Predictive Statistical Representations of Observed and Simulated Rainfall Using Generalized Linear Models

JUNHO YANG AND MIKYOUNG JUN

Department of Statistics, Texas A&M University, College Station, Texas

COURTNEY SCHUMACHER AND R. SARAVANAN

Department of Atmospheric Sciences, Texas A&M University, College Station, Texas

(Manuscript received 15 August 2018, in final form 8 March 2019)

ABSTRACT

This study explores the feasibility of predicting subdaily variations and the climatological spatial patterns of rain in the tropical Pacific from atmospheric profiles using a set of generalized linear models: logistic regression for rain occurrence and gamma regression for rain amount. The prediction is separated into different rain types from TRMM satellite radar observations (stratiform, deep convective, and shallow convective) and CAM5 simulations (large-scale and convective). Environmental variables from MERRA-2 and CAM5 are used as predictors for TRMM and CAM5 rainfall, respectively. The statistical models are trained using environmental fields at 0000 UTC and rainfall from 0000 to 0600 UTC during 2003. The results are used to predict 2004 rain occurrence and rate for MERRA-2/TRMM and CAM5 separately. The first EOF profile of humidity and the second EOF profile of temperature contribute most to the prediction for both statistical models in each case. The logistic regression generally performs well for all rain types, but does better in the east Pacific compared to the west Pacific. The gamma regression produces reasonable geographical rain amount distributions but rain rate probability distributions are not predicted as well, suggesting the need for a different, higher-order model to predict rain rates. The results of this study suggest that statistical models applied to TRMM radar observations and MERRA-2 environmental parameters can predict the spatial patterns and amplitudes of tropical rainfall in the time-averaged sense. Comparing the observationally trained models to models that are trained using CAM5 simulations points to possible deficiencies in the convection parameterization used in CAM5.

1. Introduction

One of the strongest El Niño events in recent decades occurred in the Pacific Ocean in 2015. Although it was expected to bring sufficient amount of rain to ease the California drought, California remained significantly dry. Prediction of the remote influence of El Niño relies crucially upon an accurate prediction of rainfall patterns in the tropical Pacific region. However, climate model simulations of rainfall and sea surface temperature (SST) in this region suffer from pervasive biases. The most notable biases are the excessive equatorial Pacific cold tongue and the associated double intertropical convergence zone (ITCZ; Li and Xie 2014; Oueslati and Bellon 2015). These biases are present in the state-of-the-art climate models that are part of the most recent phase (phase 5) of the Climate Model Intercomparison Project (CMIP5; Taylor et al. 2012), and there seems to be little improvement in these biases between CMIP5 and its previous phase, CMIP3 (Stocker et al. 2013). An accurate understanding of tropical rainfall is critical, as it is not just a matter of predicting local rainfall but also entails the forcing of atmospheric circulation around the globe (Hartmann et al. 1984; Schumacher et al. 2004) and the sensitivity to anthropogenic climate change (Sherwood et al. 2014).

Since 1998, high-quality measurements of rainfall over the tropics have become available via the NASA Tropical Rainfall Measurement Mission (TRMM; Kummerow et al. 1998) and Global Precipitation Measurement (GPM; Hou et al. 2014) satellites with the GPM satellite extending the dataset into higher latitudes since 2014. These high-quality datasets have improved our understanding of rainfall characteristics around the

Corresponding author: Junho Yang, junhoyang@stat.tamu.edu

DOI: 10.1175/JCLI-D-18-0527.1

globe and have been used to improve global climate model (GCM) simulations of rainfall. The most common approach to using these data for model validation is to compare the statistical properties of rainfall such as temporal means and variances, as well as the probability distribution functions of rainfall frequency and intensity, between GCMs and observations (Dai 2006). Temporal correlation properties of rainfall are also sometimes validated, but typically on daily or longer time scales, such as the rainfall anomalies associated with equatorial waves (e.g., Cho et al. 2004) or El Niño–Southern Oscillation (ENSO; e.g., Chen et al. 2007). However, satellite rainfall measurements contain spatiotemporal correlation information on subdiurnal time scales that can be analyzed to validate and improve GCM simulations of rainfall. Chen et al. (2017) analyzed observed statistical relationships between atmospheric state variables and large precipitating system to identify important predictors. In this study, we present an analysis of TRMM satellite observations and climate model output at subdiurnal time scales in the tropical Pacific using flexible statistical models. The goal is to identify purely empirical relationships between different types of rain and the instantaneous state of atmospheric variables such as temperature and humidity. These empirical relationships are evaluated for their skill in predicting the subdiurnal weather variations as well as the time-averaged spatial variance of precipitation climatology. This empirical approach is also used to evaluate the convective parameterization in a GCM.

Our observationally based analysis is complementary to studies that have used empirical approaches to develop parameterizations of subgrid processes. Kuang (2010) and Kelly et al. (2017) described the development and use of a tangent linear model for parameterizing convection using data from a cloud system resolving model. Machine learning approaches are also being increasingly used to construct fast parameterizations with training data obtained from cloud-resolving model simulations (e.g., Krasnopolsky et al. 2013; Brenowitz and Bretherton 2018; O'Gorman and Dwyer 2018; Rasp et al. 2018). Climate models parameterize convection by assuming a physically motivated predictive relationship between the resolved atmospheric state and rainfall amounts. Our statistical analysis attempts to mimic the behavior of such a convection parameterization, in that we identify a statistical predictive relationship between the atmospheric state and the subsequent occurrence/amount of different types of rainfall. However, our goal is not to forecast individual rainfall events as in a weather model, but rather to simulate the slowly varying patterns and intensity of rainfall, as in a climate model. The most important predictors of rainfall are the vertical profiles of temperature and humidity, but additional variables such as the wind field and the associated shear or convergence can also play a role. In a GCM, a snapshot of the atmospheric state at each time step is used to predict rainfall occurring through the next time step. We apply this predictive framework to analyze the empirical relationships between the atmospheric state in reanalysis data from the Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2), at a selected time and satellite measurements of rainfall over a 6-h period after that time. The empirical analysis is then repeated for a GCM simulation using the NCAR Community Atmospheric Model (CAM5), and the results are compared to the observed relationships.

The predictors for rainfall are derived from surface variables and vertical profiles of the observed atmospheric state at each location from MERRA-2, with empirical orthogonal function (EOF) decomposition applied to vertical profiles to select the dominant vertical structures of temperature, humidity, and winds. These predictors are used to predict TRMM observed rainfall occurrence and amount separately; the former is a discrete on–off relationship and the latter is a continuous relationship. We use so-called *generalized linear models* (GLMs; McCullagh and Nelder 1989; Madsen and Thyregod 2010), which can describe the relationship between rainfall and other atmospheric state variables accounting for their non-Gaussian characteristics. In particular, we fit logistic regression models for the occurrence of rain and gamma regression with log link functions for the actual rain amount.

The predictive relationships for rainfall are expected to vary with rain type. Therefore, the analysis is carried out independently for three types of observed rainfall, namely *stratiform* (STR), *deep convective* (DC), and *shallow convective* (SC), derived from radar measurements onboard the TRMM satellite. For the climate model analysis, only two rain types are used as predictands, *large-scale* and *convective*, due to the lack of separate data availability for different convective rain types from GCM data.

Deep convection is associated with strong, intermittent rain and constitutes a large portion of rainfall over tropical land and oceans (Schumacher and Houze 2003) and the extratropical storm tracks. Stratiform clouds are associated with weaker, widespread rainfall that can form either as a result of deep convective clouds, as is common in the tropics, or from large-scale lifting as found in fronts at higher latitudes (Houze 2004). Moreover, the determination of rain type can also help explain extreme rain events. For example, Ahmed

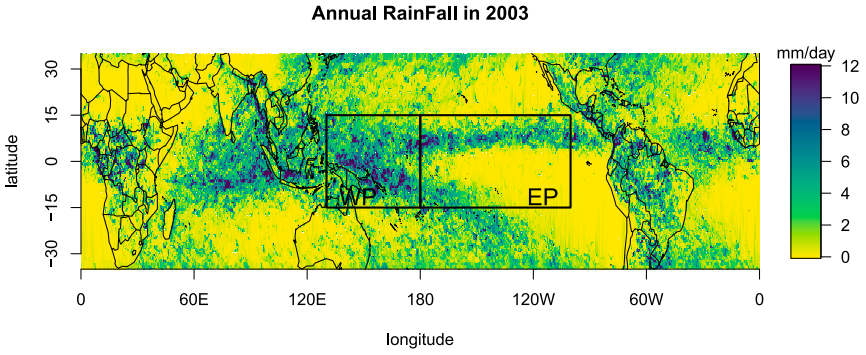

FIG. 1. Mean annual total precipitation (mm day^{-1}) in 2003 from TRMM PR data.

and Schumacher (2015) showed that the exponential increase in rainfall with tropospheric humidity over tropical oceans [originally pointed out by Bretherton et al. (2004)] is due mostly to the increase in stratiform rain area. Deep convective and stratiform conditional rain rates (i.e., rain intensity) increase in a more linear fashion, and while convective rain area also shows an exponential increase, it is much less dramatic than the stratiform rain area increase.

Climate models strongly vary in the amount of rain that they consider large-scale and convective (Dai 2006). It should be noted that large-scale rain from the climate models is not necessarily analogous to stratiform rain observed by radar. Regardless, it is of high relevance to understanding the fundamental parameterizability of each rain type.

The rest of the paper is organized as follows: In section 2, we describe the dataset used in our analysis and the predictor mode decomposition. In section 3, we introduce the statistical model for our analysis, and section 4 gives the results. We conclude with a summary and further discussion in section 5.

2. Data

We restrict our domain of interest to two tropical oceanic regions, the west Pacific (WP; 15°S–15°N, 130.25°E–180°) and the east Pacific (EP; 15°S–15°N, 180°–100.25°W), as depicted in Fig. 1. These two regions were chosen because they represent distinctly different environments where deep convection forms: the WP warm pool is a large area of warm SSTs and a moist troposphere where deep convective systems with large stratiform rain regions form year round. The EP is a more marginal environment for deep convection because of the strong equatorial cold tongue and a drier troposphere, and deep convective systems only form at the convergence of the trade winds in the ITCZ. These two regions also represent the upward and downward branches of the Pacific Walker circulation, respectively, which becomes disrupted during El Niño.

We consider two consecutive years for analysis, 2003 and 2004. Data from 2003 are used to fit the statistical models and then we independently predict rainfall for 2004 using the statistical model. There were no strong El Niño events during these two years, and we expect to observe similar weather patterns from 2003 to 2004 in each region. Future work will examine how the statistical models perform during strong ENSO events.

a. *Rainfall data*

The TRMM Precipitation Radar (PR) provides the first-ever spaceborne weather radar observations over the tropics and subtropics (from 35°S to 35°N) from late 1997 until mid-2014. We use V7 rain type information from the 2A23 product (Awaka et al. 1997; Funk et al. 2013) and V7 rain rates from the 2A25 product (Iguchi et al. 2000). Adjustments to the standard products were made according to Funk et al. (2013) to classify all shallow, isolated and shallow, nonisolated rain as convective.

The PR orbital observations were binned into 6-hourly, 0.5° grids for 2003 and 2004. The 6-hourly accumulated rainfall is expressed as a rainfall rate in units of millimeters per day. We chose this time and space resolution to ensure reasonable rain sampling from the PR's intermittent swath. When there are no PR swaths through a particular 0.5° grid during the 6-h period, that data point is assigned a missing value and is not used in the statistical model. Table 1 shows the percentage contributions of each TRMM PR rain type in 2003. Both regions have 44% stratiform rain fraction while the EP has a higher fraction of shallow rain (18% vs 11%). The relative contributions of different rain types to the overall rain amount impact the structure of heating in both regions. Another TRMM product, 3B42 (V7; Huffman et al. 2007, 2010), includes TRMM-adjusted

TABLE 1. Percent contribution of each rain type to the total rainfall as observed by the TRMM PR in the EP and WP.

