Employing Cluster Analysis to Detect Significant Cloud 3D RT Effect Indicators

MICHAEL J. FOSTER

Space Science and Engineering Center, Department of Atmospheric and Oceanic Sciences, University of Wisconsin—Madison, Madison, Wisconsin

DANA E. VERON

College of Marine and Earth Studies, University of Delaware, Newark, Delaware

(Manuscript received 16 December 2009, in final form 6 March 2010)

ABSTRACT

Three-dimensional cloud field morphology contributes to scene-averaged cloud reflectivity, but climate models do not currently incorporate methods of identifying situations where this contribution is substantial. This work represents an effort to identify atmospheric conditions conducive to the formation of cloud field configurations that significantly affect shortwave radiative fluxes. Once identified, these characteristics may form the basis of a parameterization that accounts for radiative impact of complex cloud fields. A *k*-means clustering algorithm is applied to observed cloud properties taken from the Atmospheric Radiation Measurement Program tropical western Pacific sites to identify specific cloud regimes. Results from a stand-alone stochastic model, which statistically represents shortwave radiative transfer through broken cloud fields, are compared with those of a plane-parallel model. The aggregate scenes in each regime are examined to measure the bias in shortwave flux calculations due to neglected cloud field morphology. The results from the model comparison and cluster analysis suggest that cloud fraction, vertical wind shear, and spacing between cloudy layers are all important indicators of complex cloud field geometry and that these criteria are most often met in cloud regimes characterized by moderate to strong convection. The cluster criteria are applied to output from the Community Climate System Model (version 3.0) and it is found that the presence of persistent high cirrus cloud in model simulations inhibits identification of specific cloud regimes.

1. Introduction

Recent studies have indicated that one way to improve general circulation models (GCMs) is to improve their treatment of three-dimensional cloud field geometry. There are several challenges to realizing this goal, one of which is relating large-scale (GCM-scale) fields to unresolved subgrid-scale variability in the cloud and radiation properties (Potter and Cess 2004; Randall et al. 2007). When grid-averaged values of atmospheric variables are used to represent subgrid-scale nonlinear processes, as is often the case in GCMs, errors may arise within the model simulations of climate. One example of where this is known to occur is in the calculation of cloud reflectance (Cahalan 1994; Pincus and Klein 2000). These errors in turn propagate and adversely affect atmospheric heating rates, cloud formation and dissipation, and precipitation rates, leading to biases in the cloud feedback processes. This work is part of a continuing effort to identify dynamic and thermodynamic criteria for use in determining the impact that cloud field morphology, also known as macroscale inhomogeneity, has on the shortwave radiative budget, and to develop a GCM parameterization that incorporates these effects. The first step is to identify distinct cloud regimes using GCM-resolvable atmospheric properties, and then to relate the dynamic and radiative properties of each regime using high-resolution observations and radiative transfer model simulations.

Following work by Jakob et al. (2005) and Gordon et al. (2005), a *k*-means clustering algorithm (Anderberg 1973; Jakob and Tselioudis 2003; Jakob et al. 2005) is applied to cloud and radiation data observed at the Atmospheric Radiation Measurement (ARM) (Stokes and Schwartz 1994) Cloud and Radiation Testbed (CART) facility on Nauru Island from 2001 to 2004. The radiative

Corresponding author address: Michael J. Foster, 1225 West Dayton Street, Madison, WI 53706.
E-mail: mfoster@aos.wisc.edu

DOI: 10.1175/2010JAS3414.1

© 2010 American Meteorological Society

and dynamic characteristics of the resulting clusters are analyzed using independent data from the ARM archive, as well as atmospheric profiles of temperature, humidity, and wind taken from European Centre for Medium-Range Weather Forecasting (ECMWF) climate model output. Four distinct cloud regimes are identified: 1) a convectively active regime composed of multiple coincident cloud types, 2) a suppressed regime composed primarily of boundary layer clouds, 3) a convectively active, optically thin cirrus regime with low coverage, and 4) a convectively active cirrus regime with large coverage. Hourly stochastic and plane-parallel shortwave radiative transfer calculations are performed and matched to the presence of cluster members and then evaluated using observed surface fluxes. To extend the derived relationship between cloud field structure and radiative fluxes throughout the tropics, the clustering algorithm is also applied to data from the ARM CART Manus and Darwin facilities. A comparison of the observed cloud regimes to those simulated in a GCM is performed using output from the Community Climate System Model, version 3.0 (CCSM3). Using the cluster analysis results from all three ARM facilities as well as the radiative transfer model results, a set of 3D cloud-effect indicators is identified.

Section 2 details the data used for the clustering algorithm and the specifications for the model simulations. In section 3 the resulting cloud clusters are examined and dynamic properties for each cloud regime are described. Stand-alone plane-parallel and stochastic radiative transfer calculations are discussed for each cluster in section 4. Section 5 evaluates the inclusion of the Manus and Darwin data in the cluster analysis and the development of criteria for application of a statistical cloud radiation scheme. Section 6 details the application of observationally derived cloud regimes criteria to GCM output. The implications and conclusions from this study are discussed in section 7.

2. Methodology

a. Cluster analysis

The framework for the cluster analysis is based on several studies (e.g., Jakob and Tselioudis 2003; Jakob et al. 2005; Gordon et al. 2005; Williams et al. 2005) with a few key differences. While the aforementioned studies use satellite data from the International Satellite Cloud Climatology Project (ISCCP) climatology (Rossow and Schiffer 1999), this work uses surface-based measurements from a smaller spatial domain. Cloud liquid water path (LWP), geometric cloud-top height, and cloud coverage are used to generate the histograms needed for this analysis. Averages of the observed cloud properties are taken at 5-min intervals, yielding 36 periods from which to develop 3-h histograms. The data are grouped into 10 bins each, meaning each histogram contains 1000 elements. The bins are unequally spaced and chosen to maximize the ability of the clustering algorithm to differentiate among histograms. Analysis is restricted to histograms that contain 10 or more 5-min intervals, and times with clear sky or precipitation are not included. Precipitation is identified as when the 31.4-GHz brightness temperature measured by the microwave radiometer exceeds 100 K, or when liquid water path exceeds a predetermined threshold of 500 g m^{-2}. Precipitating times are excluded for two reasons: 1) condensation formation on the microwave radiometer Teflon window may cause retrieval errors, and 2) the radiative transfer models used in this study do not currently have a method of dealing with precipitating hydrometeors.

In applying this technique the number of clusters is specified beforehand, so an objective set of criteria must be used to determine a value for k. Rossow et al. (2005) used four criteria to objectively determine k: 1) the centroid histogram patterns should not change when the initial conditions are varied, 2) the centroid patterns should significantly differ from one another, 3) the spatial–temporal correlations of the cluster members should be low, and 4) the distances between cluster centroids should be greater than the distance between the cluster members and their centroid. Beginning with $k = 2$ and incrementing its value by 1 we determined that $k = 4$ provides the optimal number of clusters following these criteria. For values of k less than 4, changes in the initial centroids lead to significant changes in the mean histogram patterns, while for values of k greater than 4 the centroid patterns do not significantly differ from one another. Even for $k = 4$, changes in the initial centroid results in small deviations in the final mean cluster patterns. Therefore, the clustering algorithm was run several times using randomly chosen initial centroids and the variance around each centroid was calculated. The final cluster set chosen was that with the smallest sum of the variance, following Gordon et al. (2005).

b. Data

All the ground-based remote sensing data in this project are from the ARM tropical western Pacific (TWP) site, composed of the Manus, Darwin, and Nauru Island observational facilities. The initial cluster analysis is performed with data from Nauru Island that include the start of 2001 to the end of 2004. Liquid water path is derived from the ARM two-channel (23.8 and 31.4 GHz) microwave radiometer line-of-site (MWRLOS) product, cloud coverage is calculated using the Long et al. (2006) algorithm and data from the Shortwave Flux Analysis value-added product (VAP), and cloud-top height is taken

TABLE 1. Description of the CRM and DSTOC model configurations. Droplet effective radius is determined by a temperature (T)-dependent function with a minimum value of 6 μm, but in reality it almost always uses the minimum value.

