Connecting the deep collection of particles with surface ocean signatures

Lu Wang[1], Jonathan Gula[2,3], Jérémy Collin[1], Laurent Mémery[1], Xiaolong Yu[4]

[1]Univ Brest, CNRS, IRD, Ifremer, Laboratoire des Sciences de l'Environnement Marin (LEMAR), IUEM, Plouzané, France

[2]Univ Brest, CNRS, IRD, Ifremer, Laboratoire d'Océanographie Physique et Spatiale (LOPS), IUEM, Plouzané, France

[3]Institut Universitaire de France (IUF), Paris, France

[4]School of Marine Sciences, Sun Yat-sen University, Zhuhai, China

Contents of this file

1. Defining the cluster number (k)

K-means seeks to iteratively minimize the within-cluster Sum of Squared Errors (SSE) which is often called cluster inertia, defined as follows:

$$\mathrm{SSE} = \sum_{i=1}^{\mathrm{k}} \sum_{x_j \in C_i} \|\mathbf{y_j} - \mathbf{c_i}\|^2 \tag{1}$$

where y_j is the jth object in cluster C_i, and c_i is the center of cluster C_i. Inertia measures how internally coherent clusters are, hence lower values are better. The Elbow Method calculates the SSE for runs of KMS clustering on the dataset using a range of values of k. The optimal value is chosen when SSE first starts to bend or level off, visible as an "elbow" in the plot of SSE-versus-k. However, this elbow cannot always be unambiguously identified.

The silhouette coefficient is a very useful method to find the number of k when the Elbow point is not shown. It is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation). The silhouette coefficient is calculated using the mean intra-cluster distance and the mean nearest-cluster distance. For a particular data point

$$\mathbf{S}(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2}$$

where $\mathbf{S}(i)$ is the silhouette coefficient for a particular data point i, $a(i)$ is the average distance between i and all the other data points in the cluster to which i belongs, and $b(i)$ is the average distance between i and all other data points belonging to outside/neighboring clusters. The value of the silhouette coefficient is between [-1,1]. The average silhouette coefficient (SIL) of all data points in the data set is used to assess the quality of clustering. A high value is desirable in the graph of average $\mathbf{S}(i)$ versus k. Ideally, the optimal k

value is picked when $\mathbf{S}(i)$ reaches its global maximum. However, the "optimal" cluster number is not always reasonable, which may cause the data structure not fully discovered. Instead, the reasonable cluster number is chosen at a knee point where a smaller peak is located.

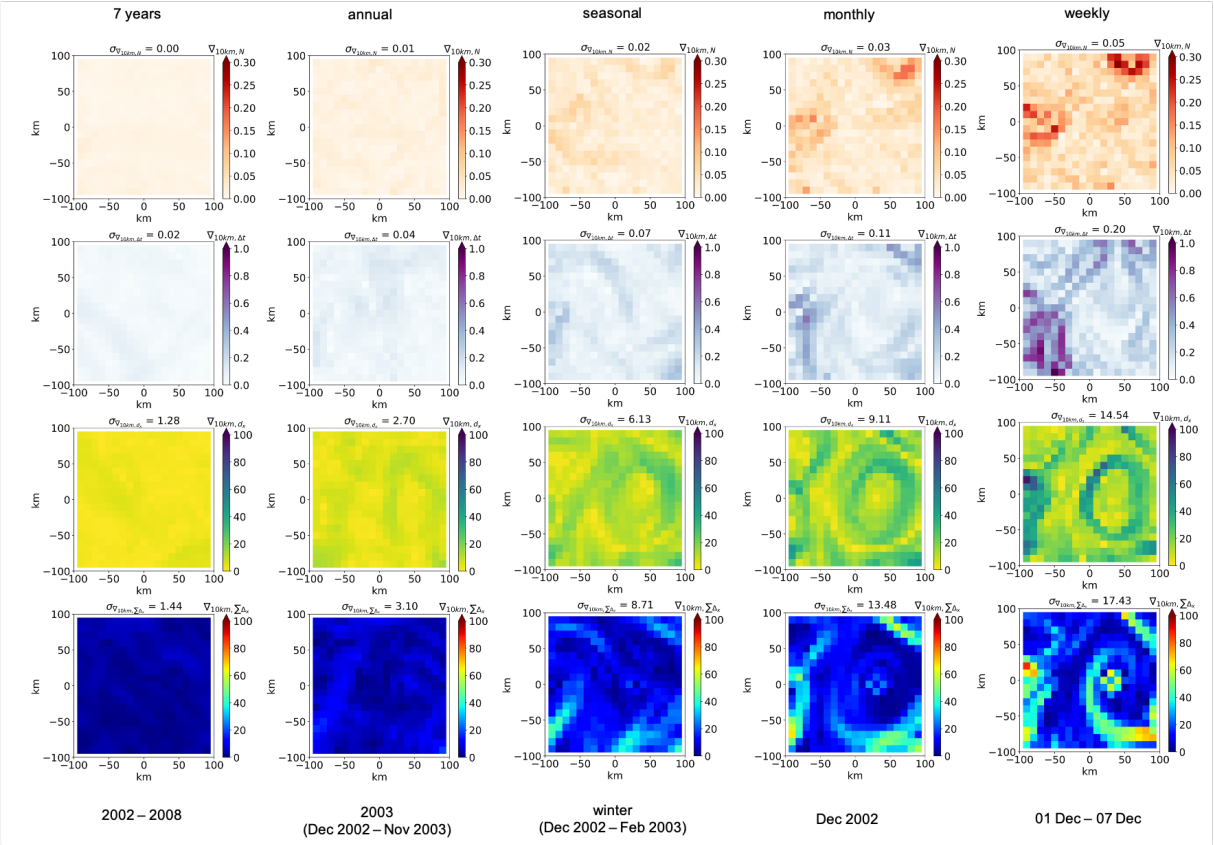

Figure S1. Spatial gradients of normalized particle distribution N, travel time anomaly Δ_t, horizontal displacement d_x, and trajectory length $\sum \Delta x$.

