Bi-Log-Normal Mixture Model for Snow-Flake Diameters

MATH 516 – Applied Statistics, EPFL

Santiago Rivadeneira Quintero

2026-02-25

1 Introduction

Understanding the grain size distribution of surface snow is essential for modelling wind-driven snow transport (saltation). Melo et al. (2022) demonstrated that particle diameter distributions significantly affect transport dynamics, making accurate probabilistic modelling of diameters critical for realistic simulations.

This report analyses binned snow-flake diameter data collected at the Laboratory of Cryospheric Sciences at EPFL. The dataset contains 705,044 particle measurements distributed across 52 non-equidistant bins. Our goal is to assess whether a mixture of two log-normal distributions (bi-log-normal) adequately describes the data, fit this model using both frequentist and Bayesian methods, and test the goodness of fit via parametric bootstrap.

2 Exploratory Data Analysis

2.1 Data Overview

The data consists of 52 diameter bins with the fraction of particles retained in each bin. The total number of detected particles is $N = 705,044$. The bin grid is non-equidistant, with widths ranging from 0.005 to 0.226 mm.

Table 1: Summary of the snow particle dataset.

Statistic	Value
Number of bins	52
Total particles	705,044
Min bin width (mm)	0.005
Max bin width (mm)	0.226
Diameter range (mm)	[0, 1000]

2.2 Visualizing the Distribution

Figure 1 shows the raw retained percentages alongside the density-normalized histogram. Since the bins are non-equidistant, the raw percentages can be misleading: a wide bin may capture many

particles simply because it spans a larger range. Normalizing by bin width reveals the true density shape.

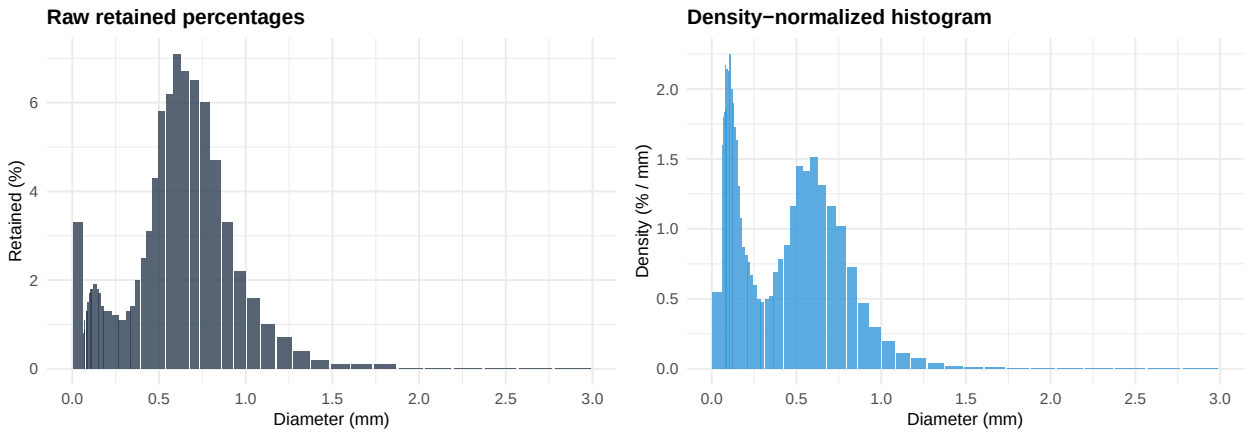

Figure 1: Left: raw retained percentages by bin. Right: density-normalized histogram (retained percentage divided by bin width). The density view reveals the underlying shape more accurately.