Rain type	East Pacific	West Pacific
Stratiform	43.7	44.5
Deep convective	38.5	44.2
Shallow convective	17.7	11.3

merged-infrared (IR) precipitation and was also binned into 6-hourly, 0.5° grids for the two years of interest.

We also consider the climate model output from the NCAR Community Earth System Model (CESM) (Hurrell et al. 2013). We focus on uncoupled integrations using the atmospheric component, the Community Atmospheric Model, version 5 (CAM5; Neale et al. 2013), which provides 0.9° × 1.25° horizontal grid resolution with 26 levels of vertical resolution. Hourly large-scale (PRECL) and convective (PRECC) precipitation rates are aggregated into 6-hourly rain rates. Convective rain comes from the CAM5 convective parameterization while the large-scale rain is produced explicitly from grid-scale variables. While large-scale rain from GCMs may mimic some aspects of stratiform rain observed by radar (such as its geographical distribution or elevated heating), it is not formed by the same physical processes so only loose comparisons can be made between the CAM5 large-scale rain and TRMM PR stratiform rain observations.

b. Atmospheric state variables

MERRA-2 environmental variables are matched to the TRMM PR and 3B42 gridded datasets. MERRA-2 is a NASA reanalysis dataset that assimilates in situ and remotely sensed data using the GEOS-5 model from 1980 onward (Rienecker et al. 2011). This dataset provides 3-hourly data at 2/3° × 1/2° horizontal grid resolution, which was interpolated using bilinear interpolation method to match the 6-hourly, 0.5° resolution of the TRMM data. MERRA-2 data for temperature, humidity, and wind vectors at 40 different pressure levels, and latent heat flux are used in our analysis.

c. Common grid analysis

Temporal correlation is also considered in our analysis. This analysis is carried out between snapshots of the atmospheric state at a reference time; that is, we choose 0000 UTC for the reference time, and accumulated rainfall during the following 6 h (i.e., from 0000 to 0600 UTC; expressed in units of mm day^{-1}). This mimics the lagged predictive relationship assumed in GCM parameterizations of rainfall, where the current atmospheric state is used to predict rainfall through the next model time step. The implicit "time step" of 6 h in our analysis is much longer than actual model time steps, but data availability limits the length of our analysis time step. Because there is generally a very weak diurnal cycle of precipitation over the tropical oceans, we only consider analysis from 0000 to 0600 UTC. A similar analysis over land would need to incorporate the full diurnal cycle because of the stronger diurnal variability.

All the grid points in a selected region (EP or WP) and all days of the year are lumped together for the purposes of the statistical analysis. This is in keeping with our goal of mimicking the convection parameterization in a climate model, which cares only about vertical profiles at a grid point, and not about its horizontal location or the time of the year. For the smaller WP region, there are 6000 grid points on the 0.5° latitude/longitude grid, each with 365 vertical profiles at time 0000 UTC for the year. This means the 2003 TRMM training data size for fitting the statistical models is 2.19 million vertical profiles for the WP region and 3.5 million vertical profiles for the EP region. (For CAM5, the model horizontal grid scale is about 1° and the training data size is, therefore, one-fourth the observed data size.)

We assessed three predictor scenarios to predict accumulated rainfall between 0000 and 0600 UTC:

- Type A: Use the atmospheric state observed at a single grid point at 0000 UTC.
- Type B: Use the atmospheric state at a single grid point as well as at four neighboring grid points at 0000 UTC.
- Type C: Use the atmospheric state at a single grid point at 0000 UTC and at four neighboring grid points at time 1800 UTC on the previous day (i.e., 6 h earlier).

We not only use predictors at the same grid point and time of the rainfall but also consider predictor values at neighboring points of the rainfall location, at either the same time or the previous 6-h period. The rationale behind this is that rainfall may not be solely affected by the atmospheric state variables at the same location and the same time point, but may be influenced by atmospheric states nearby in space or before in time due to flow advection or wave propagation.

d. Predictor mode decomposition

The atmospheric state variables considered in this study are vertical profiles of temperature T, humidity q, and zonal u and meridional v wind, along with scalar values of latent heat flux (LH), latitude ϕ, and wind shear variables. If u_x and v_x denote zonal and meridional wind speed at x hPa, we define shear variables in the following way: low-level shear $\mathrm{LS} = \sqrt{(u_{900} - u_{700})^2 + (v_{900} - v_{700})^2}$, deep

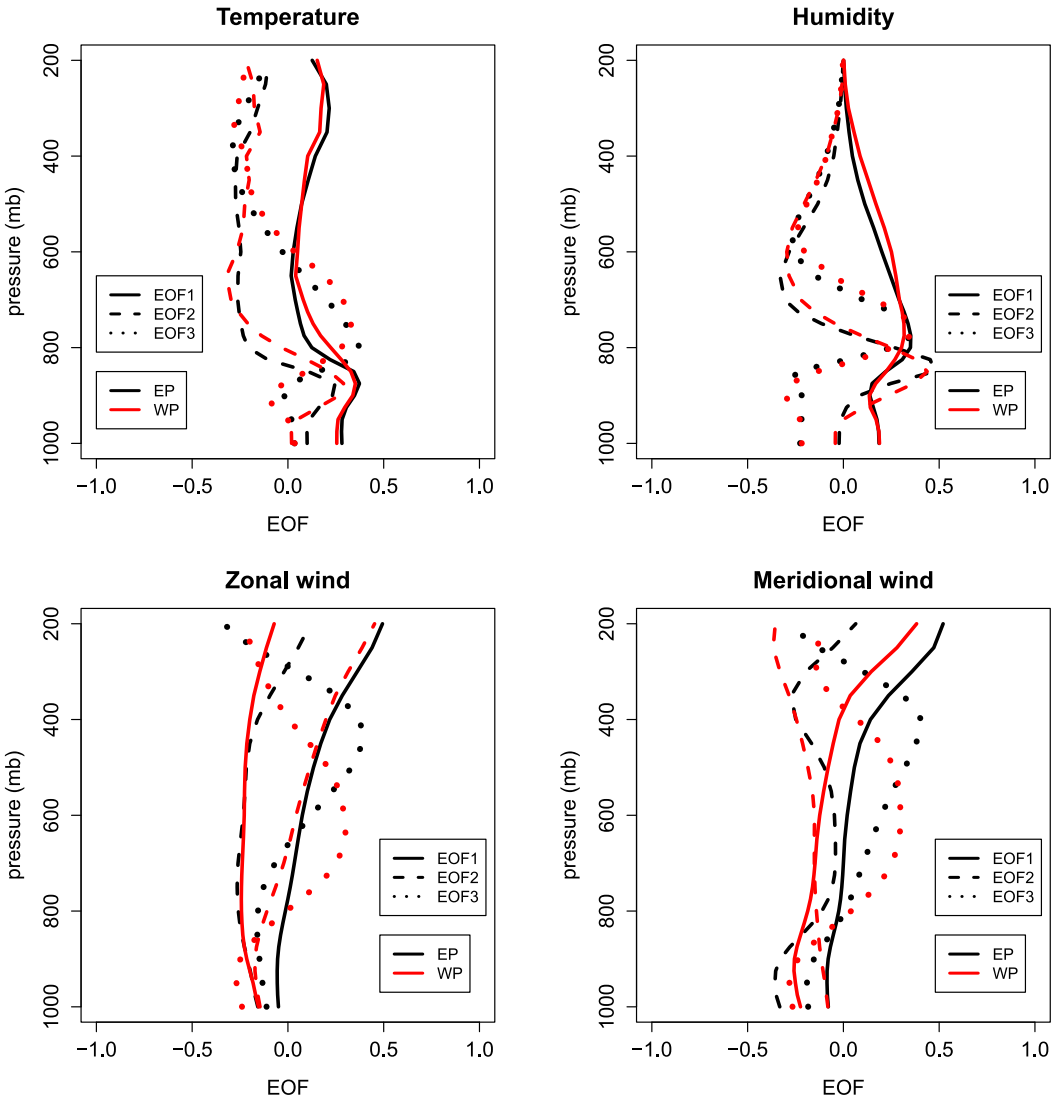

FIG. 2. First three EOFs of temperature, humidity, and zonal and meridional winds for the EP and WP from MERRA-2.

shear $\mathrm{DS} = \sqrt{(u_{900} - u_{300})^2 + (v_{900} - v_{300})^2}$, and deep direct shear $\mathrm{DDS} = (u_{300} - u_{800})$. DS and DDS were investigated to see if different aspects of deep shear might be able to help predict stratiform rain as suggested by Li and Schumacher (2011). To reduce the number of variables, we carry out EOF decomposition (Hannachi et al. 2007) in the vertical dimension for the 40 vertical levels. We lump together vertical profiles for all times and all horizontal locations in each region to compute the EOFs, and use the first three dominant modes (out of the 40 EOFs) as our predictors. As the sign of an EOF is arbitrary, we simply choose it so that the largest absolute value appears positive when displaying it.

Figure 2 shows the first three EOFs of T, q, u, and v for 2003 over the EP and WP regions from MERRA-2, and Table 2 shows the percentage of the variance explained by each EOF. Cumulatively, the first three EOFs explain 70%–90% of the variance in all cases. The EOFs of T (Fig. 2, top-left panel) show both barotropic (i.e., single sign with height) and baroclinic (i.e., changing sign with height) structure. The first EOF of T from MERRA-2, explaining about 40% of the variance, has the same sign through the depth of the troposphere with a maximum near 850 hPa. The second EOF explains about 20% of the variance and changes sign around 800 hPa, with warmer temperatures at lower levels and broad upper-level cooling in its positive phase. The third EOF explains about 10% of the variance and has a notable inversion around 850 hPa.

EOFs of q (Fig. 2, top-right panel) also show barotropic and baroclinic structures, but with a decreasing

TABLE 2. Percentage of the variance explained by the first three EOFs of T, q, u, and v over the EP and WP regions from each model (MERRA-2 and CAM5). Values in the parentheses are the cumulative percentage variability explained by each EOF.

Variable	Mode	MERRA-2 EP	MERRA-2 WP	CAM5 EP	CAM5 WP
T	First	45.0 (45.0)	41.9 (41.9)	50.2 (50.2)	59.7 (59.7)
	Second	17.3 (62.3)	21.4 (63.3)	17.0 (67.2)	16.1 (75.8)
	Third	9.4 (71.7)	11.3 (74.6)	9.6 (76.8)	6.5 (82.3)
q	First	72.0 (72.0)	69.7 (69.7)	68.3 (68.3)	74.5 (74.5)
	Second	9.0 (81.0)	9.3 (79.0)	12.3 (80.5)	9.4 (83.8)
	Third	6.4 (87.4)	6.9 (86.0)	7.7 (88.2)	7.2 (91.1)
u	First	59.9 (59.9)	45.5 (45.5)	40.8 (40.8)	45.3 (45.3)
	Second	21.4 (91.3)	38.2 (83.7)	40.4 (81.2)	33.7 (79.0)
	Third	7.5 (88.8)	6.7 (90.4)	9.0 (90.1)	9.2 (88.1)
v	First	46.5 (46.5)	39.2 (39.2)	37.0 (37.0)	38.2 (38.2)
	Second	18.3 (64.8)	27.8 (67.0)	30.5 (67.5)	29.8 (68.0)
	Third	12.9 (77.7)	12.0 (79.0)	13.4 (80.8)	13.4 (81.4)

signal at upper levels as would be expected because of the low amounts of moisture in the upper troposphere. The first EOF is particularly dominant, explaining about 70% of the variance, and indicates a very moist troposphere in its positive phase. This eigenmode bears some similarity to the vertical structure of the slowest decaying eigenmode in the linear response model of Kuang (2010), although the corresponding temperature eigenmodes are not similar. The second and the third EOFs each explain less than 10% of the variance and show peaks in humidity at low to midlevels and dry upper levels in the positive phase.