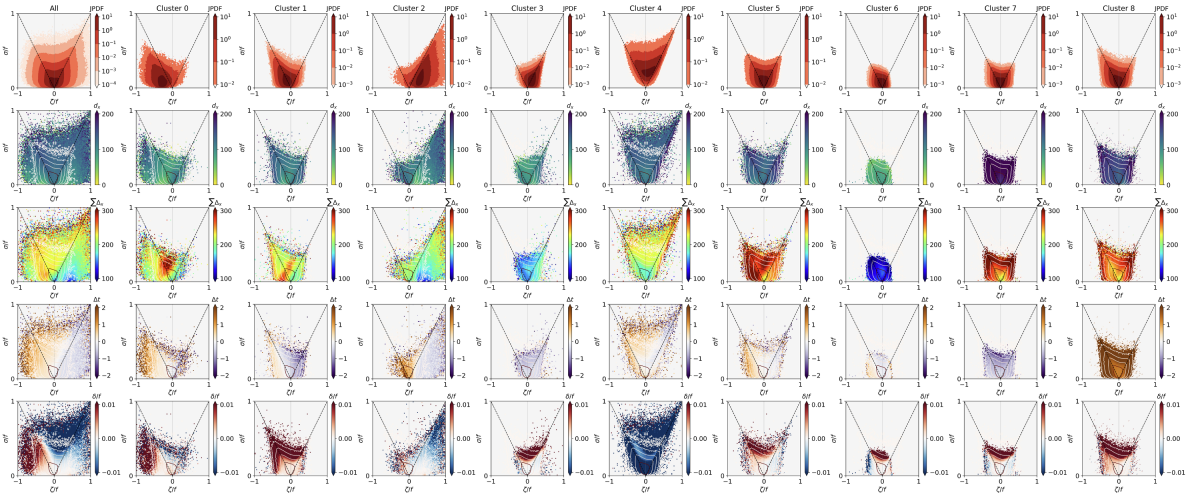

Figure S2. Vorticity-strain JPDF at 200 m for the 9 clusters (the top panel, left to right: cluster 0 - 8). The second to the final panel are horizontal displacement, trajectory length, travel time anomaly, and divergence conditioned on the vorticity-strain JPDF.

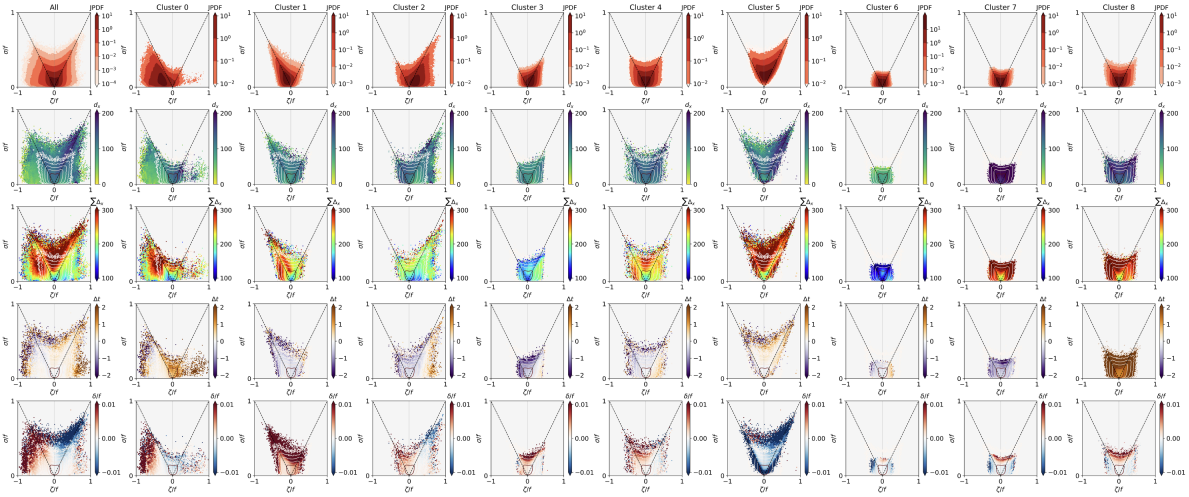

Figure S3. Vorticity-strain JPDF at 500 m for the 9 clusters.

