Preferred Intra-seasonal Circulation Patterns of the Indian Summer Monsoon and Active-Break Cycles

David Martin Straus (dstraus@gmu.edu)
George Mason University https://orcid.org/0000-0003-4463-9167

Research Article

Keywords: Indian monsoon , active and break cycles and Circulation regimes , boreal summer intra-seasonal oscillation

Posted Date: June 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-541464/v1

Preferred Intra-seasonal Circulation Patterns of the Indian Summer Monsoon and Active-Break Cycles

A new view of the active-break cycle

David M. Straus

Received: date / Accepted: date

Abstract Intra-Seasonal circulation regimes are identified from a cluster analysis of 5-day mean (pentad) anomaly fields of 850 hPa horizontal winds (u, v) from the ERA-Interim reanalysis for the boreal summer season (120 days starting 01June for the years 1979 - 2018) over the broad Indian region $(50^o - 100^oE; 5^oS - 35^oN)$. The anomalies are formed with respect to a parabolic (in time) seasonal cycle computed separately for each year, thus filtering out periods of greater than 240 days. The k-means method was applied in the phase space of the leading 6 (12) principal component modes, which explain 65% (78%) of the space-time variance, yielding k clusters. The degree of clustering is significant when compared to synthetic data sets for any value of $k > 3$.

Center for Ocean-Land-Atmosphere Studies, AOES Department, George Mason University, Fairfax, VA, 22030, USA

ORCID 0000-0003-4463-9167

E-mail: dstraus@gmu.edu

The transition matrices for $k = 4$ and $k = 5$ establish that the system is most likely to stay in the same cluster from one pentad to the next, but that the significant transitions (with 95% confidence level using a modified bootstrap method) form a cycle. The similarity between the cycle as depicted from 4 or 5 clusters is established by composites of 850 hPa $(u, v,)$, 200 hPa divergence, 500 hPa vorticity and vertical pressure velocity, and daily rainfall over India: Strong convection (with large positive divergence and vorticity) over the subtropical Indian Ocean, moves to the central Bay of Bengal and over central India, then subsequently to the northern Bay of Bengal and west Bengal, and then further north into the Himalayas. The Indian rainfall composites show a similar cycle. The phases in which strong convection is seen over central and northern India are seen for about 60% of the time for both $k = 4$ and $k = 5$ analyses. However the 4 cluster analysis also shows a preferred transition in which the convection moves equatorward from central India.

The number of complete cycles (including a return to the starting cluster) found in the 40 years of data is 7 in the 4-cluster analysis, while the number of times the system undergoes four (three) consecutive legs of the cycle is 16 (31). Fewer instances of complete cycles are found for 5 clusters (only 3), but sequences of five, four and three consecutive legs occur 10, 11 and 28 times respectively.

Composites of the tropics-wide vertically integrated diabatic heating (estimated from ERA5 reanalyses) reproduce the characteristics of the boreal summer intra-seasonal oscillation, with northwest-to-southeast oriented bands

of heating moving northward from the tropical Indian Ocean into the subtropics.

This depiction of the active-break cycle is particularly useful for diagnosing the cycle in short-range forecasts: as long as pentad anomalies can be formed, they can be assigned to one of the observed clusters described in this paper without the need for further time-filtering.

Keywords Indian monsoon · active and break cycles and Circulation regimes · boreal summer intra-seasonal oscillation

1 Introduction

The atmospheric circulation associated with the Indian monsoon has a very rich structure on intra-seasonal time scales. Fluctuations in the wind field over the broad Indian region strongly modulate the occurrence of synoptic systems and the resulting rainfall, so that prolonged episodes (several days to one week) of strong rainfall (and of no rainfall) are observed over central India and surrounding regions (Murakami et al, 1984; Annamalai and Slingo, 2001; Rajeevan et al, 2010) Since this well-known "active-break cycle" has a large impact on agriculture (Prasanna, 2020) the ability of weather and climate models to forecast this cycle (and the circulation in which it is embedded) is clearly of concern (Pattanaik et al, 2020). These active and break periods are usually defined locally in time, as a function of time lags with respect to extreme regionally-averaged rainfall, as in Rajeevan et al (2010).

This cycle is embedded in a complex of larger scale intra-seasonal fluctuations which extend over the broad Indian Ocean - western Pacific region. There is general agreement that two broad-band oscillations, the 30-60 day (boreal summer intra-seasonal oscillation, or BSISO (Murakami et al, 1984; Lau and Chan, 1986; Annamalai and Sperber, 2005; Hazra and Krishnamurthy, 2014) and the 10-20 day (quasi bi-weekly, or QBW) oscillation (Krishnamurthi and Bhalme, 1976; Krishnamurthi and Ardunay, 1980; Chatterjee and Goswami, 2004) dominate this variability. In order to identify these oscillations and determine their character, a number of complex spatio-termporal spectral analysis techniques have been used, including extended empirical orthogonal functions (Suhas et al, 2013; Sahai et al, 2015), multi-channel singular spectrum analysis (Krishnamurthy and Shukla, 2000, 2007, 2008), principal oscillation patterns (Annamalai and Slingo, 2001) and cyclo-stationary empirical orthogonal functions (Annamalai and Sperber, 2005).

Comparisons of the BSISO in simulations (Neena et al, 2017) and forecasts (Lee et al, 2015) with those from reanalyses have been carried out in the EEOF framework, although this process is somewhat cumbersome.

In contrast to the local index-based approach used to diagnose the active-break cycle, and the complex mode techniques used to isolate the BSISO and QBW, this article suggests a focus on the spatial variability of the circulation, and in particular on the preferred occurrence of certain patterns, also called "regimes". This approach can be implemented via the general technique of cluster analysis, which while very often used in studies of mid-latitude intra-

seasonal variability (see Straus et al, 2017), has not been widely applied to the Indian monsoon circulation. Chu et al (2017) use self-organizing maps (a form of cluster analysis) to characterize the intra-seasonal variability of outgoing long-wave radiation and low-level zonal winds in the very broad region extending from East Africa to the western Pacific, and are able to identify both stationary and propagating patterns. The purpose of this article is to explore the structure of preferred circulation patterns over the Indian region $(50^o - 100^o E, 0^o - 35^o N)$ and to relate transitions between them to the active-break cycle over India. This is meant to provide a framework in which the model forecasts (and simulations) can be evaluated with regard to the both the circulation regimes and, via their transitions, to the active-break cycle. Recent interest in the fidelity of predictions and simulations of the active-break cycle is evidenced in the work of Neena et al (2017) and Pattanaik et al (2020).

Cluster analysis partitions every state of the atmosphere (whether daily or time-filtered anomalies) into a small number of clusters on the basis of a common structure. The average of all states in a cluster defines a *centroid* (also called a circulation regime). The complex evolution of the circulation is "coarse-grained" and so is described by the residence time within a regime and the probability of transition into another regime. Active-break cycles within this approach emerge as preferred transition paths between regimes. A fuller picture of the cycle and its components is available by compositing a number of variables over states assigned to each regime.

In section 2, the data set and the analysis methods are described. Methods of estimating statistical significance are described in section 3. Section 4 gives a number of composite fields over the broad South Asian region for the characteristic states (centroids) from each cluster for a 4-cluster analysis and a 5-cluster analysis, while Section 5 describes the composite diabatic heating over the global tropical belt. The frequency of occurrence and lifetime of the regimes, as well as evidence for preferred transition paths between them, are given in Section 6. A Discussion is given in Section 7, and the Conclusions in Section 8. The Appendix describes some technical matters regarding cluster analysis.

2 Data Sets and Analysis Methods

2.1 Data

Once-daily fields of the horizontal wind (u, v) at 200, 500 and 850 hPa and the vertical pressure velocity (ω) at 500 hPa were obtained from the ERA-Interim reanalysis (Dee and coauthors, 2011) for the 120-day period starting 01June for the 40 years 1979 through 2018. These data were interpolated from the original (512×256) (lon x lat) Gaussian grid to a (128×64) (lon x lat) Gaussian grid. 24 consecutive five-day means (pentads) of the winds, divergence (D), vorticity (ζ) and vertical pressure velocity (ω) were formed for use in the cluster analysis and composites.

Additionally, four-times daily data of the horizontal winds (u, v), temperature T, and pressure vertical velocity ω were obtained from the ERA5 reanalysis (Hersbach and co authors, 2020) at 37 pressure levels for the purpose of estimating the diabatic heating. These fields were interpolated from the original 0.25^o grid to a (256×128) (lon x lat) Gaussian grid prior to calculating the heating. The diabatic heating, whose calculation is described in subsection 2.3 was vertically integrated into nine layers, and pentads formed.

Daily gridded precipitation data for the months of June through September for the years 1979 through 2018 were obtained from the Indian Meteorological Department over the Indian land region (Pai et al, 2014) on a 0.25^o (lon x lat) grid. This data set has been updated regularly. The daily data were averaged into pentads, as described above.

For all the fields described above, the seasonal cycle was estimated by fitting the seasonal evolution of the pentads at each grid point with a parabola, and averaging the parabola over all 40 years (Straus, 1983). Anomalies are obtained by removing this seasonal cycle.

2.2 Cluster analysis

As a first step in the cluster analysis, a principal component analysis was applied to the vector field of (u, v) at 850 hPa over the wide Indian region $(50.63^o - 98.44^oE, 1.40S^o - 34.88^oN)$. The leading 6 (12) modes explain 65% (79%) of the total space-time variance. Since the variance explained drops off rapidly with mode number, we retain only 6 modes in the cluster calculations

to be described , although we have verified that the composite circulation and rainfall for each cluster are nearly unchanged when 12 modes are retained.

The k-means cluster analysis (Deday and Simon, 1976; Desbois et al, 1982; Michelangeli et al, 1995; Straus et al, 2017) was applied in the 6-dimensional PC phase space, with the goal of identifying non-Gaussian behavior, hence preferred or unpreferred regions of phase space (Toth, 1993). This iterative method assigns each pentad in the phase space uniquely to one of k clusters (or groups of points in phase space), where k is chosen a priori. The cluster centroid is defined by the average PC coordinates of all states in the cluster. The clusters are chosen in such a way as to maximize the ratio R of the sum of squares of the PC coordinates of the centroid (weighted by the cluster population) Δ, to the intra-cluster variance S. This variance ratio measures the degree of clustering, and is used both to monitor the convergence of the algorithm, and in the assessment of its significance. Further details of the algorithm are given in Straus et al (2017). Choice of the value of k depends on the outcome of the statistical significance of the clustering, and robustness to randomly chosen half-length samples, both of which are discussed in Section 3

The cluster analysis yields an assignment of each 5-day period to one of k regimes.Composites of the full (not truncated) fields of anomalies of low-level (850 hPa) circulation u, v, ω, D, ζ and rainfall over each regime are examined. The frequency of transitions between regimes, and the possible sequences of

regimes, together give a coarse-grained approximation to trajectories in phase space.

2.3 Diabatic heating

The diabatic heating rate Q is estimated from the four-times daily ERA5 data at 37 pressure levels as a residual in the thermodynamic equation. The heating is evaluated on the (256×128) (lon x lat) Gaussian grid, using a residual method as reported in Swenson and Straus (2021): The heating rate Q is obtained from:

$$c_p \Pi \left(\frac{\partial \theta}{\partial t} + \boldsymbol{\nabla} \cdot (\mathbf{v}\theta) - \theta(\boldsymbol{\nabla} \cdot \mathbf{v}) + \omega \frac{\partial \theta}{\partial p} \right) = Q. \tag{1}$$

Here θ is potential temperature, p is pressure, $\omega = \frac{dp}{dt}$, c_p the specific heat at constant pressure per unit mass, $\mathbf{v}$ the horizontal velocity field, and $\Pi = \left(\frac{p_0}{p}\right)^{\kappa}$, with $p_0 = 1000$ hPa and $\kappa = \frac{R}{c_p}$, R being the gas constant. Q is transformed from the grid representation to spherical harmonics, truncated to a triangular (T159) representation, and transformed back to the grid.

3 Significance of Results

The assessment of significance of the cluster analysis is carried out with the use of synthetic (also called surrogate) data sets. Each synthetic data set has the same number of states as the physical data set, but the time series of each PC is replaced by a stochastically generated Gaussian time series, constructed so

that for each PC the lag-correlation statistics are close to the observed statistics for that mode. *By construction the synthetic time series of different PCs are statistically independent of each other.* This independence of the Gaussian synthetic PCs ensures that any deviations from a multi-normal distribution are due only to sampling error. The method of construction of these stochastic time series is based on a variant of the random-phase method of Christiansen (2007), as detailed in Straus (2010).

We apply the cluster analysis to 100 such synthetic data sets; the percentage of times for which the variance ratio R is less than that of the real data gives the confidence level. We have carried out this analysis for a range of k, from 2 to 6, and find that for k=4,5, or 6 the confidence level is 99%, indicating very little chance that the degree of clustering in the original data is due to sampling error. Common to many methods of clustering, the final choice for the value of k is to some extent arbitrary (Christiansen, 2007). We used both k = 4 and 5 in our analysis, but focused on the choice of 5 clusters, as explained in Section 4. Christiansen (2007) showed that the k-means method can incorrectly identify multiple clusters ($k > 1$) when presented with synthetic bivariate data drawn from skewed and platykurtic uni-modal distributions. We note that the skewness of the two leading PCs used for the cluster analysis is quite small (-0.143 and 0.125).

In order to test the robustness of the cluster centroids to sub-sampling, we repeated the cluster analysis using 1000 randomly chosen half-length data sets. For each such half-length dataset, the variance ratio R was computed and the

resulting histogram displayed as a probability distribution (not shown). When normalized by its value for the entire data set, the distribution is sharply peaked at a value near 1.01 for $k = 4$ and $k = 5$, and at a slightly higher value for $k = 6$. Another measure of the agreement between the clusters in the half-length data sets and those in the full data set is the pattern correlation between the cluster centroids. For each half-length data set and value of k, each cluster centroid was matched to one of the k centroids in the full data set based on pattern correlation. The k values of correlation were then averaged (using the Fisher transformation) to assign a single value to that half-length data set. The probability distribution of the average value is shown for various values of k in Figure 9 in the Appendix. In all cases almost the entire distribution lies above correlation values of 0.8, with a small tail below this for $k = 4$ and $k = 6$.

