Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(4), 1033-1049
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

# An Improved Water Vapor Trajectory Clustering Method and Its Application Analysis

**Jie Yu[1], Miao Cai[2*], Yuquan Zhou[2], Jianjun Ou[3]**

[1]College of Electronic Engineering, Chengdu University of Information and Technology, Chengdu, China
[2]CMA Cloud-Precipitation Physics and Weather Modification Key Laboratory (CPML), Beijing, China
[3]Shanghai by Weather Technology Co., Shanghai, China
Email: *caibird133@163.com

**How to cite this paper:** Yu, J., Cai, M., Zhou, Y.Q. and Ou, J.J. (2025) An Improved Water Vapor Trajectory Clustering Method and Its Application Analysis. *Open Journal of Applied Sciences*, **15**, 1033-1049.
https://doi.org/10.4236/ojapps.2025.154072

**Received:** March 19, 2025
**Accepted:** April 18, 2025
**Published:** April 21, 2025



## Abstract

In atmospheric water cycle research, water vapor trajectory analysis is a crucial tool for understanding the sources and transport pathways of precipitation water vapor. As a mainstream Lagrangian trajectory model, the HYSPLIT (Hybrid Single-Particle Lagrangian Integrated Trajectory) model provides water vapor trajectory data. However, its built-in trajectory clustering method has drawbacks, including long computation time and the loss of source point information. To address these issues, this study proposes an improved clustering method that incorporates a group-based computational optimization strategy and a weighted trajectory clustering approach to enhance computational efficiency and better capture dense water vapor source information. The study focuses on the cloud water resource high-value areas in Northwest China during the spring seasons from 2005 to 2015. Using the HYSPLIT model, backward water vapor trajectory tracking was conducted, followed by trajectory clustering analysis. The results demonstrate that the improved method reduces computational time costs, with experiments demonstrating an optimal reduction of up to 95.8%, while still preserving key source point information along water vapor transport pathways. Additionally, it enhances the identification of major water vapor transport routes. This improved method provides a more efficient and accurate data processing approach for large-scale water vapor trajectory analysis, offering valuable support for studying water vapor pathways in the atmospheric water cycle.

## Keywords

HYSPLIT, Improved Method, Water Vapor Trajectory, Cluster Method

## 1. Introduction

In the study of global and regional atmospheric water cycles, the primary research focus is on the role of water vapor within the cycle, such as examining the relationship between water vapor balance and precipitation and analyzing the sources of precipitation water vapor [1]-[5]. Commonly used research methods include numerical studies based on the Eulerian and Lagrangian approaches [6].

The research objectives are centered on the cyclic characteristics of the atmospheric water cycle, including precipitation recycling ratio [7] [8], water vapor source regions [9], and major water vapor transport pathways [10]. These studies extensively utilize water vapor trajectory data, which is typically obtained using the Lagrangian method [6] [9] [11] [12]. Additionally, isotopic tracing methods serve as another physical approach [13].

Currently, the mainstream Lagrangian models for generating water vapor trajectory data are HYSPLIT (Hybrid Single-Particle Lagrangian Integrated Trajectory) [14] and FLEXPART (FLEXible PARTicle dispersion model) [10]. These models, available through official websites or software platforms, efficiently provide water vapor trajectory data along with various meteorological parameters at each time step beyond just coordinate information. This capability offers essential data support for analyzing the characteristics of precipitation water vapor.

However, in studies requiring long-term and multi-origin water vapor trajectory data, the total number of trajectories can become exceptionally large. The original model software is designed to process only a limited number of trajectories. When dealing with large volumes of trajectory data, it typically outputs raw trajectory data for targeted post-processing [11].

When analyzing the main transport pathways of precipitation water vapor, it is necessary to cluster all backward trajectories of water vapor transport to the target precipitation region [15]. This clustering helps identify the major transport pathways and estimate the approximate source regions contributing to precipitation.

In the trajectory analysis method of the HYSPLIT model, clustering is primarily based on spatial variance (SV) between trajectory points and cluster spatial variance (CSV) to derive the total spatial variance (TSV). The clustering is performed by finding the combination that minimizes TSV. However, the original HYSPLIT method tends to blur the location information of dense water vapor source points and has a high computational cost. Some studies also use the K-Means clustering method [16]-[18], which similarly suffers from the issue of blurring dense source point distribution information.

To address these limitations of the original HYSPLIT clustering method, this study proposes a targeted improvement strategy. The improved method helps mitigate the loss of dense source point distribution information and reduces computational time to a certain extent.

## 2. Data and Methods

### 2.1. Global Data Assimilation System (GDAS)

The Global Data Assimilation System (GDAS1) is one of the operational systems

run by the National Centers for Environmental Prediction (NCEP) in the United States, which performs a series of computational analyses and forecasts. At the Air Resources Laboratory (ARL) of the National Oceanic and Atmospheric Administration (NOAA), NCEP model output data is archived for use in air quality transport and dispersion modeling. ARL archives the output from the Eta Data Assimilation System (EDAS) and GDAS using a 1-byte packing routine.

Both archives contain key meteorological elements such as U and V wind components, temperature, and humidity. Differences between these archives arise due to variations in the horizontal and vertical resolutions provided by NCEP and differences in specific application domains.

In this study, the GDAS1 dataset is used as the archived data source. The data files are arranged in chronological order based on meteorological time. Each file contains one week's worth of data per month, with suffixes w1 to w4 representing the first four weeks of the month, while w5 covers the remaining days of the month. Since no records are missing, data points can be precisely located within the files.

The available data period starts from December 2004 to the present. In this study, we selected a 10-year dataset from 2005 to 2010 and integrated it with the HYSPLIT model for analysis.

## 2.2. Backward Trajectory Model

HYSPLIT is a Lagrangian integrated trajectory model developed by the Air Resources Laboratory (ARL) of the National Oceanic and Atmospheric Administration (NOAA) in 1998 [19]. HYSPLIT can compute both forward and backward Lagrangian trajectories from any point in a three-dimensional space while outputting meteorological element data at specific time intervals along the trajectory, such as atmospheric temperature, potential temperature, water content, and relative humidity.

