Cloud-Precipitation Hybrid Regimes and their Projection onto IMERG Precipitation Data

Daeho Jin[a,b], Lazaros Oreopoulos[b], Dongmin Lee[c,b], Jackson Tan[a,b], Nayeong Cho[a,b]

[a] *Universities Space Research Association, Columbia, MD, USA*

[b] *NASA Goddard Space Flight Center, Greenbelt, MD, USA*

[c] *Morgan State University, Baltimore MD, USA*

Corresponding author: Daeho Jin, Daeho.Jin@NASA.gov

ABSTRACT

We extend and enhance the concept of the Cloud Regimes (CRs) developed from two-dimensional joint histograms of cloud optical thickness and cloud top pressure from the Moderate Resolution Imaging Spectroradiometer (MODIS), by adding precipitation information in order to better understand cloud-precipitation relationships. Taking advantage of the high-resolution Integrated Multi-satellitE Retrievals for GPM (IMERG) precipitation dataset, cloud-precipitation "hybrid" regimes are derived by implementing the *k*-means clustering algorithm with advanced initialization and objective measures to determine the most optimal clusters. By expressing precipitation rates within 1-degree grid cell as histograms and making choices on the relative weight of cloud and precipitation, we could obtain several editions of hybrid cloud-precipitation regimes (CPRs), and examine their characteristics.

In the deep tropics, when precipitation is weighted weakly, the cloud part of the hybrid centroids resembles the centroid of cloud-only regimes, but still tightens the cloud-precipitation relationship by decreasing the precipitation variability of each regime. As precipitation weight progressively increases, the shape of the cloudy part of the hybrid centroids becomes blunter, while the precipitation part of the centroids sharpens. In the case where cloud and precipitation are weighted equally, the CPRs representing high clouds with intermediate to heavy precipitation exhibit distinct features in the precipitation parts of the centroids, which allows us to project them onto the 30-minly IMERG domain. Such a projection can be used to overcome the temporal sparseness of MODIS cloud observations, which leads to great application potential for various convection-focused studies, including diurnal cycle analysis.

SIGNIFICANCE STATEMENT

Clouds and precipitation are related closely, but in complex ways. In this work we attempt to provide a classification of daytime cloud-precipitation co-occurrence and co-variability, with emphasis in tropical regions. We achieve such a classification using *k*-means clustering algorithm applied on cloud property and precipitation intensity histograms which yields “hybrid” clusters. These hybrid clusters reveal more detailed features of coincident daytime cloud and precipitation systems than clusters where clouds and precipitation are treated separately. Moreover, the realization that precipitation features associated with high and thick clouds have very distinct patterns enables hybrid cluster prediction based solely on precipitation information, which has the important implication that rarer cloud observations can be extended to the more frequent (including nighttime) precipitation domain.

1. Introduction

In many applications, a variable or combinations of variables that co-vary need to be sorted into groups whose members are considered similar. One option to accomplish the grouping is clustering analysis, a discipline of unsupervised machine learning. Among various algorithms that perform clustering, “*k*-means” is one of the most popular options in geophysical sciences due to its simplicity and efficiency in processing large volumes of data. Examples of recent studies where *k*-means clustering is used are the grouping of precipitation patterns to identify the South Pacific convergence zone (SPCZ; Pike and Lintner 2020), analysis of geopotential height data to identify weather patterns for subseasonal forecast (Robertson et al. 2020), and finding dominant modes in sea surface temperature data to identify two kinds of the North Pacific Meridional Mode (NPMM; Zhao et al. 2020), etc.

k-means clustering has also been applied in the last two decades to cloud grouping. Based on the gridded Level-3 2D-joint histogram of cloud top height (CTP) and cloud optical thickness (COT) retrieved from the International Satellite Cloud Climatology Project (ISCCP), dominant mixtures of clouds, later called “weather states”, were identified in the tropical western Pacific (Jakob and Tselioudis 2003), the deep tropics from 15°S to 15°N (Rossow et al. 2005), extended tropics and mid-latitudes (Oreopoulos and Rossow 2011), and globally (Tselioudis et al. 2013). The same methodology was extended to similar 2D-joint histogram of CTP and COT retrieved from the Moderate Resolution Imaging Spectroradiometer (MODIS), to obtain cloud groups referred to as “cloud regimes (CRs)” (Oreopoulos et al. 2014, 2016; Jin et al. 2020).

Clouds and precipitation are closely related to each other, albeit in complex ways, so the effort of Luo et al. (2017) to perform joint clustering of cloud and precipitation information came as a natural progression in expanding clustering applications. Using the Tropical Rainfall Measuring Mission (TRMM) Ku-band Precipitation Radar and the CloudSat W-band Cloud-Profiling Radar, they first built 2D joint histogram of height and radar reflectivity (a.k.a. H-dBZ histogram) for rather sparse coincident observations, on which they then performed *k*-means clustering analysis. They also tested another expanded version of joint histograms where CALIOP lidar products were added to capture optically thinner clouds, and obtained a larger number of meaningful joint cloud-precipitation groups. This pioneering work opened new pathways to group microphysical properties of hydrometeors by regimes with data that can also resolve vertical structures. Combined cloud-precipitation analysis, but without joint clustering, have also been performed within the framework of weather states or CRs. But in these studies precipitation variability was a dependent variable sorted for specific kinds of cloud mixtures as represented by the weather states or CRs (e.g., Lee et al. 2013; Rossow et al. 2013; Tan et al. 2015; Tan and Oreopoulos 2019).

Recently, precipitation datasets have been greatly improved in terms of quality and spatiotemporal coverage due to advances in algorithms such as the Integrated Multi-satellitE Retrievals for GPM (IMERG) product providing precipitation rates at 0.1° every 30 minutes. The combination of the IMERG precipitation and MODIS cloud products provides an unprecedented opportunity to examine cloud-precipitation joint variability not possible with previous generation datasets. We thus return in this study to the joint clustering concepts of Luo et al. (2017) aiming once again to identify dominant mixtures of cloud and precipitation patterns. While our data, Level-3 cloud and precipitation products, do not have the capability to resolve vertical variability, we can perform joint clustering with much wider coverage compared to the availability of the Level-2 reflectivity and backscatter. It turns out that the existence of a tight coupling between clouds and precipitation in some of our "hybrid" regimes allows us to take advantage of the higher temporal resolution of IMERG to greatly expand the rarer cloud information suffering the limitations of sun-synchronous satellite observations. We discuss this further in section 5 of this paper.

The remainder of the paper provides the details of data and *k*-means clustering methodology (sections 2 and 3), formally presents the cloud-precipitation hybrid regimes and discusses their characteristics in section 4. Section 6 summarizes the study and discusses possible applications of the new dataset.

2. Data

a. MODIS cloud data

Cloud properties are retrieved from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard the Terra and Aqua satellites. The MODIS cloud product (MOD08_D3 and MYD08_D3; King et al. 2003; Platnick et al. 2003, 2017b) provides Level-3 cloud observations at daily time scales with 1°×1° horizontal resolution. Among various variables in Level-3 products, we specifically use the ISCCP-like 2D joint histogram of cloud optical thickness (COT) and cloud top pressure (CTP). The histogram is composed of cloud fraction (CF) values along 7 classes of CTP and 6 classes of COT (for a total 42 histogram bins), thus providing information about pixel-level cloud variability at the 1° scale. Since the recent major version of the MODIS atmospheric datasets, known as "Collection 6" (Platnick et al. 2017a), a separate histogram for "partially cloudy" (PCL) pixels is provided, flagged as such by the so-called "clear-sky restoral" algorithm (Pincus et al. 2012; Zhang and Platnick 2011). The 2D joint histograms used in this study include the sum of the PCL and nominal joint histograms, as in Jin et al. (2018, 2020). The update from Collection 6 to Collection 6.1 used here is relatively minor (Platnick et al. 2018).

b. IMERG precipitation data

The Integrated Multi-satellitE Retrievals for GPM (IMERG) data provides seamless precipitation estimates at a 0.1° grid every half hour by unifying observations from a network of partner satellites in the Global Precipitation Measurement (GPM) constellation (Huffman et al. 2019a,b; Tan et al. 2019a). The most recent major update version V06 extends spatial coverage to the entire globe (except over frozen surfaces at high latitudes) and the temporal period back to June 2000 (the pre-GPM era of the Tropical Rainfall Measuring Mission – TRMM) onwards. The IMERG product comprises three runs (Early, Late, and Final), of which we use the Final run which is of best quality. We note that for this study we limit the data period for both cloud and precipitation from June 2014 to May 2019 in order to avoid potential risk of inconsistencies between the GPM and TRMM satellites.

c. Spatio-temporal matching between MODIS and IMERG data

The MODIS Level-3 gridded data is provided daily for each of the Terra and Aqua satellites. Observations on swath paths for a large portion of the globe take place at similar local time but varying Coordinated Universal Time (UTC). In order to temporally match MODIS cloud data and IMERG precipitation data which are segmented by UTC, we calculate the UTC of each MODIS grid cell using the assigned mean solar zenith angle in the Level-3 product, and then select the temporally closest IMERG data point. The details of this temporal matching method are described in Jin et al. (2018), and although in that paper the precipitation data was the TRMM Multi-satellite Precipitation Analysis (TMPA), the principle of the method is the same. Spatial matching is much easier: for each 1°×1° grid cell of MODIS clouds, the one hundred enclosed precipitation rates of 0.1°×0.1° resolution are assigned. Hence, we ultimately obtain 42 values of binned cloud fraction and 100 values of precipitation rates for 5 years, for each 1° grid cell that has Terra and Aqua cloud observations.

3. Application of *k*-means clustering

In this study we build our basis dataset of hybrid regimes using *k*-means. The *k*-means clustering algorithm (Anderberg 1973; MacQueen 1967) is one of the most popular unsupervised clustering algorithms. This simple algorithm can handle very large data volumes efficiently, hence it is widespread in various studies implementing clustering of geophysical variables, as noted in the Introduction. The underlying principle of the algorithm is that for input data consisting of $m\_samples \times n\_features$, feature distances are calculated between each sample and given centroids, and each sample is assigned to the centroid corresponding to the smallest distance. The mean of newly assigned samples becomes the new centroid, and the assignment is repeated until the new centroids are (nearly) identical to the centroids of the previous iteration. Eventually all data are assigned to the group with the most similar members, which minimizes the total Mean Squared Error of the grouped data. In this study, we set the threshold of centroid movement to 1.0e-6, which yields convergence in a few hundred iterations (we set no limit on the total number of iterations).

a. Preparing input data: how to balance between cloud and precipitation data

Previously, Jin et al. (2020) derived tropical cloud regimes (TCRs) using MODIS cloud 2D joint histogram data. Since the cloud histogram bin values ranged from 0 to 1 by definition, TCRs could be obtained from the *k*-means clustering algorithm without any normalization process. In order to derive hybrid regimes, the range of values of IMERG precipitation rates must be equivalent to the cloud histogram data. This was easily accomplished by transforming precipitation rates to normalized histogram bin values, similarly to the cloud data.