Preferred transition paths between the clusters are assessed with the use of the transition matrix, whose i^{th} row and j^{th} column gives the number of times the system makes a transition between cluster i and cluster j. The off-diagonal elements define transitions, whose statistical significance is determined by comparison to the results of a large number (here 1000) of scrambled data sets. In each scrambled data set the cluster assignments are resampled using a bootstrap with replacement technique: the cluster assignment of each state (pentad p of year x) is replaced with that of a randomly chosen pentad q and year y. If a block of n pentads $q + 0, q + 1, ...q + n - 1$ of year y are

all assigned to the same cluster, then the pentads associated with the entire block of n days are assigned to pentads $p, p+1, ..., p+n-1$ of year x.

The significance of the composite anomalies of the vorticity, divergence, vertical velocity and diabatic heating were assessed with a two-sided t-test. Significance is indicated for all anomalies exceeding 95% confidence.

4 Circulation Regimes

Composites of a number of anomaly fields are presented here for the cases of $k = 4$ and $k = 5$. The 850 hPa horizontal wind field (u, v), shown using the full anomalies (not truncated in PC-space), is helpful in indicating the location of anomalies in the Monsoon trough. This is also indicated by the 500 hPa vorticity ζ. Both the 200 hPa divergence D and the 500 hPa vertical pressure velocity ω indicate regions of anomalous upward motion associated with convection. Finally, we show the rainfall anomalies (available only over the Indian land region).

4.1 Four regimes

Composites for $k = 4$ are shown in Figures 1, 2, and 3. The composite figures are ordered (from (a) through (d)) to correspond to a preferred cyclical transition path $(a) \rightarrow (b) \rightarrow (c) \rightarrow (d) \rightarrow (a)$ which will be discussed in Section 6.

In Figure 1, Cluster (a) shows low-level cyclonic anomalies over the equatorial Indian Ocean associated with enhanced upper-level divergence (indicative

of deep convection), and a suppression of the Monsoon trough over India itself (with Easterly anomalies) associated with a negative divergence anomaly in the Bay of Bengal. The mid-level ζ and ω anomalies for the same cluster, shown in Figure 2a, indicate enhanced vorticity and upward motion over the equatorial Indian Ocean, with suppressed vorticity and downward motion over India and the Bay of Bengal. Consistent with this, the corresponding rainfall anomalies over India, shown in Figure 3 show suppressed rainfall over the Western Ghats and over the Indo-Gangetic plain.

Cluster (b) presents a sharp contrast: the Monsoon trough is now enhanced, with positive upper-level divergence anomalies over the Bay of Bengal and central India (Figure 1b), associated with enhanced mid-level vorticity and upward motion (Figure 2b) and strongly enhanced rainfall over the Western Ghats and over central India (Figure 3b). The enhanced rainfall over the west coast is consistent with the enhanced on-shore low-level winds, seen in Figure 3.

In cluster (c), the enhanced cyclonic low-level anomaly (and the accompanying upper-level divergence) are now located further to the northeast compared to their positions in cluster (b), and are also weaker (compare Figures 1b) and 1c). Enhanced mid-level vorticity (along with rising motion) is seen over the northern Bay of Bengal, with a large area of sinking motion and decreased vorticity further to the southwest (Figure 2c). As expected from the circulation anomalies, the rainfall is modestly enhanced over northeastern India but is suppressed over central India (Figure 3c).

Finally, in cluster (d) the enhanced low-level cyclonic anomaly and associated upper level divergence (and rising motion) has moved further northward to the eastern Himalayan mountain range (see Figures 1d) and 2d), while anomalous sinking motion and negative vorticity are seen over the Bay of Bengal and across much of India into the Arabian Sea. Over the tropical Indian Ocean and southern India, increased convection, with upper level divergence and rising motion are apparent in Figures 1d) and 2d). The rainfall anomalies (Figure 3d) are consistent with the circulation, with suppressed rainfall over much of central India, but enhanced rainfall in the Himalayas and over southern India.

4.2 Five regimes

Composites for $k = 5$ are shown in Figures 4, 5, and 6. The composite figures are ordered (from (a) through (e)) to correspond to a preferred cyclical transition path $(a) \rightarrow (b) \rightarrow (c) \rightarrow (d) \rightarrow (e) \rightarrow (a)$ which will be discussed in Section 6.

The low-level circulation and upper-level divergence anomalies for cluster (a) shown in Figure 4a are very similar to cluster (a) for $k = 4$ (compare Figure 1a), although the center of upper-level divergence over the tropics is displaced southwards slightly in Figure 4a, while a small area of positive divergence is seen over the eastern Himalayas. These observations also hold for the regions of enhanced mid-level vorticity and ω (compare Figures 5a and 2a).

Cluster (b) for $k = 5$, showing enhanced upper-level divergence, cyclonic anomalies and rising motion in both the Bay of Bengal and the Arabian Sea, with modest subsidence further to the northeast (Figures 4b and 5b) has no obvious counterpart in the results for $k = 4$. This cluster appears as a (new) transition state between cluster (a) and cluster (c). The latter (Figures 4c and 5c) shows strong divergence, upward motion and enhanced midl-level vorticity over the Bay of Bengal and across central India into the Arabian Sea, where this feature extends southwestward into the tropics. Regions of subsidence are seen in the tropics (east of 65^o) E and further north over the eastern Himalayas. Note that this cluster is very similar to clusters (b) for $k = 4$ seen in Figure 1b and 2b. Similarly, cluster (d) for $k = 5$ (Figures 4d and 5d), with positive upper-level divergence, rising motion and positive mid-level vorticity in the northern Bay of Bengal and the opposite circulation over a broad band south of 15^o N, is very similar to cluster (c) for $k = 4$ (Figures 1c and 2c). Finally, cluster (e) for $k = 5$ (Figures 4e and 5e) shows suppression of convection (downward motion) over a band extending from Myanmar westward over central India, and into the Arabia Sea, with some rising motion to the south of this, over the extreme northeastern portion of the domain. The patterns are similar to those of cluster (d) for $k = 4$ (Figures 1d and 2d); however there is a clear northward shift in the $k = 4$ cluster (d) compared to the $k = 5$ cluster (e).

The cluster (a) rainfall composite for $k = 5$ shown in Figure 6a indicates suppressed rainfall over central India and the western Ghats with enhanced

rainfall over the Himalayas and south central India, in rough agreement with cluster (a) for $k = 4$ (Figure 3a), although the positive anomalies in the $k = 5$ cluster are clearly stronger. Cluster (b) (Figure 6b) shows a weak enhancement of rainfall over central India and only a small remnant of the dryness which has moved northeastward (compared to cluster (a)). In cluster (c), strong positive anomalies of rainfall appear over central India and the western Ghats, with a suppression seen over the eastern Himalayas (Figure 6c), very similar to cluster (b) for $k = 4$ (Figure 3b). The positive rainfall anomalies weaken and move northeastward in cluster (d) (Figure 6d), matching cluster (c) for $k = 4$ (Figure 3c). At the end of the sequence, cluster e (Figure 6e) shows widespread suppression of rainfall over both central India and the western Ghats, and a small area of enhanced rainfall over the eastern Himalayas, a pattern very similar to that seen for $k = 4$ in cluster (d) (Figure 3d).

5 Diabatic Heating

Composites of the anomalies of diabatic heating (vertically integrated from 1000 to 50 hPa) are given for the four clusters ($k = 4$) over the global tropical belt in Figure 7. Again we shall discuss the clusters in the framework of a cyclical transition path $(a) \to (b) \to (c) \to (d) \to (a)$.

In cluster (a), prominent tropical Indian Ocean heating anomalies to the south of the equator are connected to a northwest-to-southeast ("tilted") band of heating extending from the Arabian Sea, across southern India into the southern Bay of Bengal and further east over the Maritime continent. A par-

allel, similarly oriented band of cooling lies about 15 degrees further to the north. Heating anomalies are also seen further north over the Himalayas. The tilted band of heating moves northward in cluster (b) so that it covers central and northern India, and extends southeastwards into the tropical western Pacific. The Himalayan area of heating in cluster (a) is replaced by cooling in cluster (b), as is the tropical Indian Ocean heating. The tilted heating bands moves further northward in cluster (c), reaching the Himalayas in cluster (d), in which heating is again seen in the tropical Indian Ocean. Generally bands of heating are separated by bands of cooling. Of interest is the alternating areas of cooling (in clusters (a) and (b)) and heating (clusters (c) and (d)) over the eastern tropical Pacific.

Heating anomalies for five clusters are shown in Figure 8, labeled (a) through (e) as in subsection 4.2. Cluster (a) is very similar to cluster (a) for the 4 cluster case (Figure 7), while cluster (d) and (e) for 5 clusters are very similar to clusters (c) and (d) for 4 clusters, respectively. It is clusters (b) and (c) in the 5-cluster case which are merged into cluster (b) in the 4-cluster case (compare Figures 8b and 7b). The far eastern Pacific cooling noted in the 4-cluster case is seen in clusters (a) and (b) in Figure 8, whereas the previously noted eastern Pacific heating is seen in clusters (d) and (e).

6 Regime lifetimes and Transition Paths

6.1 Regime lifetimes

The average lifetime in a cluster (regime) is given for each of the $k = 4$ clusters in Table 1 and for each of the $k = 5$ clusters in table 2, in both pentads and the corresponding number of days. Also given is the total time the system resides in each cluster. The average regime lifetime for $k = 4$ is 8.6 days, while that for $k = 5$ is 7.8 days. For $k = 4$ the system spends 58% of time in regimes (b) and (c), the regimes which show rising motion and upper-level divergence over the Bay of Bengal as in Figures 1b and 1c. The remaining 40% is divided nearly evenly between the regime with divergence closer to the tropics or over the Himalayas (Figures 1a and 1d.) Considering 5 clusters, the system spends 61% of time in regimes (b)-(d) with upper-level divergence over the Bay of Bengal (Figures 4b through 4d). The remaining two regimes (a) and (e) with rising motion to the north and south of India (Figures 4a and 4e) are occupied for the remaining 39% of the time.

6.2 Regime transitions

Table 3 presents the total number of transitions between individual clusters for $k = 4$. The number of transitions between cluster i and cluster j is given in the i^{th} row and jth column. The rows and columns are labeled (a) through (d) to correspond to the composites of Figures 1, 2 and 3. The diagonal elements give the number of pentads for which the system stayed in the same cluster, and

this is largest for clusters (b) and (c). The off-diagonal elements are generally smaller, and give the number of transitions. Bold values are significant at the 95% confidence level using the boot-strap method detailed in Section 3. Significant transitions include the complete cycle: $(a) \rightarrow (b)$, $(b) \rightarrow (c)$, $(c) \rightarrow (d)$, and $(d) \rightarrow (a)$. In addition, the number of transitions from cluster $(c) \rightarrow (b)$ is not only significant, but is also larger than the $(c) \rightarrow (d)$ leg of the complete cycle. Table 4 shows the number of episodes during which 3, 4 or 5 legs of the cycle occur consecutively, and the average length of each episode. Such episodes are counted as long as residence in one of the clusters is followed (after one or more pentads) by residence in the next cluster in the cycle. All possible sequences within the entire cycle are counted; for example the entry in the table for 3 legs includes all sequence $(a) \rightarrow (b) \rightarrow (c)$, $(b) \rightarrow (c) \rightarrow (d)$, $(c) \rightarrow (d) \rightarrow (a)$, and $(d) \rightarrow (a) \rightarrow (b)$.

The entire cycle from (a) through (d) back to (a) occurs only 7 times in the 40 years; 3 consecutive legs occur on average for over three quarters of the years. The average cycle length for the entire cycle is about 48 days.

6.3 Five clusters

Table 5 presents the total number of transitions between individual clusters for $k = 5$, in the same form as Table 3. The *only* significant transitions encompass the complete cycle: $(a) \rightarrow (b)$, $(b) \rightarrow (c)$, $(c) \rightarrow (d)$, and $(d) \rightarrow (e)$, and $(e) \rightarrow (a)$. Table 6 shows the number of episodes during which 3, 4, 5 or 6 legs

of the cycle occur consecutively, and the average length of each episode. The episodes are counted in a way analogous to those in Table 4.

The entire cycle from (a) through (e) back to (a) occurs only 3 times in the 40 years, with an average length of 42 days. Subsets of the cycle occur increasingly more often as the number of required legs of the cycle decreases; 3 legs occur 70% of the time, taking about 23 days.

7 Discussion

7.1 Comparison of 4 and 5 clusters

In general, the cyclical sequence of states for $k = 5$ from $(a) \rightarrow (b) \rightarrow (c) \rightarrow (d) \rightarrow (e) \rightarrow (a)$ is very similar to the sequence for $k = 4$ of $(a) \rightarrow (b) \rightarrow (c) \rightarrow (d) \rightarrow (a)$. The states in the $k = 4$ sequence seem to occur at slightly earlier times than the corresponding ones in the $k = 5$ sequence. For example, the appearance of rising motion and upper-level divergence over the Himalayas appear at the "end" of the cycle using 4 clusters, but appear both at the "end" of the $k = 5$ cycle and the "beginning" of the next one. Thus we see slightly stronger rising motion over the Himalayas in Figure 5a compared to 2a, as well as the weaker rising motion over the equatorial Indian Ocean in in Figure 5e compared to 2d. Remnants of positive rainfall anomalies over the Himalayas and south India are seen in the $k = 5$ cluster (a) composite(Figure 6a) which appear at end of the "previous" cycle for the $k = 4$ cluster (d) 3d).

418 7.2 Relationship to active / break episodes and intra-seasonal oscillations

419 The sequences of rainfall composites presented here can be compared to com-
420 posites of the active and break episodes presented by Rajeevan et al (2010)
421 (hereafter R). These authors define the dates of these isolated episodes on the
422 basis of rainfall anomalies averaged over the Monsoon core region exceeding
423 $\pm 1\sigma$ (standard deviation) for 3 consecutive days, using a high resolution data
424 set of precipitation over India. The active phase lag-0 composite of R, showing
425 large positive rainfall anomalies over the Monsoon core region of India, over
426 the western Ghats, along with suppressed rainfall over southern India and
427 to the northeast over the Himalayas, is very similar to cluster (c) for $k = 5$
428 (Figure 6c), as well as with cluster (b) for $k = 4$ (Figurefig:rf4b). The break
429 composite presented in R shows completely the reverse pattern, similar to the
430 $k = 5$ cluster (a) (and to some extent cluster (e)) in Figure 6, and cluster (d)
431 for $k = 4$ (Figure 3d). It should be noted that the longer life-span of break
432 phases compared to active phases found by R is not reflected in the average
433 persistence times seen for the clusters. For example, table 2 shows that for
434 $k = 5$ cluster (c) (active phase) has the same average persistence (7.8 days) as
435 clusters (a) and (e), which encompass the break phase.