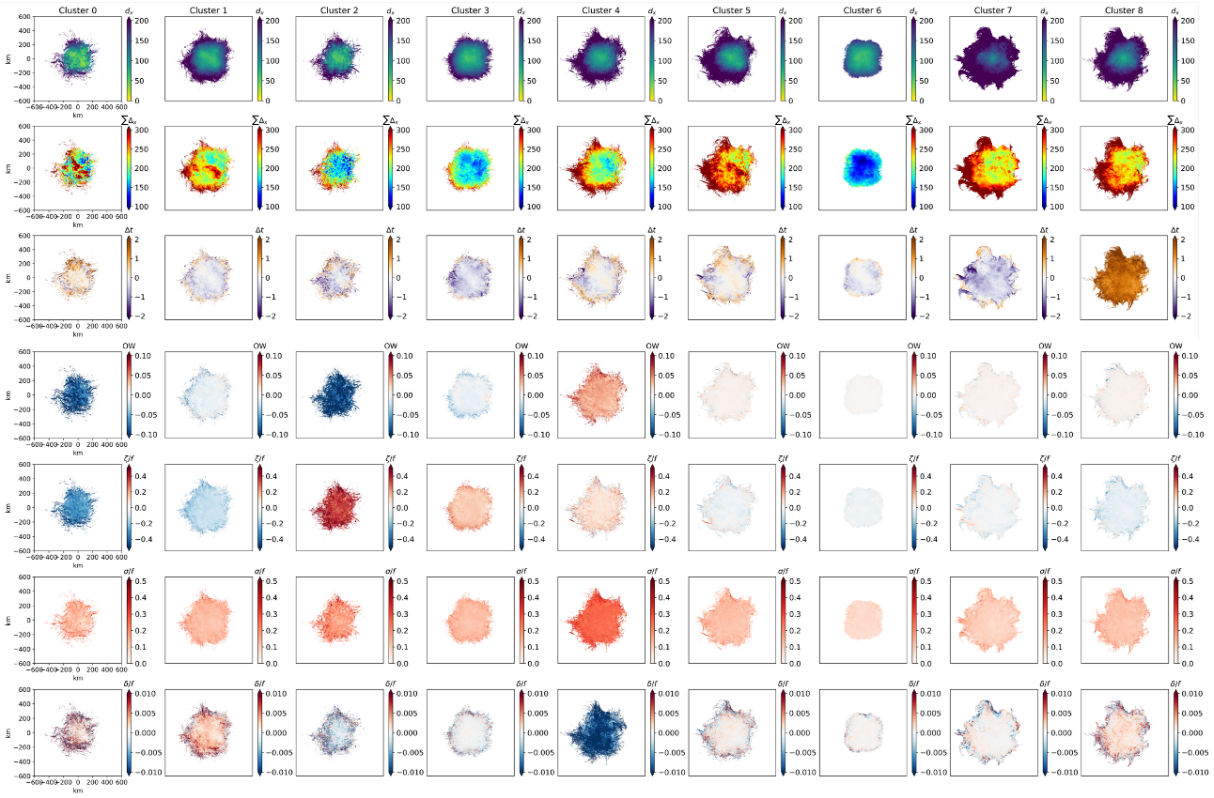

Figure S4. Maps of variables on particle initial positions for the 9 clusters (from top to bottom). The variables (from left to right) are horizontal displacement, trajectory length, travel time anomaly, and Okubo-Weiss parameter, relative vorticity, strain and divergence at 200 m.

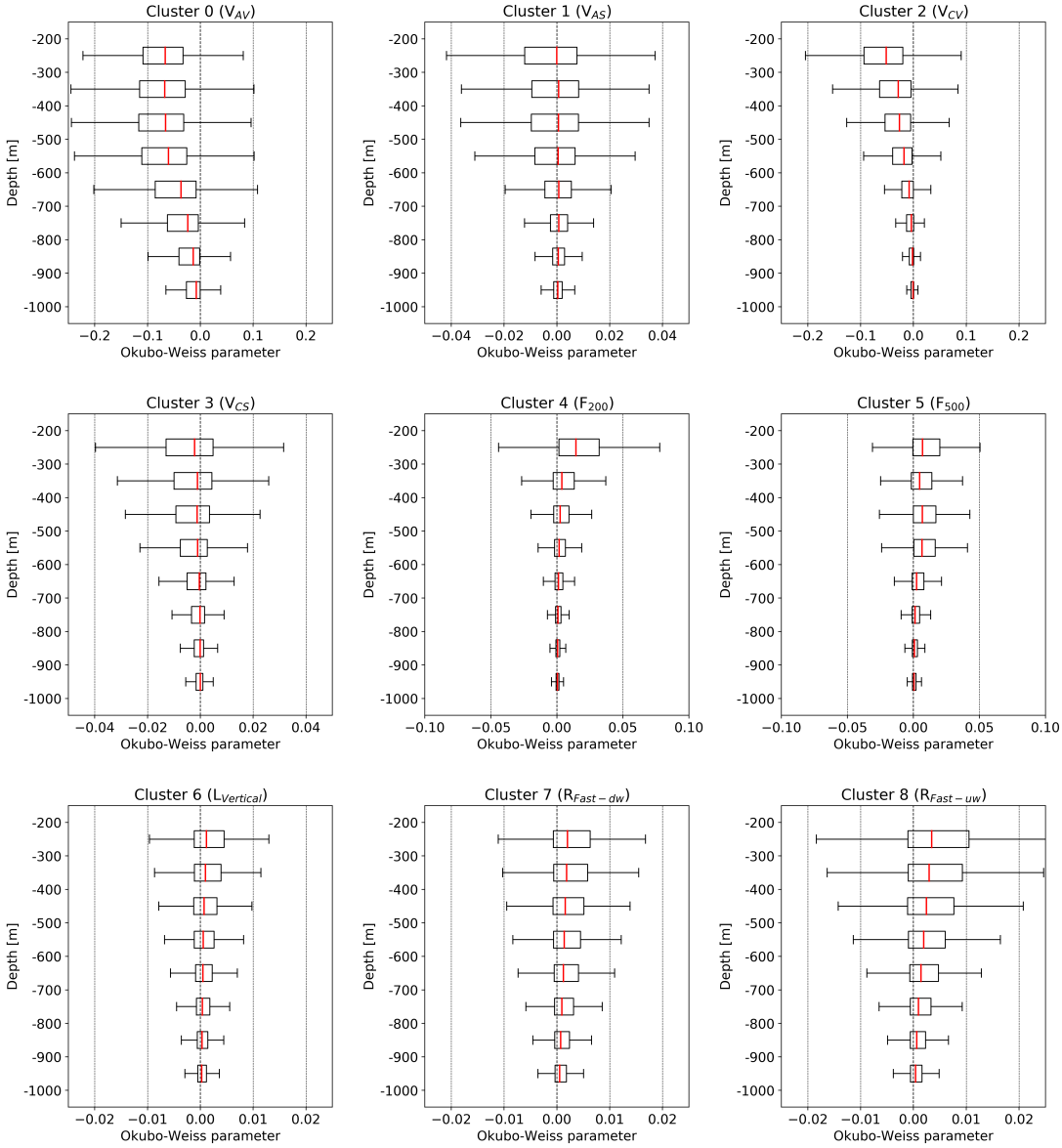

Figure S5. Full-period (7 years) Okubo-Weiss parameter along particle trajectory in different depth ranges.