While the EOFs of T and q are remarkably similar between the EP and WP, there is greater variability difference between regions in the EOFs of u and v (Fig. 2, bottom panels). In addition, the EOF profiles of wind are dominated by upper-level variations, as would be expected for wind profiles because of the stronger winds in the upper troposphere, and exhibit greater differences between the EP and WP.

The CAM5 EOFs of T (Fig. 3 and Table 2) are qualitatively similar to the MERRA-2 EOFs, with a more "smeared out" vertical structure. The exception is the third CAM EOF of T in the EP, which has more structure in the midtroposphere compared to the third MERRA-2 EOF. The CAM5 EOFs of q show considerable similarity to all three MERRA-2 EOFs. The first two CAM5 EOFs of u and v are similar to the corresponding MERRA-2 EOFs, but the third EOF shows some differences. Thus, using EOFs to simplify the atmospheric variables for our rain predictions appears warranted across platforms and MERRA-2 and CAM5 have similar variability in vertical profiles of T, q, and wind.

3. Statistical methods

Our statistical model consists of two parts: logistic regression and gamma regression. For each rain type, logistic regression determines whether it rains or not and the gamma regression determines how intense the rain will be. Different models are needed because atmospheric conditions conducive to rain initiation may be different from the atmospheric conditions that help produce large rain amounts.

a. Logistic regression

The first part determines the probability of rain through a logistic regression, which is a common example of a GLM that is used when the response variable is binary: 0 for no rain and 1 for rain. A *logit* transformation,

$$f(x) = \log\left(\frac{x}{1-x}\right), 0 \le x \le 1,$$

of the probability of rain is expressed as a linear combination of relevant predictors (in our case, temperature, humidity, and others, as described in section 2b). That is, if $p(\mathbf{s})$ denotes the probability of rain at a grid point $\mathbf{s}$, we write

$$\log\left[\frac{p(\mathbf{s})}{1-p(\mathbf{s})}\right] = \beta_0 + \beta_1 X_1(\mathbf{s}) + \cdots + \beta_p X_p(\mathbf{s}),$$

where $X_i(\mathbf{s})$ denotes predictors at the grid point $\mathbf{s}$.

We fit the logistic regression model for each rain type separately using data from 2003. When we predict if there is rain or no rain at a given grid point for 2004, we specify a cutoff probability p_c that is allowed to be different for each rain type. That is, if $\hat{p}(\mathbf{s})$ is the predicted probability of rain for 2004 at a grid point $\mathbf{s}$, we predict rain only over those locations where $\hat{p}(\mathbf{s}) \ge p_c$.

b. Gamma regression

The second part of the model predicts the amount of rain under the GLM framework with a gamma distribution

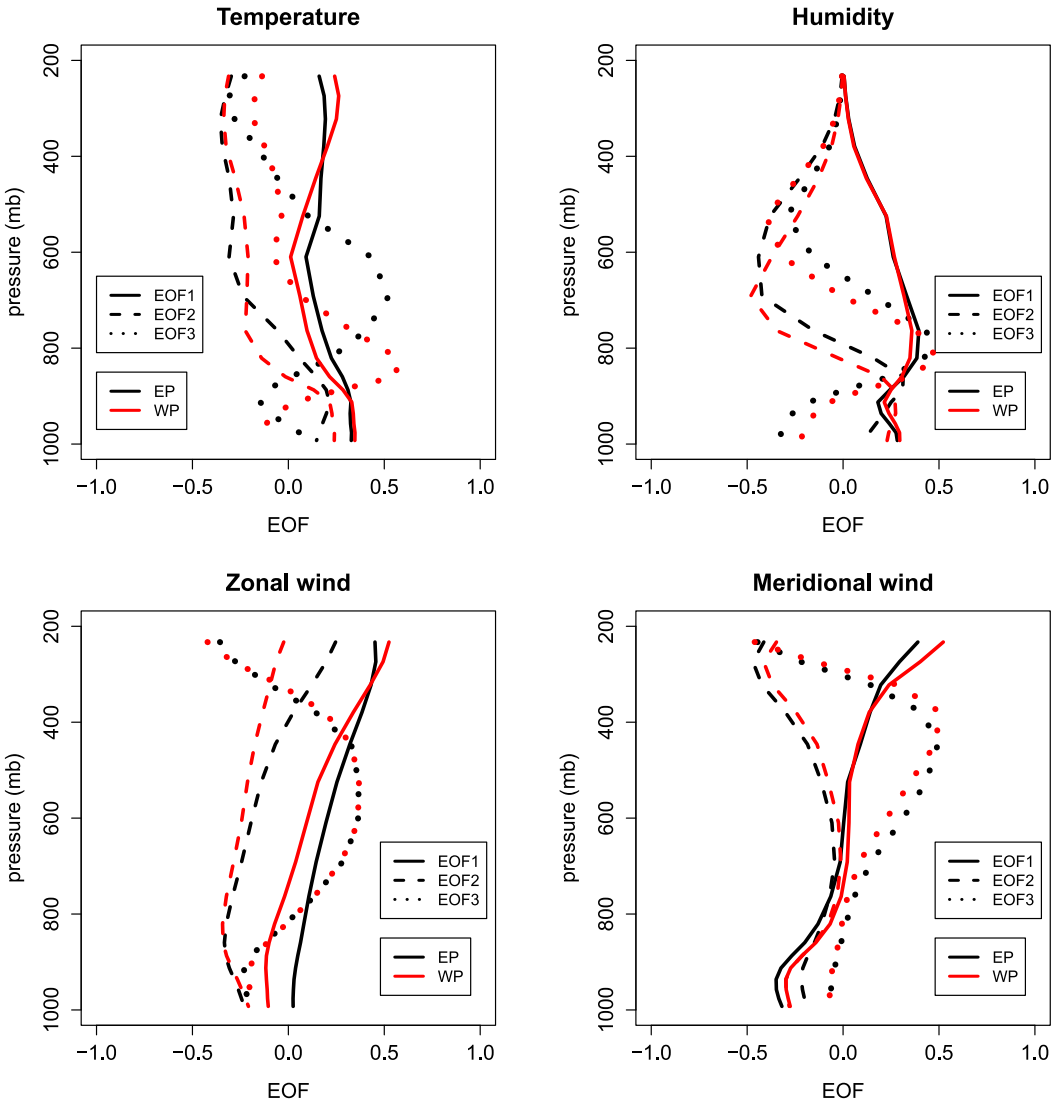

FIG. 3. As in Fig. 2, but for CAM5 data.

and a log link function. The gamma distribution only takes a positive value, and has positive skewness, which is useful for describing precipitation amounts (Wilks 1995; Husak et al. 2007). Suppose $Y(\mathbf{s})$ is a rainfall amount over a grid point $\mathbf{s}$ for a certain rain type with an expected value $\mathbb{E}[Y(\mathbf{s})] = \mu_{\mathbf{s}}$. Then we assume that Y follows a gamma distribution $G(\alpha_{\mathbf{s}}, \beta_{\mathbf{s}})$, with a probability density function, $f(x) = \{1/[\Gamma(\alpha_{\mathbf{s}})\beta_{\mathbf{s}}^{\alpha_{\mathbf{s}}}]\}x^{\alpha_{\mathbf{s}}-1}e^{-x/\beta_{\mathbf{s}}}$, $0 < x < \infty$. The mean $\mu_{\mathbf{s}}$ is given by $\alpha_{\mathbf{s}}\beta_{\mathbf{s}}$ and the variance, $\mathrm{Var}[Y(\mathbf{s})]$, given by $\alpha_{\mathbf{s}}\beta_{\mathbf{s}}^2$. We use subscripts for the two parameters α and β to denote that these parameter values depend on $\mathbf{s}$. Then,

$$\log\mu_{\mathbf{s}} = \log(\alpha_{\mathbf{s}}\beta_{\mathbf{s}}) = \eta_0 + \eta_1 X_1(\mathbf{s}) + \cdots + \eta_p X_p(\mathbf{s}).$$

Note that η_i terms are common across all the grids and thus do not depend on $\mathbf{s}$.

c. *Relative importance of predictors*

One of our main interests is to quantify the importance of each atmospheric state variable in predicting rainfall. Convective parameterizations typically use physical arguments to motivate the choice of predictors, which are often derived from vertical profiles of temperature and humidity. In our analysis, we make no a priori assumptions about the predictors, allowing the data to dictate what is important. In the statistics literature, Pratt (1987) and Thomas et al. (2008) proposed methods for quantifying the relative importance of predictors in multivariate linear models or logistic regression model settings. For each of the statistical models in this study, we employ similar ideas to quantify the relative importance of predictors; we report the scaled

TABLE 3. Observed and predicted proportion of gridpoint rainfall occurrences based on logistic regression for the TRMM PR and 3B42 data over the east and west Pacific. Numbers in parentheses are cutoff probabilities p_c used for each region and each rain type.

		East Pacific		West Pacific	
Rain type	Prediction	obs = 0	obs = 1	obs = 0	obs = 1
Stratiform	pred = 0	0.811	0.065	0.625	0.121
(0.3, 0.4)	pred = 1	0.062	0.062	0.125	0.129
Deep convective	pred = 0	0.821	0.051	0.582	0.078
(0.2, 0.25)	pred = 1	0.083	0.045	0.226	0.115
Shallow convective	pred = 0	0.660	0.085	0.444	0.119
(0.3, 0.35)	pred = 1	0.136	0.120	0.243	0.194
3B42	pred = 0	0.771	0.072	0.688	0.123
(0.3, 0.4)	pred = 1	0.078	0.079	0.097	0.092

t statistic (or Wald statistic, which is a squared *t* statistic). The *t* statistic for a particular variable is associated with the change in the residual sum of squares after removing this variable. The statistical programming language R (http://www.r-project.org) is used to carry out the analysis.

4. Results

We first discuss results for the two statistical models (logistic and gamma regressions) applied to TRMM PR and 3B42 data. We then apply the same statistical analysis to CAM5 output. Last, we discuss the cases of false negatives in the prediction. For all cases, the statistical models are fitted separately for each rain type and the two domains, EP and WP, using data from 2003. We then make independent predictions for 2004 using the fitted statistical models.

a. TRMM observations

Table 3 shows prediction results using the logistic regression (i.e., rain vs no rain). Each number represents the proportion of the number of pixels that fall into each category. The cutoff probability, described in section 3, is chosen to yield the best prediction result for each rain type separately. Overall, the best prediction results were obtained when type B predictors were used (i.e., a grid point and its four nearest neighbor points at time *t*), although for some cases types B and C were comparable. This suggests that the spatial dependence is an important constraint on the model. Note that the types B and C require the same number of predictors.