In transforming precipitation data into a histogram, one issue to consider is how to choose the number of bins. Too small a number of bins results in excessive smoothing, which makes notable precipitation patterns indistinguishable. Conversely, too large a number of bins increases noise and prevents us from obtaining meaningful clusters. Since it is known that similar clouds can have varying precipitation rates (e.g., Jin et al. 2018, 2020), we gravitated towards a rather coarser binning of the precipitation histograms. After some testing, we settled on an approximately logarithmically-spaced 6-bin precipitation histogram with bin boundaries at 0.03, 0.1, 0.33, 1, 3.33, 10, 999mm/h. We note that these histogram bin boundaries exclude no-rain counts for consistency with the cloud histogram, and also very small precipitation rates below 0.03mm/h.

The second issue we had to address was the relative weight between cloud and precipitation when applying the clustering algorithm. If we combine cloud and precipitation histograms without any weighted treatment, the relative importance of cloud compared to precipitation in the *k*-means clustering calculation is 7 to 1 because the cloud histogram consists of 42 bins while the precipitation histogram consists of 6 bins (for a total of 48 bins). With Euclidean distance adopted as the measure to assign data to one of centroids in the *k*-means algorithm, the number of bins translates linearly to relative importance. In this sense, it is possible to make both cloud and precipitation equally important by combining the 42-bin cloud histogram with the precipitation histogram replicated seven times for a total of 84 bins that come from two equal 42-bin contributions from the cloud and precipitation side. In this study, a total of 3 different versions of weights for cloud and precipitation were tested, namely 7:1, 7:3, 7:7. Only the 7:1 and 7:7 versions will be shown in the manuscript itself, with the 7:3 version shown in the Supplementary Material Part A. We also derive a new set of cloud-only regimes to be used as a reference by following the same procedures, described in the next subsection, as for the hybrid regimes.

In terms of regional coverage, we performed the *k*-means algorithm separately for the deep tropics (15°S-15°N) and for much larger portion of the globe that expands to midlatitudes (50°S-50°N). The two domains for 5 years for both Terra and Aqua data result in populations of ~34 million and ~116 million data points once missing values are excluded. In this study, we focus on the deep tropical results only, while the near-global results are shown in the Supplementary Material Part B.

b. Initializing with k-means++ algorithm

The *k*-means clustering algorithm is, by definition, deterministic to the initial values, namely the centroids chosen initially. If more than one of the initial centroids are chosen from potentially the same cluster group (i.e., they are similar to each other), the end result of the clustering may not be optimal. To reduce the probability of this happening, and to improve the performance of the *k*-means clustering, a "*k*-means++" algorithm was developed for smarter initialization (Arthur and Vassilvitskii 2007). The *k*-means++ employs a weighted random selection method, where the distance from a pre-selected initial centroid is set as the weight of the data member. If two or more initial centroids are already selected, the minimum distance is selected as the weight. This process ensures that the farthest (largest Euclidean distance) data member from pre-selected centroid(s) has the highest possibility to be chosen, thus ultimately making the initial centroids well-separated from each other. We employ the *k*-means++ algorithm to initialize the *k*-means clustering scheme with 50 different sets of initial centroids (i.e., 50 realizations) for each candidate number *k* of clusters, in order to potentially achieve the best *k*-means clustering results (see next subsection).

c. Criteria for choosing the optimal number of clusters

The *k*-means clustering algorithm requires the number of clusters, *k*, as a preset to be decided by the user. By the nature of *k*-means clustering, a larger number of clusters always decreases the magnitude of "error", measured by the "Within-Cluster(intra-cluster) Variance (WCV)", since the larger *k* the less diverse the members of a group are. At the same time, a large *k* has the undesirable effect of diminishing the level of data compression, which is another way of saying that too many clusters make the grouping less practical and useful. An appropriate value of *k* therefore represents a compromise between the amount of error and the level of compression.

Several methods exist to determine the optimal value of *k*. One of the most basic and intuitive methods is the so-called “elbow” method. By observing the percentage of explained variance as a function of the number of clusters, the value of *k* is selected when the marginal gain of explained variance is small with another cluster added. An issue with this method is that characterizing the gain as marginal is subjective and ambiguous. In many cases the “elbow” point is not obvious, which makes this method unreliable (e.g., Ketchen and Shook 1996).

The Calinski-Harabasz criterion (CHC; Caliński and Harabasz 1974) is another popular method to determine the most optimal *k*. The basic idea of CHC is to maximize the overall “between-cluster(inter-cluster) variance (BCV),” which indicates maximum separation among clusters, while minimizing the error expressed by WCV. A CHC metric is defined as

$$CHC_k = \frac{BCV}{(k-1)} \Big/ \frac{WCV}{(N-k)}$$

where *N* is the total number of data points, and *k* is the number of clusters. The BCV and WCV are defined as

$$BCV = \sum_{i=1}^{k} n_i \|\mu_i - \mu\|^2$$

$$WCV = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

where n_i is the number of data points in cluster *i* (C_i), μ_i is the mean of data points in cluster *i* (a.k.a. centroid), and μ is the overall mean of all data points. The value of *k* yielding the maximum CHC represents the best choice for cluster number *k*.

The Davies-Bouldin criterion (DBC; Davies and Bouldin 1979) also pursues the maximum separation of clusters with minimum errors in the clusters as CHC, but uses different measures. The DBC metric is defined as

$$DBC_k = \frac{1}{k}\sum_{i=1}^{k} max_{j\neq i}\{R_{i,j}\}$$

$$R_{i,j} = \frac{S_i + S_j}{D_{i,j}}$$

where $R_{i,j}$ is the ratio of within-cluster scatter of the i^{th} and j^{th} clusters (S_i, S_j) to the separation between the i^{th} and j^{th} clusters ($D_{i,j}$). S_i and $D_{i,j}$ are defined as

$$S_i = \left(\frac{1}{n_i}\sum_{x\in C_i}\|x-\mu_i\|^2\right)^{1/2}$$

$$D_{i,j} = \left\|\mu_i - \mu_j\right\|$$

Here, the within-cluster scatter (S_i) represents average distance between each data point and centroid, and the separation measure $D_{i,j}$ is the Euclidean distance between two centroids. For a given k, by choosing the maximum ratio for each cluster, DBC measures the worst-case scenario for each cluster. The minimum value of DBC represents therefore the most optimal number of clusters.

Figure 1 shows the dependence on k of these criteria in the case of 6 precipitation histogram bins with weight number 1 (i.e., 48-element combined array, referred to as “Cld42+Pr6x1”). The left panel (Fig. 1a) shows maximum BCV and minimum WCV as a function of k. The elbow method can be applied to both BCV and WCV. (We note that, because the explained variance is defined as BCV divided by total variance and total variance is a fixed number, it is essentially the same to apply the elbow method to either explained variance or BCV.) However, both BCV and WCV change smoothly as k increases, and it is hard to find an “elbow” in the above figure. In the right panel (Fig. 1b), DBC clearly indicates that 16 is the optimal k while CHC monotonically decreases as k increases. The CHC metric heavily depends on the total population of data points (N) by definition, and in the case of huge N ($N\approx$ 34M for our deep tropics domain), variability of CHC is dominated by the term, $N/(k\text{-}1)$, which results in monotonical decrease with k in a reasonable range. Taking all these into account, DBC is chosen as the primary criterion for selecting the optimal number of clusters, and the trial producing the globally minimum DBC value determines the final set of regimes composed of k centroids. Table 1 shows the values of k that came out of this procedure for the four (i.e., including zero) precipitation weights, for both the narrow and extended domains in latitudes. Figures similar to Fig. 1 for the other cases are shown in the Supplementary Materials.

4. Details of tropical hybrid regimes

a. Cloud-only regimes

A set of cloud-only regime is derived as the baseline with which the cloud-precipitation hybrid regimes can be compared to. Jin et al. (2020) previously derived a set of cloud regimes with k=10 in the same deep tropics domain, with the last regime being decomposed to 4 sub regimes, for a total of 13 regimes, using the concept of "nested clustering" (Luo et al. 2017; Mason et al. 2014; Oreopoulos et al. 2016). Here, the data period is shortened from 14 years to 5 years to accommodate the availability of precipitation observations, and DBC is employed to select the final set of regimes without invoking nested clustering.

Figure 2 shows that the (deep) tropics cloud-only regime (TCR; please note that for economy we drop the tropical "T" designation in the following figures) set is composed of 8 high-cloud regimes, 5 low-cloud regimes, and one mixed semi-clear regime (TCR14). Each TCR, except TCR14, has a unique distinct peak of bin cloud fraction. This is a notable difference from the previous TCR set reported by Jin et al. (2020), particularly for high clouds with relatively large optical thickness. Figure 1 in Jin et al. (2020) showed three TCRs relevant to convective activity, with peaks of similarly large cloud fraction values at two neighboring histogram bins. These blunt peaks seem to have now split into two TCRs. For example, the old TCR1 had the largest cloud fraction bin values across the cirrostratus (Cs) and cumulonimbus (Cb) bins, according to the traditional ISCCP cloud types (Rossow and Schiffer 1999), and these have now split into peaks that occur in TCR1 and TCR2. By comparing the assignments of each grid cell to old and new TCRs, we confirm that the most grid cells previously assigned to old TCR1, TCR2, and TCR3 in Jin et al. (2020) are now assigned to TCR1 to TCR6. Among them, the first three TCRs dominate precipitation, and TCR1 having the optically thickest and highest cloud dwarfs the other regimes in mean precipitation rate.

b. Hybrid regime2 with precipitation weight of 1 (Cld42+Pr6x1)

We first introduce the tropical cloud-precipitation (hybrid) regime (TCPR) set that corresponds to the precipitation weight of 1 (i.e., cloud-to-precipitation weight ratio is 7:1 with 48-element array; Cld42+Pr6x1). By adding precipitation information this way, the optimal number of clusters according to the DBC increases from 14 to 16 in our tropical domain (Table 1 and Fig. 3). This TCPR set is composed of 9 high cloud regimes, 5 low cloud regimes, and 2 mixed regimes (including a semi-clear regime, TCPR16). A notable difference in centroids when rainfall information added is the newly occurring TCPR10. This regime represents high and low mixed clouds with intermediate cloud fraction and substantial precipitation. In order to investigate the origin of this version of TCPR10, we introduce a regime coincidence distribution matrix (Fig. 4) showing the RFO of new regimes (i.e., Cld42+Pr6x1; x-axis) for the grid cells assigned to one of the cloud-only regimes (y-axis). This graphical matrix indicates that grid cells assigned to TCPR10 belonged previously to various TCRs (e.g., TCR3, 5, 7, 10, 14, etc.). In terms of population, the biggest contributor is TCR14 which is semi-clear regime with RFO 38% (Fig. 2). Because of the split of TCR14 due to the addition of rainfall information, the similar semi-clear hybrid regime (TCPR16) has now a lower RFO value (32.8% in Pr6x1 TCPR16 vs. 37.9% for TCR14) and a lower cloud fraction (26% vs. 32%).

