CrossMark

A long-term tropical mesoscale convective systems dataset based on a novel objective automatic tracking algorithm

Xiaomeng Huang[1,2] · Chenqi Hu[1] · Xing Huang[1,3] · Yang Chu[4] · Yu-heng Tseng[5] · Guang Jun Zhang[1,6] · Yanluan Lin[1]

Abstract

Mesoscale convective systems (MCSs) are important components of tropical weather systems and the climate system. Long-term data of MCS are of great significance in weather and climate research. Using long-term (1985–2008) global satellite infrared (IR) data, we developed a novel objective automatic tracking algorithm, which combines a Kalman filter (KF) with the conventional area-overlapping method, to generate a comprehensive MCS dataset. The new algorithm can effectively track small and fast-moving MCSs and thus obtain more realistic and complete tracking results than previous studies. A few examples are provided to illustrate the potential application of the dataset with a focus on the diurnal variations of MCSs over land and ocean regions. We find that the MCSs occurring over land tend to initiate in the afternoon with greater intensity, but the oceanic MCSs are more likely to initiate in the early morning with weaker intensity. A double peak in the maximum spatial coverage is noted over the western Pacific, especially over the southwestern Pacific during the austral summer. Oceanic MCSs also persist for approximately 1 h longer than their continental counterparts.

1 Introduction

Tropical convection is a key component of the global climate system because it transports mass, momentum and heat vertically. The convection sometimes organizes into mesoscale convective systems (MCSs), which are prominent features occurring worldwide. They have captured much attention due to their extensive impacts, including strong winds, precipitation, flooding, hail and lightning (Houze 2004). MCSs that occur over ocean regions can be precursors of tropical cyclones (Bister 1996; Gray 1998; Teng et al. 2014). Moreover, they play a critical role in connecting atmospheric convection with large-scale atmospheric circulation patterns (Chen et al. 1996; Laing and Fritsch 2000; Moncrieff 2010). MCSs are composed of convective cores and stratiform anvils, whose vertical and horizontal structures evolve during different phases of their life cycle (Houze 1982). In particular, MCSs are embedded within tropical waves (Jakob 2003), synoptic-scale super-clusters and the Madden-Julian Oscillation (Nakazawa 1988), affecting the atmosphere and atmosphere–ocean coupling across a range of scales (Moncrieff 2013). The complex structure of MCSs and their interactions with their surroundings hinder the monitoring and forecasting of MCSs. To obtain a better understanding of the role of MCSs in the climate system, building a comprehensive long-term dataset of MCSs over the whole tropics is highly desirable.

The emergence of Earth-orbiting satellites and geostationary observations make it possible to observe MCSs continuously at global scales. Detecting MCSs using satellite data has a long history. In general, the detection of MCSs involves two major steps: the identification of MCSs and

✉ Yanluan Lin
  yanluan@tsinghua.edu.cn

1   Ministry of Education Key Laboratory for Earth System Modeling, and Department for Earth System Science, Tsinghua University, Beijing 100084, China

2   Laboratory for Regional Oceanography and Numerical Modeling, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China

3   National Supercomputing Center in Wuxi, Wuxi 214072, China

4   Department of Electrical and Electronic Engineering, Imperial College, London, UK

5   Institute of Oceanography, National Taiwan University, Taipei, Taiwan

6   Scripps Institution of Oceanography, La Jolla, CA, USA

the tracking of their evolution (e.g., Machado and Rossow 1993; Machado et al. 1998; Schröder et al. 2009; Hennon et al. 2011; Fiolleau and Roca 2013a). In the identification step, each MCS is located in satellite observations using a set of criteria and constraints. The tracking step then aims to determine the track of each MCS via comparison of various properties of the MCSs at successive times.

At the identification stage, the convective clusters in each satellite snapshot are captured and labelled. Boer and Ramanathan (1997) proposed an iterative multi-threshold cloud identification scheme to improve the delineation of cloud boundaries. This scheme identifies both the convective cores and stratiform anvils associated with MCSs, and thus delineates their "true" spatial extents (Zhang et al. 1999; Wilcox and Ramanathan 2001; Roca et al. 2002; Xu et al. 2005). However, the multi-threshold identification method tends to produce smaller clusters than the single-threshold method. Williams and Houze (1987) used a specified threshold to objectively identify cloud clusters. Mapes and Houze (1993) further discussed the choices of threshold and developed a single-threshold method that has been commonly adopted in climatological research on MCSs (Pope et al. 2008; Goyens et al. 2012; Fiolleau and Roca 2013b).

The tracking stage represents a special type of multiple object tracking problem because MCSs continuously evolve and deform during their lifetimes. MCSs differ significantly from each other in terms of their morphology and intensity, and the properties vary rapidly during the lifetimes of MCSs. In addition, an individual MCS can split into two or more MCSs or merge with other MCSs. As a result, successively tracking each targeted convective system has been a challenge. Manually tracking was applied in early studies, but this process is labor intensive and subjective with large uncertainties. To overcome these limitations, various objective automatic methods have been proposed over the past few decades, including the area-overlapping method (Williams and Houze 1987), the overlap of ellipsoidal equivalents (Boer and Ramanathan 1997), the centroid tracking method (Johnson et al. 1998), and the maximum spatial correlation method (Carvalho and Jones 2001).

The most widely used automated method is the area-overlapping method (Williams and Houze 1987; Arnaud et al. 1992; Mathon and Laurent 2001). This method assumes that the MCSs in successive frames represent the same entity if there are sufficient common overlapping pixels in their images. Some methods use a searching radius (Dixon and Wiener 1993; Johnson et al. 1998) to find the MCS in the next image rather than searching for overlapping pixels. The main problem for the fixed-search-radius methods is that MCSs can move at different speeds and thus a fixed radius is not reliable for all MCSs. It makes the tracking process more difficult by setting the searching radius. The area-overlapping method is conceptually straightforward and works

reasonably well for tracking large and slow-moving MCSs. However, it assumes that the location and area of MCSs do not change significantly with time. The method tends to fail for small and fast-moving MCSs, especially when the temporal resolution of available satellite observations is low. In addition, the cloud tracking algorithm (CTA) is another type of overlapping method, which performs overlapping of ellipsoidal equivalents (Boer and Ramanathan 1997). Previously, the cloud masks needed to be extracted and stored to calculate the overlapping area. Now, the cloud masks are no longer required for the CTA to reduce its computational costs and memory load. However, the CTA may not be applicable to data with larger sampling intervals, in that MCSs may travel too far in several hours to assure accurate association. This property also explains why small systems cannot be tracked with these overlapping schemes (Boer and Ramanathan 1997).

To remedy the limitations of conventional area-overlapping methods, Kalman Filter (KF) based methods are used in this work. The KF is an optimal estimator that can predict the state of a process and use measurements to correct its predictions. One of its many successful applications is object tracking (e.g., Reid 1979; Xing et al. 2009). Through estimating the speed and direction of the target systems, the KF method can robustly track small and fast-moving systems, which is not well-represented or even missed by conventional area-overlapping methods. As a result, the complete life cycles of MCSs can be better captured.

Previous endeavors in MCS identification and tracking were generally limited in terms of their temporal and spatial coverage. To the best of our knowledge, Hennon et al. (2011) provides the only publicly available, long-term tropical cloud cluster (TCC) dataset. However, the MCSs that develop or move over land areas are not included in their TCC dataset. In this study, we combine the overlapping method with KF-based approaches to perform MCS tracking. Moreover, we apply this novel method to long-term global satellite infrared brightness temperature observations to construct a long-term tropical MCS dataset that covers both land and ocean regions.

The paper is structured as follows. Section 2 introduces the satellite data and the design of our new tracking algorithm. The MCS dataset and selected applications are described in Sect. 3. Section 4 summarizes the results and discusses further potential applications of this dataset.

