AGU PUBLICATIONS

Geophysical Research Letters

Supporting Information for

Evaluation of stratocumulus evolution under contrasting temperature advections in CESM2 through a Lagrangian framework

Haipeng Zhang[1,2], Youtong Zheng[3,4], 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

Text S1. Predict LCF/CLWP variations using XGBoost

The XGBoost model is adopted to predict LCF or CLWP variations (ΔLCF/ΔCLWP) through thirteen input features consisting of the initial values (at 0 hr) of LCF/CLWP and six dominant meteorological factors and the 36-hr variations in these meteorological factors: $LCF_0/CLWP_0$, EIS_0, SST_0, $(q_{700})_0$, $(dq)_0$, $(\omega_{700})_0$, $(U_{925})_0$, ΔEIS, ΔSST, Δq_{700}, $\Delta(\mathrm{d}q)$, Δw_{700}, and ΔU_{925}, where Δ denotes temporal variations between 36 hr and 0 hr along the trajectory, EIS is estimated inversion strength, SST is sea surface temperature, q_{700} is free-tropospheric moisture, dq is moisture difference between 700 hPa and 1000 hPa, ω_{700} is large-scale subsidence rate at 700 hPa, and U_{925} is surface wind speed.

When predicting ΔLCF/ΔCLWP, XGBoost models are developed under CADV and WADV conditions using observational data and CESM2 outputs, respectively. Each model is developed on millions of selected trajectory samples (see Section 2.3), with 80% used for training and the remaining 20% for testing the model's performance. All the features are standardized by removing their means and scaling them to unit variance. The predictant variables, ΔLCF/ΔCLWP, are linearly transformed to between 0 and 1. Two crucial hyperparameters, n_estimators and max_depth, are fine-tuned to avoid overfitting by making the mean square root (MSE) difference between the training and test datasets remain within a 20% range. The performance metrics for the XGBoost models are summarized in Table S1.

As a comparison, a multiple linear regression model (MLR) is used to predict ΔLCF/ΔCLWP. The model inputs and data pre-processing remain the same as those for the XGBoost model. The performance metrics for the MLR models are summarized in Table S2.

Table S1. Summary of MSE and explained variance score for the XGBoost models

	MSE for training dataset	MSE for test dataset	Explained variance score
Models for predicting ΔLCF			
Based on observations			
Under CADV	0.0191	0.0228	0.470
Under WADV	0.0261	0.0314	0.415
Based on CESM2 outputs			
Under CADV	0.0355	0.0425	0.460
Under WADV	0.0371	0.0444	0.405
Models for predicting ΔCLWP			
Based on observations			
Under CADV	0.000180	0.000216	0.655
Under WADV	0.000377	0.000450	0.608
Based on CESM2 outputs			
Under CADV	0.000693	0.000828	0.730
Under WADV	0.00202	0.00241	0.555

Note: The explained variance score represents the proportion of the variance in the predictant variable that can be explained by the predictor variables in XGBoost. The higher the score, the more the built model is able to explain the variation in the data.

Table S2. Summary of MSE and explained variance score for the MLR models

	MSE for training dataset	MSE for test dataset	Explained variance score
Models for predicting ΔLCF			
Based on observations			
Under CADV	0.0248	0.0248	0.314
Under WADV	0.0333	0.0330	0.255
Based on CESM2 outputs			
Under CADV	0.0453	0.0460	0.311
Under WADV	0.0459	0.0456	0.264
Models for predicting ΔCLWP			
Based on observations			
Under CADV	0.000223	0.000198	0.572
Under WADV	0.000573	0.000412	0.400
Based on CESM2 outputs			
Under CADV	0.00109	0.000988	0.574
Under WADV	0.00258	0.00264	0.431

Figures:

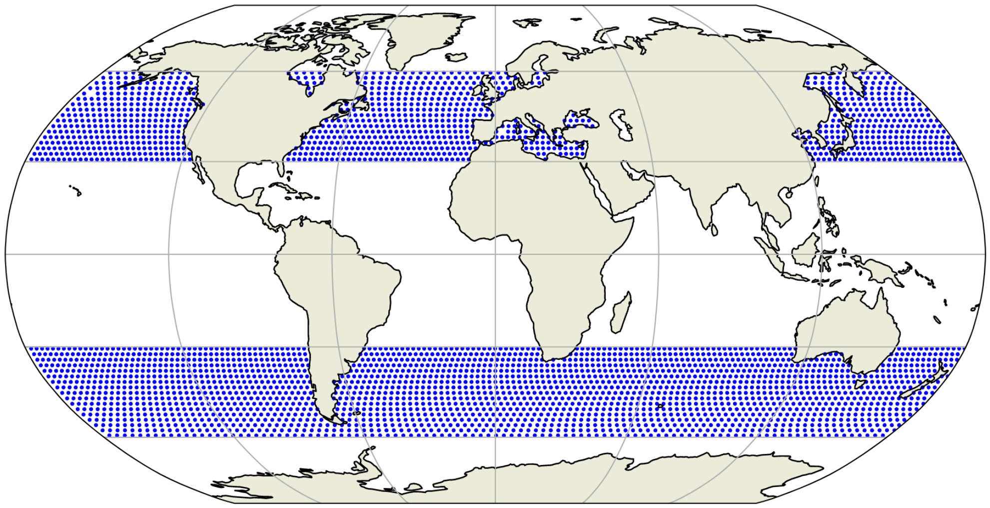

Figure S1. Global distribution of starting points of trajectories sampled over the midlatitude oceans.

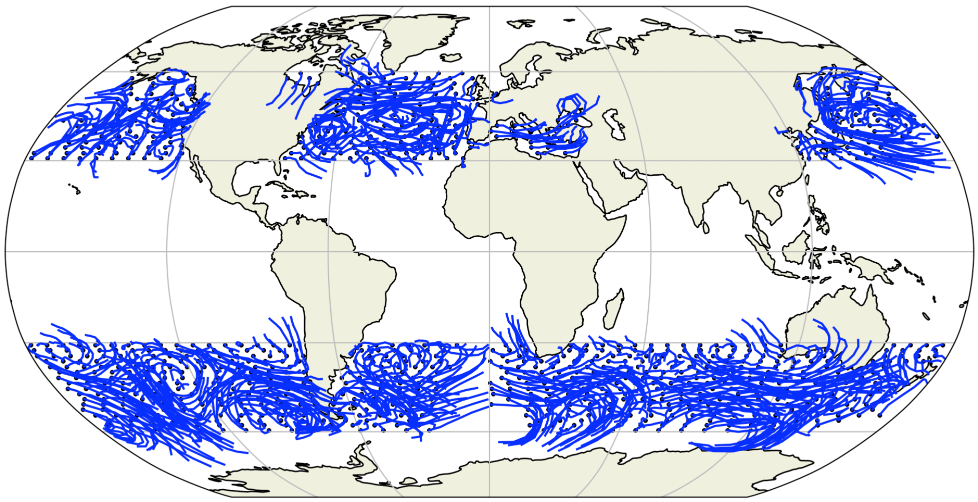

Figure S2. Subsets of forward trajectories (36 hours) over the midlatitude oceans starting at 12:00 pm on January 1st, 2010. The black points denote the starting points.

