A Gaussian Mixture Method for Specific Differential Phase Retrieval at X-band Frequency

Guang Wen[1], Neil I. Fox[1,*], and Patrick S. Market[1]

[1]School of Natural Resources, University of Missouri, 332 ABNR Building, Columbia, Missouri, USA, 65201

Correspondence: Neil I. Fox (foxn@missouri.edu)

Abstract. Specific differential phase K_{dp} is one of the most important polarimetric radar variables, but the variance $\sigma^2(K_{dp})$, regarding the errors in the calculation of the range derivative of differential phase shift Φ_{dp}, is not well characterized due to the lack of a data generation model. This paper presents a probabilistic method based on Gaussian mixture model for K_{dp} estimation at X-band frequency. The Gaussian mixture method can not only estimate the expected values of K_{dp} by differentiating the expected values of Φ_{dp}, but also obtain $\sigma^2(K_{dp})$ from the product of the square of the first derivative of K_{dp} and the variance of Φ_{dp}. Additionally, ambiguous phase and backscattering differential phase shift are corrected via the mixture model. The method is qualitatively evaluated with a convective event of a bow echo observed by the X-band dual-polarization radar in the University of Missouri. It is concluded that K_{dp} estimates are highly consistent with the gradients of Φ_{dp} in the leading edge of the bow echo, and large $\sigma^2(K_{dp})$ occurs with high variation of K_{dp}. Furthermore, the performance is quantitatively assessed by two-year radar-gauge data, and the results are compared to linear regression model. It is clear that K_{dp}-based rain amounts have good agreement with the rain gauge data, while the Gaussian mixture method gives improvements over linear regression model, particularly for far ranges.

1 Introduction

Apart from radar reflectivity (Z_H) and differential reflectivity (Z_{DR}), polarimetric radars also obtain differential phase shift (Φ_{dp}) to reflect the forward scattering property of hydrometeor scatterers (Seliga and Bringi, 1978; Sachidananda and Zrnić, 1986). Its range derivative, also called specific differential phase (K_{dp}), has some advantages over Z_H and Z_{DR} (Zrnić and Ryzhkov, 1996), including insensitivity to attenuation, clutter, partial beam blockage and radar absolute calibration. The specific differential phase has played a key role in various meteorological applications—such as hydrometeor classification (Lim et al., 2005; Park et al., 2009), raindrop size distribution retrieval (Bringi et al., 2002; Williams et al., 2014) and quantitative precipitation estimation (Ryzhkov et al., 2005; Cifelli et al., 2011)—since K_{dp} is a phase variable independent of Z_H and Z_{DR} and almost linearly proportional to rain rate.

A linear regression model has been developed to derive K_{dp} from the slope of the range profile of Φ_{dp} for S-band radars. In Hubbert et al. (1993), Φ_{dp} is first processed by a light filter that attenuates the Φ_{dp} magnitudes within a scale of 375 m by 10 dB, and then heavily smoothed in 1.5 km by 10 dB. An iterative filtering technique is used for eliminating non-zero backscattering differential phase shift (Hubbert and Bringi, 1995). The filtered Φ_{dp} measurements are finally fitted into a first-order polynomial to estimate the Φ_{dp} slope in a given window. Liu et al. (1993) supply the accuracy of mean K_{dp} as $\pm$0.25 deg km^{-1} using 128 pulses, while Aydin et al. (1995) indicate that the accuracy is within $\pm$0.5 deg km^{-1} for a heavy rainfall event using 64 pulses. On the other hand, Ryzhkov and Zrnić (1996) produce two kinds of K_{dp} for S-band radars: one is obtained over 16 range gates (2.4 km) for $Z_H \leq$40 dBZ, and the other is produced over 48 gates (7.2 km) for Z_H>40 dBZ. Negative K_{dp} is incorporated into the rain rate algorithm to avoid bias in low rain rate. The analyses of 15 storms show that the standard error of K_{dp} is 0.04~0.10 deg km^{-1} for heavily filtered K_{dp} and 0.12~0.30 deg km^{-1} for lightly filtered K_{dp} using either 128 or 64 pulses. Under the complex terrain, the K_{dp} retrieval algorithm needs to be modified to obtain the accurate rainfall rate. Vulpiani et al. (2012) has opened new scenarios for the operational K_{dp} processing in the Italian C-band radar network. In addition, Gorgucci et al. (1999) note that the nonuniform rainfall path produces large errors in the K_{dp} estimates, while the errors increase as the radar reflectivity varies in dimensions.

X-band dual-polarization radars have drawn increasing attention in the radar meteorology community in recent years on account of low cost, fine resolution and high sensitivity to light precipitation (Chandrasekar et al., 2012; Lim et al., 2013; Berne and Krajewski, 2013; Kalogiros et al., 2014; Oue et al., 2016). In the literature, X-band algorithms have been proposed for K_{dp} estimation. For example, the linear regression method is adapted for the X-band radar data, and used to retrieve rainfall (Matrosov et al., 2006). The ambiguous Φ_{dp} is naturally corrected by examining the complex values of the range profiles of Φ_{dp} exponentials, and K_{dp} is then estimated by a regularization framework based on a cubic spline smoothing (Wang and Chandrasekar, 2009). In this method, the bias and variance are adjustable through the smoothing parameter, giving high spatial resolutions of K_{dp} estimates. Comparing to S-band frequency, the Φ_{dp} measurements at X-band frequency are affected by backscattering differential phase shift δ_{co}. Nevertheless, the constraints of $K_{dp} - Z_H - Z_{DR}$ and $\delta_{co} - Z_{DR}$ can be used to improve the estimation of K_{dp} and δ_{co} (Otto and Russchenberg, 2011; Reinoso-Rondinel et al., 2018). Algorithms of linear programming (Giangrande et al., 2013) and Kalman filter (Schneebeli et al., 2014) have also been applied to the K_{dp} estimation, yielding good performance for rainfalls and snowfalls. It is noticeable that the Kalman filter method minimizes the Gaussian error function to obtain the mean profile of K_{dp}. It gives a significant improvement on the K_{dp} mean, particularly in the small-scale structure with high peaks.

The recent algorithms are focused on the improvement of estimating the mean K_{dp}, whereas its variance $\left(\sigma^2(K_{dp})\right)$ is not well characterized due to the lack of a data generation model. The K_{dp} variance is often inherited from the Φ_{dp} variance $\left(\sigma^2(\Phi_{dp})\right)$ leading to large relative errors for low K_{dp} with a fixed path length. In this study, we propose a probabilistic method based on Gaussian mixture model for K_{dp} estimation at X-band frequency. The Gaussian mixture method can not only estimate the expected values of K_{dp} by differentiating the conditional expectation of Φ_{dp}, but also yield $\sigma^2(K_{dp})$ by regarding the errors in the calculation of the first derivative of Φ_{dp}. It is found that $\sigma^2(K_{dp})$ is closely related to the square of the first derivative of K_{dp} and $\sigma^2(\Phi_{dp})$, while large $\sigma^2(K_{dp})$ is associated with high variation of K_{dp} estimates. When compared to the existing methods, our method considers the joint probability density function of the data as the non-linear Gaussian mixture, leading to better performance for the multimodal data. The K_{dp} variance can be used to calculate the variances of Z_H, Z_{DR} and rain rate, and to study the streamflow trends in the hydrological model.

The paper is organized as follows. Section 2 provides background information about K_{dp} and the Gaussian mixture model. Section 3 describes the radar and gauge data. Section 4 presents the methodology. We first remove the residual clutter using data masks (section 4.1), and then derive the joint probability density function to estimate the expected value of Φ_{dp} and $\sigma^2(\Phi_{dp})$ (section 4.2). Next, we correct the ambiguous phase and δ_{co} via the mixture model (section 4.3). Last, we calculate the expected value and variance of K_{dp} (section 4.4), and improve the K_{dp} profile by reducing $\sigma^2(K_{dp})$ (section 4.5). To evaluate the algorithm, section 5 gives a case study and a comparison between radar and gauge. Section 6 summarizes the paper.

2 Background

The specific differential phase is the first derivative of differential phase shift Φ_{dp} along the radar range, giving a way to estimate K_{dp} by radar measurement of Φ_{dp}. Furthermore, the probability density function of Φ_{dp} can be modelled as a Gaussian mixture, which is often obtained via an expectation-maximization (EM) approach. The mean and variance of the Gaussian mixture may lead to the improvement of the K_{dp} estimation.

In this section, we introduce the physical interpretation of K_{dp} and the regression model for estimating K_{dp}. Since the Gaussian mixture is adopted as the data generation model, we also give a brief description of mathematical definition of the Gaussian mixture model and the EM approach.

2.1 Specific differential phase (K_{dp})

For linear polarization, K_{dp} is proportional to the integral of the raindrop size distribution and the real part of the difference of forward scattering amplitudes at orthogonal polarizations. It is mathematically formulated as

$$K_{dp} = \frac{0.18\lambda}{\pi} \int_0^\infty N(D) \cdot \Re\left[f_{hh}(0,D) - f_{vv}(0,D)\right] dD \quad (\text{deg km}^{-1}), \tag{1}$$

where λ is radar wavelength in millimeters, D is raindrop size in millimeters, $N(D)$ is size spectrum in m^{-3}mm^{-1}, $f_{hh,vv}(0,D)$ is forward scattering amplitudes at horizontal and vertical polarizations, respectively.

By considering the Rayleigh-Gans scattering from identical and horizontally-oriented oblate spheroids, such as raindrops, the forward scattering amplitudes are proportional to the inverse square of radar wavelength, i.e., $f_{hh,vv}(0,D) \propto 1/\lambda^2$, leading to the fact that K_{dp} is inversely proportional to radar wavelength, i.e., $K_{dp} \propto 1/\lambda$. Therefore, the values of K_{dp} at X-band are often larger than that at S-band by a factor of 3, indicating that X-band radar can provide better K_{dp} data than S-band radar when retrieving the rainfall rate. The conclusion is still valid even if the Mie effect is taken into account (Bringi and Chandrasekar, 2001; Chandrasekar et al., 2006).

However, K_{dp} cannot be detected by polarimetric radar directly, whereas its integral Φ_{dp} is measurable. Hence, K_{dp} can be estimated as the range derivative of the profile of Φ_{dp}, i.e., $K_{dp} = \frac{\Delta\Phi_{dp}}{2\Delta r}$,where r is the radar range in kilometers. An alternative approach to estimating K_{dp} is to apply a regression fit to the profile of Φ_{dp}, and the first order polynomial is usually considered as the fitting function (Balakrishnan and Zrnić, 1990; Ryzhkov and Zrnić, 1995). Subsequently, if the Φ_{dp} measurements are equally spaced in range by Δr, K_{dp} is then estimated by

$$K_{dp} = \frac{\sum_{i=1}^{n} \Phi_{dp}(r_i)\left[i - \frac{(n+1)}{2}\Delta r\right]}{\frac{1}{6}n(n-1)(n+1)\Delta r^2}, \tag{2}$$

where n is the number of gates. Equation (2) shows that the accuracy of K_{dp} estimates is determined by the number of gates (n), and the accuracy of Φ_{dp}. By assuming $\sigma^2(\Phi_{dp})$ is relatively stable for all gates along a ray and noting that $\Phi_{dp}(r_i)$ is the only variable in Eq. (2), $\sigma^2(K_{dp})$ is formulated as

$$\sigma^2(K_{dp}) = \frac{\sigma^2(\Phi_{dp})}{\frac{1}{3}\Delta r^2\left[n(n-1)(n+1)\right]}. \tag{3}$$

In Eq. (3), $\sigma^2(K_{dp})$ is proportional to $\sigma^2(\Phi_{dp})$, which is related to the spectrum width, cross-correlation coefficient, and the dwell time (Sachidananda and Zrnić, 1986; Hubbert et al., 1993), and inversely proportional to n^3. This method has been widely used in the existing radar system (Cifelli et al., 2018; Chandrasekar et al., 2018; Chen et al., 2017c, b). The details of the regression-based estimation of K_{dp} are given in Bringi and Chandrasekar (2001) and Appendix A.

Moreover, it is notable that the backscattering phase shift is not negligible at X-band, thus the total propagation phase shift (Ψ_{dp}) consists of Φ_{dp} and the backscattering differential phase, δ_{co}, i.e., $\Psi_{dp} = \Phi_{dp} + \delta_{co}$. The backscattering phase shift is often showed as a sudden jump over one or few range gates in a monotonically increasing Ψ_{dp} profile of rain (May et al., 1999a), with a value much larger than standard deviation $\sigma(\Psi_{dp})$. The presence of δ_{co} over a small number of consecutive gates can be eliminated by a simple filter (Hubbert and Bringi, 1995).

The specific differential phase is a unique polarimetric variable in terms of statistical errors in the rain rate estimation, since it is the range derivative of the phase measurement Φ_{dp}. The errors in the calculation of the first derivative also needs to be taken into account. In this study, we consider a Gaussian mixture as the data generation model, which plays an important role in the estimation of K_{dp} and $\sigma^2(K_{dp})$.

2.2 Gaussian mixture model

The Gaussian mixture is a statistical model for data probability density estimation, assuming that the data points are generated by a mixture of a finite number of Gaussian distributions associated with their weights (McLachlan and Peel, 2000; Sung, 2004). Intuitively, it is used to model the multimodal data, with each Gaussian component corresponding to a subpopulation of the data. The mathematical formulation is given as

$$f(\mathbf{z}) = \sum_{i=1}^{m} w_i \mathcal{N}(\mathbf{z}; \mu_i, \Sigma_i), \tag{4}$$

where m is the number of components in the Gaussian mixture, w_i is a weight with $\sum_{i=1}^{m} w_i = 1$, and $\mathcal{N}(z;\mu_i,\Sigma_i)$ is the ith Gaussian distribution with mean μ_i and covariance Σ_i, i.e., $\mathcal{N}(z;\mu_i,\Sigma_i) = \left|(2\pi)^k \Sigma_i\right|^{-1/2} \exp\left[-\frac{1}{2}(z-\mu_i)^T \Sigma_i^{-1}(z-\mu_i)\right]$, where k is the data dimension.

It is prevalent to use an Expectation–Maximization (EM) algorithm to estimate the parameters, w, μ, and Σ, by constructing the lower bound of log-likelihood based on Jensen inequality (Dempster et al., 1977). The EM algorithm is divided into two steps, namely, an expectation (E) step and a maximization (M) step. In the E step, a degree of membership toward to the jth cluster is calculated, i.e.,

$$Q_j^i = p\left(y^{(i)} = j | x^{(i)}; w, \mu, \Sigma\right), \tag{5}$$

where i is the ith data with a total number of n data points, and y is a latent variable that determines the corresponding cluster. Here, Q gives a tight lower bound for the log-likelihood, equivalent to maximizing the expectation. In the M step, the exact form of the lower bound based on Jensen inequality is expressed as

$$\mathcal{L}(w,\mu,\Sigma) = \sum_i \sum_j Q_j^i \log \frac{\exp\left[-\frac{1}{2}(x^i-\mu)^T \Sigma^{-1}(x^i-\mu)\right] w_j}{\sqrt{|(2\pi)^k \Sigma|}\, Q_j^i}. \tag{6}$$

By maximizing the lower bound with respect to each parameter, w_j, μ_j, and Σ_j are updated as (Petersen and Pedersen, 2012)

$$w_j = \frac{\sum_i Q_j^i}{n}, \tag{7}$$

$$\mu_j = \frac{\sum_i Q_j^i x^{(i)}}{\sum_i Q_j^i}, \tag{8}$$

$$\Sigma_j = \frac{\sum_i Q_j^i (x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T}{\sum_i Q_j^i}, \text{respectively.} \tag{9}$$

Notably, the M step increases the log-likelihood monotonically, and the covariance retains positive definite with sufficiently large data samples. Finally, the E step and M step are iteratively operated until the log-likelihood converges to a value with the difference between two successive steps below a certain threshold. In addition, the EM algorithm requires a specification of the number of clusters, m, prior to the E and M steps, and an inappropriate choice of m may lead to meaningless values of the parameters. To tackle this problem, the Bayesian information criterion is often calculated to select the optimal m, while a Dirichlet process may also be used to model a prior probability to construct an infinite Gaussian mixture.