2 Data and methods

We used the European Union Cloud Archive User Service (CLAUS) project dataset (Hodges et al. 2000), a global dataset based on the calibrated International Satellite Cloud Climatology Project (ISCCP) B3 radiance data (Rossow and

Schiffer 1999), in this work. CLAUS has been widely used to detect convective activity (e.g., Yang and Slingo 2001; Nguyen and Duvel 2008; Dias et al. 2012; Dong et al. 2016). The available CLAUS data provides global brightness temperatures (BTs) at 3-h intervals sampled at 30 km (or 1/3°) scale, which provides a good indication of convection.
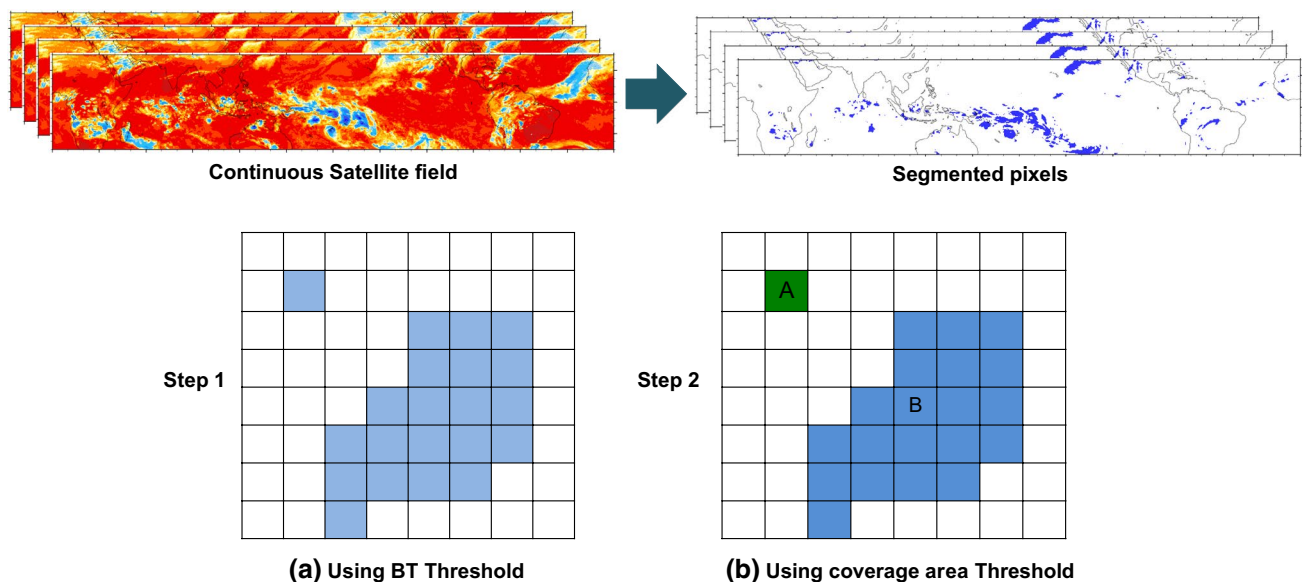
2.1 MCS identification

Low BTs generally correspond to the cold cloud shields of convective systems. The criteria for identifying MCSs are generally based on BT thresholds and minimum area coverage thresholds. The identification of MCSs is illustrated in Fig. 1. Utilizing 3-hourly satellite data, pixels satisfying the pre-defined criteria (defined below) are isolated from the surrounding continuous BT field. The bottom panel of Fig. 1 gives an example of how a potential MCS is detected. All pixels with BT values smaller than the threshold are extracted in the first step. The adjacent pixels are then considered to be part of a coherent region of interest (ROI). ROIs larger than the area coverage threshold are considered to be potential MCSs. In step 2, ROI B is found to be a potential MCS, while ROI A is discarded since it is too small to meet the area coverage threshold (Fig. 1).

One difficulty in the identification of MCSs is the lack of consensus on the definition of MCSs. BT thresholds ranging from 255 to 208 K have been proposed (Mapes and Houze 1993; Chen et al. 1996; Machado et al. 1998). For the minimum area coverage thresholds, values from 100 to 100,000 km^2 (Maddox 1980; Morel and Senesi 2002a; Kolios and Feidas 2010) have been applied to satellite data with different resolutions and channels. An overview of the thresholds for identifying MCSs is provided in Table 1 of Goyens et al. (2012). Although this overview is not exhaustive, it showed that a wide range of threshold values have been used in previous studies. Some studies have used a variable threshold of BT in different ocean basins to account for differences in climatological background convection temperatures (Hennon et al. 2011, 2013). Imposing stricter criteria will certainly exclude some potential MCSs, whereas the use of less strict thresholds may result in the inclusion of some spurious MCSs.

Goyens et al. (2012) reviewed the criteria used in the identification of MCSs and suggested broad adoption of a BT threshold of 233 K to indicate tropical atmospheric convection. Fiolleau and Roca (2013b) also concluded that a threshold of 233 K is reasonable, and this value has been widely used in many previous investigations. Therefore, in this work, the BT threshold is set to 233 K, and the minimum area coverage is set to 5000 km^2 to produce a prototype dataset without loss of generality. Considering the wide range of BT and area coverage thresholds used, we provide a flexible interface that enables users to provide their own criteria for MCS identification to generate datasets of interest. Our focus is thus on the applicability of the algorithm instead of the specific criteria used.

Fig. 1 The schematic of the MCS identification method. Top panel: Getting the continuous brightness temperature (BT) distribution from each satellite snapshot (left panel) and extracting pixels from the background field (right panel). Bottom panel: An illustration of how we identify a potential MCS within a sample domain during one time step. **a** All pixels that satisfy the BT threshold are identified and marked in light blue. **b** Adjacent pixels are linked as coherent regions (region A is shown in green and region B is shown in dark blue) if their sizes are larger than the prescribed area coverage threshold

Table 1 Features of MCSs tracked by the area-overlapping (AOL) method and our new method (KF) in 2000

Feature Method	Total number	Mean lifetime (h)	Mean size (km^2)	Mean speed (km/h)
AOL	46,512	10.74	83,679	40.9
KF	58,234	11.85	61,486	42.3
KF_T228	50,668	11.75	50,766	39.6
KF_A10000	38,847	12.73	82,420	41.9
KF_25%	58,219	11.85	60,162	40.6

Total number indicates the cumulative number of MCSs occurring over the whole year

The latter three variables indicate the mean lifetime, size and moving speed of each MCS. The size and moving speed are calculated by averaging over each time step of the lifespan of each MCS. The base comparison of MCSs records between AOL and KF are detected from the BT threshold of 233 K, the area coverage threshold of 5000 km^2 and the overlapping threshold 15%. The sensitivity experiments for changing the criteria are KF_T228, KF_A10000 and KF_25% with a lower BT threshold of 228 K, a larger area coverage threshold of 10,000 km^2 and a higher overlapping threshold 25% separately

2.2 MCS tracking

Once all potential MCSs have been identified at consecutive times, we need to determine their trajectories. The evolution of MCSs is a continuous process, but satellite observations are discrete and limited by their temporal resolution. For each potential MCS identified at the current time t, the potential MCSs at the next time step (t + 1) are searched to identify matches. The critical technical challenge in tracking strategies is how to match the potential MCSs, i.e., how to identify the same potential MCSs in successive time frames.

Physically, the location of a single MCS in two consecutive frames is constrained by the theoretical maximum distance the MCS can travel. Thus, the conventional area-overlapping tracking method assumes that MCSs in successive frames belong to the same entity if there are sufficient common overlapping pixels in their images. An overlapping rate threshold is pre-defined. For each potential MCS at time (t + 1), if more than one potential MCS at time t meets the requirements, the algorithm selects the potential MCS with the greatest degree of overlap. However, this method fails if there are no overlapping pixels between the pairs of successive frames; for example, small and fast-moving MCSs may not overlap when a 3-hour interval is used. On the other hand, the KF method assumes that the moving state of a potential MCS at time t evolved from its prior state at time (t-1). The probability of moving state s_t can be represented as

$$P\left(s_t | s_1 : s_{t-1}\right) = P(s_t | s_{t-1}) \qquad (1)$$

The potential MCS evolves from the prior state at time (t-1) according to the following equation,

$$s_t = A s_{t-1} + q \qquad (2)$$

where the moving state vector $s_t = \left(x_t \ y_t \ \dot{x}_t \ \dot{y}_t \right)$ contains the coordinates (x_t, y_t) and velocity $(\dot{x}_t, \dot{y}_t)$ of a MCS at time t. The state transition matrix A describes the dynamics from

time t − 1 to t. The process noise term q is assumed to be drawn from the standard normal distribution.

After initiation, in each tracking time step t, the KF method first predicts the movement state of the potential MCS, then updates its estimation by maximizing the posterior probability s_t of the observed position of the target potential MCS(see Appendix A for more details).The distance between the position of a potential MCS and the position of the predicted potential MCS at time (t + 1) are then calculated to determine the most appropriate potential MCS for the continuation (right panel in Fig. 2). One uncertainty of the KF algorithm stems from the measured positions of the potential MCSs. We determine the position of a potential MCS by averaging the coordinates of the coldest 10 pixels inside the cold cloud shield. If a potential MCS contains fewer than 10 pixels, then the geographic information of all the pixels are averaged to establish the position. Both methods introduce uncertainties into the position determination because of the irregular shape and inhomogeneous spatial distribution of cloud systems. The larger the potential MCS, the more difficult it is to accurately determine its exact position. Note that a larger MCS is more likely to overlap with itself in consecutive frames.