The statistical model performs best in the EP, where it is able to predict the no-rain cases of stratiform and deep convective rainfall from the TRMM PR well, being correct over 90% of the time (i.e., obs = 0, pred = 0 vs obs = 0, pred = 1 values in Table 3). However, it is only able to predict the occurrence of stratiform and deep convective rain correctly less than 50% of the time (i.e., obs = 1, pred = 1 vs obs = 1, pred = 0 values in Table 3). In the WP, the predictive value for stratiform and deep convective rain is lower than the EP for no-rain cases (78% on average) but slightly higher for rain cases (about 56%). The difference in predictability between the two regions is consistent with the fact that the WP environmental conditions are more conducive to deep convective cloud systems and that it is easier to predict rain versus no rain cases for stratiform and deep convective rain compared to the EP, which has a more marginal environment for deep convective clouds and less rain occurrence overall. Shallow convective rainfall is more ubiquitous than stratiform and deep convective rain and occurs more often than the other two rain types. As such, it generally has better predictability for rain occurrence compared to the deep rain types (about 60%) but lower predictability for the no-rain situations (about 75%).

Table 3 also shows that the prediction of rain occurrence from 3B42 (rain types are not separated in this product) is similar to the PR stratiform and deep convective rain predictions in the EP. The 3B42 no-rain prediction is better than the PR in the WP, but rain cases are not predicted as well and may be due to an underestimate of rain occurrence overall by 3B42 compared to the PR (i.e., 21.5% vs 25%).

When a grid point has rain, rainfall amounts are fitted to a gamma regression model. Figures 4 and 5 show comparisons between the observed and predicted 2004 rain amount for each PR rain type and the total 3B42 rainfall in the EP and WP, respectively. For all rain types, predicted rain maps match the observed rain amounts fairly well. The predicted rain field is somewhat smoothed but the ITCZ and cold tongue are well delineated in the EP (Fig. 4) and rain covers the entire WP (Fig. 5). It is worth noting that although the statistical model is able to predict the averaged properties of rainfall, it performs poorly in predicting rainfall on a day-to-day basis. This means that the statistical model

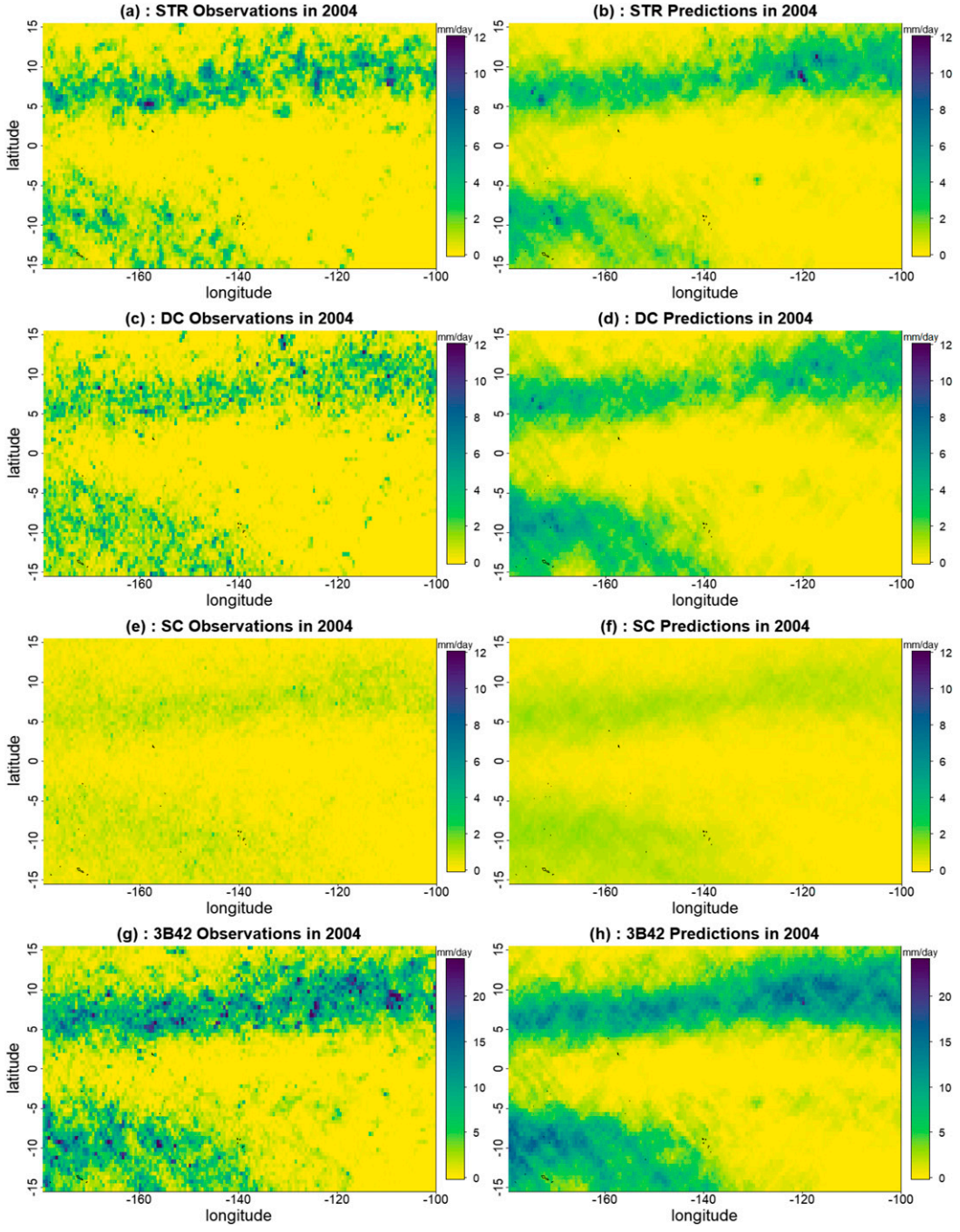

FIG. 4. (left) TRMM observed and (right) predicted rainfall (mm day^{-1}) in the EP: (a),(b) stratiform, (c),(d) deep convective, (e),(f) shallow convective, and (g),(h) 3B42 total rain.

may be more suitable for use in a climate model than in a weather forecast model.

Figure 6 shows the observed and predicted rain amount distributions for both TRMM PR and 3B42 datasets over the EP and WP in the original scale and base-10 log scale. The observed distributions (solid lines) maximize at 2 mm day^{-1} or less in both regions, while the predicted distributions (dashed lines) have maximum rain rates shifted right of the observed peaks. The log-scaled plots (Fig. 6, right panels) suggest that the observed distributions have longer tails than the predicted distributions in both regions. The largest shift

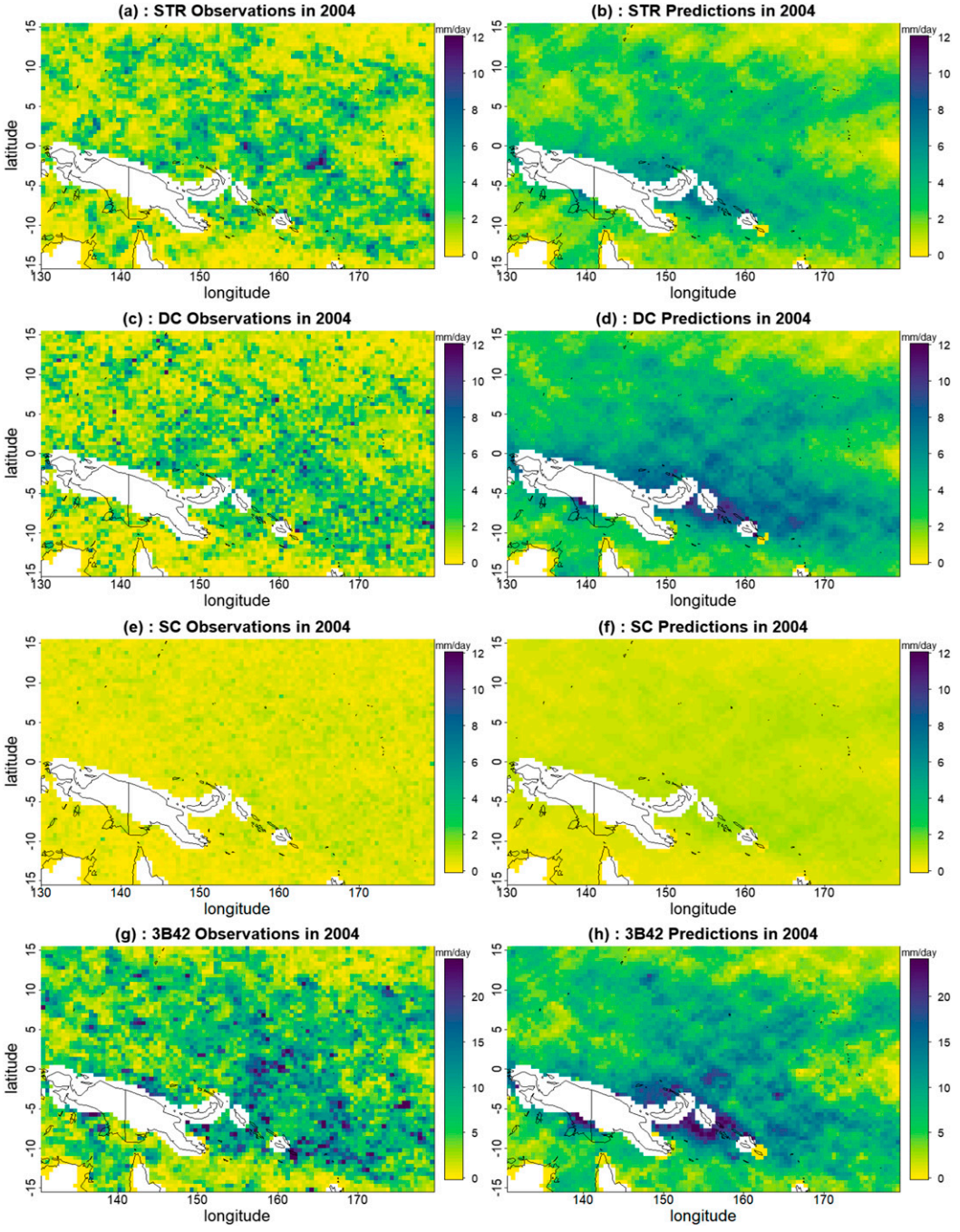

FIG. 5. As in Fig. 4, but over the WP.

occurs in the 3B42 prediction (red dashed line), where the peak density occurs beyond 15 mm day^{-1}. The smallest shift occurs for the shallow convective rain type (green dashed line); however, the predicted peak is much larger and sharper than the observed shallow convective rain distribution. The predicted stratiform (blue dashed line) and deep convective (orange dashed line) distributions indicate a moderate shift to higher rain rates compared to observations, but the deep convective rain rates shift farther right and maintain a shape that is closer to observations. The large density of weak stratiform rain rates is spread out at higher rates in the prediction. It is common to fit high-frequency rainfall amounts using a gamma distribution due to their skewed