One of interpretations of the Gaussian mixture is to view each distribution as a cluster with a Gaussian probability density, while the individual data point is attributed to a specific cluster or a weight toward the cluster, regarded as unsupervised learning (Hastie et al., 2009). The clustering procedures based on Gaussian mixture have been applied to the identification of storm structure (Veneziano and Villani, 1996), and the particle identification at S-band (Wen et al., 2015, 2016b, 2017) and X-band (Wen et al., 2016a) frequencies. Furthermore, the Gaussian mixture model can be extended to fit a set of unknown parameters in the prior probability of the Bayesian framework, forming a Bayesian Gaussian mixture model (Li et al., 2012). The prior is then multiplied with the known conditional probability of data given the parameters to be estimated, yielding the posterior probability with a new set of parameters. The expectation of the posterior is often used to retrieve the conditional mean of the new parameters based on least square criteria.

For the regression problem, the characteristics of the Gaussian mixture imply that the direct modeling of a regression function is very difficult. Nevertheless, the joint probability of the measurements and the estimated parameters may be modeled as a Gaussian mixture, leading to a regression function derived from the joint density model. Due to the asymptotic consistency of a Gaussian mixture model, it is capable of estimating a general density function in $\mathbb{R}^n$ in any shape (Sung, 2004). Moreover, the speed of calculating unknown parameters within a Gaussian mixture linearly depends on the number of the training data points, and the computation of the outputs is independent of the size of the training data. Consequently, regression based on a Gaussian mixture can be achieved very rapidly, compared to Gaussian process regression that grows with the data size. In addition, the Gaussian mixture can also be used to solve the regression problem with multiple dimensions, and a subset of dimensions can be selected to handle the missing data (Wen et al., 2015).

3 Data

As part of the Missouri Experimental Project to Stimulate Competitive Research (EPSCoR), an X-band dual-polarization radar in the University of Missouri (MZZU) was deployed at the South Farm Research Center (38.906°N, 92.269°W) in the midwest of America in the summer of 2015. The details of the radar characteristics are described in Simpson and Fox (2017). The primary objective is to provide the observations of precipitation near the surface by means of low-cost and fine-scale X-band radar, and to fill the observational gaps of the S-band radar network in Saint Louis (KLSX), Kansas City (KEAX), and Springfield (KSGF). Within the MZZU radar coverage, the Hinkson Creek located near Columbia, MO, flows through a catchment basin and eventually merges into the Missouri river, forming a typical urban watershed (Hubbart and Zell, 2013). The radar can provide timely flash flooding warning for the Hinkson Creek watershed and surrounding areas.

In this study, we analyze the data collected by the X-band MZZU dual-polarization radar. The maximum unambiguous range of the MZZU radar is 94.64 km with a resolution of 260 m in range and 1° in azimuth. During the observational periods, the radar operates in a volumetric scanning mode of nine elevations at 0.8°, 2°, 3°, 4°, 5°, 6°, 7°, 8.5°, and 10°, updated every 4 minutes. The raw radar data are organized and processed by an open-source software package called Python ARM Radar Toolkit (Py-ART: Helmus and Collis, 2016). Moreover, to validate the K_{dp} estimation algorithm, we also use the data from tipping-bucket rain gauges in the Missouri Mesonet weather station network, including Bradford Farm (38.897°N, 92.218°W), Sanborn Field (38.942°N, 92.320°W), Auxvasse (39.089°N, 91.999°W), and Williamsburg (38.907°N, 91.734°W). The horizontal distances between the rain gauges and the radar center are 4.4 km, 6.0 km, 30.8 km, and 46.2 km, respectively. The first elevations at Bradford and Sanborn may be affected by ground clutter, since the radar beams are very close to the ground, with heights of 314.6 m and 336.9 m ASL, respectively, including the radar tower. Therefore, the second elevation at 2° is selected for validation. In contrast, the first elevations at Auxvasse and Williamsburg reach about 723.8 m and 999.0 m ASL, which are less contaminated by ground clutter. Furthermore, the point measurement of rain gauge is different from the volumetric measurement of radar, imposing additional errors on the comparison between radar and gauge (Anagnostou et al., 1999). The radar-based rain rate is then derived by averaging K_{dp} over three successive range gates and three successive azimuthal rays with a total of 9 values centered over each gate in order to obtain good consistency between the instruments. In addition, the rain gauges are carefully calibrated in terms of instrumentation failure, clogging, and other discrepancies between the devices (Simpson and Fox, 2017), and well documented to provide long-term data for rainfall observations.

Table 1 summarizes the characteristics of rainfalls observed at Bradford, Sanborn, Auxvasse and Williamsburg between April 2016 and June 2018. It is clear that the hourly rain amounts are dominated by light rain, with similar means of 2.0–2.1 mm at the four sites, indicating uniformly distributed rainfalls within the experimental region. On the other hand, the standard deviations of Bradford and Williamsburg are 3.5 mm and 3.7 mm, respectively, a little larger than that of 3.3 mm at Sanborn and Auxvasse. Moreover, Sanborn gives the highest hourly rain amount, the lowest total rain amount, and the second lowest duration out of the four sites, due to the effects of urban heat island (Hubbart et al., 2014). The second highest maximum hourly rain amount is recorded at Williamsburg, however, the total rain amount and duration are also the highest among the four sites, implying that convective rain is the most frequent at Williamsburg. In contrast, stratiform rain is more common at Bradford, since the gauge records the lowest maximum hourly rain amount and duration, and the second total hourly rain amount. In addition, it can be seen that Auxvasse also provides useful data for the comparisons between gauges and between radar and gauge, though the statistics are all ranked in the middle of the four sites. Overall, the rain gauge data at Bradford, Sanborn, Auxvasse and Williamsburg are representative and sufficiently large, leading to a valid dataset for testing the K_{dp} and K_{dp}-based rain amounts.

4 K_{dp} retrieval

As discussed in section 2, the joint probability density function (PDF) based on a Gaussian mixture can be used to derive the regression model for K_{dp} estimation. The Gaussian mixture method (GMM) not only estimates the expected values of K_{dp} by differentiating the conditional expectation of Φ_{dp}, but also gives an estimation of K_{dp} variance by regarding the errors in the calculation of the first derivative of Φ_{dp}. In this section, we describe GMM for the K_{dp} estimation using MZZU radar data. Figure 1 illustrates the flowchart of GMM (Fig. 1.b), comparing to that of the linear regression model (LR; Fig. 1.a).

From the chart of LR in Fig. 1.a, we can see that after the radar measurements are collected, the Ψ_{dp} is unfolded, and then the clutter is removed. After these corrections, an iterative filtering method is applied to the Ψ_{dp} profile. An adaptive method is finally used to estimate the K_{dp} profile according to the values of Z_H. The Gaussian mixture model, on the other hand, processes Ψ_{dp} differently. First of all, the clutter is masked out according to the thresholds of Z_H and the variation of Ψ_{dp}. Secondly, the range r and Ψ_{dp} are fitted into a Gaussian mixture to yield the joint PDF, while the Ψ_{dp} mean and the Ψ_{dp} variance are obtained by taking the first raw and second central moments of the conditional PDF of Ψ_{dp} given r. Thirdly, some specific clusters in the Gaussian mixture PDF are adjusted to solve the problems of ambiguous Ψ_{dp} and backscattering differential phase shift δ_{co} in order to derive the PDF of Φ_{dp}. Fourthly, a raw K_{dp} profile is calculated from the first derivative of the expected values of Φ_{dp}, and the associated variances are obtained via a Taylor series expansion. Finally, the raw K_{dp} profile is smoothed, and consequently, the variances are reduced. In addition, new Φ_{dp} with lower variances can be re-constructed from the K_{dp} estimates.

4.1 Data masking

The presence of clutter in the Ψ_{dp} measurements may severely affect the K_{dp} estimation, producing significantly large variations on the estimates. It is well known that the effect of clutter can be reduced by applying a spectrum filter to the time-series data (e.g., May and Strauch, 1998; Hubbert et al., 2009). However, some residual clutter echoes are still shown on the radar measurements including Ψ_{dp} (Wen et al., 2017). Therefore, the clutter needs to be well handled in GMM, prior to the deviation of the regression model based on the joint PDF.

In LR, the clutter is often eliminated by some criteria based on Ψ_{dp} or ρ_{hv}. For instance, we use the thresholds of local standard deviation of Ψ_{dp} less than 10° to classify valid points. Further, ten consecutive range gates of valid points signify the beginning of a rain cell, and five consecutive gates of invalid points finish the associated rain cell. Overall, the thresholds give a fairly good performance on the MZZU radar, however, the clutter may be incorrectly identified in the regions of high reflectivity or for the echoes mixed by weather and clutter, which are often associated with large Ψ_{dp} variation.

In contrast, GMM adopts sophisticated procedures, as depicted in Figure 2. It is clear that there are five stages in the data masking, beginning with the input of raw Ψ_{dp} and ending with masked data. At the first stage, the raw data are fitted to a Gaussian mixture initialized by the k-means clustering, while the covariance is set to be diagonal for simplicity. The clusters with no more than 5 points are promptly masked out, before they pass to the second stage. Stages two, three and four of the process all involve the clusters. At the second stage, the clusters are validated according to two sets of thresholds with respect to mean reflectivity. For the MZZU radar, the ratio of the standard deviations, $\sigma(\Psi_{dp})/\sigma(r)$, less than 14.2° km^{-1}, and $\sigma(\Psi_{dp})$ less than 4.1° are used for Z_H less than 41 dBZ. To reduce the mis-classification in the hail regions, the thresholds are increased for higher Z_H, resulting in $\sigma(\Psi_{dp})/\sigma(r) < 47.9°$ km^{-1} and $\sigma(\Psi_{dp}) < 6.3°$. Next, the entire Ψ_{dp} profile is divided into multiple rain cell segments by considering the gaps between two consecutive clusters. Similar to the first stage, the segments containing no more than 5 points are excluded from the output of masked data. Following this, the dominant is determined for each segment by comparing the weight accumulations of weather and clutter clusters. For a clutter segment with mean height below 200 m, the clusters within the segment are re-evaluated by thresholds of $\sigma(\Psi_{dp})/\sigma(r) < 2.0°$ km^{-1} and $\sigma(\Psi_{dp}) < 0.8°$, on the other hand, the clusters in a weather segment are re-examined using $\sigma(\Psi_{dp})/\sigma(r) < 34.7°$ km^{-1} and $\sigma(\Psi_{dp}) < 6.1°$. This step can efficiently identify the clutter-contaminated weather echoes, which are often associated with large variances. At the last stage, some isolated points along the azimuth are obscured in the final results.

Figure 3 illustrates two examples for data masking, including a convective case (Fig. 3.a) and a stratiform case (Fig. 3.b). The data points in the two cases show steadily increasing trends related to anisotropic media along the wave propagation path. However, between 1.3 and 15 km at an azimuth of 252° in the convective case (Fig. 3.a), the data present significant fluctuations with the minimum value at about 0° but the maximum value at 180°. Since the dynamic range of Ψ_{dp} is from 0 to 180° for the MZZU radar, the measurements near the ground are likely to be the clutter returns, verifying the results of data masking. After 15 km, the Ψ_{dp} points start from about 50° and go all the way up to 180°. Notwithstanding this trend, the points sharply decrease to about 10° at about 40 km, indicating the occurrence of phase folding. The data masking can effectively detect the phase folding, and provide valid masked data for deriving the joint PDF. On the other hand, the weather echoes are more frequently observed at 1° in azimuth in the stratiform case (Fig. 3.b). By taking a closer inspection on the Ψ_{dp} data, we can discern that the points largely fluctuate between 40 and 80 km due to low signal-to-noise ratio. In LR, these points may be incorrectly discarded based on $\sigma(\Psi_{dp})$ thresholds, leading to some missing data in the stratiform regions. In contrast, the data masking accurately identifies weather echoes characterized by a number of vertically-oriented density ellipses. The continuous and uniformly-distributed regimes are consistent with the physical interpretation of stratiform precipitation. In addition, the data masking is also sensitive to sudden jumps at the beginning of the Ψ_{dp} data, which may be caused by δ_{co}.

4.2 Ψ_{dp} density estimation

In the previous section, it is shown that the Ψ_{dp} profile varies along the range r. It rises quickly for horizontally-oriented anisotropic scatterers, and conversely, it falls steadily for vertically-oriented particles (Marzano et al., 2010). To estimate the relationship between r and Ψ_{dp}, we consider r as an independent variable, denoted as x, and Ψ_{dp} as a dependent variable, denoted as y. If the minimization of mean square error is required, the regression function is obtained by taking the average value of y at fixed x, equivalent to estimating the expected values of y conditioned on x, i.e.,

$$\bar{y}(x) = E(y|x) = \int y\, p(y|x,\beta)\, dx, \tag{10}$$

where β is a set of unknown variables, for example, $\beta = (m, w, \mu, \Sigma)$ for the mixture model. Since the Gaussian mixture can be used to model any shapes of probability density with a rapid speed, the (x,y) points are then assumed to follow a joint PDF of Gaussian mixture, as defined in Eq. (4). Moreover, the properties of the multivariate Gaussian distribution in each cluster determine the Gaussianity of the marginal distribution of either variable and the conditional distribution of one variable given the other (Bishop, 2006). Therefore, the conditional PDF of y given x is expressed as

$$p(y|x,\beta) = \sum_{i=1}^{m} w_i^{y|x} \mathcal{N}\left(y; \mu_i^{y|x}, \Sigma_i^{y|x}\right), \text{ with} \tag{11}$$

$$\mu_i^{y|x} = \mu_i^{y} + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(x - \mu_i^{x}), \tag{12}$$

$$\Sigma_i^{y|x} = \Sigma_i^{yy} - \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}\Sigma_i^{xy}, \tag{13}$$

$$w_i^{y|x} = \frac{f_i(x)}{f(x)} = \frac{w_i \mathcal{N}(x; \mu_i^{x}, \Sigma_i^{xx})}{\sum_{j=1}^{m} w_j \mathcal{N}(x; \mu_j^{x}, \Sigma_j^{xx})}, \tag{14}$$

where w_i, $\mu_i = (\mu_i^x, \mu_i^y)^T$ and $\Sigma_i = \begin{pmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{pmatrix}$ are obtained by the EM algorithm. In Eq. (14), $f(x)$ is the marginal PDF of x with the parameters identical to the mixture, and $f_i(x)$ is the weighted marginal PDF of each cluster, i.e., $f(x) = \sum_{i=1}^{m} f_i(x)$. By substituting Eq. (11) into Eq. (10) and noting the linearity of the mathematical expectation, the expected value of y conditioned on x is then obtained as

$$E(y|x) = \sum_{i=1}^{m} \frac{f_i(x)}{f(x)} (a_i x + b_i), \text{ with} \tag{15}$$

$$a_i = \Sigma_i^{yx} (\Sigma_i^{xx})^{-1}, \tag{16}$$

$$b_i = \mu_i^y - \Sigma_i^{yx} (\Sigma_i^{xx})^{-1} \mu_i^x, \tag{17}$$

and the conditional variance is given as (see Appendix B)

$$\sigma^2(y|x) = \sum_{i=1}^{m} w_i^{y|x} \left[\Sigma_i^{y|x} + \left(\mu_i^{y|x} \right)^2 \right] - \left(\sum_{i=1}^{m} w_i^{y|x} \mu_i^{y|x} \right)^2 . \tag{18}$$

Equations (15) and (18) play an important role in the joint PDF-based regression analysis, called the regression and skedastic functions (Spanos, 1999). In Eq. (15), it can be seen that the regression function in GMM consists of multiple linear kernels, which is similar to LR. However, the weighting function $w_i^{y|x}$ is not determined by the local structure but the marginal PDF of global data x. Comparing to LR, GMM is more flexible to capture the data information, while it still retains a finite set of parameters. Moreover, Eq. (18) readily estimates the point-wise variances $\sigma^2(y|x)$ that characterize the random errors in the measurements, whereas these errors $\sigma(\Psi_{dp})$ are often considered as small and stable values in LR.

