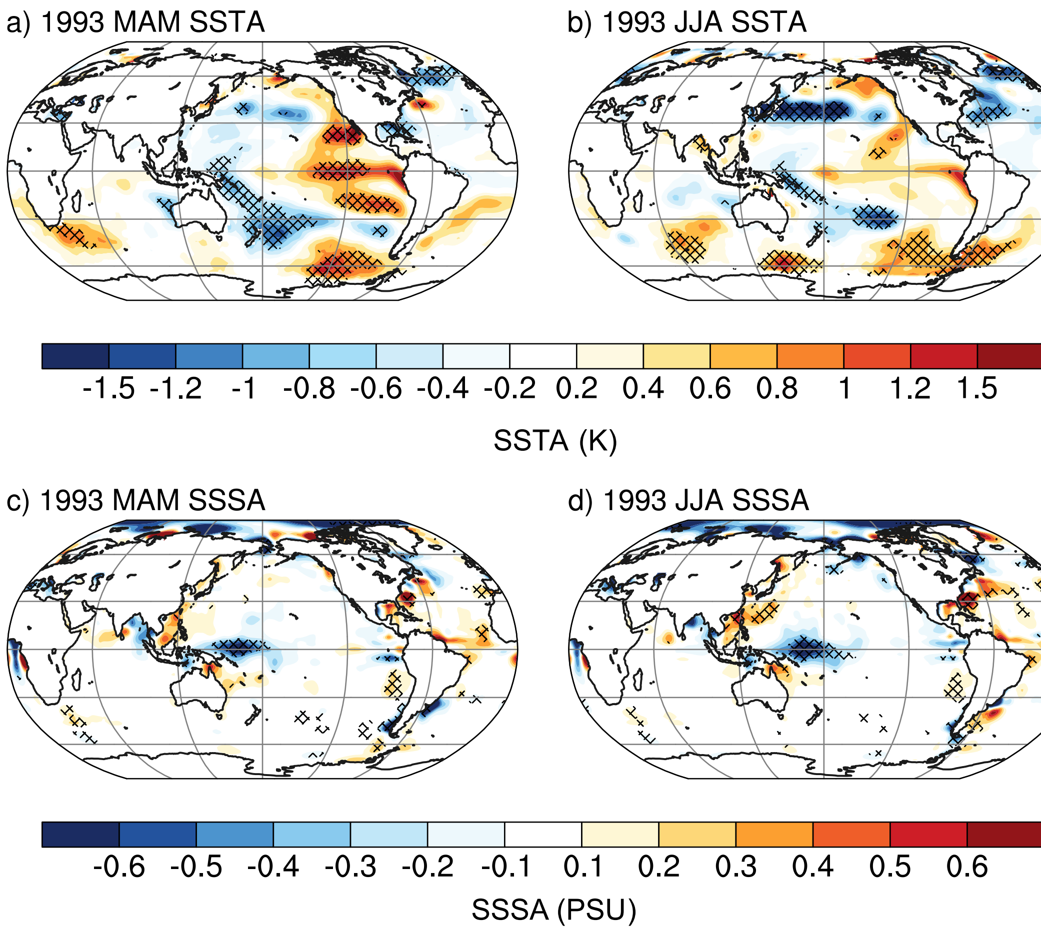

Figure S5 Global SSTA (shaded, unit: K) and SSSA (shaded, unit: PSU) in the year 1993: a) MAM SSTA, b) JJA SSTA, c) MAM SSSA, d) JJA SSSA. The hatched areas indicate where the SSTA and SSSA is significant at the 0.05 level.