The other contributor to the increased k from the cloud-only regimes to hybrid regimes is the split of TCR8 into TCPR8 and TCPR9. TCR8 in Fig. 2 represents a cirrus (Ci)-dominant regime with a cloud fraction peak in the bin of highest cloud top (lowest CTP) and smallest optical thickness; it is now split into two versions of Ci-dominated regimes with total cloud fractions of 58% (TCPR8) and 78% (TCPR9). While neither TCPR8 nor TCPR9 seem to be producing substantial rainfall, the precipitation histogram component of the centroid shows that TCPR8 has a slightly elevated chance of intermediate intensity precipitation.

It is also worth noting that significant fractions of grid cells occupied by TCR3 are now assigned to TCPR5 in addition to TCPR3. TCPR3 and TCPR5 show clearly different precipitation characteristics: the estimated average precipitation rate of TCPR3 is 1.2mm/h with the peak of precipitation histogram around 1mm/h while the average rate of TCPR5 is 0.2mm/h. A possible interpretation is that TCR3 has grid cells of similar clouds with varying precipitation intensities from light to intermediate, and grid cells of lighter precipitation are shifted to TCPR5 by the addition of precipitation information. Similar phenomena of lighter rain grid cells shifted to other hybrid regimes are also found for TCR1, TCR5, and TCR6 indicating that within-regime precipitation variability decreases in the hybrid regimes because outliers with weak precipitation in cloud-only regimes are now removed. On the other hand, regimes dominated by low clouds show great consistency between the cloud-only and hybrid regime sets because there are barely any precipitation features that would make them distinguishable.

c. Hybrid regimes with precipitation weight of 7 (Cld42+Pr6x7)

As the relative weight of precipitation increases from 1 to 3, the patterns of the cloud joint histogram component of the centroids lose peak sharpness, and some regimes even show blunt peaks across two adjacent levels of CTP (see Supplementary Material Part A). As the relative weight of precipitation further increases to 7, namely when cloud and precipitation histograms matter equally in the (84-element) combined arrays subjected to *k*-means clustering, the patterns of the mean joint cloud histogram exhibit even blunter peaks, and some hybrid regimes now share quite similar cloud patterns (e.g., TCPR3 and TCPR4; TCPR7 and TCPR8 and TCPR9 in Fig. 5). This suggests that precipitation rather than cloud has now a greater impact in determining the assignment to certain TCPRs, and a previous regime of the no or small precipitation weight set can be split into multiple regimes depending on the shape of the precipitation histogram. Indeed, the optimal number of clusters in the Cld42+Pr6x7 case (a.k.a. equal-weight set) increases to 19, with 13 high cloud regimes, 4 low cloud regimes and 2 mixed regimes (including the semi-clear regime).

A similar matrix of regime coincidence distribution between precipitation weight number 1 and 7 is displayed in Fig. 6. In the Cld42+Pr6x1 set, TCPR1, TCPR2 and TCPR3 represents high and thick clouds producing intermediate to heavy precipitation. All these three TCPRs are now split into 3 or more TCPRs in the equal-weight set because of the increased impact of precipitation on the clustering. As a result, centroids of equal-weight set show distinct patterns in the precipitation histogram component of the centroid, something that can be interpreted as decreased variability in precipitation intensity and increased variability in cloud type mixtures in the grid cells belonging to a specific TCPR of the equal-weight set. Also noteworthy is that TCPR10 of Cld42+Pr6x1 which was diagnosed as representing mixed clouds with intermediate precipitations is now split into 4 different

TCPRs. In terms of cloud histogram pattern, TCPR14 of Fig. 5 shares some similarity with TCPR10 of Fig. 3, but TCPR14 of the equal-weight set has notably smaller high-cloud fractions and intermediate-precipitation fractions. The decomposition of TCPR10 in Cld42+Pr6x1 is a major contributor to the increased number of clusters from 16 to 19.

In summary, we find that the information added by precipitation helps to also distinguish clouds with a greater degree of detail in terms of cloud-precipitation relationship. In the set where the added precipitation information matters the least, namely the 7:1 weight ratio (Cld42+Pr6x1), the cloud histogram patterns are mostly consistent with the cloud-only regimes. Still, the added precipitation information rearranges some outlier grid cells in cloud-only regimes (in terms of precipitation properties), thus resulting in tighter relationships between cloud and precipitation in the new regimes. The enhanced weight of precipitation obviously decreases the influence of cloud patterns in the resulting centroids, even to the degree where similar cloud histogram patterns (albeit with distinct precipitation histogram patterns) appear in the equal-weight set. These cloud and precipitation pattern changes occur mostly in regimes dominated by high-clouds; regimes dominated by low clouds are not changing much by increasing the precipitation weight indicating the lack of diversity in precipitation properties, at least according to IMERG.

5. Projection onto IMERG domain

a. Can cloud be predicted from precipitation?

Cloud and precipitation are closely related, but at the same time there is significant precipitation variability within similar clouds, and vice versa. In the previous section, we showed two sets of tropical cloud-precipitation hybrid regimes, representing the dominant mixtures of specific cloud types and corresponding precipitation intensities (other variants of relative weights and an extension that includes extratropics are shown in the Supplemental Materials). In this section, we examine the feasibility of "predicting" clouds from solely precipitation information using these hybrid regimes. The reason we want to predict clouds is because the cloud observations suffer from substantial amounts of missing grid cells due to the swath width of the MODIS granules, and are much sparser temporally compared to the IMERG precipitation dataset. An extended dataset of cloud information with higher temporal resolution could be useful for various research endeavors.

The availability of cloud-precipitation hybrid regimes simplifies a potential cloud prediction scheme because clouds in a grid cell are represented by the limited number of classes (regimes) derived from the clustering analysis. (Additional information about the clouds besides what hybrid regime they belong would obviously not be available.) Hence, the problem at hand is predicting one of the hybrid regimes based on only the precipitation information of a grid cell. The simplest way to assign a hybrid regime to grid cell at a time when no cloud information is available is to adopt the Euclidean distance criterion used in the *k*-mean clustering, but now applied only on the observed IMERG precipitation histogram and the precipitation component of the hybrid regime centroid. Of course, this assignment by precipitation is only possible when a reasonable amount of precipitation is detected; identification of hybrid regime occurrence in a grid cell where barely any rain occurs is impossible.

The performance of hybrid regime prediction by matching observed and centroid precipitation histograms is summarized in Fig. 7 for the case of the equal-weight set (Cld42+Pr6x7) in the extended tropical domain of 20°S to 20°N. Figure 7 is a Fig.4-like regime coincidence distribution matrix between original TCPRs (y-axis) observed at the time of Terra and Aqua daytime overpasses and predicted TCPRs by precipitation-only (x-axis) for the same grid cells. Among the 19 regimes, those with precipitation fraction (= sum of 6 bins of precipitation histogram) below 10% are merged into the "Others" class. Overall, the prediction results are quite impressive; among regimes having significant amounts of precipitation, five regimes (CPR1, 2, 4, 8, 9) have precipitation-based prediction accuracy of TCPR occurrence above 95%. Furthermore, the accuracy of TCPR3 and TCPR7 prediction is also quite high, over 90%. This means that the precipitation signatures of members belonging to these hybrid regimes are unique enough to allow them to be differentiated from members of other regimes. These regimes commonly have precipitation fractions above 50%. While for TCPR5, exhibiting only 20% prediction accuracy, the estimated mean precipitation is greater than that of TCPR9, precipitation fraction is just 23%, less than half of TCPR9's (Fig. 5). A small total precipitation fraction usually means that histogram bin values are also small, which makes them hard to be distinguished from other regimes under our adopted Euclidean distance criterion. In addition, we also examined the accuracies geographically (by longitudes), and found that prediction accuracies are quite stable regardless of longitudes with only small drops of accuracy in the central Africa and South America in the case of TCPR1 and TCPR7 (see Supplementary Material Part A).

The equal-weight set shows that the regimes having intermediate-to-heavy precipitation intensity can by predicted well by the precipitation-only histogram constructed by the 0.1° IMERG data, a result likely due to the significant impact of precipitation on the clustering process. We also tested the case of small precipitation weight, and as expected, the prediction accuracy was markedly lower, as shown in Fig. 8. In the case of Cld42+Pr6x1, 7 regimes pass the criterion of precipitation fraction above 10% among the 16 regimes. The highest accuracy, 81%, is achieved by TCPR1 which has the heaviest precipitation and thickest clouds representing a group of convective cores (Fig. 3). The second highest prediction accuracy, 50%, is achieved by TCPR10 (mixed cloud types and a large fraction of light rain), while all other prediction skills are below 50%. In the case of TCPR2 and TCPR3 both of which have intermediate precipitation intensity, precipitation histogram patterns are too similar (Fig. 3) for them to be separable in the regime prediction. Still, Fig. 8 suggests a hint of different precipitation characteristics between TCPR2 and TCPR3, where the light precipitations tails of TCPR2's rainfall distribution gives rise to TCPR10 assignment for 14% of the grid cells, while TCPR3 being biased towards heavy precipitation results in assignment of 13% of the grid cells to TCPR1.