The core concept of HYSPLIT assumes that particles move with the airflow, and their movement trajectories are obtained by integrating their position vectors over time and space. The final position is calculated based on the initial position ( $P$ ) and the average velocity of the first estimated position ( $P'$ ):

$$P'\left(t+\Delta t\right)=P\left(t\right)+V\left(P,t\right) \tag{1}$$

$$P'\left(t+\Delta t\right)=P\left(t\right)+0.5\times\left[V\left(P,t\right)+V\left(P',t+\Delta t\right)\right]\Delta t \tag{2}$$

where $\Delta t$ is the time step, which is variable, and $U_{\max}$ represents the maximum wind speed. It is required that $\Delta t$ be chosen such that the movement distance of the air mass within one time step does not exceed 0.75 grid points. The time step used in this study is 6 hours.

## 2.3. Trajectory Clustering Method in HYSPLIT

The HYSPLIT website tutorial introduces the clustering methods used in the clustering analysis:

https://www.ready.noaa.gov/documents/Tutorial/html/traj_cluseqn.html.
The main clustering process is as follows:

During the clustering process, the spatial variance (SV) between each endpoint (P) along the trajectory within the cluster is calculated, as shown in Equation (3).

$$SV_{i,j} = \sum_k \left( P_{j,k} - M_{i,k} \right)^2 \tag{3}$$

The sum is taken over the number of endpoints along the trajectory, where $P$ and $M$ represent the position vectors of the individual trajectory and the average trajectory of the cluster, respectively. The Cluster Spatial Variance (CSV) is the sum of the spatial variances of all the trajectories within the cluster, as shown in Equation (4). The Total Spatial Variance (TSV) is the sum of the CSVs of all the clusters, as shown in Equation (5).

$$CSV_i = \sum_j SV_{i,j} \tag{4}$$

$$TSV_i = \sum_i CSV_i \tag{5}$$

The clustering process begins by assigning each trajectory to its own cluster, resulting in a total of N clusters. In each iteration, as two clusters merge, the number of clusters decreases by one. Therefore, after the second iteration, there will be N − 1 clusters (one cluster containing two trajectories, and the remaining clusters each containing one trajectory). This process continues until only one cluster remains.

In each iteration, the Total Spatial Variance (TSV) for each potential cluster combination is calculated, which involves adding the trajectories from cluster 1 to cluster 2, from cluster 3 to cluster n, and so on, until all remaining clusters are combined and their TSV values are calculated. The combination with the minimum TSV is selected as the optimal cluster and is merged. The result is passed to the next iteration. This process continues until only one cluster remains or the target number of clusters is reached.

The number of trajectories influences the computational time, so a large number of trajectories results in longer clustering times. The original method mentioned later in the text refers to the clustering method introduced in this section.

## 2.4. Improved Clustering Method

Section 2.3 mentions the conventional clustering TSV method, but it has the following shortcomings: 1) it blurs the position of dense trajectory source points, and 2) it requires significant computational resources and time. To achieve better clustering results, the following solutions are proposed for these two issues.

For the problem of large computational resource consumption and slow calculation speed, the solution is to group the trajectory data. Grouping and batch processing find the cluster combinations with the smallest TSV locally, whereas without grouping, the combinations are determined globally. Based on subsequent grouping experiments and comparative analysis, a set of grouping rules is proposed, which helps retain key information from the original trajectories while re-

ducing computation time.

In the original method, source point information is lost, mainly due to using equal-weight spatial variance as the cost function to select the optimal merging combination. To solve this issue, the proposed solution is to assign exponentially decreasing weights to all points from the source to the endpoint when calculating the spatial variance. The purpose of this is to reduce the impact of variance caused by intermediate transport points during clustering and to increase the variance cost for points closer to the initial source.

At the same time, to better preserve the source point distribution information, the number of original trajectories for the clustered average trajectory is introduced as a new weight in the calculation of TSV. The goal is to merge sparse trajectories into clusters with a larger number of nearby trajectories.

Figure 1 shows the flowchart of the improved method. Suppose the initial total number of trajectories is 20,050, which will be divided into 40 groups. However, trajectories within a group may not be consistent, so the excess trajectories will be placed into already assigned groups, making the last group contain 550 trajectories. The 40 groups will undergo clustering one by one, which constitutes the first batch of clustering. Specifically, based on the original clustering method in HYSPLIT, weights are introduced. In the original method, both CSV and TSV are calculated with equal weights. The improved method introduces nonlinear weight W (as shown in Equation (6)) when calculating CSV, and the number of trajectories within a cluster is introduced as a weight N (as shown in Equation (7)) when calculating TSV.

$$CSVW_i = \sum_j W_j \cdot SV_{i,j} \tag{6}$$

$$TSVN_i = \sum_i N_i \cdot CSV_i \tag{7}$$

On this basis, clustering is performed for each group of trajectories. When the total number of clusters within a single group reaches the stopping threshold (assumed to be 200), the clustering for that group ends, resulting in 40 groups with 200 clusters of trajectories. For each group of 40, the average trajectory of each cluster is calculated, resulting in 200 average trajectories per group. These are then merged to form a total of 8000 average trajectory groups. Since these 8000 average trajectories exceed the minimum number of trajectories required for the grouping (1200 trajectories), the next round of grouping calculations is performed. After three rounds of grouping calculations, a total of 1200 average trajectories are achieved. At this point, the 1200 average trajectories will no longer be grouped, but will be directly clustered until the final target cluster classification result is reached.

To facilitate human differentiation, this study visualizes only the 50 cluster average trajectories, setting 50 clusters as the final number of cluster classifications.