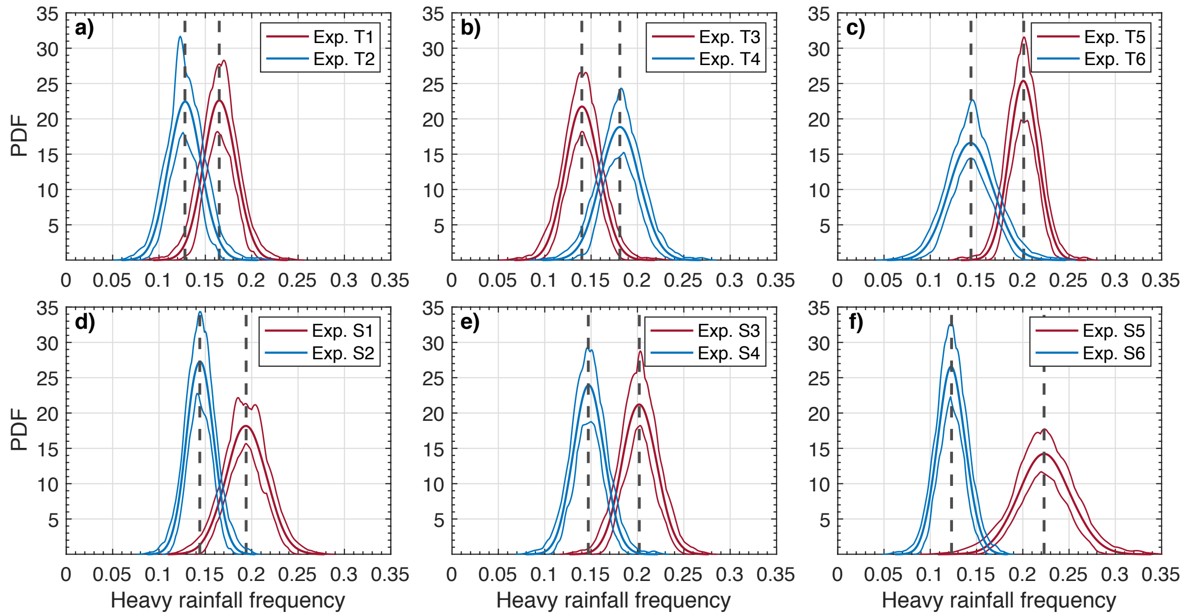

Figure S6 Bayesian inference on the heavy rainfall frequency based on pre-season SSTA (a-c) and SSSA (d-f) indices. The thick curves are the probability density function fitted to heavy rainfall frequency (π_3) sampled in accordance with SSTA or SSSA indices (see Table S1 for detail). The thin curves are the upper and lower bound of the fitted PDF, and the vertical dashed lines are the mode of the PDF (i.e., where the maximum PDF is obtained).

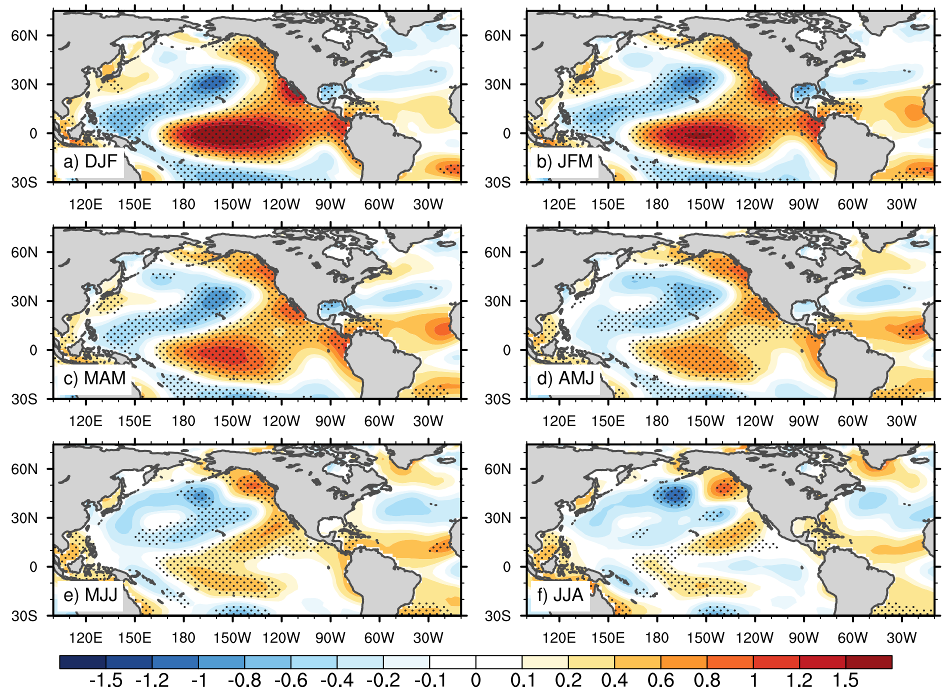

Figure S7 Seasonal evolution of SSTA (shaded; unit: K) composite on tropical western Pacific SSSA dipole index (Figure S2d). The grid cells with SSTA significant at $\alpha = 0.05$ level based on 1000-trial Monte-Carlo simulations are stippled.