To summarize, we demonstrated that we can predict cloud patterns through the prediction of hybrid regimes from precipitation-only information when using the set of hybrid regimes derived with equal weighting between cloud and precipitation (Cld42+Pr6x7). A total of 7 hybrid regimes can be predicted highly accurately when their precipitation features include intermediate to heavy rainfall intensity and their cloudiness corresponds to high-thick cloud patterns. In practical terms this means that through the process of assigning regimes by precipitation histogram Euclidean distance, we can transform the 30-minute full tropical coverage IMERG data into occurrence maps of these 7 regimes at 1-degree resolution and at the same 30-minute temporal resolution, i.e. we have achieved a projection of TCPRs onto the IMERG domain. In the following subsection, we present an application example of this newly built hybrid regime occurrence maps.

b. Analysis example: Diurnal cycle of hybrid regimes

Due to the reliance of cloud optical thickness retrievals on the availability of solar insolation, 2D joint histogram data of cloud is available once daily for each of Terra and Aqua, at around 10:30am and 1:30pm local solar time (LST), respectively. Hence, even a combined analysis of Terra and Aqua can provide only limited information on cloud variability around noon in LST. The occurrence map of hybrid regimes projected onto IMERG domain according to our method described in the previous subsection radically improves the temporal resolution (30-min), thus enabling examination of the diurnal cycle of the hybrid regimes for which we have good prediction capability, based on the assumption that nighttime cloud-precipitation relationship remains the same as in daytime. Figures 9 and 10 show the RFO of TCPR1 and TCPR2 of the Cld42+Pr6x7 set in the longitude-LST phase space, respectively. We note that LST is calculated by adding the regionally-dependent factor, longitude×(24/360) to UTC as in Tan et al. (2019b).

TCPR1 of the Cld42+Pr6x7 set represents deep convective cores with the heaviest precipitation. Previously, Jin et al. (2018, 2020) showed that the regime corresponding to the heaviest precipitation most frequently occurs in the tropical warm pool oceans. Figure 9 is consistent with the previous studies, and shows the highest RFO in the east and west of the Maritime Continent. Moreover, the temporal evolution indicates that the most active hour of TCPR1 occurrence is in the early morning, 2am to 8am in this region, consistent with Fig. 11 in Yang and Smith (2006), but deviating from the findings of Kikuchi and Wang (2008) who noted oceanic peak between 6am to 9am. Other than the warm pool region, TCPR1 also notably occurs in the Amazon basin, and is slightly more active in the early morning than other local times, which is consistent with the precipitation diurnal cycle driven by dynamical processes (Vernekar et al. 2003). Regardless of the longitude, a hint of local RFO minimum appears just before noon, a feature that actually becomes clearer when examining TCPR2 in Fig. 10.

TCPR2 of the Cld42+Pr6x7 set also responds to quite heavy precipitation, with the peak of cloud fraction occurring at the same CTP level, but for slightly optically thinner clouds

(Fig. 5), suggesting a combination of convective cores and thick anvils. Figure 10 shows that TCPR2 also frequently occurs in the tropical warm pool oceans and Amazon basin, like TCPR1. However, the active hours are clearly different from TCPR1. For example, in addition to the early morning times as in Fig. 9, TCPR2 also frequently occurs just before noon and in the afternoon between 2pm and 6pm in the warm pool region. In the Amazon basin, the most active hour is shifted to afternoon, the time previous studies noted as the most active hours of continental convection driven by thermodynamic processes (Giles et al. 2020; Janowiak et al. 2005).

In Fig. 10 we can see RFO local minima troughs four times a day: 12am-2am, 8am-10am, 12pm-2pm, and 8pm-10pm. In these time windows, the occurrence of TCPR2 decreases abruptly, which may suggest an artifact in the IMERG dataset. Similar trough-like patterns are also detected with other TCPRs, notably for TCPR4, TCPR8, and TCPR9, as for TCPR2, and less prominently for TCPR3 and TCPR7, as for TCPR1 (see Supplementary Material Part A). The troughs, especially spaced in two pairs 12-h apart, points to the possibility of an artifact stemming from particular sensors on board sun-synchronous satellites used in IMERG. In particular, these times match the overpass times of several cross-track scanning sounders in the constellation which generate double-peaks in precipitation distributions over ocean (You et al. 2020). However, troughs of the same diurnal cycle analysis over land-only are still notable (but with weakened signal; see Supplementary Material Part A), indicating that there may be other unidentified factors at play or that the troughs represent true diurnal signals in fact.

In summary, through the projection of hybrid regimes onto IMERG domain, the temporal resolution for some of regimes with the greatest precipitation contribution and most likely associated with convection, is greatly improved. In addition to the diurnal cycle analysis shown in this subsection, this diurnally-extended dataset of cloud-precipitation hybrid regimes has enormous potential to examine other features of convective systems. We should also note that this projection method works not only for the deep tropical regimes, but also for the hybrid regimes of extended latitudes when the higher weights of precipitation are used in the clustering procedure (see Supplementary Material Part B).

6. Summary and Conclusion

We generated hybrid cloud and precipitation regimes (CPRs) by applying the *k*-means clustering algorithm, with advanced initialization and objective measures to select the optimal number of clusters *k*, on coincident cloud and precipitation data from MODIS and IMERG. We discussed how multiple versions of hybrid CPR sets can be obtained depending on the relative weighting of the cloud and precipitation information and the boundaries of the geographical domain.

Given that precipitation was represented by a rather coarse 6-bin histogram and clouds were represented by a 42-element joint histogram, a naïve concatenation of cloud and precipitation arrays implies a 7 to 1 ratio in cloud versus precipitation weighting. When performing joint clustering with this 48-element array, the patterns of the cloud histogram centroids looked quite similar to those of cloud-only centroids, indicating a weak influence of precipitation on the clustering. However, for the cloud regime associated with intermediate to heavy rainfall intensity, some outliers with relatively lighter rainfall were moved to other regimes of corresponding rainfall intensity, making the precipitation variability of hybrid regimes generally tighter. As the weight of precipitation in the joint clustering progressively increased (by replicating the precipitation histograms as needed), the precipitation histogram component of the hybrid centroids became more unique from those of the other centroids while the cloud histogram parts of the centroids started losing peak sharpness. In the set of equal-weight between cloud and precipitation, three CPRs of high clouds with light-to-intermediate precipitation intensity even shared quite similar cloud histogram patterns (but with distinct precipitation histogram patters, of course). Compared to the high cloud regimes experiencing dramatic changes by varying the weight of precipitation, low cloud regimes remained relatively unchanged among different sets, because their weak rainfall did not impact the clustering process.

Given that the precipitation histogram part of centroid became progressively more distinct from that of the other centroids as precipitation weight increased, we tested whether we can predict a specific CPR based on only the precipitation information of the grid cell. This attempt was motivated by the fact that IMERG dataset has much higher temporal resolution with nearly no missing data at 30-minute intervals compared to temporally sparse MODIS cloud observations. We found that, in the case of equal-weight set, seven high cloud regimes with intermediate-to-heavy precipitation can be predicted with over 90% accuracy by the precipitation information only. This result suggests that a projection of certain CPRs onto the IMERG domain is possible, opening thus a broad path for a variety of studies that require diurnally-resolved cloud information.

In a previous study by Jin et al. (2020), three cloud-only regimes related to tropical convective activities were selected, to study various features of convective systems at synoptic scales. However, their investigation was limited to snapshots of convective systems near 1:30pm LST due to the limitation of MODIS cloud observation availability and with morning Terra observations filling swath gaps based on persistence assumptions. The IMERG-based projection method enabled by hybrid regimes as mentioned above can expand their study in various directions. For example, thanks to 30-minute temporal resolution without gaps, diurnal cycle of convective systems can be examined in a manner demonstrated in Figs. 9 and 10. In addition, it is also possible to examine the life cycle of large-scale convective systems by systematically tracking them. While the prediction skill using IMERG precipitation is not perfect at all instances, the expansion of hybrid regimes to temporally high resolution is a significant advancement that can contribute to better understanding of large-scale tropical convective systems.

Acknowledgments.

We acknowledge funding from NASA's Precipitation Measurement Missions program. Resources supporting this work were provided by the NASA High-End Computing (HEC) Program through the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center.

Data Availability Statement.

IMERG precipitation data used in this study is openly available from the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) at https://doi.org/10.5067/GPM/IMERG/3B-HH/06 as cited in Huffman et al. (2019b). *Daily MODIS L3 cloud histogram data for Terra (MOD08_D3) and Aqua (MYD08_D3) are openly available from the Level-1 and Atmosphere Archive & Distribution System (LAADS) Distributed Active Archive Center (DAAC) in the Goddard Space Flight Center at* *https://doi.org/10.5067/MODIS/MOD08_D3.061 and* *https://doi.org/10.5067/MODIS/MYD08_D3.061 as cited in Platnick et al. (2017b). The MODIS cloud regime and MODIS-IMERG cloud-precipitation hybrid regime datasets* *derived in 15°S-15°N domain is available at https://data.nasa.gov/Earth-Science/Cloud-Precipitation-Hybrid-Regimes-MODIS-IMERG-in-/ee3g-swmf.*

REFERENCES

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

Arthur, D., and S. Vassilvitskii, 2007: k-means++: the advantages of careful seeding. *SODA 07 Proc. Eighteenth Annu. ACM-SIAM Symp. Discrete Algorithms*, 1027–1035.

Caliński, T., and J. Harabasz, 1974: A dendrite method for cluster analysis. *Commun. Stat. - Theory Methods*, **3**, 1–27, https://doi.org/10.1080/03610927408827101.

Davies, D. L., and D. W. Bouldin, 1979: A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-1**, 224–227, https://doi.org/10.1109/TPAMI.1979.4766909.

Giles, J. A., R. C. Ruscica, and C. G. Menéndez, 2020: The diurnal cycle of precipitation over South America represented by five gridded datasets. *Int. J. Climatol.*, **40**, 668–686, https://doi.org/10.1002/joc.6229.

Huffman, G. J., and Coauthors, 2019a: Algorithm Theoretical Basis Document (ATBD) version 06. NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG). https://gpm.nasa.gov/sites/default/files/document_files/IMERG_ATBD_V06_0.pdf.

——, E. F. Stocker, D. T. Bolvin, E. J. Nelkin, and J. Tan, 2019b: GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/GPM/IMERG/3B-HH/06.

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

Janowiak, J. E., V. E. Kousky, and R. J. Joyce, 2005: Diurnal cycle of precipitation determined from the CMORPH high spatial and temporal resolution global precipitation analyses. *J. Geophys. Res.*, **110**, D23105, https://doi.org/10.1029/2005JD006156.

Jin, D., L. Oreopoulos, D. Lee, N. Cho, and J. Tan, 2018: Contrasting the co-variability of daytime cloud and precipitation over tropical land and ocean. *Atmospheric Chem. Phys.*, **18**, 3065–3082, https://doi.org/10.5194/acp-18-3065-2018.

——, ——, ——, J. Tan, and K. Kim, 2020: Large-Scale Characteristics of Tropical Convective Systems Through the Prism of Cloud Regime. *J. Geophys. Res. Atmospheres*, **125**, e2019JD021157, https://doi.org/10.1029/2019JD031157.