## 3. Results and Analysis

### 3.1. Model Initialization

Reference to Figure 2 shows the multi-year average distribution of cloud water

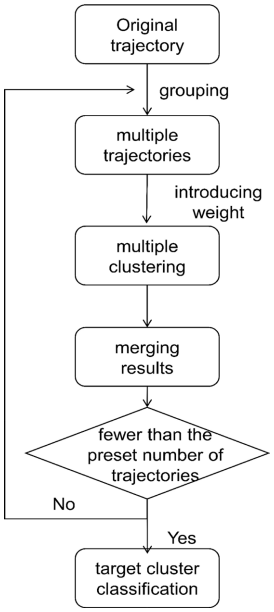


**Figure 1.** Flowchart of the improved clustering method calculation.

resources (CWRs) in the northwest region of China during the spring. A CWRs minimum threshold of 550 mm is selected, and 23 grid points located in the northern part of the northwest region (Target Area 1; denoted as T1) and 14 grid points located in the southeastern part (Target Area 2; denoted as T2) are selected. To study the water vapor sources and main transport paths of higher CWRs in the northwest region, backward water vapor tracking is conducted for the high-value CWRs grid points. The selected grid points are used as the initial coordinates for backward water vapor trajectory tracking.

For the water vapor tracking initialization scheme, we refer to Shi *et al.* [20] and select six vertical height levels: 100 m, 500 m, 1500 m, 3000 m, 5000 m, and 9000 m above the ground. The 100 m, 500 m, and 1500 m levels are low levels, 3000 m and 5000 m are mid-levels, and 9000 m is the high level. As a result, T1 and T2 regions have 120 and 102 points, respectively. The tracking initialization time is selected from the spring of 2005 to 2015 (March-May) at 00:00, 06:00, 12:00, and 18:00 (UTC) each day. The global water vapor cycle update period is approximately 10 days [21]. Therefore, a 10-day backward tracing is performed for each spatial point, outputting the spatial three-dimensional information (latitude, longitude, and altitude) and water content at 6-hour intervals. In this way, each trajectory can obtain information from 240 points. However, each trajectory undergoes a validity check, and if invalid values are detected, the trajectory is removed.

The criteria for filtering the water vapor movement trajectory are as follows: To identify the trajectories that contribute to cloud water over the target region, the water content must decrease by at least 0.1 g/kg from the moment the trajectory enters T1 or T2 to its endpoint, indicating that precipitation has occurred along

the trajectory or that the water vapor has condensed into cloud water in the target region. The number of obtained water vapor trajectories is shown in Table 1. It can be observed that the number of low- and mid-level trajectories is much larger than that of high-level trajectories. Since the water vapor trajectories in region T2 are more complex than those in region T1, this paper primarily analyzes the trajectories in region T2.

Table 1. Number of precipitation water vapor trajectories in T1 and T2 regions (unit: number of trajectories).

| | T1 | T2 |
|---|---|---|
| Low level | 60,362 | 53,808 |
| Mid level | 38,364 | 27,425 |
| High level | 999 | 2983 |

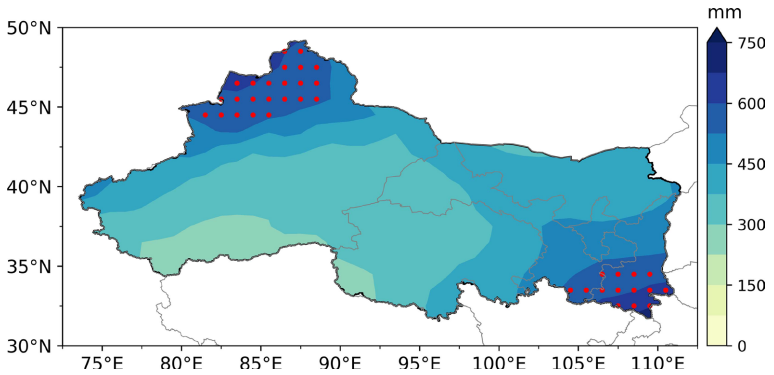


Figure 2. Multi-year average distribution of spring CWRs in the Northwest China region and high-value areas (red grid points).

### 3.2. Limitations of the Original Clustering Method

Due to the large number of selected backward-traced trajectories, as shown in Figure 3, tracing the mid-level water vapor in the T2 region (medium quantity) results in numerous source points, and the paths from the source points to the target area endpoints are in the thousands. To obtain a clear representation of the main water vapor transport paths, clustering operations must be performed on the large number of trajectory data. If trajectories are randomly merged, there will be poor differentiation between clusters. Figure 4 shows the averaged trajectory paths after randomly merging 200 trajectories per group. As can be seen, the clusters are hard to distinguish. According to the original clustering method provided by HYSPLIT, the method works well when clustering a small number of trajectories, but when clustering a large number of trajectories, the following issues arise:

- Fuzzy water vapor dense source area location information: due to the high number and density of source points, the obtained source points are gridded into $1^{\circ} \times 1^{\circ}$ cells (see Figure 4), and the number of source points in each grid is counted. From the grid distribution, it is evident that the main source points are located near the Himalayas and central China. However, the original clus-

tering method results in fuzziness in these features, as discussed in Section 3.3.2.

- Long computation time and high memory resource usage: in the conventional clustering method, for N trajectories, merging each pair of trajectories requires $N(N-1)/2$ calculations. For example, with 27,425 trajectories in the T2 region, approximately 380 million calculations are needed, which not only consumes a large amount of computing resources but also takes a long time to compute.