436 Krishnamurthy and Shukla (Krishnamurthy and Shukla, 2008), herafter
437 KS8, show rainfall composites based on 8 phases of both the 45-day oscillation
438 and 28-day oscillation obtained from application of the MSSA technique to
439 observed outgoing long-wave radiation. The composites for phases 7 and 8
440 of the 45-day oscillation (from Figure 4 of KS8), with large negative OLR

over India, correspond well to the circulation in regime (c) for $k = 5$ shown in Figures 4 and 5. The opposite phases of KS8 (1 and 2) correspond to the circulation in regime (a) in Figures 4 and 5. The correspondence of our results with the 28-day oscillation of KS8 is less clear. This oscillation shows a dipolar structure (in latitude) of OLR over the Indian longitudes, in contrast to the more tripolar patterns seen in both the 45-day oscillation and our cluster composites. The association of our results with the 45-day oscillation is consistent with the average length of time it takes for the entire preferred sequence of $k = 4$ or $k = 5$ clusters to complete, 42 - 48 days, with partial sequences completing in about 34 days (Tables 4 and 6).

8 Summary and Conclusions

The cluster analysis presented here categorizes the evolution of the full circulation in time, rather than to focus on only those components which are oscillatory. The preferred transition paths between the circulation regimes clearly reflect aspects of both the active/break episodes and the 45-day oscillation. Yet one of the significant transitions (from cluster (c) to (b) in Figures 2 and 3) is characterized by *equatorward* propagation, and there are a number of other (not statistically significant) transitions between clusters. Such transitions occur during periods when regular northward propagation is not seen.

The evaluation of forecast and reforecast anomalies using the clustering approach presented has a number of advantages, including: (i) the focus on large-scale circulation patterns that are frequently seen in reanalyses; (ii) a

natural way to describe the components of the active/break cycle in forecasts, and (iii) its applicability to short forecasts, for which application of time-filtering is not an option. The assignment of each forecast pentad anomaly to one of the observed regimes can be done on the basis of pattern correlation or mean-squared error, and leads to a coarse-grained representation of the (re)-forecast evolution. Ensembles of such (re)-forecasts can then meaningfully summarized in terms of the evolution of regime probabilities. To evaluate the fidelity of long simulations, another option is to apply the clustering procedure described in this paper to determine the realism of the preferred circulation states as well as their transition paths. In this way the presence (or absence) of a simulated active-break cycle can be obtained without the use of spectral time filtering.

Appendix: Clustering details

(a) Cluster analysis algorithm

Let S be the sum of squared Euclidean distances between each state (expressed in PC space) and the centroid of the cluster to which it is assigned. Let the cluster separation Δ be the sum of squared distances of each cluster centroid about the origin, weighted by the cluster population. Since one can easily show that $(N-1)\sigma^2 = S + \Delta$, where N is the number of states and σ^2 the variance of all N points (which is fixed for the given data set), maximizing the ratio $R = \Delta/S$ is equivalent to minimizing S.

The iterative algorithm to seek the minimum of S is initiated by randomly identifying k initial points called seeds, whose coordinates form the first estimate of the cluster centroids. Each of the N points in the data set is then assigned to the seed to which it is closest, so that k groups (clusters) are formed. Each of the k centroids (average coordinates) are then recomputed based on all members in that cluster, and with this new definition of the centroids, the cluster assignments are recomputed. This iterative process of re-defining clusters proceeds until S has converged. The process is repeated in 100 trials using different seeds; the cluster assignments from the trial with minimum value of S are taken to be the final assignments. Following a method suggested by Dr. F. Molteni (personal communication) one can ensure more rapid convergence by choosing the k seed points so that they are truly representative of the entire set of points. In PC space, this is accomplished by ensuring that each seed has a norm less than a maximum value in the full PC space, that each pair of seeds has a minimum distance from each other, and that each pair of seeds belongs to a different sector in the plane of the leading two PCs.

(b) Synthetic time series for PCs

The temporal correlation structure of any time series can be captured in a synthetic time series by computing the Fourier harmonic coefficients of the original time series for all frequencies, retaining the original amplitudes, but randomizing the phases for each frequency. This random-phase approach (Christiansen,

2007) preserves the spectrum. Since the spectrum is the Fourier transform of the temporal autocovariance function (Jenkins and Watts, 1968), the synthetic time series should retain the original temporal correlation information.This approach is modified when applied to daily data for many winters, thus to discontinuous segments: the random-phase calculation is applied to the time series of seasonal means, and separately to the deviations of the daily data from those seasonal means. Finally the two synthetic time series are added (Straus, 2010).

Acknowledgements The authors gratefully acknowledge the Financial support given by the Earth System Science Organization, Ministry of Earth Sciences, Government of India (Grant no. IITM/MM-II/UnivGMU USA/2018/I) to conduct this research under the Monsoon Mission.

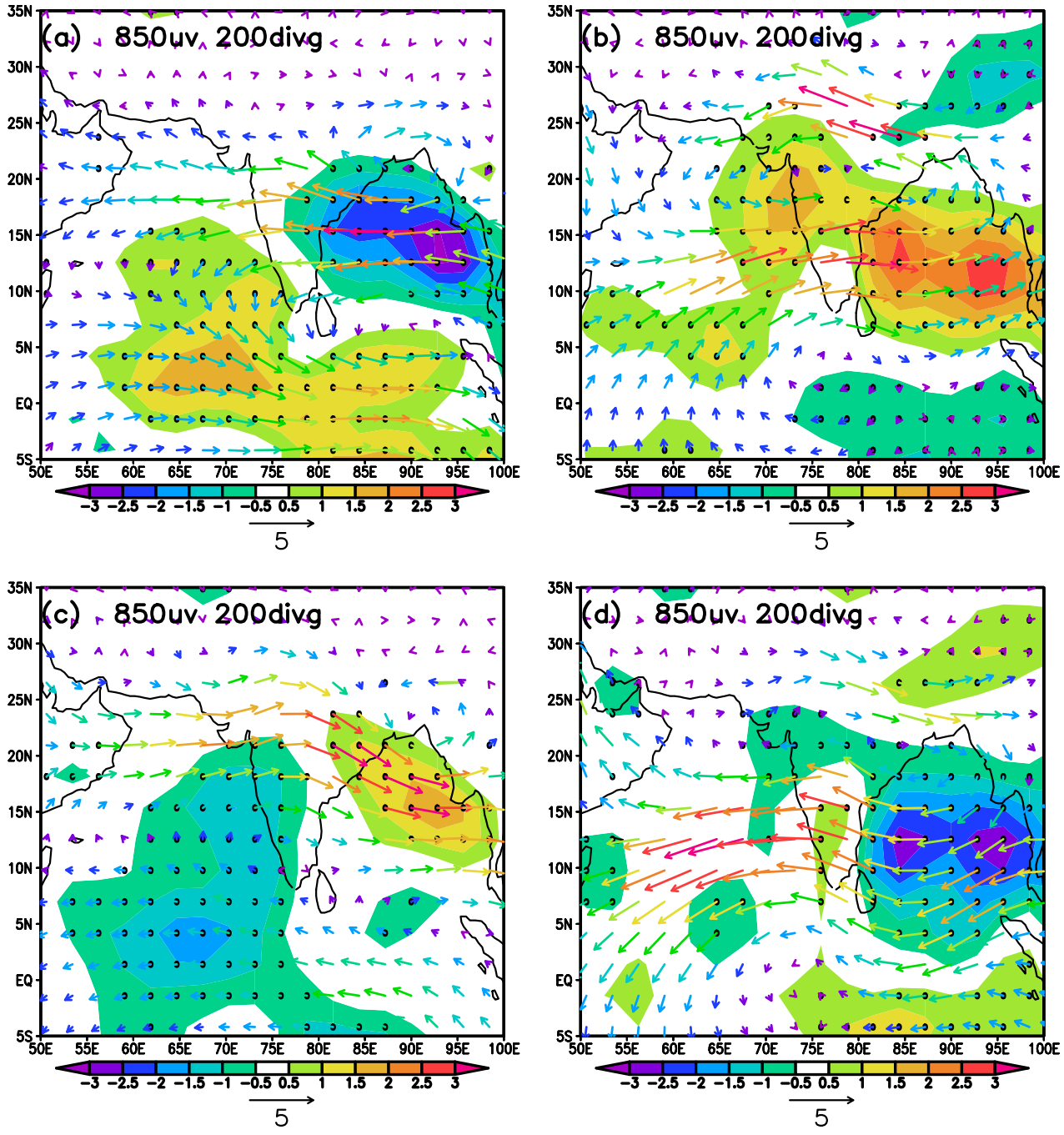

Fig. 1 Composite anomalies of horizontal winds (u, v) at 850 hPa for each of 4 clusters ($k = 4$), shown in vectors. Composite anomalies of divergence at 200 hPa shown in shading, in units of $10^{-6}s^{-1}$. Significance of the divergence at the 95% level is indicated by the stippling. The clusters are labeled (a), (b), (c), and (d) corresponding to the panels.

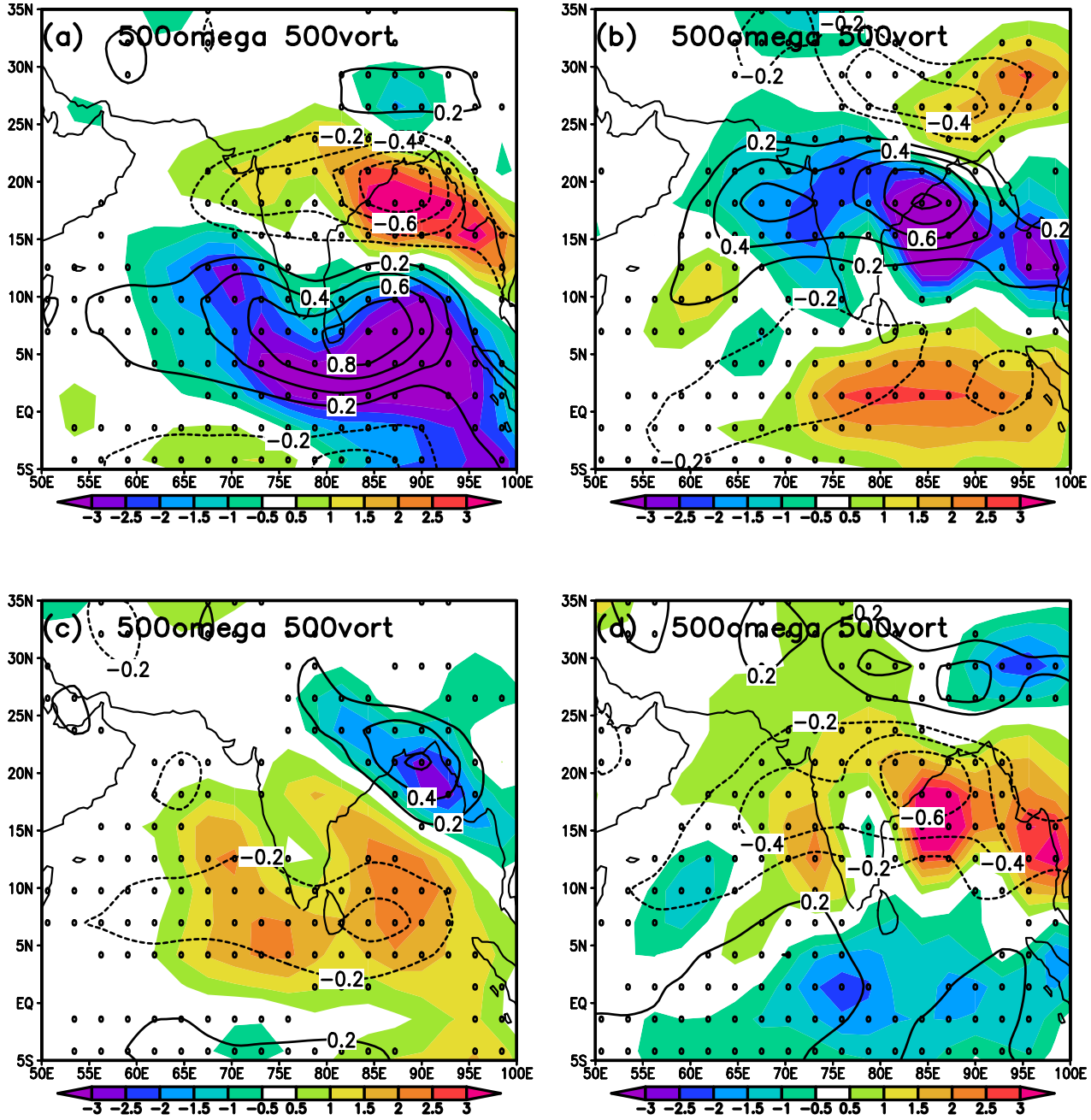

Fig. 2 Composite anomalies of vorticity ζ at 500 hPa for each of 4 clusters ($k = 4$), shown in contours in units of $10^{-5} sec^{-1}$. Composite anomalies of vertical pressure velocity (ω) at 500 hPa shown in shading, in units of $1.0^{-2} Pa\, s^{-1}$. Significance of the vorticity at the 95% level is indicated by the stippling. The clusters are labeled (a), (b), (c), and (d) corresponding to the panels.