Figure 4 compares the Ψ_{dp} profiles given by Eqs. (15) and (18) with that obtained by LR. Figure 4.a gives the same example as Fig. 3.a, but the EM algorithm is configured differently. In the Ψ_{dp} density estimation, the mixture with full covariance yields density ellipses of random shapes. Furthermore, the algorithm repeats the fitting procedures three times to avoid the local maxima of the log-likelihood. Meanwhile, the choice of the cluster number relies on the Bayesian information criterion calculated for each m, starting at 10 clusters. It can be seen that the mixture composed by density ellipses well characterizes the data points, since the root-mean-square error is small relative to the expected values. Between 15 and 35 km, the narrow ellipses result in Ψ_{dp} with a rising trend consistent with LR. On the other hand, the mixture has very small variances, giving a high confidence for the fitted parameters. From 35 km, the ellipses become wider, and the associated variances increase due to low signal-to-noise ratio at the edge of radar echoes. What is notable, however, is that the Ψ_{dp} profile dramatically increases to a large value, whereas LR remains a relatively steady trend. It indicates the importance of the Ψ_{dp} unfolding for the Ψ_{dp} density estimation.

Figure 4.b presents another example of the density estimation. It is clear that the Ψ_{dp} profiles produced by GMM and LR both rise considerably along the range, and the trends for the two methods are very similar with a strong correlation of 0.998. The profile starts at about 50°, and remains relatively stable before rising dramatically between 35 and 55 km. By 65 km, Ψ_{dp} has more than doubled, and then, there is a steady increase for Ψ_{dp} reaching about 130° at the end of the profile, which is around 70° up on the ranges of 0 and 35 km, and 10° more than recorded at the ranges of 55 and 65 km. If we examine Ψ_{dp} measured at X-band frequency, we can see that some points fall out of the dash lines corresponding to one standard error (i.e., 95% interval). Most notably, between 18 and 20 km, the Ψ_{dp} profile shows a sudden slump, indicating the occurrence of δ_{co}. In conclusion, the expected value and the variance of Ψ_{dp} can be obtained from the joint PDF, but the mixture needs to be tuned in terms of Ψ_{dp} unfolding and δ_{co} elimination in order to obtain the PDF of Φ_{dp}.

4.3 Ψ_{dp} unfolding and δ_{co} elimination

According to the continuity and consistency of the phase data, we can discern that some issues exist in the density estimation, such as ambiguous Ψ_{dp} and δ_{co}. Since Ψ_{dp} is an range accumulative measurement of propagation phase, depending on the initial $\Psi_{dp}(0)$. The measurements may exceed the dynamic range of 0–180° when the wave propagates through a rain medium. This situation is even more significant at X-band frequency than S-band due to the inverse relation of the wavelength and the rate of phase shift. Nevertheless, it can be noted that Ψ_{dp} gives a non-negative trend along the range for rain, and therefore, the ambiguous Ψ_{dp} may be corrected accordingly (Wang and Chandrasekar, 2009).

In LR, Ψ_{dp} is first averaged over a small window for weather data, and a linear fit is then performed to obtain the increment for the range gate next to the window. In the following stage, a reference is predicted by summing up the average and the increment, and compared to the observed value at the same gate. If the difference between the predicted and observed values is larger than 90°, the observed Ψ_{dp} is then increased by 180°. Finally, the correction process is iteratively operated until the last gate.

On the other hand, the Ψ_{dp} unfolding is more straightforward in GMM. Figure 5 shows the flowchart of the Ψ_{dp} unfolding and the δ_{co} elimination. After obtaining the PDF of Ψ_{dp}, the initial step of the Ψ_{dp} unfolding selects the density ellipses with at least 6 data points. Next, the second step calculates the difference of the means μ_i between the two consecutive density ellipses along the range. At this point, the PDF of Ψ_{dp} is ready to be corrected for ambiguous Ψ_{dp}. In the final step, the mean of the latter density ellipse is added up 180°, if the former mean is larger than the latter one by 80°.

As illustrated in Fig. 4.a, the profile Ψ_{dp} reaches 180° at about 38 km, and then becomes ambiguous between 38 and 42 km. In LR, the Ψ_{dp} values at these locations are interpolated according to the trend of the previous few gates, and the maximum value is 180°. In contrast, the corrected density ellipses in GMM show an upward trend between 38 and 42 km, while the Ψ_{dp} profile reaches a maximum value of about 195°, indicating the effectiveness of the Ψ_{dp} unfolding in the region of heavy rain.

In addition to ambiguous Ψ_{dp}, the estimation of the joint PDF may also be affected by non-zero δ_{co}, which is defined as the phase difference between the horizontal and vertical polarizations upon the backscattering of the particles in a radar resolution volume. This effect occurs more frequently at X-band frequency than S-band due to Mie scattering (Trömel et al., 2013). The δ_{co} is shown as a sudden phase change over a small number of gates in a monotonically increasing trend for rain. According to this manifestation, the magnitude and gate number of the Ψ_{dp} perturbation can be used to eliminate δ_{co} (Matrosov et al., 2002; Otto and Russchenberg, 2011).

The linear regression model often adopts an iterative filter technique, which generates a new Φ_{dp} profile from either the raw data or the filtered one based on a threshold (Hubbert and Bringi, 1995). If the filtering alters the data by 4°, the new profile selects the filtered data, otherwise the raw data are remained. The new profile is then used as input in the next iteration until the convergence condition is satisfied.

As shown in Fig. 5, the δ_{co} elimination is embedded into the process of the Ψ_{dp} unfolding. For two consecutive density ellipses, the latter density ellipse is removed if its mean is larger than the former one by 85°. Prior to this step, the mean of the first density ellipse in the mixture should be below 90° to reduce the δ_{co} effect at the first few gates. Since δ_{co} occurs over a small number of range gates, a mixture pruning is also employed to remove the density ellipses with weights less than 0.0501, equivalent to 2% of the data.

It is clear from Fig. 4.b that δ_{co} has occurred at multiple locations in the data. The Ψ_{dp} profile starts at a high value and drops somewhat over the first two gates. Notably, there is a narrow gap between 18 and 20 km, which is non-zero δ_{co}. These data are characterized by a density ellipse with a slightly decreasing trend in GMM, and the resulting expected values are consistent with the filtered data in LR. Between 70 and 90 km, a few isolated points beyond the density ellipses are associated with δ_{co}. Both of the two methods can produce Φ_{dp} following the main trend of the data, which suggests that the process is effective for the δ_{co} elimination.

4.4 K_{dp} density estimation

As discussed in section 2.1, K_{dp} is the first derivative of Φ_{dp} with respect to the range r. According to the mean value and dominated convergence theorems, the derivative of the expected value of Φ_{dp} conditioned on r is equal to the expected value of the derivative of Φ_{dp} with respect to r, i.e., K_{dp} (see Appendix C). Following the notation in section 4.2, we denote K_{dp} as y'. Therefore, the expected value of K_{dp} is obtained by taking the derivative of Eq. (15), yielding

$$E\left(y'|x\right) = \frac{1}{f^2(x)}\left\{\sum_{i=1}^{m}\sum_{j=1}^{m} f_i(x)f_j(x)\left[\left(\frac{x-\mu_j^x}{\Sigma_j^{xx}} - \frac{x-\mu_i^x}{\Sigma_i^{xx}}\right)(a_i x + b_i) + a_i\right]\right\}. \tag{19}$$

The variance of y' conditioned on x can be approximated by the first-order Taylor series expansion (see Appendix D), i.e.,

$$\sigma^2\left(y'|x\right) = \left[E''(y|x)\right]^2 \sigma^2(y|x), \tag{20}$$

where $\sigma^2(y|x)$ is given in Eq. (18). By taking the derivative of Eq. (19), $E''(y|x)$ is expressed as

$$E''(y|x) = 2\left[\sum_{i=1}^{m} a_i \left(w_i^{y|x}\right)'\right] + \sum_{i=1}^{m}(a_i x + b_i)\left(w_i^{y|x}\right)''. \tag{21}$$

From Eq. (C8) in Appendix C, it is clear that

$$\left(w_i^{y|x}\right)' = \frac{g_i(x)}{f^2(x)} = \frac{1}{f^2(x)}\sum_{j=1}^{m} f_j(x)f_i(x)\left(\frac{x-\mu_j^x}{\Sigma_j^{xx}} - \frac{x-\mu_i^x}{\Sigma_i^{xx}}\right), \tag{22}$$

where $g_i(x)$ is the summation term. Subsequently, the second derivative of $w_i^{y|x}$ is given as

$$\left(w_i^{y|x}\right)'' = \frac{g_i'(x)f(x) - 2f'(x)g_i(x)}{f^3(x)}, \quad \text{where} \tag{23}$$

$$f'(x) = -\sum_{j=1}^{m}\left(\frac{x-\mu_j^x}{\Sigma_j^{xx}}\right) f_j(x), \tag{24}$$

$$g_i'(x) = \sum_{j=1}^{m} f_j(x)f_i(x)\left[\left(\frac{x-\mu_i^x}{\Sigma_i^{xx}}\right)^2 - \left(\frac{x-\mu_j^x}{\Sigma_j^{xx}}\right)^2 + \frac{1}{\Sigma_j^{xx}} - \frac{1}{\Sigma_i^{xx}}\right]. \tag{25}$$

Equations (19) and (20) are the regression and skedastic functions for the K_{dp} estimation. In Eq. (19), it is clear that the expected value of K_{dp} can be divided into two components, including Eqs. (C7) and (C11). On one hand, Eq. (C7) is related to the changing rate a_i weighted by the marginal distribution of each cluster in the mixture, equivalent to a linearly weighted combination of small portions of data. If a data point is dominated by a specific cluster, i.e., the weight of a cluster is significantly larger than the others, K_{dp} is determined by the coefficients of the cross-correlation and auto-correlation of r, and independent of the means and auto-correlation of Φ_{dp}, yielding a constant value within the dominated cluster. On the other hand, Eq. (C11) shows that the weighting function also contributes to the K_{dp} estimates by considering the Gaussian derivative of the Φ_{dp} estimates in two or three adjacent clusters along the range. The sign of K_{dp} is then determined by the marginal means and variances of the clusters, weighted by the difference of their contributions to Φ_{dp}.

In Eq. (20), it can be seen that $\sigma^2(K_{dp})$ is proportional to $\sigma^2(\Phi_{dp})$, which is similar to Eq. (3) in LR. However, $\sigma^2(\Phi_{dp})$ varies along the range due to the random errors of the Φ_{dp} estimates in GMM, whereas $\sigma^2(\Phi_{dp})$ is stable in LR. In addition, the statistical errors with respect to signal processing may be included in Eq. (20) as an additive term, independent of Φ_{dp}. Moreover, the radar gate spacing and gate number for the K_{dp} estimation are often fixed in Eq. (3), indicating $\sigma^2(K_{dp})$ is also stable in LR. In contrast, $\sigma^2(K_{dp})$ in GMM is closely related to the first derivative of K_{dp} in Eq. (20). As the changing rate of K_{dp} increases, the random errors associated with the K_{dp} estimates rise dramatically.

Figure 6.b illustrates K_{dp} and its variance estimated from Φ_{dp} in Fig. 6.a, which is the same case as given in Figs. 3.a and 4.a. It is apparent that the K_{dp} estimates present a large fluctuation, while the associated variances are significant. In GMM, K_{dp} starts from about 0.5 deg km^{-1}, and then fluctuates between 17 and 20 km and between 24 and 42 km. In the profile, there are six local peaks with the maximum at about 8.5 deg km^{-1}. Meanwhile, the K_{dp} variances vary as the K_{dp} estimates change. Between 15 and 17 km and between 20 and 24 km, the K_{dp} estimates stand at a value, leading to small K_{dp} variances in these regions. When short excursions are present, such as that between 18 and 20 km, K_{dp} variances increase significantly due to the contribution of the first derivative of K_{dp} in Eq. (20). Furthermore, the large Φ_{dp} variances between 35 and 42 km also result in an increase of the K_{dp} variances. In contrast, LR gives less fluctuation in K_{dp} estimates with two peaks at about 20 and 34 km. The comparison of K_{dp} obtained by the two methods may suggest that a smoothing procedure is required to reduce the significant variance in GMM.

4.5 K_{dp} smoothing

As discussed previously, the K_{dp} variance is small for high K_{dp}, but relatively large for low K_{dp}. Therefore, an adaptive estimation is adopted in LR. For radar reflectivity (Z_H) less than 20 dBZ, the gate number n in Eq. (2) is set as 15, while n is 8 for 20 $\leq Z_H < 35$ dBZ, and 2 for $Z_H \geq 35$ dBZ, respectively. On the other hand, GMM also applies an adaptive technique based on finite impulse filter (FIR) to the expected values of K_{dp} in order to reduce the associated variances. Figure 7 shows the time responses of the FIR with the cutoff frequency of 0.053 and the Gaussian window of 28, which yield the best performance for the MZZU radar. The impulse response (Fig. 7.a) is peaked at the center, and gradually decreases towards the two ends. Furthermore, the step response (Fig. 7.b) gives the accumulation of the impulse response, indicating that the magnitudes around the center change faster than that at the two ends. If a longer window is required, the order of the FIR is increased accordingly.