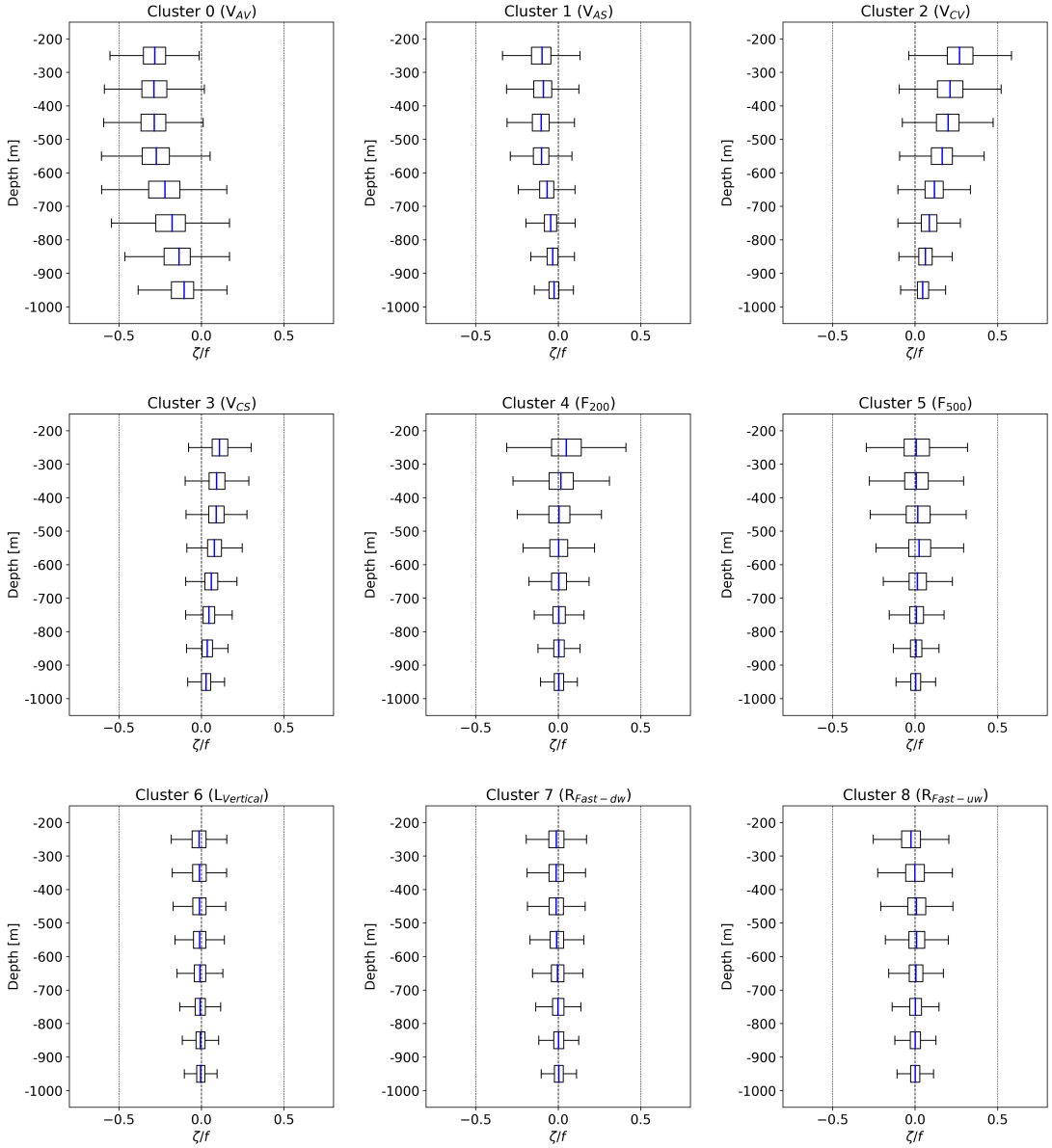

Figure S6. Full-period Relative vorticity along particle trajectory in different depth ranges.

