https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


# Quality Control of Historical Temperature Data for Pure Rotational Raman Lidar Using Density-Based Clustering

Rongzheng Cao[1], Siying Chen[1,2], Wangshu Tan[1], Yixuan Xie[1], He Chen[1,2], Pan Guo[1], Rui Hu[1], Yinghong Yu[1], Jie Yu[1], Shusen Yao[3]

[1]School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China
[2]Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing 314019, China
[3]Medical Supplies Center of PLA General Hospital, Beijing 100853, China

*Correspondence to*: Siying Chen (csy@bit.edu.cn), Shusen Yao (yssvip@sina.com)

**Abstract.** This paper is the first to use two density-based clustering algorithms, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Ordering Points To Identify the Clustering Structure (OPTICS), to screen the historical detection data of pure rotational Raman (PRR) temperature measurement lidar. To address the issues of threshold radius in DBSCAN and output value processing in OPTICS, three automated processing methods suitable for PRR temperature lidar detection data characteristics are proposed. These methods are the k-distance Fast Change Region (k-FCR) Method based on the DBSCAN, the Reachability Distance (RD) Method based on the OPTICS, and the Predecessor Divergence (PD) Method based on the OPTICS. Using these three methods, quality control was conducted on the historical data detected by a PRR temperature lidar from March 2021 to May 2024, demonstrating the effectiveness of these methods in automated quality control of historical data and the complementary nature of their quality control effects. Under the reliable threshold set in this paper, compared with the traditional Signal-to-Noise Ratio (SNR) method, the RD method increased the True Positive Rate (TPR) by 23.7%, the PD method increased the True Negative Rate (TNR) by 6.0%, and the k-FCR method increased the TPR by 72.1% at the cost of some TNR loss. The influence of the SNR of data points and the number of continuous observation profiles on the quality control results is also explored, providing further references for the selection and application of different quality control methods. The methods provided in this paper will allow relevant researchers to filter PRR lidar data of atmospheric temperature according to their own needs, and these methods can also be applied to the automated processing of future atmospheric temperature data from detection networks.

## 1 Introduction

Temperature is a crucial parameter in characterizing the state of the Earth's atmosphere. Detailed and high-resolution temperature structure observations are urgently needed for studying atmospheric energy balance, dynamics, and chemistry (Gerding et al., 2008). The troposphere, being the region most closely related to human activities (Stull, 2012), requires precise temperature detection for studying the atmospheric transport of pollutants (Burkow and Kallenborn, 2000; Beyer et

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


al., 2003; Klonecki et al., 2003) and for short to medium term weather forecasting (Adam et al., 2016; Lawrence et al., 2019; Thundathil, 2023).

Lidar is currently one of the primary methods for tropospheric temperature detection, capable of providing continuous high temporal and spatial resolution atmospheric temperature information at the same location, which offers unique advantages over other detection methods (He et al., 2018a). Among these, Pure Rotational Raman (PRR) temperature measurement lidar has proven effective in detecting vertical temperature profiles from the troposphere to the stratosphere (Wandinger, 2005). PRR lidar obtains temperature information of atmospheric molecules by detecting weak rotational Raman scattering signals. Due to the weak nature of these signals, they are prone to data quality variations caused by hardware stability or the presence of aerosols and clouds(Wandinger, 2005). However, the quality assessment of PRR Lidar data has long relied on the Signal-to-Noise Ratio (SNR) (e.g., SNR > 10 dB is considered reliable data (Yan et al., 2019)), which is a coarse method and does not account for the characteristics of the temperature profiles themselves. Additionally, using other data sources for data quality verification is also common (Aspey et al., 2006), but it is not conducive to long-term independent detection by the PRR lidar system.

Relying on the temporal and spatial continuity of atmospheric parameters, there have been studies reporting the use of density-based clustering methods for quality control of wind lidar data in the frequency domain (Alcayaga, 2020). Atmospheric temperature, similar to wind fields, also exhibits temporal and spatial continuity, making density-based clustering methods potential for screening PRR lidar temperature detection data. Density-based clustering classifies data based solely on its features without the aid of external data sources, and it is a form of unsupervised learning. The most commonly used density-based clustering methods are Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996; Khan et al., 2014) and Ordering Points To Identify the Clustering Structure (OPTICS) (Ankerst et al., 1999). DBSCAN innovatively defines “density reachability” and “density connectivity” of data points by specifying two parameters: minimum number of points (minPts) and radius ε. In the feature space, points with more than minPts neighbors within a radius ε are considered dense and called “core points”. These neighboring points are considered directly density-reachable from this core point, and the algorithm computes the transitive closure of this reachability relationship (Schubert and Gertz, 2018). OPTICS optimizes the DBSCAN algorithm to adapt to variations in data density and only requires specifying minPts. However, its output values (primarily reachability distances) need further interpretation to obtain clustering results. Given that minPts has an empirical range of values (Ester et al., 1996; Ankerst et al., 1999), the selection of the threshold radius ε in DBSCAN and the processing of the output value sequence in OPTICS are the main obstacles to automated classification.

There have been many studies reported on automated solutions for these two issues (Ester et al., 1996; Ankerst et al., 1999; Sander et al., 2003; Brecheisen et al., 2004; Esmaelnejad et al., 2010; Sawant, 2014; Karami and Johansson, 2014; Schubert and Gertz, 2018). However, most of these studies target multiple classification objectives, with significant differences in data density between valid data and noise, making them unsuitable for PRR lidar data characteristics. PRR lidar temperature data gradually transitions from valid data to noise, and in most cases, there is no significant change in data density between the

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


two. Moreover, for atmospheric temperature detection data, it is only necessary to distinguish between good and poor data quality, making it a binary classification problem. L. Alcayaga made a beneficial attempt at using the DBSCAN for wind lidar (Alcayaga, 2020), but its effectiveness was limited for PRR lidar. Currently, there have been no reported attempts using the OPTICS algorithm for atmospheric lidar data screening. Therefore, new automated methods addressing the aforementioned two issues specifically for PRR lidar detection data are still needed.

Given the above context, this paper is going to utilize two density-based clustering algorithms, DBSCAN and OPTICS, to process PRR lidar temperature detection data for the first time. Addressing the threshold radius issue in DBSCAN and the output value processing issue in OPTICS, we propose automated processing methods tailored to the characteristics of PRR lidar temperature detection data. Additionally, we introduce the application of the predecessor, a by-product of the OPTICS algorithm, to OPTICS classification discrimination for the first time. By employing this approach, we aim to achieve better quality control results and realize the automation of quality control for PRR lidar temperature detection data. This paper, part of the Atmospheric detectIon lidaR Quality and UncErtainty aSsessmenT (AIR-QUEST) series, focuses on the quality control of PRR lidar data and represents the first application of machine learning methods for PRR lidar data quality control.

The density-based clustering algorithms and atmospheric temperature data used in this study are introduced in Sect. 2. Clustering process and automated segmentation point determination methods are introduced in Sect. 3. The final clustering results and further analysis of these results are presented in Sect. 4.

## 2. Clustering algorithms and temperature data

### 2.1 Density-based clustering algorithms

The two algorithms have been explained in detail in the existing literatures (Ester et al., 1996; Khan et al., 2014; Ankerst et al., 1999; Schubert and Gertz, 2018). This section briefly introduces the execution steps of the two density-based clustering algorithms.

#### 2.1.1 DBSCAN

DBSCAN is a density-based spatial clustering algorithm that defines three types of points in the data:

(1) Core points: Points that contain more than MinPts within a specified radius ε.

(2) Border points: Points that have fewer than MinPts within the ε radius but are within the ε neighborhood of a core point.

(3) Noise points: Points that are neither core points nor border points.

The DBSCAN algorithm starts with an arbitrary unvisited point and checks its ε neighborhood. If this point is a core point, a new cluster is formed around it, and the cluster is recursively expanded (i.e., the process is repeated for all unvisited neighboring points). If the point is not a core point, it is marked as noise (it may later be reclassified as a border point). Figure 1 visually demonstrates this process of density-based cluster expansion.

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


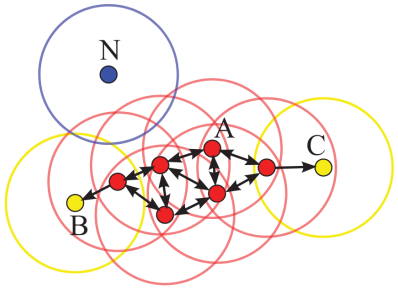


**Figure 1. Illustration of the DBSCAN cluster model (Schubert et al., 2017).**

### 2.1.2 OPTICS

The OPTICS algorithm is an extension of DBSCAN designed to overcome its limitations in handling datasets with varying density regions. Based on the concepts in DBSCAN, OPTICS defines two types of distances:

(1) Core distance: For a point P, if there are at least MinPts points in its ε neighborhood, its core distance is the distance to the MinPts-th closest point to P.

(2) Reachability distance: The reachability distance from point A to point B is the maximum of the actual distance from B to A and the core distance of A.

The OPTICS algorithm begins by selecting an arbitrary, unprocessed point as the starting point, with its initial reachability distance set to infinity. From this starting point, the algorithm calculates the reachability distance to all other points, updating the reachability distances for all unprocessed points. A point's reachability distance is only updated if the new distance is smaller than its current value. The point with the smallest reachability distance is then selected, its reachability distance is recorded, and it becomes the new starting point. This process is repeated until all points are processed. Ultimately, the algorithm produces a sequence of reachability distances.

## 2.2 Atmospheric temperature data

The atmospheric temperature data used in this study includes PRR lidar data, ERA5 data, and radiosonde data, which are introduced in this section.