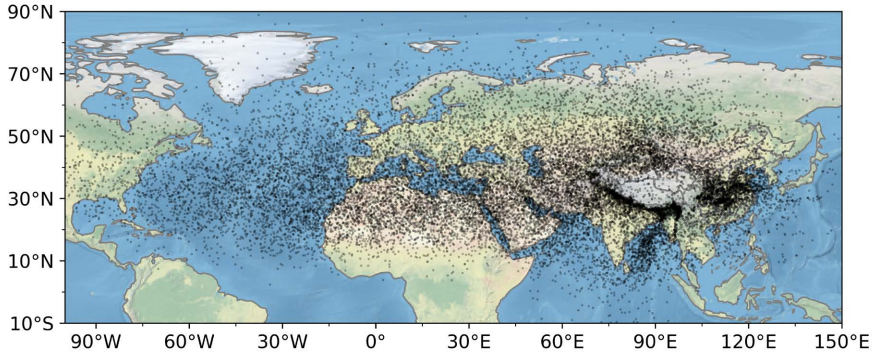


**Figure 3.** Distribution of source points for mid-level water vapor tracing in the T2 region.

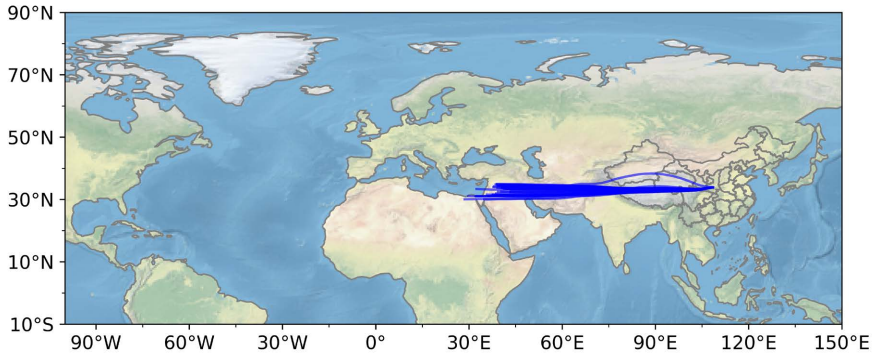


**Figure 4.** Random clustering results of mid-level water vapor in the T2 region, with blue lines representing the clustered average trajectories of 200 random trajectories.

## 3.3. Comparison of Improvement Effects

### 3.3.1. Impact of Grouping

The impact of grouping on the clustering results using the original method needs further validation. The specific grouping scheme is as follows: divide the large group into multiple smaller groups, and then perform clustering on each small group one by one (since this section only discusses the effect of grouping, the original clustering algorithm is used), until the target number of clusters is reached. After a batch of grouping clustering is finished, the trajectories within each group are averaged to obtain a cluster average trajectory, and then combined into a new large group. If the number of trajectories in the large group is still large, a new round of grouping clustering can be performed. Therefore, the experiment mainly has two variables: one is the minimum number of trajectories per group (M), and

the other is the number of clusters to be formed (N).

First, let's discuss the setting of the minimum number of trajectories per group (M). According to the principle of controlling variables, N is set as a fixed value, and here it is temporarily set to 100. Meanwhile, the final target number of clusters is set to 50. The experiment uses the original water vapor trajectories from the middle layer of the T2 region as shown in Table 1, and randomly selects 2500 trajectories from them. The experiment will test M values from 200 to 700 with a step size of 100, and also provide the clustering results of these 2500 trajectories without grouping, comparing the results with those of the grouped cases.

Figure 5 shows the clustering result after 10 clusters are formed using the original clustering method without grouping. The red trajectories represent the cluster average trajectories with more than 60 trajectories in the cluster. Figure 6 shows the grouping clustering results for different M values, and Table 2 lists the corresponding computation times for different M values (due to differences in computer performance, the same computer was used for all calculations). It is visually evident that the clustering results vary under different M values. From an overall clustering perspective, when M is set to 400, the result is closest to that of the ungrouped clustering. It can also be observed that as M decreases, the time required for computation decreases, but the time reduction efficiency decreases non-linearly. In contrast, the original method without grouping requires 9502.5 seconds, demonstrating that the grouping approach significantly reduces time costs. As shown in Table 2, the time-saving ratio ranges from 88.8% to 98.1%.

Based on Shannon's theorem, the information entropy of each group can be analyzed. When M is greater than 400, the trajectories within each group need to be clustered into N clusters, *i.e.*, 100 clusters. This causes the information entropy of the trajectories within each group to decrease significantly. As M increases, the amount of original trajectory information retained in each group decreases. On the other hand, when M is less than 400, although the information compression rate within each group is low, the overall data sampling rate is insufficient, which leads to unsatisfactory final clustering results.

In summary, in the ungrouped case, the original method searches for the optimal solution globally, while the grouped clustering scheme approximates the search for the optimal solution in local groups. Therefore, different M values will produce different clustering effects, but eventually, a reasonable M parameter will approach the optimal solution.

Table 2. Computation time required for different M values.

| M Value (Tracks) | Time (Seconds) | Time Reduction Rate (%) |
|---|---|---|
| 700 | 1070.2 | 88.8 |
| 600 | 603.8 | 93.6 |
| 500 | 534.6 | 94.4 |
| 400 | 393.6 | 95.8 |
| 300 | 251.3 | 97.4 |
| 200 | 177 | 98.1 |

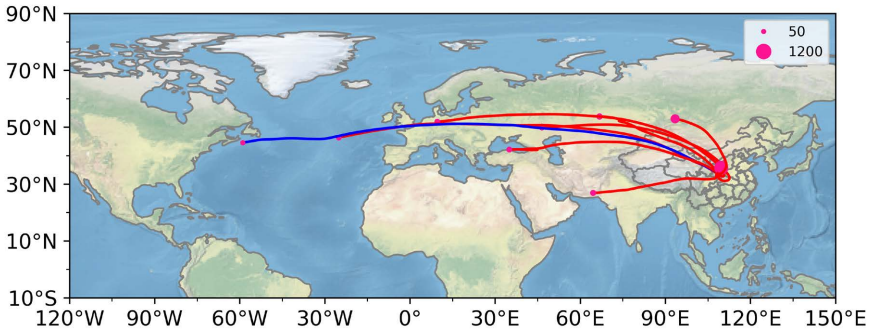


**Figure 5.** Clustering results of the original method without grouping. The red lines represent clusters with more than 60 trajectories.