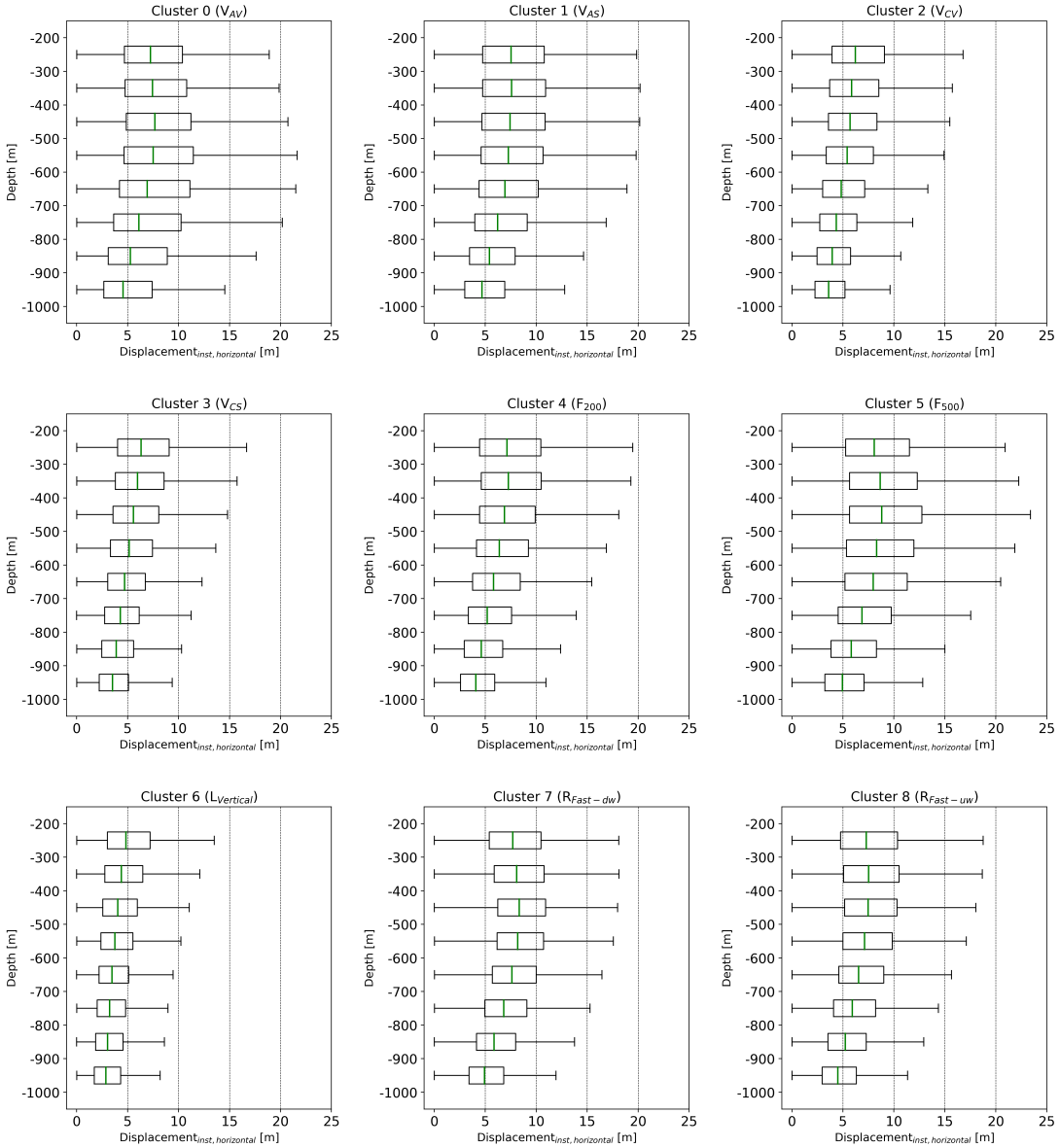

Figure S7. Full-period instantaneous horizontal displacement along particle trajectory in different depth ranges.