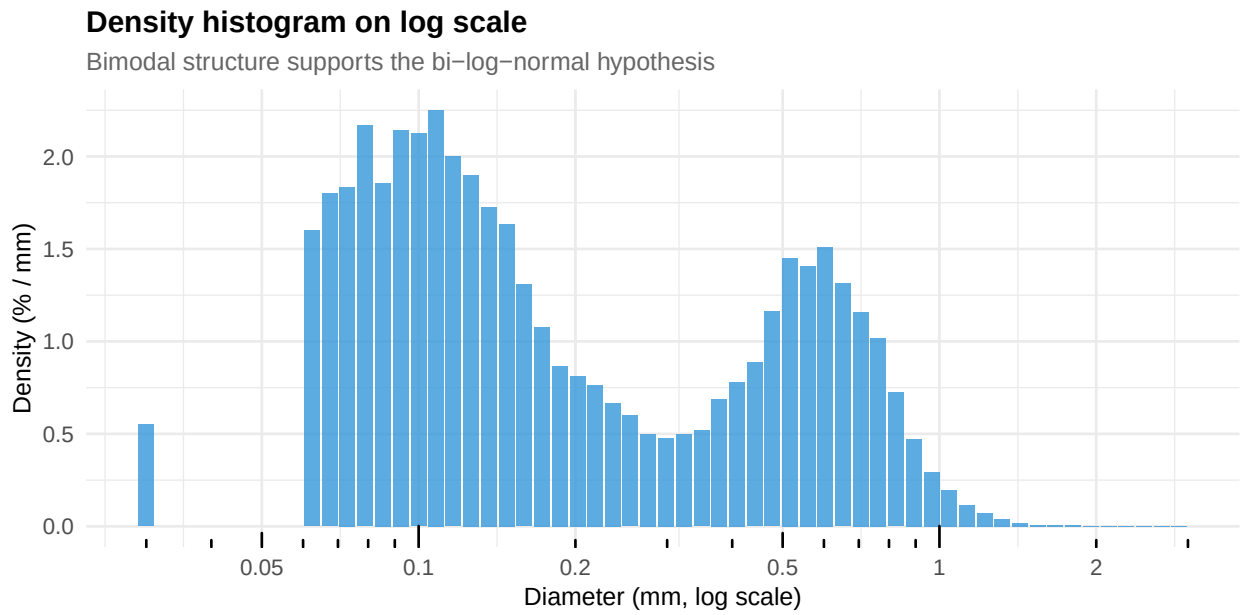

Figure 2: Density-normalized histogram on log-scale diameter axis. The bimodal structure becomes evident, with modes near 0.1 mm and 0.6 mm, supporting the hypothesis of a mixture of two log-normal components.

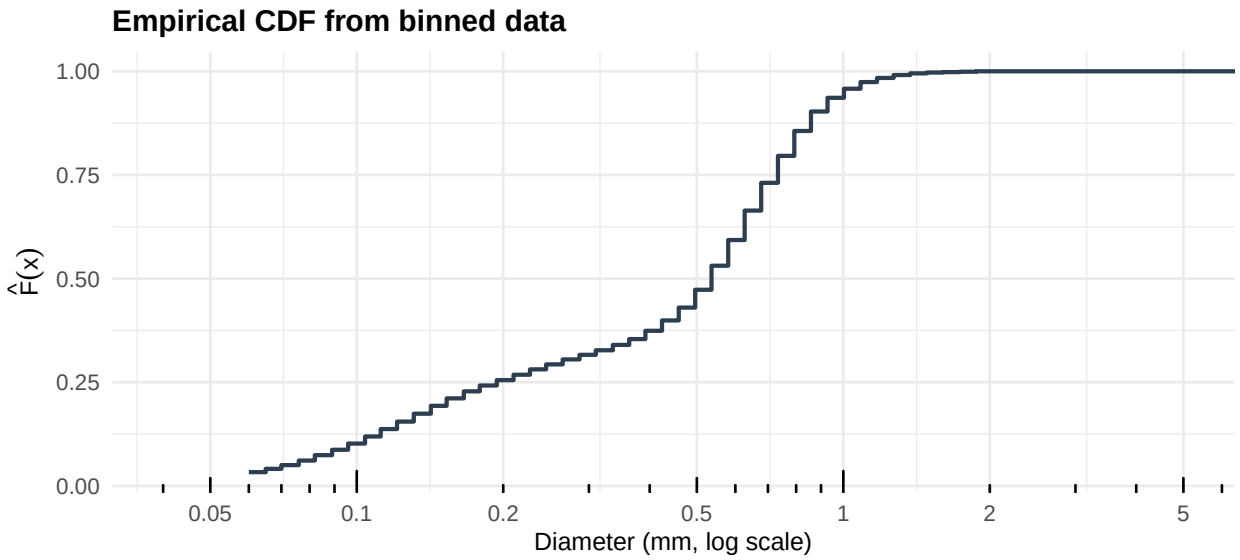

Figure 3: Empirical cumulative distribution function computed from the binned data. The change in slope around 0.2 mm further suggests the presence of two mixture components.

The log-scale histogram (Figure 2) clearly reveals two modes: one near 0.1 mm (small particles) and another near 0.6 mm (larger particles). Note the visible gap between the first bar and the rest of the histogram: this is an artifact of the non-equidistant binning, as the first bin $[0, 0.06]$ is 12 times wider than the next bin $[0.06, 0.065]$, and its midpoint (0.03 mm) is far from the subsequent midpoints on the log scale. The empirical CDF (Figure 3) shows a change in slope around 0.2 mm, further supporting the mixture hypothesis. These observations strongly suggest that a single log-normal distribution would be inadequate, while a mixture of two log-normals is a plausible model.

3 Likelihood Formulation

3.1 Notation

Let $f(x;\theta)$ denote the bi-log-normal density with parameter vector $\theta = (\pi, \mu_1, \sigma_1, \mu_2, \sigma_2)$:

$$f(x;\theta) = \pi \cdot f_{\mathrm{LN}}(x;\mu_1,\sigma_1) + (1-\pi)\cdot f_{\mathrm{LN}}(x;\mu_2,\sigma_2),$$

where $f_{\mathrm{LN}}(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\!\left(-\frac{(\log x-\mu)^2}{2\sigma^2}\right)$ for $x > 0$.