	CRM	DSTOC
No. of vertical layers	32	32
Moisture profile	NCEP-derived	NCEP-derived
O_3, T, CO_2 profiles	McClatchey et al. (1972)	McClatchey et al. (1972)
Radiative transfer solver	Delta-Eddington method (Briegleb 1992)	Discrete ordinate with approximate iterative technique (Wiscombe and Evans 1977)
Shortwave spectrum	18 unequally spaced bands	38 unequally spaced bands
Cloud ice/water partitioning	Single-moment T-dependent	Single-moment T-dependent
Cloud overlap	Random	Random
Droplet effective radius	Usually 6 μm	Usually 6 μm
Ice crystal effective radius	23 μm	23 μm
Aerosol	0.07 visible extinction optical depth	None

from the Active Remotely Sensed Cloud Locations (ARSCL) VAP, which combines measurements from the Vaisala ceilometer, micropulse lidar, and millimeter wavelength cloud radar (Clothiaux et al. 2000; Kollias et al. 2005). It is important to note that many of these instruments employ zenith narrow-beam single-point measurements, so that the size and source of the spatial domain being examined is strongly dependent on the magnitude and direction of the wind. Diagnostic data derived from ECMWF model runs generated for TWP ARM sites provide profiles of wind speed and direction, which, along with cloud boundary data taken from the ARSCL VAP, are used to calculate wind shear between cloudy layers. The ECMWF diagnostic data also provide profiles of temperature and moisture that are used to calculate convective available potential temperature (CAPE). In addition to providing cloud profiles for the model simulations, the ARSCL product also contains cloud boundaries used to derive the geometric extent of clear sky between noncontiguous cloudy layers, Doppler velocity, and vertical profiles of hydrometeor reflectivity used to allocate cloud liquid water in each cloudy vertical layer.

The primary input variables for the radiative transfer models, hourly averaged layered cloud fraction and cloud liquid water content, are taken from the ARSCL VAP and MWR measurements. The model simulations use climatological values for surface albedo, ice crystal effective radius, and water droplet effective radius, which can be found in Table 1. The model simulations also use climatological values for cloud ice content, derived from the aforementioned ECMWF diagnostic data. The plane-parallel model uses an aerosol optical depth, calculated by taking the mean value of sun photometer measurements over the 4-yr period. Downwelling shortwave broadband radiation measurements are provided by the Shortwave Flux Analysis VAP, using data derived from ARM Sky Radiation (SKYRAD) radiometers. Global hemispheric shortwave irradiance is measured with an unshaded pyranometer with a hemispheric field of view, while diffuse shortwave irradiance is measured with a shaded pyranometer. Shortwave surface fluxes are averaged hourly and compared to coincident radiative transfer calculations to assess model performance. A comparison of these two models may be found in Foster and Veron (2008).

c. *Model description*

1) STOCHASTIC MODEL

The stochastic radiative transfer model used in this study, known as DSTOC, is based on a model originally developed by Malvagi et al. (1993). DSTOC is a more generalized form than the Malvagi et al. (1993) version and has been modified for use in a number of studies (Byrne et al. 1996; Lane et al. 2002; Lane-Veron and Somerville 2004; Foster and Veron 2008). The model requires input of cloud fraction and cloud liquid and ice water content for each cloudy layer, along with values of effective liquid droplet and ice crystal radii, surface albedo, and solar zenith angle. Liquid water path is allocated to each model liquid cloud layer using hydrometeor reflectivity measurements taken from the ARSCL VAP and then converted into cloud water content using cloud fraction and layer thickness, from which volume extinction and absorption coefficients are derived (Byrne et al. 1996; Lane-Veron and Somerville 2004).

DSTOC generates ensemble-averaged radiative fluxes for multiple clear- and cloudy-sky scenarios sharing the same statistics for each spectral band and model layer. Solving for an ensemble of stochastic realizations generates the statistical variances of cloud-field properties required to calculate the nonlinear reflectance of clouds. The standard time-independent radiative transfer equations are modified to contain two additional terms that

describe the cloud-field geometry using conditional linear probabilities (Byrne et al. 1996) as shown below:

$$\Omega \cdot \nabla(p_i I_i) + \sigma_i p_i I_i = \sigma_{si} \int_{4\pi} f_i(\Omega\Omega') p_i I_i(\Omega')\, d\Omega' + \frac{p_j \bar{I}_j}{\lambda_j} - \frac{p_i \bar{I}_i}{\lambda_i}, \quad i = 0, 1, \quad j \neq i, \tag{1}$$

where $i = 0$ denotes clear sky and $i = 1$ denotes cloud, Ω is the direction of flow, p is the local probability of material i being present, I is the specific intensity of radiation, ρ is the macroscopic total cross section, ρ_s is the macroscopic scattering cross section, $f(\Omega\Omega')$ is the single-scatter angular distribution function, $\bar{I}$ is the conditional ensemble-averaged intensity, and χ is the scale length for transition from one material to another. The statistical line theory used to calculate the distribution of clear sky and cloud requires a probability distribution function of cloud chord lengths; currently this function is governed by Markovian statistics. The distribution is approximated hourly using National Centers for Environmental Prediction (NCEP)-derived layered horizontal wind speed and corresponding layered cloud fraction taken from the ARSCL VAP. The product of these variables is treated as the integration of a Markovian distribution of cloud chord lengths. This distribution is dependent on the size of the model horizontal domain, which in turn is related to the scale of the inhomogeneity in the cloud field; in this case the domain size is approximately 20 km per side.

2) COLUMN RADIATION MODEL

The Column Radiation Model (CRM) is a stand-alone version of the plane-parallel radiative transfer code employed in the NCAR Community Climate Model (CCM3; Kiehl et al. 1998). The CRM is representative of shortwave radiative transfer codes used in many present-day GCMs. The CRM utilizes the delta-Eddington approximation described in Briegleb (1992) to solve the radiative transfer equation. The shortwave spectrum is divided into 18 unequally spaced bands with wavelengths ranging from 0.2 to 5.0 μm. Absorption sources include ozone, carbon dioxide, water vapor, and oxygen as well as cloud water and ice. Scattering sources include molecules, cloud, and aerosols, with isotropic scattering assumed between vertical layers. Profiles of temperature, carbon dioxide, and ozone at 32 unequally spaced vertical layers match those used by DSTOC and are taken from the McClatchey et al. (1972) climatology whereas the moisture profile is derived from NCEP model results. A double-moment microphysical function is normally used in the CRM to partition cloud water and ice, but it was replaced in this study by a temperature-dependent single-moment function that assumes clouds below −15°C are ice clouds and above are water clouds in order to more closely match the stochastic model microphysics. A series of CRM runs were performed utilizing the double-moment and single-moment functions, and the resulting difference was negligible. The in-cloud microphysical properties for each layer are homogenous with adjacent cloudy layers overlapped randomly. For cloudy layers with temperatures above −15°C the liquid water path is allocated among the layers based on radar reflectivities. For temperatures below −15°C, the clouds in both models are treated as ice with a climatological ice content derived from ECMWF diagnostic data.

3) COMMUNITY CLIMATE SYSTEM MODEL

The CCSM is a fully coupled global climate model composed of four primary components simulating the earth's atmosphere, ocean, land surface, and sea ice. Version 3 of the Community Atmosphere Model (CAM3) represents the sixth generation of the atmospheric component of the CCSM. CAM3 is a three-dimensional global spectral model capable of being run either in a stand-alone or coupled mode; it uses 26 vertical layers (Collins et al. 2006). There are significant improvements in the treatment of cloud microphysical and condensation processes introduced in CAM3. These include the separate treatment of cloud water and ice condensate, advection of these variables in large-scale circulations, improvement in convective parameterizations, and consistent treatment of cloud particles including sedimentation and radiative properties (Boville et al. 2006).

The simulations used in this study are among those generated for the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) Special Report on Emission Scenarios (SRES) and are T85 resolution, which equates approximately to 1.4° × 1.4° grid cell size or 256 × 128 horizontal grid points globally. They can be found at the Earth System Grid Web site (http://www.earthsystemgrid.org). The SRES results used are those from the A1F1 scenario, which represents a continued heavy reliance on the burning of fossil fuels throughout the twenty-first century. This scenario was chosen for a number of reasons: 1) the aerosol and greenhouse gas emissions are not drastically changed for the purposes of an idealized scenario (e.g., a complete freeze at year 2000 levels); 2) the output is written at 6-h intervals and uses instantaneous as opposed to mean values; and 3) additional cloud-related output is available for this scenario, such as cloud fraction at each vertical level.