### 2.2.1 PRR lidar data

The PRR lidar data used in this paper is sourced from historical data detected by the PRR-Mie lidar system from March 2021 to May 2024. This system is located at Beijing Institute of Technology in Beijing, China (116.31°E, 39.96°N) (Chen et al., 2011, 2016). It can simultaneously detect temperature and aerosols in the lower and middle troposphere at night. Detailed system and data parameters are shown in Table 1. The system uses a dual-grating spectrometer to simultaneously receive Stokes and anti-Stokes Raman signals, and the temperature is retrieved using the signal intensity ratio of the high- and low-

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


quantum-number channels. During testing, the PRR channel achieved a suppression ratio of elastic scattering signals by up to $10^7$ orders of magnitude (He et al., 2019).

| Parameters | | | Values |
|---|---|---|---|
| PRR Lidar system | Transmitter | Wavelength | 532.17 nm |
| | | Laser pulse energy | 120 mJ |
| | | Repetition frequency | 20 Hz |
| | Receiver | Diameter | 40 cm |
| | | Angle of view | 0.9 mrad |
| | | Optics reflectivity | 0.92 |
| | | Quantum efficiency | 0.12 |
| L0/L2 data | | Vertical resolution | 30 m |
| | | Integral time | 17 min |
| | | Total number of range gates | 1024/500 |

**Table 1. Major parameters for PRR lidar system.**

The historical data used in this study consists of 182 sets of daily detection data from March 2021 to May 2024. During this period, the system underwent two hardware adjustments: the addition of a vibrational Raman channel for water vapor detection from June to September 2022, and the upgrade of the PMT in the PRR channel in January 2024. The former reduced the signal energy of the PRR channel, while the latter increased the quantum efficiency of the PRR channel. Therefore, this period's data can be used to verify the quality control algorithm's effectiveness under different signal conditions. The Level 2 (L2) data were retrieved from Level 0 (L0) data follows the procedures outlined in references (Chen et al., 2016; He et al., 2018b), with daily calibration using radiosonde data around 20:00 Beijing time.

### 2.2.2 ERA5 data

Reanalysis data has high temporal and spatial resolution, allowing for good spatiotemporal matching with target data and is commonly used as a reference in atmospheric data analysis (Hersbach et al., 2020). This study uses hourly ERA5 data provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) as the reference data. ERA5 is the fifth generation of reanalysis data created by ECMWF. Reported studies (Virman et al., 2021; Sun et al., 2022) show that ERA5 exhibits a temperature bias (less than 3K) near the surface compared to radiosonde measurements, with the temperature difference rapidly decreasing with altitude.

### 2.2.3 Radiosonde data

Radiosonde data is considered in-situ observation data and is regarded as the closest measurement to the true atmospheric state (Luers and Eskridge, 1998). However, since radiosondes are only launched at 0:00 and 12:00 UTC daily, the temporal resolution is very low, providing only one radiosonde profile during the PRR lidar observation period. This study uses radiosonde data to compensate for the poor performance of ERA5 near-surface temperature data and to indirectly validate

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


the quality control results. Figure 2 shows that the ERA5 data used in this study has a high level of consistency with the radiosonde data, with an $R^2$ value of 0.9983 and a maximum deviation of about 3 K near the surface, which is consistent with the literatures (Virman et al., 2021; Sun et al., 2022). Considering that the threshold for reliable data filtering in this study is a bias of less than 5 K from the ERA5 data, it is reasonable to use ERA5 data to validate the PRR lidar temperature retrievals.

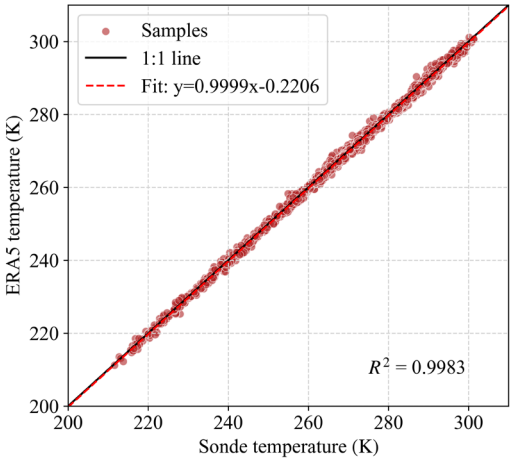


**Figure 2. Scatter plot of ERA5 data and sonde data used in this paper (12:00 UTC per observation day)**

## 3 Clustering process and automated segmentation point determination methods

### 3.1 Clustering process

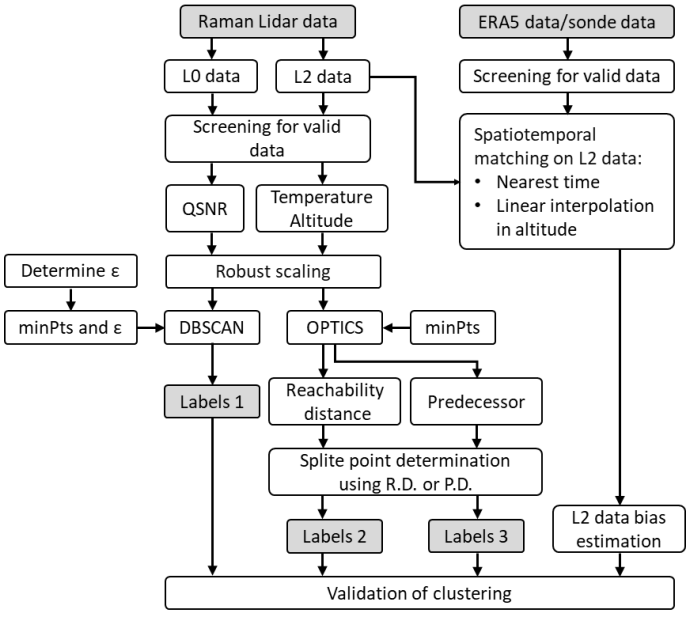


**Figure 3. Clustering process and validation of clustering result**

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


The methods of quality control using clustering and validation of clustering results used in this paper are shown in Fig. 3. These include:

(1) Filtering out invalid values from the L0 and L2 data.

Considering the height range of the PRR temperature lidar used in this study, data points with temperatures less than 180 K and greater than 320 K are excluded.

(2) Extracting data quality parameter QSNR from the L0 data and obtaining retrieval temperature and height from the L2 data.

QSNR is a new parameter defined in this paper and its construction method needs to be introduced here. Firstly, considering only the signal photon noise, the SNR of the dual channels in the L0 data is:

$$\mathrm{SNR}_{\mathrm{PRR1}} = \frac{N_{\mathrm{PRR1}}}{\sqrt{N_{\mathrm{PRR1}}}}, \tag{1}$$

$$\mathrm{SNR}_{\mathrm{PRR2}} = \frac{N_{\mathrm{PRR2}}}{\sqrt{N_{\mathrm{PRR2}}}}, \tag{2}$$

$$\mathrm{SNR}_{\mathrm{g}} = \sqrt{\mathrm{SNR}_{\mathrm{PRR1}}\mathrm{SNR}_{\mathrm{PRR2}}}, \tag{3}$$

In the formula, $N_{\mathrm{PRR1}}$ and $N_{\mathrm{PRR2}}$ represent the denoised signal photon counts for the high-quantum-number and low-quantum-number signal channels, respectively. The background noise is derived from the mean signal above 20 km. Since the PRR lidar has two channels, Eq. (3) uses the geometric mean of the SNRs of the two channels, denoted as $\mathrm{SNR}_{\mathrm{g}}$, to represent the average SNR of both channels. The geometric mean is sensitive to low SNR values, which allows for a more accurate representation of the actual retrieval situation.

During the retrieval process of the L2 data, variable range gate smoothing is performed, where more range gate integration is applied to farther range gates. This results in the SNR recorded in the L0 data being lower than the actual SNR of the temperature retrieval results in the L2 data. However, since most L2 data do not record the smoothing process, the SNR calculated from the L0 data remains the best choice for representing the quality of the original data.

However, the dual-channel signals of the PRR lidar are not directly used for temperature retrieval. Instead, the ratio $Q$ of high- to low-quantum-number signals channels is constructed for temperature retrieval:

$$Q = \frac{N_{\mathrm{PRR1}}}{N_{\mathrm{PRR2}}}, \tag{4}$$

Therefore, directly using SNR to describe data quality is not appropriate. This paper derives a new data quality parameter QSNR using the uncertainty propagation formula (Shimaoka K, Kinoshita M, Fujii K, et al., 2008):

$$\mathrm{QSNR} = Q\sqrt{(\mathrm{SNR}_{\mathrm{PRR1}})^{-2} + (\mathrm{SNR}_{\mathrm{PRR2}})^{-2}}, \tag{5}$$

The derivation process is provided in Appendix A. The logarithm of QSNR is used as one of the data features input into the clustering algorithm. Taking the logarithm is done to prevent its rapid decrease with altitude from affecting the clustering algorithm. Since it is difficult to establish a universally accepted threshold for QSNR, this paper still uses $\mathrm{SNR}_{\mathrm{g}}$ when an SNR threshold is needed (with $\mathrm{SNR}_{\mathrm{g}} > 10$ dB indicating good quality of the original data).

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


(3) Constructing the data feature matrix and scaling the data features.

Temperature and height obtained from L2 data, along with QSNR, are unfolded into one dimension to construct the data feature matrix. System and atmospheric conditions may change across different detection events, leading to overall changes in the data features obtained. Therefore, this study uses data generated during a single continuous detection session (typically between 19:00 and 24:00 local time) for each clustering analysis.

Data feature scaling is performed using robust scaling, which involves dividing by the median and then scaling according to the interquartile range. This approach ensures that different features contribute uniformly to the results, reduces the sensitivity of the clustering algorithm to outliers, and enhances the robustness of the algorithm.

(4) Determining the parameters required for the DBSCAN and OPTICS algorithms and inputting the data feature matrix into the algorithms.