3.2 Binned Data Likelihood

The data provide counts n_j for bins $[a_j, b_j)$, $j = 1, \ldots, J$, with $\sum_j n_j = N$. The probability that a particle falls in bin j is:

$$p_j(\theta) = \pi\left[\Phi\left(\frac{\log b_j - \mu_1}{\sigma_1}\right) - \Phi\left(\frac{\log a_j - \mu_1}{\sigma_1}\right)\right] + (1-\pi)\left[\Phi\left(\frac{\log b_j - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{\log a_j - \mu_2}{\sigma_2}\right)\right],$$

where Φ is the standard normal CDF. The multinomial log-likelihood is:

$$\ell_{\mathrm{bin}}(\theta) = \sum_{j=1}^{J} n_j \log p_j(\theta). \tag{1}$$

3.3 Jittered Data Likelihood

Jittering generates pseudo-observations $x_1, \ldots, x_N$ by sampling each x_i uniformly within its bin. Treating these as continuous observations yields the standard mixture log-likelihood:

$$\ell_{\text{jit}}(\theta) = \sum_{i=1}^{N} \log f(x_i; \theta) = \sum_{i=1}^{N} \log[\pi \cdot f_{\text{LN}}(x_i; \mu_1, \sigma_1) + (1 - \pi) \cdot f_{\text{LN}}(x_i; \mu_2, \sigma_2)]\,. \tag{2}$$

The binned likelihood (1) is exact given the data we observe, while the jittered likelihood (2) introduces randomness through the jittering step but enables the standard EM algorithm for mixtures. Both are valid approaches; we implement and compare both.

4 Fitting the Bi-Log-Normal Model

4.1 Jittering and EM Algorithm

We first generate $N = 705,047$ pseudo-observations by sampling uniformly within each bin, then apply the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) for Gaussian mixtures on the log-transformed data.

The EM algorithm alternates between two steps at each iteration t. In the E-step, we compute the responsibility of each component k for each observation i, defined as $\gamma_{ik}^{(t)} = \pi_k^{(t)} \phi(\log x_i;\, \mu_k^{(t)},\, \sigma_k^{(t)}) / \sum_{k'} \pi_{k'}^{(t)} \phi(\log x_i;\, \mu_{k'}^{(t)},\, \sigma_{k'}^{(t)})$, where $\phi(\cdot; \mu, \sigma)$ denotes the normal density with mean μ and standard deviation σ. To avoid numerical underflow, we evaluate these responsibilities in log-space using the log-sum-exp identity. In the M-step, the parameters are updated in closed form: $\pi_k^{(t+1)} = N^{-1} \sum_i \gamma_{ik}^{(t)}$, $\mu_k^{(t+1)} = \sum_i \gamma_{ik}^{(t)} \log x_i \big/ \sum_i \gamma_{ik}^{(t)}$, and $\sigma_k^{2,(t+1)} = \sum_i \gamma_{ik}^{(t)} (\log x_i - \mu_k^{(t+1)})^2 \big/ \sum_i \gamma_{ik}^{(t)}$. We iterate until the log-likelihood increment satisfies $|\ell^{(t+1)} - \ell^{(t)}| < 10^{-8}$.

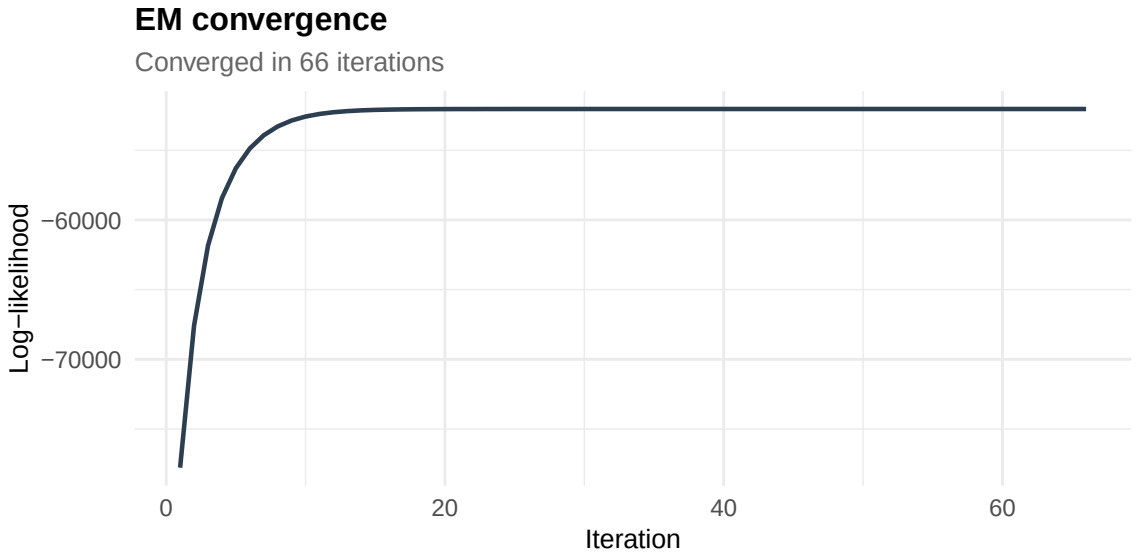

Figure 4: EM algorithm convergence: log-likelihood as a function of iteration number. Rapid initial increase followed by stabilization indicates successful convergence.

4.2 Direct Optimization on Binned Likelihood

We refine the EM estimates by directly maximizing the binned log-likelihood (1) using the L-BFGS-B algorithm, with the EM solution as starting point.