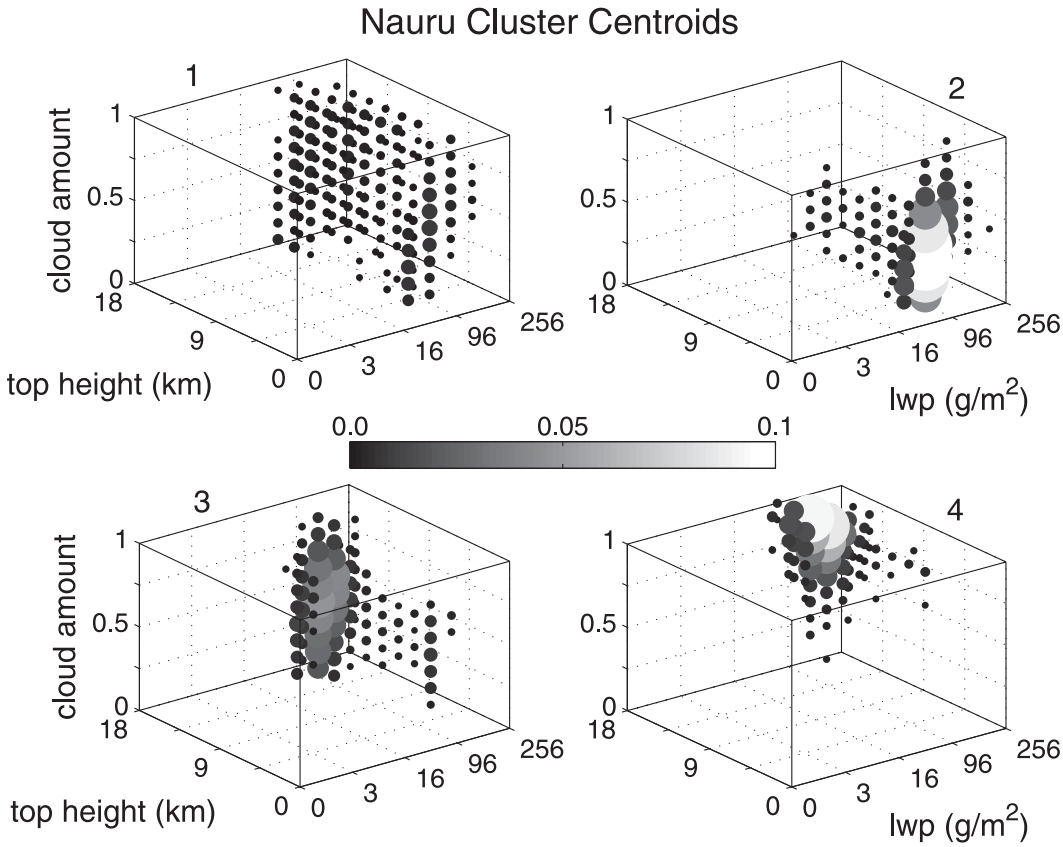

FIG. 1. Three-dimensional centroids—(top left) 1, (top right) 2, (bottom left) 3, and (bottom right) 4—resulting from the *k*-means clustering of LWP, cloud-top height, and total cloud coverage as measured by surface instrumentation on Nauru Island from the beginning of 2001 to the end of 2004. The larger the circle and lighter the shade of gray, the greater the relative frequency of occurrence.

3. Cluster analysis results

a. Cluster description

The *k*-means clustering algorithm is applied to the histograms described in section 2a using $k = 4$ for the number of clusters. The resulting four clusters are shown in Figure 1. Cluster 1 contains multiple coincident cloud types ranging from boundary layer to cirrus. Because of the presence of multiple cloudy layers, this regime has the most variance in cloud coverage, cloud-top height, cloud geometric thickness, and liquid water path. This cluster also has a positive (upward) mean Doppler velocity, indicating it may be a convectively active regime, and cluster 1 seems to be the only cluster that contains significant amounts of midlevel cloud. The second cluster is dominated by low boundary layer clouds with low total coverage. This is a stable regime with little convective activity. The third cluster represents a regime dominated by high cirrus, but it has relatively low liquid water path and total cloud coverage. This is a convectively active regime. The fourth and final cluster is also convectively active and composed of high cirrus. However, this cluster has higher liquid water path than cluster 3 and much larger cloud coverage, suggesting the presence of deep convective activity and cirrus outflow. The relative frequencies of occurrence of the regimes are 64%, 15%, 13%, and 8%, respectively. Figure 2 displays mean and standard deviation values for some key cluster characteristics.

It is worth noting that the results from this analysis and one of the primary studies on which it is based, that of Jakob et al. (2005), share certain similarities. Both studies identify four major cloud regimes, three of which are composed primarily of high-top clouds while one contains low-top clouds. Two of the regimes are convectively active and one is suppressed. From there, however, differences can be seen. Cluster 1 may be thought of as a "mixed" cluster with a high relative frequency of occurrence (RFO) and large variance in cloud properties, which does not correspond with any of the regimes identified in Jakob et al. (2005). This may be attributable to differences in how the cluster histograms are generated. This study

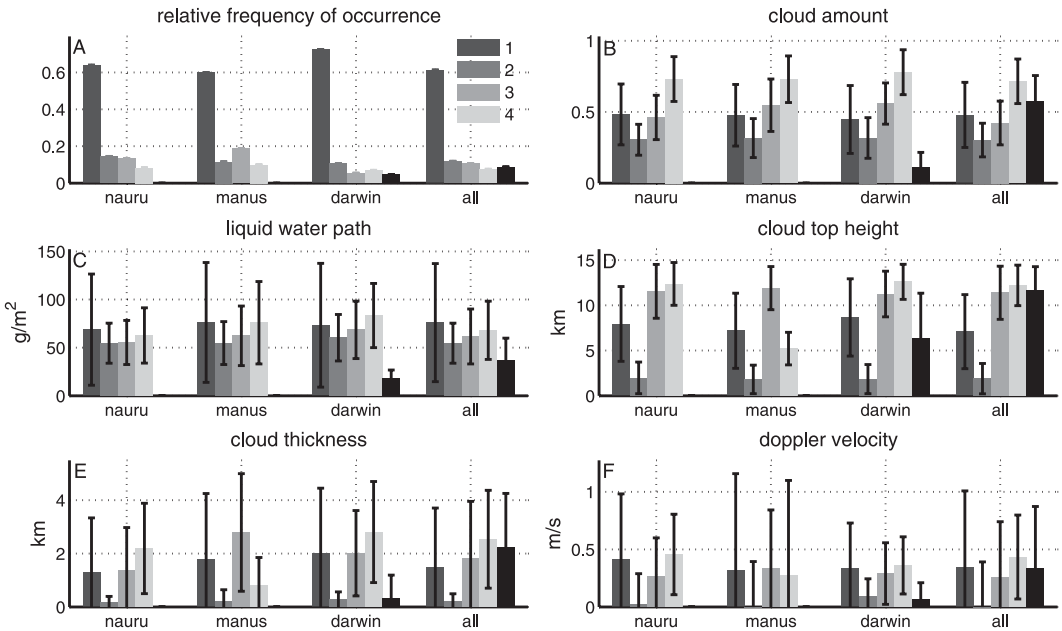

FIG. 2. Mean and 1 std dev of atmospheric variables for the clusters generated by the *k*-means clustering algorithm at the ARM TWP facilities from the beginning of 2001 to the end of 2004. The legend in (a) displays the shades of gray of the four clusters generated at the Nauru facility and found throughout the other facilities. The fifth cluster found at the Darwin facility and in combining all three TWP facilities do not correspond with any of the four Nauru clusters.