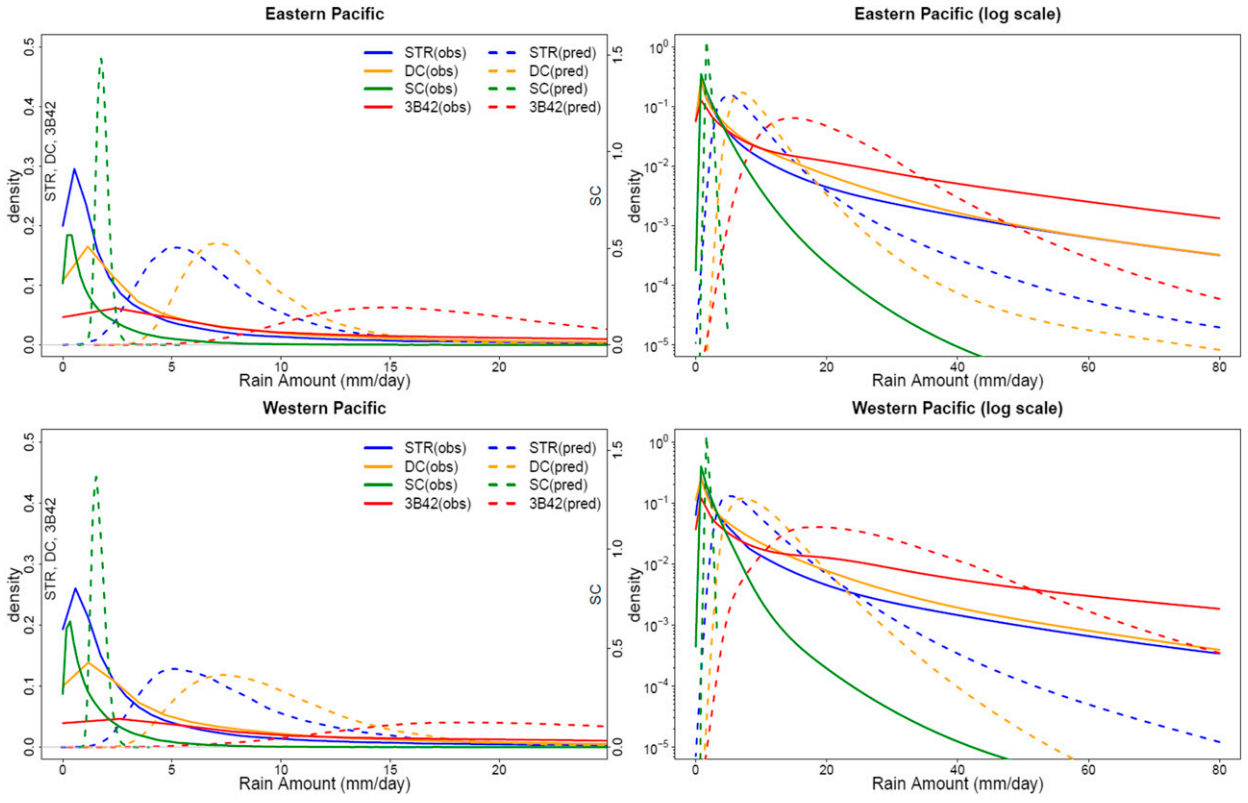

FIG. 6. Observed (black) and predicted (gray) TRMM PR and 3B42 rain rate distributions over the (top) EP and (bottom) WP in the (left) original scale and (right) base-10 log scale.

distribution (Katz 1999; Husak et al. 2007). However, as shown in Fig. 6, there is a significant discrepancy between distributions of the observed and predicted rain amounts at high intensities [also discussed in Karl et al. (1995)].

We now examine the relative contributions of different atmospheric state predictors to this skill (Fig. 7). For stratiform and deep convective rain, the dominant contributor (i.e., 25%–35%) to the predictions of both rainfall occurrence and rainfall amount is the first EOF of q, which is characterized by an overall moister atmosphere in its positive phase (Fig. 2). However, the first EOF of q only explains 10%–15% of the shallow convective rain amount in the gamma regression while the second EOF of q describes 15%–20%. This difference is because the second EOF of humidity exhibits midtropospheric drying in its positive phase, and midlevel dryness tends to be associated with a higher likelihood of congestus clouds (e.g., Jensen and Del Genio 2006). The second EOF of q also describes relatively more of the shallow convective rain occurrence compared to the other rain types. The third EOF of q, representing dry low levels in its positive phase, contributes 5%–15% to the prediction of the rain occurrence and rain amount for all rain types.

The second EOF of T tends to be the second most important predictor, especially for rain occurrence, and is characterized by a baroclinic structure changing sign around 850 hPa indicating a sharp lapse rate above the height of the trade wind inversion. However, quite a bit of spread of relative importance exists for this EOF between regions and rain types. For example, this predictor is not as important in the WP compared to the EP and while it does well at explaining EP stratiform rain occurrence (20%), it has little to do with predicting EP stratiform rainfall amount (although it remains important in helping describe more than 20% of the EP deep convective rain occurrence and amount). Another notable difference for this temperature EOF is that it contributes very little to the 3B42 prediction of rain occurrence compared to the PR. The third EOF of T, which has a strong inversion around 800 hPa in its positive phase, contributes to about 10% to rain occurrence predictions in both regions, but 10%–15% to rain predictions in the WP and only 5% in the EP.

Interestingly, the first EOF of T, characterized by a barotropic vertical structure, appears to be least important among the thermodynamic (T, q) predictors. Among the dynamic (wind related) predictors, low-level

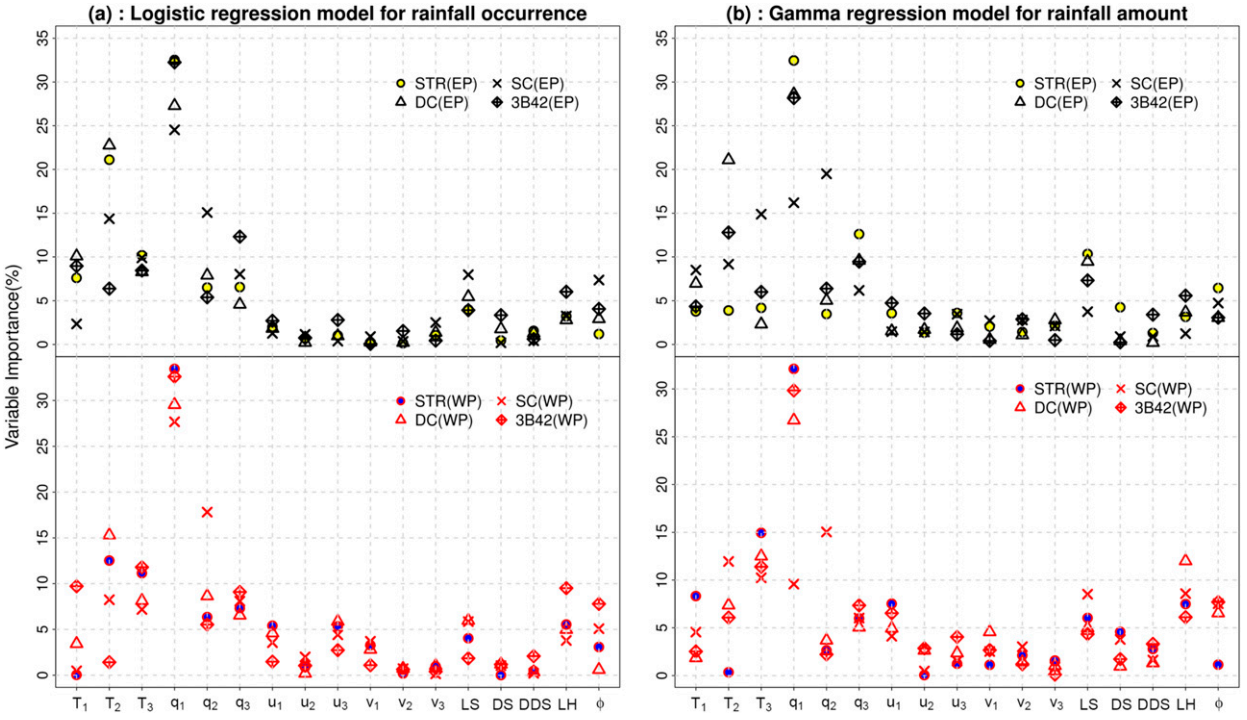

FIG. 7. Relative importance of predictors for (a) logistic and (b) gamma regression using TRMM PR and 3B42 data. The sum of every variable for each rain type is set to 100.

shear makes the biggest contribution (about 5%–10%), which is comparable to some thermodynamic predictors. The surface latent heat flux also makes a reasonable contribution, but this may not be independent of the contribution from other thermodynamic predictors.

b. *CAM5 simulations*

We also fit the two statistical models to CAM5 simulations. Note that we use this fit to predict the CAM5 simulated rainfall, not the TRMM observed rainfall. In essence, we are treating the simulated rainfall and atmospheric state as synthetic observations that we try to predict using a statistical model. As noted previously, the first two CAM5 EOFs (Fig. 3) are qualitatively similar to the MERRA-2 EOFs (Fig. 2). This indicates that CAM5 is able to simulate the vertical structure of the predictors fairly well. However, the temporal characteristics of rainfall in the CAM5 data turn out to be quite different. The prediction results using logistic regression for CAM5 data are shown in Table 4. When compared to the TRMM results (Table 3), the most striking difference is the ratio of no-rain to rain cases (i.e., CAM5 simulates rain over 80% of the time, whereas TRMM observes rain less than 25% of the time over the time and space scales being considered). These numbers are consistent with previous studies that found that climate models rain far more frequently than in observations (e.g., Dai 2006; Stephens et al. 2010).

Our statistical model is able to predict the occurrence of CAM5 simulated rain correctly about 80% of the time in the EP region, but is only able to predict the no-rain cases correctly about 40% of the time of nonraining events (obs = 0). This is the opposite of the predictions using TRMM data where there was more skill in predicting the no-rain cases. The fact that the logistic model applied to CAM5 has lower prediction skill is surprising because TRMM observational data has measurement errors not present in CAM5 simulations. The poorer fits for CAM5 data suggest that the relationship between the atmospheric state and rainfall in CAM5 may be more nonlinear, or harder to capture with a GLM model, as compared to observations.

TABLE 4. As in Table 3, but for CAM5 data.

		East Pacific		West Pacific	
Rain type	Prediction	obs = 0	obs = 1	obs = 0	obs = 1
Large-scale (0.7, 0.7)	pred = 0	0.079	0.165	0.054	0.146
	pred = 1	0.110	0.645	0.135	0.665
Convective (0.7, 0.75)	pred = 0	0.129	0.181	0.070	0.183
	pred = 1	0.083	0.607	0.103	0.644

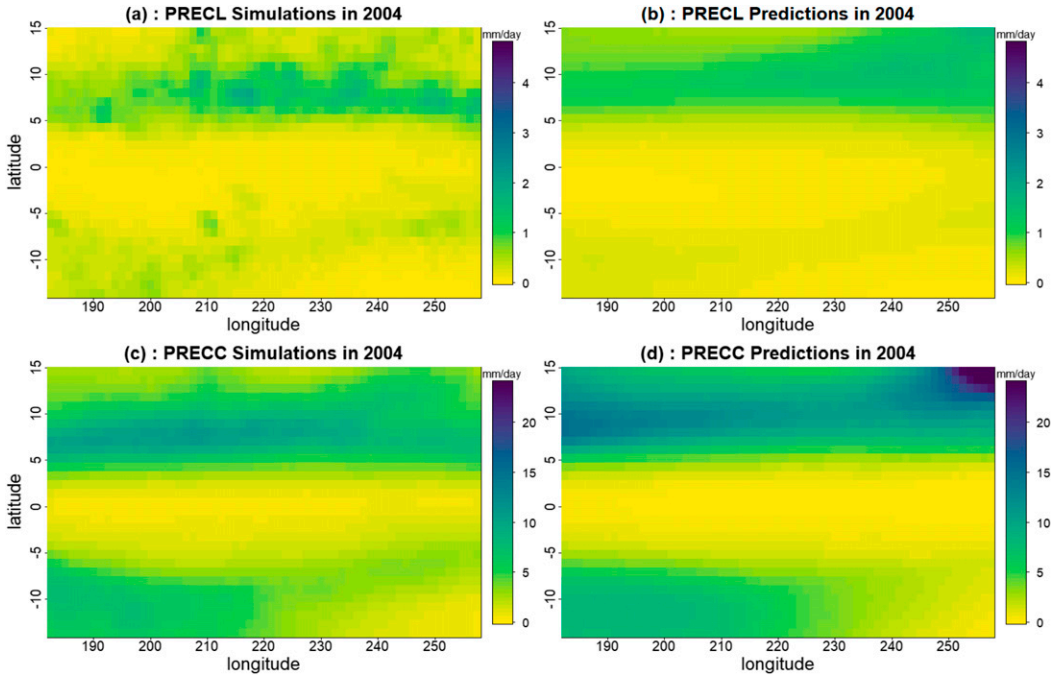

FIG. 8. As in Fig. 4, but for CAM5 data over the EP, showing simulated CAM5 rainfall for (a) large-scale and (c) convective rain and predicted CAM5 rainfall for (b) large-scale and (d) convective rain.