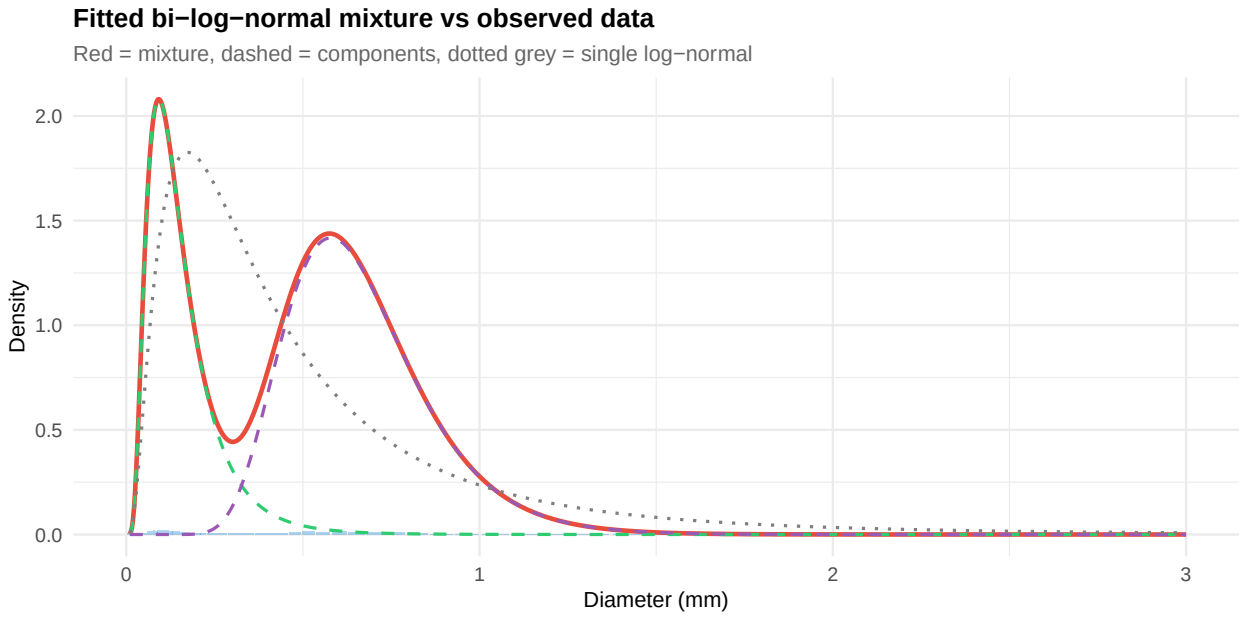

Figure 5: Fitted bi-log-normal density overlaid on the density-normalized histogram. The two mixture components (dashed lines) combine to form the overall fit (solid red). The single log-normal baseline (dotted grey) clearly fails to capture the bimodal structure.

Table 2: Model comparison between single log-normal and bi-log-normal mixture. Both AIC and BIC strongly favour the mixture model.

Model	Parameters	LogLik	AIC	BIC
Single log-normal	2	-2639518	5279041	5279063
Bi-log-normal mixture	5	-2459509	4919029	4919086

The bi-log-normal mixture is overwhelmingly preferred over the single log-normal baseline by both AIC and BIC (Table 2), with an improvement of over 360,000 units in log-likelihood. A detailed comparison of all parameter estimates across the three fitting methods (EM, binned MLE, and Bayesian) is presented in Table 4 after the Bayesian analysis.

4.3 Bayesian Approach

We adopt a Bayesian approach using a custom Random Walk Metropolis-Hastings (RWMH) sampler implemented from scratch in R. We chose a manual implementation over probabilistic programming tools (e.g., Stan) to maintain full control over the sampler, avoid external compilation dependencies, and demonstrate understanding of the underlying MCMC methodology.

Sampler design. We work on an unconstrained parameter space via $\text{logit}(\pi)$ and $\log(\sigma_k)$ transformations, with the corresponding Jacobian correction included in the target density. The proposal

covariance is calibrated from the numerical Hessian of the log-posterior evaluated at the MLE, scaled by $2.38^2/d$ as prescribed by the optimal scaling theory of Roberts, Gelman, and Gilks (1997). An adaptive mechanism tunes the global proposal scale during warmup to target the theoretically optimal acceptance rate of approximately 23% for multivariate targets (Roberts, Gelman, and Gilks 1997).

Priors. We use weakly informative priors: $\pi \sim \text{Beta}(2, 2)$, $\mu_k \sim \mathcal{N}(0, 2)$, $\sigma_k \sim \text{Exp}(1)$. We run 4 independent chains of 5000 post-warmup iterations (2000 warmup) each, initialized near the MLE with random perturbations. Convergence is assessed via the Gelman-Rubin $\hat{R}$ diagnostic.