Both the DBSCAN and OPTICS algorithms require the determination of minPts within a neighborhood. In the original DBSCAN paper (Ester et al., 1996), it was suggested that the k-distance graph does not change significantly when minPts exceeds 4. However, for PRR lidar detection data, too small a minPts value can lead to excessive tolerance for noise-dominated regions, resulting in a flattening of the k-distance elbow (Fig. 4a), making it difficult to identify data quality segmentation points through the elbow. Therefore, this paper uses a minPts value of 10 for the DBSCAN algorithm. In the literature introducing the OPTICS algorithm (Ankerst et al., 1999), a minPts value of 10–20 is considered suitable for most situations. However, for PRR lidar detection data, too small a minPts value retains too much detail in the reachability distance curve (Fig. 4b), complicating the subsequent identification of peaks and valleys. Therefore, this paper uses a minPts value of 20 for the OPTICS algorithm. The automated determination method for the threshold radius required by DBSCAN are introduced in Sect. 3.2.1.

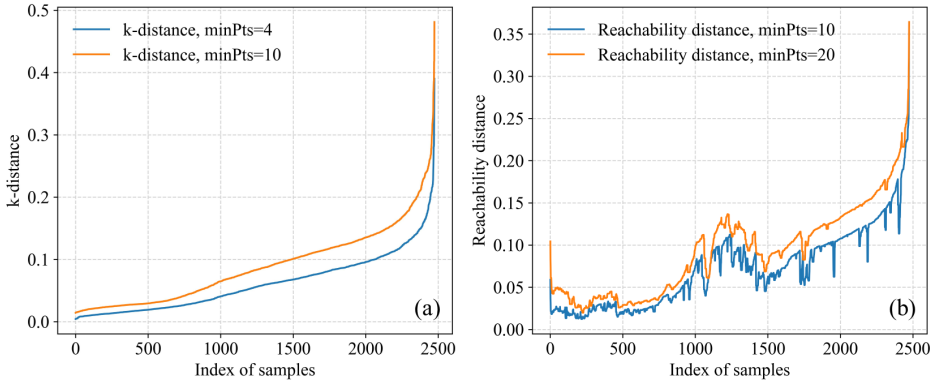


**Figure 4. (a) DBSCAN's k-distance and (b) OPTICS's reachability distance for different minPts values, with data from March 25, 2021.**

(5) DBSCAN directly outputs quality control labels, while the OPTICS output results need processing to obtain quality control labels.

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


The output of the OPTICS algorithm includes both the reachability distances and the predecessor of the sample points. For any given point, its reachability distance is defined as the distance from its predecessor to that point. This paper designs different automated processing methods for reachability distances and predecessors to obtain data quality segmentation points, which are introduced in Sect. 3.2.2 and 3.2.3.

(6) Use spatiotemporally matched ERA5 data to validate the quality control labels.
The radiosonde data displayed in Fig. 2 is used to cross-validate the quality of the ERA5 data.

### 3.2 Automated determination methods of data quality segmentation points

The automated determination of data quality segmentation points is the most critical component of the automated quality control method for PRR lidar temperature data based on density clustering algorithms. This section proposes three new methods for determining data quality segmentation points tailored to the characteristics of PRR lidar data.

#### 3.2.1 Determination of threshold radius $\varepsilon$ in DBSCAN

Calculate the distance to the $k$th nearest point for all data points ($k$ = 10) and sort to obtain the k-distance curve, as shown in Fig. 5a. This method partially draws on the approach from the literature (Alcayaga, 2020), but it has been improved to suit the characteristics of PRR lidar data. It is referred to as the k-distance Fast Change Region method (k-FCR). The steps are as follows:

(1) Smooth the k-distance curve using a sliding window. This step is to facilitate the identification of sudden change position using the first derivative in step (4), avoiding the impact of oscillations in the k-distance curve. In this paper, the smoothing window size $N_{\mathrm{DB}}$ is set to 20 data points.

(2) Identify the knee point on the k-distance curve, which corresponds to the radius $\varepsilon_{\mathrm{knee}}$. This paper uses the KneeLocator function in Python, which smooths the curve first and then identifies the point where the rate of change of the first derivative is maximal.

(3) Calculate the proportion $f_{\mathrm{SNR}}$ of points with $\mathrm{SNR_g}$ greater than 10, with the corresponding radius $\varepsilon_{\mathrm{SNR}}$:

$$\varepsilon_{\mathrm{SNR}} = \min(\text{k-distance}) + f_{\mathrm{SNR}}\varepsilon_{\mathrm{knee}}, \tag{6}$$

(4) The region between $\varepsilon_{\mathrm{knee}}$ and $\varepsilon_{\mathrm{SNR}}$ is the k-distance fast change region. The median of the first derivative of the k-distance within this region is $m_{\mathrm{dk}}$. When the first derivative of the k-distance at a certain point exceeds $m_{\mathrm{dk}}$, and the first derivatives of the k-distances at the next $M$ consecutive points also exceed $m_{\mathrm{dk}}$, that point is considered a sudden change point within the fast change region. And that point is the threshold radius $\varepsilon_{\mathrm{DB}}$. In this study, $M$ is set to 10% of the data points in the fast change region. The value of $M$ also influences the classification results. In most cases, the fast change region does not have a distinct data elbow point (Fig. 5a is a special case), so the starting point of the faster changing area can only be approximately determined using the first derivative.

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


Traditional methods often use the logarithm of the k-distance to find sudden change points. However, the k-FCR method uses the k-distance directly because PRR lidar temperature data tends to exhibit instability near the atmospheric boundary layer. These fluctuations are reasonable but become prominent when using the logarithm of the k-distance, which can affect the automated determination of $\varepsilon_{\mathrm{DB}}$. Experiments show that when data quality significantly deteriorates, it still can be clearly reflected in the k-distance (Fig. 5).

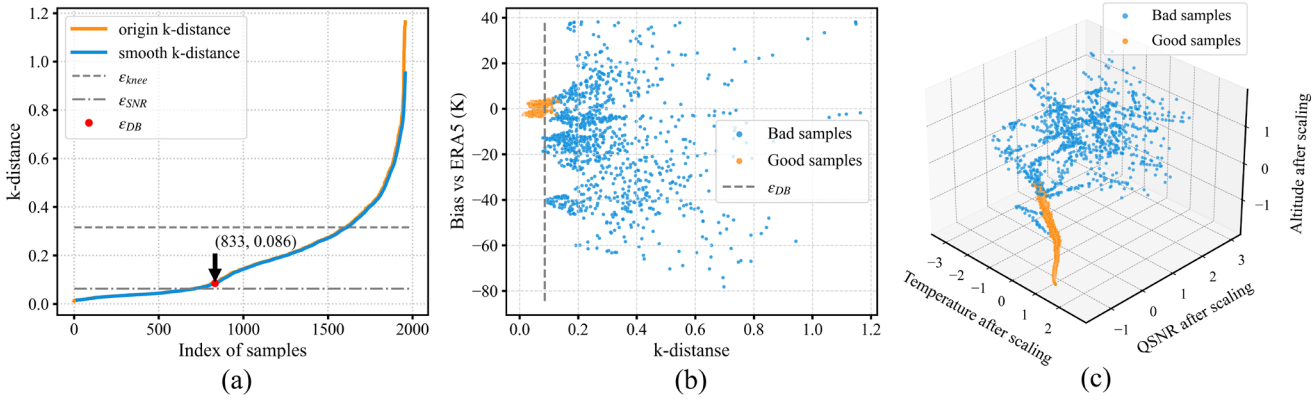


**Figure 5. Classification results using the k-FCR method based on the DBSCAN algorithm. (a) k-distance curve and threshold radius selection point. (b) k-distance of samples and temperature deviation relative to ERA5, with orange points representing reliable data and blue points representing unreliable data according to the classification results. (c) Three-dimensional clustering results, with the x, y, and z axes representing the three data features after scaling used for clustering: temperature, QSNR, and altitude. The colour of the data points has the same meaning as in (b). Data is from August 30, 2021.**

Figure 5 shows the results of applying the above steps to the data from August 30, 2021, a representative day with a large amount of noise. Figure 5a shows that the k-FCR method can accurately identify the fast-changing region, which excludes both the stable k-distance region in the front (indicating clearly better data quality) and the abrupt k-distance change region in the back (indicating clearly worse data quality). The selected $\varepsilon_{\mathrm{DB}}$ closely approximates the starting point of the sudden change in the fast-changing region. Then the $\varepsilon_{\mathrm{DB}}$ is put into the DBSCAN algorithm yields the filtering results shown in Fig. 5b and 5c. The results indicate that the algorithm effectively filters the data, with the selected points clustering together in space, reflecting the true data's self-similarity (Mandelbrot and Mandelbrot, 1982) and all located in areas with minimal deviation from the ERA5 data. Similar results were obtained for other dates, proving the method's applicability for PRR lidar temperature data quality control.

### 3.2.2 Determination of segmentation point using reachability distance in OPTICS

The reason for validating the OPTICS algorithm alongside the DBSCAN algorithm is that DBSCAN does not adequately capture the internal structural variations under the original sequence of data features. For PRR lidar data, the transition from reliable data to noise is gradual in most cases, which causes the k-distance to increase slowly. This can result in a sudden change in k-distance only when the data quality has already significantly deteriorated. Since the k-distance curve is based on

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


sorted data, it is challenging to reflect changes in data features from the original sequence on the k-distance curve. Therefore, this paper aims to find a method that can more accurately reflect changes in the original data sequence.

The reachability distance of each data point is the primary output of the OPTICS algorithm and is commonly used as a basis for classification. Peaks in the reachability distance indicate changes in data features, while valleys indicate stability. Traditional classification methods seek to identify peaks, with the flat valleys between peaks serving as the classification results. However, due to the characteristics of PRR lidar temperature data, the reachability distance gradually increases, with multiple peaks and valleys scattered throughout. As results, flat valleys are uncommon. However, the goal of this paper is to identify the dividing point between good and bad data rather than to classify the data into multiple categories. Therefore, a new Reachability Distance (RD) method is proposed for the OPTICS algorithm classification:

(1) Smooth the reachability distance curve using Gaussian smoothing, then identify all peaks and valleys. Gaussian smoothing removes minor fluctuations in the reachability distance while preserving the overall trend. In this study, the standard deviation $\sigma_{\mathrm{RD}}$ for Gaussian smoothing is set to 20 data points, same as $N_{\mathrm{DB}}$.