Figures 8 and 9 show comparisons between simulated and predicted rain amounts from CAM5 using the gamma regression. The simulated CAM5 rain maps (Figs. 8 and 9, left panels) differ from the observed TRMM maps (Figs. 4 and 5, left panels) in that the EP rain feature south of the equator is more zonal in the model (an indicator of the double-ITCZ problem endemic to climate models) and the equatorial cold tongue that intrudes into the model WP. This comparison illustrates an important point—the CAM5 simulation of rain exhibits large biases as compared to observation. Indeed, TRMM-trained statistical model predictions of the rainfall (Figs. 4 and 5, right panels) appear more realistic than CAM5 simulations (Figs. 8 and 9, left panels). One caveat regarding this comparison is that we are comparing a single year of TRMM-based prediction to a single year of CAM5 simulation. However, gross features such as the unrealistic double ITCZ pervasive in many climate model simulations are not present the TRMM-based predictions. Thus, the TRMM-based predictions, although imperfect, could potentially perform better than convection parameterization in current climate models.

The statistical model fitted to CAM5 data is able to reproduce the spatial patterns of simulated CAM5 rainfall, but with less skill than in the case of the model fitted to TRMM data. The predictions are better over the EP region, although there are some artifacts in the northeast corner. In the WP regions, the predicted spatial patterns are much more diffuse and less skillful, especially for the convective rainfall.

The different performance of the gamma regression in the EP and WP is further highlighted by the rain rate distribution plots in Fig. 10. The CAM5 simulated and predicted rain distributions generally match well in the EP, although the predicted rain distributions shift slightly right to higher values. In the WP, the predicted rain shifts are more dramatic, especially for PRECC (red dashed line in Fig. 10), which has a large peak near 7 mm day^{-1} and a sharp dropoff at very high rain rates. This peak at 7 mm day^{-1} and sharp dropoff is also seen in the TRMM PR deep convective rain rate prediction in Fig. 6, so it likely results because the assumed gamma model is not appropriate for deep convective rain.

Next, we compare the relative contribution of different atmospheric state predictors to the prediction skill for the simulated rain in CAM5 (Fig. 11). As in the case of TRMM data (Fig. 7), the biggest contribution to prediction skill comes from the first EOF of q, although it is less dominant. The baroclinic second EOF of T, again, appears to be the second most important predictor. Interestingly, the barotropic first EOF of T appears to make a bigger contribution in the CAM5 simulations. This suggests that the convective parameterization in CAM5 may be more sensitive to barotropic vertical temperature structure. Also, the contributions

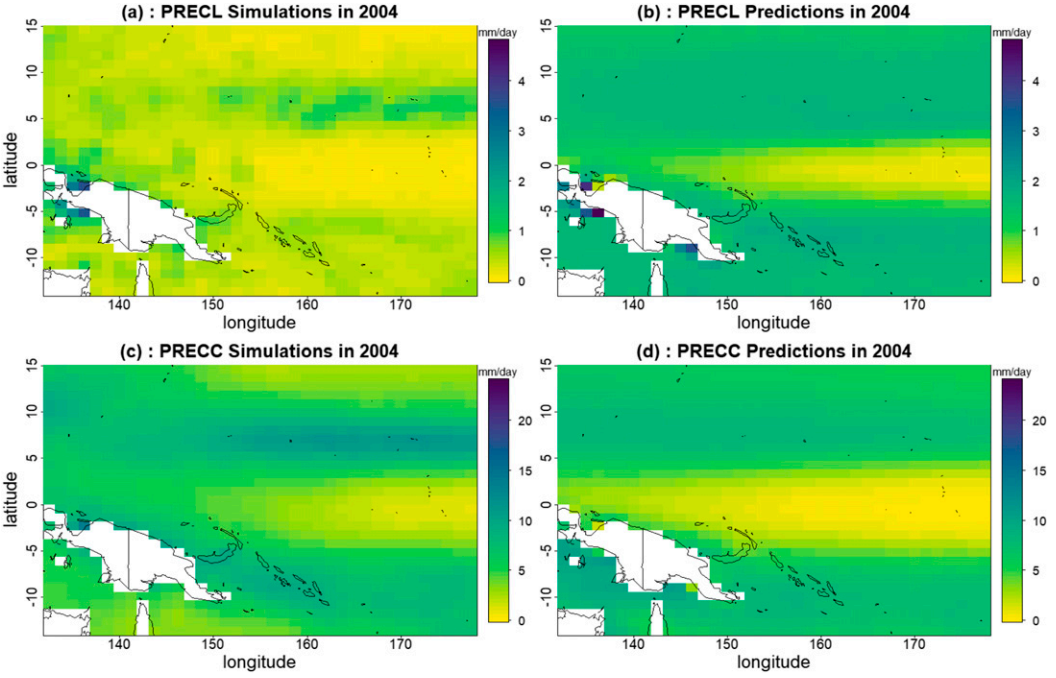

FIG. 9. As in Fig. 8, but over the WP.

from the wind-related predictors do not stand out from the noise level. This is perhaps not surprising as convective parameterizations do not explicitly use vertical shear as predictors. Another noteworthy feature is that latitude as a predictor makes a much stronger contribution in CAM5 simulations compared to TRMM data. This suggests that the CAM5 convective parameterization may be more sensitive to latitude-dependent environmental properties. When latitude is removed as a predictor, the relative importance of the first EOF of q increases most (not shown).

c. *False negative diagnosis*

To further investigate the performance of the logistic regression model, we determined the main atmospheric conditions that resulted in false negatives (i.e., rain is observed but the model failed to predict rain). We compare the *kernel density estimators*, a statistical method to smooth the density function nonparametrically (Silverman 1986), for true positive cases (i.e., rain is observed and the model correctly predicts rain) and false negative cases. That is, Fig. 12 shows comparisons of density curves for the first EOF of q (denoted q_1) and the third EOF of T (denoted T_3) as these were the EOFs that exhibited the largest difference between the true positive and false negative cases.

The top set of panels in Fig. 12 show that the false negative cases (dashed lines) occur when conditions are more humid than in the true positive cases (solid lines) for all rain types. This is especially true for the shallow convective rain occurrence, which has a large negative skew toward negative q_1 values for true positive cases. The logistic regression assumes a linear relationship between humidity and the logit transform of the probability of rain occurrence, but this relationship is likely nonlinear as suggested by the exponential pickup of rain amounts with column integrated humidity observed across the tropics (Bretherton et al. 2004; Ahmed and Schumacher 2015, 2017). A next step would be to use a different statistical model or combination of models that better represent this nonlinearity and non-Gaussianity.

The bottom set of panels in Fig. 12 show that the false negative cases tend to occur when conditions are less stable than in the true positive cases, again suggesting a nonlinear relationship between stability and the logit transform of the probability of rain occurrence. The offset toward more negative T_3 for false negative cases is especially pronounced over the WP. Thus, there are geographical variations in the nonlinear relationship between the environment and rain occurrence that should be considered in future work.

5. Conclusions

The goal of this study was to find empirical relationships between the environment and rain production in

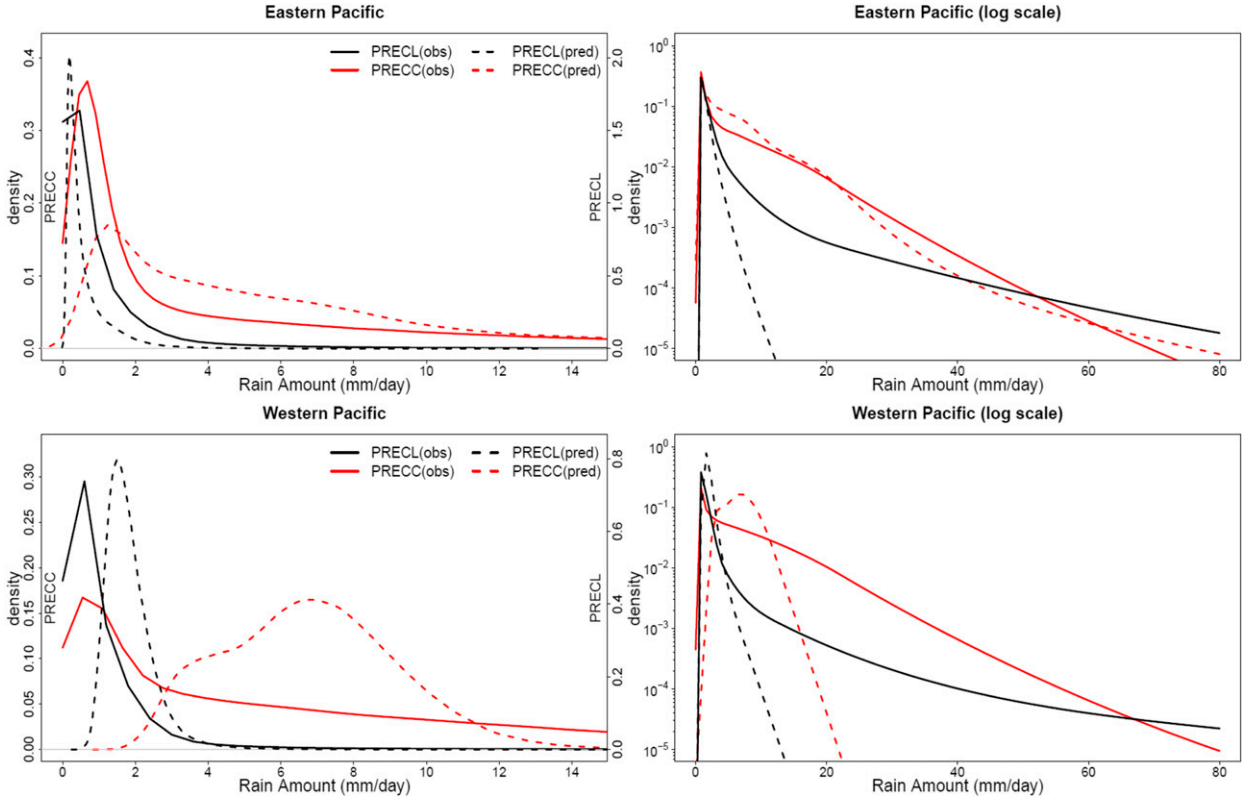

FIG. 10. As in Fig. 6, but for CAM5 data.

the tropical east and west Pacific at subdaily time scales using a set of generalized linear models. The motivation was to see if statistically derived relationships from observations may be used to improve the parameterization of rainfall in climate models. We chose to separate the prediction into rain occurrence and rain amount since these are distinct processes in the atmosphere and are treated differently in the trigger and closure assumptions in convective parameterizations. The logistic regression was used for rain occurrence because this is an on–off decision: either it rains or it does not. The gamma regression was used for rain amount because it can be