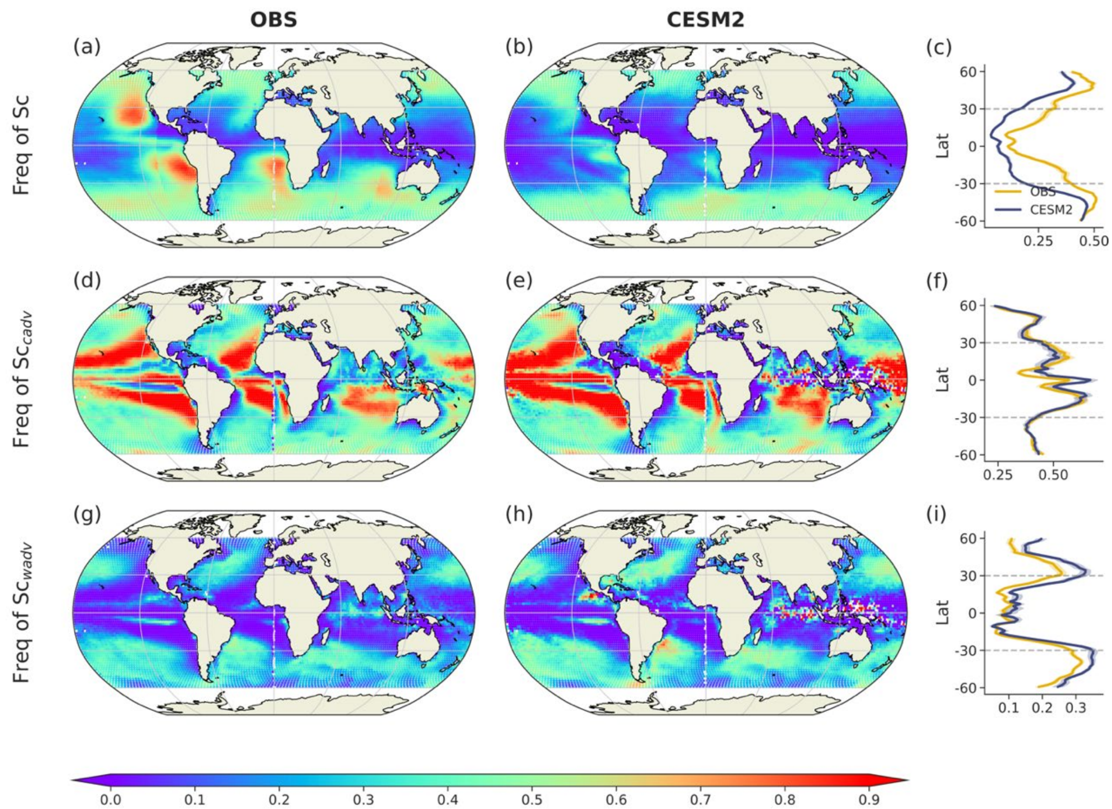

Figure S3. Global maps of the occurrence frequency of stratocumulus (Sc) clouds for (a) CERES SYN Ed4 product and (b) CESM2, with (c) illustrating their zonal mean distribution. Panels (d-f) are the same as (a-c), but for the frequency of clouds experiencing cold-air advection, normalized by all-type Sc occurrence frequency. Similarly, panels (g-i) display the occurrence frequency of Sc clouds experiencing warm-air advection. The shading in panels (c, f, i) represents one standard error (SE), which is calculated as SE=$\sigma/\sqrt{n}$ with σ the standard deviation and n the total sample size. Note that to ensure the even sample size of Sc clouds globally the criteria for single-layer clouds is not implemented for this map.