Ketchen, D. J., and C. L. Shook, 1996: The Application of Cluster Analysis in Strategic Management Research: An Analysis And Critique. *Strateg. Manag. J.*, **17**, 441–458, https://doi.org/10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G.

Kikuchi, K., and B. Wang, 2008: Diurnal Precipitation Regimes in the Global Tropics. *J. Clim.*, **21**, 2680–2696, https://doi.org/10.1175/2007JCLI2051.1.

King, M. D., and Coauthors, 2003: Cloud and aerosol properties, precipitable water, and profiles of temperature and water vapor from MODIS. *IEEE Trans. Geosci. Remote Sens.*, **41**, 442–458, https://doi.org/10.1109/TGRS.2002.808226.

Lee, D., L. Oreopoulos, G. J. Huffman, W. B. Rossow, and I.-S. Kang, 2013: The Precipitation Characteristics of ISCCP Tropical Weather States. *J. Clim.*, **26**, 772–788, https://doi.org/10.1175/JCLI-D-11-00718.1.

Luo, Z. J., R. C. Anderson, W. B. Rossow, and H. Takahashi, 2017: Tropical cloud and precipitation regimes as seen from near-simultaneous TRMM, CloudSat, and CALIPSO observations and comparison with ISCCP: Tropical Clouds From Radars and Lidar. *J. Geophys. Res. Atmospheres*, **122**, 5988–6003, https://doi.org/10.1002/2017JD026569.

MacQueen, J., 1967: Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1 of, Oakland, CA, USA., 281–297.

Mason, S., C. Jakob, A. Protat, and J. Delanoë, 2014: Characterizing Observed Midtopped Cloud Regimes Associated with Southern Ocean Shortwave Radiation Biases. *J. Clim.*, **27**, 6189–6203, https://doi.org/10.1175/JCLI-D-14-00139.1.

Oreopoulos, L., and William. B. Rossow, 2011: The cloud radiative effects of International Satellite Cloud Climatology Project weather states. *J. Geophys. Res.*, **116**, D12202, https://doi.org/10.1029/2010JD015472.

——, N. Cho, D. Lee, S. Kato, and G. J. Huffman, 2014: An examination of the nature of global MODIS cloud regimes. *J. Geophys. Res. Atmospheres*, **119**, 8362–8383, https://doi.org/10.1002/2013JD021409.

——, ——, ——, and ——, 2016: Radiative effects of global MODIS cloud regimes. *J. Geophys. Res. Atmospheres*, **121**, 2299–2317, https://doi.org/10.1002/2015JD024502.

Pike, M., and B. R. Lintner, 2020: Application of Clustering Algorithms to TRMM Precipitation over the Tropical and South Pacific Ocean. *J. Clim.*, **33**, 5767–5785, https://doi.org/10.1175/JCLI-D-19-0537.1.

Pincus, R., S. Platnick, S. A. Ackerman, R. S. Hemler, and R. J. P. Hofmann, 2012: Reconciling Simulated and Observed Views of Clouds: MODIS, ISCCP, and the Limits of Instrument Simulators. *J. Clim.*, **25**, 4699–4720, https://doi.org/10.1175/JCLI-D-11-00267.1.

Platnick, S., M. D. King, S. A. Ackerman, W. P. Menzel, B. A. Baum, J. C. Riedi, and R. A. Frey, 2003: The MODIS cloud products: algorithms and examples from terra. *IEEE Trans. Geosci. Remote Sens.*, **41**, 459–473, https://doi.org/10.1109/TGRS.2002.808301.

——, and Coauthors, 2017a: The MODIS Cloud Optical and Microphysical Products: Collection 6 Updates and Examples From Terra and Aqua. *IEEE Trans. Geosci. Remote Sens.*, **55**, 502–525, https://doi.org/10.1109/TGRS.2016.2610522.

——, M. D. King, and P. A. Hubanks, 2017b: *MODIS Atmosphere L3 Daily Product (C6.1)*. NASA MODIS Adaptive Processing System, Goddard Space Flight Center, [doi:10.5067/MODIS/MOD08_D3.061; doi:10.5067/MODIS/MYD08_D3.061],.

——, and Coauthors, 2018: MODIS cloud optical properties: User guide for the collection 6/6.1 Level-2 MOD06/MYD06 product and associated Level-3 datasets, Version 1.1. https://atmosphere-imager.gsfc.nasa.gov/sites/default/files/ModAtmo/MODISCloudOpticalPropertyUserGuideFinal_v1.1_1.pdf.

Robertson, A. W., N. Vigaud, J. Yuan, and M. K. Tippett, 2020: Toward Identifying Subseasonal Forecasts of Opportunity Using North American Weather Regimes. *Mon. Weather Rev.*, **148**, 1861–1875, https://doi.org/10.1175/MWR-D-19-0285.1.

Rossow, W. B., and R. A. Schiffer, 1999: Advances in Understanding Clouds from ISCCP. *Bull. Am. Meteorol. Soc.*, **80**, 2261–2287, https://doi.org/10.1175/1520-0477(1999)080<2261:AIUCFI>2.0.CO;2.

——, 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, https://doi.org/10.1029/2005GL024584.

——, A. Mekonnen, C. Pearl, and W. Goncalves, 2013: Tropical Precipitation Extremes. *J. Clim.*, **26**, 1457–1466, https://doi.org/10.1175/JCLI-D-11-00725.1.

Tan, J., and L. Oreopoulos, 2019: Subgrid Precipitation Properties of Mesoscale Atmospheric Systems Represented by MODIS Cloud Regimes. *J. Clim.*, **32**, 1797–1812, https://doi.org/10.1175/JCLI-D-18-0570.1.

——, C. Jakob, W. B. Rossow, and G. Tselioudis, 2015: Increases in tropical rainfall driven by changes in frequency of organized deep convection. *Nature*, **519**, 451–454, https://doi.org/10.1038/nature14339.

——, G. J. Huffman, D. T. Bolvin, and E. J. Nelkin, 2019a: IMERG V06: Changes to the Morphing Algorithm. *J. Atmospheric Ocean. Technol.*, **36**, 2471–2482, https://doi.org/10.1175/JTECH-D-19-0114.1.

——, ——, ——, and ——, 2019b: Diurnal Cycle of IMERG V06 Precipitation. *Geophys. Res. Lett.*, **46**, 13584–13592, https://doi.org/10.1029/2019GL085395.

Tselioudis, G., W. Rossow, Y. Zhang, and D. Konsta, 2013: Global Weather States and Their Properties from Passive and Active Satellite Cloud Retrievals. *J. Clim.*, **26**, 7734–7746, https://doi.org/10.1175/JCLI-D-13-00024.1.

Vernekar, A. D., B. P. Kirtman, and M. J. Fennessy, 2003: Low-Level Jets and Their Effects on the South American Summer Climate as Simulated by the NCEP Eta Model. *J. Clim.*, **16**, 297–311, https://doi.org/10.1175/1520-0442(2003)016<0297:LLJATE>2.0.CO;2.

Yang, S., and E. A. Smith, 2006: Mechanisms for Diurnal Variability of Global Tropical Rainfall Observed from TRMM. *J. Clim.*, **19**, 5190–5226, https://doi.org/10.1175/JCLI3883.1.

You, Y., V. Petkovic, J. Tan, R. Kroodsma, W. Berg, C. Kidd, and C. Peters-Lidard, 2020: Evaluation of V05 Precipitation Estimates from GPM Constellation Radiometers Using KuPR as the Reference. *J. Hydrometeorol.*, **21**, 705–728, https://doi.org/10.1175/JHM-D-19-0144.1.

Zhang, Z., and S. Platnick, 2011: An assessment of differences between cloud effective particle radius retrievals for marine water clouds from three MODIS spectral bands. *J. Geophys. Res.*, **116**, https://doi.org/10.1029/2011JD016216.

Zhao, J., J. Kug, J. Park, and S. An, 2020: Diversity of North Pacific Meridional Mode and Its Distinct Impacts on El Niño-Southern Oscillation. *Geophys. Res. Lett.*, **47**, e2020GL088993, https://doi.org/10.1029/2020GL088993.

TABLES

Table 1. Optimal values of *k* according to the DBC metric for the two domains and four precipitation weights.

Deep Tropics (15°S-15°N)		Low-to-Mid Latitudes (50°S-50°N)	
Cloud-only	k=14	Cloud-only	k=15
Pr_wt=1	k=16	Pr_wt=1	k=20

Pr_wt=3	k=16	Pr_wt=3	k=19
Pr_wt=7	k=19	Pr_wt=7	k=22

FIGURES

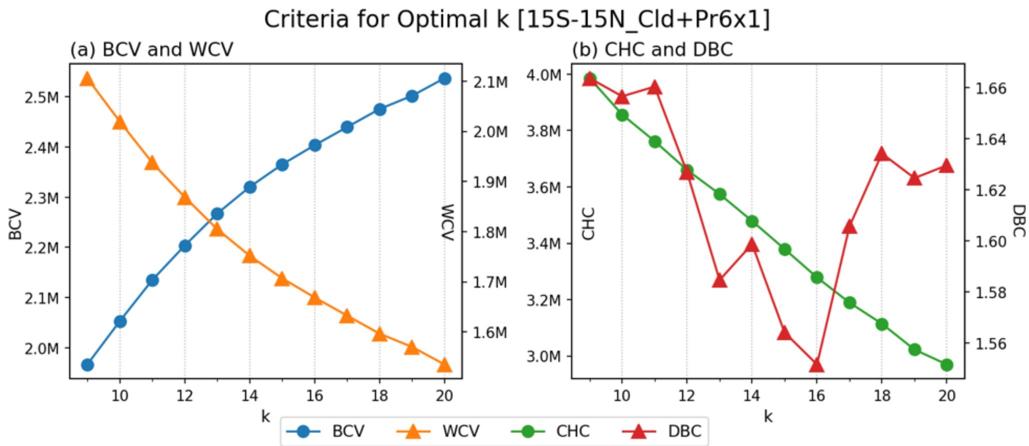

Figure 1. Criteria for selecting optimal number of clusters (*k*) are displayed as a function of *k* for the case of 7:1 weighting in the combined cloud-precipitation array (Cld42+Pr6x1). (a) Between-cluster variance (BCV; blue circles) and within-cluster variance (WCV; orange triangles), and (b) Calinski-Harabasz criterion (CHC; green circles) and Davies-Bouldin criterion (DBC; red triangles). We note that for the same *k*, a set of initial centroids (i.e., one realization) selected as the best by one criterion can be different from that selected for another criterion.