uses surface instruments with zenith-pointing single-point measurements, as opposed to satellite sensors with relatively large footprints. This means that the histograms evolve at a single point over each 3-h window, while the ISCCP histograms are generated from a snapshot of a spatial domain. What this implies for the mixed cluster is that at least some of the member histograms represent transitional periods from one regime to another. Different averaging intervals and time periods have been tested to minimize this effect, and it was found that the 3-h period worked best. The large relative frequency of occurrence of cluster 1 also begs the question of whether it may be broken up into smaller clusters. This was not observed; even when repeating the analysis with k values up to 12 it was found that cluster 1 consistently existed and contained a large number of member histograms.

b. Spatial domain

The magnitude and direction of the wind at the Nauru facility are analyzed to determine the size and location of the spatial domain being observed by the zenith-pointing surface instruments. The mean wind speed throughout the observed vertical column ($\sim$200 m $\times$ 200 m $\times$ 20 km) by cluster is 5.0, 4.9, 5.4, and 4.9 m s^{-1} for clusters 1, 2, 3, and 4, respectively. The wind direction is dominated by the easterly trade winds. This indicates that the size and source of the spatial domain is consistent for all four identified cloud regimes: roughly 20 km $\times$ 20 km size and approaching from the east of Nauru.

c. Convective available potential energy

The next step is to determine characteristic dynamic properties of the four cloud regimes. Vertical velocity may serve as one indicator of convective activity, as may CAPE. CAPE is calculated for the entire atmospheric column and for the area below 5 km where clouds are composed primarily of water. We examine the relationship between CAPE and cloud-top height by cloud regime. Using the low-level time derivative of CAPE (Zhang 2009) and cloud-top height, Fig. 3 shows relative frequency-of-occurrence histograms for each of the four cloud regimes described earlier. Clusters 3 and 4 display similar patterns of high cloud top with CAPE mean values of 46 and 41 J kg^{-1} (see Table 2), respectively, and a relatively stable distribution across the range of values. Interestingly, although cluster 2 is composed primarily of stable low-level boundary layer clouds, it contains a relatively high mean value of CAPE at 53 J kg^{-1}. Cluster 1 displays a wide array of cloud-top heights with a CAPE mean value of 46 J kg^{-1}. To add context, we calculate convective inhibition (CINH), which may be thought of as the energy that an air parcel must overcome before deep convection may develop. The mean CINH values below 5 km are 279, 261, 288, and 299 J kg^{-1} for clusters 1–4, respectively. This suggests that complex thermodynamic processes may be occurring for clouds in the second cloud regime, as these clouds coincide with the highest values of CAPE and the lowest values of CINH.

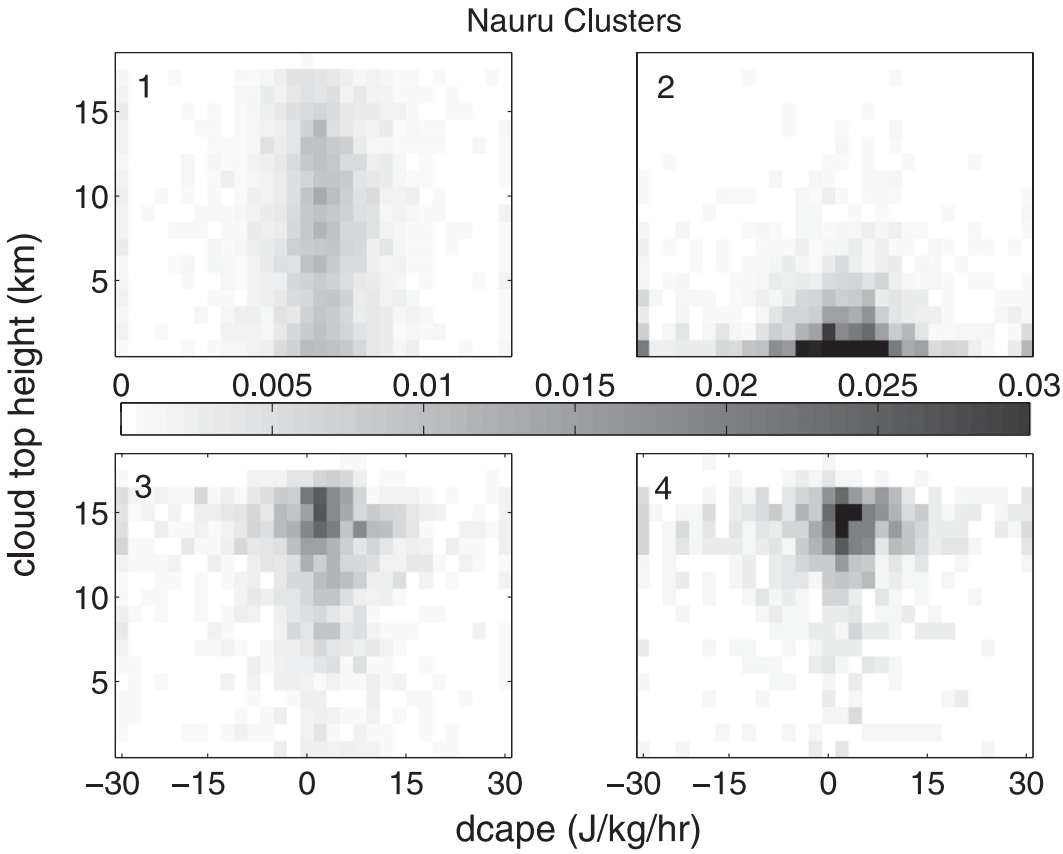

FIG. 3. Time derivative of CAPE and cloud-top height RFO histograms for each of the four Nauru clusters.

Signal attenuation due to the presence of low thick clouds may be part of the reason more convective cloud is not detected in cluster 2. Other possible explanations for the limited cloud-top heights of cluster 2, at least for the case of cumulus congestus clouds, include entrainment of dry air and the presence of weak stable layers near the freezing level (Redelsperger et al. 2002).

d. Wind shear and vertical spacing

Wind shear and vertical spacing between cloudy layers may be useful factors in identifying appropriate conditions for complex cloud fields because they are both related to cloud overlap (Naud et al. 2008). Vertical spacing between cloudy layers is defined as the geometric extent of clear sky between the top height of a cloud detected by the ARSCL VAP and the bottom height of a cloud located above it, assuming the layers are noncontiguous. Wind shear is calculated between these same heights. Table 2 contains wind shear and vertical cloud spacing information for each cluster, and Fig. 4 shows relative frequency-of-occurrence histograms for wind shear and cloud spacing based on cluster membership. Clusters 3 and 4 display similar patterns, with relatively high wind shear and small spacing between cloudy layers when compared to the stable boundary layer clouds found in cluster 2. Cluster 3 has wind shear greater than 0.003 s^{-1} (moderate strength) 30% of the time while cluster 4 has moderate or higher wind shear strength 32% of the time.

TABLE 2. Mean and 1 std dev values of CAPE, CINH, geometric extent of clear sky between noncontiguous cloudy layers, and wind shear between noncontiguous cloudy layers for the four clusters generated at Nauru.

	CAPE (J kg^{-1})		CINH (J kg^{-1})		Cloud spacing (m)		Wind shear (s^{-1})	
Cluster	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev
1	46	39	279	118	2514	2250	2.40×10^{-3}	2.00×10^{-3}
2	53	40	261	108	3758	3789	1.70×10^{-3}	1.50×10^{-3}
3	46	39	288	116	2059	1555	2.60×10^{-3}	2.20×10^{-3}
4	41	37	299	119	1895	1145	2.70×10^{-3}	2.20×10^{-3}