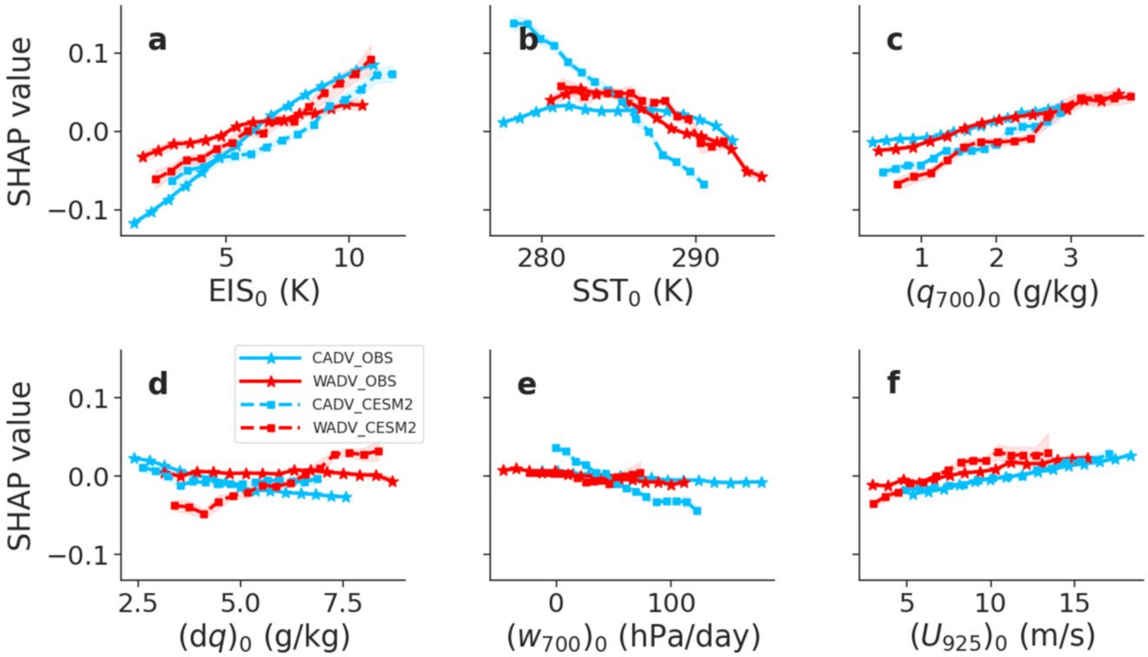

Figure S4. Dependence plot of SHAP values on initial values of meteorological factors when predicting 36-hr variations in LCF (ΔLCF). SHAP values are sorted into 15 equal bins based on the values of each influential factor. The error bars represent the expanded standard error of 10 for visualization.

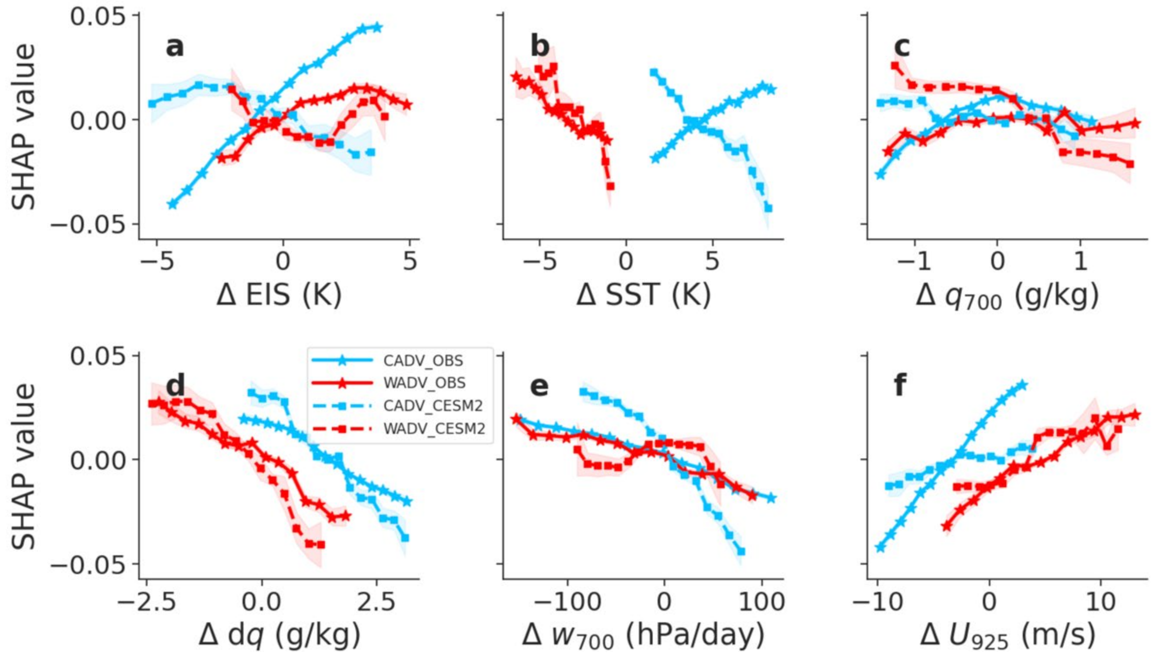

Figure S5. Dependence plot of SHAP values on 36-hr variations (denoted by Δ) in meteorological factors when predicting ΔLCF. SHAP values are sorted into 15 equal bins based on the values of each influential factor. The error bars represent the expanded standard error of 10 for visualization.