In this study, we gradually increase the order number to calculate the difference between the K_{dp} profiles obtained by the FIR filters with two adjacent order numbers. The optimal order of the FIR filter is then set when the relative square error of the two K_{dp} is below 0.001. For profiles with sufficiently large data points, the order number is between 29 and 33 for the MZZU radar.

To obtain the reduced variance, we consider the filter as a number of weighting functions, denoted as $h_i(x)$, and subsequently, the smoothed data become

$$y = \sum_{i=1}^{n} h_i * x_i \tag{26}$$

where y is a smoothed data point, x_i is the original data within the smoothing window, and n is the window length. By taking the variance on both sides of Eq. (26), we have

$$\sigma^2(y) = \sum_{i=1}^{n} h_i^2 \sigma^2(x_i). \tag{27}$$

Therefore, the variance of the smoothed data is the weighted sum of the variances of the original data within the smoothing window. Since the FIR coefficients are much less than unity, $\sigma^2(y)$ is smaller than $\sigma^2(x)$ at the same gate. Furthermore, the K_{dp} estimates with the reduced variances can be used to re-construct Φ_{dp} to obtain smaller Φ_{dp} variances. For a fixed gate spacing Δr, the re-constructed Φ_{dp} for the jth range gate is

$$\Phi_{dp}^{j} = \sum_{i=1}^{j} K_{dp}^{i} \Delta r, \quad \text{and} \tag{28}$$

$$\sigma^2(\Phi_{dp}^{j}) = \sum_{i=1}^{j} \sigma^2(K_{dp}^{i}) \Delta r^2. \tag{29}$$

The red curves in Figs. 6.a and 6.b illustrate the re-constructed Φ_{dp} and the smoothed K_{dp} using FIR, respectively. The smoothed K_{dp} in Fig. 6.b is more consistent with the LR results compared to the original K_{dp} produced by the GMM. In the first few kilometers, the smoothed K_{dp} gradually rises, and then peaks at about 21 km. With no fluctuations, the smoothed K_{dp} falls gradually, followed by a growth before reaching a plateau at 33 km. After a slight decrease between 33 and 36 km, K_{dp} rises dramatically, which is very different from LR. Meanwhile, the variances are small at the beginning, but get larger as K_{dp} is climbing up. Between 20 and 33 km, the K_{dp} estimates do not change very much, leading to small variances in this region. But after 33 km, the variances begin to increase and retain large values until the end of the profile. Overall, the smoothed K_{dp} is stable, producing a profile considerably consistent with LR, and the variances are significantly reduced comparing to the original data. In addition, the re-constructed Φ_{dp} (Fig. 6.a) constantly increases with few local fluctuations, while the associated variances are smaller than the Φ_{dp} variances in GMM.

5 Evaluation

In this section, a case study is first presented to qualitatively analyze the storm structure and evolution based on K_{dp} . The radar-gauge dataset is then used to provide a quantitative evaluation for the K_{dp} estimation in terms of root mean squared error

(RMSE), normalized bias (NB) and Pearson correlation coefficient (ρ_{RG}), which are defined as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(R_i - G_i\right)^2}{N}}, \tag{30}$$

$$\mathrm{NB} = \frac{\sum_{i=1}^{N}(R_i - G_i)}{\sum_{i=1}^{N} G_i}, \tag{31}$$

$$\rho_{RG} = \frac{\sum_{i=1}^{N}(R_i - \bar{R})(G_i - \bar{G})}{\sqrt{\sum_{i=1}^{N}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{N}(G_i - \bar{G})^2}}, \tag{32}$$

where N is the sample size, R_i is the individual radar hourly rain amount, G_i is the gauge data, and $\bar{R}$ and $\bar{G}$ are the sample means for radar and gauge, respectively. The radar hourly rain amount is calculated based on the CASA radar rainfall algorithm, which is given as (Wang and Chandrasekar, 2010; Chen and Chandrasekar, 2015)

$$R(K_{dp}) = 18.15 K_{dp}^{0.79}, \tag{33}$$

where R is the instantaneous rain rate in mm h^{-1}. It is noted that the radar collects instantaneous measurements every 4–5 minutes, whereas RGs obtain the precipitation accumulations over 60 minutes. Therefore, it is necessary to average 12–15 consecutive radar scans to derive the hourly rain amounts.

5.1 Case study

On 24 March 2016, a severe storm developed in central Missouri and moved eastward across Columbia, MO, causing strong winds and heavy precipitation at the surface. When the storm became mature, the radars at Kansas City and St. Louis observed the storm structure at high levels, since each radar was about 150 km away from the storm. Notably, the Kansas City radar showed positive and negative Doppler velocities in a small area (not shown), indicating the occurrence of a downburst. On the other hand, the MZZU radar illustrated a bow echo of Z_H close to the radar center (Figs. 8.b). In addition to Z_H, the GMM-based K_{dp} (Figs. 8.d, e and f) was also obtained to investigate the storm structure near the surface.

Figure 8 illustrates that the convective storm evolves from a strong and large echo to a bow shape echo, and then dissipates at far range. At 0304 UTC (Fig. 8.a), a cell with strong Z_H moves into the radar area, while K_{dp} is moderate with a maximum of about 3 deg km^{-1} (Fig. 8.d). As the cell is transforming to a bow shape, the radar echo becomes intensive, and forms a rain band with embedded convective cores (Fig. 8.b). It is clear to see that K_{dp} reaches over 10 deg km^{-1} in these core regions (Fig. 8.e), indicating very heavy precipitation at the surface. With the fast movement of the storm, the downburst has been weaken, and the storm starts to dissipate (Fig. 8.c). At 0441 UTC, it can be seen that K_{dp} is gradually reduced at the far range, while its maximum is much less than that at the mature stage.

In this storm, the bow echo is shown as a number of convective cores embedded in a rain band, while the downbursts occured at the leading edge near the echo center. The bow echo can be considered as a mesoscale convection with a horizontal dimension of more than 60 km. To gain a further insight, Fig. 9 shows raw Φ_{dp} and K_{dp} for the bow echo. In Fig. 9.a, raw Φ_{dp} presents large gradients along the leading edge, rising from about 50° to over 140°. Due to the sharp increase, Φ_{dp} exceeds the maximum dynamic range, leading to ambiguity in the areas of X:-20~-18 km and Y:12~18 km and X:-40~-23 km and Y:-8~-5 km. In addition, the echoes behind the convective cores occasionally vanish as a result of signal attenuation. Nevertheless, LR (Fig. 9.b) produces continuous K_{dp} by Φ_{dp} unfolding and linear interpolation according to the trends of the profiles, but some missing data still exist within the storm, due to low signal-to-noise ratio. In contrast, GMM (Fig. 9.c) corrects these data with the expected values derived from the joint PDF, and simultaneously obtains the statistical errors in the production of K_{dp}. It is evident that the GMM method can efficiently handle the missing data via the mixture model, which is another advantage over the LR model. Furthermore, the statistical errors are not very large in these areas, since the missing data are filled by the distribution of the entire data profile. Additionally, the GMM K_{dp} estimates are generally a few deg km^{-1} higher than the LR ones, particularly for the regions of high Z_H.

By taking a closer look at GMM K_{dp}, we can see that the bow echo is generally characterized by K_{dp} of above 2.5 deg km^{-1}, while five pockets of high K_{dp} are identified. In the bow head, the first pocket presents very high K_{dp} associated with a rapid growth of Φ_{dp}. Behind this pocket, there is a region of negative K_{dp}, whereas LR generally yields positive values. It may be due to a reduction of cross-correlation coefficient caused by low signal-to-noise ratio, since the signals have been significantly attenuated after propagating through the pocket. In the middle of the second and third pockets in the bow center, LR and GMM both show lower K_{dp} comparing to the two pockets, while K_{dp} is substantially consistent with the gradient of Φ_{dp} in the area. By considering the high Z_H in Fig. 8, these moderate K_{dp} values may indicate less anisotropic scatterers, such as small hail in the process of wet growth. Similarly, a hail signature with maximum Z_H of above 66 dBZ and small K_{dp} of 1~2 deg km^{-1} can also be identified in the middle of the fourth and fifth pockets in the bow tail. Along with the expected values of GMM K_{dp}, Fig. 9.d depicts the statistical errors $\sigma(K_{dp})$ in the calculation of the expected values. The five pockets of high K_{dp} are generally associated with small $\sigma(K_{dp})$ of a few tenths deg km^{-1}. However, the estimates behind the top four pockets yield very large $\sigma(K_{dp})$ with a maximum above 10 deg km^{-1}, and the expected values of K_{dp} are sometimes below 0 deg km^{-1}, such as X:-25~-20 km and Y:11~20 km. In contrast, a region of high $\sigma(K_{dp})$ appears in front of the bottom pocket, superimposed on the high Z_H area associated with hail. In conclusion, the GMM K_{dp} estimates of high confidence give good agreement with the gradients of Φ_{dp} in the leading edge of the bow echo, while large $\sigma(K_{dp})$ are expected at the region of high variation of the K_{dp} estimates.

Moreover, the computational time is crucial for the real-time application of the K_{dp} retrieval algorithms. For the data in Fig. 9, the GMM takes about 7.058/4.068 seconds to process the K_{dp} with/without the data masking, whereas the LR reduces the time to about 2.037 seconds. It indicates that the LR has the advantages of simplicity and efficiency. Nevertheless, the GMM can obtain more information from the radar data, which is useful for the model studies.

5.2 Statistical analysis

In order to quantitatively evaluate the accuracy of GMM K_{dp}, hourly accumulated rain amounts are derived from the X-band rainfall rate algorithm (Chen and Chandrasekar, 2015), and compared to the rain gauge data collected at Bradford, Sanborn, Auxvasse and Williamsburg between 1 April 2016 and 2 June 2018. The scatterplots presented in Fig. 10 illustrate the comparison between GMM-based radar and gauge rain amounts, and the accompanying table (Table 2) gives RMSE, NB and ρ_{RG} results obtained by GMM and LR.

Consistent with the data in Table 1, the rainfall at the four sites is predominately made up of light rain with hourly rain amounts no more than 2.5 mm h^{-1}. Nevertheless, according to Fig. 10, moderate rain with amounts between 2.6 and 8 mm h^{-1} gives a considerable contribution to the total rain events, followed by a small portion of heavy rain with amounts more than 8 mm h^{-1}. When we study the scatterplots and statistics for each of the four sites, it is apparent that Bradford (Fig. 10.a and b) and Sanborn (Fig. 10.c and d) are more concentrated on the red line than Auxvasse (Fig. 10.e and f) and Williamsburg (Fig. 10.g and h), since Bradford and Sanborn are closer to the radar. Accordingly, RMSEs for Bradford and Sanborn (Table 2) are relatively small, about 13%~35% lower than Auxvasse and Williamsburg. Furthermore, it can be seen that Bradford and Sanborn show negative bias associated with negative NBs, indicating an underestimation of rain amounts by GMM K_{dp}. In contrast, a slight overestimation may be concluded for Auxvasse and Williamsburg by considering the point trends and the positive NBs. Additionally, Sanborn claims the highest ρ_{RG} out of the four sites, yielding the best consistency between radar and gauge.

When compared to LR statistics as given in Table 2, it is clear that GMM improves the RMSEs, NBs and ρ_{RG} for Auxvasse and Williamsburg. Notably, the GMM-based NB for Auxvasse reaches a very small value of 0.04, one fifths of LR-based NB. For Bradford, RMSE is reduced by GMM, but the absolute value of NB is slightly increased, while ρ_{RG} remains the same. On the other hand, for Sanborn, the GMM-based RMSE, NB and ρ_{RG} get worse by a few hundredths of millimeters, which may be due to the local complex terrain near the radar. Overall, the rainfall estimates of GMM K_{dp} give a better performance than that of LR in terms of RMSE, NB and ρ_{RG} at the far ranges.

It is clear that the rain rates based on the GMM K_{dp} have a moderate consistency with the rain gauge data. To improve the results, some advanced rain rate algorithms can be considered, such as the rain-ice separation technique in the IFloodS campaign (Chen et al., 2017b) and the radar-gauge comparison method in the MC3E campaign (Giangrande et al., 2014). Nevertheless, the GMM has the advantage over the existing methods, since it can yield the variance of K_{dp}. Furthermore, the variance of R can also be obtained by the K_{dp} mean and the K_{dp} variance via the R–K_{dp} relation, leading to the variability in the error characteristics of R. Thus, the variances can be used to study the streamflow trends in the hydrological model.

6 Summary and discussions

In this study, we proposed a probabilistic method based on Gaussian mixture model to estimate the specific differential phase K_{dp}, which is the range derivative of differential phase shift Φ_{dp}. The Gaussian mixture method (GMM) not only obtained the expected values of K_{dp} by differentiating the conditional expectation of Φ_{dp}, but also yielded the variance $\sigma^2(K_{dp})$ regarding the errors in the calculation of the first derivative of Φ_{dp}.

As an initial step of GMM, the data masking was performed to eliminate the residual clutter in the measurements of the total differential phase (Ψ_{dp}). The data of r and Ψ_{dp} were first fitted into a simplified Gaussian mixture to generate a number of clusters, which were validated against the two sets of the $\sigma(\Psi_{dp})$ and $\sigma(\Psi_{dp})/\sigma(r)$ thresholds given by radar reflectivity Z_H.

The clusters were then combined to form the rain cell segments, and the segments were classified by comparing the weight accumulations of weather and clutter clusters. Next, the clusters within each segment were re-evaluated by the thresholds according to the segment types. Finally, the azimuthally isolated points were masked out.

Secondly, the joint probability density function (PDF) was obtained by fitting the data of r and Ψ_{dp} into a mixture model with full covariance, where the cluster number m, weight w, mean μ and covariance Σ were optimized via the Expectation-Maximization (EM) algorithm. Subsequently, the PDF of Ψ_{dp} conditioned on r was also a mixture with parameters related to the joint PDF. Finally, the Ψ_{dp} mean was estimated by the conditional expectation, and the statistical errors $\sigma^2(\Psi_{dp})$ were given by the conditional variance, which was not always constant, but varied with w and the marginal PDF of r.

Thirdly, the ambiguous Ψ_{dp} and backscattering differential phase shift δ_{co} were corrected by examining the two adjacent density ellipses in the mixture. On one hand, if the former density ellipse had a mean larger than the latter one by 80°, the latter mean was added to 180° for Ψ_{dp} unfolding. On the other hand, if the former mean was smaller than the latter one by 85°, the latter density ellipse was removed as δ_{co}. Moreover, for δ_{co} elimination, the first density ellipse mean was assumed as below 90°, while the density ellipses with small weights were also removed.

Fourthly, the joint PDF of r and Φ_{dp} was used in the calculations of K_{dp} and $\sigma^2(K_{dp})$. Since K_{dp} was the range derivative of Φ_{dp}, the expected values of K_{dp} were then obtained via the derivative of the expected value of Φ_{dp}. Moreover, by taking the first-order Taylor series expansion, $\sigma^2(K_{dp})$ was the product of the square of the first derivative of K_{dp} and $\sigma^2(\Phi_{dp})$, yielding non-constant values of $\sigma^2(K_{dp})$.