The tracking procedures of the area-overlapping method and the KF method are compared schematically in Fig. 2. Each panel in each column represents one of the three distinct steps of the tracking algorithm. All the potential MCSs are identified at two successive times in the first step. The major difference between the two tracking approaches is highlighted in the middle panel of Fig. 2. The area-overlapping method compares the potential MCSs at time t with each potential MCS at time (t + 1) by evaluating the overlap percentage (left panel). In this case, the potential MCS C2 from t_2 is associated with the two candidate potential MCSs in t_1, A1 and B1. For the area-overlapping method, the overlapping rates A(C2, B1) and A(C2, A1) between each

Fig. 2 A schematic illustration of the differences between the area-overlapping (left) and KF (right) tracking methods. Step 1 determines all potential MCSs from the identification stage. Step 2 associates each potential MCS in the subsequent time step with all potential MCSs in the current time step. Step 3 links the same potential MCSs into a single trajectory. The area-overlapping method matches C2 with B1 based on their larger overlapping rate. The KF method selects B1 because the predicted position of B1 (pB2) is closer to the position of C2 (pC2)
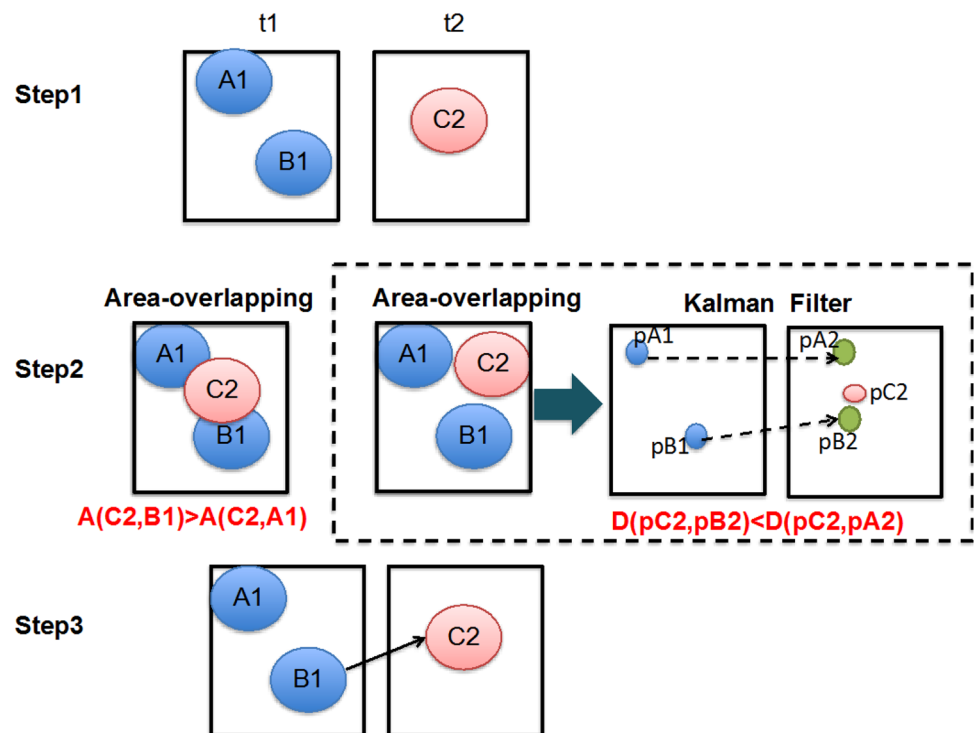
candidate potential MCS in t_1 with the candidate potential MCS in t_2 are calculated. In the right panel, the area-overlapping method fails since there is no overlap between the two potential MCSs at two successive times, due either to the small coverage or rapid movement of the potential MCSs. The KF method is able to predict the positions of potential MCSs at the next time step $(t+1)$. By comparing the Euclidean distance of the predicted position and the actual position of potential MCSs at $(t+1)$, the closest potential MCSs that satisfy the distance threshold are considered to be the same system. The KF method predicts the position of A1 and B1 at t2 (pA2 and pB2) from their previous positions at t1 (pA1 and pB1). The distances between the position of C2 and pA2 and pB2 are then compared.

The bottom panel in Fig. 2 provides further illustration of the trajectories derived from the area-overlapping and KF methods. Even if the area-overlapping method fails or cannot identify a reasonable track given the limitations imposed by temporal resolution, the KF method can help track the MCSs and thus obtain a better coverage of the MCS life cycle (see Fig. 3). If a potential MCS persists over at least 3 successive frames (Nguyen and Duvel 2008), i.e., it has a life duration longer than 6 h, it will be recorded as one MCS. On average, 4,718 MCSs were recorded in each month after this duration requirement was imposed, and this number accounts for approximately 32% of the total number of potential MCSs tracked.

The advantages of the KF method are illustrated in Fig. 3 over the coast of western Africa on 2nd Jan 1993. In this example, the area-overlapping method failed to track the MCSs with IDs 918 and 993. In addition, it failed to track the MCS with ID 1048 in the last time frame. MCS 918 initiated at 15:00 on 2nd Jan. (Fig. 3b) and had propagated southwest by the next time step (Fig. 3c). The area-overlapping method mislabels MCS 918 as two different systems due to the small overlapping area of MCS 918 between two successive frames. As a result, the area-overlapping method misses MCSs 918 and 993 because they failed to persist over 3 successive time steps. MCS 1048 initiated at 12:00 on 2nd Jan and had a long lifespan; it had dissipated by 03:00 3rd Jan (Fig. 3a–f). However, the area-overlapping method lost its track (Fig. 3f), resulting in a shorter lifespan for it (Fig. 3h). In contrast, the KF method is able to capture the complete life cycle of these MCSs (Fig. 3g).

Statistically, the KF method can keep track of potential MCSs, whereas the overlapping method failed 31% of the time during 1985–2008. The area-overlapping method resulted in a large percentage (64%) of potential MCSs lasting for only one time step. On the other hand, the percentage of potential MCSs seen in only one time step was reduced by 48% after the KF method was used.

Table 1 shows the basic features of the MCSs tracked by the KF and area-overlapping methods during year 2000. The total number of MCSs that persisted at least 3 successive frames using the KF method is 58,234 over the whole year, which is approximately 25% greater than that obtained using the area-overlapping method. The average life span of the MCSs determined by the KF method is approximately

Fig. 3 Cloud top BTs obtained from the CLAUS dataset from 12:00 UTC 2nd Jan to 03:00 UTC 3rd Jan 1993 are shown in (**a–f**). The pixels contained within red contours have BTs less than 233 K. The trajectories of MCSs that form on 2nd Jan 1993 and persist over at least 3 frames, as determined using the new KF-based method and the area-overlapping (AOL) method, are shown in (**g, h**), respectively. Each dot displays the centroid of a MCS in each time step, and the initiation time is labeled in (**g, h**). The brackets in (**g**) indicate MCS IDs from the output dataset file
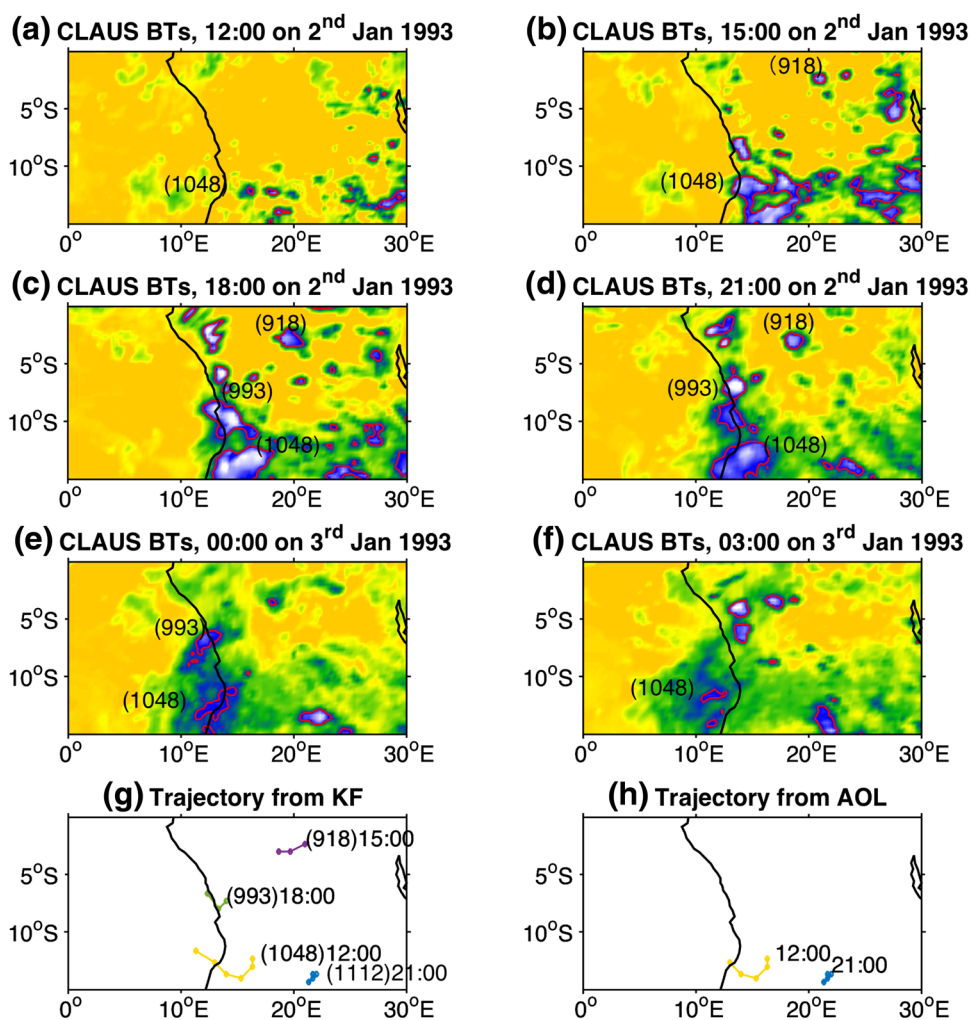
11.8 h, which is about one hour longer than that obtained using the area-overlapping method (10.7 h). The average moving speed of MCSs is 11.75 m s^{-1} (42.3 km/h), which is slightly larger than that of MCSs tracked only by the area-overlapping method. Overall, the KF method tracked more small MCSs with a mean size of 61,486 km^2, which is smaller in size than that obtained without the KF method (83,679 km^2). These results indicate that the KF method performs better in capturing small and fast-moving MCSs, and thus is a preferred approach for MCS tracking. Next, we show some preliminary results with a focus on the diurnal variations of MCSs over land and ocean regions based on the MCS dataset.