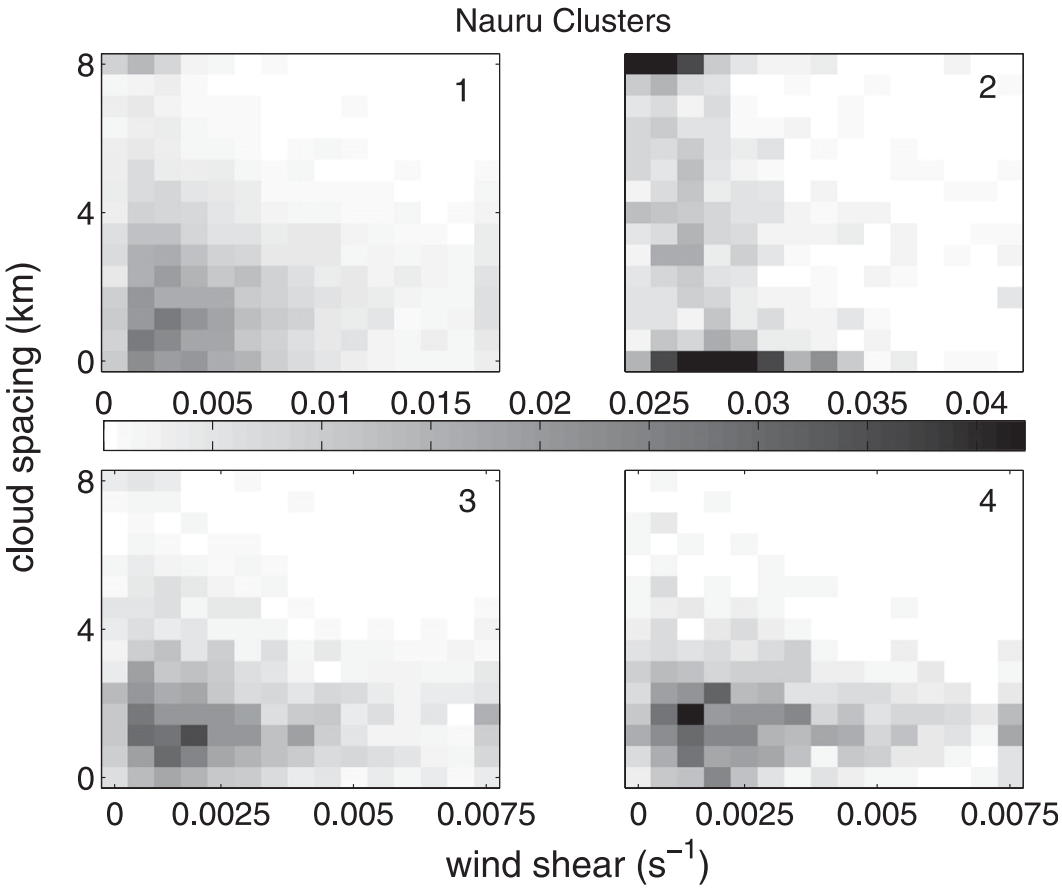

FIG. 4. Mean wind shear between cloudy layers and mean spacing between cloudy layers RFO histograms for each of the four clusters.

Moderate or higher wind shear is present only 13% of the time for cluster 2 and 26% of the time for cluster 1. The small values for spacing between cloudy layers could indicate that a maximum overlap assumption would be appropriate for these clusters, although Naud et al. (2008) found that large values for wind shear sometimes warrant a minimum overlap assumption. Cluster 2 has the lowest values for wind shear and the highest concentration of both very low and very high (8 km and greater) spacing between cloudy layers. However, 36% of the time there is only a single cloud layer present in cluster 2, which is not included in these histograms. Single cloud layers are present only 9%, 5%, and 3% of the time for clusters 1, 3, and 4, respectively. The times with very high cloud spacing are due to the coincident presence of low boundary layer clouds and high cirrus and account for only 9% of cluster 2. Cluster 1 contains a wide array of different cloud spacing and wind shear. The large size and wide range of values in cluster 1 make it difficult to characterize cloud field morphology or overlap type.

4. Model performance

The DSTOC and CRM models are run using the same data sources for forcing data as are employed by the clustering algorithm (ARSCL VAP, MWRLOS VAP, Shortwave Flux Analysis VAP) from the beginning of 2001 to the end of 2004. Limiting the runs by the availability of instrument measurements and sunlit hours reduces the number of hours run from the 35 064 possible to just under 8000, or a little less than half the daytime hours. For this study we focus on the models' ability to simulate downwelling solar surface irradiance (SSI) relative to measurements taken from the Shortwave Flux Analysis VAP. It is expected that DSTOC will simulate SSI closer to that observed as compared to the CRM for scenes where cloud field configuration significantly affects shortwave fluxes (Foster and Veron 2008). Figure 5 displays the performance of DSTOC and the CRM when compared to one another and observations, separated by cluster membership. In this context the term ''outperform'' is defined as one model

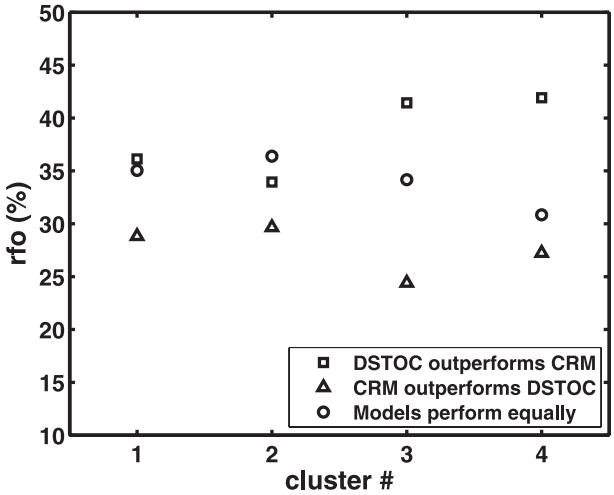

FIG. 5. RFO plot for each of the four clusters comparing the DSTOC and CRM model abilities to generate downwelling solar surface irradiance close to that observed. One model outperforms another when results are closer to observations by at least 5% of the observed value.

generating SSI that is at least 5% closer to that observed than the other model. For example, if the observed SSI is 100 W m^{-2} then a model must generate SSI at least 5 W m^{-2} closer to 100 W m^{-2} than the other model to be considered as outperforming it. When both models are within this 5% range they are considered to be performing equally well. The occurrence of one model outperforming the other varies considerably among the clusters. Cluster 1 shows DSTOC outperforming the CRM 36% of the time and the models performing equally 35% of the time. Cluster 2 shows similar performance by both models. Clusters 3 and 4 show considerable differences between model performance, with DSTOC outperforming the CRM 41% and 42% of the time, respectively (the CRM outperforms DSTOC 24% and 27% of the time). Overall the stochastic model performs best for clusters 3 and 4, which share frequent moderate-to-strong vertical wind shear and small spacing between noncontiguous cloudy layers.

5. Expansion of cluster analysis spatial domain

a. Incorporating the Manus and Darwin facilities

For these objectively derived cloud regimes to be useful as criteria for a GCM parameterization, they must occur over an area significantly larger than a single island in the tropical western Pacific. For this reason it is helpful to expand the spatial domain of the cluster analysis to include the Darwin and Manus facilities. The time period examined at Manus coincides with that of Nauru, from the beginning of 2001 to the end of 2004, but the Darwin site is newer than its counterparts and does not have coincident measurements of liquid water path, cloud coverage, and cloud top during this period. Therefore, the cluster analysis at Darwin is performed for the period from the beginning of 2006 through July 2007. This may have some effect on the resulting clusters at Darwin because of interannual variability from events such as the El Niño–Southern Oscillation.

b. Manus

Figure 2 details some properties for the four clusters generated at the Manus facility. Three of the four clusters share features similar to those generated at Nauru, and the shades of gray for the corresponding Manus and Nauru clusters match to indicate this. Specifically, cluster 1 at Manus is a frequently occurring mixed cluster composed primarily of multiple coincident cloud types. Cluster 2 at Manus is composed primarily of low boundary layer clouds with relatively low cloud coverage, very low geometric thickness, and very small or negative (downward) Doppler velocities indicating more stable conditions. Cluster 3 at Manus is composed primarily of high cirrus clouds with medium coverage and a positive median Doppler velocity, indicating the possibility of convective activity. Finally, cluster 4 at Manus is similar to cluster 4 at Nauru in that both clusters primarily contain clouds with nearly overcast to completely overcast coverage that is optically thick. One important difference between cluster 4 at Manus and Nauru is that the mean value of cloud-top height for Manus is around 5 km while the mean cloud-top height at Nauru is close to 13 km. It is possible that the presence of weak stable layers or the entrainment of dry air above the boundary layer (Redelsperger et al. 2002) is more prominent at Manus, limiting cloud-top heights for convective clouds and generating more cumulus congestus.