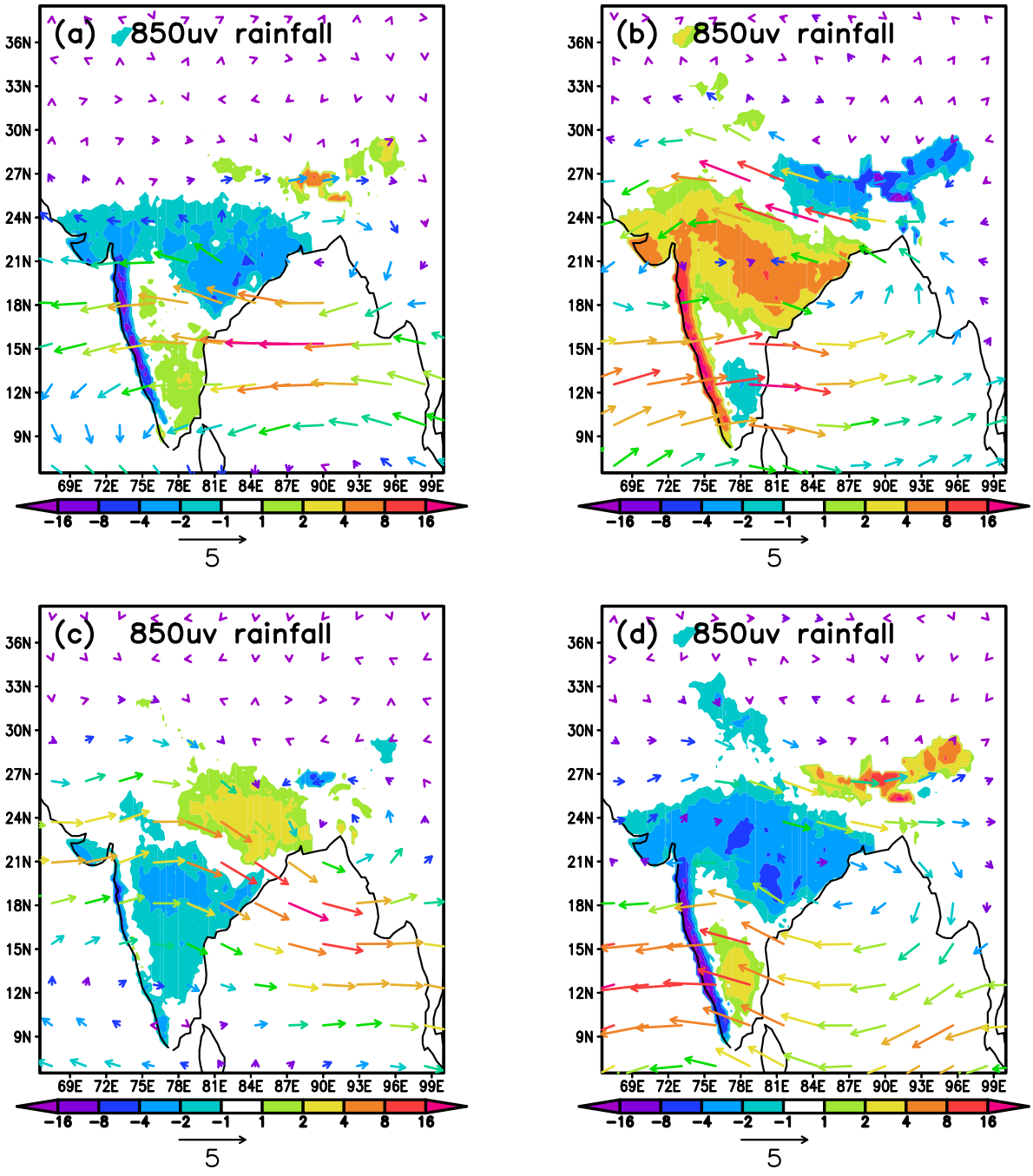

Fig. 3 Composite anomalies of horizontal winds (u, v) at 850 hPa for each of 4 clusters ($k = 4$), shown in vectors. Composite anomalies of rainfall shown in shading, in units of $10^{-3}\ d^{-1}$. Only values of the precipitation significant at the 95% level are plotted. The clusters are labeled (a), (b), (c), and (d) corresponding to the panels.

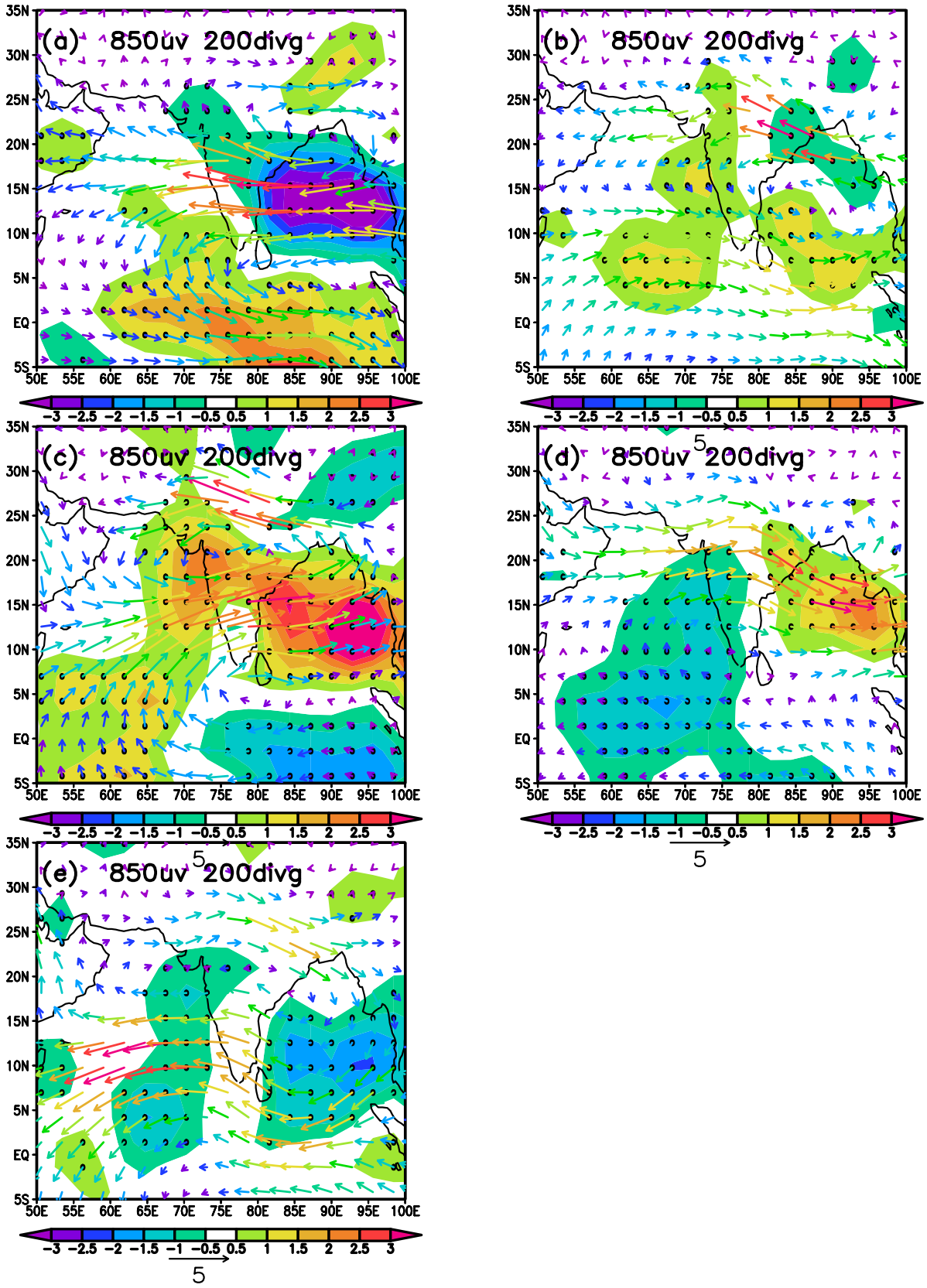

Fig. 4 Composite anomalies of horizontal winds (u, v) at 850 hPa for each of 5 clusters $(k = 5)$, shown in vectors. Composite anomalies of divergence at 200 hPa shown in shading, in units of $10^{-6}s^{-1}$. Significance of the divergence at the 95% level is indicated by the stippling. The clusters are labeled (a), (b), (c),(d) and (e) corresponding to the panels.

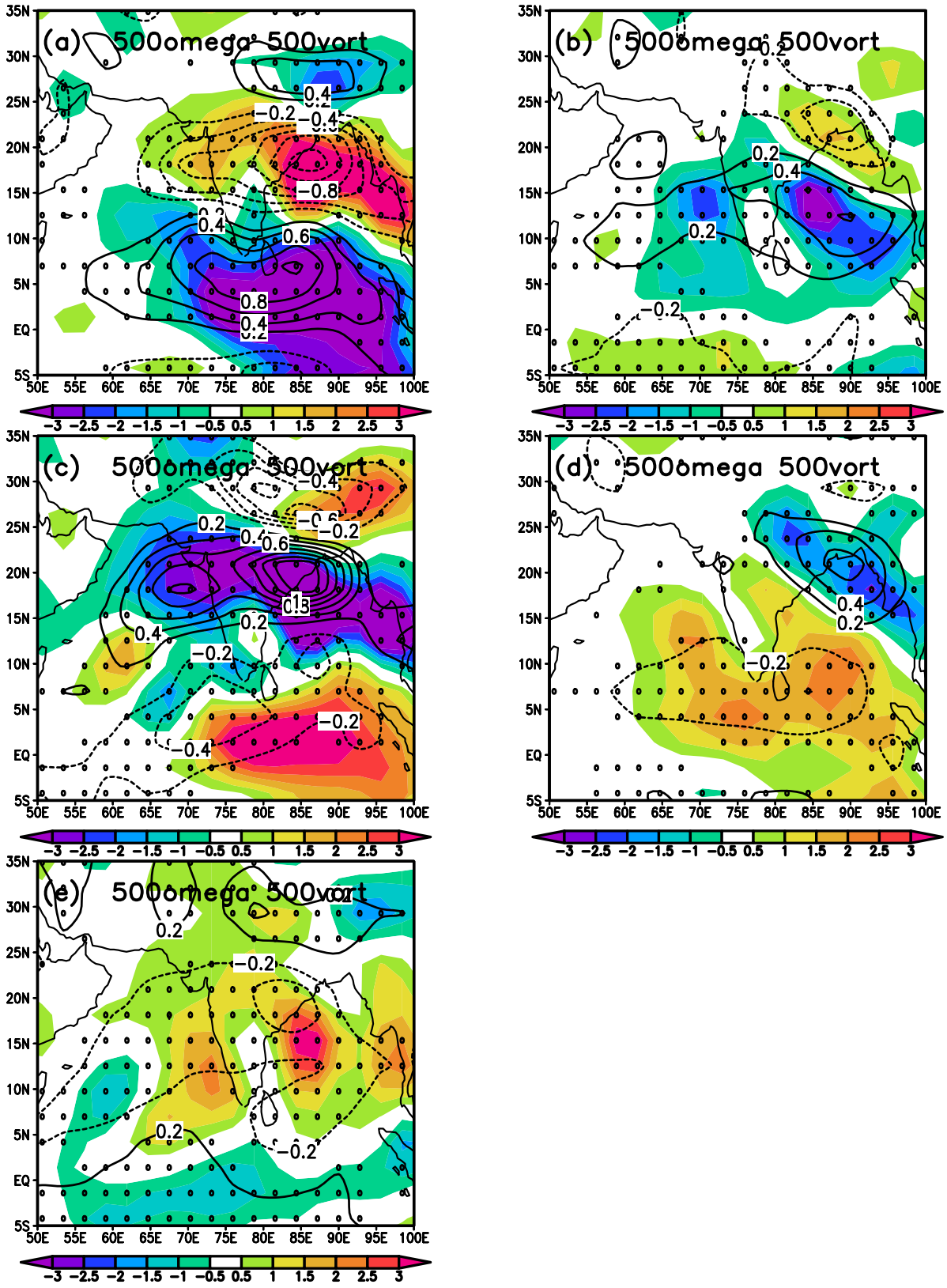

Fig. 5 Composite anomalies of vorticity ζ at 500 hPa for each of 5 clusters ($k = 5$), shown in contours in units of $10^{-5} sec^{-1}$. Composite anomalies of vertical pressure velocity (ω) at 500 hPa shown in shading, in units of $1.0^{-2} Pa\ s^{-1}$. Significance of the vorticity at the 95% level is indicated by the stippling. The clusters are labeled (a), (b), (c), (d) and (e) corresponding to the panels.

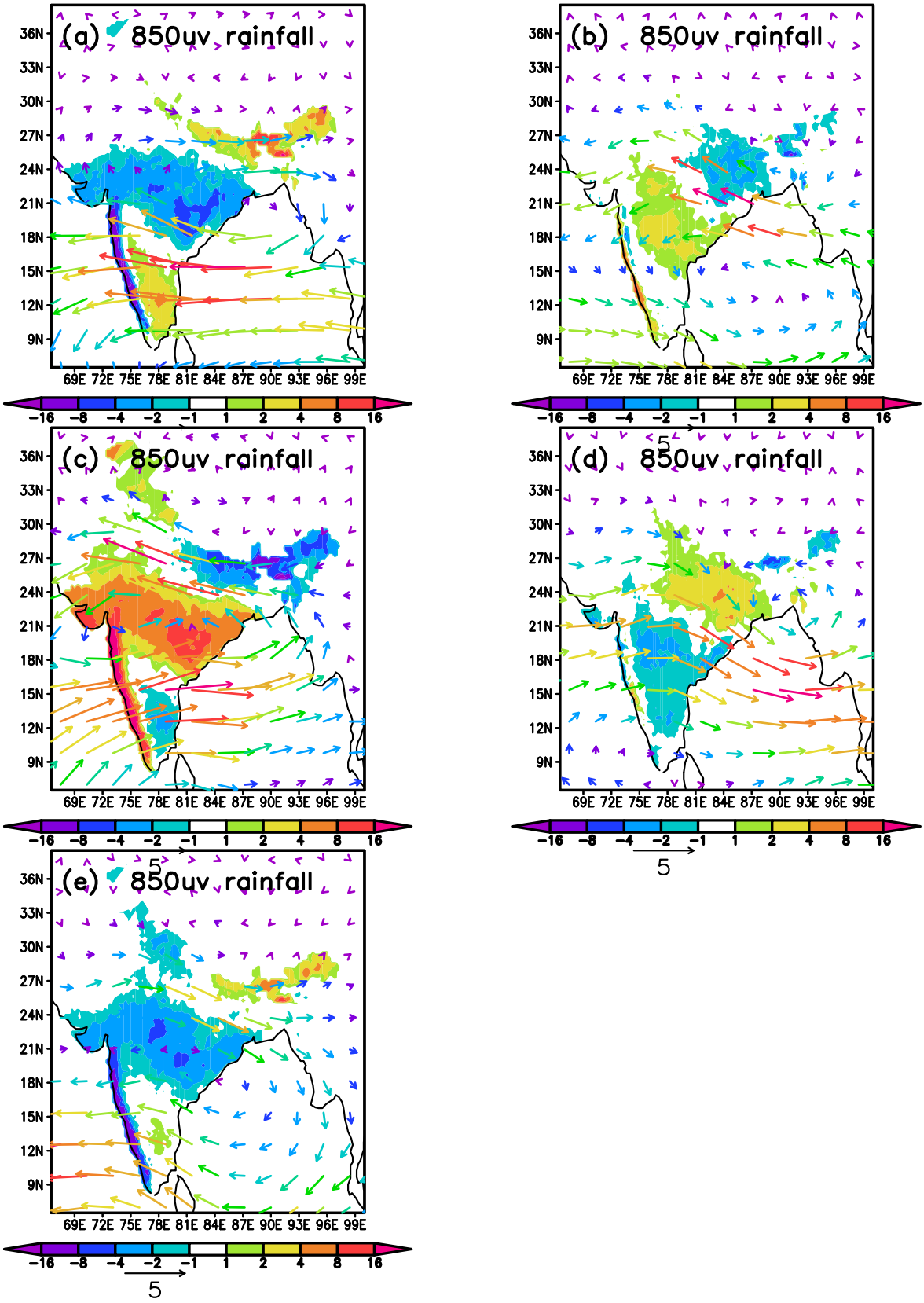

Fig. 6 Composite anomalies of horizontal winds (u, v) at 850 hPa for each of 5 clusters ($k = 5$), shown in vectors. Composite anomalies of rainfall shown in shading, in units of $10^{-3}\ d^{-1}$. Only values of the precipitation significant at the 95% level are plotted.The clusters are labeled (a), (b), (c),(d) and (e) corresponding to the panels.