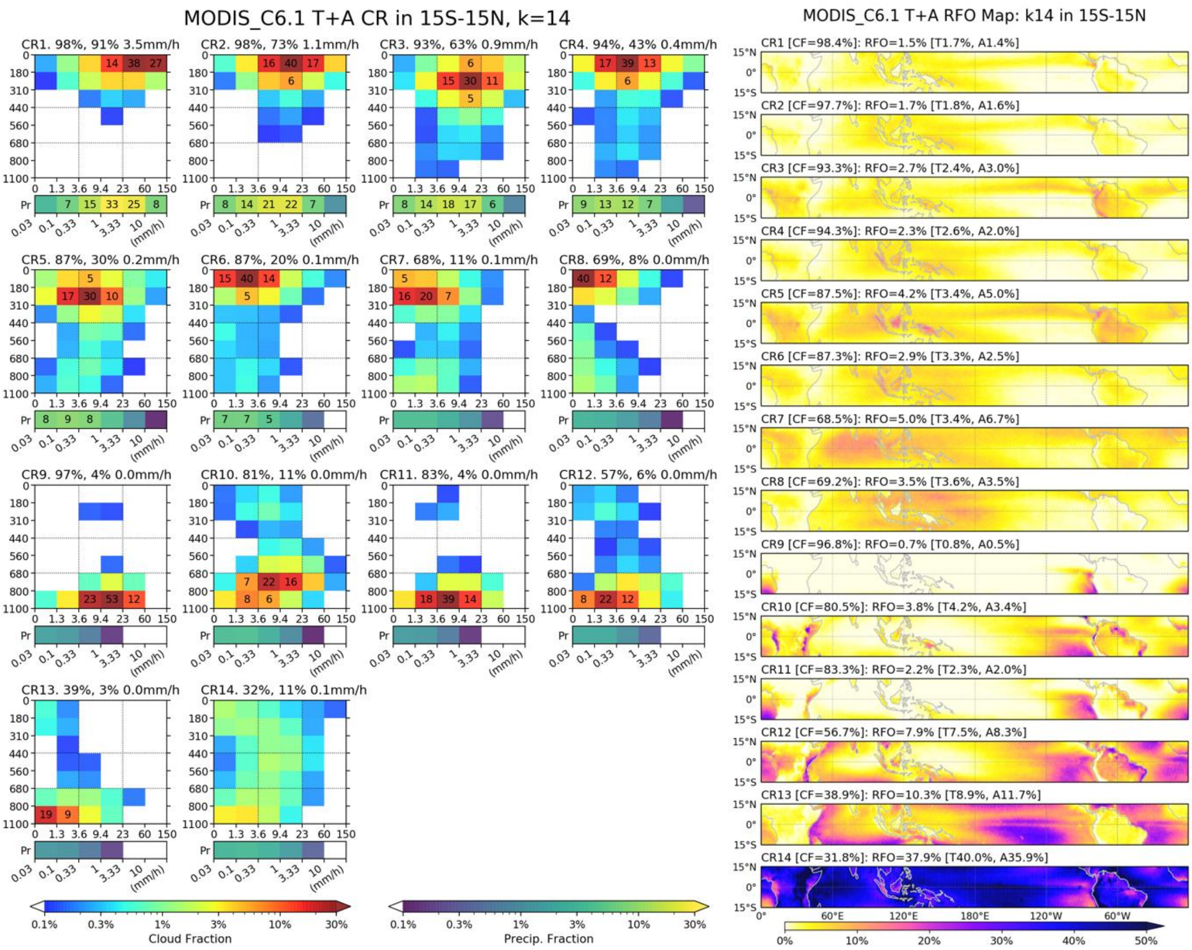

Figure 2. Deep tropics cloud-only regime centroids (mean histograms, left) and geographical distribution of relative frequency of occurrence (RFO, right). Bin cloud fraction values exceeding 5% are shown explicitly on the centroid panels. The precipitation histograms shown below the cluster centroids are composite means for each cloud regime. In addition to the total cloud fraction, total precipitation fraction which is the sum of all precipitation histogram bin values, and estimated mean precipitation rate based on the histogram are also given on the panel title. Above the RFO panels, individual Terra and Aqua RFOs are provided in brackets.

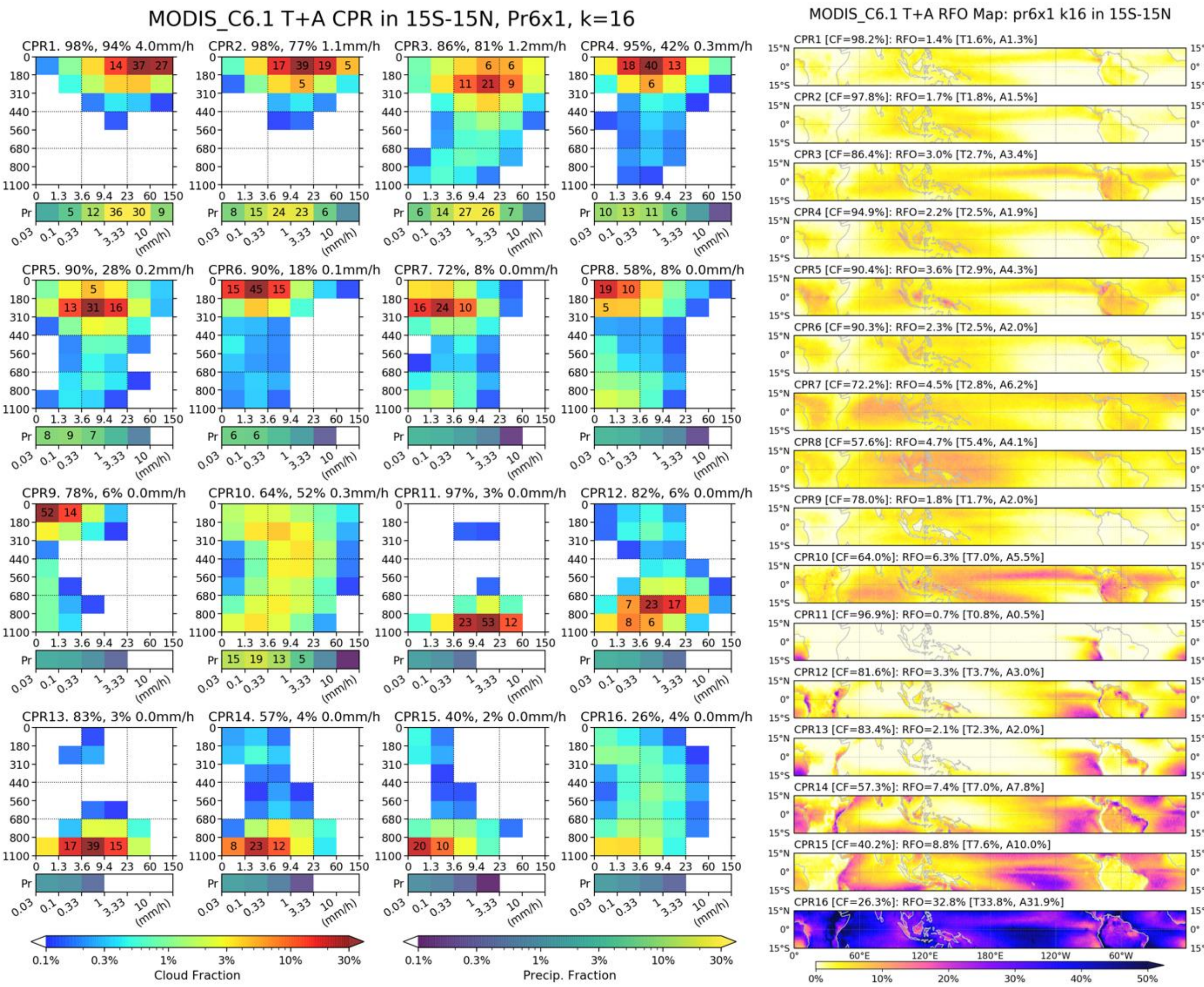

Figure 3. Similar to Fig. 2 but now the precipitation part of the centroids was also derived from clustering where precipitation contributed with a weight number of 1 (Cld42+Pr6x1; i.e., 7:1 ratio in 48-element combined array in clustering).

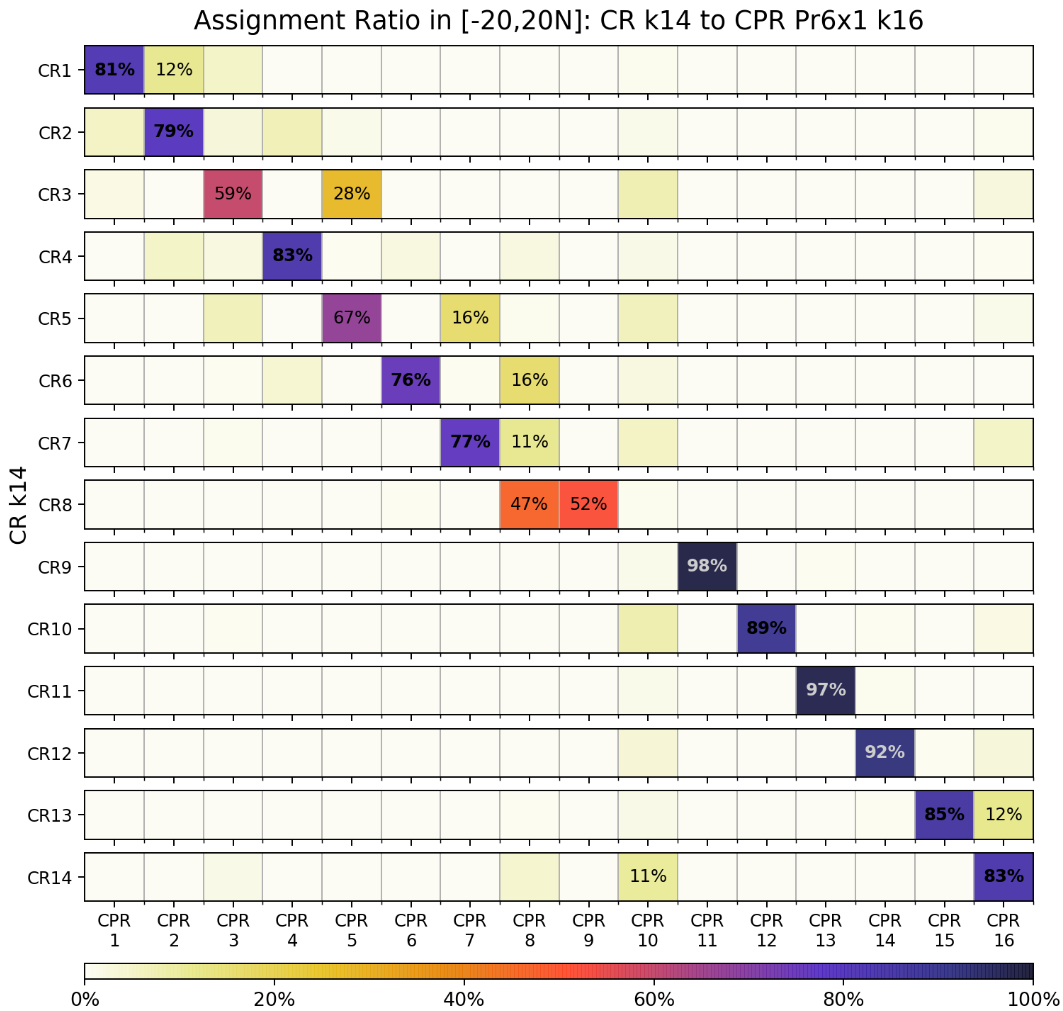

Figure 4. Regime coincidence distribution matrix comparing assignment frequencies on the same grid cell between the cloud-only regimes of Fig. 2 (y-axis) and the hybrid regimes Cld42+Pr6x1 of Fig. 3 (x-axis). The values of the matrix are normalized across rows, and values above 10% are explicitly shown. Please note that while the regimes were derived with data in 15°S-15°N, regime assignment was performed in the extended domain, 20°S-20°N for both Terra and Aqua, because tropical phenomena often extend beyond the 15° latitude boundaries.