(2) Identify significant peaks based on the characteristics of PRR lidar data. A significant peak must meet two conditions: Firstly, the peak's height must be greater than the previous peak (Eq. (7)) but less than the mean smoothed reachability distance (smooth-RD) plus three standard deviations (SD) of smooth-RD (Eq. (8)). This ensures that selected significant peaks are in the increasing trend while excluding extreme values at the tail end.

$$\text{smooth-RD}(p_i) > \text{smooth-RD}(p_{i-1}), \qquad p_i \text{ is index of peaks} \tag{7}$$

$$\text{smooth-RD}(p_i) < \text{mean}(\text{smooth-RD}(\text{all})) + 3\text{Std}(\text{smooth-RD}(\text{all})), \tag{8}$$

Secondly, the peak's height must be greater than the mean of all preceding smooth-RD plus one SD of the interval (Eq. (9)), and also greater than the mean smooth-RD between this peak and the previous significant peak plus one SD of this interval (Eq. (10)). This ensures that selected significant peaks represent abrupt changes in all preceding smooth-RD while avoiding selecting too many significant peaks in the same region.

$$\text{smooth-RD}(p_i) > \text{mean}(\text{smooth-RD}[0, p_i]) + \text{Std}(\text{smooth-RD}[0, p_i]), \tag{9}$$

$$\text{smooth-RD}(p_i) > \text{mean}(\text{smooth-RD}[p_s(\text{previous}), p_i]) + \text{Std}(\text{smooth-RD}[p_s(\text{previous}), p_i]),$$
$$p_s \text{ is the index of significant peaks} \tag{10}$$

(3) Calculate the median smooth-RD of all significant peaks and find its last intersection with the smooth-RD curve, which serves as the data quality segmentation point *key* (Fig. 6a).

The RD method is a fuzzy decision approach that uses the median of significant peaks in the reachability distance to determine the data quality segmentation point. This point may not precisely reflect the transition from good to bad data quality. However, due to the gradual transition characteristics of PRR lidar temperature data, there are rarely abrupt declines in data quality, and the segmentation point is reasonable within a certain range. Figure 6 shows the results of applying the RD method to data from May 17, 2021. From Fig. 6a and 6b, it can be seen that significant peaks in the stable regions of reachability distance generally reflect reasonable fluctuations within the reliable data. Significant peaks in the regions where reachability distance increases rapidly are more likely to indicate a step change in data quality deterioration. The median of

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


the significant peaks typically lies between the stable and rapidly increasing states of the reachability distance. From this perspective, the segmentation points provided by the RD method, though not precisely accurate, are reasonable. Figure 6b and 6c show that the reliable data identified by the RD method are mostly located in high-density, low-deviation regions, while the unreliable data are found in sparse, high-deviation regions. This also aligns with the self-similarity of true data. The results for other dates are similar, demonstrating the applicability of the RD method for quality control of PRR lidar temperature data.

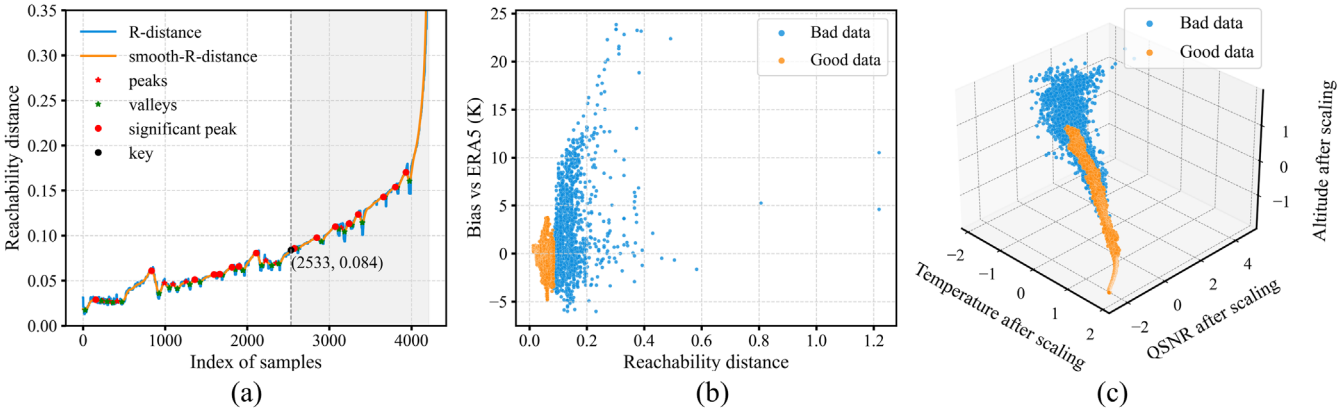


**Figure 6. Classification results using the RD method based on the OPTICS algorithm. (a) Reachability distance curve, significant peaks, and the data quality segmentation point key. (b) and (c) are similar to Fig. 5. Data is from May 17, 2021.**

### 3.2.3 Determination of segmentation point using predecessor in OPTICS

The two aforementioned methods for the automated quality control of PRR lidar temperature data share common issues: they both require manual setting of smoothing intensity, such as $N_{DB}$ in the k-FCR method and $\sigma_{RD}$ in the RD method, and the classification results are sensitive to the smoothing intensity. Therefore, this paper aims to find a method that requires less human interaction. Additionally, the k-FCR and RD methods may fail in certain special cases (see Sect. 4), because both methods rely on the shape of the k-distance or reachability distance curves to find data quality segmentation points. However, unusual shapes can occasionally appear in actual data, making it difficult for these methods to adapt to all possible curve shapes. Consequently, there is an urgent need for a robust method that performs well across different datasets.

Predecessors are by-products of the OPTICS algorithm when outputting reachability distances. Since the OPTICS algorithm updates the reachability distance sequence by recording the smallest reachability distance for each point, it is possible for multiple data points to have the same predecessor point. Therefore, OPTICS traces back and records the predecessors. Although predecessors are rarely used for classification, they show potential for use in the classification of PRR lidar temperature data.

When constructing the data feature matrix, data points at the same range gate at different times are adjacent, with data points having smaller index numbers generally showing lower altitudes and better data quality. The OPTICS algorithm always looks for the nearest points in feature space to update the reachability distance sequence. Due to the self-similarity of true

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


data, data points with good quality have their predecessors with nearby index numbers, while data points with poor quality show unstable divergence in their predecessors. As shown in Fig. 7, the predecessors converge and steadily increase before index 1000, while the points after index 1000 tend to diverge. Utilizing this characteristic, this paper proposes the Predecessor Divergence (PD) method:

(1) Apply a sliding three-standard-deviation filter to the predecessors (Fig. 7a). The outliers identified are replaced with the mean of the predecessor within the sliding window. The sliding window $N_{PD}$ is set to 5% of the total number of data points for a single continuous detection campaign. This step prevents extreme values in the predecessors from affecting the results, as the goal is to identify the point where the predecessors start to diverge significantly, rather than isolated jumps.

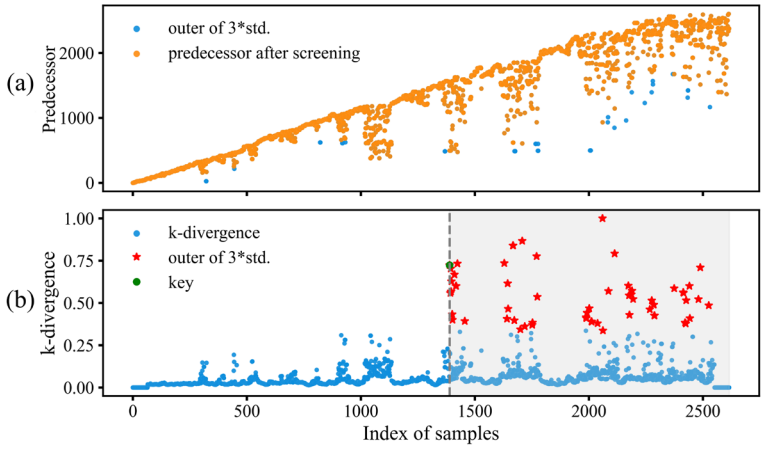


**Figure 7. (a) Predecessor, with blue points representing filtered out values and orange points representing retained values. (b) k-divergence constructed from predecessors, with red points indicating those exceeding the threshold and green points indicating the data quality segmentation points. Data is from May 29, 2021.**

(2) Perform robust scaling on both the predecessor values and their indices to balance their influence on the k-distance. Then, calculate the k-distance from the center point of the sliding window to all other points within the window, sort these distances, and take the mean of the top 50%. This produces a sequence of k-distance averages. Normalize the sequence to have a minimum value of 0 and a maximum value of 1, constructing a parameter called *k-divergence* to represent local predecessor divergence. The sliding window size is the same as the $N_{PD}$ in step (1). Through our practice, we find that the PD method is not sensitive to the sliding window size but should use a smaller window to retain rich details in the predecessors.

(3) Apply a global three-standard-deviation filter to the k-divergence. The first point that exceeds the threshold is considered the data quality segmentation point *key* (Fig. 7b). To avoid interference from outliers, check if there are any other points exceeding the threshold within $N_{PD}$ before and after this point. If this point is isolated, it is ignored, and the search for the data quality segmentation point continues.

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


From Fig. 7b, it can be seen that the k-divergence accurately represents the divergence of predecessor, with the divergence of predecessor and the increase in k-divergence occurring simultaneously. The PD method tolerates a certain degree of divergence in predecessor, and only detects regions where the divergence exceeds the overall average divergence level, indicating a rapid change in data quality. Therefore, the *key* point is suitable as data quality segmentation points. As shown in Fig. 8a and 8b, the PD method is stringent in selecting data quality segmentation points, classifying only data points with relatively high spatial density as reliable. The data identified as reliable by the PD method have a smaller overall deviation from ERA5 compared to the other two methods. Additionally, the PD method is less sensitive to the size of the smoothing window, reducing the need for manual adjustments. More importantly, thanks to the abrupt nature of predecessor divergence, the PD method overcomes the challenge of finding "sudden change points" in a gradual trend, thereby providing more precise data quality segmentation points.