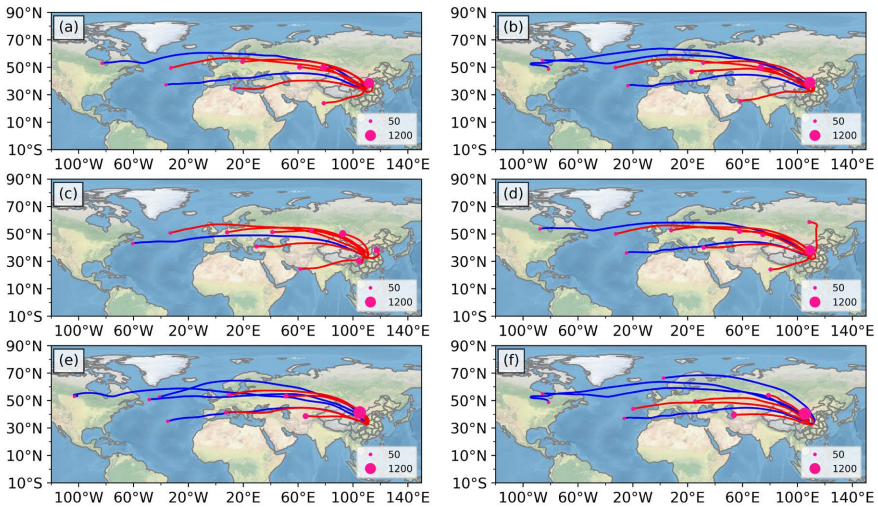


**Figure 6.** Clustering results of the original method under different grouping conditions. (a) M = 700, (b) M = 600, (c) M = 500, (d) M = 400, (e) M = 300, (f) M = 200, with N = 100 for all cases. The red lines represent clusters with more than 60 trajectories.

Next is the discussion of the N parameter. In the previous section, 400 was identified as a suitable value for M. Now, with M set to 400, we will discuss the N parameter. With a fixed M value, as the N value approaches M, the trajectory information compression rate within the original group is minimized. Conversely, as N decreases, the compression rate increases. The experimental setup for N is similar to the one for M, with N values ranging from 50 to 250 in steps of 50.

As shown in **Figure 7**, when M is fixed at 400, there are significant differences in clustering results with different N values, especially in the larger clusters, which are a key focus of the study. In the ungrouped result, a large cluster appears near the vapor convergence center (T2 area), containing thousands of trajectories. Using the starting point of the cluster's average trajectory as a reference, it can be seen that when N is 50 or 400, the starting point of the large cluster is closest to the ungrouped scenario, and when N is 100, the overall clustering distribution is closer to the ungrouped result. However, when N is between 150 and 250, the starting point of the large cluster shifts significantly to the northwest. Overall,

when N is 100, the clustering distribution is more similar to the ungrouped result.

Table 3 shows the time required for calculation at different N values. It can be seen that as N decreases from 250 to 50, the time required decreases, but the time reduction is relatively limited. This may be due to the fact that as N becomes smaller, more batches need to be calculated, thus limiting the time savings. However, N should not be too small. As N decreases, the information entropy corresponding to the trajectories within the group decreases more significantly, so N should not be set too low.

**Table 3.** Time required for calculation at different N values with M constant.

| N Values (Tracks) | Time (seconds) |
|---|---|
| 250 | 448.4 |
| 200 | 440.6 |
| 150 | 371.2 |
| 100 | 393.6 |
| 50 | 317.2 |

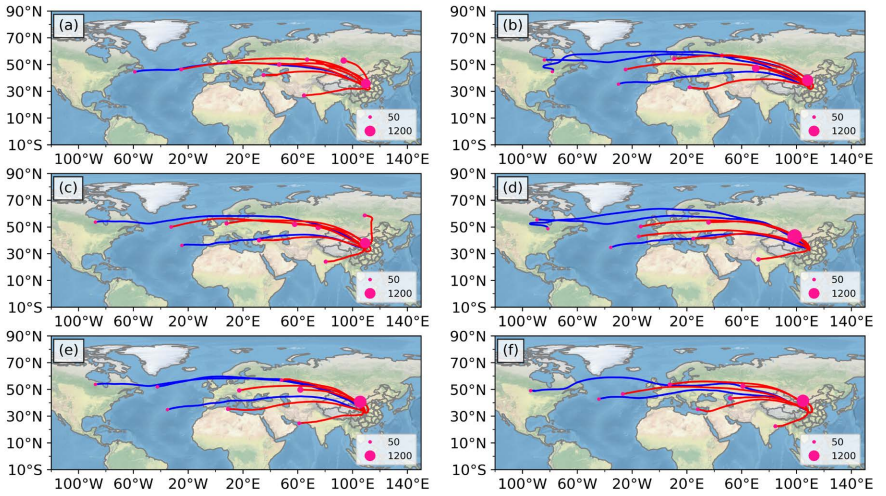


**Figure 7.** Shows the clustering results with M = 400 and different N values. (a) Without grouping, (b) N = 50, (c) N = 100, (d) N = 150, (e) N = 200, (f) N = 250. The red lines represent clusters with more than 60 trajectories.

Based on the above discussion, the selection of M and N values can be flexible within the defined range, and this range is determined by how well the grouping approach approximates the ungrouped results. As mentioned earlier, grouping involves searching for an optimal solution in a local context, while ungrouped data searches for an optimal solution in a global context. The purpose of grouping is to accelerate the computation process, and with limited computational resources, it is necessary to evaluate computational efficiency and determine a suitable combination of M and N values. Through experimentation, the goal is to maximize trajectory sampling rate (for M), reduce computational time costs (for

both M and N), and retain more original information (for N). In this study, the combination of M = 400 and N = 100 will be used for subsequent processing.

### 3.3.2. The Impact of Introducing Weights

In this study, two weight factors are introduced. One is an exponential weight (with base 2) assigned during the calculation of Cluster Spatial Variance (CSV), and the other is the number of trajectories within a cluster assigned to the corresponding cluster's CSV when calculating the Total Spatial Variance (TSV), thereby altering the classification results. To investigate the impact of the introduced weights on the results, 10,000 random trajectories from the T2 low-level water vapor trajectory data are selected for comparison between the improved method and the original method. At the same time, to shorten the computation time, the following discussions use the grouping approach with a grouping scheme of M = 400 and N = 100.

