AGU PUBLICATIONS

Geophysical Research Letters

Supporting Information for

Improving low-cloud fraction prediction through machine learning

Haipeng Zhang[1,2], Youtong Zheng[3,4], and Zhanqing Li[1,2]

[1]Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD, USA

[2]Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA

[3]Department of Atmospheric and Earth Science, University of Houston, Houston, TX, USA

[4]Institute of Climate and Atmospheric Science, University of Houston, Houston, TX, USA

Contents of this file

Texts S1 to S2
Tables S1 to S4
Figures S1 to S4

Texts

Text S1. Building XGB10 and XGB7

To determine the best combination of the hyperparameters in XGB10 and XGB7, the optimization space of six key parameters (learning_rate, n_estimators, subsample, colsample_bytree, and colsample_bylevel) was explored using a Bayesian optimization technique (Snoek et al., 2012). This technique is highly efficient for parameter tuning, allowing finding the maximum value of a target function in as few iterations as possible based on Bayesian inference and the Gaussian process. To avoid overfitting on the training set, we used 5-fold cross-validation in each iteration during the optimization processes. The mean squared error (MSE) for the validation set reached the minimum, usually within 10 iterations. The investigated ranges of each parameter for hyperparameter optimization in XGB10 and XGB7 were summarized in Table S2. To address the uncertainties caused by random seeds in optimization processes, we performed Bayesian optimization ten times. This resulted in ten optimized parameter sets, creating ten ensemble members each for XGB10 and XGB7, as summarized in Tables S3 and S4, respectively. The ensemble-mean prediction results and absolute SHAP values from each model were used for evaluation in this study.

Text S2. SHAP Explainability analysis

SHAP (SHapley Additive exPlanations; Lundberg et al., 2018; Lundberg & Lee, 2017) is a high-fidelity and unified approach to exploring and interpreting the tree-based ML model (e.g., XGBoost) behavior. It explains a model's individual output as a sum of the contributions of each feature (or predictor) and the mean predicted value through an explanation model, which can be expressed as:

$$y = \bar{y} + \sum_i \phi_i, \qquad (1)$$

where y is the final prediction for one case, $\bar{y}$ is the average prediction across all cases, and ϕ_i is the contribution of the i-th feature to the prediction for this case (called SHAP values). Based on cooperative game theory, Lundberg & Lee, (2017) first proposed the KernelSHAP algorithm to calculate SHAP values by sampling the predictions of a machine learning model by replacing feature values with random values from the feature distribution. But KernalSHAP is computationally slow and ignores feature dependence. A more efficient and exact algorithm for tree ensemble models (TreeSHAP) was developed using the conditional expectation to estimate feature effects with no feature independence assumption required (Lundberg et al., 2018, 2020). The features with larger absolute SHAP values contribute more to a prediction than those with smaller values.

Tables:

Table S1. Summary of meteorological factors used as predictors in XGB10 and XGB7

Model	Predictor	Description
XGB10	θ_{1000}, θ_{850}, θ_{700}	Potential temperature at 1000, 850, and 700 hPa (K)
	RH_{1000}, RH_{850}, RH_{700}	Relative humidity at 1000, 850, and 700 hPa (%)
	U_{1000}	Horizontal wind speed at 1000 hPa (m/s)
	ω_{700}	Vertical velocity at 700 hPa (Pa/s)
	LHF	Latent heat flux (W/m^2)
	PWV	Column-integrated precipitable water vapor (kg/m^2)
XGB7	RH_{700}, U_{1000}, ω_{700}, LHF	Same as those predictors used in XGB10
	LTS	Lower-tropospheric stability ($\theta_{700} - \theta_{1000}$) (K)
	Δq	The moisture contrast between the boundary layer and free troposphere ($q_{1000} - q_{700}$) (g/kg)
	T_{adv}	Horizontal temperature advection (K/day)

Table S2. Investigated range of hyperparameters for XGB10 and XGB7

Parameter	Investigated range
max_depth	3-7
n_estimators	100-1000
learning_rate	0.01-1.0
subsample	0.5-1.0
colsample_bytree	0.5-1.0
cosample_bylevel	0.5-1.0

Table S3. Optimized hyperparameters and training/test errors for ten XGB10 members