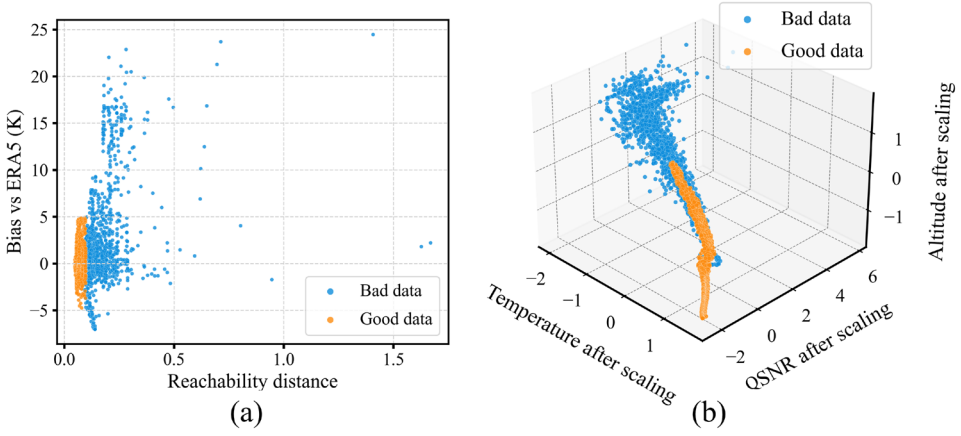


**Figure 8. Classification results using the PD method based on the OPTICS algorithm. (a) and (b) are similar to Fig. 6. Data is from May 29, 2021.**

## 4. Results and analysis

This section presents and analyses the quality control results of the four methods (including the traditional SNR method) applied to 182 sets of daily detection data from March 2021 to May 2024, discussing the differences of results and the underlying reasons for these differences.

| Sample number | 999 | 1992 | 2988 | 3996 | 4995 | 5661 | 6930 |
|---|---|---|---|---|---|---|---|
| **DBSCAN** | 3.98 | 13.01 | 13.46 | 18.94 | 24.43 | 27.91 | 44.85 |
| **k-FCR** | 2.99 | 12.95 | 13.95 | 15.95 | 17.94 | 15.95 | 43.85 |
| **OPTICS** | 215.80 | 456.17 | 769.39 | 1004.45 | 1323.94 | 1516.64 | 1976.05 |
| **RD** | 0.99 | 0.99 | 1.99 | 1.04 | 1.99 | 2.99 | 3.99 |
| **PD** | 139.01 | 525.45 | 1185.84 | 2281.08 | 3211.20 | 4540.00 | 6529.95 |

**Table 2. Runtime (ms) of Different Algorithms or Quality Control Methods**

The processing time of data quality control methods is a significant concern, especially when dealing with large volumes of historical data. Table 2 provides representative data volumes and processing times for 182 sets of daily detection data as a

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


reference. It can be seen that the runtime of the k-FCR and RD methods is mainly influenced by the size of the fast change region and the number of significant peaks, and does not increase linearly with the number of samples. The PD method, due to the need to calculate k-divergence for all points, has the longest processing time.

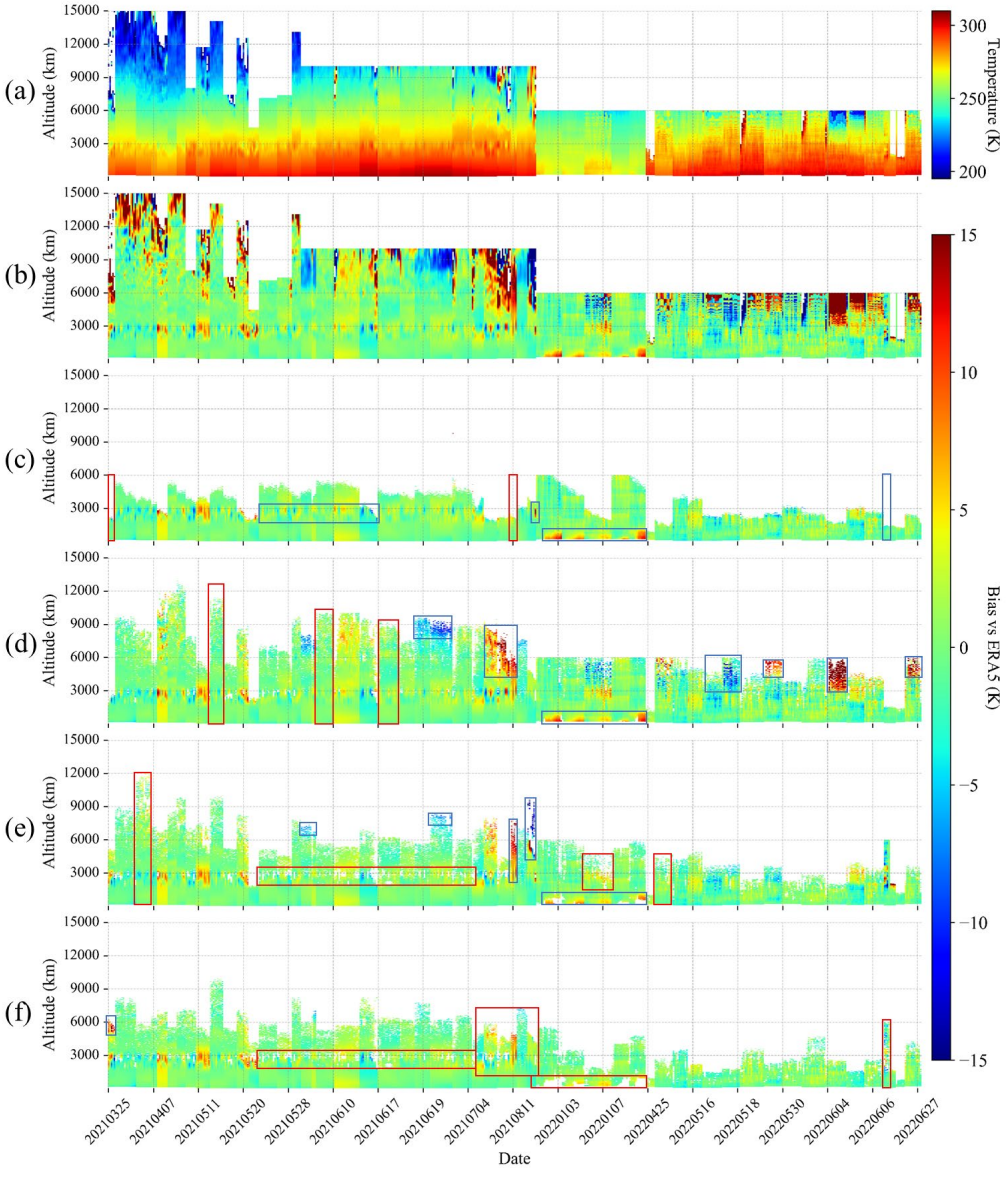


**Figure 9. Classification results of different methods, showing only the reliable data after quality control. The red box highlights a typical area where the method performs relatively well, while the blue box indicates a typical area where the method performs poorly. (a) Original temperature data, (b) Deviation of PRR temperature data from ERA5, (c) SNR method results, (d) k-FCR method results, (e) RD method results, (f) PD method results. Data is from March 2021 to June 2022.**

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


### 4.1 Comparison of quality control results of different methods

Considering the volume of all detection data, displaying it in a single spatiotemporal graph would make it difficult to discern details. Therefore, this section analyses the processing results of the first phase of historical data as a typical case. The results for the other two phases of historical data are summarized and presented in tables.

Figure 9 shows the results of different quality control methods applied to the historical detection data from the first phase of the system (March 2021 to June 2022, a total of 53 detection days). The red boxes indicate typical areas where the method performs relatively well, and the blue boxes indicate areas where the method performs relatively poorly. Since the characteristics of the processing results of the same algorithm are similar, the results of the other phases are not repeatedly displayed in this paper, but can be found in the supplementary materials.

As seen in Fig. 9c, the SNR method relies entirely on the original signal (L0 data) characteristics. Because there is a smoothing process during the retrieval, the filtering criterion of SNR less than 10 is quite strict for L2 data. The SNR method results in neat data edges, with most data edges still belonging to reliable data. However, high SNR does not always mean reliable retrieval temperature data, and the SNR method cannot filter out low-quality data that meets the SNR threshold. Additionally, the SNR method cannot utilize the L2 data characteristics to retain more amount valid data.

From Fig. 9d, it is evident that the k-FCR method employs a lenient quality control strategy. Especially in the red box areas, compared to other methods, the k-FCR method has a significant advantage in retaining more amount reliable data. Additionally, on detection days with overall good data quality, it shows a good capability in filtering out anomalous data. However, for detection days with generally poor data quality, where there is a gradual transition from good to bad data, the k-FCR method is insensitive to such overall changes. As a result, the k-FCR method may fail in such scenarios, retaining a considerable amount of anomalous data.

Figure 9e shows the quality control effect of the RD method. Although the RD method employs a fuzzy decision-making process, its actual effect is entirely acceptable. Compared to the PD method (Fig. 9f), it can retain more amount valid data and shows significant improvement in areas where the DBSCAN method fails (blue box in Fig. 9d). However, due to its compromise in quality control strategy, it often has issues with incomplete removal of anomalous data. In the blue box, it can be seen that the RD method has detected and removed most of the unreliable points in the anomalous region, but removing unreliable points at the boundary of the anomalous region becomes difficult, especially during the high aerosol concentration summer season. Overall, the RD method has good applicability in most cases.