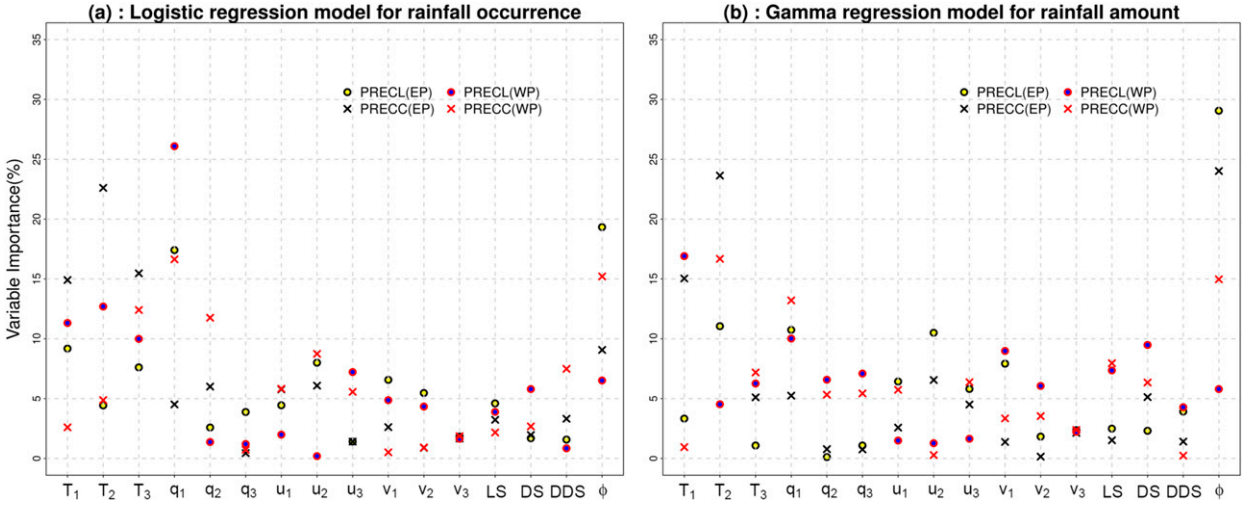

FIG. 11. As in Fig. 7, but for CAM5 data.

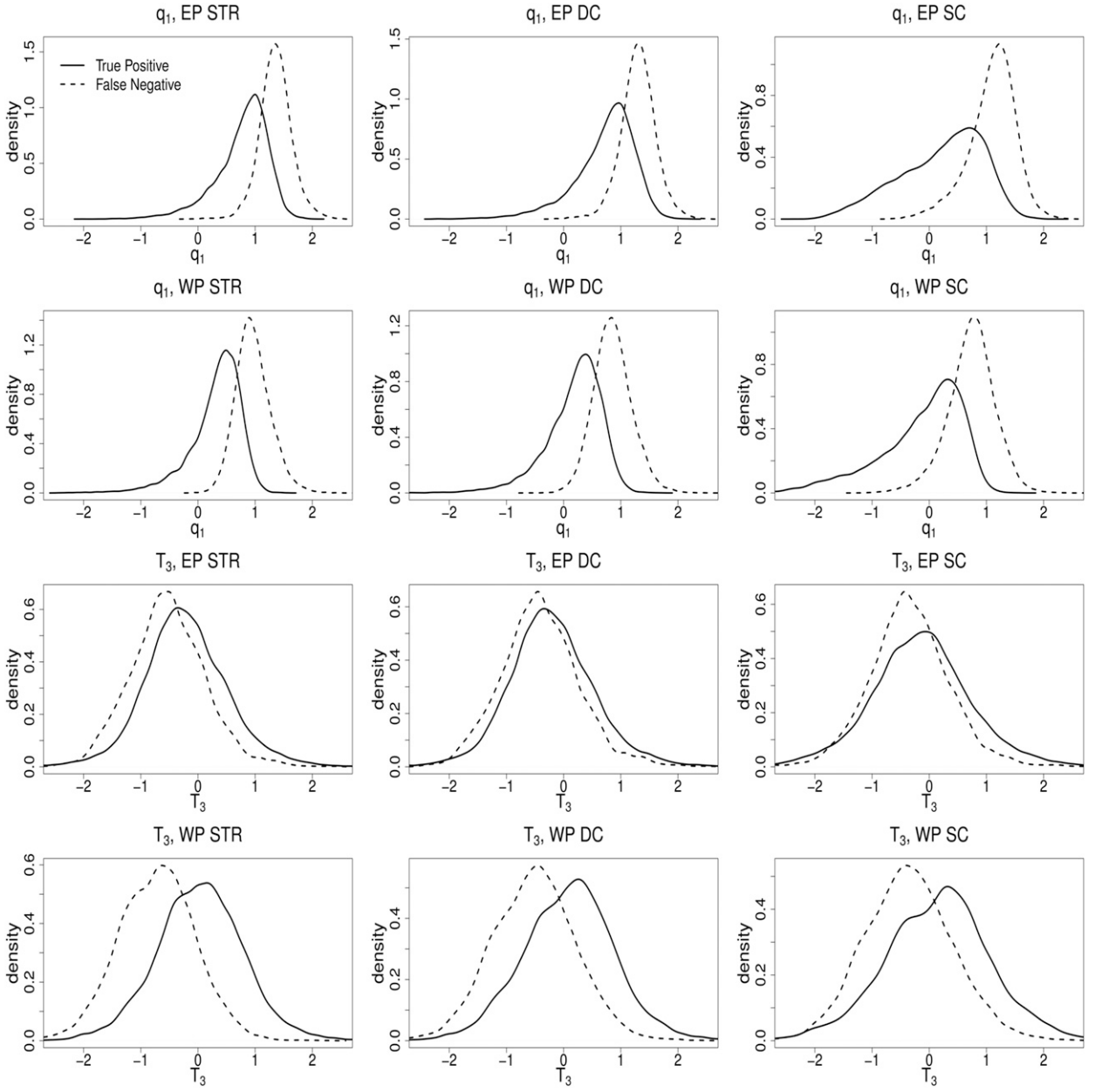

FIG. 12. Density plots of q_1 (first EOF of humidity) and T_3 (third EOF of temperature) for true positive (solid line) and false negative (dashed line) cases from the logistic regression results using TRMM PR data.

applied to positive valued skewed distributions which rain rates tend to have (i.e., lower rates are more common with a long tail for higher rain rates).

We also chose to separate the prediction by rain type as not all tropical rain is created equal. Shallow convection is ubiquitous across the tropical oceans, while deep convective and stratiform rain is confined to regions with high sea surface temperatures and high rain accumulations. In addition, while deep convection can occur in isolation, stratiform rain either forms from aged convection or is fed by active convective towers so cannot form in isolation. Our observations of rain came from the TRMM PR because it can differentiate between these rain types. We also included analysis with the total rain field from TRMM 3B42 since most tropical precipitation datasets are derived from infrared satellite observations that cannot easily be separated into rain type.

GCM model precipitation was separated into two types, large-scale rain (i.e., the rain that is determined

explicitly from grid-scale variables) and convective rain (i.e., the rain that is determined from the convective parameterization). Large-scale and convective rain from a model are not the same as stratiform and convective rain observed by radar, but GCMs have widely varying rain contributions from each rain type so we deemed it important to examine the environmental–rain relationships for separate rain types in GCMs as well.

The main predictors were profiles of temperature, humidity, and zonal and meridional wind from either MERRA-2 or CAM5. The profiles were determined by applying EOF analysis to each environmental field over the EP and WP and the first three EOFs for each variable were included in the analysis. The CAM5 profiles were remarkably similar to the MERRA-2 profiles, providing confidence in the robustness of the EOF technique for data compression. Some surface variables were used as predictors as well, but they generally were not important to the prediction, stressing the fact that the vertical structure of the atmosphere plays a highly relevant role in rain production. The statistical models were trained on 2003 data and used to predict 6-hourly rain occurrence and rate on 0.5° grids in 2004. These years were chosen because of the absence of a strong ENSO event, but we plan to test the performance of the statistical models during a strong El Niño in future work.

In the observational analysis with TRMM and MERRA-2 data, the first EOF profile of humidity—representing overall humid conditions throughout the troposphere—contributed most to the prediction for both statistical models. The second EOF of temperature—representing greater atmospheric instability—was the second most important predictor for both statistical models although more variability existed in the performance of this parameter between regions and rain types. Low-level shear was the most important wind predictor. These results appear to be consistent with the dominant statistical predictors for large precipitating systems identified by Chen et al. (2017), which include total precipitable water vapor, low and midlevel humidity, and low-level wind shear. The importance of q_1 and T_2 to the logistic regression based on CAM5 data was similar to the TRMM and MERRA-2 analysis; however, q_1 was only of moderate importance in the CAM5 gamma regression analysis, while T_2 was more dominant. CAM5 results also showed a large contribution from latitude not seen in the observational analysis, and as expected from the convective parameterization, wind predictors were unimportant.

The logistic regression on the observational data generally performed well at predicting whether it will rain or not, but did better in the EP compared to the WP in minimizing false negatives and false positives. Analysis of conditions when false negatives occur showed that environment–rain relationships can sometimes be nonlinear and that this nonlinearity is geographically dependent, which needs to be taken into account when applying statistical models to the prediction of rain. Future analysis will also include applying this technique to tropical land and midlatitude land and ocean regions to create more global empirical relationships. The logistic regression also performed better for stratiform and deep convective rain than for shallow convective rain. This is likely due to the fact that shallow convective rain is more ubiquitous across the tropical ocean and can occur in a wider range of environmental conditions.

The CAM5 logistic regression had higher false negative percentages in both regions compared to the TRMM and MERRA-2 results, but lower false positive percentages in the WP. There was also no strong difference in performance between large-scale and convective rain occurrence predictions for CAM5. It is worth noting that it rains much more often in CAM5 than TRMM (i.e., 80% vs 10%–30% occurrence rates per rain type, respectively). This is a well-known issue in GCMs and may impact the performance of a chosen statistical model, although the logistic regression as formulated in this study still performs well for CAM5.

The gamma regression applied to the observational data produced reasonable geographical rain amount distributions for each rain type in each region, but rain rate probability distributions were not predicted as well. In particular, predicted rain rates shifted to higher median values compared to observed values suggesting the need for a different, higher order model or combination of models (such as gamma plus extreme value) to predict rain rate distributions more accurately. The gamma regression did not perform as well on CAM5 data, especially in the WP, with predicted geographical rain patterns much less nuanced and larger increases in the prediction of high rain rates compared to the original data.

Overall, the statistical models fitted to TRMM observations were able to predict the main features of the spatial patterns and amplitudes of tropical rainfall, albeit in the time-averaged sense. Indeed, the TRMM-fitted statistical model prediction of the spatial structure of rainfall appears more realistic than the CAM5 simulations using a convection parameterization in that it avoids simulating spurious features like the double ITCZ (Li and Xie 2014; Oueslati and Bellon 2015). This gives us hope that with further extensions, the statistical model may be used as a substitute for convection parameterizations. In particular, the statistical model needs to be extended to predict not just the rainfall

amounts, but also the vertical profile of moistening/drying as well as diabatic heating tendencies. With this extension, our observationally based approach can yield an empirical parameterization for convection that complements recent studies that use machine learning approaches to develop convection parameterizations trained on high-resolution model simulations (e.g., O'Gorman and Dwyer 2018; Rasp et al. 2018). Another extension is to apply this methodology outside the tropics, using the new GPM measurements that extend into the midlatitudes. Ideally, though, the statistical model fit should not be based on the geographical location but on rain types, mimicking the structure of a physically based rainfall parameterization.