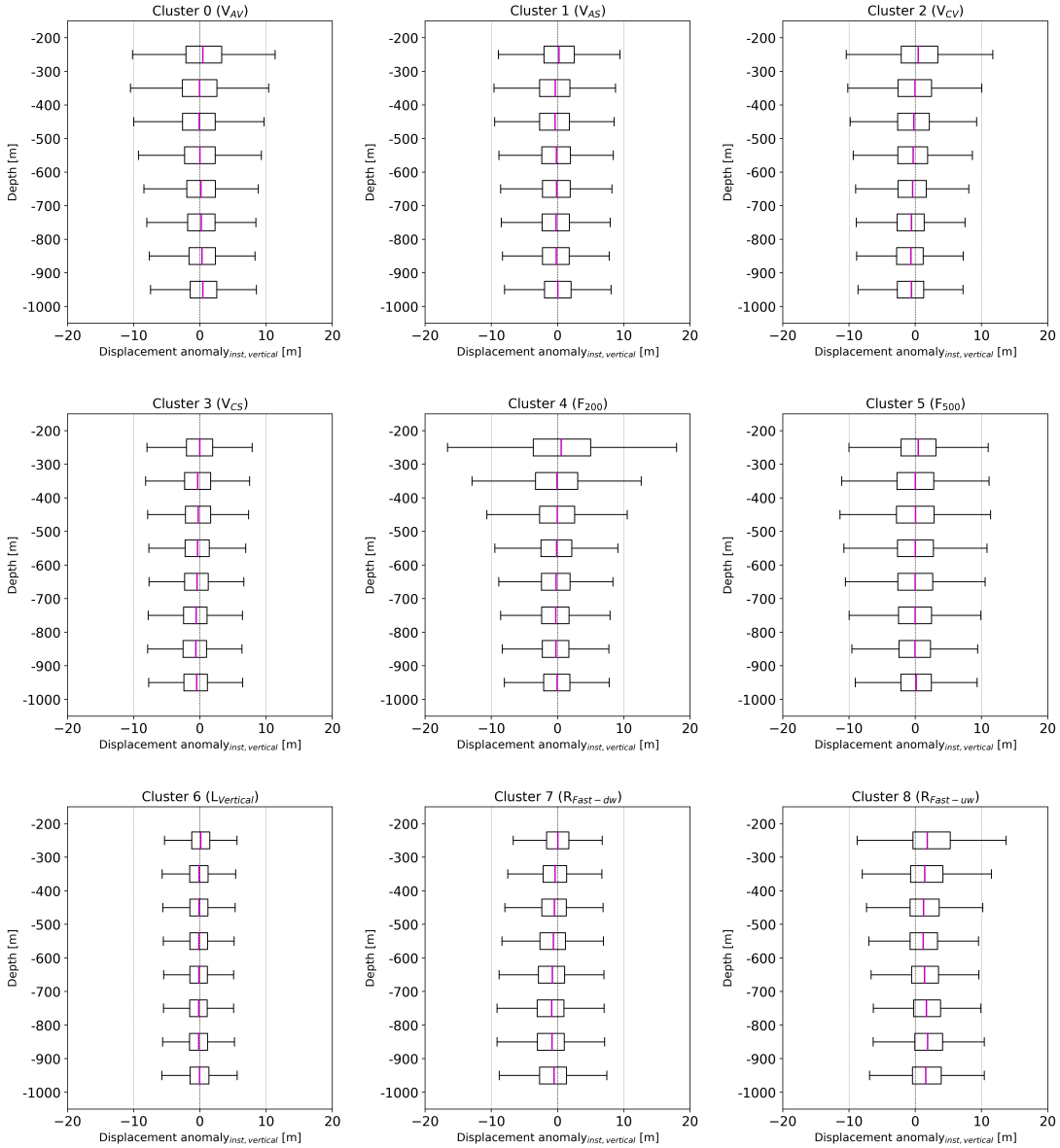

Figure S8. Full-period instantaneous vertical displacement anomaly along particle trajectory in different depth ranges.

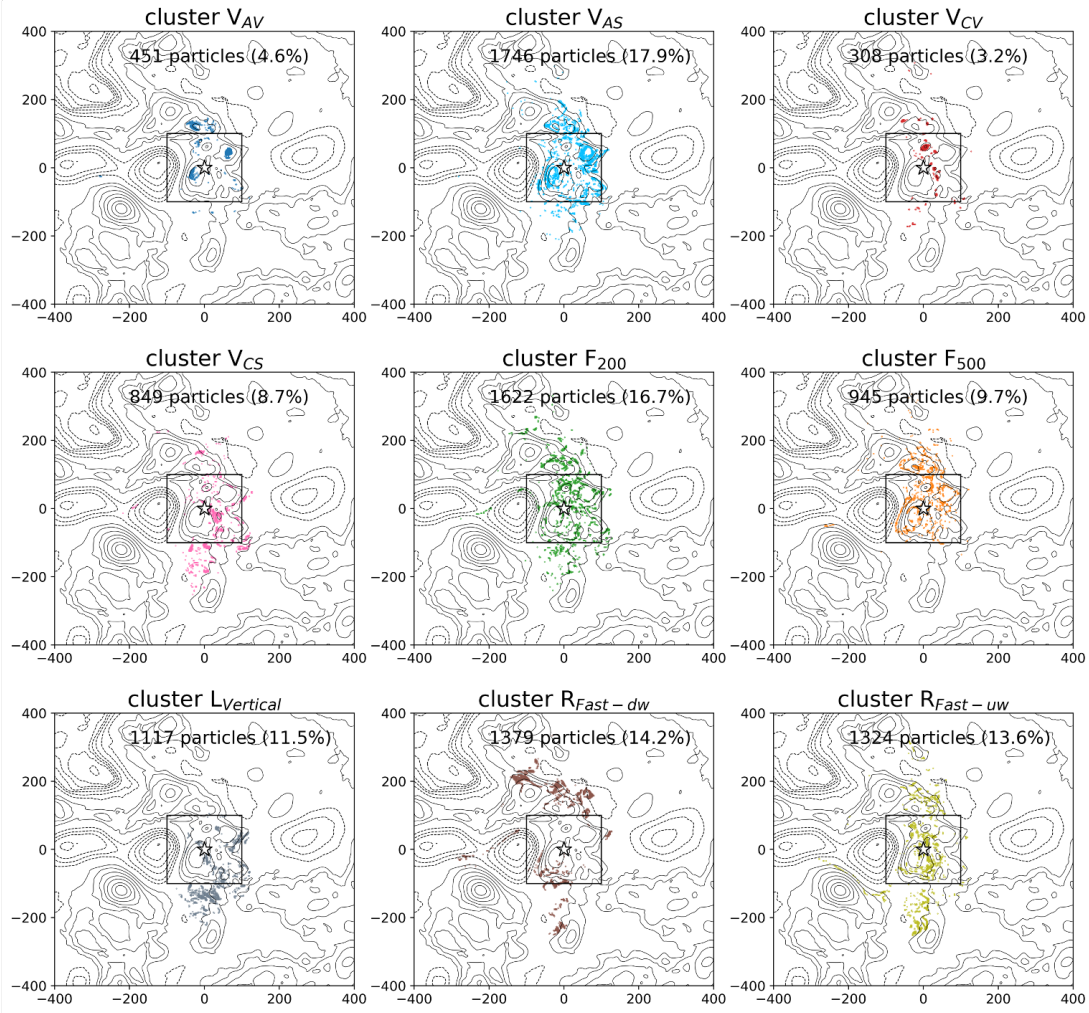

Figure S9. Separate distribution of clusters identified for particles released on 1 February 2003 reaching the 200 × 200 km target zone, with PAP site location (star marker) and the 200 × 200 km target zone (black square). The amount of particles in each cluster and its percentage are shown on top of each map.