In the final step, the expected values of K_{dp} were smoothed to reduce the associated $\sigma^2(K_{dp})$. An FIR filter was implemented, and iteratively applied to the data to search for an optimal window length. Subsequently, the reduced $\sigma^2(K_{dp})$ was obtained by the sum of the original $\sigma^2(K_{dp})$ weighted by the FIR coefficient squares within the window. Additionally, new Φ_{dp} were re-constructed from the smoothed K_{dp}, while $\sigma^2(\Phi_{dp})$ was also reduced.

The experimental results with a severe storm observed by the X-band polarimetric radar in the University of Missouri (MZZU) revealed the advantages of GMM. By studying the structure and evolution of a bow echo in the storm, it was concluded that the GMM K_{dp} was consistent with the gradients of raw Φ_{dp} along the leading edge of the bow echo, while large $\sigma^2(K_{dp})$ occurred with high variation of K_{dp}. The GMM method produced results similar to the LR method, with the ability to handle the missing data. Moreover, the hourly rain amounts based on K_{dp} were compared to the rain gauge data, showing fairly good agreement between radar and gauge measurements. The rain amounts obtained by GMM K_{dp} gave improvements over the linear regression model, particularly for the far ranges.

The potential applications of GMM K_{dp} and $\sigma^2(K_{dp})$ include quantitative precipitation estimation (Cifelli et al., 2011; Chen et al., 2017a) and attenuation correction (Park et al., 2005). For quantitative precipitation estimation, the relationship between K_{dp} and rain rate R is almost linear, since K_{dp} is about the fourth-order moment of raindrop size distribution, and R is the 3.67th-order moment. As illustrated in Fig. 10, the $R(K_{dp})$ algorithm is consistent with the in situ measurements. To further investigate the R errors, it is necessary to consider the K_{dp} errors in the calculation of the first derivative of Φ_{dp}. The standard deviation $\sigma(K_{dp})$ is then related to $\sigma(R)$ by a factor of R/K_{dp} (Bringi and Chandrasekar, 2001). In a similar manner, K_{dp} is linearly proportional to specific attenuation A_H and specific differential attenuation A_{DP} (Bringi and Hendry, 1990).

Therefore, the errors of radar reflectivity Z_H and differential reflectivity Z_{DR} may also be proportional to $\sigma(K_{dp})$ after the attenuation correction, and eventually contribute to the R errors via $R(Z_H)$ and $R(Z_H, Z_{DR})$. Moreover, the error estimates can be used to provide streamflow trends in hydrological model. In the future study, the algorithm will also be extended to other frequencies, such as C-band (Vulpiani et al., 2012; May et al., 1999b) and S-band (Bringi and Chandrasekar, 2001). The thresholds in the data masking, the Ψ_{dp} unfolding and the δ_{co} elimination will be adjusted according to the radar specifications. Nevertheless, the steps for the calculations of the PDFs of Ψ_{dp} and K_{dp} will be remained.

Appendix A: Regression-based estimation of K_{dp}

Let the total differential phase Ψ_{dp} be y, and the range gate r be x. The Ψ_{dp} profile over small range segments can be approximated by a first-order polynomial, i.e,

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{A1}$$

where β_0 and β_1 are the coefficients in the linear approximation, and ϵ is an error function. It can be assumed that ϵ is independent and individual distributed with zero mean and variance of $\sigma_\epsilon^2 = \sigma^2$.

In the linear regression, it is easy to find that

$$\beta_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}. \tag{A2}$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y in the segment, respectively. Since

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x}) y_i - \sum_i (x_i - \bar{x}) \bar{y} \tag{A3}$$

and

$$\sum_i (x_i - \bar{x}) \bar{y} = \bar{y} \left(\sum_i x_i - N\bar{x} \right) = \bar{y}(N\bar{x} - N\bar{x}) = 0, \tag{A4}$$

we have

$$\beta_1 = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}, \tag{A5}$$

where N is the number of the gates in the segment.

It is noted that the range gate r is equally spaced with an interval of Δr, Ψ_{dp} is the two-way propagation phase shift, and K_{dp} is the one-way specific differential phase. The K_{dp} is then estimated by

$$K_{dp} = \frac{\sum_{i=1}^{n} \Psi_{dp}(r_i) \left[i - \frac{(n+1)}{2} \Delta r \right]}{\frac{1}{6} n(n-1)(n+1) \Delta r^2}. \tag{A6}$$

At S-band, the backscattering differential phase shift δ_{co} is often negligible, and thus Ψ_{dp} and Φ_{dp} are interchangeable, leading to Eq. (2).

By taking the variance on both sides of Eq. (A5) and noting ϵ is the only variable, we have

$$\sigma^2(\beta_1) = \sigma^2\left(\frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon)}{\sum_i (x_i - \bar{x})^2}\right) \tag{A7}$$

$$= \frac{\sum_i (x_i - \bar{x})^2 \sigma_\epsilon^2}{\left[\sum_i (x_i - \bar{x})^2\right]^2} \tag{A8}$$

$$= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \tag{A9}$$

Similar to Eq. (A6), we have

$$\sigma^2\left(K_{dp}\right) = \frac{\sigma^2\left(\Psi_{dp}\right)}{\frac{1}{3}\Delta r^2\left[n(n-1)(n+1)\right]}. \quad \# \tag{A10}$$

Appendix B: Variance of Φ_{dp}

We consider the range r as an independent variable, denoted as x, and Φ_{dp} as a dependent variable, denote as y. The joint distribution of $z = (x, y)$ follows a Gaussian mixture as

$$p(z) = \sum_{i=1}^{m} w_i \mathcal{N}(z; \mu_i, \Sigma_i), \tag{B1}$$

where w_i, μ_i and Σ_i are the weight, mean and covariance for each component, respectively. The probability of y conditioned on x is also a Gaussian mixture with parameters $w_i^{y|x}$, $\mu_i^{y|x}$ and $\Sigma_i^{y|x}$, leading to the conditional expectation as

$$E(y|x) = \int y \sum_{i=1}^{m} w_i^{y|x} \mathcal{N}\left(y; \mu_i^{y|x}, \Sigma_i^{y|x}\right) dy, \tag{B2}$$

$$= \sum_{i=1}^{m} w_i^{y|x} \int y\, \mathcal{N}\left(y; \mu_i^{y|x}, \Sigma_i^{y|x}\right) dy, \tag{B3}$$

$$= \sum_{i=1}^{m} w_i \mu_i^{y|x}. \tag{B4}$$

and the second-order moment as

$$E(y^2|x) = \int y^2 \sum_{i=1}^{m} w_i^{y|x} \mathcal{N}\left(y; \mu_i^{y|x}, \Sigma_i^{y|x}\right) dy, \tag{B5}$$

$$= \sum_{i=1}^{m} w_i^{y|x} \int y^2 \mathcal{N}\left(y; \mu_i^{y|x}, \Sigma_i^{y|x}\right) dy, \tag{B6}$$

$$= \sum_{i=1}^{m} w_i^{y|x} \left[\Sigma_i^{y|x} + \left(\mu_i^{y|x}\right)^2\right]. \tag{B7}$$

Therefore, the conditional variance is expressed as

$$\sigma^2(y|x) = E(y^2|x) - \left[E(y|x)\right]^2 \tag{B8}$$

$$= \sum_{i=1}^{m} w_i^{y|x} \left[\Sigma_i^{y|x} + \left(\mu_i^{y|x}\right)^2\right] - \left(\sum_{i=1}^{m} w_i^{y|x} \mu_i^{y|x}\right)^2. \quad \# \tag{B9}$$

Appendix C: Conditional expectation of K_{dp}

First, we need to show that the derivative of the expected value of random variable y as a function of random variable x is equal to the expected value of the derivative of the expected value of y. By the definition, the derivative of y is expressed as

$$E'[y(x)] = \lim_{h\to 0} \frac{1}{h}\left\{E[y(x+h)] - E[y(x)]\right\} \tag{C1}$$

$$= \lim_{h\to 0} E\left[\frac{y(x+h)-y(x)}{h}\right] \tag{C2}$$

$$= \lim_{h\to 0} E\left\{y'[\tau(h)]\right\}, \tag{C3}$$

where $\tau(h) \in (x, x+h)$ exists by the mean value theorem. By assuming $|y'[\tau(h)]| \le Z$, we can use the dominated convergence theorem to obtain

$$E'[y(x)] = E\left\{\lim_{h\to 0} y'[\tau(h)]\right\} \tag{C4}$$

$$= E\left[y'(x)\right]. \tag{C5}$$

According to the conclusion in Eq. (C5), the expected value of y' is expressed as

$$E(y'|x) = E'(y|x) = \sum_{i=1}^{m} w_i^{y|x}\left(\mu_i^{y|x}\right)' + \sum_{i=1}^{m}\left(w_i^{y|x}\right)'\mu_i^{y|x}. \tag{C6}$$

Since $\left(\mu_i^{y|x}\right)' = a_i$ and $w_i^{y|x} = \frac{f_i(x)}{f(x)}$, the first term is equal to

$$\sum_{i=1}^{m} w_i^{y|x}\left(\mu_i^{y|x}\right)' = \sum_{i=1}^{m} a_i \frac{f_i(x)}{f(x)}. \tag{C7}$$

Meanwhile, the second term is given as

$$\sum_{i=1}^{m}\left(w_i^{y|x}\right)'\mu_i^{y|x} = \sum_{i=1}^{m}\frac{f_i'(x)f(x) - f_i(x)f'(x)}{f^2(x)}(a_i x + b_i). \tag{C8}$$

Based on the properties of Gaussian function, the derivatives of $f_i(x)$ and $f(x)$ are expressed as

$$f_i'(x) = -\frac{x-\mu_i^x}{\Sigma_i^{xx}} f_i(x), \quad \text{and} \tag{C9}$$

$$f'(x) = -\sum_{j=1}^{m}\frac{x-\mu_j^x}{\Sigma_j^{xx}} f_j(x). \tag{C10}$$

By substituting Eqs. (C9) and (C10) into Eq. (C8), the second term is transformed as

$$\sum_{i=1}^{m}\left(w_i^{y|x}\right)'\mu_i^{y|x} = \frac{1}{f^2(x)}\left[\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\frac{x-\mu_j^x}{\Sigma_j^{xx}} - \frac{x-\mu_i^x}{\Sigma_i^{xx}}\right) f_i(x) f_j(x)(a_i x + b_i)\right]. \tag{C11}$$

By substituting Eqs. (C7) and (C11) into Eq. (C6), we obtain

$$E\left(y'|x\right) = \frac{1}{f^2(x)}\left\{\sum_{i=1}^{m}\sum_{j=1}^{m} f_i(x) f_j(x)\left[\left(\frac{x-\mu_j^x}{\Sigma_j^{xx}} - \frac{x-\mu_i^x}{\Sigma_i^{xx}}\right)(a_i x + b_i) + a_i\right]\right\} \quad \#. \tag{C12}$$

Appendix D: Variance of K_{dp}

The first-order Taylor expansion is defined as

$$g(y) = g(\theta) + g'(\theta)(y - \theta) + \epsilon, \tag{D1}$$

where $\theta = E(y)$ is the mean of random variable y, and ϵ is the sum of the higher-order Taylor series. By considering the conclusion in Eq. (C5), it can be noted that the expected values of the coefficients associated with the derivatives in Eq. (D1) are zeros if the series is expanded at the mean value of y. By taking mathematical expectations on both sides of Eq. (D1), it is transformed as

$$E[g(y)] \approx g(\theta) + g'(\theta)[E(y) - \theta] \tag{D2}$$

$$= g(\theta). \tag{D3}$$

From Eqs. (D1) and (D3), the variance of $g(y)$ is approximated as

$$\sigma^2[g(y)] \approx E\left\{[g(y) - g(\theta)]^2\right\} \tag{D4}$$

$$\approx E\left\{[g'(\theta)(y - \theta)]^2\right\} \tag{D5}$$

$$= g'(\theta)^2 \sigma^2(y) \tag{D6}$$

Let $g(y)$ be y', and then we have

$$\sigma^2(y'|x) = [E''(y|x)]^2 \sigma^2(y|x). \tag{D7}$$

From Eq. (B9), we can see that

$$\sigma^2(y|x) = \sum_{i=1}^{m} w_i^{y|x} \Sigma_i^{y|x} + \sum_{i=1}^{m} w_i^{y|x} \left(\mu_i^{y|x}\right)^2 - \left(\sum_{i=1}^{m} w_i^{y|x} \mu_i^{y|x}\right)^2. \tag{D8}$$

By taking the derivative of Eq. (C6), the second derivative of the expected value of y conditioned on x becomes

$$E''(y|x) = \sum_{i=1}^{m} w_i^{y|x} (\mu_i^{y|x})'' + 2\sum_{i=1}^{m} (\mu_i^{y|x})'(w_i^{y|x})' + \sum_{i=1}^{m} \mu_i^{y|x} (w_i^{y|x})'' \tag{D9}$$

$$= 2\sum_{i=1}^{m} a_i (w_i^{y|x})' + \sum_{i=1}^{m} (a_i x + b_i)(w_i^{y|x})'', \tag{D10}$$

since $(\mu_i^{y|x})' = a_i$ and $(\mu_i^{y|x})'' = 0$. From Eq. (C8), the first derivative of the weighting function in the conditional probability is

$$\left(w_i^{y|x}\right)' = \frac{1}{f^2(x)} \sum_{j=1}^{m} f_j(x) f_i(x) \left(\frac{x - \mu_j^x}{\Sigma_j^{xx}} - \frac{x - \mu_i^x}{\Sigma_i^{xx}}\right). \tag{D11}$$

Let $g_i(x)$ be the summation term. The second derivative is then expressed as

$$\left(w_i^{y|x}\right)'' = \frac{g_i'(x)f(x) - 2f'(x)g_i(x)}{f^3(x)}, \quad \text{where} \tag{D12}$$

$$g_i(x) = \sum_{j=1}^{m} f_j(x) f_i(x) \left(\frac{x-\mu_j^x}{\Sigma_j^{xx}} - \frac{x-\mu_i^x}{\Sigma_i^{xx}} \right). \tag{D13}$$

According to the properties of Gaussian mixture, the first derivative of the marginal distribution of x is

$$f'(x) = -\sum_{j=1}^{m} \left(\frac{x-\mu_j}{\Sigma_j} \right) f_j(x). \tag{D14}$$

Similarly, the first derivative of $g(x)$ if given as

$$g_i'(x) = \sum_{j=1}^{m} \left[-\frac{x-\mu_j^x}{\Sigma_j^{xx}} \left(\frac{x-\mu_j^x}{\Sigma_j^{xx}} - \frac{x-\mu_i^x}{\Sigma_i^{xx}} \right) f_j(x) f_i(x) - \frac{x-\mu_i^x}{\Sigma_i^{xx}} \left(\frac{x-\mu_j^x}{\Sigma_j^{xx}} - \frac{x-\mu_i^x}{\Sigma_i^{xx}} \right) f_j(x) f_i(x) + \left(\frac{1}{\Sigma_j^{xx}} - \frac{1}{\Sigma_i^{xx}} \right) f_j(x) f_i(x) \right] \tag{D15}$$

$$= \sum_{j=1}^{m} \left[\left(\frac{x-\mu_i^x}{\Sigma_i^{xx}} - \frac{x-\mu_j^x}{\Sigma_j^{xx}} \right) \left(\frac{x-\mu_i^x}{\Sigma_i^{xx}} + \frac{x-\mu_j^x}{\Sigma_j^{xx}} \right) f_j(x) f_i(x) + \left(\frac{1}{\Sigma_j^{xx}} - \frac{1}{\Sigma_i^{xx}} \right) f_j(x) f_i(x) \right] \tag{D16}$$

$$= \sum_{j=1}^{m} \left[\left(\frac{x-\mu_i^x}{\Sigma_i^{xx}} \right)^2 - \left(\frac{x-\mu_j^x}{\Sigma_j^{xx}} \right)^2 \right] f_j(x) f_i(x) + \left(\frac{1}{\Sigma_j^{xx}} - \frac{1}{\Sigma_i^{xx}} \right) f_j(x) f_i(x) \tag{D17}$$

$$= \sum_{j=1}^{m} f_j(x) f_i(x) \left[\left(\frac{x-\mu_i^x}{\Sigma_i^{xx}} \right)^2 - \left(\frac{x-\mu_j^x}{\Sigma_j^{xx}} \right)^2 + \frac{1}{\Sigma_j^{xx}} - \frac{1}{\Sigma_i^{xx}} \right]. \quad \# \tag{D18}$$

Data availability. The MZZU radar data can be made available upon request to the authors. The rain gauge data are available online: http://agebb.missouri.edu/weather/stations/index.php.