Figure 9f shows the quality control effect of the PD method. The PD method employs the strictest quality control strategy and has the best robustness. Compared to other methods, the PD method achieves the most thorough removal of unreliable data on most detection days, even during the complex data conditions in summer. Additionally, compared to the SNR method, the PD method often retains more amount reliable data while also removing anomalous points within the reliable data. Of course, the PD method also has suboptimal performance on certain detection days (e.g., March 25, 2021, and August 11, 2021).

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


Overall, among the three methods proposed in this paper, the PD method achieves the most stringent removal of unreliable data, the k-FCR method achieves the maximum retention of reliable data, and the RD method falls between the two. The traditional SNR method performs similarly to the PD method but still lags behind overall.

To more intuitively observe the quality control effects of different methods on all data, this paper sets the threshold for reliable data to deviations from ERA5 of less than or equal to 5 K and less than or equal to 10 K, respectively. The true positive rate (TPR) and true negative rate (TNR) for all data points using different methods are then calculated, as shown in Fig. 10. Unlike other fields, in the quality control of PRR lidar temperature data, TPR and TNR are our primary concerns.

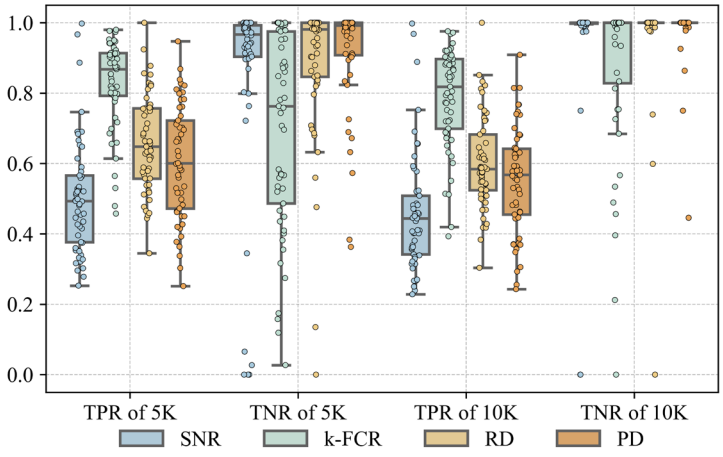


**Figure 10. True positive rate (TPR) and true negative rate (TNR) of quality control results for different methods, with thresholds for reliable data set to deviations from ERA5 of less than or equal to 5 K and less than or equal to 10 K. Data is from March 2021 to June 2022.**

| | Phase1 (202103 – 202206) | | | | Phase2 (202209 – 202312) | | | | Phase3 (202401 – 202405) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **TPR @5 K** | **TNR @5 K** | **TPR @10 K** | **TNR @10 K** | **TPR @5 K** | **TNR @5 K** | **TPR @10 K** | **TNR @10 K** | **TPR @5 K** | **TNR @5 K** | **TPR @10 K** | **TNR @10 K** |
| **SNR** | 0.493 | 0.967 | 0.444 | 1 (Q1=0.998) | 0.513 | 0.894 | 0.442 | 1 (Q1=0.971) | 0.566 | 0.905 | 0.548 | 1 (Q1=0.981) |
| **k-FCR** | 0.868 | 0.763 | 0.818 | 0.936 | 0.901 | 0.528 | 0.856 | 0.874 | 0.880 | 0.702 | 0.832 | 1 (Q1=0.878) |
| **RD** | 0.648 | 0.981 | 0.584 | 1 (Q1=1) | 0.628 | 0.861 | 0.557 | 1 (Q1=0.948) | 0.634 | 0.915 | 0.608 | 1 (Q1=1) |
| **PD** | 0.601 | 0.992 | 0.568 | 1 (Q1=1) | 0.496 | 0.939 | 0.429 | 1 (Q1=1) | 0.483 | 0.942 | 0.452 | 1 (Q1=1) |

**Table 3. Quality control results (median) for different methods across various phases of historical detection data**

From the Fig. 10, it can be seen that the statistical results are generally consistent with expectations. The k-FCR method has the highest TPR among the three methods (median 0.868 @ 5 K), indicating that it retains the most amount reliable data. However, its TNR is the lowest among the three methods (median 0.763 @ 5 K), meaning that its quality control results still include a considerable amount of unreliable data. In contrast, the PD method, with its strict filtering strategy, excels in TNR (median 0.992 @ 5 K) but loses a lot of correct data, resulting in a lower TPR (median 0.599 @ 5 K). Compared to the PD method, the SNR method performs a little worse in both TPR (median 0.493 @ 5 K) and TNR (median 0.967 @ 5 K) and

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


even slightly underperforms the RD method in TNR (median 0.981 @ 5 K). This may because the characteristics of the historical detection data in this phase are well-suited to density clustering methods. As a compromise quality control method, the RD method (median 0.648 @ 5 K) achieves a higher TPR than the PD method while obtaining a TNR only slightly worse than the PD method. Particularly at a threshold of 10 K, the TNR of the RD method (lower quartile Q1 = 1 @ 10 K) is almost identical to that of the PD method (Q1 = 1 @ 10 K) and slightly higher than that of the SNR method (Q1 = 0.998 @ 10 K). Overall, TPR and TNR are conflicting metrics, and no single quality control method can optimize both parameters simultaneously. The choice of method depends on the specific data requirements and the necessary trade-offs.

This paper provides statistics on the quality control results for PRR temperature lidar data across three phases (Sect. 2.2.1), as shown in Table 3. From the table, it can be seen that the overall quality control results for the other two phases are similar to Phase 1. However, the TPR for the PD method is slightly lower than the traditional SNR method in both Phase 2 and Phase 3. The TNR for the RD method is slightly lower than the SNR method in Phase 2 but slightly higher in Phase 3. Overall, the best phase in terms of historical detection data quality is Phase 1, while the worst is Phase 2. Furthermore, we can observe that the RD and PD methods, which are based on the OPTICS algorithm, perform better in phases with higher data quality, such as Phase 1 and Phase 3. This is because high-quality continuous detection data facilitates the OPTICS algorithm in constructing stable structure of data features, leading to more accurate data quality identification. Additionally, across all three phases, the PD method consistently achieves the most thorough removal of unreliable data, significantly outperforming other methods.

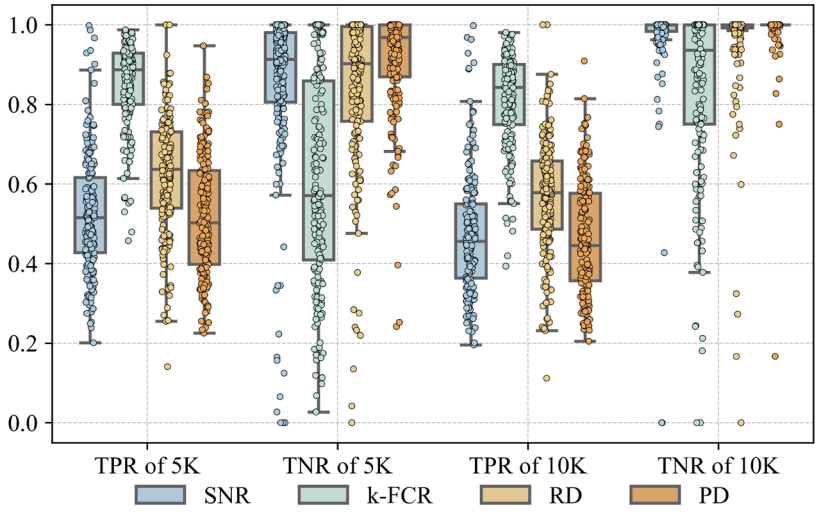


**Figure 11. Similar to Fig. 10, the data is from March 2021 to May 2024.**

Similarly, box plots of the quality control results for all historical detection data across the three phases are shown in Fig. 11. It can be observed that the PD method achieves a higher TNR compared to the SNR method (higher by 6% @ 5 K) while maintaining a similar TPR, even though the TNR of the SNR method is already quite high. On the other hand, the RD

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


method achieves a significantly higher TPR than the SNR method (higher by 23.7% @ 5 K) while maintaining a TNR comparable to the SNR method. This indicates that the PD and RD methods proposed in this paper offer significant improvements over the traditional SNR method in quality control of PRR lidar historical data. As for the k-FCR method, its quality control strategy is different from the other methods. It aims to retain more amount reliable data after a certain degree of unreliable data removal. The box plots show that the k-FCR method meets its intended goal, achieving a TPR that is 72% higher than the SNR method @ 5 K.