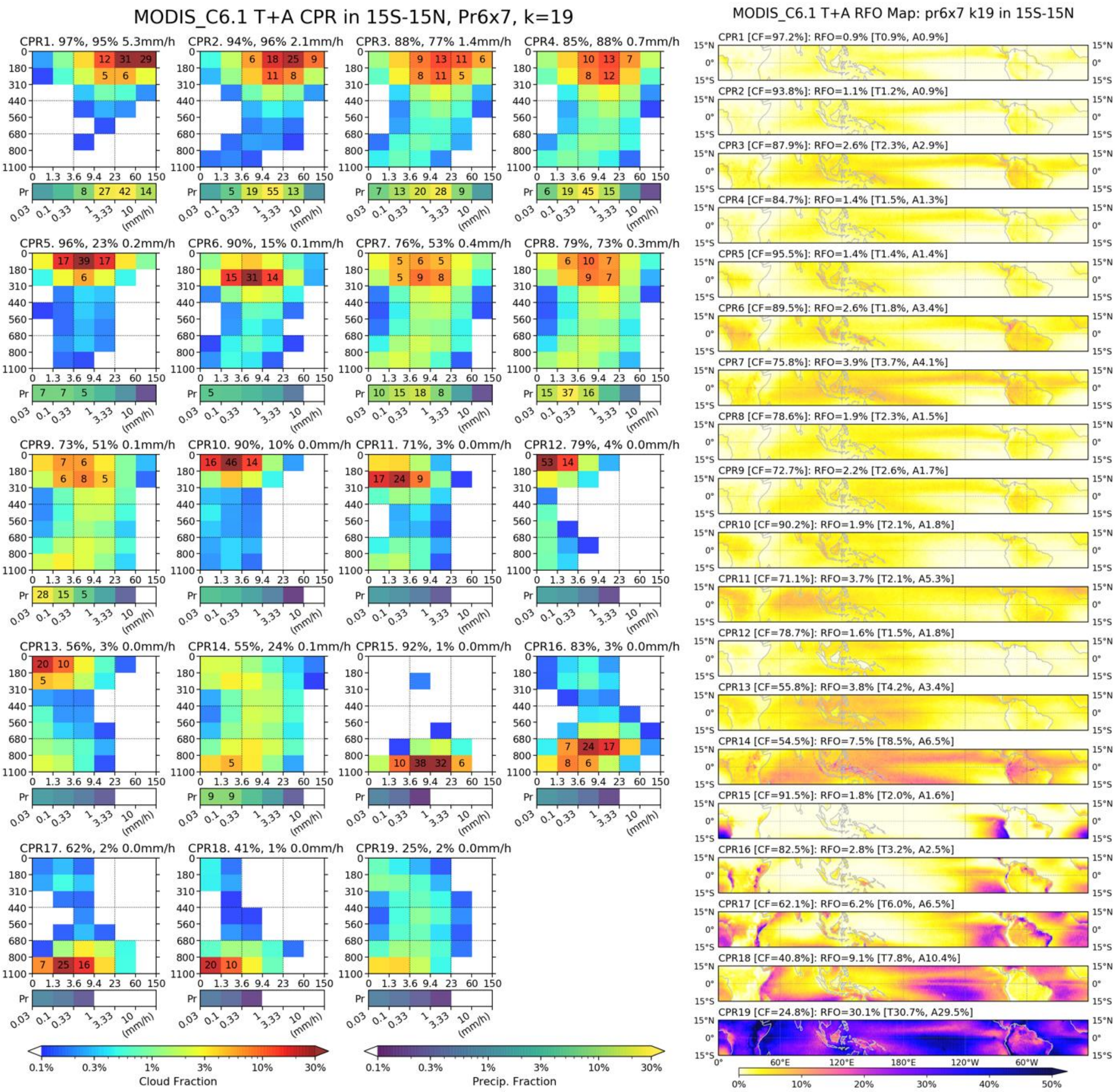

Figure 5. As Fig. 3 but with precipitation contributing with weight number 7 (Cld42+Pr6x7; i.e., 7:7 ratio in 84-element combined array in clustering).

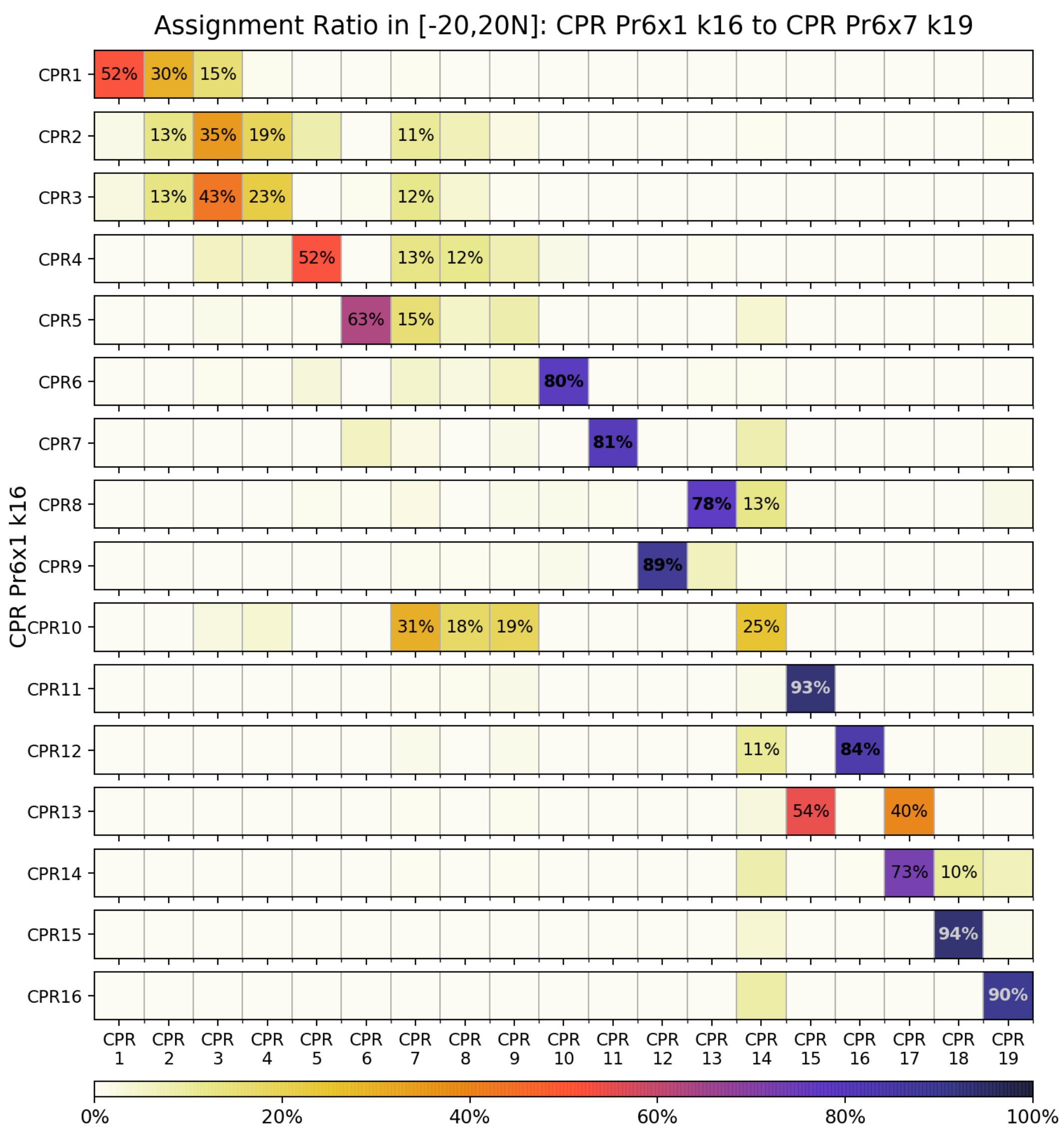

Figure 6. As Figure 4 but between Cld42+Pr6x1 (y-axis; Fig. 3) and Cld42+Pr6x7 (x-axis; Fig. 5).