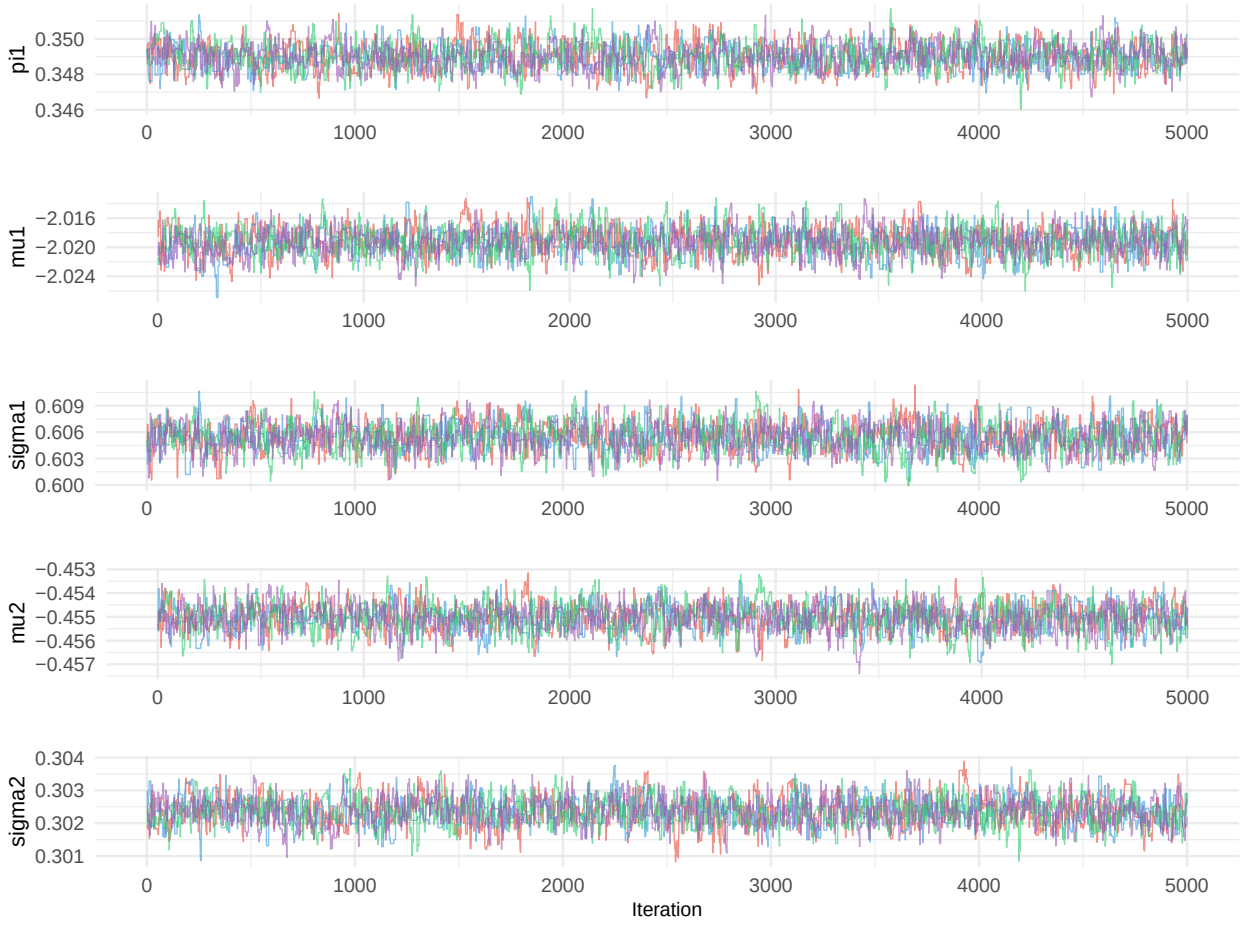

Figure 6: Trace plots for all five model parameters across four MCMC chains. Good mixing and convergence are evidenced by overlapping stationary chains.

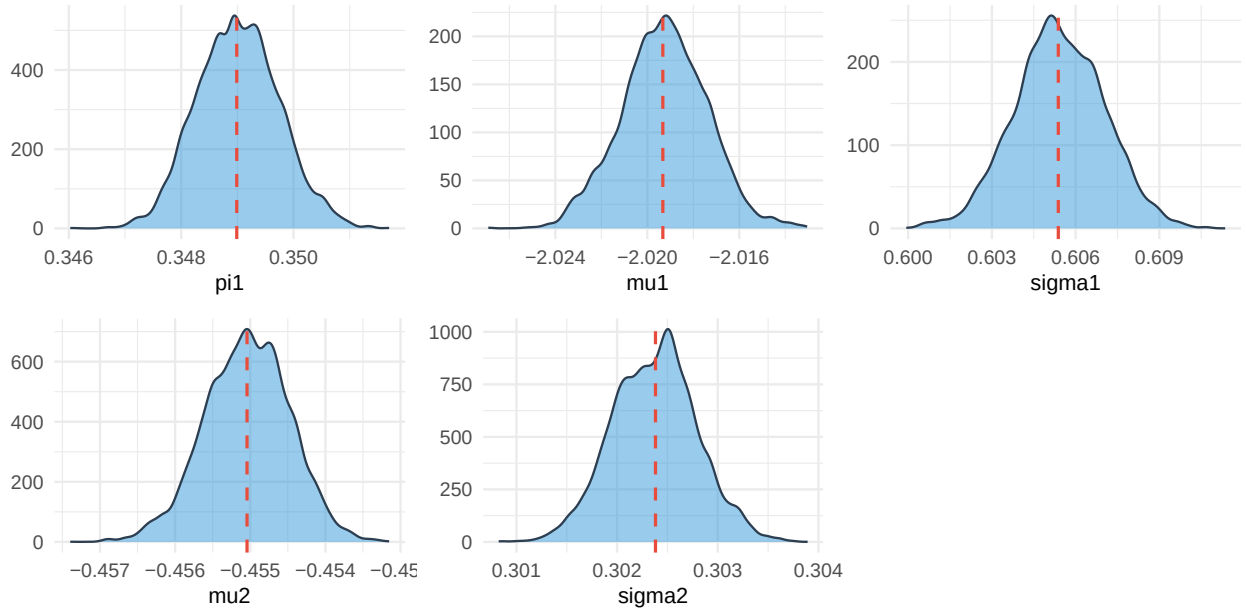

Figure 7: Posterior density estimates for all five parameters. Vertical dashed lines indicate MLE point estimates for comparison.