One caveat on the use of empirical parameterizations is that they implicitly assume stationarity of climate. This means that an empirical may be suitable for simulating interannual climate variability, but not for longer time scales when the predictive relationships may themselves change (e.g., see O'Gorman and Dwyer 2018). Another caveat is that our statistical model has a predictive skill for the slowly varying (time averaged) spatial structure of rainfall, but not for day-to-day weather variability. Our goal is to use the empirical parameterization to simulate climate phenomena like El Niño, where the averaged rainfall patterns drive the time evolution of the coupled ocean–atmosphere system.

The predictor mode analysis identifies some interesting differences between the statistical fit to TRMM observations and the fit to CAM5 simulations. Our results suggest that the CAM5 convection parameterization is more sensitive to the barotropic vertical temperature structure and latitude compared to observations. Also, the observational fit suggests that wind shear can contribute to predictive skill—an effect that is not considered in model parameterizations. It is also surprising that the statistical model exhibits greater predictive skill when fitted to TRMM observations as opposed to CAM5 simulations, despite that fact that there is no measurement error to degrade prediction skill in the CAM5 simulated data. This suggests that the true relationship between atmospheric state and rainfall may perhaps be more linear than is represented by the CAM5 parameterization.

Further research is needed to extend the statistical framework to better handle nonlinear predictive relationships, and also to test the performance of these statistical relationships in an actual climate modeling framework (e.g., as in Brenowitz and Bretherton 2018).

Acknowledgments. The authors acknowledge Aaron Funk for providing the MERRA-2 and TRMM PR datasets binned to the same grid size. The authors thank the Editor, the Associate Editor, and four anonymous referees for their comments that helped to improve the manuscript significantly. M. Jun's research was partially supported by NIH P42ES027704 and NSF DMS-1613003, R. Saravanan's research was supported by NSF Grant AGS-1462127 and NOAA Grant NA16OAR4311082, and C. Schumacher's research was supported by NASA PMM Grant NNX16AE34G.

REFERENCES

Ahmed, F., and C. Schumacher, 2015: Convective and stratiform components of the precipitation–moisture relationship. *Geophys. Res. Lett.*, **42**, 10 453–10 462, https://doi.org/10.1002/2015GL066957.

——, and ——, 2017: Geographical differences in the tropical precipitation-moisture relationship and rain intensity onset. *Geophys. Res. Lett.*, **44**, 1114–1122, https://doi.org/10.1002/2016GL071980.

Awaka, J., T. Iguchi, H. Kumagai, and K. Okamoto, 1997: Rain type classification algorithm for TRMM precipitation radar. *Proc. 1997 Int. Geoscience and Remote Sensing Symp.*, IEEE, Singapore, 1633–1635, https://doi.org/10.1109/IGARSS.1997.608993.

Brenowitz, N. D., and C. S. Bretherton, 2018: Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.*, **45**, 6289–6298, https://doi.org/10.1029/2018GL078510.

Bretherton, C. S., M. E. Peters, and L. E. Back, 2004: Relationships between water vapor path and precipitation over the tropical oceans. *J. Climate*, **17**, 1517–1528, https://doi.org/10.1175/1520-0442(2004)017<1517:RBWVPA>2.0.CO;2.

Chen, B., C. Liu, and B. E. Mapes, 2017: Relationships between large precipitating systems and atmospheric factors at a grid scale. *J. Atmos. Sci.*, **74**, 531–552, https://doi.org/10.1175/JAS-D-16-0049.1.

Chen, Y., A. D. Del Genio, and J. Chen, 2007: The tropical atmospheric El Niño signal in satellite precipitation data and a global climate model. *J. Climate*, **20**, 3580–3601, https://doi.org/10.1175/JCLI4208.1.

Cho, H.-K., K. P. Bowman, and G. R. North, 2004: Equatorial waves including the Madden–Julian oscillation in TRMM rainfall and OLR data. *J. Climate*, **17**, 4387–4406, https://doi.org/10.1175/3215.1.

Dai, A., 2006: Precipitation characteristics in eighteen coupled climate models. *J. Climate*, **19**, 4605–4630, https://doi.org/10.1175/JCLI3884.1.

Funk, A., C. Schumacher, and J. Awaka, 2013: Analysis of rain classifications over the tropics by version 7 of the TRMM PR 2A23 algorithm. *J. Meteor. Soc. Japan*, **91**, 257–272, https://doi.org/10.2151/jmsj.2013-302.

Hannachi, A., I. Jolliffe, and D. Stephenson, 2007: Empirical orthogonal functions and related techniques in atmospheric science: A review. *Int. J. Climatol.*, **27**, 1119–1152, https://doi.org/10.1002/joc.1499.

Hartmann, D. L., H. Hendon, and R. Houze Jr., 1984: Some implications of the mesoscale circulations in tropical cloud clusters for large-scale dynamics and climate. *J. Atmos. Sci.*, **41**, 113–121, https://doi.org/10.1175/1520-0469(1984)041<0113:SIOTMC>2.0.CO;2.

Hou, A. Y., and Coauthors, 2014: The Global Precipitation Measurement Mission. *Bull. Amer. Meteor. Soc.*, **95**, 701–722, https://doi.org/10.1175/BAMS-D-13-00164.1.

Houze, R. A., 2004: Mesoscale convective systems. *Rev. Geophys.*, **42**, RG4003, https://doi.org/10.1029/2004RG000150.

Huffman, G. J., and Coauthors, 2007: The TRMM Multi-satellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeor.*, **8**, 38–55, https://doi.org/10.1175/JHM560.1.

——, R. Adler, D. Bolvin, and E. Nelkin, 2010: The TRMM Multi-Satellite Precipitation Analysis (TMPA). *Satellite Rainfall Applications for Surface Hydrology*, F. Hossain and M. Gebremichael, Eds., Springer, 3–22.

Hurrell, J. W., and Coauthors, 2013: The Community Earth System Model: A framework for collaborative research. *Bull. Amer. Meteor. Soc.*, **94**, 1339–1360, https://doi.org/10.1175/BAMS-D-12-00121.1.

Husak, G. J., J. Michaelsen, and C. Funk, 2007: Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *Int. J. Climatol.*, **27**, 935–944, https://doi.org/10.1002/joc.1441.

Iguchi, T., T. Kozu, R. Meneghini, J. Awaka, and K. Okamoto, 2000: Rain-profiling algorithm for the TRMM precipitation radar. *J. Appl. Meteor.*, **39**, 2038–2052, https://doi.org/10.1175/1520-0450(2001)040<2038:RPAFTT>2.0.CO;2.

Jensen, M. P., and A. D. Del Genio, 2006: Factors limiting convective cloud-top height at the ARM Nauru Island climate research facility. *J. Climate*, **19**, 2105–2117, https://doi.org/10.1175/JCLI3722.1.

Karl, T. R., R. W. Knight, and N. Plummer, 1995: Trends in high-frequency climate variability in the twentieth century. *Nature*, **377**, 217–220, https://doi.org/10.1038/377217a0.

Katz, R., 1999: Extreme value theory for precipitation: Sensitivity analysis for climate change. *Adv. Water Resour.*, **23**, 133–139, https://doi.org/10.1016/S0309-1708(99)00017-2.

Kelly, P., B. Mapes, I. Hu, S. Song, and Z. Kuang, 2017: Tangent linear superparameterization of convection in a 10 layer global atmosphere with calibrated climatology. *J. Adv. Model. Earth Syst.*, **9**, 932–948, https://doi.org/10.1002/2016MS000871.

Krasnopolsky, V. M., M. S. Fox-Rabinovitz, and A. A. Belochitski, 2013: Using ensemble of neural networks to learn stochastic convection parameterizations for climate and numerical weather prediction models from data simulated by a cloud resolving model. *Adv. Artif. Neural Syst.*, **2013**, 485913, https://doi.org/10.1155/2013/485913.

Kuang, Z., 2010: Linear response functions of a cumulus ensemble to temperature and moisture perturbations and implications for the dynamics of convectively coupled waves. *J. Atmos. Sci.*, **67**, 941–962, https://doi.org/10.1175/2009JAS3260.1.

Kummerow, C., W. Barnes, T. Kozu, J. Shiue, and J. Simpson, 1998: The Tropical Rainfall Measuring Mission (TRMM) sensor package. *J. Atmos. Oceanic Technol.*, **15**, 809–817, https://doi.org/10.1175/1520-0426(1998)015<0809:TTRMMT>2.0.CO;2.

Li, G., and S.-P. Xie, 2014: Tropical biases in CMIP5 multimodel ensemble: The excessive equatorial Pacific cold tongue and double ITCZ problems. *J. Climate*, **27**, 1765–1780, https://doi.org/10.1175/JCLI-D-13-00337.1.

Li, W., and C. Schumacher, 2011: Thick anvils as viewed by the TRMM precipitation radar. *J. Climate*, **24**, 1718–1735, https://doi.org/10.1175/2010JCLI3793.1.

Madsen, H., and P. Thyregod, 2010: *Introduction to General and Generalized Linear Models*. CRC Press, 316 pp.

McCullagh, P., and J. A. Nelder, 1989: *Generalized Linear Models*. 2nd ed. Chapman & Hall, 511 pp.

Neale, R. B., J. Richter, S. Park, P. H. Lauritzen, S. J. Vavrus, P. J. Rasch, and M. Zhang, 2013: The mean climate of the Community Atmosphere Model (CAM4) in forced SST and fully coupled experiments. *J. Climate*, **26**, 5150–5168, https://doi.org/10.1175/JCLI-D-12-00236.1.

O'Gorman, P. A., and J. G. Dwyer, 2018: Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *J. Adv. Model. Earth Syst.*, **10**, 2548–2563, https://doi.org/10.1029/2018MS001351.

Oueslati, B., and G. Bellon, 2015: The double ITCZ bias in CMIP5 models: Interaction between SST, large-scale circulation and precipitation. *Climate Dyn.*, **44**, 585–607, https://doi.org/10.1007/s00382-015-2468-6.

Pratt, J. W., 1987: Dividing the indivisible: Using simple symmetry to partition variance explained. *Proc. Second International Tampere Conf. in Statistics*, Tampere, Finland, Department of Mathematical Sciences, University of Tampere, 245–260.

Rasp, S., M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent sub-grid processes in climate models. *Proc. Natl. Acad. Sci. USA*, **115**, 9684–9689, https://doi.org/10.1073/pnas.1810286115.

Rienecker, M. M., and Coauthors, 2011: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Climate*, **24**, 3624–3648, https://doi.org/10.1175/JCLI-D-11-00015.1.

Schumacher, C., and R. A. Houze Jr., 2003: Stratiform rain in the tropics as seen by the TRMM precipitation radar. *J. Climate*, **16**, 1739–1756, https://doi.org/10.1175/1520-0442(2003)016<1739:SRITTA>2.0.CO;2.

——, ——, and I. Krauchunas, 2004: The tropical dynamical response to latent heating estimates derived from the TRMM Precipitation Radar. *J. Atmos. Sci.*, **61**, 1341–1358, https://doi.org/10.1175/1520-0469(2004)061<1341:TTDRTL>2.0.CO;2.

Sherwood, S. C., S. Bony, and J.-L. Dufresne, 2014: Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, **505**, 37–42, https://doi.org/10.1038/nature12829.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 176 pp.

Stephens, G. L., and Coauthors, 2010: Dreary state of precipitation in global models. *J. Geophys. Res.*, **115**, D24211, https://doi.org/10.1029/2010JD014532.

Stocker, T. F., and Coauthors, 2013: Technical summary. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 33–115.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Thomas, D. R., P. Zhu, B. D. Zumbo, and S. Dutta, 2008: On measuring the relative importance of explanatory variables in a logistic regression. *J. Mod. Appl. Stat. Methods*, **7**, 21–38, https://doi.org/10.22237/jmasm/1209614580.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.