Member ID	**colsample _bylevel**	**colsample _bytree**	**learning_rate**	**max_depth**	**n_estimators**	**subsample**	**MSE for trainning set**	**MSE for test set**
01	0.6800	0.7890	0.6129	6	857	0.9990	0.0701	0.0739
02	0.7347	0.5853	0.5122	6	914	0.7131	0.0702	0.0740
03	0.8588	0.9309	0.6960	6	598	0.9372	0.0703	0.0741
04	0.7295	0.9263	0.3387	6	844	0.7752	0.0699	0.0737
05	0.9966	0.9833	0.2730	6	861	0.9342	0.0699	0.0737
06	0.7742	0.8292	0.4596	6	737	0.7562	0.0700	0.0738
07	0.8666	0.7718	0.2392	6	606	0.6046	0.0704	0.0741
08	0.8208	0.7708	0.3453	6	922	0.7314	0.0699	0.0737
09	0.7125	0.9662	0.3404	6	904	0.9243	0.0699	0.0737
10	0.5488	0.7314	0.5591	6	621	0.8558	0.0704	0.0742

Table S4. Optimized hyperparameters and training/test errors for ten XGB7 members

Member ID	colsample _bylevel	colsample _bytree	learning_rate	max_depth	n_estimators	subsample	MSE for trainning set	MSE for test set
01	0.8515	0.7095	0.4283	6	475	0.7146	0.0816	0.0842
02	0.6609	0.9254	0.4318	6	837	0.6013	0.0813	0.0840
03	0.5205	0.9203	0.4727	6	999	0.8017	0.0813	0.0840
04	0.8626	0.7625	0.6250	6	637	0.9427	0.0814	0.0841
05	0.8614	0.9740	0.3785	6	781	0.5637	0.0813	0.0840
06	0.9062	0.5898	0.3028	6	619	0.9218	0.0815	0.0841
07	0.9456	0.7859	0.6579	5	871	0.8049	0.0815	0.0841
08	0.9559	0.5823	0.4667	6	994	0.6967	0.0814	0.0840
09	0.7334	0.9058	0.7428	6	843	0.8529	0.0813	0.0841
10	0.9634	0.7916	0.8672	4	937	0.7346	0.0818	0.0843

Figures:

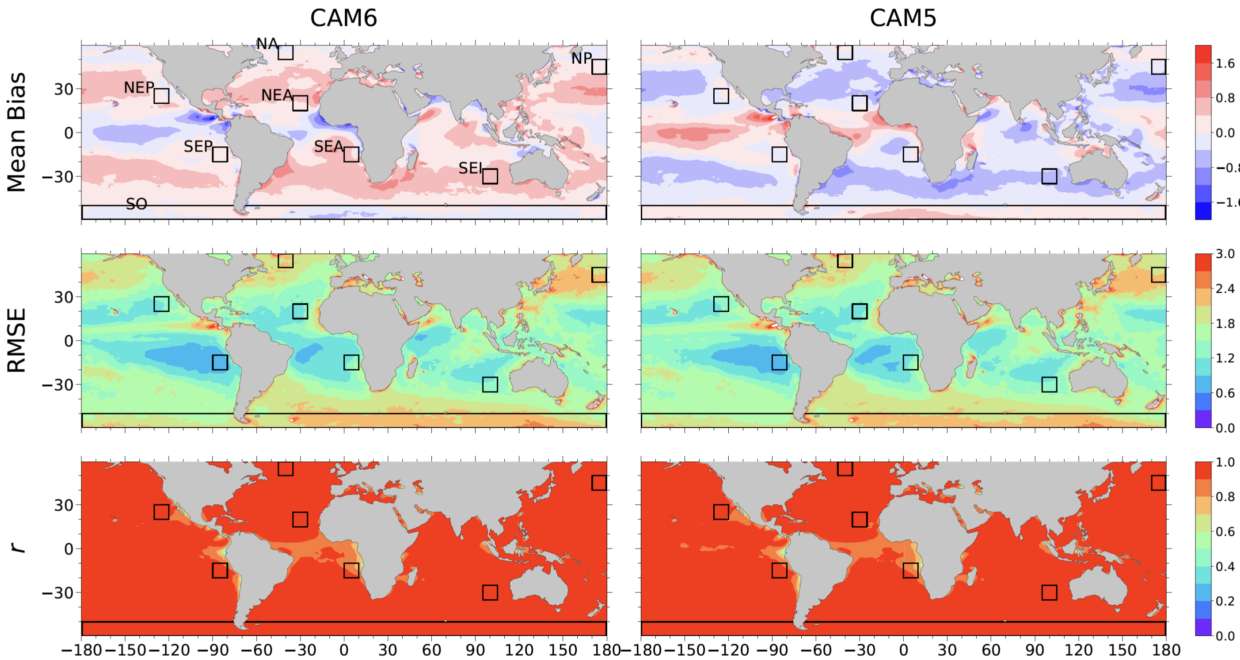

Figure S1. Comparison of near-surface zonal wind speeds between model nudging experiment outputs and ERA5 regarding three metrics: the mean bias, root mean squared error (RMSE), and correlation coefficients (*r*) (from top to bottom rows, respectively). The first column shows the comparison for CAM6 outputs, with the second column for CAM5 outputs.

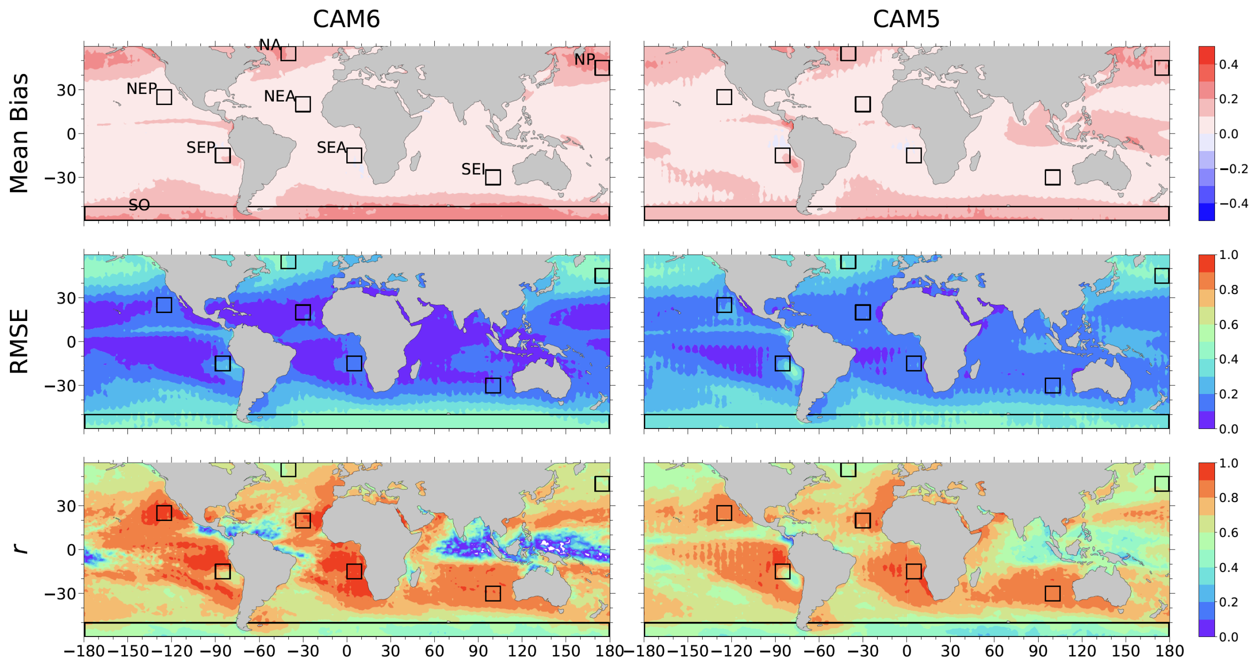

Figure S2. Same as Figure S1 but for comparisons between non-COSP low-cloud fraction (LCF) and COSP-enabled LCF in CAM6 and CAM5.