Author contributions. NF designed the experiment and provided the radar data, GW developed the Gaussian mixture model and prepared the manuscript, GW and NF performed the validation, and NF and PM reviewed the paper.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgements. The authors would like to express our sincere thanks to the anonymous reviewers for their valuable comments and suggestions. This work was supported by Missouri Experimental Project to Stimulate Competitive Research (EPSCoR) of National Science Foundation, under Award Number IIA-1355406. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Anagnostou, E. N., Krajewski, W. F., and Smith, J.: Uncertainty Quantification of Mean-Areal Radar-Rainfall Estimates, Journal of Atmospheric and Oceanic Technology, 16, 206–215, 1999.

Aydin, K., Bringi, V., and Liu, L.: Rain-rate estimation in the presence of hail using S-band specific differential phase and other radar parameters, Journal of Applied Meteorology, 34, 404–410, 1995.

Balakrishnan, N. and Zrnić, D. S.: Estimation of rain and hail rates in mixed-phase precipitation, Journal of the Atmospheric Sciences, 47, 565–583, 1990.

Berne, A. and Krajewski, W. F.: Radar for hydrology: Unfulfilled promise or unrecognized potential?, Advances in Water Resources, 51, 357–366, 2013.

Bishop, C. M.: Pattern recognition and machine learning, Springer, 2006.

Bringi, V. and Chandrasekar, V.: Polarimetric Doppler weather radar: Principles and applications, Cambridge Univ Pr, 2001.

Bringi, V. and Hendry, A.: Technology of polarization diversity radars for meteorology, in: Radar in Meteorology, pp. 153–190, Springer, 1990.

Bringi, V. N., Huang, G.-J., Chandrasekar, V., and Gorgucci, E.: A Methodology for Estimating the Parameters of a Gamma Raindrop Size Distribution Model from Polarimetric Radar Data: Application to a Squall-Line Event from the TRMM/Brazil Campaign, Journal of Atmospheric and Oceanic Technology, 19, 633–645, 2002.

Chandrasekar, V., Lim, S., and Gorgucci, E.: Simulation of X-band rainfall observations from S-band radar data, Journal of Atmospheric and Oceanic Technology, 23, 1195–1205, 2006.

Chandrasekar, V., Wang, Y., and Chen, H.: The CASA quantitative precipitation estimation system: a five year validation study, Natural Hazards and Earth System Sciences, 12, 2811–2820, 2012.

Chandrasekar, V., Chen, H., and Philips, B.: Principles of High-Resolution Radar Network for Hazard Mitigation and Disaster Management in an Urban Environment, Journal of the Meteorological Society of Japan. Ser. II, advpub, 2018.

Chen, H. and Chandrasekar, V.: The quantitative precipitation estimation system for Dallas–Fort Worth (DFW) urban remote sensing network, Journal of Hydrology, 531, 259–271, 2015.

Chen, H., Chandrasekar, V., and Bechini, R.: An improved dual-polarization radar rainfall algorithm (DROPS2. 0): Application in NASA IFloodS field campaign, Journal of Hydrometeorology, 18, 917–937, 2017a.

Chen, H., Chandrasekar, V., and Bechini, R.: An Improved Dual-Polarization Radar Rainfall Algorithm (DROPS2.0): Application in NASA IFloodS Field Campaign, Journal of Hydrometeorology, 18, 917–937, 2017b.

Chen, H., Lim, S., Chandrasekar, V., and Jang, B.-J.: Urban Hydrological Applications of Dual-Polarization X-Band Radar: Case Study in Korea, Journal of Hydrologic Engineering, 22, E5016 001, 2017c.

Cifelli, R., Chandrasekar, V., Lim, S., Kennedy, P. C., Wang, Y., and Rutledge, S. A.: A new dual-polarization radar rainfall algorithm: Application in Colorado precipitation events, Journal of Atmospheric and Oceanic Technology, 28, 352–364, 2011.

Cifelli, R., Chandrasekar, V., Chen, H., and Johnson, L. E.: High resolution radar quantitative precipitation estimation in the San Francisco Bay area: Rainfall monitoring for the urban environment, Journal of the Meteorological Society of Japan. Ser. II, 96, 141–155, 2018.

Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B, 39, 1–38, 1977.

Giangrande, S. E., McGraw, R., and Lei, L.: An application of linear programming to polarimetric radar differential phase processing, Journal of Atmospheric and Oceanic Technology, 30, 1716–1729, 2013.

Giangrande, S. E., Collis, S., Theisen, A. K., and Tokay, A.: Precipitation Estimation from the ARM Distributed Radar Network during the MC3E Campaign, Journal of Applied Meteorology and Climatology, 53, 2130–2147, 2014.

Gorgucci, E., Scarchilli, G., and Chandrasekar, V.: Specific Differential Phase Estimation in the Presence of Nonuniform Rainfall Medium along the Path, Journal of Atmospheric and Oceanic Technology, 16, 1690–1697, 1999.

Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning., Springer Series in Statistics, Springer, New York, 2nd edn., 2009.

Helmus, J. and Collis, S.: The Python ARM Radar Toolkit (Py-ART), a library for working with weather radar data in the Python programming language, Journal of Open Research Software, 4, 2016.

Hubbart, J., Kellner, E., Hooper, L., Lupo, A., Market, P., Guinan, P., Stephan, K., Fox, N., and Svoma, B.: Localized climate and surface energy flux alterations across an urban gradient in the central US, Energies, 7, 1770–1791, 2014.

Hubbart, J. A. and Zell, C.: Considering streamflow trend analyses uncertainty in urbanizing watersheds: a baseflow case study in the central United States, Earth Interactions, 17, 1–28, 2013.

Hubbert, J. and Bringi, V. N.: An Iterative Filtering Technique for the Analysis of Copolar Differential Phase and Dual-Frequency Radar Measurements, Journal of Atmospheric and Oceanic Technology, 12, 643–648, 1995.

Hubbert, J., Chandrasekar, V., Bringi, V. N., and Meischner, P.: Processing and Interpretation of Coherent Dual-Polarized Radar Measurements, Journal of Atmospheric and Oceanic Technology, 10, 155–164, 1993.

Hubbert, J. C., Dixon, M., Ellis, S. M., and Meymaris, G.: Weather Radar Ground Clutter. Part I: Identification, Modeling, and Simulation, Journal of Atmospheric and Oceanic Technology, 26, 1165–1180, 2009.

Kalogiros, J., Anagnostou, M. N., Anagnostou, E. N., Montopoli, M., Picciotti, E., and Marzano, F. S.: Evaluation of a new polarimetric algorithm for rain-path attenuation correction of X-band radar observations against disdrometer, IEEE Transactions on Geoscience and Remote Sensing, 52, 1369–1380, 2014.

Li, Z., Zhang, Y., and Giangrande, S. E.: Rainfall-Rate Estimation Using Gaussian Mixture Parameter Estimator: Training and Validation, Journal of Atmospheric and Oceanic Technology, 29, 731–744, 2012.

Lim, S., Chandrasekar, V., and Bringi, V. N.: Hydrometeor classification system using dual-polarization radar measurements: Model improvements and in situ verification, Geoscience and Remote Sensing, IEEE Transactions on, 43, 792–801, 2005.

Lim, S., Cifelli, R., Chandrasekar, V., and Matrosov, S.: Precipitation classification and quantification using X-band dual-polarization weather radar: Application in the Hydrometeorology Testbed, Journal of Atmospheric and Oceanic Technology, 30, 2108–2120, 2013.

Liu, L., Bringi, V., Caylor, I., and Chandrasekar, V.: Intercomparison of multiparameter radar signatures from Florida storms, in: Preprints, 26th Int. Conf. on Radar Meteorology, Norman, OK, Amer. Meteor. Soc, pp. 733–735, 1993.

Marzano, F. S., Botta, G., and Montopoli, M.: Iterative Bayesian Retrieval of Hydrometeor Content From X-Band Polarimetric Weather Radar, Geoscience and Remote Sensing, IEEE Transactions on, 48, 3059–3074, 2010.

Matrosov, S. Y., Clark, K. A., Martner, B. E., and Tokay, A.: X-band polarimetric radar measurements of rainfall, Journal of Applied Meteorology, 41, 941–952, 2002.

Matrosov, S. Y., Cifelli, R., Kennedy, P. C., Nesbitt, S. W., Rutledge, S. A., Bringi, V., and Martner, B. E.: A comparative study of rainfall retrievals based on specific differential phase shifts at X-and S-band radar frequencies, Journal of Atmospheric and Oceanic Technology, 23, 952–963, 2006.

May, P. T. and Strauch, R. G.: Reducing the Effect of Ground Clutter on Wind Profiler Velocity Measurements, Journal of Atmospheric and Oceanic Technology, 15, 579–586, 1998.

May, P. T., Keenan, T. D., Zrnić, D. S., Carey, L. D., and Rutledge, S. A.: Polarimetric Radar Measurements of Tropical Rain at a 5–cm Wavelength, Journal of Applied Meteorology, 38, 750–765, 1999a.

May, P. T., Keenan, T. D., Zrnić, D. S., Carey, L. D., and Rutledge, S. A.: Polarimetric Radar Measurements of Tropical Rain at a 5-cm Wavelength, Journal of Applied Meteorology, 38, 750–765, 1999b.

McLachlan, G. and Peel, D.: Finite mixture models, vol. 299, JohnWiley & Sons, New York, 2000.

Otto, T. and Russchenberg, H. W.: Estimation of specific differential phase and differential backscatter phase from polarimetric weather radar measurements of rain, IEEE Geoscience and Remote Sensing Letters, 8, 988–992, 2011.

Oue, M., Galletti, M., Verlinde, J., Ryzhkov, A., and Lu, Y.: Use of X-band differential reflectivity measurements to study shallow Arctic mixed-phase clouds, Journal of Applied Meteorology and Climatology, 55, 403–424, 2016.

Park, H. S., Ryzhkov, A. V., Zrnić, D. S., and Kim, K.-E.: The hydrometeor classification algorithm for the polarimetric WSR-88D: Description and application to an MCS, Weather and Forecasting, 24, 730–748, 2009.

Park, S., Bringi, V., Chandrasekar, V., Maki, M., and Iwanami, K.: Correction of radar reflectivity and differential reflectivity for rain attenuation at X band. Part I: Theoretical and empirical basis, Journal of Atmospheric and Oceanic Technology, 22, 1621–1632, 2005.

Petersen, K. B. and Pedersen, M. S.: The matrix cookbook (version: November 15, 2012), 2012.

Reinoso-Rondinel, R., Unal, C., and Russchenberg, H.: Adaptive and high-resolution estimation of specific differential phase for polarimetric X-band weather radars, Journal of Atmospheric and Oceanic Technology, 35, 555–573, 2018.

Ryzhkov, A. and Zrnić, D.: Assessment of rainfall measurement that uses specific differential phase, Journal of applied meteorology, 35, 2080–2090, 1996.

Ryzhkov, A. V. and Zrnić, D. S.: Comparison of Dual-Polarization Radar Estimators of Rain, Journal of Atmospheric and Oceanic Technology, 12, 249–256, 1995.

Ryzhkov, A. V., Schuur, T. J., Burgess, D. W., Heinselman, P. L., Giangrande, S. E., and Zrnic, D. S.: The Joint Polarization Experiment: Polarimetric Rainfall Measurements and Hydrometeor Classification, Bulletin of the American Meteorological Society, 86, 809–824, 2005.

Sachidananda, M. and Zrnić, D. S.: Differential propagation phase shift and rainfall rate estimation, Radio Sci., 21, 235–247, 1986.

Schneebeli, M., Grazioli, J., and Berne, A.: Improved estimation of the specific differential phase shift using a compilation of Kalman filter ensembles, IEEE Transactions on Geoscience and Remote Sensing, 52, 5137–5149, 2014.

Seliga, T. A. and Bringi, V. N.: Differential reflectivity and differential phase shift: Applications in radar meteorology, Radio Science, 13, 271–275, 1978.

Simpson, M. J. and Fox, N. I.: X-band dual-polarized radar quantitative precipitation estimate analyses in the Midwestern United States, Atmos. Meas. Tech. Discuss., https://doi.org/https://doi.org/10.5194/amt-2017-439, 2017.

Spanos, A.: Probability theory and statistical inference: econometric modeling with observational data, Cambridge University Press, 1999.

Sung, H.: Gaussian mixture regression and classification, Thesis, Rice University, 2004.

Trömel, S., Kumjian, M. R., Ryzhkov, A. V., Simmer, C., and Diederich, M.: Backscatter differential phase–Estimation and variability, Journal of Applied Meteorology and Climatology, 52, 2529–2548, 2013.

Veneziano, D. and Villani, P.: Identification of rain cells from radar and stochastic modelling of space-time rainfall, Meccanica, 31, 27–42, 1996.

Vulpiani, G., Montopoli, M., Passeri, L. D., Gioia, A. G., Giordano, P., and Marzano, F. S.: On the Use of Dual-Polarized C-Band Radar for Operational Rainfall Retrieval in Mountainous Areas, Journal of Applied Meteorology and Climatology, 51, 405–425, 2012.

Wang, Y. and Chandrasekar, V.: Algorithm for estimation of the specific differential phase, Journal of Atmospheric and Oceanic Technology, 26, 2565–2578, 2009.

Wang, Y. and Chandrasekar, V.: Quantitative precipitation estimation in the CASA X-band dual-polarization radar network, Journal of Atmospheric and Oceanic Technology, 27, 1665–1676, 2010.

Wen, G., Protat, A., May, P. T., Wang, X., and Moran, W.: A Cluster-Based Method for Hydrometeor Classification Using Polarimetric Variables. Part I: Interpretation and Analysis, Journal of Atmospheric and Oceanic Technology, 32, 1320–1340, 2015.