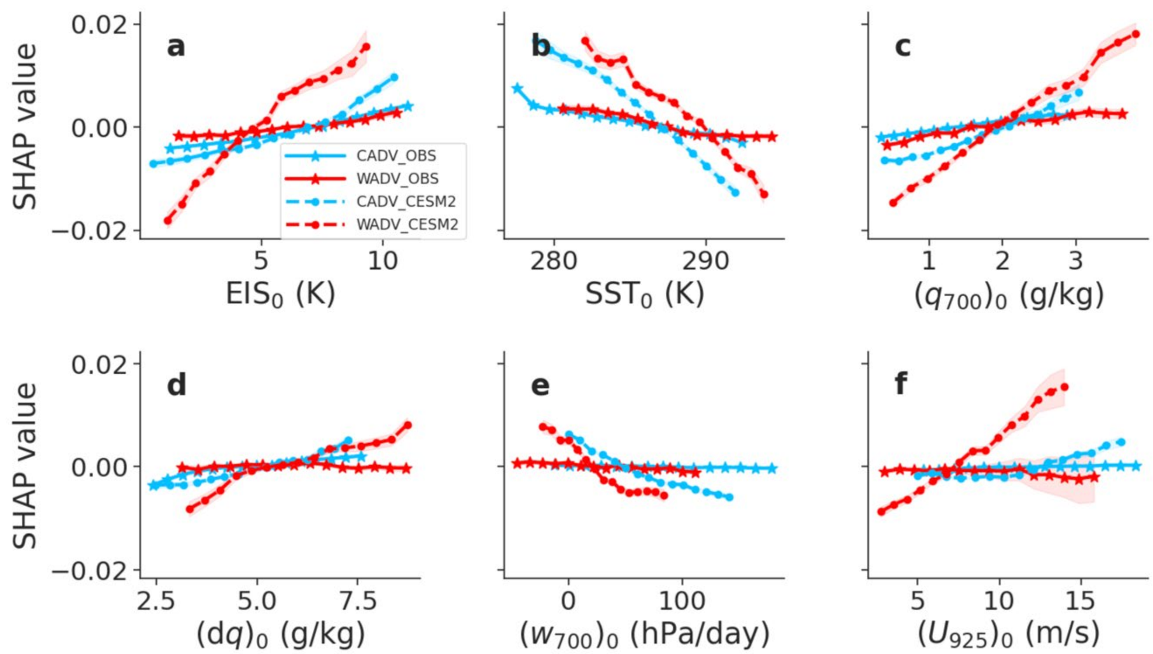

Figure S6. Same as Figure S4, but for ΔCLWP.