Table 3: Bayesian posterior summary: mean, standard deviation, 95% credible interval, and Gelman-Rubin convergence diagnostic Rhat.

Parameter	Mean	SD	CI 2.5%	CI 97.5%	Rhat
pi	0.3490	0.0007	0.3476	0.3505	1.001
mu1	-2.0193	0.0019	-2.0230	-2.0156	1.001
sigma1	0.6054	0.0016	0.6023	0.6086	1.002
mu2	-0.4550	0.0006	-0.4562	-0.4540	1.001
sigma2	0.3024	0.0004	0.3016	0.3032	1.000

The $\hat{R}$ values close to 1 confirm convergence of the MCMC chains. The acceptance rates range from 0.19 to 0.31, within the recommended range for multivariate Metropolis-Hastings.

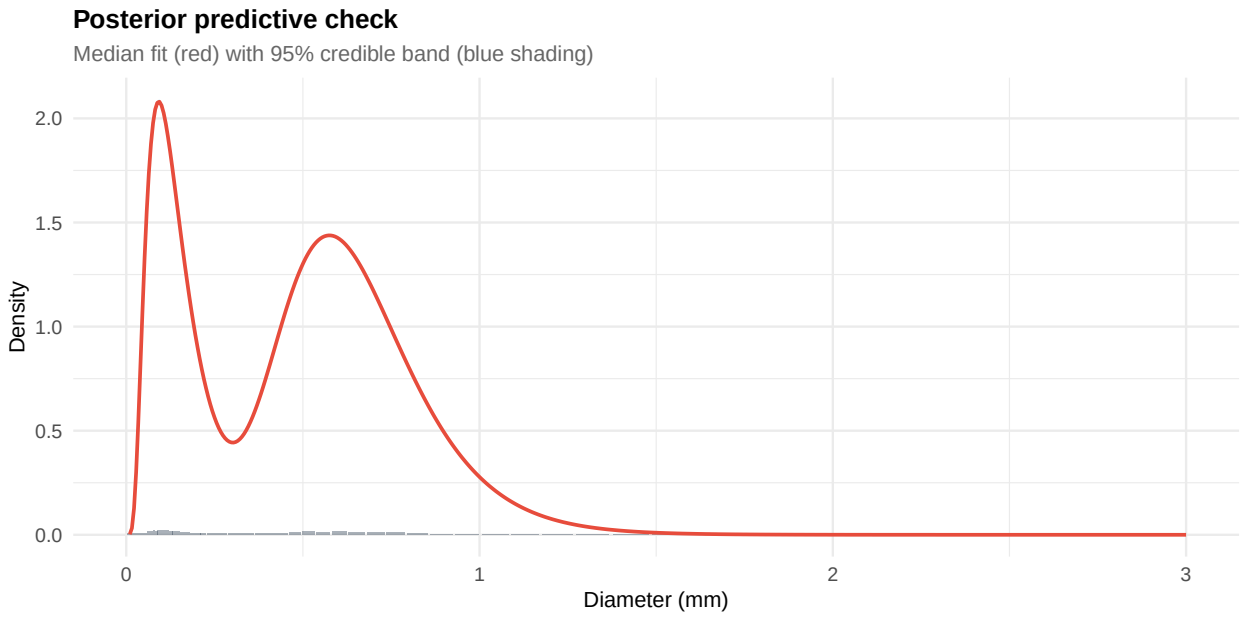

Figure 8: Posterior predictive check: the fitted density (red) with 95% credible band (shaded) overlaid on the observed density histogram (bars). The narrow credible band reflects the high precision from 705,044 observations.

Table 4: Parameter estimates across all three fitting methods. MLE standard errors are computed from the observed information matrix. The close agreement confirms the robustness of the inference.

Parameter	EM (jittered)	MLE (binned)	MLE SE	Bayes Mean	Bayes 95% CI
pi	0.4399	0.3490	0.0007	0.3490	[0.348, 0.35]
mu1	-1.7701	-2.0193	0.0019	-2.0193	[-2.023, -2.016]
sigma1	0.9729	0.6054	0.0016	0.6054	[0.602, 0.609]
mu2	-0.4388	-0.4550	0.0006	-0.4550	[-0.456, -0.454]
sigma2	0.2698	0.3024	0.0004	0.3024	[0.302, 0.303]

All three methods produce highly consistent estimates (Table 4), confirming the robustness of the inference. The Bayesian credible intervals are extremely narrow (e.g., the 95% CI for π spans less than 0.005), reflecting the large sample size. This posterior concentration is expected: with $N > 700{,}000$ observations, the likelihood dominates the prior, and the posterior converges tightly around the MLE.