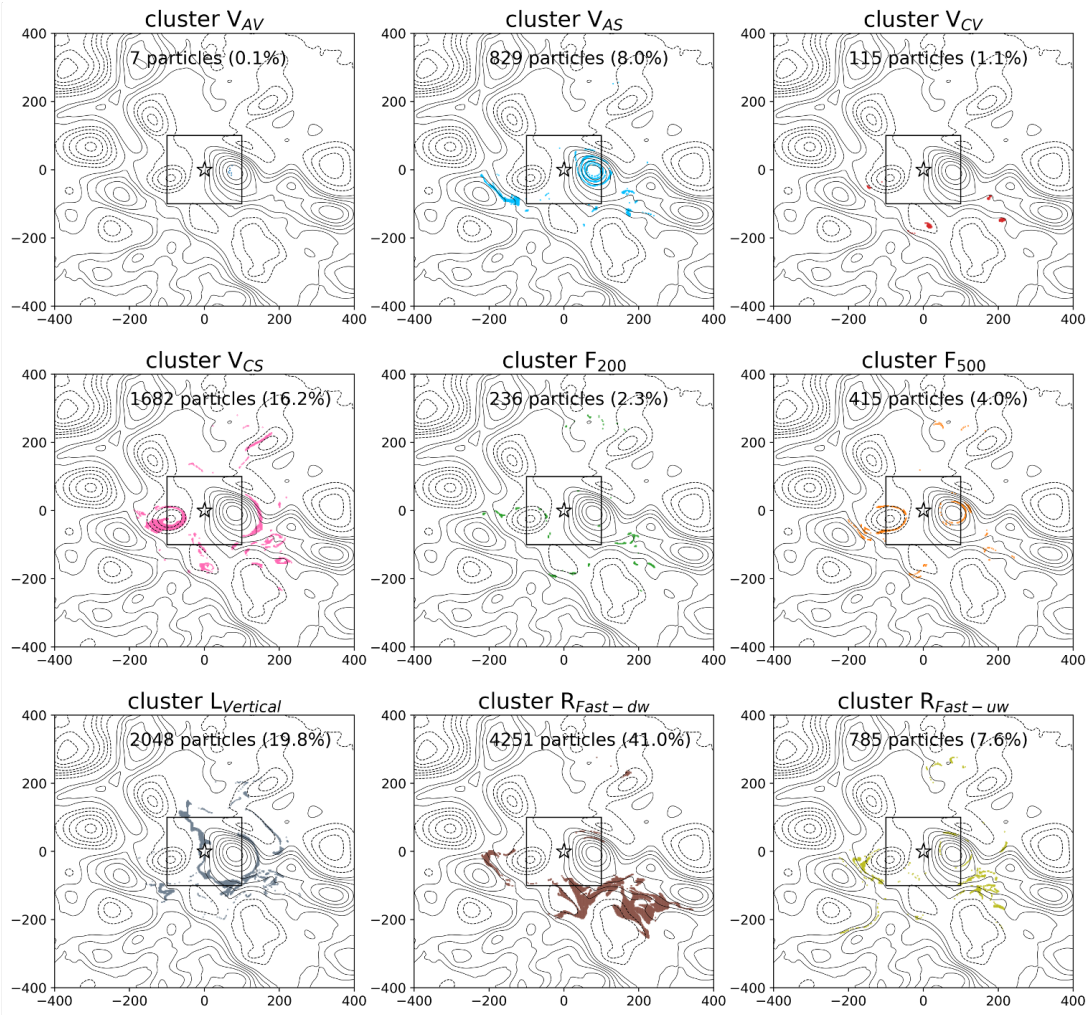

Figure S10. Same as in Figure S9, but for particles released on 10 September 2005 reaching the target zone.