3 Application examples

3.1 Dataset description

Using the CLAUS dataset, we generated a long-term tropical MCSs dataset that covers the period from 1985 to 2008 with a BT threshold of 233 K and an area coverage threshold of 5000 km^2. As mentioned above, users of the dataset can easily change the criteria used in the algorithm to generate MCS records that meet their needs. The dataset contains basic trajectory information along with other characteristics of each MCS, including intensity, area, eccentricity and lifetime (Table 2). The MCSs identified in Jan 1985 may have initiated in the preceding month, and the MCSs identified in Dec 2008 may have dissipated in the following month, which lie outside the time period covered by the CLAUS dataset. To ensure that all MCSs contain their entire lifespans, the records in Jan 1985 and Dec 2008 are discarded. Houze (2004) concluded that tropical cyclones might spin up from MCSs. Our MCSs dataset have not removed the tropical cyclones to keep the integrity of MCSs data that could be used to investigate the relationship between MCSs and tropical cyclones (Kouadio et al. 2010; Yuan and Houze 2010). The number of tropical cyclones is very small compared to that of MCSs, and thus has negligible effects on the following MCSs statistical analysis.

Table 2 Description of the output records for each tracked MCS

Variable	Unit	Description
ID		The unique ID of a MCS
Lifetime	h	Persistence time of a MCS
gLat	N	Latitude of the geometric center of a MCS in degrees north
gLon	E	Longitude of the geometric center MCS in degrees east
wLat	N	Latitude of the centroid (weighted center) of a MCS in degrees north
wLon	E	Longitude of the centroid (weighted center) of a MCS in degrees east
Size	km^2	MCS area coverage
Eccentricity		Ratio of the minor axis to the major axis of the ellipse that provides the best fit to a MCS
BTavg	K	Average BT of all pixels within a MCS
BTmin	K	Lowest BT of all pixels within a MCS
Time	UTC	MCS occurrence time
Speed	km/h	The propagation speed of a MCS from the present time step to the next time step
Direction	degree	Movement direction of a MCS from the present time step to the next time step, measured clockwise from the east

To make the data accessible to a wide user community, the MCS dataset is available online at https://doi.pangaea.de/10.1594/PANGAEA.877914. In addition, we performed a series of sensitivity experiments using different criteria to verify our algorithm. As discussed before, different criteria lead to different MCSs records, and influence the climatological features of MCSs, including intensity, coverage area and lifetime. A small BT threshold (KF_T228 in Table 1) leads to less MCSs being identified, with shorter lifetime and smaller size. A larger coverage area threshold (KF_A10000 in Table 1) also generates less MCSs, but with longer lifetime and larger size. A larger overlapping threshold (KF_25% in Table 1) in the tracking stage has minor influence in the MCSs number, lifetime and size. The detection results confirmed the robustness and effectiveness of our new tracking method. These datasets derived from different criteria can also be obtained from the same website. The detailed implementation, including parallel optimization of the algorithm, is presented in Appendix B.
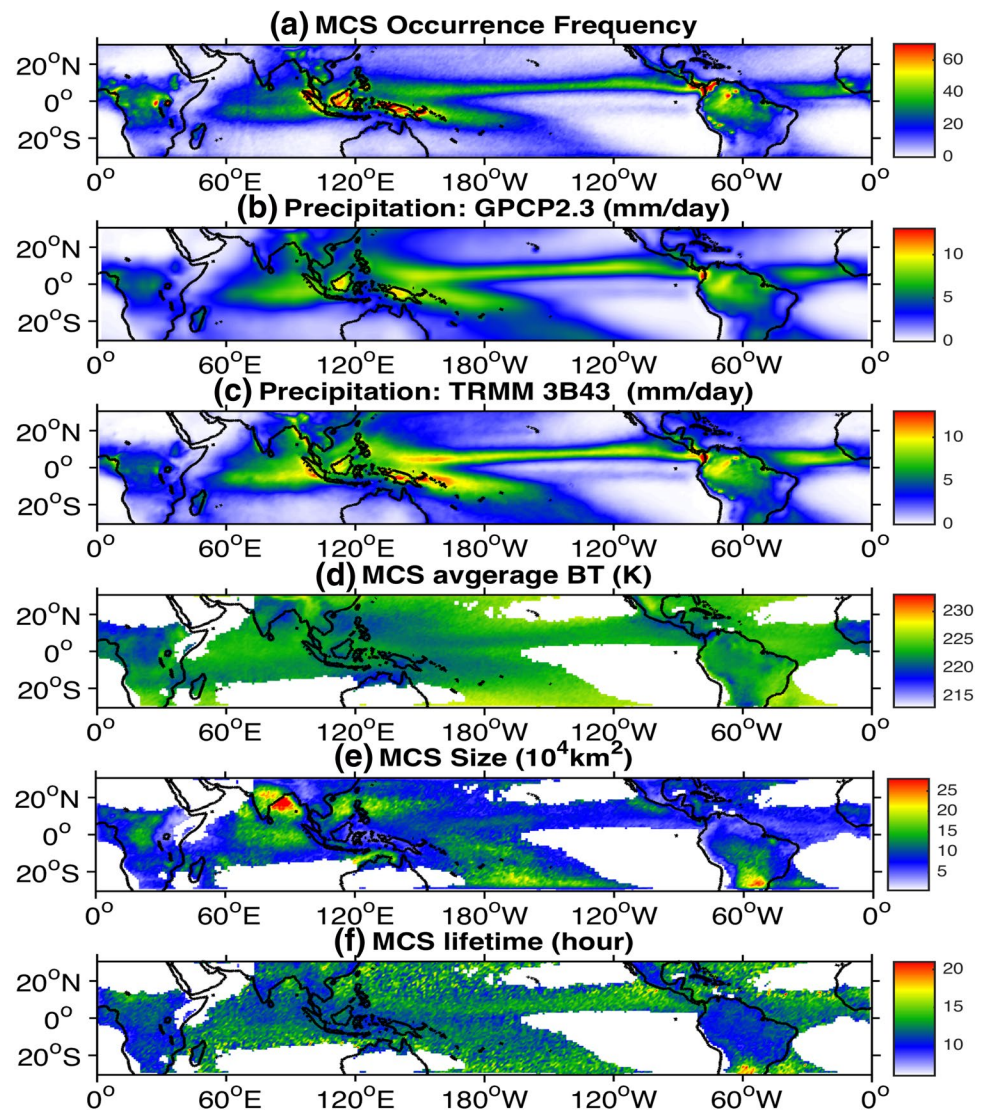
3.2 Spatial distribution of MCSs

The annual mean geographical distributions of MCS occurrence frequency, intensity, size, and lifetime are displayed in Fig. 4 along with precipitation from the Global Precipitation Climatology Project (GPCP) (Arkin 1989; Huffman et al. 1997) and the Tropical Rainfall Measuring Mission (TRMM). The GPCP v2.3 monthly product (Adler et al. 2003) includes the precipitation rate over the full period (1985–2008) at a spatial resolution of 2.5°. The TRMM v7 (Huffman et al. 2007) 3B43 monthly rainfall data is from 1998 to 2008 with a higher spatial resolution of 0.25°. All of the precipitation datasets have been interpolated to a 1° grid.

As expected, the MCS occurrence corresponds closely to the GPCP and TRMM precipitation distributions (Fig. 4b, c). The location of each MCS is defined by its centroid at each time step. Similar to the precipitation distributions, the regions with the most frequent occurrence of MCSs are located over tropical Africa, Amazonia and the Maritime Continent. MCSs are also prevalent over the tropical warm pool and intertropical convergence zone (ITCZ) and monsoon trough. Few MCSs appear in the tropical southeastern Pacific and the southern Atlantic, where relatively low SSTs and stable atmospheric conditions inhibit the development of convection. The distribution of MCS occurrence from our new dataset agrees well with other MCSs datasets, such as those determined using ice scattering signatures (Mohr and Zipser 1996), mesoscale convective complexes (MCCs, Laing and Fritsch 1997), TCCs (Hennon et al. 2013) and deep convection weather states (Tan et al. 2015) in the tropics. The similarities in the spatial distribution of MCCs with those in other datasets indicates the robustness and reliability of the new method. We also examined the spatial distribution of the occurrence of MCSs after changing the criteria used in the identification stage, and the correlation coefficient between different MCS datasets is greater than 0.95.