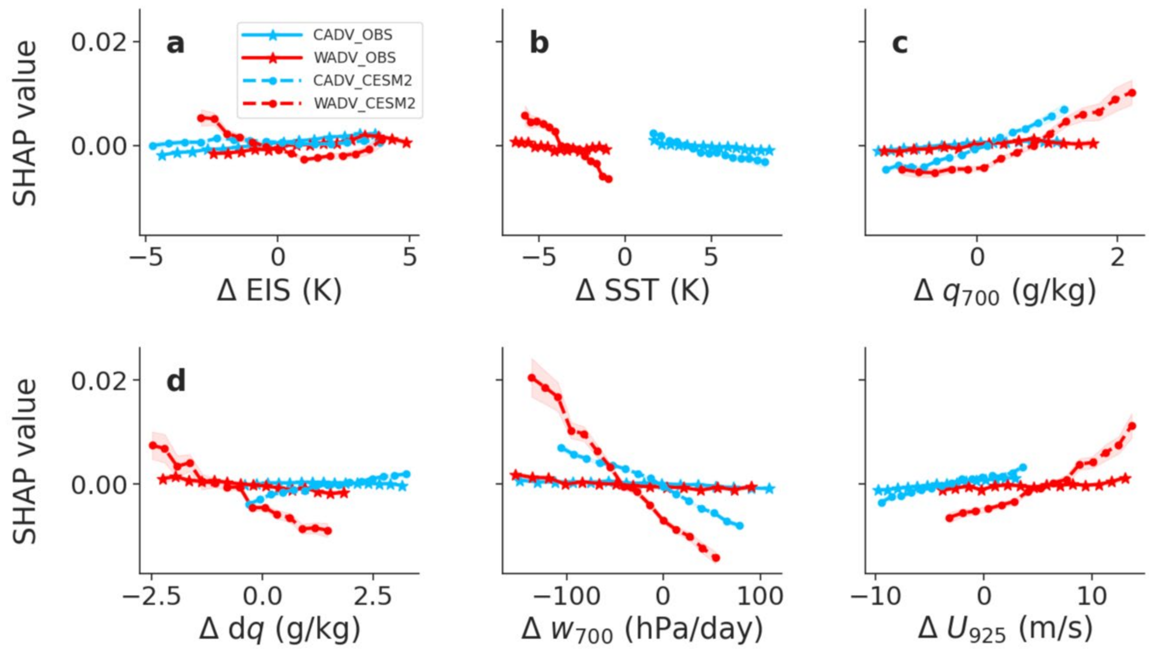

Figure S7. Same as Figure S5, but for ΔCLWP.

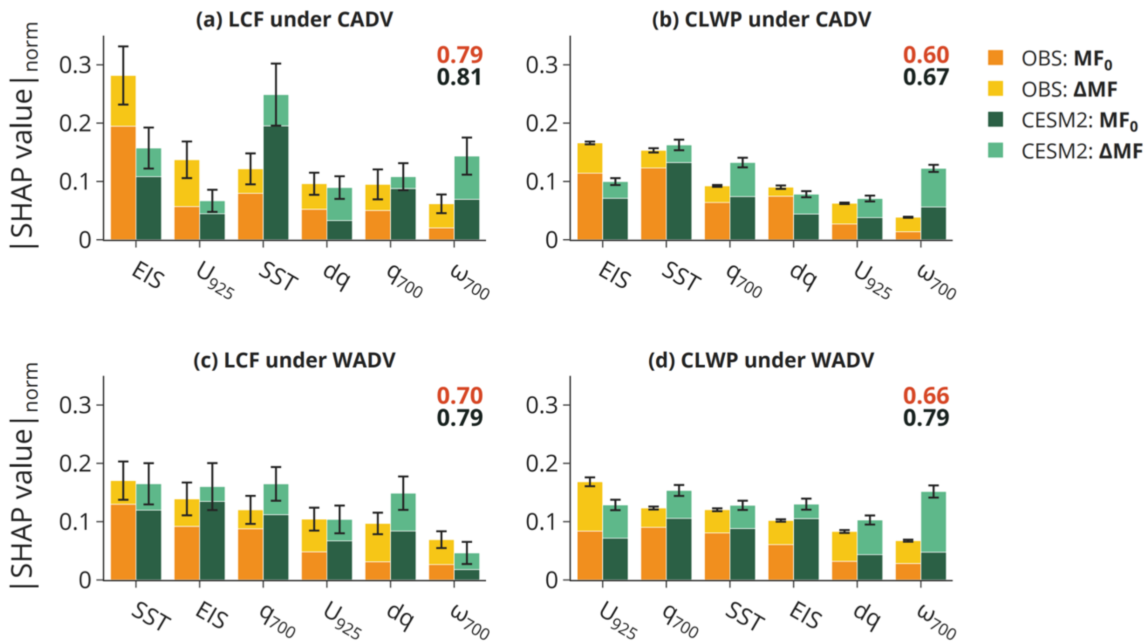

Figure S8. Barplot of normalized SHAP absolute values for predictors when predicting 36-hr variations in LCF (or ΔLCF) through observational data (orange) or CESM2 modeling results (green) under CADV conditions (a) and WADV conditions (c), respectively. Panels (b) and (d) show the same ones, but for predicting changes in CLWP. The predictors here include initial values (MF_0; dark color) and 36-hr variations (ΔMF; light color) in each meteorological factor. Each error bar exhibits the standard deviation of a given meteorological factor, calculated as $\sigma = \sqrt{\sigma_1 + \sigma_2}$, where σ_1 and σ_2 are the standard deviations of initial values and 36-hr variations in this factor, respectively. The red number in each panel shows the sum of SHAP values of all meteorological factors from observations, with the black number from CESM2 results.