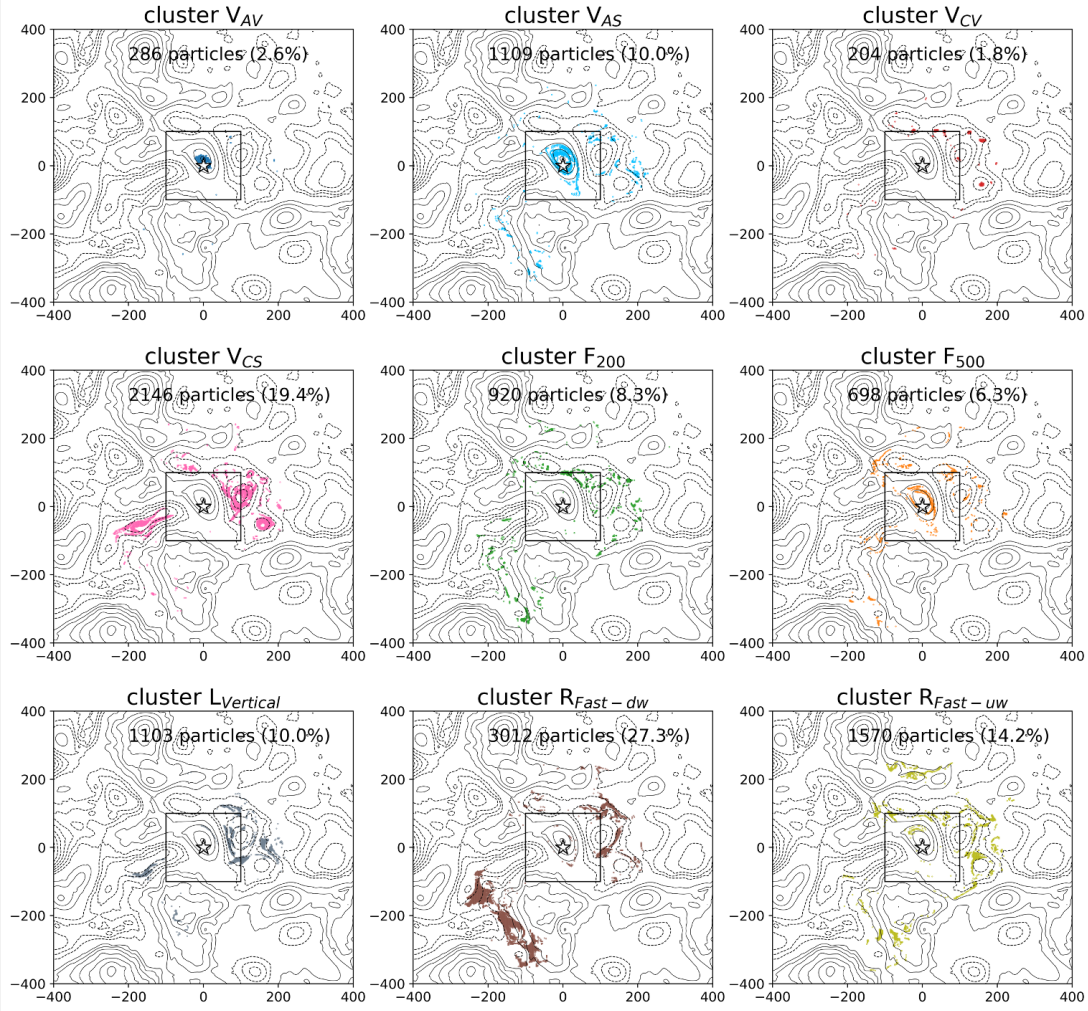

Figure S11. Same as in Figure S9, but for particles released on 25 March 2007 reaching the target zone.

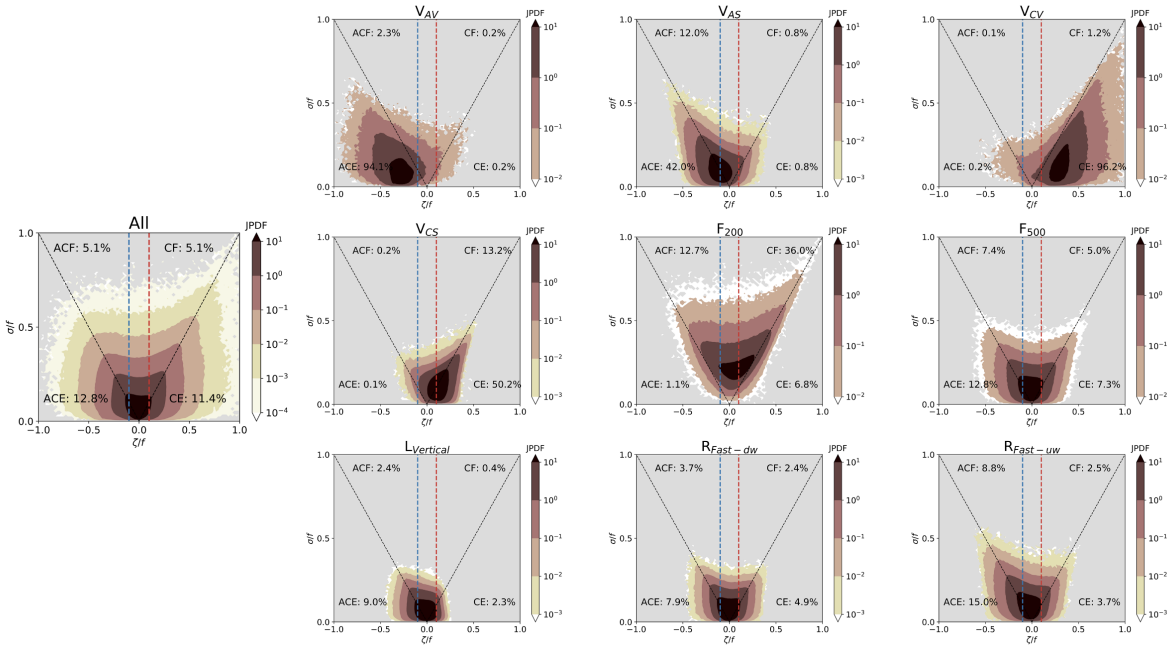

Figure S12. Vorticity-strain JPDF at 200 m for all particles and the nine clusters. By applying a criterion $|\zeta|/f > 0.1$ to define eddy/frontal regions, the lines of $\sigma = |\zeta|$ (black dashed lines) and $|\zeta|/f = 0.1$ (red/blue lines) decompose the eddy/frontal regions into four parts: cyclonic fronts (CF), anticyclonic fronts (ACF), cyclonic eddies (CE) and anticyclonic eddies (ACE). The numbers are the proportion of particles within each region in the JPDF.

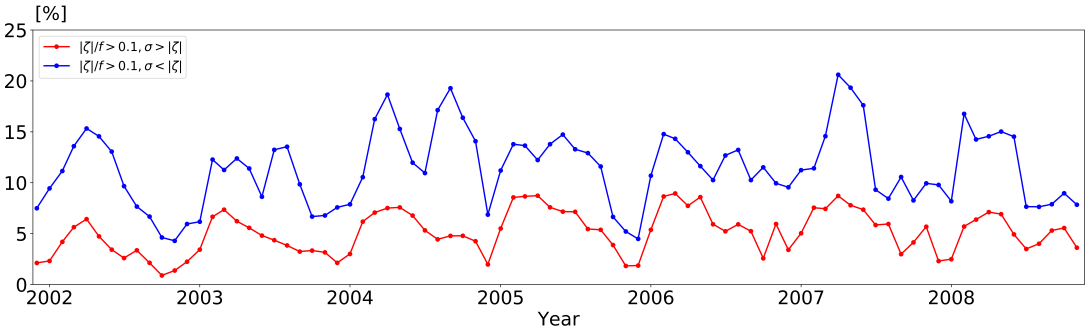

Figure S13. Monthly time series of the fraction of particles initialized in eddy/frontal regions, using the same criteria as in Figure S12. The red line represents the frontal region (CF+ACF: $|\zeta|/f > 0.1$ and $\sigma > |\zeta|$), and the blue line represents the eddy region (CE+ACE: $|\zeta|/f > 0.1$ and $\sigma < |\zeta|$).