Wen, G., Oue, M., Protat, A., Verlinde, J., and Xiao, H.: Ice particle type identification for shallow Arctic mixed-phase clouds using X-band polarimetric radar, Atmospheric Research, 182, 114–131, 2016a.

Wen, G., Protat, A., May, P. T., Moran, W., and Dixon, M.: A Cluster-Based Method for Hydrometeor Classification Using Polarimetric Variables. Part II: Classification, Journal of Atmospheric and Oceanic Technology, 33, 45–60, 2016b.

Wen, G., Protat, A., and Xiao, H.: An Objective Prototype-Based Method for Dual-Polarization Radar Clutter Identification, Atmosphere, 8, 72, http://www.mdpi.com/2073-4433/8/4/72, 2017.

Williams, C. R., Bringi, V., Carey, L. D., Chandrasekar, V., Gatlin, P. N., Haddad, Z. S., Meneghini, R., Joseph Munchak, S., Nesbitt, S. W., and Petersen, W. A.: Describing the shape of raindrop size distributions using uncorrelated raindrop mass spectrum parameters, Journal of Applied Meteorology and Climatology, 53, 1282–1296, 2014.

Zrnić, D. S. and Ryzhkov, A.: Advantages of Rain Measurements Using Specific Differential Phase, Journal of Atmospheric and Oceanic Technology, 13, 454–464, 1996.

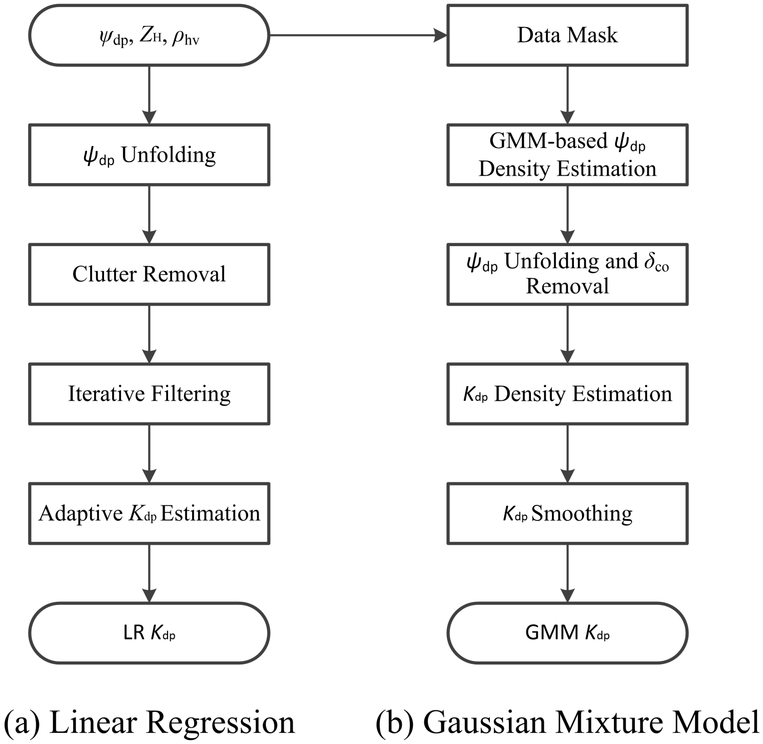
ψ_{dp}, Z_H, ρ_{hv}
ψ_{dp} Unfolding
Clutter Removal
Iterative Filtering
Adaptive K_{dp} Estimation
LR K_{dp}
Data Mask
GMM-based ψ_{dp} Density Estimation
ψ_{dp} Unfolding and δ_{co} Removal
K_{dp} Density Estimation
K_{dp} Smoothing
GMM K_{dp}
(a) Linear Regression
(b) Gaussian Mixture Model

Figure 1. Flowcharts of K_{dp} estimation algorithms used in the MZZU radar: (a) linear regression model, and (b) Gaussian mixture method.

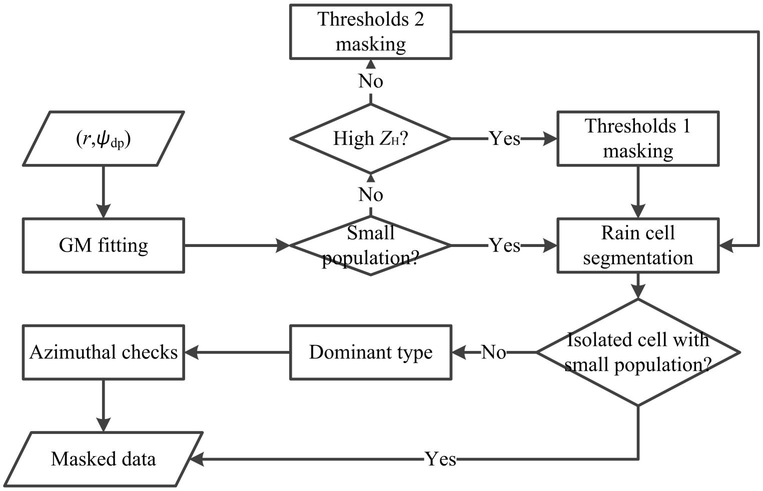

Figure 2. Flowchart of data masking.

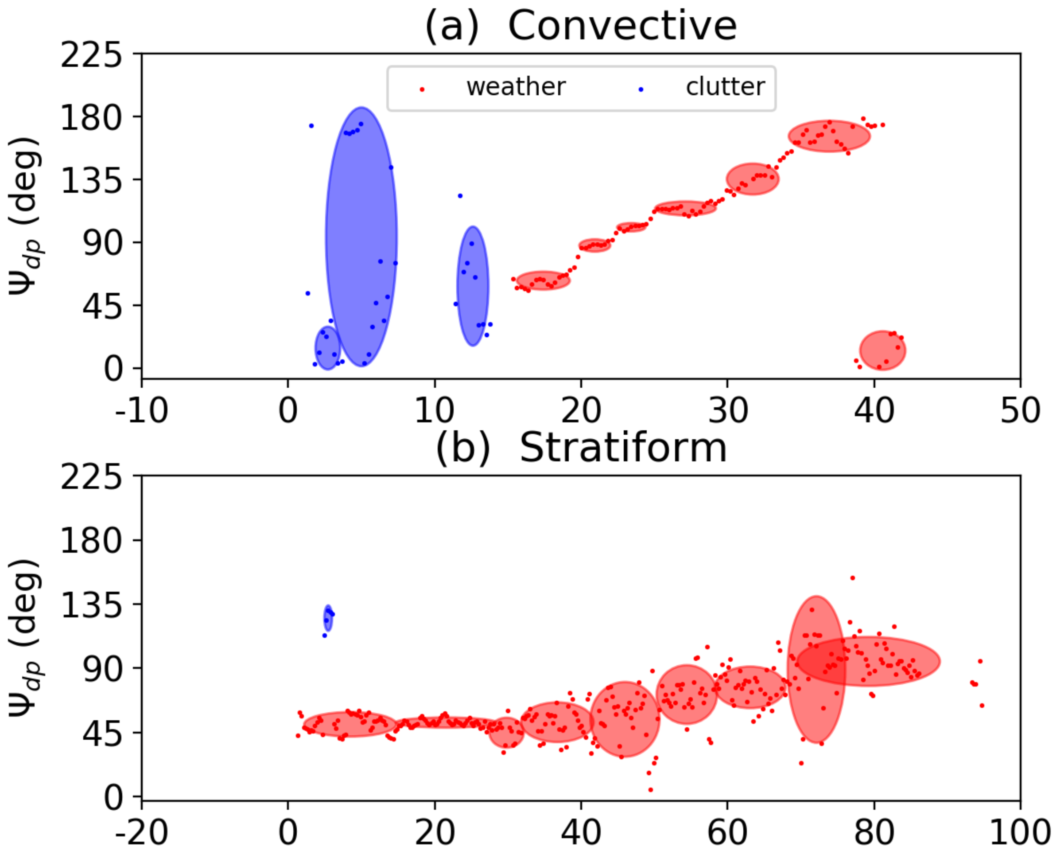

Figure 3. Examples of data masking: (a) a convective case (azimuth 252°), and (b) a stratiform case (azimuth 1°). The blue points and ellipses represent the clutter data and clusters, respectively, while the red color corresponds to the weather echoes.

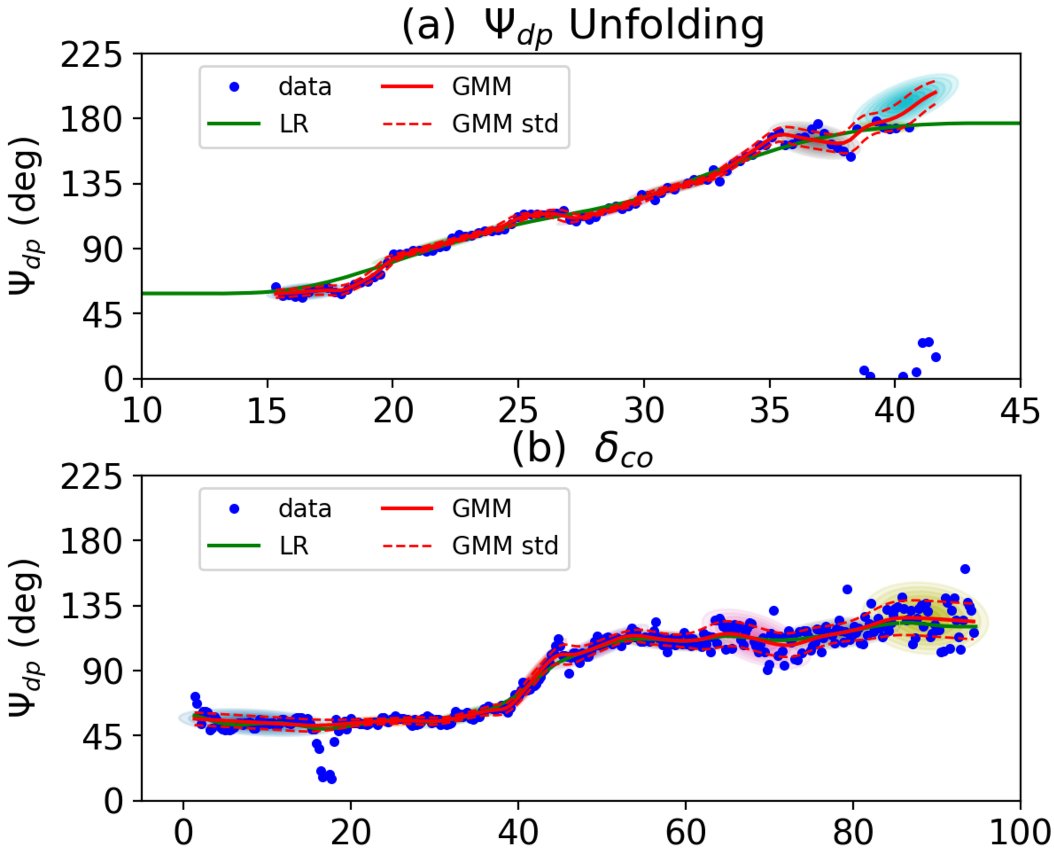

Figure 4. Examples of Ψ_{dp} density estimation: (a) a Ψ_{dp} unfolding case, and (b) a δ_{co} case. The blue points are the Ψ_{dp} data, the green curve represents the Ψ_{dp} profile obtained by the linear regression model (LR), and the red curve indicates the Ψ_{dp} profile produced by the Gaussian mixture method (GMM). The dash lines are the standard deviations, while the colored ellipses show the components of the Gaussian mixture.

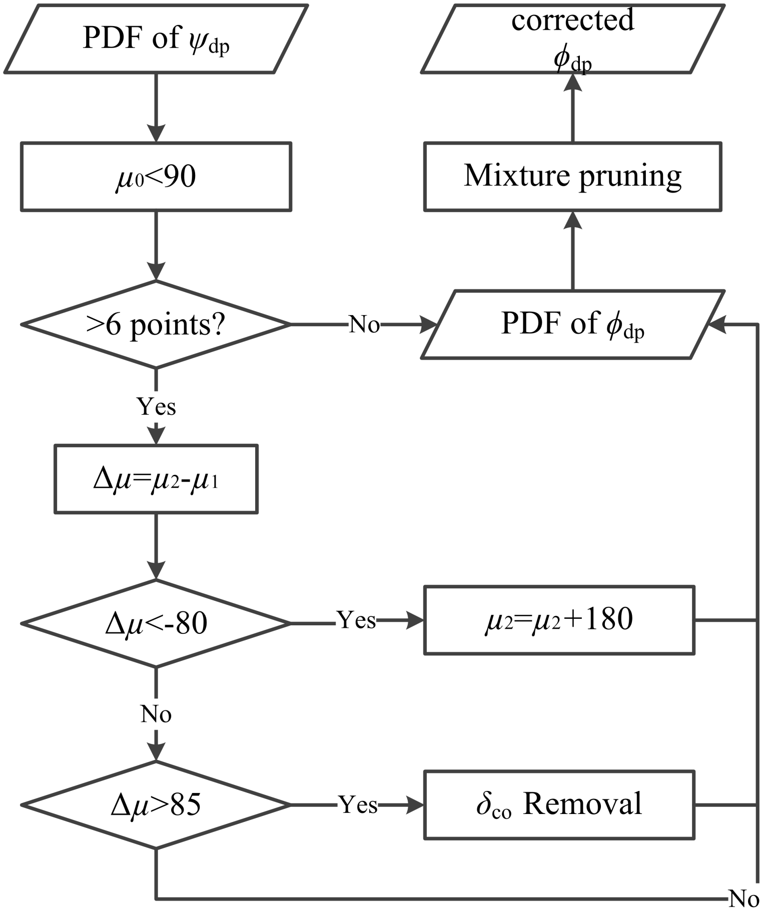

Figure 5. Flowchart of the Ψ_{dp} unfolding and the δ_{co} elimination. The μ_0 is the mean of the first density ellipse. The μ_1 and μ_2 are the means of the two consective density ellipses along the range. The μ_1 is the mean of the former one, and the μ_2 is the mean of the latter one.