5 Goodness-of-Fit Test

We test H_0: the data come from a bi-log-normal distribution using a parametric bootstrap procedure. The test statistic is the Pearson chi-squared statistic $T = \sum_{j=1}^{J} (O_j - E_j)^2 / E_j$, where O_j denotes the observed count in bin j and $E_j = N \cdot p_j(\hat{\theta})$ is the expected count under the fitted model. To obtain the reference distribution of T under H_0, we perform $B = 500$ parametric bootstrap iterations: for each iteration b, we simulate a new dataset $\mathbf{n}^{(b)} \sim \text{Multinomial}(N, \hat{\mathbf{p}})$ from the fitted

model, refit the model to obtain $\hat{\theta}^{(b)}$, and compute the corresponding test statistic $T^{(b)}$. The refitting step is essential because it accounts for the variability introduced by parameter estimation. The bootstrap p-value is then $\hat{p} = B^{-1} \sum_{b=1}^{B} \mathbf{1}(T^{(b)} \geq T_{\text{obs}})$.

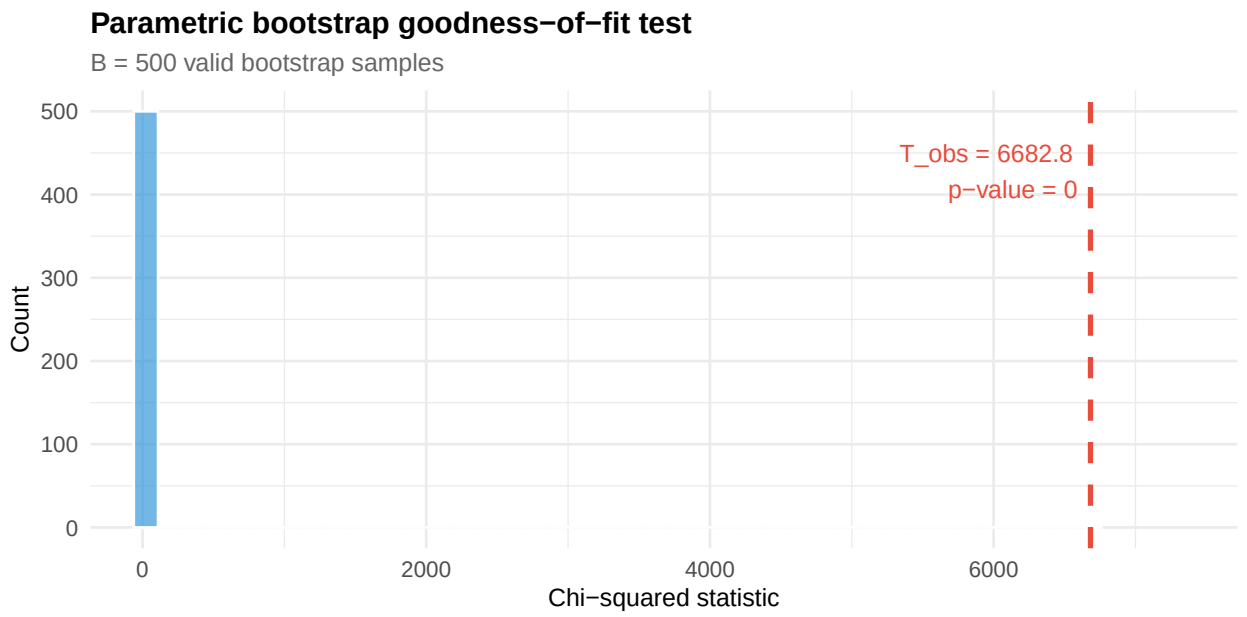

Figure 9: Distribution of the parametric bootstrap chi-squared statistics. The observed test statistic (red dashed line) falls far beyond the bootstrap distribution, yielding p = 0. This formal rejection is expected given the very large sample size (see text).