The intensity of MCSs can be represented by the averaged BTs of their pixels (Table 2, BTavg). Note that BTs are a measure of convective activity with lower values for more intense convection. For this study, our intensity definition is BTavg, therefore lower values mean greater intensity. Overall, MCSs are more intense over land than over oceans, and the most intense MCSs occur in central Africa (Fig. 4d). Strong MCSs also occur in the coastal areas of India, the South China Sea, mid-America and northern Australia. In terms of size, larger MCSs are more inclined to develop over ocean regions (Fig. 4e), such as the Bay of Bengal, the

(a) MCS Occurrence Frequency

(b) Precipitation: GPCP2.3 (mm/day)

(c) Precipitation: TRMM 3B43 (mm/day)

(d) MCS avgerage BT (K)

(e) MCS Size (10^{4}km^2)

(f) MCS lifetime (hour)

northern Indian Ocean, the South China Sea, the northern Australian marginal sea and the sub-tropical South Pacific. The largest MCSs over the Bay of Bengal are generally associated with monsoon depressions, and can sometimes develop into synoptic scale systems accompanied by large stratiform precipitation areas (Houze and Churchill 1987). Another region with large MCSs is the southern part of the Amazon basin, consistent with the precipitation distribution shown in Fig. 4b, c.

Globally, MCSs that initiate over the ocean last longer than MCSs that develop over land (Fig. 4f). The characteristics of MCS identified here are generally in line with previous surveys that have documented that MCSs that develop over oceans are larger, shallower and longer-lasting than those over the continents (Mohr and Zipser 1996; Laing and Fritsch 1997). Overall, the features of MCSs identified in this dataset are similar to those seen in

previous studies, but the time period has been extended to over 20 years.

3.3 Comparison of MCSs over land and ocean

The differences in the diurnal variations of convective systems over land and ocean areas have been noted in many previous studies (Chen and Houze 1997; Liu and Zipser 2008; Inoue et al. 2009; Kolios and Feidas 2010). To better characterize this distinction, we stratified the life cycle evolution of MCSs into four stages: initiation, maximum intensity, maximum spatial area coverage, and dissipation. The initiation stage represents the time when a MCS has just begun to develop and become identified; the maximum intensity stage refers to the time when the intensity of the MCS is greatest in terms of brightness temperature over its lifespan; the maximum spatial area coverage stage refers to

the time when the MCS reaches its largest horizontal size; and the dissipation stage refers to the last occurrence time of the MCS.

Here, we classify the continental and oceanic regions based on several well-known MCS genesis locations. Three continental regions are examined, specifically Africa (0–50E), the Maritime Continent (80–160E) and America (260-320E) with ocean grids excluded. Similarly, five ocean basins are examined, including the Indian Ocean (50–100E), the western Pacific (120–160E), the central Pacific (160–220E), the eastern Pacific (220–280E), and the Atlantic (300–360E) with land grid points excluded. Figure 5 presents the regional boundaries separating MCS for different formation locations. The diurnal variation of the occurrence frequency of the four MCS stages over land and ocean areas are compared in Fig. 6. Note that the diurnal variation analysis is performed using 1-h intervals after converting the original data in Universal Time Coordinated (UTC) at 3-h intervals to local solar time (LST).

The MCSs over land display a prominent diurnal cycle with the highest initiation frequency in the afternoon (1500–1700 LST), followed by the maximum intensity in 1–2 h at 1700–1800 LST (Fig. 6a, b). The maximum spatial coverage is achieved in another hour and lasts for 3–4 h before dissipating around midnight (Fig. 6c, d). It should be noted that the area coverage threshold used in the MCSs dataset is 5000 km^2. In contrast, no distinct diurnal variation or preferred time exists for the initiation of MCSs over ocean areas. Oceanic MCSs generally reach their maximum intensity in the early morning (~0500–0600 LST). Interestingly, there are two peaks in the spatial coverage of oceanic MCSs, one occurring at 0500–0600 LST, and the other occurring in the afternoon at 1400–1600 LST. This double-peak feature has also been noted in previous studies (Dai 2001; Tsakraklides and Evans 2003). The Oceanic MCSs tend to dissipate more frequently in the late afternoon (1500–1700 LST). The distinct differences in the diurnal variation between the oceanic and land MCSs are closely related to the larger diurnal thermal variations over land than over the ocean. Over coastal areas and the Maritime Continent, the MCS activities are also strongly regulated by the local occurrence of sea/land breezes (Dai 2001; Goyens et al. 2012). Overall, the average lifespan of the oceanic MCSs (12.4 h) is ~1 h longer than that of their land-based counterparts (11.3 h).

A detailed comparison of the diurnal variations of MCSs over the three land regions suggests subtle differences (Fig. 7). MCSs in Africa display the largest diurnal variations in terms of intensity and maximum coverage, followed by America and the Maritime Continent. MCSs also last slightly longer over Africa (11.7 h) than over the other two regions (11.1 h), and this might be partially related to the greater intensity of MCSs in Africa as noted above (Fig. 4). MCSs tend to initiate and develop ~1 h earlier in America than the other two regions. Consistently, MCSs over Africa reach their maximum intensity and size 1–2 h earlier than those over the other two regions. Finally, all of the MCSs over the three land regions show the highest frequency of dissipation at midnight (Fig. 7d).
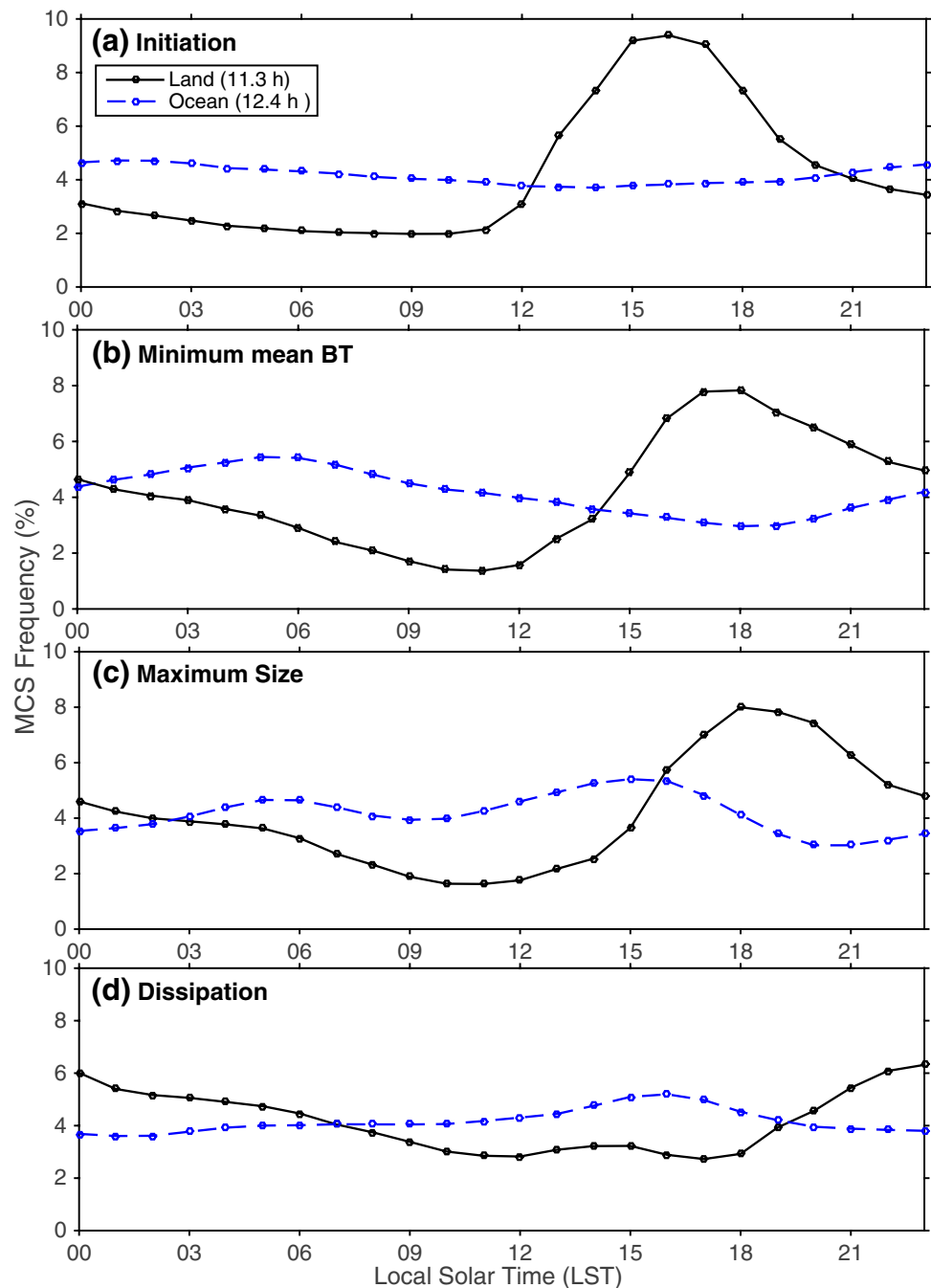
Similarly, the MCSs over the five ocean regions are compared in Fig. 8. Overall, the diurnal signals of MCSs over the five regions are close to each other. MCSs initiate in the western Pacific has a slightly different diurnal cycle with another initiation peak in the afternoon around 1500 LST. This feature may be related to the scattered islands located in the western Pacific, where the offshore MCSs might be influenced by the MCSs over the Maritime Continent.