c. Darwin

Figure 2 also contains information about the five clusters generated at the Darwin facility. Four of the clusters generated at Darwin share have characteristics similar to those of the Nauru clusters. Similar to the description of the Manus clusters, clusters 1–4 at Darwin correspond with clusters 1–4 at Nauru, and three of the clusters correspond closely with those found at Manus. The fifth cluster at Darwin has a small relative frequency of occurrence (0.05); very small cloud amount with a mean of 0.08; small Doppler velocity, geometric thickness, and liquid water path; and a mean top height of just over 5 km. It does not correspond with any cluster from either Nauru or Manus.

d. All ARM TWP facilities

Finally, Fig. 2 contains information about the five clusters generated using the histograms from all three

ARM TWP facilities: Manus, Darwin, and Nauru. When the clustering algorithm is run with $k = 4$, the clusters generated match those at the Nauru site. With values of k above 5, smaller clusters appear that are generally just small variations on these four primary cloud regimes. For example, depending on the selection of the initial centroids, several iterations split the low-level boundary layer regime into two clusters containing mean values of cloud fraction of 0.2 and 0.4 instead of a single cluster with a mean value of 0.3. The selection of $k = 5$ generates the four clusters at Nauru and a fifth that is very similar to the convective regimes in cluster 3, but with greater cloud coverage and lower mean liquid water path. This regime is composed primarily of cirrus outflow from neighboring convective towers.

The results of this analysis indicate that regional variations exist in the composition and frequency of specific cloud regimes, but that these differences tend to be small. Certain cloud regimes, such as convectively active high coverage cirrus, low-level boundary layer stratus, and medium coverage cirrus exist at several locations with large enough frequency to suggest that improving the radiative treatment of cloud-field geometry for these regimes could prove beneficial for many parts of the tropics. The multilayer regime, which exists with very large frequency at all sites, is more difficult to characterize because it has the largest variance among its cloud properties.

6. Identification of observed cloud regimes in CAM3

a. *Liquid water path, cloud-top height, and cloud coverage*

The results from the previous sections suggest that cloud fraction, vertical wind shear, and spacing between cloudy layers are all important indicators of complex cloud field geometry, and that these criteria are most often met in the cloud clusters characterized by moderate to strong convection. A reasonable next step, then, is to compare the criteria/clusters obtained from these high-resolution, single-point observations to those obtained from relatively low-resolution GCM simulations that span the entire tropics. Previous studies have used clusters derived from satellite observations over the tropics, but, as mentioned earlier, differences in the variables, spatial domain size, and time period make it difficult to directly link these clusters to those derived here.

Distributions of cloud-top height, total cloud coverage, and liquid water path are used for the clustering algorithm in the previous chapter because of challenges in directly comparing hourly observations to model-generated atmospheric properties. This should therefore not only evaluate the ability of CAM3 to produce observed atmospheric conditions but also assess the potential to identify the cloud regimes generated from the clustering algorithm using GCM-derived cloud properties. Figure 6 shows distributions of cloud-top height, liquid water path, and cloud coverage generated by CAM3 and observed at the ARM Nauru facility. CAM3 output for the SRES scenarios is written every 6 h, so observations taken at matching times (~3000 h) are used for this comparison. While the modeled distributions of liquid water path are quite similar to those observed, there are marked differences in the distributions of cloud-top height. CAM3 simulates clouds with top heights above 14 km 98% of the time, compared to only 68% of the time for the ARM data. This suggests that cloud regimes with low- to midlevel cloud tops will occur with little frequency in the CAM3 simulations. The distribution of cloud coverage shows greater frequency of overcast conditions simulated by CAM3 than that derived from the ARM data. This difference in cloud coverage is greatest for low clouds (>740 hPa), where the mean cloud fraction for CAM3 is 0.26, compared to 0.13 for the ARM data. Mean high cloud fraction (<400 hPa) is 0.50 for CAM3 and 0.45 for ARM data, while mean midlevel cloud fraction is 0.28 and 0.25, respectively.

There are a number of possible reasons for the persistent high cirrus in the CAM3 simulations. The transition from CCM3 to CAM2 introduced a cold bias near the tropical tropopause, resulting in a dry bias for stratospheric water vapor (Boville et al. 2006). In CAM3 the treatment of subvisible cirrus clouds was improved by separating the treatment of cloud ice and liquid particles and including additional sources and sinks, such as large-scale advection of cloud and gravitational settling of cloud particles. The result was that the radiative imbalance causing the cold bias in CAM2 was largely removed. It is possible that an increase in the stratospheric water vapor could generate more subvisible cirrus, causing the issues seen here. It is also possible that surface instruments used for the ARSCL VAP may not detect some optically thin high cirrus cloud, as the heights being discussed are near the upper limits of lidar range, and cirrus clouds are common in this region. This question is addressed using Cloud–Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) data, which use active lidar and passive infrared measurements to probe thin cloud properties globally. Specifically, measurements from the Cloud–Aerosol Lidar with Orthogonal Polarization (CALIOP) are collected for 1 yr over the Nauru site and a distribution of cloud-top height is generated in Fig. 7. Although the CALIOP measurements detect cloud tops between 12 and 16 km more often than

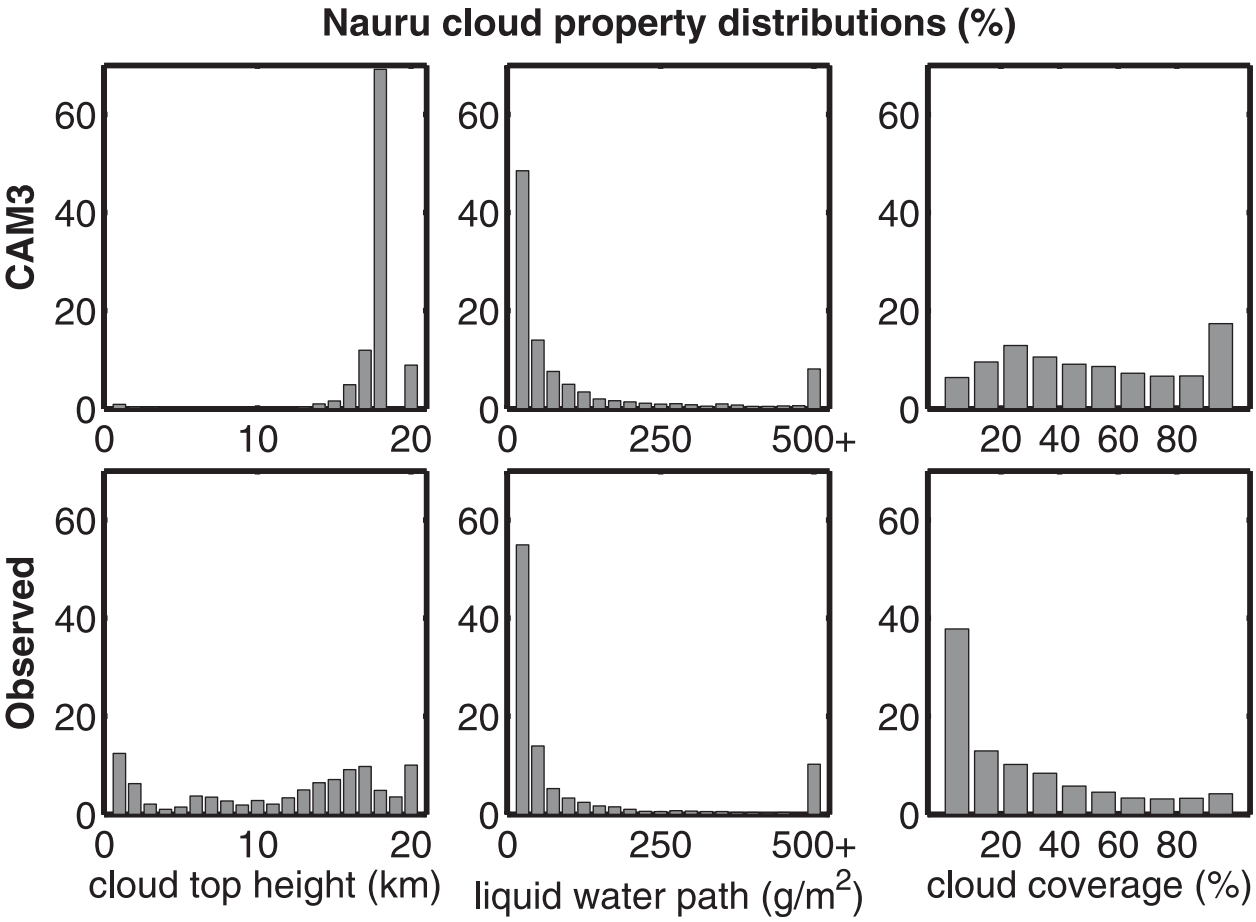

FIG. 6. Distribution of cloud-top height, LWP, and cloud coverage for observed vs model-generated 6-hourly data from the beginning of 2001 to the end of 2004 at the Nauru Island ARM facility. (top) CAM3 simulations and (bottom) data from ARM ARSCL and MWRLOS VAPs. The location of Nauru Island is 0.521°S, 166.891°E; the closest CAM3 grid cell is centered at 0.7004°S, 167.3438°E.