Figure 8 shows a grid map of the 10,000 test trajectory source points, where each grid point corresponds to the number of original trajectory source points belonging to that grid. From the figure, two areas with dense source points are visible: one near 80˚E, 50˚N, and the other near 110˚E, 40˚N. Figure 9(a) shows the clustering result of 50 clusters calculated by the new method, while Figure 9(b) shows the clustering result from the original method. It is evident that the new method captures all three dense source point areas within the yellow box in Figure 8, whereas the original method only captures the most densely concentrated area and fails to effectively capture regions with lower densities. For the moderately dense area near 80˚E, 50˚N, the original method divides this area into two clusters, while the new method reasonably merges it into a single cluster.

In summary, the new method performs better in terms of sensitivity to dense source point regions compared to the original method.

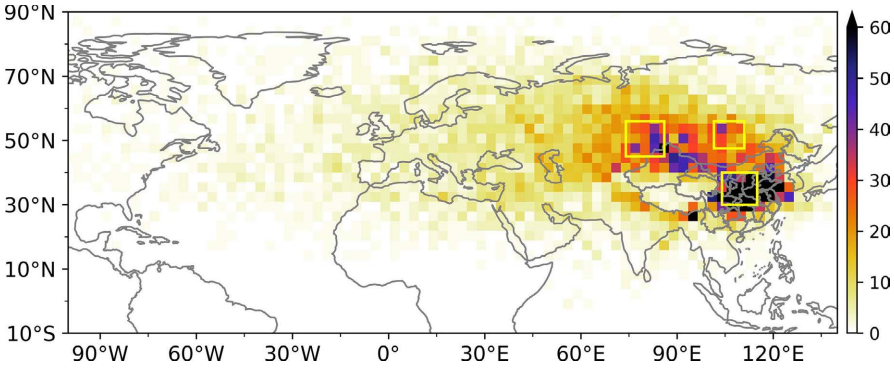


Figure 8. Experiment vapor trajectory source point gridding result, unit: count.

### 3.3.3. Clustering Results of T2 Region Mid-Layer Vapor Trajectories

Figure 10 shows the source point distribution of the entire mid-layer vapor trajectories in the T2 region. It can be seen that the source points are mainly concentrated in central China and the northwest of China, with some distribution in the southern part of the Tibetan Plateau, and the rest are more evenly distributed

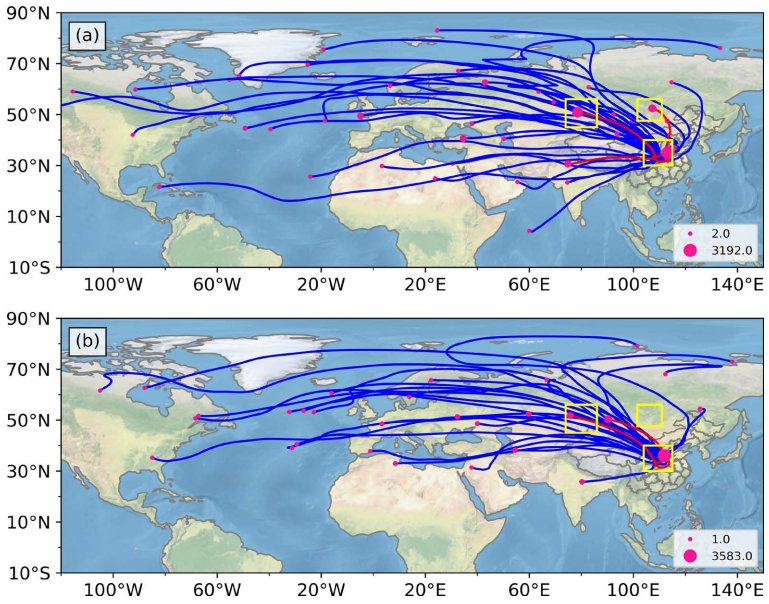


**Figure 9.** Comparison of the effects after introducing weights, (a) results with introduced weights; (b) results with the original method. Red represents the cluster average trajectory with more than 800 trajectories within the cluster.

across the Eurasian continent. From the overall source point distribution, the vapor shows a clear west-to-east transport trend, which is related to the location of the T2 region in the westerly wind belt.

Figure 11 presents the clustering results of 50 clusters for mid-layer water vapor trajectories in the T2 region using the improved method and the K-means method. It can be observed that both methods perform similarly in identifying large clusters (with more than 800 trajectories), successfully detecting major water vapor source regions in central China, the southern Tibetan Plateau, northwestern China, and even farther areas in Central Asia. For the dense source region in central China, the improved method provides more accurate results than the K-means method.

Although both methods can identify the general location of these sources, there is a significant difference in the number of trajectories within the same large cluster, with the improved method exhibiting a stronger clustering effect. The largest clusters differ by nearly threefold between the two methods. Further analysis reveals that while both methods identify similar dense source regions, the K-means method produces multiple large clusters within the same source region, particularly in central China, the southern Tibetan Plateau, and northwestern China, where at least two separate clusters are detected. These overlapping clusters should ideally be merged into a single cluster.

In summary, the improved method demonstrates superior performance in identifying dense source regions and primary transport pathways, whereas the K-means method is better suited for the broad-scale identification of water vapor source regions.