### 4.2 Influence of profile number in single day and SNR on different methods

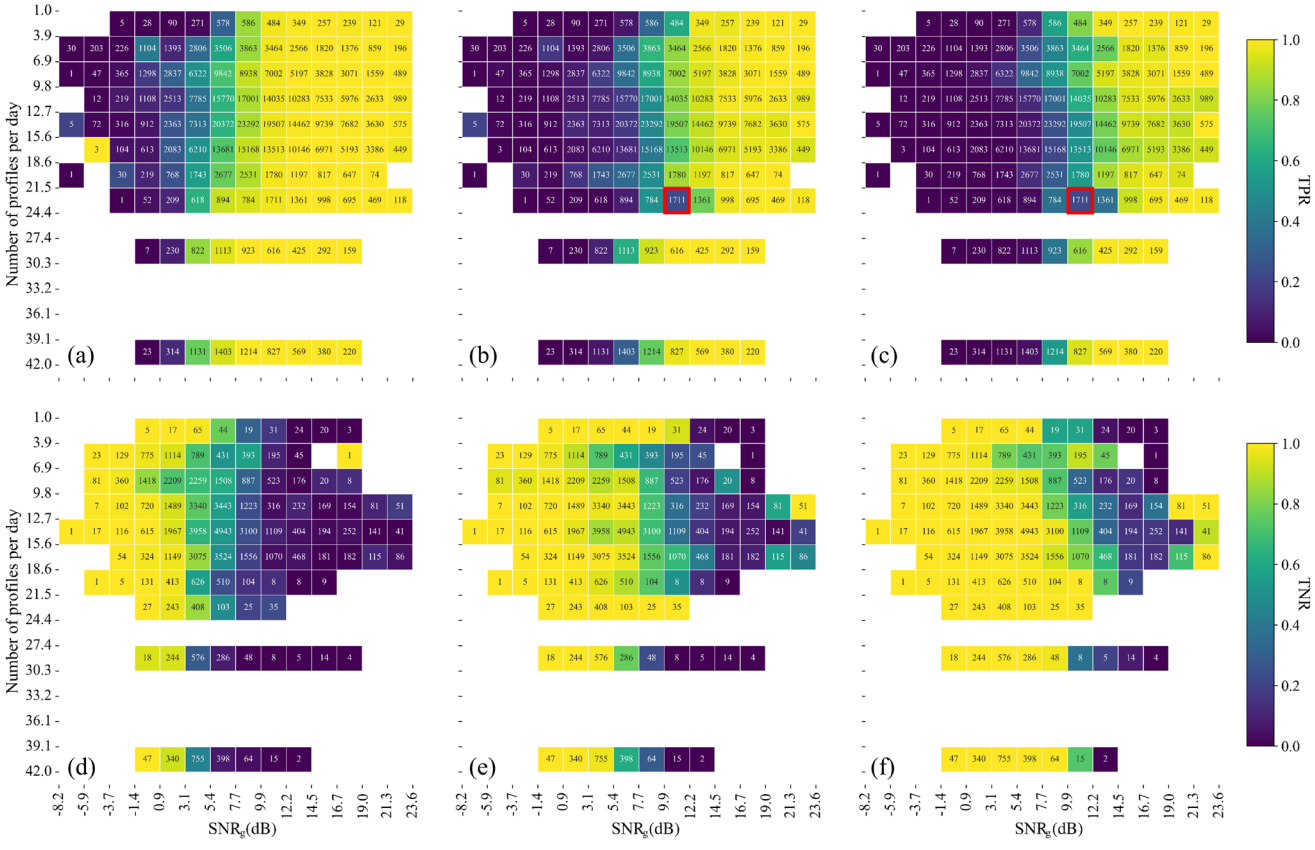


**Figure 12. The effect of the number of daily profiles and SNR$_g$ (dB) on TPR and TNR. The numbers within the cells represent the number of data points falling within that cell. For TPR cells, the numbers represent the number of reliable data points (deviation from ERA5 less than or equal to 5 K). For TNR cells, the numbers represent the number of unreliable data points (deviation from ERA5 greater than 5 K). The number of data points includes those on the top and left edges of the cell, with the last row including the bottom edge and the last column including the right edge. (a) TPR for the k-FCR method, (b) TPR for the RD method, (c) TPR for the PD method, (d) TNR for the k-FCR method, (e) TNR for the RD method, (f) TNR for the PD method.**

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


Based on the quality control results from Sect. 4.1, it is evident that for the same method, the number of profiles on a single detection day and the SNR of different data points can both impact the quality control results. This section analyses these effects to provide references for choosing a quality control method, using the $SNR_g$ from Eq. (3).

Figure 12 shows the average TPR and TNR of data points for different numbers of daily profiles and SNRs. For the k-FCR method, as shown in Fig. 12a and 12d, the boundary between high and low TPR shifts significantly towards lower SNR regions as the number of daily profiles increases. A similar pattern is observed for TNR (considering only the dark regions), except for an anomalous decrease in TNR when the number of profiles is between 6.9 and 9.8. This boundary shift means that as the number of daily profiles increases, the k-FCR method tends to be more lenient in data quality filtering, resulting in a higher retention rate of reliable data (TPR) and a lower rejection rate of unreliable data (TNR). This may be because a higher number of daily profiles causes the unreliable data clusters to form a considerable data density, and the k-FCR method tends to remove the sparsest data points. As a result, the boundary line gradually shifts towards the lower SNR region (sparser data density in feature space). In terms of SNR, compared to the other two methods, the boundary between high and low TPR and TNR for the k-FCR method shifts significantly to the left, indicating a higher tolerance for SNR and better data retention at low SNR levels. This is similar to the information obtained in Table 3, where the k-FCR method's TPR is higher when signal quality is poorer (Phase 2). However, this increase is actually meaningless, as the TNR significantly decreases simultaneously, indicating that more unreliable data is being retained.

For the RD and PD methods based on OPTICS (Fig. 12b, 12c, 12e, and 12f), their sensitivity to the internal structure of the data may lead to a filtering tendency towards specific data regions, disrupting the original trend. A clear example is the red-boxed data in Fig. 12b and 12c, where the TPR for RD and PD methods is significantly lower than the surrounding cells. This selection tendency for specific data blocks, combined with an already insignificant change trend, makes the TPR of RD and PD methods insensitive to changes in the number of daily profiles. From Fig. 12e and 12f, it can be seen that the boundary between high and low TNR for the RD and PD methods shifts to the right as the number of daily profiles increases. This indicates that with more daily profiles, the RD and PD methods can more effectively remove unreliable data from higher SNR data. Combined with their performance in TPR, where the TPR does not show significant changes with an increasing number of daily profiles, the TNR significantly improves. This means that the RD and PD methods perform better with a higher number of daily profiles. This is likely because a higher number of daily data profiles helps to form distinct structural differences between reliable and unreliable data within the data, enhancing the ability to remove unreliable data without affecting the retention rate of reliable data. Of course, there are exceptions when the number of profiles is between 1 and 3.9, mainly due to the small sample size.

From the perspective of SNR, the boundary between high and low TPR and TNR in the PD method is the furthest to the right among the three methods, while the boundary for the RD method lies between the PD method and the k-FCR method. This is consistent with the conclusions in Sect. 4.1. Notably, for SNR greater than 19dB, both the RD and PD methods show an increase in TNR, which is particularly evident in the PD method, while the k-FCR method does not exhibit this phenomenon. This may be because, at very high SNRs, the differences in data feature structure between reliable and unreliable data

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


become more pronounced, and the OPTICS algorithm can sensitively capture these differences. In contrast, the DBSCAN algorithm is less sensitive to such differences. This observation aligns with existing studies that describe these algorithms' characteristics, specifically that DBSCAN performs poorly in reflecting internal structural changes within data point clusters, affecting the performance of the k-FCR method based on DBSCAN. This is precisely why this paper also proposes two additional quality control methods based on the OPTICS algorithm.

Additionally, although the reachability distance used in the RD method and the predecessor used in the PD method represent entirely different aspects, the quality control results of both methods show a certain degree of similarity. This suggests a potential implicit connection between the reachability distance and the predecessor derived from the OPTICS algorithm.

## 5. Conclusion

This paper is the first to apply the DBSCAN and OPTICS density clustering algorithms to the quality control of PRR temperature lidar historical data. Utilizing three newly proposed methods (k-FCR, RD, and PD), it is possible to automatically determine data quality segmentation points and achieve automated quality control of historical detection data. The PD method, in particular, introduces the concept of predecessor divergence as a criterion for data quality, resulting in excellent quality control performance.

Quality control effectiveness was validated on data from 182 detection days between March 2021 and May 2024. Overall, the three methods exhibit complementary strengths in quality control. The k-FCR method retains the most amount reliable data but requires tolerance for some residual unreliable data. The PD method is the most thorough in eliminating unreliable data but results in the loss of some reliable data. The RD method strikes a balance between the two. In terms of robustness, the PD method performs the best, showing good applicability to the historical data validated in this paper without the need for fuzzy decision-making. In contrast, the k-FCR and RD methods involve some degree of manual interaction, and their quality control effectiveness is sensitive to this interaction. Compared to the traditional SNR method, the PD method achieves a higher TNR with a comparable TPR, while the RD method achieves a higher TPR with a comparable TNR. The k-FCR method's advantage lies in achieving a high TPR. This paper also conducted a statistical analysis of the impact of the number of daily profiles and SNR of data points on quality control results. The findings show that higher numbers of daily profiles increase the tolerance of the k-FCR method to low SNR, while the RD and PD methods more accurately eliminate unreliable data. These indirectly indicate the dependency of density clustering algorithms on the size of sample data. The choice of method should be based on the specific data needs and conditions in practical applications.

Given the continuity of atmospheric parameter variations, it is anticipated that the quality control methods proposed in this paper will also be applicable to the historical data quality control of Rayleigh temperature lidars. With certain optimizations, they may also be suitable for Doppler wind lidars. As detection technology advances and device costs decrease, atmospheric lidars are gradually transitioning from research instruments to observation tools. Unattended observation network has become a development trend. In the future, the large volumes of atmospheric environment measurement data will inevitably

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


face the challenge of automated data quality control. This paper provides an approach for the automated quality control of historical data for PRR temperature lidar and also makes a useful attempt at the automated quality control of the data from atmospheric environment network in future, thereby making a contribution to the further utilization of atmospheric environment data.

## Appendix A

According to the publication (Shimaoka K, Kinoshita M, Fujii K, et al., 2008), for the measurement model $y = f(x_1, x_2, \dots, x_n)$, uncertainty propagation follows the method below:

$$u_y = \sqrt{\sum_{n=1}^{N}\left(\frac{\partial y}{\partial x_n}\right)^2 + 2\sum_{m=1}^{N-1}\sum_{n=m+1}^{N}\frac{\partial y}{\partial x_n}\frac{\partial y}{\partial x_m}r_{nm}u_n u_m}\,, \tag{A1}$$

In the formula, $y$ and $u_y$ represent the output of model and its uncertainty, $x_{n,m}$ and $u_{n,m}$ are the input parameters of model and their uncertainties, $r_{n,m}$ is the correlation coefficient, and $\partial y/\partial x_{n,m}$ is the partial derivative of $y$ with respect to $x_{n,m}$. Applying Eq. (A1) to Eq. (4) yields:

$$u_Q = \sqrt{\left(\frac{u_{N_{\mathrm{PRR1}}}}{u_{N_{\mathrm{PRR2}}}}\right)^2 + \left(\frac{N_{\mathrm{PRR1}}}{N_{\mathrm{PRR2}}^2}u_{N_{\mathrm{PRR2}}}\right)^2}\,, \tag{A2}$$

Considering only photon noise, $u_{N_{\mathrm{PRR}i}} = \sqrt{N_{\mathrm{PRR}i}}$ $(i = 1,2)$, Eq. (A2) can be simplified to:

$$u_Q = \frac{N_{\mathrm{PRR1}}}{N_{\mathrm{PRR2}}}\sqrt{\left(\frac{\sqrt{N_{\mathrm{PRR1}}}}{N_{\mathrm{PRR1}}}\right)^2 + \left(\frac{\sqrt{N_{\mathrm{PRR2}}}}{N_{\mathrm{PRR2}}}\right)^2}\,, \tag{A3}$$

Evidently, Eq. (A3) is identical to Eq. (5).