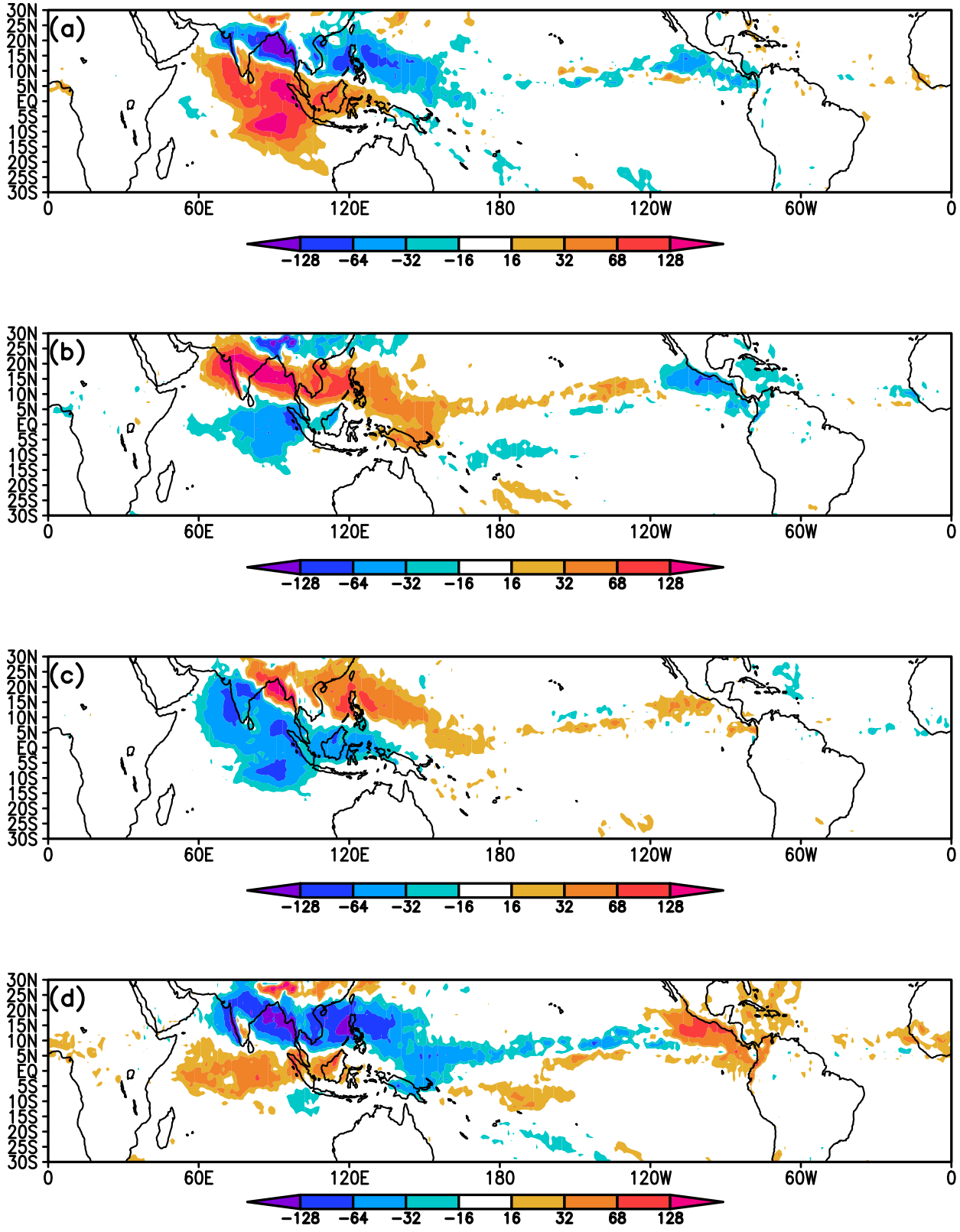

Fig. 7 Composite anomalies of vertically integrated diabatic heating for each of 4 clusters ($k = 4$), in units of $W\ m^{-2}$. The clusters are labeled (a), (b), (c), and (d) corresponding to the panels.

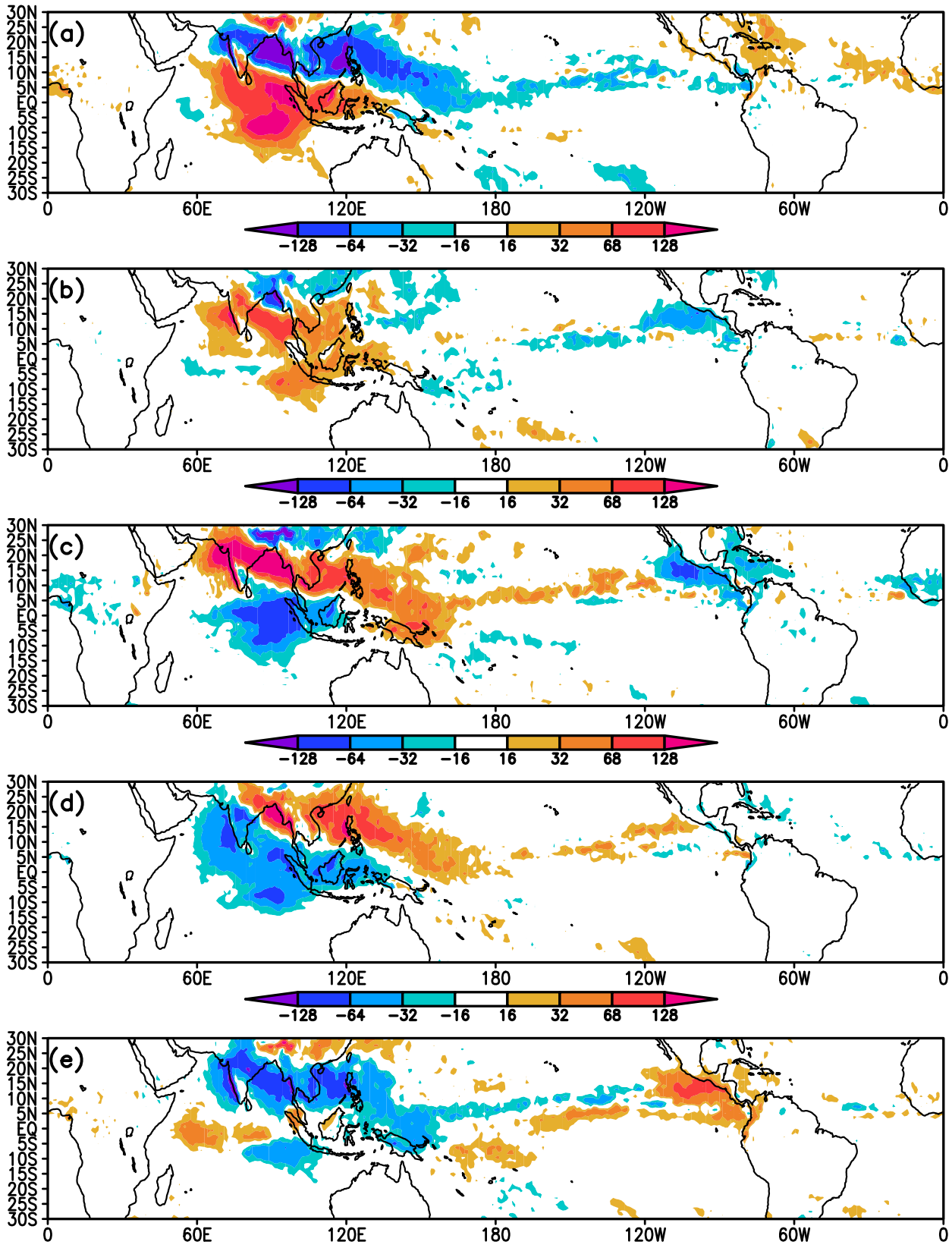

Fig. 8 Composite anomalies of vertically integrated diabatic heating for each of 5 clusters ($k = 5$), in units of $W\ m^{-2}$. The color scale applies to all panels. The clusters are labeled (a), (b), (c), (d) and (e) corresponding to the panels.

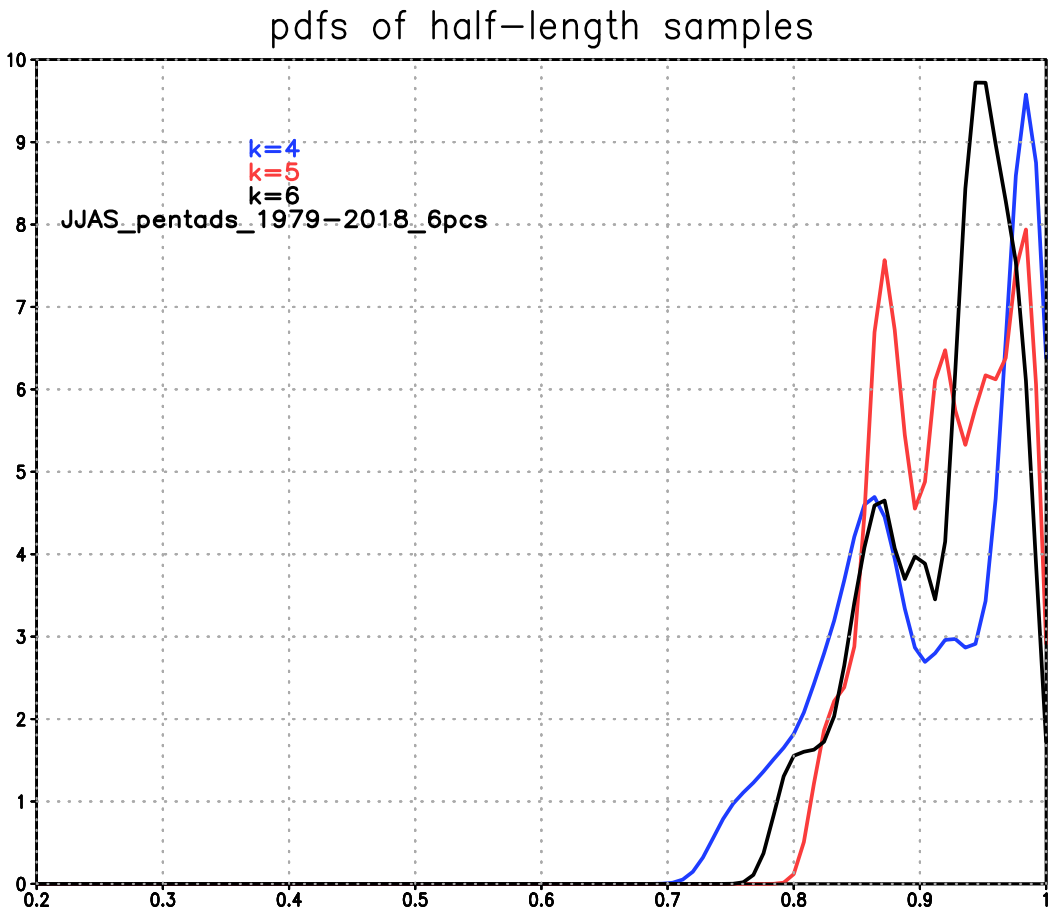

Fig. 9 Probability distributions (normalized) of the average pattern correlations of the cluster centroids obtained from 100 randomly chosen half-length samples with the centroids from the full data set, shown for three values of k. For each value of k, the centroids from the half-length samples are matched with those of full data set and the correlation averaged over the k centroids.

Table 1 Statistics of Cluster Episodes for 4 clusters: Average Length and Total Time

Cluster	**Ave Length**		**Total Time**	
	(pentads)	(days)	(pentads)	(percent)
(a)	1.627	8.1	205	21.4
(b)	1.838	9.2	283	29.5
(c)	1.731	8.7	277	28.9
(d)	1.696	8.5	195	20.3

Table 2 Statistics of Cluster Episodes for 5 clusters: Average Length and Total Time

Cluster	**Ave Length**		**Total Time**	
	(pentads)	(days)	(pentads)	(percent)
(a)	1.557	7.8	165	17.2
(b)	1.444	7.2	205	21.4
(c)	1.558	7.8	148	15.4
(d)	1.648	8.2	234	24.4
(e)	1.564	7.8	208	21.7

Table 3 Transition Table for 4 clusters. Elements give number of transitions from cluster M (row) to column (N). The off-diagonal elements significant at the 95% confidence level are given in **bold** font.

COLUMN ROW M	(a)	(b)	(c)	(d)
(a)	79	**62**	28	29
(b)	26	129	**89**	26
(c)	43	**59**	117	**44**
(d)	**47**	27	35	80

Table 4 Analysis of complete and partial cycles: $(a) \rightarrow (b) \rightarrow (c) \rightarrow (d)$

	Number		**Avg. Length**	
No. of legs	per year	total	pentads	days
3	0.77	31	5.16	25.8
4	0.40	16	6.75	33.8
5	0.17	7	. 9.75	47.9

Table 5 Transition Table for 5 clusters. Elements give number of transitions from cluster M (row) to column (N). The off-diagonal elements significant at the 95% confidence level are given in **bold** font.