From the clustering results of the improved method in Figure 11, it can be seen

that the main transport paths of the T2 region's mid-layer vapor, from nearest to farthest, are as follows: 1) One vapor source from central China, which occupies a high share of vapor, indicating that a significant portion of the vapor in the T2 region comes from local land evapotranspiration, resulting in recycling; 2) One vapor source from the southern Tibetan Plateau, where the vapor is hindered by the high mountainous terrain of the Tibetan Plateau. The 10-day backward tracing in this study also shows that the plateau can retain large amounts of moisture from the Indian Ocean region for more than ten days, which is consistent with some earlier studies [11]; 3) One vapor source starting from the Tianshan Mountains in northwest China, where some vapor is intercepted by the topography of the plateau and carried southward, while the other part is intercepted by the northwest of China. The Tianshan Mountains' terrain influences the vapor source region starting from this mountain range; 4) Other vapor transport paths follow the westerly jet stream, originating from the northern part of Africa, the Middle East, and even farther from the Atlantic Ocean.

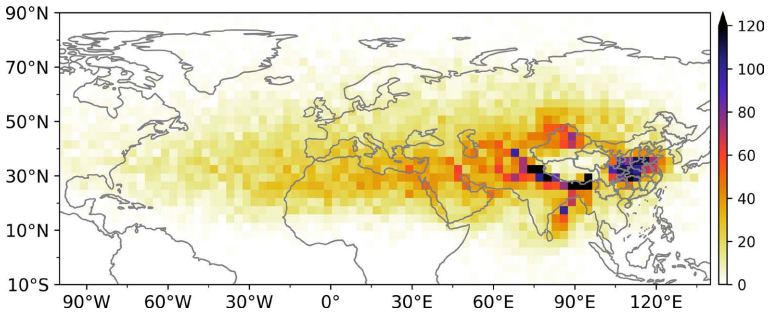


**Figure 10.** Gridded distribution of vapor trajectory source points in the middle layer, unit: count.

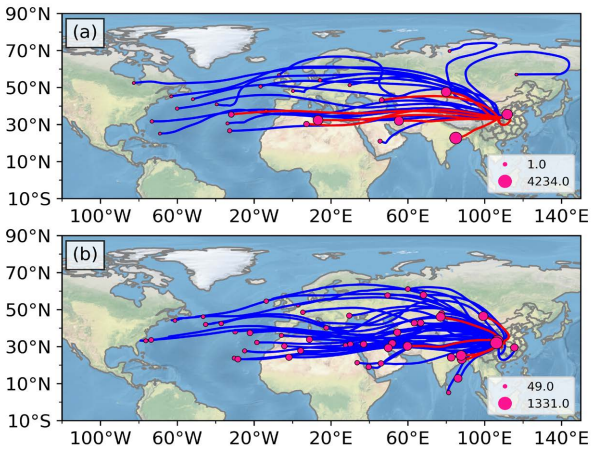


**Figure 11.** Clustering results of the improved method and the K-means method for vapor trajectories in the middle layer of T2. (a) Represents the improved method; (b) Represents the K-means method. Note: the red color indicates the cluster average trajectory with more than 800 trajectories within the cluster, representing a large cluster, clusters composed of anomalous trajectories have been removed from the figure.

## 4. Discussion and Conclusion

### 4.1. Conclusion

This paper focuses on the water vapor source analysis of the high cloud water resource areas in Northwest China. It traces the water vapor over these areas and obtains a large number of vapor trajectories. Based on this, a targeted improvement to the original HYSPLIT trajectory clustering method is proposed, and the clustering effect of the improved method is investigated. The results indicate:

1) The grouping calculation proposed in the improved scheme can significantly reduce the computation time of the total cluster variance clustering method, addressing the issue of high computational cost in the original method. The experiments show that when the minimum trajectory number per group is set to M = 400 and the endpoint cluster number to N = 100, the clustering time cost can be reduced by 98.5% while maintaining clustering results similar to the original method.

2) The improved scheme introduces weights. By adding nonlinear weights in the calculation of cluster variance and total variance and incorporating the number of trajectories within the cluster as a weight, the method effectively captures locally dense source point distribution information, reducing the original scheme's loss of dense source point information.

3) This study performed water vapor tracing in the high-value cloud water resource area in the southeastern part of Northwest China. The clustering results from the improved method reveal the main transmission paths of middle-layer water vapor in T2 area: the water vapor is mainly transported via local recycling from Central China, the southern part of the Tibetan Plateau, and the Tianshan Mountains in Northwest China, as well as some from the Middle East, North Africa, and the Atlantic, converging through the westerly jet stream to form cloud water resources over the target study area.

### 4.2. Discussion

This study represents an improvement based on the original HYSPLIT clustering method, driven by research objectives. The most significant effect of the proposed improvement is the substantial reduction in computation time, accompanied by a shift from global optimization to local optimization in the clustering results. The original clustering method did not support parallel computation, whereas the improved scheme with grouping enables parallel computation, further reducing computation time and cost. It is worth noting that parallel computing was not used in this paper.

In the analysis of source point distribution in this paper, it can be observed that some grid points have exceptionally dense distributions of source points. Based on the principle of retaining source point information, pre-clustering measures can be applied. Simple density thresholds or density-based clustering algorithms (e.g., DBSCAN) could be used for pre-clustering.

When classifying large numbers of trajectories into clusters and calculating the

cluster-averaged trajectories, only approximate vapor transport paths from the source area to the target area can be obtained. To gain more precise paths, future work could consider applying grid processing to the latitude and longitude mesh, thereby associating vapor trajectories with specific grid borders. This would allow for finer water vapor transport paths and be helpful for localized, detailed research.

The study period and region in this paper focus on the southeastern part of Northwest China from 2005 to 2010. However, this study is not limited to this specific period and region—the proposed improvement method can also be applied to research in other regions and time periods. The proposed method can provide stronger scientific support for weather forecasting, extreme weather prediction, climate model optimization, and regional water cycle studies, offering important assistance in addressing climate change and disaster prevention and mitigation.

## Acknowledgements

This work was supported by the Scientific Research Plan Project of Bayingol Mongolian Autonomous Prefecture in Xinjiang (Grant No. 202318).