## Data availability

The Raman lidar data and radiosonde data can be provided for non-commercial research purposes upon request (csy@bit.edu.cn). The ERA5 data can be downloaded from https://cds.climate.copernicus.eu/cdsapp#!/dataset/ reanalysis-era5-pressure-levels?tab=overview (last access: 20 June 2024).

## CRediT authorship contribution statement

**Rongzheng Cao**: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Siying Chen**: Writing – review & editing, Project administration, Funding acquisition, Conceptualization. **Wangshu Tan**: Writing – review & editing, Methodology, Formal analysis, Project

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


administration. **Yixuan Xie**: Writing – review & editing, Data curation. **He Chen**: Writing – review & editing, Data curation. **Pan Guo**: Writing – review & editing, Data curation. **Rui Hu**: Data curation. **Yinghong Yu**: Data curation. **Jie Yu**: Software. **Shusen Yao**: Writing – review & editing.

## Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We are very grateful to the ECMWF for ERA5 data.

## References

Adam, S., Behrendt, A., Schwitalla, T., Hammann, E., and Wulfmeyer, V.: First assimilation of temperature lidar data into an NWP model: impact on the simulation of the temperature field, inversion strength and PBL depth, Quarterly Journal of the Royal Meteorological Society, 142, 2882–2896, https://doi.org/10.1002/qj.2875, 2016.

Alcayaga, L.: Filtering of pulsed lidar data using spatial information and a clustering algorithm, Atmospheric Measurement Techniques, 13, 6237–6254, https://doi.org/10.5194/amt-13-6237-2020, 2020.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J.: OPTICS: ordering points to identify the clustering structure, SIGMOD Rec., 28, 49–60, https://doi.org/10.1145/304181.304187, 1999.

Aspey, R. A., McDermid, I. S., Leblanc, T., Walsh, D., and Howe, J.: New Raman water vapor and temperature lidar at JPL Table Mountain Facility: optimization, validations, and Sonde intercomparison, in: Lidar Technologies, Techniques, and Measurements for Atmospheric Remote Sensing II, 72–81, 2006.

Beyer, A., Wania, F., Gouin, T., Mackay, D., and Matthies, M.: Temperature Dependence of the Characteristic Travel Distance, Environ. Sci. Technol., 37, 766–771, https://doi.org/10.1021/es025717w, 2003.

Brecheisen, S., Kriegel, H.-P., Kröger, P., and Pfeifle, M.: Visually Mining Through Cluster Hierarchies, in: Proceedings of the 2004 SIAM International Conference on Data Mining, Proceedings of the 2004 SIAM International Conference on Data Mining, 400–411, https://doi.org/10.1137/1.9781611972740.37, 2004.

Burkow, I. C. and Kallenborn, R.: Sources and transport of persistent pollutants to the Arctic, Toxicology letters, 112, 87–92, 2000.

Chen, H., Chen, S., Zhang, Y., Guo, P., Chen, H., and Chen, B.: Overlap determination for temperature measurements from a pure rotational Raman lidar, JGR Atmospheres, 121, 2805–2813, https://doi.org/10.1002/2015JD024163, 2016.

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


Chen, S., Qiu, Z., Zhang, Y., Chen, H., and Wang, Y.: A pure rotational Raman lidar using double-grating monochromator for temperature profile detection, Journal of Quantitative Spectroscopy and Radiative Transfer, 112, 304–309, https://doi.org/10.1016/j.jqsrt.2010.07.002, 2011.

Esmaelnejad, J., Habibi, J., and Yeganeh, S. H.: A Novel Method to Find Appropriate ε for DBSCAN, in: Intelligent Information and Database Systems, Berlin, Heidelberg, 93–102, https://doi.org/10.1007/978-3-642-12145-6_10, 2010.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, in: kdd, 226–231, 1996.

Gerding, M., Höffner, J., Lautenbach, J., Rauthe, M., and Lübken, F.-J.: Seasonal variation of nocturnal temperatures between 1 and 105 km altitude at 54° N observed by lidar, Atmos. Chem. Phys., 8, 7465–7482, https://doi.org/10.5194/acp-8-7465-2008, 2008.

He, J., Chen, S., Zhang, Y., Guo, P., and Chen, H.: A Novel Calibration Method for Pure Rotational Raman Lidar Temperature Profiling, J. Geophys. Res. Atmos., 123, 10,925-10,934, https://doi.org/10.1029/2018JD029062, 2018a.

He, J., Chen, S., Zhang, Y., Guo, P., and Chen, H.: A Novel Calibration Method for Pure Rotational Raman Lidar Temperature Profiling, J. Geophys. Res. Atmos., 123, 10,925-10,934, https://doi.org/10.1029/2018JD029062, 2018b.

He, J., Chen, S., Zhang, Y., Guo, P., and Chen, H.: Accurate inversion of tropospheric bottom temperature using pure rotational Raman lidar in polluted air condition, Optics Communications, 452, 88–94, https://doi.org/10.1016/j.optcom.2019.07.030, 2019.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., and Schepers, D.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, 2020.

Karami, A. and Johansson, R.: Choosing DBSCAN Parameters Automatically using Differential Evolution, IJCA, 91, 1–11, https://doi.org/10.5120/15890-5059, 2014.

Khan, K., Rehman, S. U., Aziz, K., Fong, S., and Sarasvady, S.: DBSCAN: Past, present and future, in: The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014), The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014), 232–238, https://doi.org/10.1109/ICADIWT.2014.6814687, 2014.

Klonecki, A., Hess, P., Emmons, L., Smith, L., Orlando, J., and Blake, D.: Seasonal changes in the transport of pollutants into the Arctic troposphere-model study, J. Geophys. Res., 108, 2002JD002199, https://doi.org/10.1029/2002JD002199, 2003.

Lawrence, H., Bormann, N., Sandu, I., Day, J., Farnan, J., and Bauer, P.: Use and impact of Arctic observations in the ECMWF Numerical Weather Prediction system, Quarterly Journal of the Royal Meteorological Society, 145, 3432–3454, https://doi.org/10.1002/qj.3628, 2019.

Luers, J. K. and Eskridge, R. E.: Use of Radiosonde Temperature Data in Climate Studies, Journal of Climate, 11, 1002–1019, https://doi.org/10.1175/1520-0442(1998)011<1002:UORTDI>2.0.CO;2, 1998.

Mandelbrot, B. B. and Mandelbrot, B. B.: The fractal geometry of nature, WH freeman New York, 1982.

Sander, J., Qin, X., Lu, Z., Niu, N., and Kovarsky, A.: Automatic Extraction of Clusters from Hierarchical Clustering Representations, in: Advances in Knowledge Discovery and Data Mining, vol. 2637, edited by: Whang, K.-Y., Jeon, J., Shim,

https://doi.org/10.5194/egusphere-2024-2650
Preprint. Discussion started: 21 October 2024


K., and Srivastava, J., Springer Berlin Heidelberg, Berlin, Heidelberg, 75–87, https://doi.org/10.1007/3-540-36175-8_8, 2003.

Sawant, K.: Adaptive methods for determining DBSCAN parameters, International Journal of Innovative Science, Engineering & Technology, 1, 329–334, 2014.

Schubert, E. and Gertz, M.: Improving the Cluster Structure Extracted from OPTICS Plots, in: LWDA, 318–329, 2018.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X.: DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN, ACM Trans. Database Syst., 42, 19:1-19:21, https://doi.org/10.1145/3068335, 2017.

Shimaoka K, Kinoshita M, Fujii K, et al.: Evaluation of measurement data-supplement 1 to the guide to expression of uncertainty in measurement-propagation of distributions using a monte carlo method., JCGM, 2008.

Stull, R. B.: An introduction to boundary layer meteorology, Springer Science & Business Media, 2012.

Sun, N., Zhong, L., Zhao, C., Ma, M., and Fu, Y.: Temperature, water vapor and tropopause characteristics over the Tibetan Plateau in summer based on the COSMIC, ERA-5 and IGRA datasets, Atmospheric Research, 266, 105955, https://doi.org/10.1016/j.atmosres.2021.105955, 2022.

Thundathil, R. M.: Convective-scale data assimilation of thermodynamic lidar data into the weather research and forecasting model, 2023.

Virman, M., Bister, M., Räisänen, J., Sinclair, V. A., and Järvinen, H.: Radiosonde comparison of ERA5 and ERA-Interim reanalysis datasets over tropical oceans, Tellus A: Dynamic Meteorology and Oceanography, 73, 1–7, https://doi.org/10.1080/16000870.2021.1929752, 2021.

Wandinger, U.: Raman Lidar, in: Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere, edited by: Weitkamp, C., Springer, New York, NY, 241–271, https://doi.org/10.1007/0-387-25101-4_9, 2005.

Yan, Q., Wang, Y., Gao, T., Gao, F., Di, H., Song, Y., and Hua, D.: Optimized retrieval method for atmospheric temperature profiling based on rotational Raman lidar, Appl. Opt., 58, 5170, https://doi.org/10.1364/AO.58.005170, 2019.