COLUMN ROW M	(a)	(b)	(c)	(d)	(e)
(a)	59	**45**	8	12	34
(b)	**30**	63	**47**	33	23
(c)	6	18	53	**54**	13
(d)	19	39	23	92	**48**
(e)	**43**	31	16	36	75

Table 6 Analysis of complete and partial cycles: $(a) \rightarrow (b) \rightarrow (c) \rightarrow (d) \rightarrow (e)$

	Number		**Avg. Length**	
No. of legs	per year	total	pentads	days
3	0.70	28	4.57	22.9
4	0.28	11	6.45	32.3
5	0.25	10	6.90	34.5
6	0.88	3	. 8.33	41.7

9 Declarations

9.1 Funding

The authors gratefully acknowledge the Financial support given by the Earth System Science Organization, Ministry of Earth Sciences, Government of India (Grant no. IITM/MM-II/UnivGMU USA/2018/I) to conduct this research under the Monsoon Mission.

9.2 Conflicts of interest

The authors have no relevant financial or non-financial interests to disclose.

9.3 Data and Code Availability

The datasets generated during and/or analysed during the current study, as well as the Fortran codes, are available from the corresponding author on reasonable request.

References

Annamalai H, Slingo J (2001) Active/break cycles: diagnosis of the intraseasonal variability of the asian summer monsoon. Clim Dyn 18:85–102, DOI https://doi.org/10.1007/s003820100161

534 Annamalai H, Sperber KR (2005) Regional heat sources and the active and
535 break phases of boreal summer intraseasonal (30–50 day) variability. J At-
536 mos Sci 62:2726–2748
537 Chatterjee P, Goswami BN (2004) Structure, genesis and scale selection of the
538 tropical quasi-biweekly mode. Quart J Royal Meteor Soc 130:1171–1194,
539 DOI https://doi.org/10.1256/qj.03.133
540 Christiansen B (2007) Atmospheric circulation regimes: Can cluster analysis
541 provide the number? J Climate 20:2229–2250
542 Chu JE, Wang B, Lee JY, Ha KJ (2017) Boreal summer intraseasonal phases
543 identified by nonlinear multivariate empirical orthogonal function-based self-
544 organizing map (esom) analysis. J Climate 30:3513–3528
545 Deday E, Simon JC (1976) Clustering analysis. In: Fu K (ed) Communication
546 and Cybernetics 10 Digital Pattern Recognition, Springer-Verlag, pp 47–94
547 Dee DP, coauthors (2011) The ERA-interim reanalysis: configuration and per-
548 formance of the data assimilation system. Q J R Meteor Soc 137:553–597
549 Desbois M, Seze G, Swejwach G (1982) Automatic classification of cloudfs
550 on METEOSAT imagery: Application to high-level clouds. J Appl Meteor
551 21:401–412
552 Hazra A, Krishnamurthy V (2014) Space?time structure of diabatic heating
553 in monsoon intraseasonal oscillation. J Climate 28:2234–2255
554 Hersbach H, co authors (2020) The ERA5 gobal reanalysis. Quart J Royal
555 Meteor Soc 146:1999–2049, DOI https://doi.org/10.1002/qj.3803

556 Jenkins GM, Watts DG (1968) Spectral Analysis and its applications. Holden-
557 Day, San Francisco, 525 pp.
558 Krishnamurthi TN, Ardunay P (1980) The 10 to 20-day westward propagating
559 mode and "Breaks in the Monsoons". Tellus 32:15–26, DOI https://doi.org/
560 10.1111/j.2153-3490.1980.tb01717.x
561 Krishnamurthi TN, Bhalme HN (1976) Oscillations of a monsoon system. Part
562 i. observational aspects. J Atmos Sci 33:1937–1954, DOI https://doi.org/10.
563 1175/1520-0469(1976)033%3C1937:OOAMSP%3E2.0.CO;2
564 Krishnamurthy V, Shukla J (2000) Intraseasonal and interannual variability of
565 rainfall over India. J Climate 13:4366–4377, DOI https://doi.org/10.1175/
566 1520-0442(2000)013⟨0001:IAIVOR⟩2.0.CO;2
567 Krishnamurthy V, Shukla J (2007) Intraseasonal and seasonally persisting
568 patterns of indian monsoon rainfall. J Climate 20:3–20, DOI https://doi.
569 org/10.1175/JCLI3981.1
570 Krishnamurthy V, Shukla J (2008) Seasonal persistence and propagation of
571 intraseasonal patterns over the Indian monsoon region. Clim Dyn 30:353–
572 369, DOI 10.1007/s00382-007-0300-7
573 Lau KM, Chan PH (1986) Aspects of the 40?50 day oscillation during the
574 northern summer as inferred from outgoing longwave radiation. Mon Wea
575 Rev 114:1354–1367
576 Lee SS, Wang B, Waliser DE, Neena JM, Lee JY (2015) Predictability and
577 prediction skill of the boreal summer intraseasonal oscillation in the In-
578 traseasonal Variability Hindcast Experiment. Clim Dyn 45:2123–2135, DOI

579 10.1007/s00382-014-2461-5

580 Michelangeli PA, Vautard R, Legras B (1995) Weather regimes: Reccurrence
581 and quasi-stationarity. J Atmos Sci 52:1237–1256

582 Murakami T, Nakazawa T, He J (1984) On the 40 – 50 day oscillation during
583 1979 Northern Hemisphere summer. Part 1: Phase propagation. J Meteor
584 Soc Japan 62:440–468

585 Neena JM, Waliser D, Jiang X (2017) Model performance metrics and process
586 diagnostics for boreal summer intraseasonal variability. Clim Dyn 48:1661–
587 1683, DOI 10.1007/s00382-016-3166-8

588 Pai DS, Sridhar L, Rajeevan M, Sreejith OP, Satbhai NS, Mukhopadhyay B
589 (2014) Development of a new high spatial resolution (0.25 deg x 0.25 deg)
590 long period (1901-2010) daily gridded rainfall data set over India and its
591 comparison with existing data sets over the region. MAUSAM pp 1–18

592 Pattanaik DR, Sahai AK, Muralikrishna R, Mandal R, Dey A (2020) Active-
593 break transitions of monsoons over india as predicted by coupled model en-
594 sembles. Pure Appl Geophys 177:4391–4422, DOI https://doi.org/10.1007/
595 s00024-020-02503-2

596 Prasanna V (2020) Impact of monsoon rainfall on the total foodgrain yield
597 over india. J Earth Syst Sci 123:1129–1145

598 Rajeevan M, Gadgil S, Bhate J (2010) Active and break spells of the indian
599 summer monsoon. J Earth Syst Sci 119:229–247

600 Sahai AK, Abhilash S, Chattopadhyay R, Borah N, Joseph S, Sharmila S, Ra-
601 jeevan M (2015) High-resolution operational monsoon forecasts: an objective

602 assessment. Clim Dyn 44:3129–3140, DOI 10.1007/s00382-014-2210-9
603 Straus DM (1983) On the role of the seasonal cycle. J Atmos Sci 40:303–313
604 Straus DM (2010) Synoptic-eddy feedbacks and circulation regime analysis.
605 Mon Wea Rev 138:4026–4034
606 Straus DM, Molteni F, Corti S (2017) Atmospheric Regimes: The Link Be-
607 tween Weather and Climate. In: Franzke CLE, O'Kane TJ (eds) Nonlinear
608 and Stochastic Climate Dynamics, Cambridge University Press, pp 105–134
609 Suhas E, Neena JM, Goswami BN (2013) An indian monsoon intraseasonal
610 oscillations (MISO) index for real time monitoring and forecast verification.
611 Clim Dyn 40:2605–2616, DOI 10.1007/s00382-012-1462-5
612 Swenson ET, Straus DM (2021) A modelling framework for a better under-
613 standing of the tropically-forced component of the indian monsoon variabil-
614 ity. J Earth Syst Sci 130, DOI https://doi.org/10.1007/s12040-020-01503-z
615 Toth Z (1993) Preferred and unpreferred circulation types in the northern
616 hemisphere wintertime phase space. J Atmos Sci 50:2868–2888

Figures

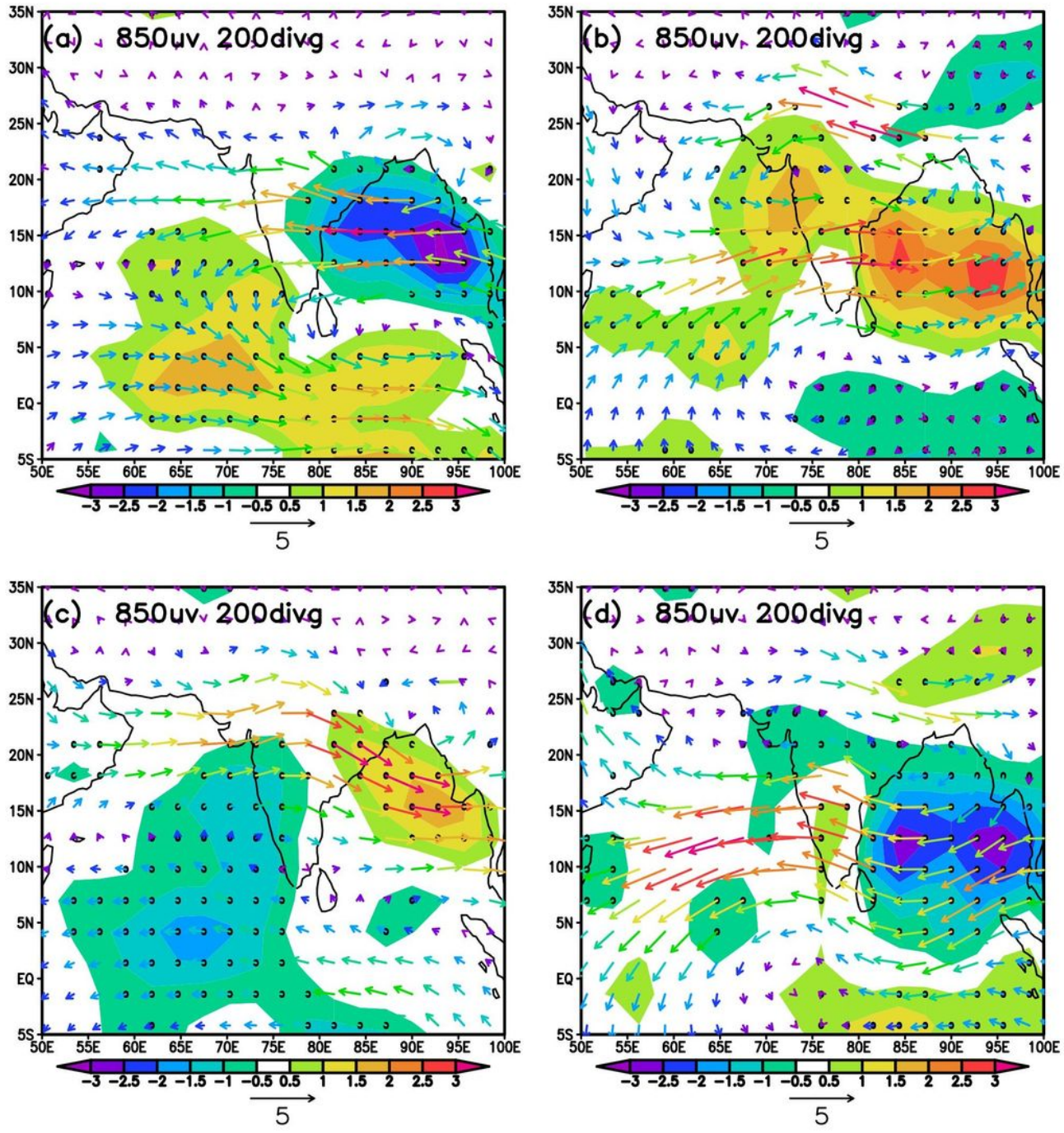

Figure 1

Please see the Manuscript PDF file for the complete figure caption Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its

authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

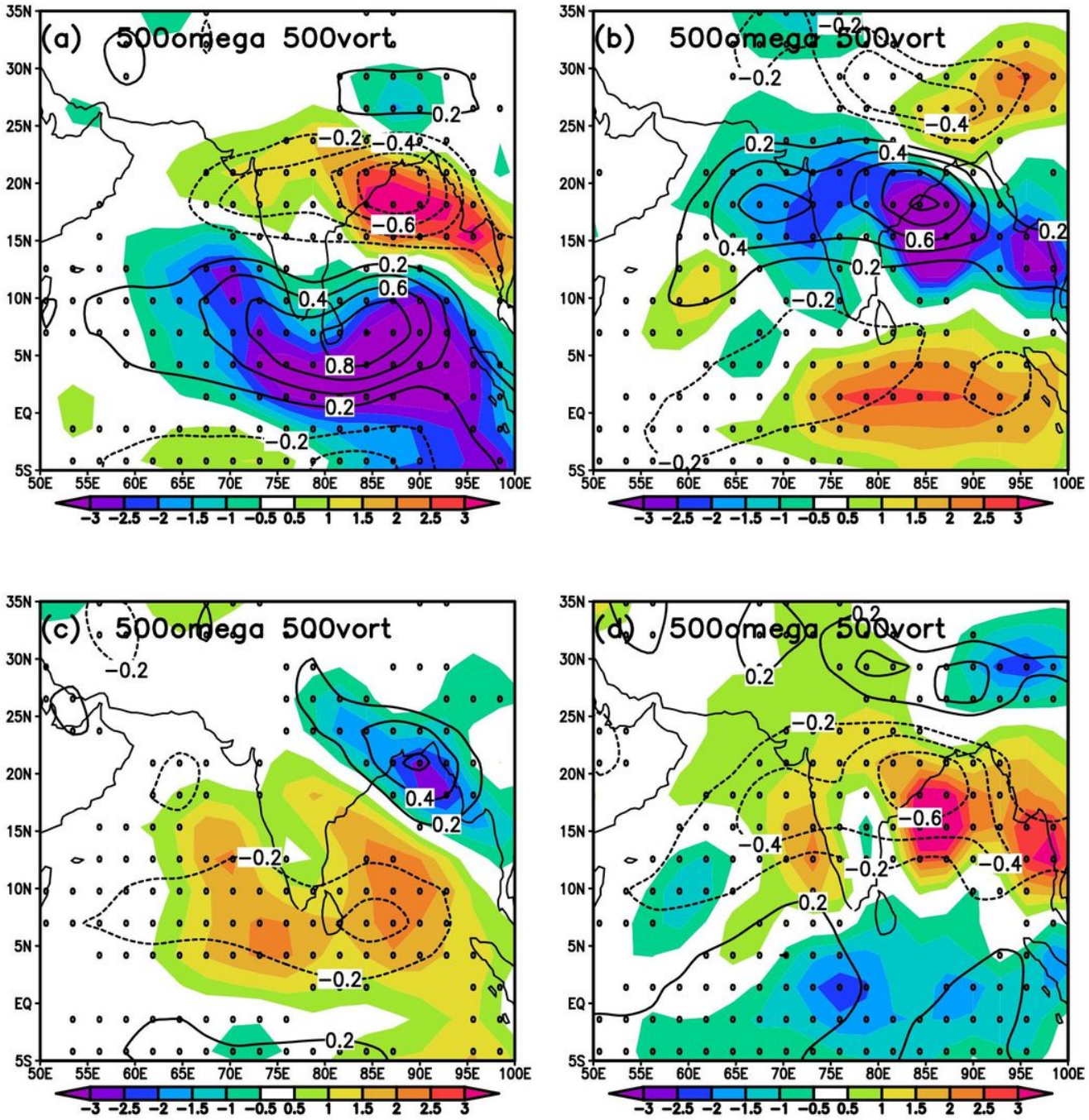

Figure 2

Please see the Manuscript PDF file for the complete figure caption Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its

authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

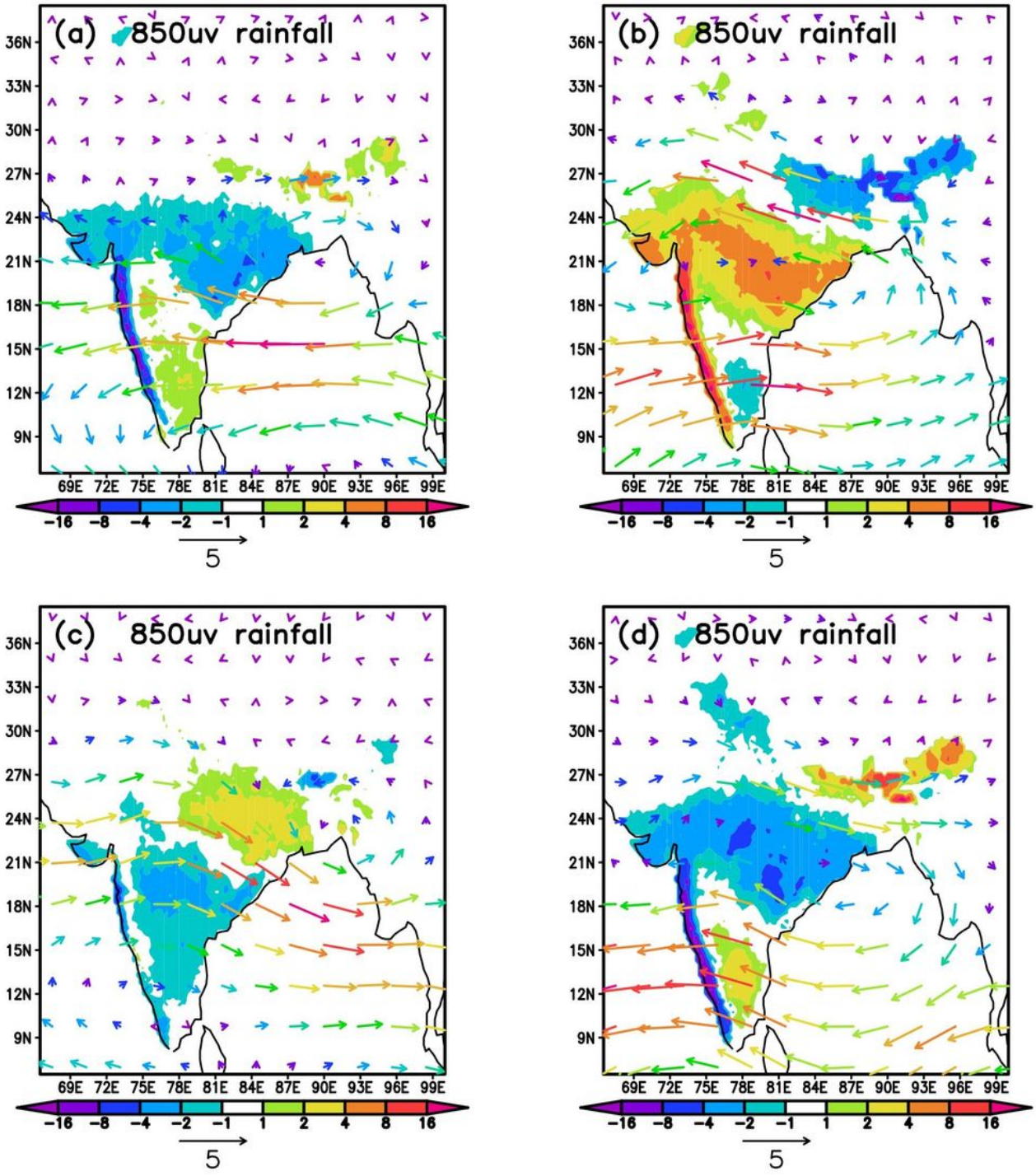

Figure 3

Please see the Manuscript PDF file for the complete figure caption Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its

authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

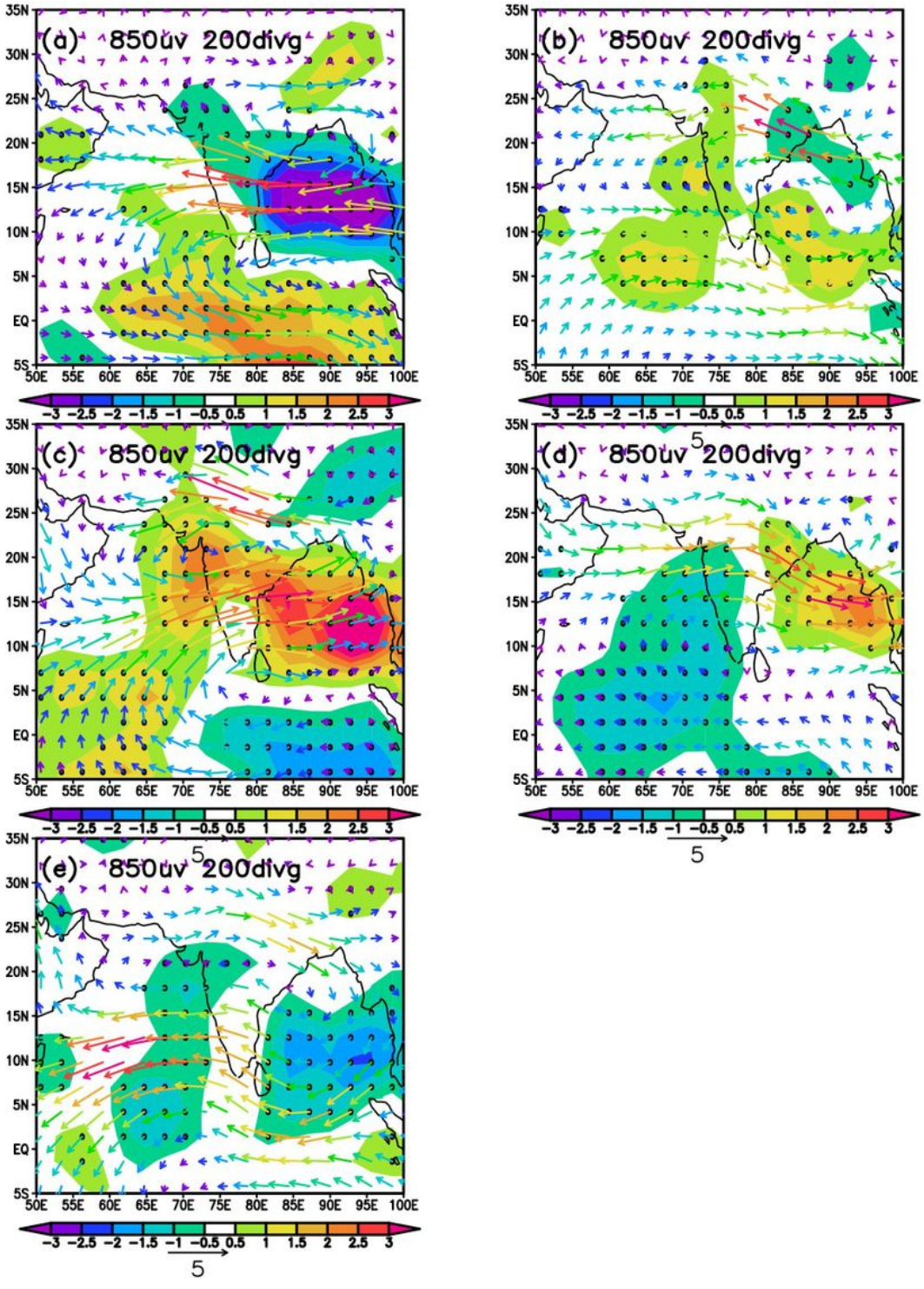

Figure 4

Please see the Manuscript PDF file for the complete figure caption Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its

authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

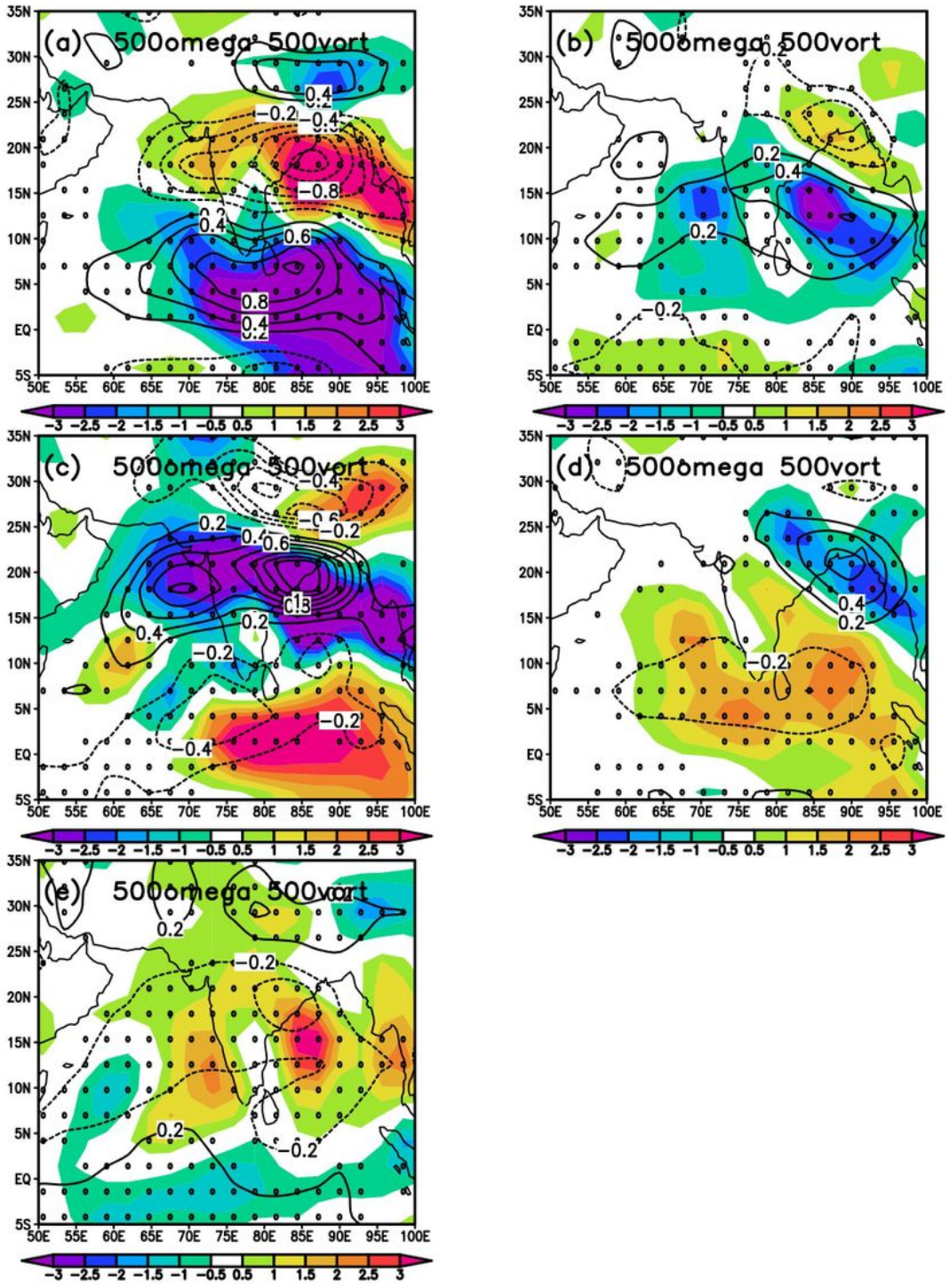

Figure 5

Please see the Manuscript PDF file for the complete figure caption Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its

authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

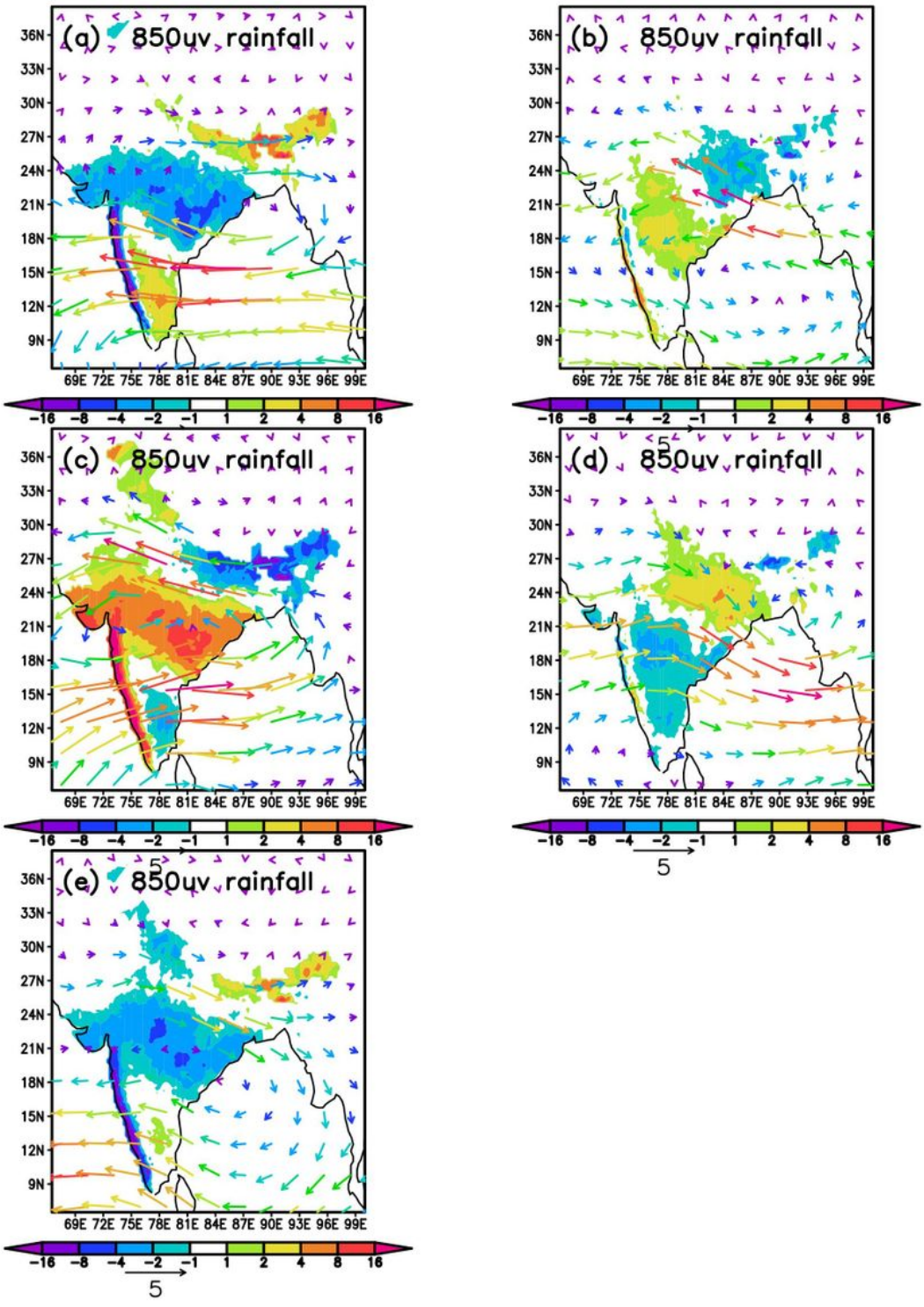

Figure 6

Please see the Manuscript PDF file for the complete figure caption Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its

authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

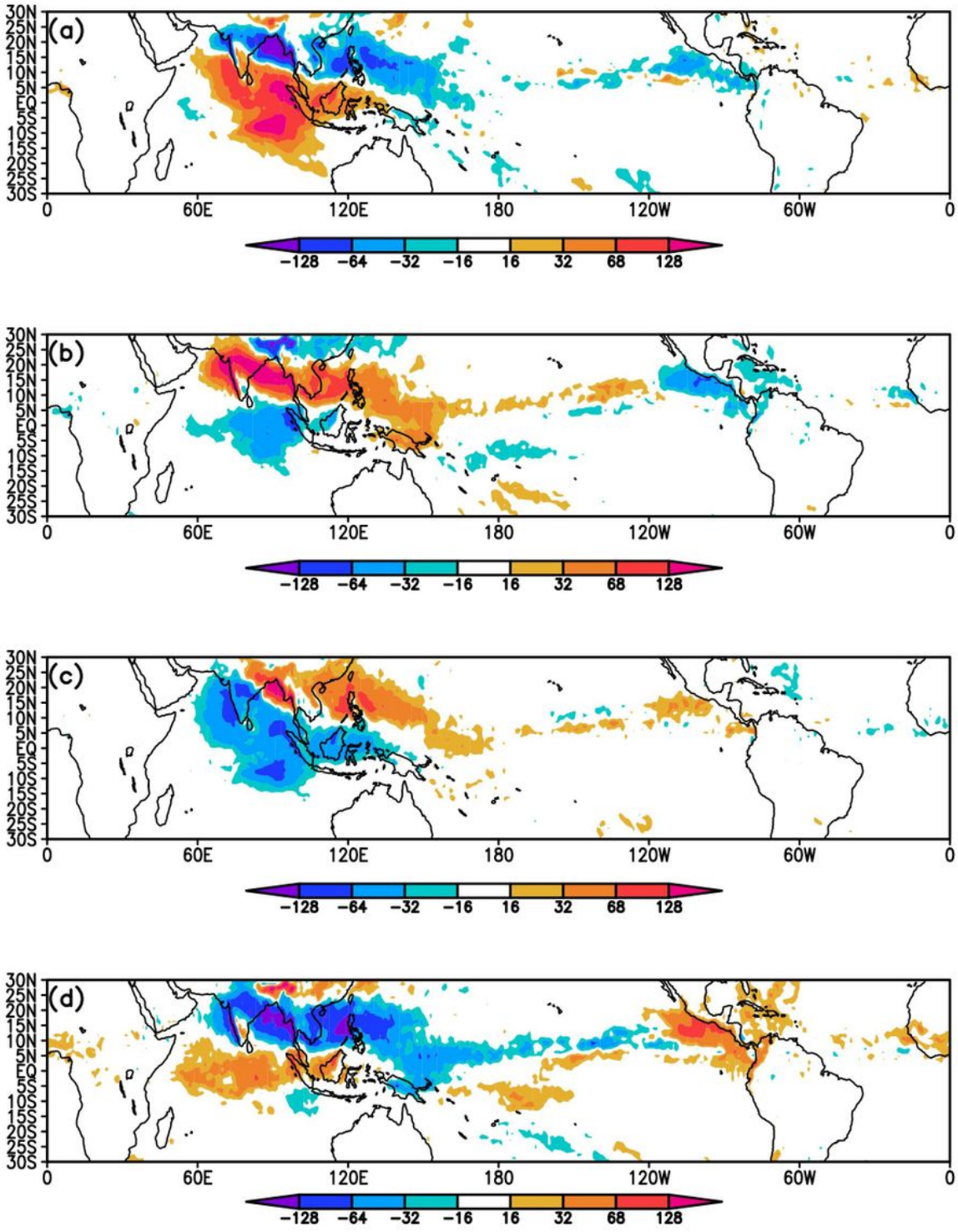

Figure 7

Please see the Manuscript PDF file for the complete figure caption Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its

authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

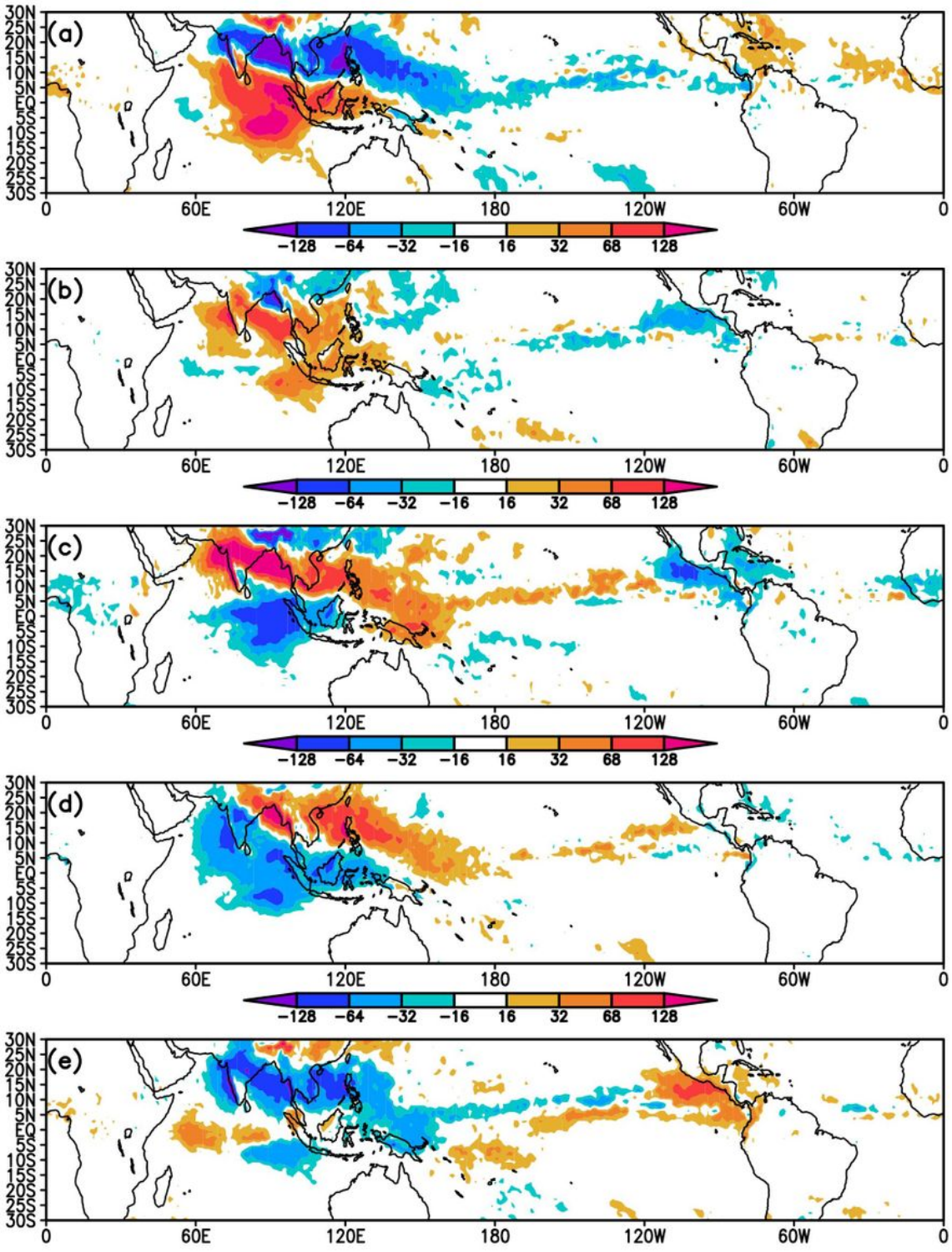

Figure 8

Please see the Manuscript PDF file for the complete figure caption Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its

authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

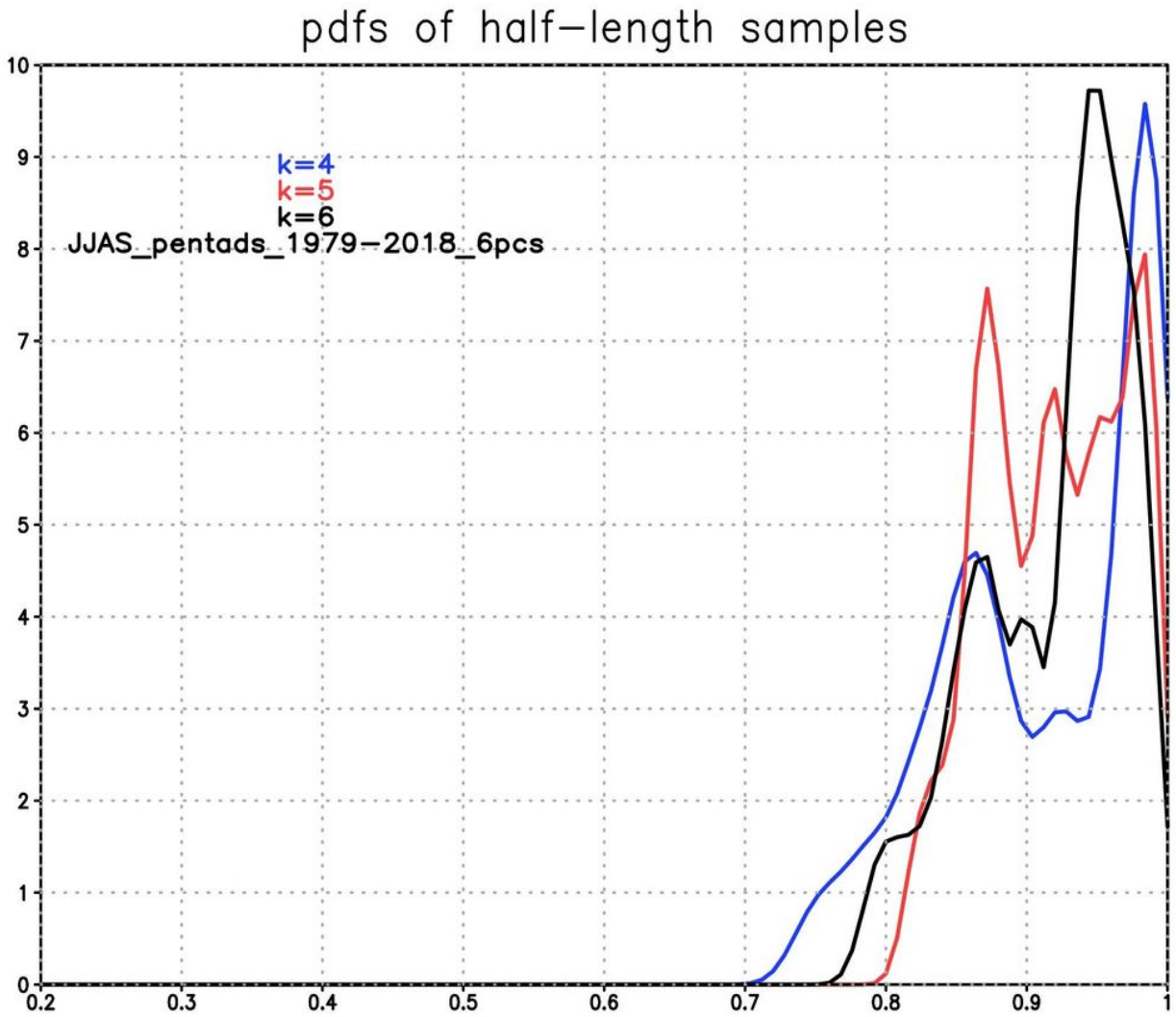

Figure 9

Please see the Manuscript PDF file for the complete figure caption