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Whiteman, D.N., Rush, K., Rabenhorst, S., Welch, W., Cadirola, M., McIntire, G., *et al.* (2010) Airborne and Ground-Based Measurements Using a High-Performance Raman Lidar. *Journal of Atmospheric and Oceanic Technology*, **27**, 1781-1801. https://doi.org/10.1175/2010jtecha1391.1

[2] Ma, J. and Guo, W. (2025) A Coupling Model for Global Average Water Vapor and Temperature Change. *Climate Dynamics*, **63**, Article No. 141. https://doi.org/10.1007/s00382-025-07638-3

[3] Feng, L. and Zhou, T. (2012) Water Vapor Transport for Summer Precipitation over the Tibetan Plateau: Multidata Set Analysis. *Journal of Geophysical Research*: *Atmospheres*, **117**, D20114. https://doi.org/10.1029/2011jd017012

[4] Chen, Y., Wen, J., Liu, R., Zhou, J. and Liu, W. (2022) The Characteristics of Water Vapor Transport and Its Linkage with Summer Precipitation over the Source Region of the Three Rivers. *Journal of Hydrometeorology*, **23**, 441-455.

[5] Park, C., Son, S. and Kim, H. (2021) Distinct Features of Atmospheric Rivers in the Early versus Late East Asian Summer Monsoon and Their Impacts on Monsoon Rainfall. *Journal of Geophysical Research*: *Atmospheres*, **126**, e2020JD033537. https://doi.org/10.1029/2020jd033537

[6] Gimeno, L., Stohl, A., Trigo, R.M., Dominguez, F., Yoshimura, K., Yu, L., *et al.* (2012) Oceanic and Terrestrial Sources of Continental Precipitation. *Reviews of Geophysics*, **50**, RG4003. https://doi.org/10.1029/2012rg000389

[7] Brubaker, K.L., Entekhabi, D. and Eagleson, P.S. (1993) Estimation of Continental Precipitation Recycling. *Journal of Climate*, **6**, 1077-1089. https://doi.org/10.1175/1520-0442(1993)006<1077:eocpr>2.0.co;2

[8] Eltahir, E.A.B. and Bras, R.L. (1994) Precipitation Recycling in the Amazon Basin. *Quarterly Journal of the Royal Meteorological Society*, **120**, 861-880. https://doi.org/10.1002/qj.49712051806

[9] Martinez, J.A. and Dominguez, F. (2014) Sources of Atmospheric Moisture for the La Plata River Basin. *Journal of Climate*, **27**, 6737-6753. https://doi.org/10.1175/jcli-d-14-00022.1

[10] Stohl, A., Forster, C., Frank, A., Seibert, P. and Wotawa, G. (2005) Technical Note: The Lagrangian Particle Dispersion Model FLEXPART Version 6.2. *Atmospheric Chemistry and Physics*, **5**, 2461-2474. https://doi.org/10.5194/acp-5-2461-2005

[11] Sun, B. and Wang, H. (2014) Moisture Sources of Semiarid Grassland in China Using the Lagrangian Particle Model Flexpart. *Journal of Climate*, **27**, 2457-2474. https://doi.org/10.1175/jcli-d-13-00517.1

[12] Dominguez, F., Kumar, P., Liang, X. and Ting, M. (2006) Impact of Atmospheric Moisture Storage on Precipitation Recycling. *Journal of Climate*, **19**, 1513-1530. https://doi.org/10.1175/jcli3691.1

[13] Worden, J., Bowman, K., Noone, D., Beer, R., Clough, S., Eldering, A., *et al.* (2006) Tropospheric Emission Spectrometer Observations of the Tropospheric HDO/$H_2O$ Ratio: Estimation Approach and Characterization. *Journal of Geophysical Research: Atmospheres*, **111**, D16309. https://doi.org/10.1029/2005jd006606

[14] Stein, A.F., Draxler, R.R., Rolph, G.D., Stunder, B.J.B., Cohen, M.D. and Ngan, F. (2015) NOAA's HYSPLIT Atmospheric Transport and Dispersion Modeling System. *Bulletin of the American Meteorological Society*, **96**, 2059-2077. https://doi.org/10.1175/bams-d-14-00110.1

[15] Dorling, S.R., Davies, T.D. and Pierce, C.E. (1992) Cluster Analysis: A Technique for Estimating the Synoptic Meteorological Controls on Air and Precipitation Chemistry—Method and Applications. *Atmospheric Environment. Part A. General Topics*, **26**, 2575-2581. https://doi.org/10.1016/0960-1686(92)90110-7

[16] Xin, F., Peng, D., Liu, R. and Liu, S.C. (2022) Moisture Sources for the Weather Pattern Classified Extreme Precipitation in the First Rainy Season over South China. *International Journal of Climatology*, **42**, 6027-6041. https://doi.org/10.1002/joc.7576

[17] Lolis, C. and Türkeş, M. (2016) Atmospheric Circulation Characteristics Favouring Extreme Precipitation in Turkey. *Climate Research*, **71**, 139-153. https://doi.org/10.3354/cr01433

[18] Agarwal, A., Maheswaran, R., Sehgal, V., Khosa, R., Sivakumar, B. and Bernhofer, C. (2016) Hydrologic Regionalization Using Wavelet-Based Multiscale Entropy Method. *Journal of Hydrology*, **538**, 22-32. https://doi.org/10.1016/j.jhydrol.2016.03.023

[19] Draxier, P.R. and Hess, G.D. (1998) An Overview of the HYSPLIT_4 Modelling System for Trajectories, Dispersion and Deposition. *Australian Meteorological Magazine*, **47**, 295-308.

[20] Shi, Y., Jiang, Z., Liu, Z. and Li, L. (2020) A Lagrangian Analysis of Water Vapor Sources and Pathways for Precipitation in East China in Different Stages of the East Asian Summer Monsoon. *Journal of Climate*, **33**, 977-992. https://doi.org/10.1175/jcli-d-19-0089.1

[21] Bosilovich, M.G., Schubert, S.D. and Walker, G.K. (2005) Global Changes of the Water Cycle Intensity. *Journal of Climate*, **18**, 1591-1608. https://doi.org/10.1175/JCLI3357.1