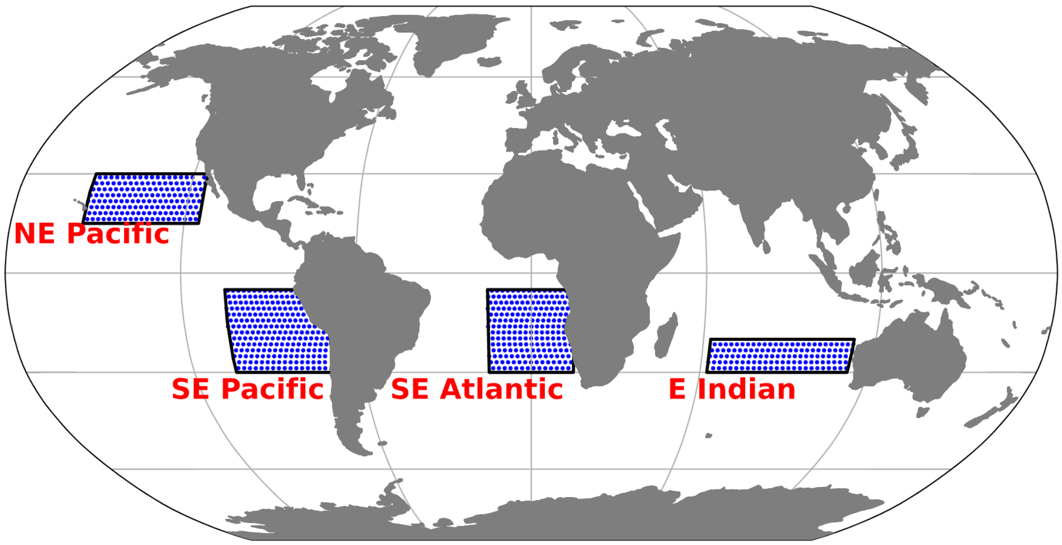

Figure S3. Starting points of trajectories sampled over the four selected regions where the stratocumulus-to-cumulus transition dominates, following Eastman & Wood (2018): Northeast Pacific (-155 to -115ºE, 15 to 30ºN), Southeast Pacific (-105 to -70ºE, -30 to -5ºN), Southeast Atlantic (-15 to 15ºE, -30 to -5ºN), and East Indian (62.5 to 112.5ºE, -30 to -20ºN).

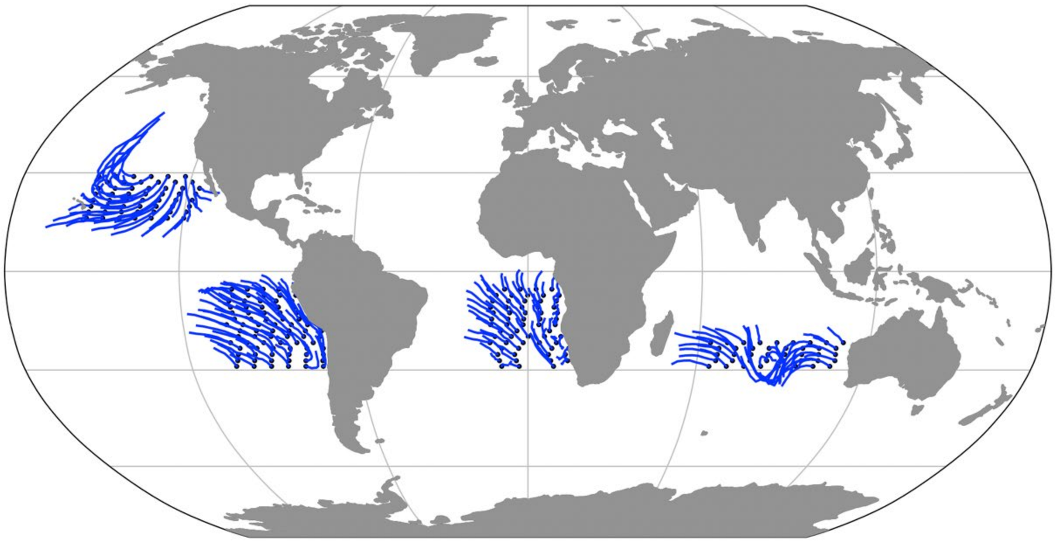

Figure S4. Subsets of forward trajectories (36 hours) over the four subtropical regions starting at 6:00 p.m. on March 31, 2004. The black points denote the starting points.

Reference

Eastman, R., & Wood, R. (2018). The competing effects of stability and humidity on subtropical stratocumulus entrainment and cloud evolution from a Lagrangian perspective. *Journal of the Atmospheric Sciences*, *75*(8), 2563–2578. https://doi.org/10.1175/JAS-D-18-0030.1

Lundberg, S. M., & Lee, S.-I. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-Decem*(Section 2), 4766–4775. Retrieved from http://arxiv.org/abs/1705.07874

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. Retrieved from http://arxiv.org/abs/1802.03888

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Snoek, J., Larochelle, H., Adams, R. P., & Jeffery, C. (2012). Practical Bayesian optimization of machine learning algorithms. *Religion and the Arts*, *17*(1–2), 57–73. https://doi.org/10.48550/ARXIV.1206.2944