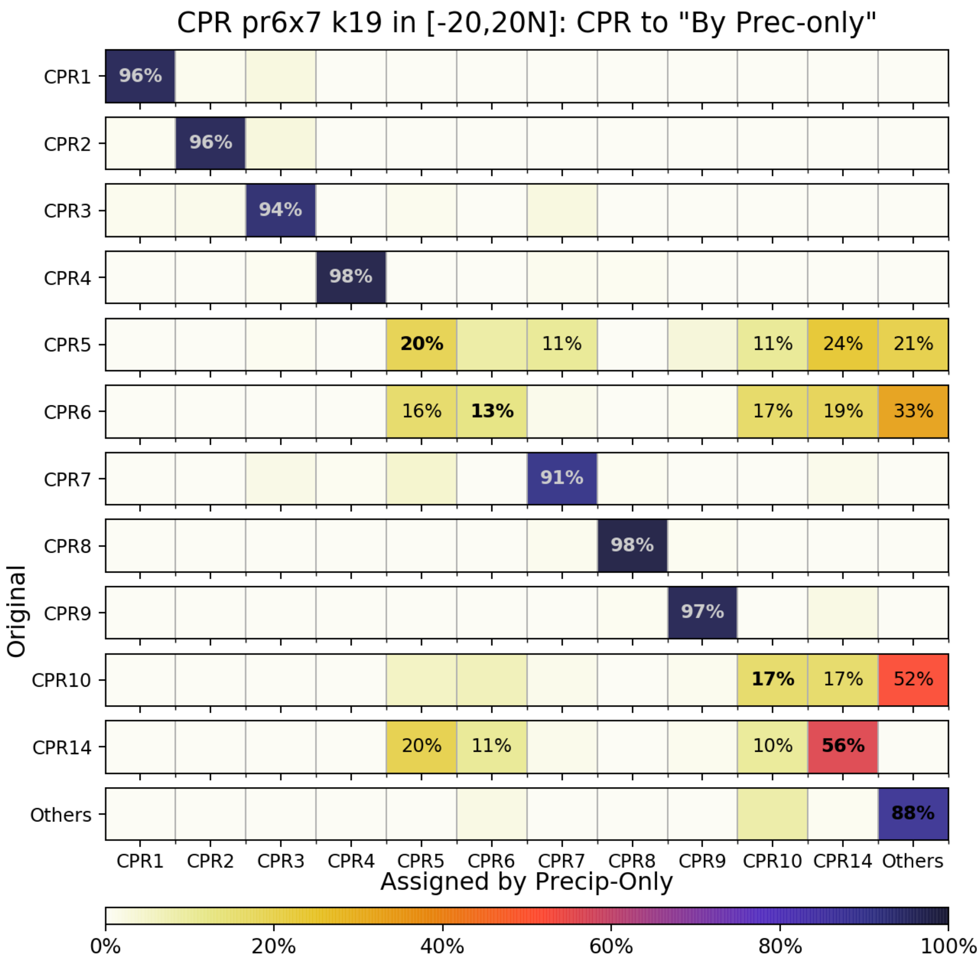

Figure 7. As Fig. 4 but between original Cld42+Pr6x7 (y-axis) and regimes assigned by precipitation only (x-axis). Regimes with precipitation fractions below 10% have been combined in the "Others" category.

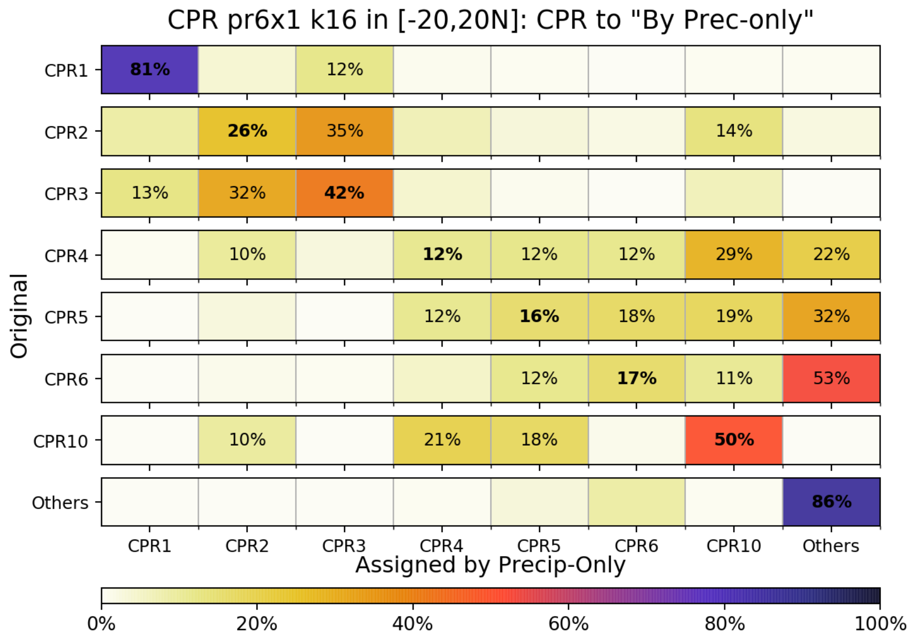

Figure 8. Same as Figure 7 but for the set of Cld42+Pr6x1.

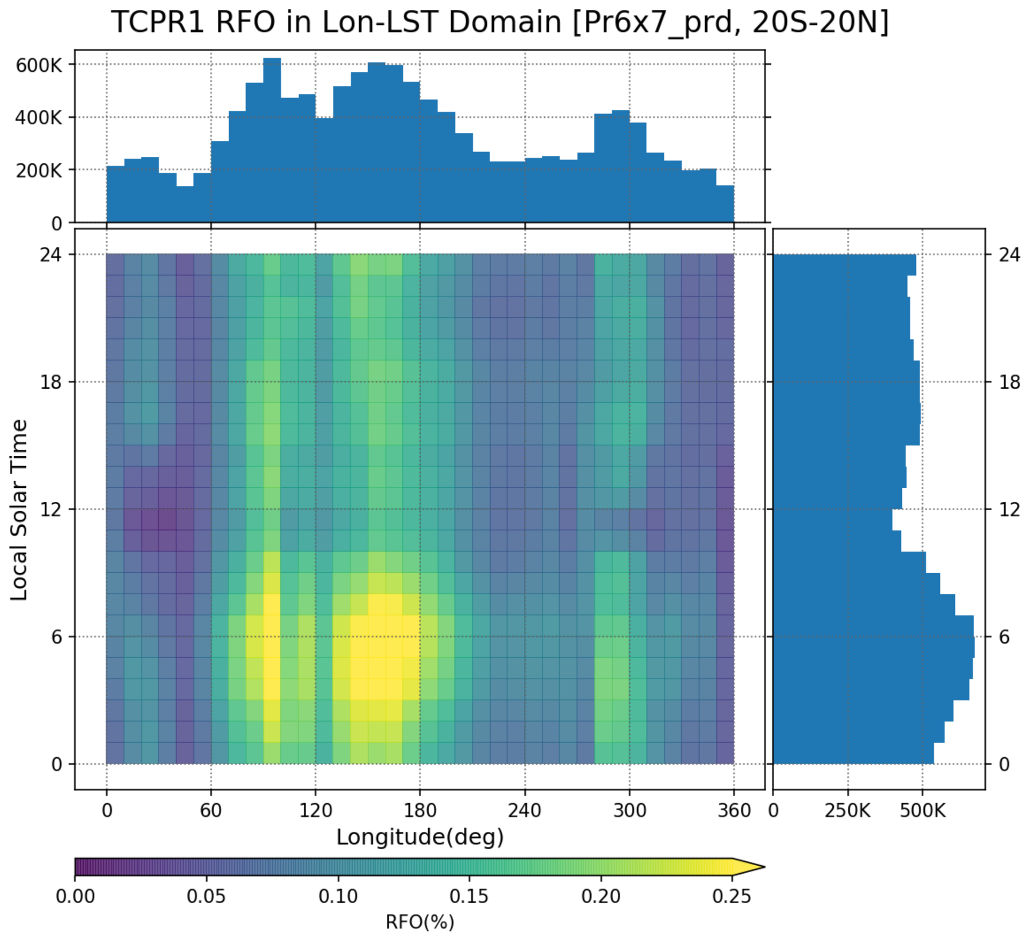

Figure 9. RFO of TCPR1 of the Cld42+Pr6x7 set predicted by precipitation-only in a longitude (x-axis) and local solar time (LST; y-axis) phase space. Bin resolutions are 10° in longitude, and 1-hour in time. The top and right panels show RFO marginal histograms (sums across rows and columns before normalization) for the same resolution of longitude and LST.

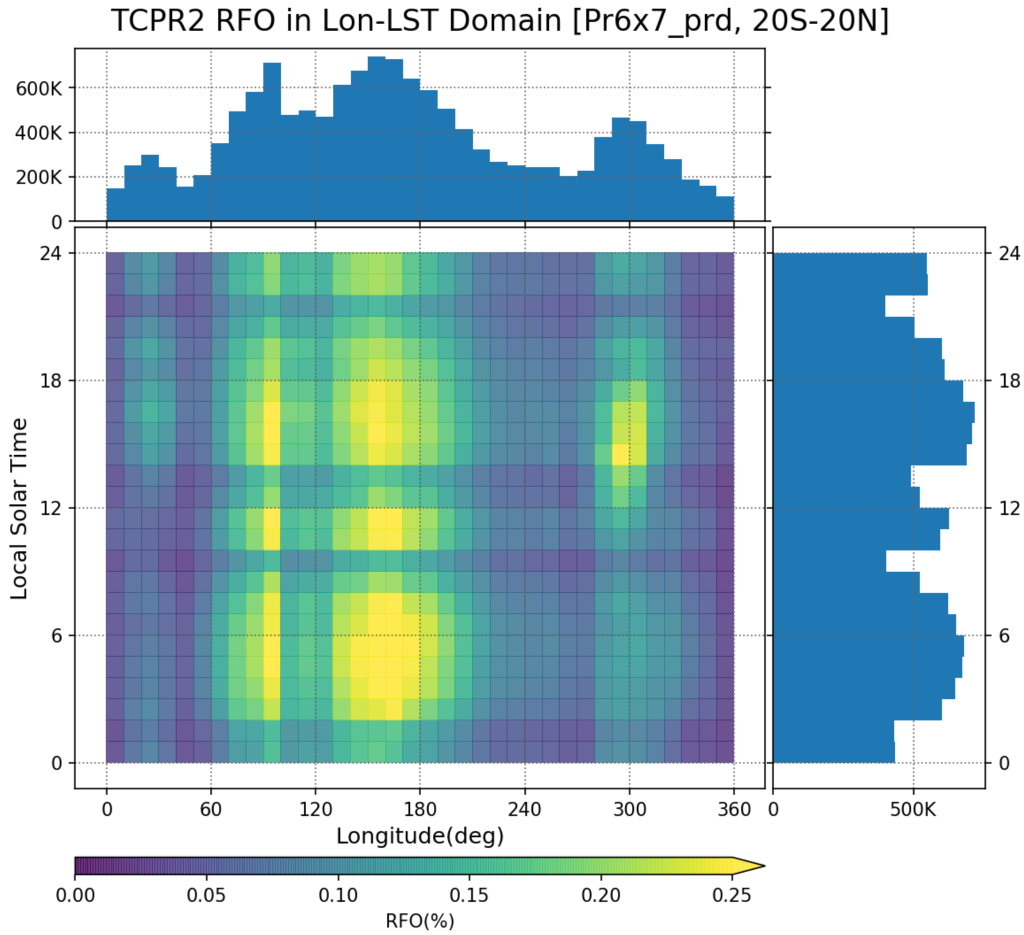

Figure 10. Same as Fig. 9, but for TCPR2.