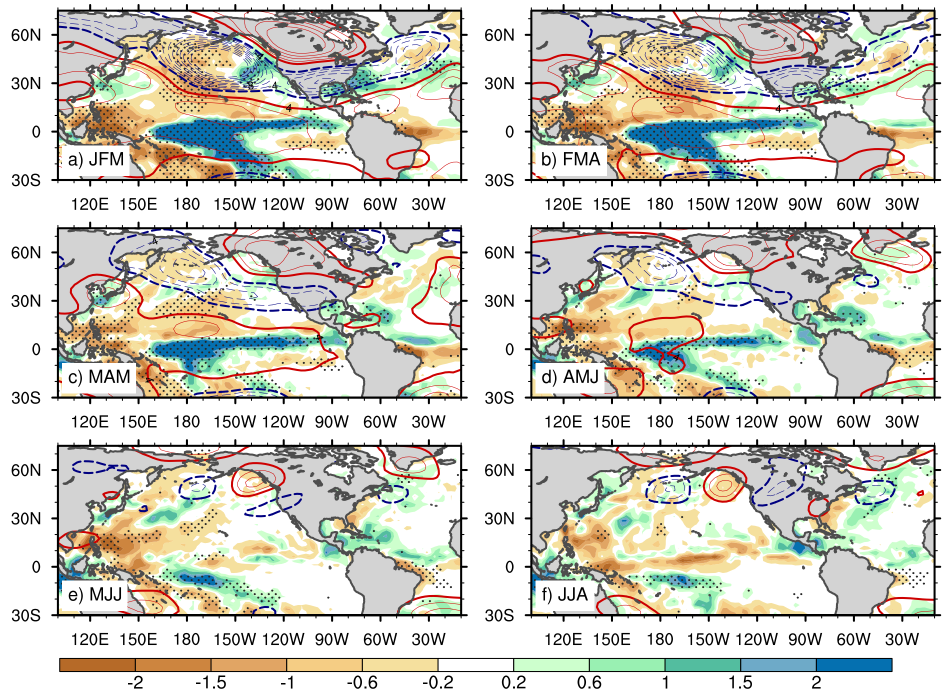

Figure S8 Seasonal evolution of precipitation (shaded; unit: mm day^{-1}) and 500hPa geopotential height (contours; interval is 10 geopotential meter) anomalies composite on the North Pacific salinity dipole index. The stippled grid cells are where precipitation anomalies are significant at $\alpha = 0.05$ level. The solid (dashed) contours represent positive (negative) anomalies in 500hPa geopotential height, and only the anomalies significant at $\alpha = 0.05$ level are shown. The significance test is performed using the 1000-trial Monte Carlo simulation.

Table S1 | Bayesian inference on US Midwest heavy rainfall frequency (π_3) based on the daily precipitation subsampled according to pre-season (MAM) SSTA and SSSA indices.

	Experiment ID	Indices	Years selected
SSTA	Exp. T1	Nino3.4>1.5 STD	1983, 1987, 1992, 1998, 2016
	Exp. T2	Nino3.4<-1.5 STD	1950, 1971, 1974, 1985, 1989, 1999, 2000, 2008
	Exp. T3	GOM SSTA > 1.5 STD	1948, 1949, 2012, 2015, 2017
	Exp. T4	GOM SSTA < -1.5 STD	1968, 1988, 1993, 1996, 1998
	Exp. T5	Nino3.4>1.0 STD &GOM SSTA<-1.0 STD	1983, 1987, 1993, 1998, 2010
	Exp. T6	Nino3.4<-1.0 STD &GOM SSTA>1.0 STD	1955, 1975, 2011
SSSA	Exp. S1	Pacific Dipole>1.5 STD	1973, 1993, 2010, 2016
	Exp. S2	Pacific Dipole<-1.5 STD	1962, 1971, 2000, 2008, 2011
	Exp. S3	NAtl SSSA>1.5 STD	1957, 1962, 1993, 1994, 2000, 2017
	Exp. S4	NAtl SSSA<-1.5 STD	2003, 2005, 2007, 2010
	Exp. S5	NPac Dipole >1.0 STD & NAtl SSSA>1.0 STD	1992, 1993, 2014, 2016
	Exp. S6	NPac Dipole <-1.0 STD & NAtl SSSA<-1.0 STD	1968, 1971, 1974, 1976, 1996