ARSCL, they detect cloud tops between 18 and 20 km only approximately 1% of the time compared to 53% in CAM3 simulations. In this respect the CALIOP distribution more closely matches that generated from the surface-based lidar measurements, suggesting that the persistent high cirrus found in the CAM3 simulations are not physically based. Finally it should be noted that the observed distribution of cloudiness does not include clear-sky conditions, but approximately 46% of the time no cloud is detected over Nauru. The CAM3 simulations predict clear sky less than 1% of the time.

b. *Identification of clusters*

Figure 8 shows relative frequency-of-occurrence maps throughout the tropics for the four clusters generated at the ARM Nauru facility and the summed occurrence of all four clusters. The maps are generated using output from the CAM3 simulations. The three variables used in the *k*-means clustering algorithm—cloud-top height, cloud coverage, and liquid water path—are used as criteria for the CAM3 output. Values of cloud-top height, cloud coverage, and liquid water path that are within 2 standard deviations of the mean values for each cluster are considered to have met the required criteria for that cluster. A cluster is considered present when all three variables fall within the specified range of values. Cluster 3, the moderate coverage cirrus regime, maintains a consistent presence throughout the tropics, with an occurrence of between 30% and 40% over much of the ocean. There is a small drop in occurrence around New Zealand, which is the approximate location of the ARM TWP facilities. Cluster 3 criteria are met in the CAM3 output 24%, 33%, and 37% of the time at Manus, Nauru, and Darwin, respectively. Cluster 4, the other convectively active cluster with large mean cloud-top height, occurs in the CAM3 output 17%, 18%, and 12% of the time at the respective sites, while the criteria for clusters 1 and 2 are

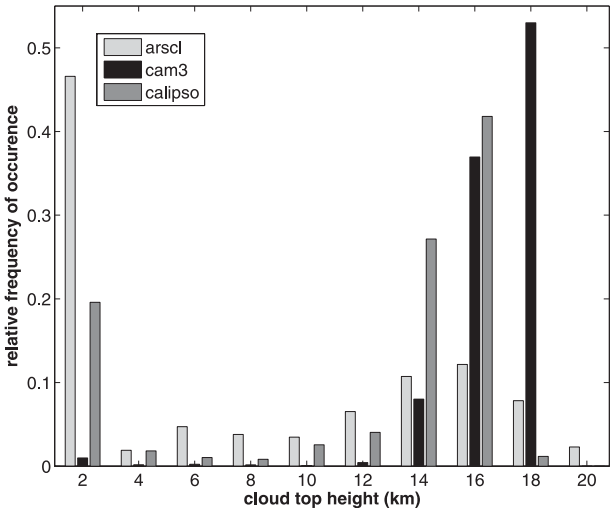

FIG. 7. Distribution of cloud-top height over Nauru Island for the year 2007. The measurements are taken from the CALIOP instrument located on the CALIPSO satellite. The RFO is calculated from nearly 6000 CAM3 and ARSCL data points and nearly 3000 CALIPSO data points.

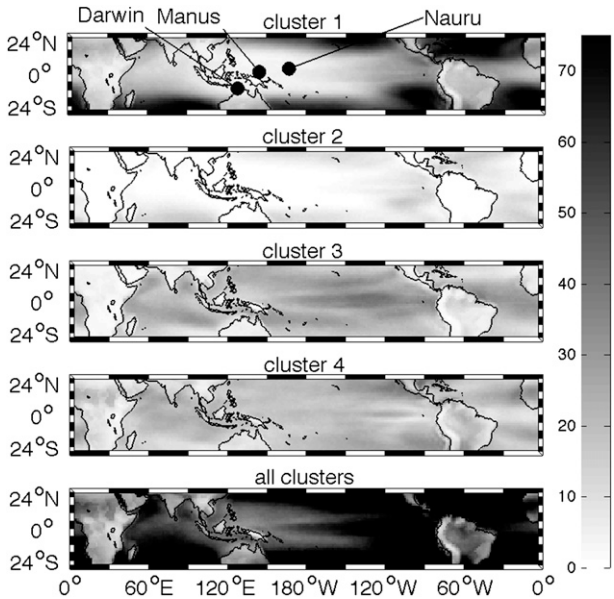

FIG. 8. RFO maps generated using CAM3 simulations in the tropics from 2001 to 2004. The gray bar represents the percent occurrence of the clusters described in section 3.

met less than 2% of the time at any of the sites. Clusters 1 and 2, the boundary layer and mixed cloud regimes, occur most frequently bordering the subtropics and off the coasts of South America and infrequently over the Pacific and Indian Oceans. The infrequent occurrence over the tropical warm pool is due to persistent high cirrus generated in the CAM3 simulations. Figure 8 shows total frequency-of-occurrence maps for all clusters. The areas of 100% occurrence and greater are due to cluster overlap, which is a result of using the inclusive criteria of two standard deviations from the mean values of liquid water path, cloud coverage, and cloud-top height. Although the border areas near the subtropics and the western coast of Brazil experience high frequency of occurrence, most of the tropics meet cluster criteria less than 50% of the time, and this can be attributed almost solely to clusters 3 and 4.

Much of the reason why the CAM3 cluster occurrence does not match that generated using the surface-based measurements at the ARM TWP facilities can be attributed to the CAM3 cloud-top height. The persistent presence of cirrus clouds with top heights above 14 km makes it difficult for the clusters with low- and midlevel cloud-top heights (clusters 1 and 2) to form. This makes clustering based on cloud-top height problematic. There are a number of possible approaches to this problem. One would be to avoid using cloud-top height and choose another variable such as cloud-base height or thickness. However, preliminary results (not shown) indicate significant differences in the observed cloud thicknesses versus those generated by CAM3, and it is uncertain how well cloud-base height may determine the presence of specific cloud regimes. Another possibility would be to look for the presence of cloud within the range of values for cloud-top height for each cluster, but this may lead to considerable overlap among the clusters. An approach currently being explored is to remove the subvisible cirrus generated by CAM3 and detected by ARSCL when determining cloud-top height.

7. Conclusions

The application of the *k*-means clustering algorithm to the surface-based measurements of atmospheric state taken from Nauru Island has produced four cloud regimes with distinct characteristics. Three of the four regimes show signs of being convectively active, while the other regime appears to be composed primarily of stable boundary layer clouds. Although the convectively active regimes have generally higher cloud tops then the stable regime, we find only a weak relationship between cloud-top height and CAPE. Examination of wind shear and vertical spacing between cloudy layers suggest that the convectively active regimes tend toward large values of wind shear and smaller spacing between cloudy layers than does the stable regime.

To estimate the macroscale inhomogeneity associated with these four cloud regimes, downwelling surface shortwave fluxes are calculated by a stochastic radiative

transfer model and compared against that of a plane-parallel RT model for the regimes described above. It is found that the stochastic model is able to capture more of the variability of the observed solar radiative cloud forcing at the surface than the CRM for all four clusters, and that this difference is particularly noticeable for clusters 3 and 4. These clusters represent regimes that often contain cirrus clouds with large ranges of cloud coverage and liquid water path. Clusters 3 and 4 also have the highest mean values for wind shear between cloudy layers and lowest mean values for spacing between cloudy layers. The combination of cloud-top height, wind shear, and spacing between cloudy layers directly affects the depth of the cloud field and the vertical and horizontal spacing between clouds, making them good indicators of significant macroscale inhomogeneity in the cloud fields.