The robust and evident diurnal variations in continental MCSs suggests that they respond quickly to diurnal solar heating and are susceptible to perturbations caused by their surroundings. MCSs that occur over oceans are possibly triggered by other mechanisms, such as equatorial waves and convective self-aggregation, and they are generally sustained by the presence of a warm moist boundary layer (Chen and Houze 1997).
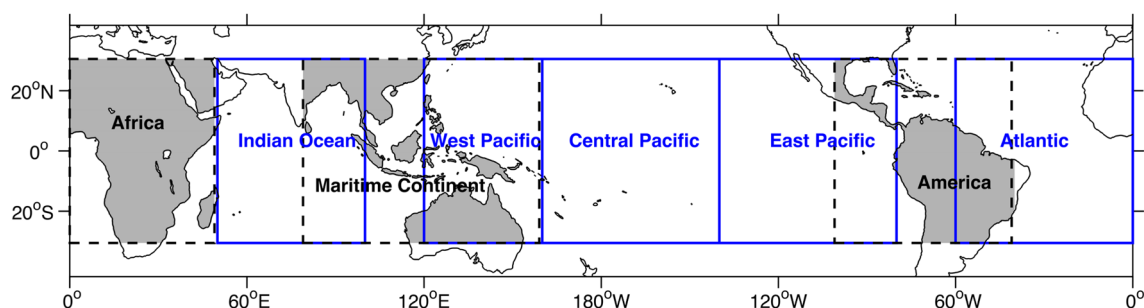
Fig. 5 Region boundaries used for this study. The shaded areas with black dashed lines are the continental regions. The blue lines show the boundaries of oceanic area

4 Conclusions

A novel algorithm that combines the conventional area-overlapping method with a Kalman filter has been developed to track MCSs. Since the new approach takes into account the estimated location of potential MCSs through a KF-based method, it can track small and fast-moving MCSs commonly missed by the area-overlapping method. In addition, a flexible interface has been designed so that users can apply their own criteria for the identification of MCSs. A parallel system is used to process the large volumes of satellite data efficiently. The algorithm code is publically available and thus facilitates further refinement of the method.

A 20-year tropical MCS dataset has also been generated based on satellite data with a temporal resolution of 3 h. This dataset builds upon earlier studies by substantially increasing the temporal scale and geographical coverage of the survey. A few examples are provided. The first one is the global distribution of occurrence frequency, intensity, and size of MCSs. It shows that MCSs occur more frequently and are more intense over land than over oceans. However, oceanic MCSs are generally larger and last longer than their
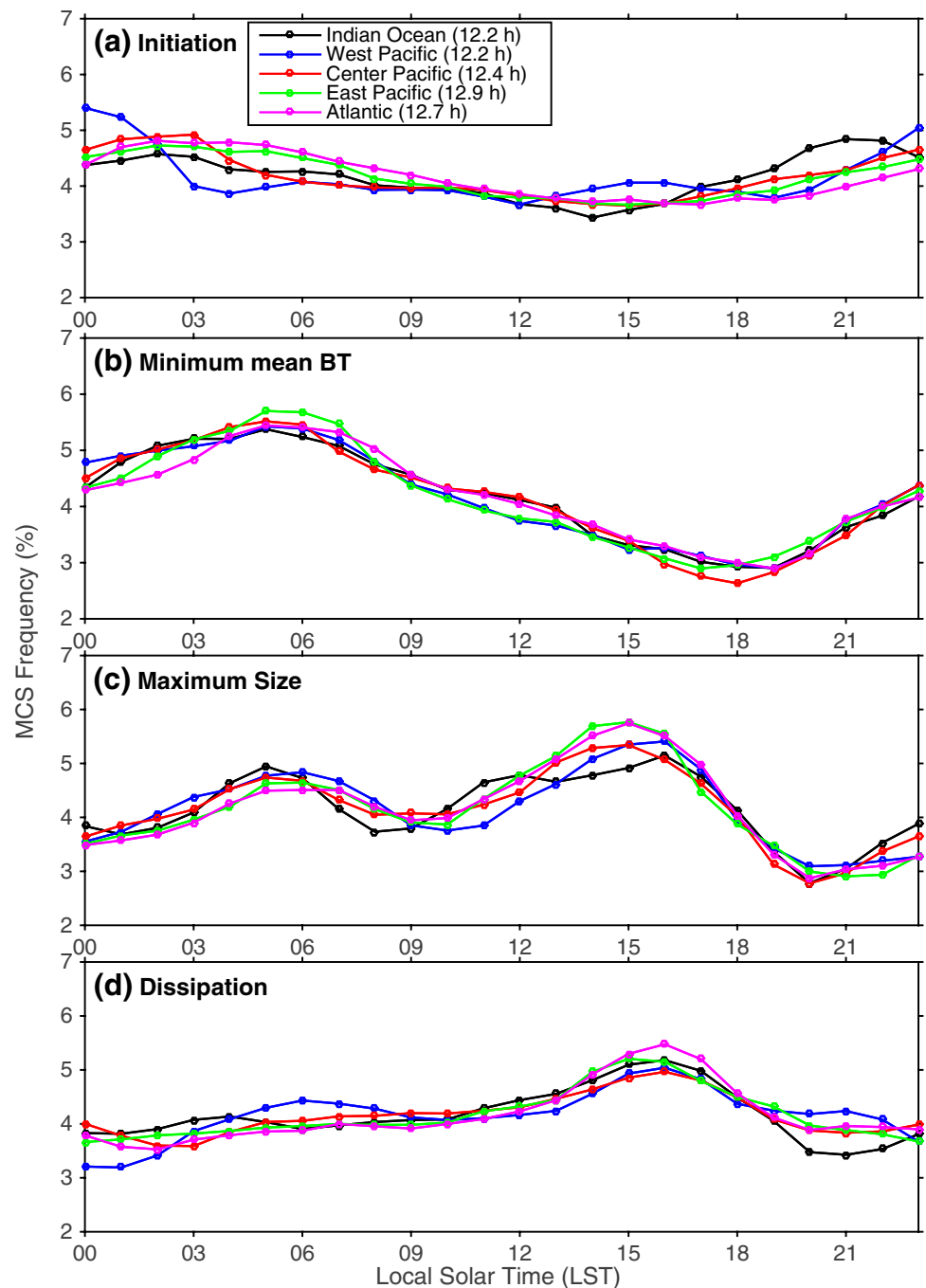
Fig. 7 Same as Fig. 6, but showing the three tropical continental regions
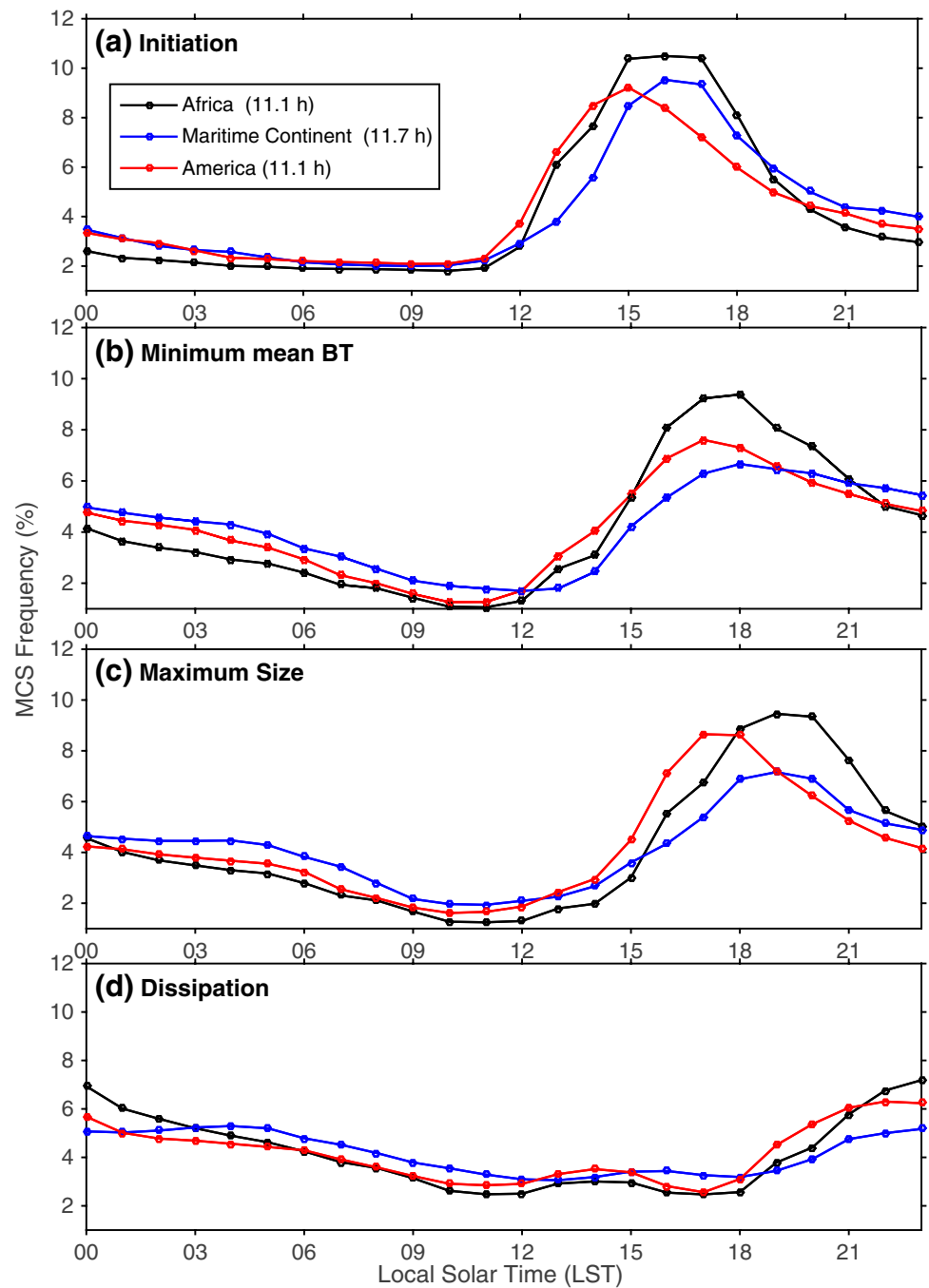
land counterparts except in South America, India and some part of tropical Africa. Another example is the diurnal cycle of MCS. The MCSs over land have a prominent diurnal cycle; initiating frequently in mid to late afternoon, reaching their maximum intensity and horizontal extent in the early evening, and dissipating around midnight. In contrast, oceanic MCSs have a much weaker diurnal cycle, with no preferred time for initiation. They generally reach their maximum intensity in the early morning. There are two