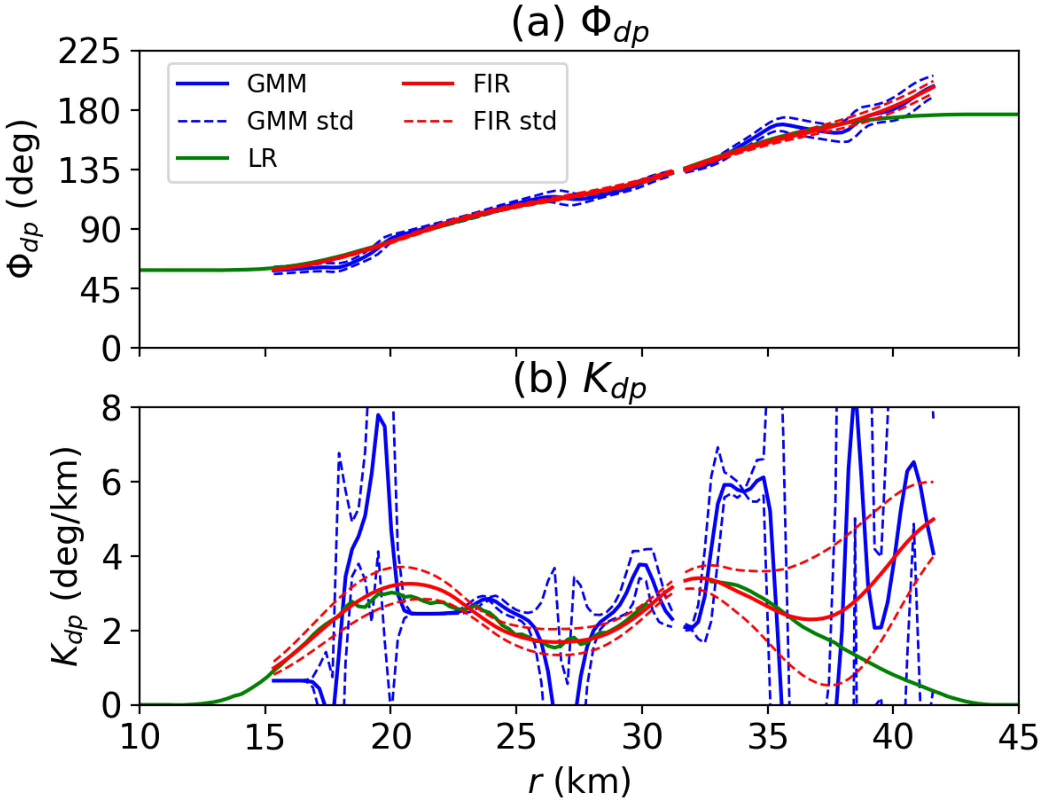

Figure 6. Examples of K_{dp} estimation: (a) Φ_{dp}, and (b) K_{dp}. The blue curves are the Φ_{dp} and K_{dp} estimates obtained by the Gaussian mixture method (GMM), the green curves represent the estimates derived from the linear regression model (LR), and the red curves indicate the reconstructed Φ_{dp} and smoothed K_{dp} profiles (FIR). The dash lines are the standard deviations.

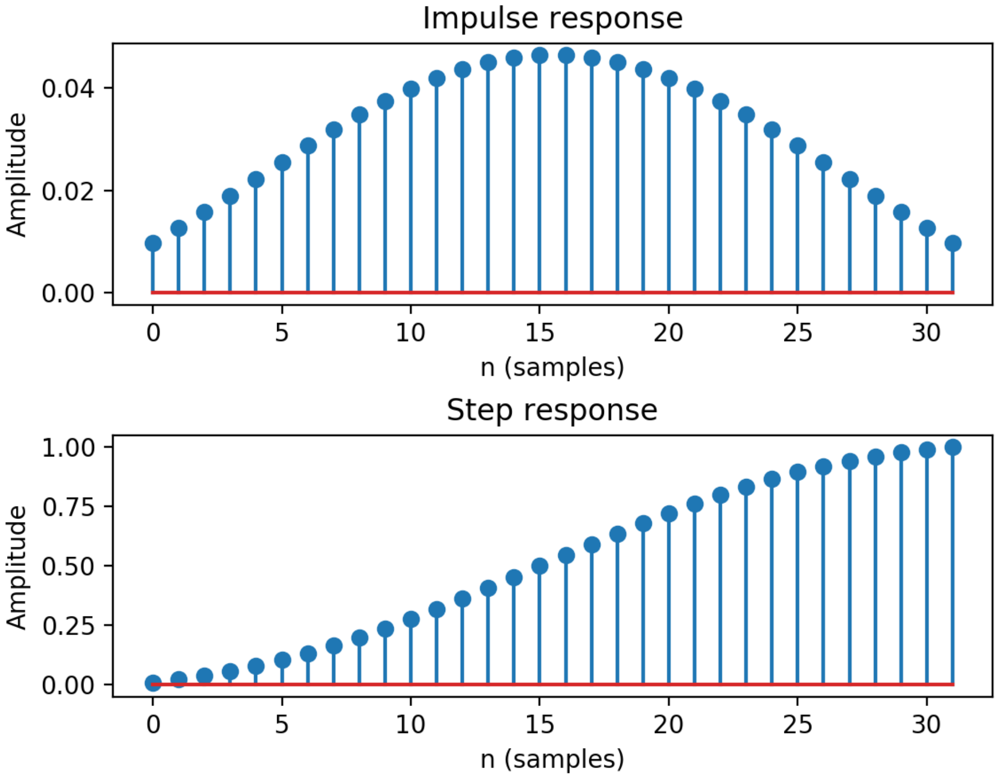

Figure 7. Responses of finite impulse filter: (a) Impulse response, and (b) step response.

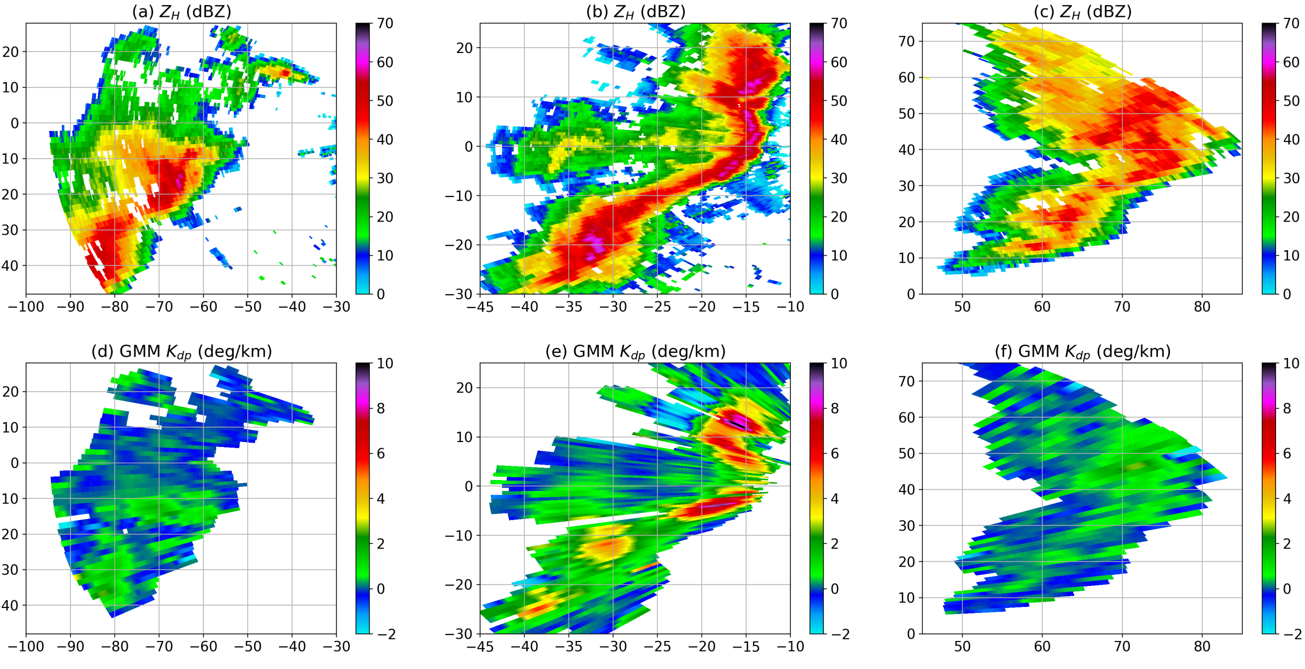

Figure 8. A case study for GMM: (a) raw Z_H at the development stage (03:04 UTC), (b) raw Z_H at the mature stage (03:39 UTC), and (c) rawZ_H at the dissipation stage (04:41 UTC). (d), (e) and (f) are the same as (a), (b) and (c), respectively, but for K_{dp}. The data were collected at a elevation of 0.85° by the MZZU radar between 0304 and 0441 UTC on 24 March 2016.

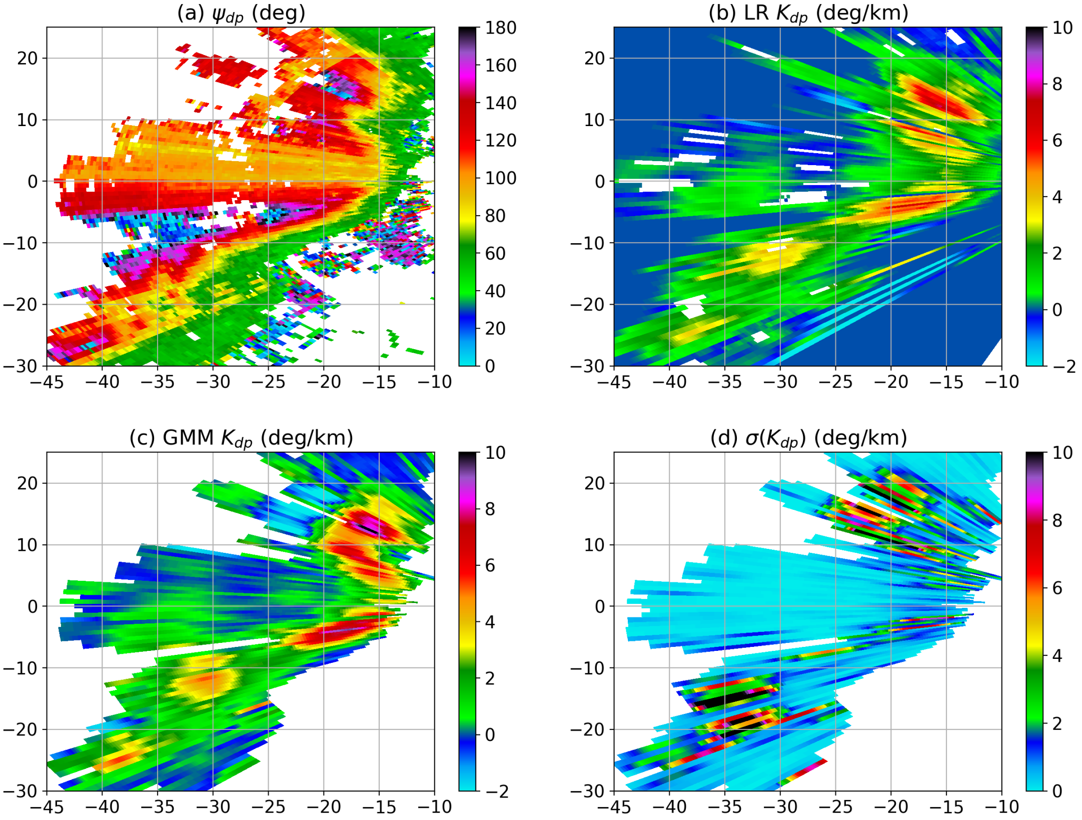

Figure 9. K_{dp} estimation for the mature stage: (a) raw Ψ_{dp}, (b) LR-based K_{dp}, (c) GMM-based K_{dp}, and (d) GMM-based $\sigma(K_{dp})$. The data were collected at 0339 UTC on 24 March 2016.

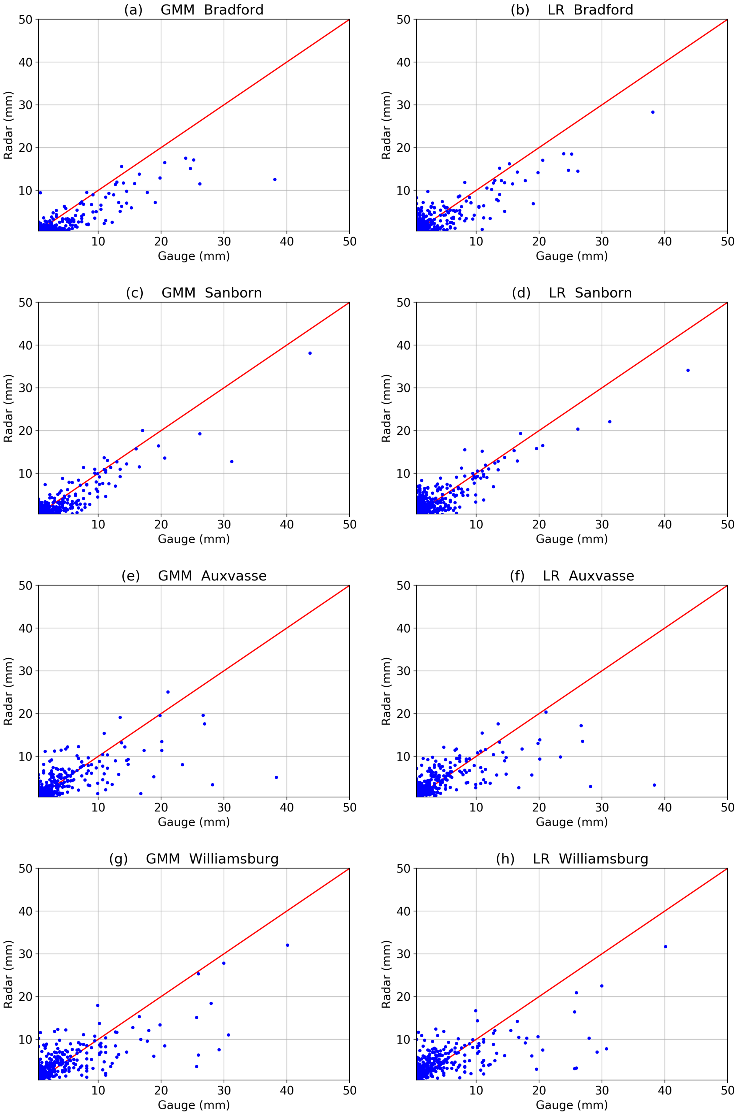

Figure 10. Comparison between hourly radar and gauge data derived from GMM K_{dp} and LR K_{dp}. (a) GMM Bradford, (b) LR Bradford, (c) GMM Sanborn, (d) LR Sanborn, (e) GMM Auxvasse, (f) LR Auxvasse, (g) GMM Williamsburg and (h) LR Williamsburg. The data were collected between 1 April 2016 and 2 June 2018.

Table 1. Characteristics of hourly rain gauge data at Bradford, Sanborn, Auxvasse, and Williamsburg between April 2016 and June 2018. Mean: mean values, Std: standard deviation, Max: maximum values, Total: sums of rain amounts, and Duration: sum of rainfall time.

Sites	Mean (mm)	Std (mm)	Max (mm)	Total (mm)	Duration (h)
Bradford	2.1	3.5	38.1	2224.9	1080
Sanborn	2.0	3.3	43.7	2181.4	1082
Auxvasse	2.0	3.3	38.4	2284.3	1144
Williamsburg	2.1	3.7	40.1	2495.9	1191

Table 2. Statistics for the comparison between radar and gauge. RMSE: root mean squared error, NB: normalized bias, ρ_{RG}: Pearson correlation coefficient; LR: linear regression model, GMM: Gaussian mixture method.

Algorithm	Sites	RMSE (mm)	NB	ρ_{RG}
LR	Bradford	2.87	-0.28	0.84
	Sanborn	1.97	-0.08	0.89
	Auxvasse	3.25	0.21	0.67
	Williamsburg	3.55	0.20	0.70
GMM	Bradford	2.71	-0.31	0.84
	Sanborn	2.06	-0.13	0.88
	Auxvasse	3.14	0.04	0.69
	Williamsburg	3.20	0.14	0.76