Other variables such as cloud-top height require additional interpretation to add insight to this analysis. For example, cluster 4 has high cloud tops, the strongest vertical velocity, and the smallest spacing between cloudy layers of all the clusters, yet the stochastic model outperforms the CRM more frequently under cluster 3 conditions. The most likely explanation for this is the large total cloud coverage and relatively high optical depth of the first cluster, since optically thick overcast or nearly overcast skies are likely to minimize the importance of radiative effects due to interactions between clouds. Cluster 1 is difficult to characterize radiatively and generally because its high occurrence and variety of different cloud types generate a large range of values for all cloud properties.

The cluster analysis is expanded to include the ARM Manus and Darwin facilities. It is found that although there are small changes in the relative frequency of occurrence and the number and structure of the clusters, the primary cloud regimes found at Nauru may also be found at these other locations. This suggests it should be possible to apply a parameterization developed at Nauru to other areas in the tropics.

Finally, output from a CAM3 climate change scenario generated for use in IPCC AR4 is evaluated to assess how well the modeled distributions of cloud variables such as cloud-top height, liquid water path, and cloud coverage relate to the observed ones, which in turn may be used to assess how well GCM output corresponds to the observed cloud clusters. CAM3 reproduces a distribution of liquid water path that is very similar to that observed but displays marked differences in the distributions of cloud-top height and cloud coverage. Specifically, CAM3 generates persistent, often overcast, cirrus cloud with very high top heights, while the surface-based observations at Nauru detect cirrus approximately 70% of the time with fewer occurrences of overcast conditions. Distributions of CALIPSO cloud-top heights over this region suggest that the persistent cirrus generated by CAM3 is not physically based.

REFERENCES

Anderberg, M. R., 1973: *Cluster Analysis for Applications*. Academic Press, 359 pp.

Boville, B. A., P. J. Rasch, J. J. Hack, and J. R. McCaa, 2006: Representation of clouds and precipitation processes in the Community Atmosphere Model version 3 (CAM3). *J. Climate,* **19,** 2184–2198.

Briegleb, B. P., 1992: Delta-Eddington approximation for solar radiation in the NCAR Community Climate Model. *J. Geophys. Res.,* **97,** 7603–7612.

Byrne, R. N., R. C. J. Somerville, and B. Subasilar, 1996: Broken-cloud enhancement of solar radiation absorption. *J. Atmos. Sci.,* **53,** 878–886.

Cahalan, R. F., 1994: Bounded cascade clouds albedo and effective thickness. *Nonlinear Processes Geophys.,* **1,** 156–167.

Clothiaux, E. E., T. P. Ackerman, G. G. Mace, K. P. Moran, R. T. Marchand, M. A. Miller, and B. E. Martner, 2000: Objective determination of cloud heights and radar reflectivities using a combination of active remote sensors at the ARM CART sites. *J. Appl. Meteor.,* **39,** 645–665.

Collins, W. D., and Coauthors, 2006: The formulation and atmospheric simulation of the Community Atmospheric Model Version 3 (CAM3). *J. Climate,* **19,** 2144–2161.

Foster, M. J., and D. E. Veron, 2008: Evaluating the stochastic approach to shortwave radiative transfer in the tropical western Pacific. *J. Geophys. Res.,* **113,** D22205, doi:10.1029/2007JD009581.

Gordon, N. D., J. R. Norris, C. P. Weaver, and S. A. Klein, 2005: Cluster analysis of cloud regimes and characteristic dynamics of midlatitude synoptic systems in observations and a model. *J. Geophys. Res.,* **110,** D15S17, doi:10.1029/2004JD005027.

Jakob, C., and G. Tselioudis, 2003: Objective identification of cloud regimes in the tropical western Pacific. *Geophys. Res. Lett.,* **30,** 2082, doi:10.1029/2003GL018367.

——, ——, and T. Hume, 2005: The radiative, cloud, and thermodynamic properties of the major tropical western Pacific cloud regimes. *J. Climate,* **18,** 1203–1215.

Kiehl, J. T., J. J. Hack, G. B. Bonan, B. A. Boville, D. L. Williamson, and P. J. Rasch, 1998: The National Center for Atmospheric Research Community Climate Model: CCM3. *J. Climate,* **11,** 1131–1149.

Kollias, P., E. E. Clothiaux, B. A. Albrecht, M. A. Miller, K. P. Moran, and K. L. Johnson, 2005: The Atmospheric Radiation Measurement program cloud profiling radars: An evaluation of signal processing and sampling strategies. *J. Atmos. Oceanic Technol.,* **22,** 930–948.

Lane, D. E., K. Goris, and R. C. J. Somerville, 2002: Radiative transfer through broken cloud fields: Observations and model validation. *J. Climate,* **15,** 2921–2933.

Lane-Veron, D. E., and R. C. J. Somerville, 2004: Stochastic theory of radiative transfer through generalized cloud fields. *J. Geophys. Res.,* **109,** D18113, doi:10.1029/2004JD004524.

Long, C. N., T. P. Ackerman, K. L. Gaustad, and J. N. S. Cole, 2006: Estimation of fractional sky cover from broadband shortwave radiometer measurements. *J. Geophys. Res.,* **111,** D11204, doi:10.1029/2005JD006475.

Malvagi, F., R. N. Byrne, G. C. Pomraning, and R. C. J. Somerville, 1993: Stochastic radiative transfer in a partially cloudy atmosphere. *J. Atmos. Sci.,* **50,** 2146–2158.

McClatchey, R. A., R. W. Fenn, J. E. A. Selby, F. E. Volz, and J. S. Garing, 1972: Optical properties of the atmosphere. 3rd ed. Environmental Research Paper 411, Air Force Cambridge Research Laboratories, 113 pp.

Naud, C. M., A. Del Genio, G. G. Mace, S. Benson, E. E. Clothiaux, and P. Kollias, 2008: Impact of dynamics and atmospheric state on cloud vertical overlap. *J. Climate,* **21,** 1758–1770, doi:10.1175/2007JCLI1828.1.

Pincus, R., and S. A. Klein, 2000: Unresolved spatial variability and microphysical process rates in large-scale models. *J. Geophys. Res.,* **105,** 27 059–27 065.

Potter, P. L., and R. D. Cess, 2004: Testing the impact of clouds on the radiation budgets of 19 atmospheric general circulation models. *J. Geophys. Res.,* **109,** D02106, doi:10.1029/2003JD004018.

Randall, D. A., and Coauthors, 2007: Climate models and their evaluation. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 589–662.

Redelsperger, J. L., D. B. Parsons, and F. Guichard, 2002: Recovery processes and factors limiting cloud-top height following the arrival of a dry intrusion observed during TOGA COARE. *J. Atmos. Sci.,* **59,** 2438–2457.

Rossow, W. B., and R. A. Schiffer, 1999: Advances in understanding clouds from ISCCP. *Bull. Amer. Meteor. Soc.,* **80,** 2261–2287.

——, G. Tselioudis, A. Polak, and C. Jakob, 2005: Tropical climate described as a distribution of weather states indicated by distinct mesoscale cloud property mixtures. *Geophys. Res. Lett.,* **32,** L21812, doi:10.1029/2005GL024584.

Stokes, G. M., and S. E. Schwartz, 1994: The Atmospheric Radiation Measurement (ARM) Program: Programmatic background and design of the Cloud and Radiation Test Bed. *Bull. Amer. Meteor. Soc.,* **75,** 1201–1221.

Williams, K. D., C. A. Senior, A. Slingo, and J. F. B. Mitchell, 2005: Towards evaluating cloud response to climate change using clustering technique identification of cloud regimes. *Climate Dyn.,* **24,** 701–719.

Wiscombe, W. J., and J. W. Evans, 1977: Exponential-sum fitting of radiative transmission functions. *J. Comput. Phys.,* **24,** 416–444.

Zhang, G. J., 2009: Effects of entrainment on convective available potential energy and closure assumptions in convection parameterization. *J. Geophys. Res.,* **114,** D07109, doi:10.1029/2008JD010976.