peaks in the spatial extent, with a primary peak in the afternoon and a secondary peak around the time of maximum intensity in the early morning. The preferred dissipation time is about an hour after the time of the primary peak in horizontal extent. The new MCS dataset also paves the way for improving the simulation of convective processes and the hydrological cycle in climate models. Finally, the MCS dataset presented here can be complemented with additional observations.

Fig. 8 Same as Fig. 6, but showing the five tropical oceanic areas

Appendix a: kalman Filter for MCS tracking

Using Kalman Filter for MCS tracking, we specify the location (x_t, y_t) to be the coordinate of the target MCS at time t, and $(\dot{x}_t, \dot{y}_t)$ the corresponding velocity. The state (s_t) of the moving MCS can be summarized by

$$s_t = \begin{pmatrix} x_t \\ y_t \\ \dot{x}_t \\ \dot{y}_t \end{pmatrix} \tag{3}$$

Its coordinate and velocity in the next time step is described by

$$s_{t+1} = As_t + q \tag{4}$$

where

$$A = \begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and $q \sim N(0, Q)$ is the Gaussian noise with a covariance of Q. We can also observe the coordinate of the target MCS with some measurement noise. The measured coordinate is represented by

$$e_t = \begin{pmatrix} x_t^e \\ y_t^e \end{pmatrix} \tag{5}$$

The relation between measured coordinate and real state is

$$e_t = Hs_t + r \tag{6}$$

where

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

and $r \sim N(0, R)$ is the Gaussian noise with covariance R.

In probabilistic terms, the overall model is

$$p(s_t|s_{t-1}) = N(s_t|As_{t-1}, Q) \tag{7}$$

$$p(e_t|s_t) = N(e_t|Hs_t, R) \tag{8}$$

In practice using Kalman Filter to track the target includes two steps: first predicting the next coordinate, and then updating model parameters based on observation.

The Kalman Filter prediction step

Assume the posterior distribution of time t-1 is Gaussian

$$p(s_t|e_{1:t-1}) = N(s_t|m_{t-1}, P_{t-1}) \tag{9}$$

Then the predicted state at time t is

$$
\begin{aligned}
&p(s_t|e_{1:t-1}) \\
&= \int p(s_t|s_{t-1})\, p(s_{t-1}|e_{1:t-1})\, ds_{t-1} \\
&= \int N(s_t|As_{t-1}, Q)\, N(s_{t-1}|m_{t-1}, P_{t-1}) \\
&= N(s_t|Am_{t-1}, AP_{t-1}A^T + Q) \\
&= N(s_t|m_t^-, P_t^-)
\end{aligned}
\tag{10}
$$

where

$$m_t^- = Am_{t-1}$$

and

$$P_t^- = AP_{t-1}A^T + Q$$

thus, we can get the predicted coordinate (x_t, y_t) and velocity $(\dot{x}_t, \dot{y}_t)$.

The Kalman Filter update step

Using the observation at time t, we can update the posterior distribution of coordinate (x_t, y_t) and velocity $(\dot{x}_t, \dot{y}_t)$ by

$$
\begin{aligned}
&p(s_t|e_t, e_{1:t-1}) \\
&= \frac{1}{Z} p(e_t|s_t) p(s_t|e_{1:t-1}) \\
&= N(s_t|m_t, P_t)
\end{aligned}
\tag{11}
$$

where

$$
\begin{aligned}
u_t &= HP_t^- H^T + R_t \\
v_k &= P_k^- H^T u_t^{-1} \\
m_t &= m_t^- + v_t[e_t - Hm_t^-] \\
P_t &= P_t^- - v_t u_t v_t^T
\end{aligned}
$$

which can be used recursively for the next time step.

Appendix b: algorithm implementation and parallel optimization

The MCSs tracking algorithm is implemented using MATLAB and performed using parallel computation. Thus, it is efficient in addressing massive quantities of satellite data and significantly reduces the calculation time. The motivation for optimizing computing performance is to efficiently explore various parameter settings for different detection needs and to handle the massive input datasets that can be obtained from either satellite observations or numerical simulations.

We can estimate the computational complexity of each main component in the MCS tracking process using big O notation. For a tracking process covering T consecutive time

steps, let N represent the average number of potential MCS tracks in each time step, and let A represent the average area of each potential MCS in terms of the number of pixels. Then, for each time step, the computational complexity of the overlapping tracking algorithm is $O(N^2 A \log A)$, that of the KF tracking algorithm is $O(N^2)$, and that of the Hungarian assigning algorithm is $O(N^3)$. Therefore, the computational complexity of the algorithm as a whole is $O(T \times (N^3 + N^2 A \log A))$.

Several computational performance optimization techniques have been used in our system implementation. First, we vectorize the MATLAB program to improve its single-thread performance. Second, we use parallelization with multiple threads during the detection stage to improve the single process performance. Third, and most importantly, we use a data parallelism strategy: for consecutive images covering T time steps, we use P processors to track MCSs simultaneously by slicing the whole-time period into P consecutive parts. For example, if we are tracking MCSs over 20 years, we can use 240 processors, and each processor tracks MCSs in one month of the dataset. This data parallelism is straightforward due to the temporally continuous nature of MCS tracks. The merging overhead is constant for each process and controllable according to the maximum possible time length of a MCS track. Thus, we can decrease the computational complexity on each processor to $O(T \times (N^3 + N^2 A \log A) \div P)$.

We implemented the system on a distributed computer cluster with a number of processors equal to the total number of tracking months. Only several hours are required to track all MCSs in the global satellite dataset from 1985 to 2008, depending on the different criteria chosen for the MCS identification. In addition, it is possible to further exploit the spatially continuous nature of MCS tracks to slice each image into several parts. These improvements have not been implemented in the current version of the system. The open-source MCS identification and tracking system that permits modification of the criteria for different research applications is available from https://doi.org/10.1594/PANGAEA.877914. All the codes used in the analysis presented in this paper and in the production of the figures can be obtained from the same website.

References

Adler RF, Huffman GJ, Chang A et al (2003) The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979-present). J Hydrometerol 4:1147–1167

Arkin PA (1989) The global precipitation climatology project. Adv Space Res 9:311–316. https://doi.org/10.1016/0273-1177(89)90178-6

Arnaud Y, Desbois M, Maizi J (1992) Automatic tracking and characterization of african convective systems on meteosat pictures. J Appl Meteorol 31:443–453. doi:https://doi.org/10.1175/1520-0450(1992)031<0443:ATACOA>2.0.CO;2

Bister MH (1996) Development of tropical cyclones from mesoscale convective systems. Bull N Y Acad Med 43:949–58

Boer ER, Ramanathan V (1997) Lagrangian approach for deriving cloud characteristics from satellite observations and its implications to cloud parameterization. J Geophys Res Atmos 102:21383–21399. https://doi.org/10.1029/97JD00930

Carvalho LMV, Jones C (2001) A satellite method to identify structural properties of mesoscale convective systems based on the maximum spatial correlation tracking technique (MASCOTTE). J Appl Meteorol 40:1683–1701. doi:https://doi.org/10.1175/1520-0450(2001)040<1683:ASMTIS>2.0.CO;2

Chen SS, Houze RA (2010) Diurnal variation and life-cycle of deep convective systems over the tropical pacific warm pool. Quart J R Meteorol Soc 123:357–388. https://doi.org/10.1002/qj.49712353806

Chen SS, Houze RAJ, Mapes BE (1996) Multiscale variability of deep convection in realation to large-scale circulation in TOGA COARE. J Atmos Sci 53:1380–1409

Dai A (2001) Global precipitation and thunderstorm frequencies. part II: diurnal variations. J Clim 14:1092–1111. doi:https://doi.org/10.1175/1520-0442(2001)014<1112:GPATFP>2.0.CO;2

Dias J, Tulich SN, Kiladis GN (2012) An object-based approach to assessing the organization of tropical convection. J Atmos Sci 69:2488–2504

Dixon M, Wiener G (1993) TITAN: thunderstorm identification, tracking, analysis, and nowcasting—a radar-based methodology. J Atmos Oceanic Technol 10:785–797

Dong W, Lin Y, Wright JS et al (2016) Summer rainfall over the southwestern Tibetan Plateau controlled by deep convection over the Indian subcontinent. Nat Commun 7:10925. https://doi.org/10.1038/ncomms10925

Fiolleau T, Roca R (2013a) An algorithm for the detection and tracking of tropical mesoscale convective systems using infrared images from geostationary satellite. IEEE Trans Geosci R Sens 51:4302–4315

Fiolleau T, Roca R (2013b) Composite life cycle of tropical mesoscale convective systems from geostationary and low earth orbit satellite observations: method and sampling considerations. Quart J R Meteorol Soc 139:941–953

Goyens C, Lauwaet D, Schröder M et al (2012) Tracking mesoscale convective systems in the sahel: relation between cloud parameters and precipitation. Int J Climatol 32:1921–1934

Gray WM (1998) The formation of tropical cyclones. Meteorol Atmos Phys 67:37–69

Hennon CC, Helms CN, Knapp KR, Bowen AR (2011) An objective algorithm for detecting and tracking tropical cloud clusters: implications for tropical cyclogenesis prediction. J Atmos Oceanic Technol 28:1007–1018

Hennon CC, Papin PP, Zarzar CM et al (2013) Tropical cloud cluster climatology, variability, and genesis productivity. J Clim 26:3046–3066

Hodges KI, Chappell DW, Robinson GJ, Yang G (1999) An improved algorithm for generating global window brightness temperatures from multiple satellite infrared imagery. J Atmos Oceanic Technol 17:1296–1312

Houze RAJ (1982) Cloud clusters and large-scale vertical motions in the tropics. J Meteorol Soc Jpn 60:396–410

Houze RAJ (2004) Mesoscale convective systems. Rev Geophys 42:125–133

Houze RAJ, Churchill DD (2010) Mesoscale organization and cloud microphysics in a bay of bengal depression. J Atmos Sci 44:1845–1868

Huffman GJ, Adler RF, Arkin P et al (1997) The global precipitation climatology project (gpcp) combined precipitation dataset. Bull Am Meteor Soc 78:5–20

Huffman GJ, Adler RF, Bolvin DT, Nelkin EJ (2010) The TRMM multi-satellite precipitation analysis (TMPA). Satellite rainfall applications for surface hydrology. Springer, The Netherlands, pp 3–22

Inoue T, Vila D, Rajendran K et al (2009) Life cycle of deep convective systems over the eastern tropical Pacific observed by TRMM and GOES-W. J Meteorol Soc Jpn Ser II 87:381–391. https://doi.org/10.2151/jmsj.87A.381

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

Johnson JT, MacKeen PL, Witt A et al (1998) The storm cell identification and tracking algorithm: an enhanced WSR-88D algorithm. Weather Forecast 13:263–276

Kolios S, Feidas H (2010) A warm season climatology of mesoscale convective systems in the Mediterranean basin using satellite data. Theor Appl Climatol 102:29–42

Kouadio YK, Machado LAT, Servain J (2010) Tropical Atlantic Hurricanes, Easterly Waves, and West African Mesoscale Convective Systems. Adv Meteorol. https://doi.org/10.1155/2010/284503

Laing AG, Fritsch JM (1997) The global population of mesoscale convective complexes. Quart J R Meteorol Soc 123:389–405

Laing AG, Fritsch JM (2000) The large-scale environments of the global populations of mesoscale convective complexes. Mon Weather Rev 128:2756–2776

Liu C, Zipser EJ (2008) Diurnal cycles of precipitation, clouds, and lightning in the tropics from 9 years of TRMM observations. Geophys Res Lett. https://doi.org/10.1029/2007GL032437

Machado LAT, Rossow WB (1993) Structural characteristics and radiative properties of tropical cloud clusters. Mon Weather Rev 121:3234–3260

Machado LAT, Rossow WB, Guedes RL, Walker AW (1998) Life cycle variations of mesoscale convective systems over the Americas. Mon Weather Rev 126:1630–1654

Maddox RA (1980) Meoscale convective complexes. Bull Am Meteor Soc 61:1374–1387

Mapes BE, Houze RA Jr (1993) Cloud clusters and superclusters over the oceanic warm pool. Mon Weather Rev 121:1398–1416

Mathon V, Laurent H (2001) Life cycle of Sahelian mesoscale convective cloud systems. Quarter J R Meteorol Soc 127:377–406

Mohr KI, Zipser EJ (1996) Mesoscale convective systems defined by their 85-GHz ice scattering signature: size and intensity comparison over tropical oceans and continents. Mon Weather Rev 124:2417–2437

Moncrieff MW (2010) The multiscale organization of moist convection and the intersection of weather and climate. Why Does Climate Vary? American Geophysical Union, Climate Dynamics. https://doi.org/10.1029/2008GM000838 Washington, DC

Morel C, Senesi S (2002) A climatology of mesoscale convective systems over Europe using satellite infrared imagery. I: methodology. Quart J R Meteorol Soc 128:1953–1971

Nakazawa T (1988) Tropical super clusters within intraseasonal variations over the western Pacific. J Meteorol Soc Jpn 66:823–839

Nguyen H, Duvel JP (2008) Synoptic wave perturbations and convective systems over equatorial Africa. J Clim 21:6372–6388

Pope M, Jakob C, Reeder MJ (2008) Convective system of the north Australian monsoon. J Clim 21:5091–5112

Reid D (1979) An algorithm for tracking multiple targets. IEEE Trans Automatic Control 24:843–854

Roca R, Viollier M, Picon L, Desbois M (2002) A multisatellite analysis of deep convection and its moist environment over the Indian Ocean during the winter monsoon. J Geophys Res Atmos 107(D19). https://doi.org/10.1029/2000JD000040

Rossow WB, Schiffer RA (1999) Advances in understanding clouds from ISCCP. Bull Am Meteor Soc 80:2261–2287

Schröder M, König M, Schmetz J (2009) Deep convection observed by the Spinning Enhanced Visible and Infrared Imager on board Meteosat 8: spatial distribution and temporal evolution over Africa in summer and winter 2006. J Geophys Res: Atmos, 114(D5). https://doi.org/10.1029/2008JD010653

Tan J, Jakob C, Rossow WB, Tselioudis G (2015) Increases in tropical rainfall driven by changes in frequency of organized deep convection. Nature 519:451–454. https://doi.org/10.1038/nature14339

Teng HF, Lee CS, Hsu HH (2014) Influence of ENSO on formation of tropical cloud clusters and their development into tropical cyclones in the western North Pacific. Geophys Res Lett 41:9120–9126

Tsakraklides G, Evans JL (2003) Global and regional diurnal variations of organized convection. J Clim 16:1562–1572

Wilcox EM, Ramanathan V (2001) Scale dependence of the thermodynamic forcing of tropical monsoon clouds: results from TRMM observations. J Clim 14:1511–1524

Williams M, Houze RA Jr (1987) Satellite-observed characteristics of winter monsoon cloud clusters. Mon Weather Rev 115:505–519

Xing J, Ai H, Lao S (2009) Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. 2009 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2009.5206745

Xu KM, Wong T, Wielicki BA, Parker L, Eitzen ZA (2005) Statistical analyses of satellite cloud object data from CERES. Part I: methodology and preliminary results of the 1998 El Niño/2000 La Niña. J Clim 18:2497–2514

Yang GY, Slingo J (2001) The diurnal cycle in the tropics. Mon Weather Rev 129:784–801

Yuan J, Houze RA Jr (2010) Global variability of mesoscale convective system anvil structure from A-train satellite data. J Clim 23:5864–5888

Zhang GJ, Zurovac-Jevtic D, Boer ER (1999) Spatial characteristics of the tropical cloud systems: comparison between model simulation and satellite observations. Tellus A: Dyn Meteorol Oceanogr 51:922–936. https://doi.org/10.1034/j.1600-0870.1999.00026.x