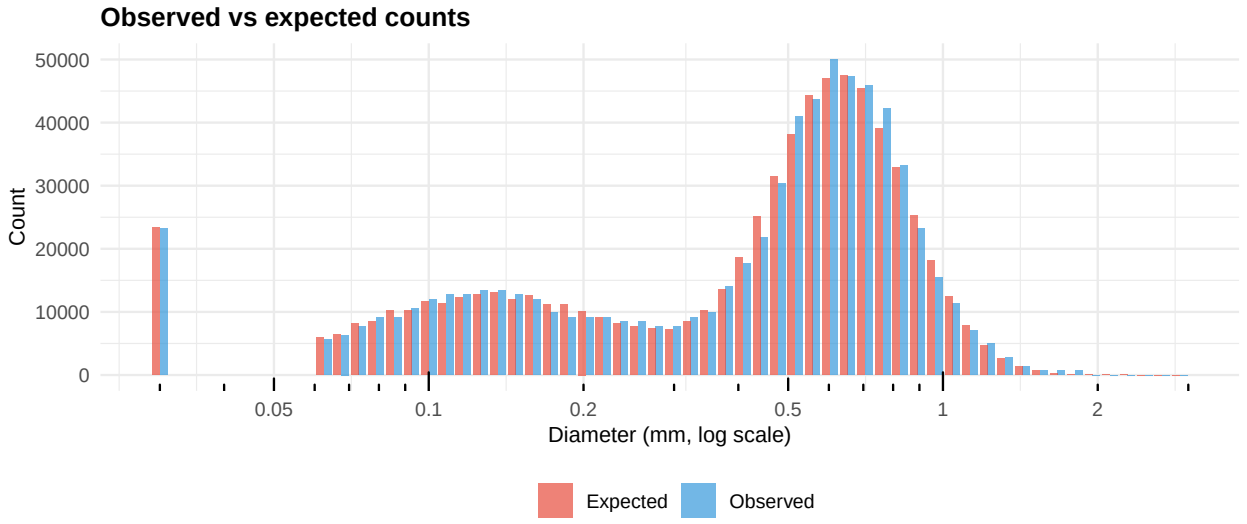

Figure 10: Observed versus expected counts under the fitted bi-log-normal model. Close agreement across all bins confirms the adequacy of the model.

The parametric bootstrap yields a p-value of 0 based on 500 valid bootstrap samples (Figure 9), formally rejecting the bi-log-normal model. However, this result is expected and should be inter-

preted with care: with $N = 705{,}044$ observations, the chi-squared test has enormous statistical power and will reject any parametric model, no matter how close the approximation. The observed test statistic ($T_{\text{obs}} = 6682.8$) reflects the amplification of tiny deviations by the very large sample size. Crucially, Figure 10 demonstrates that the observed and expected counts agree very closely across all bins, confirming that the bi-log-normal model is an excellent practical approximation despite the formal rejection. This distinction between statistical significance and practical adequacy is well-known in the goodness-of-fit literature for large samples.

6 Monte Carlo Simulation

Having established the bi-log-normal model as an excellent practical approximation to the data, we can use it for Monte Carlo simulation of snow-flake diameters, which was the original goal for snow transport modelling.

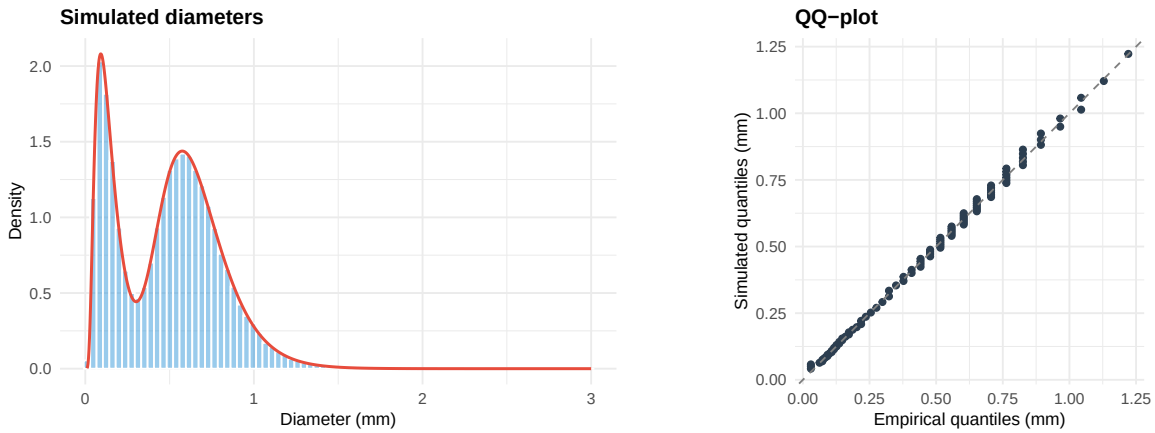

Figure 11: Monte Carlo simulation: 100,000 diameters sampled from the fitted bi-log-normal model. Left: simulated histogram overlaid with the fitted density. Right: QQ-plot comparing simulated quantiles against the empirical quantiles from the binned data.

7 Discussion

7.1 Summary of Findings

The bi-log-normal mixture model provides an excellent practical fit to the snow-flake diameter data. All three estimation methods (EM on jittered data, direct MLE on binned data, Bayesian MCMC) yield consistent parameter estimates, confirming the robustness of the analysis. The first component captures the smaller particles (mode near 0.09 mm) while the second captures the larger particles (mode near 0.58 mm). Although the parametric bootstrap formally rejects the model (as expected with $N > 700{,}000$), the observed-versus-expected comparison confirms excellent practical adequacy.

7.2 Comparison of Methods

The three estimation approaches offer complementary strengths. The EM algorithm on jittered data converges in 66 iterations and yields closed-form M-step updates, but its estimates depend on the particular jittering realization and operate on an approximate (continuous) likelihood rather than the exact binned one. The direct MLE on the binned likelihood avoids both issues: it opti-

mizes the exact multinomial log-likelihood and is deterministic given the same starting point. The Bayesian MCMC approach produces the most complete inferential output, providing full posterior distributions and 95% credible intervals for all five parameters; its computational cost is approximately 30 seconds for 4 chains of 7000 iterations each. In practice, the binned MLE point estimates are recommended for simulation purposes, while the Bayesian credible intervals are valuable for reporting parameter uncertainty.

7.3 Limitations

Several limitations should be acknowledged. First, the original continuous diameter measurements are not available; the binned format inherently discards information about the within-bin distribution, which limits the precision of any fitted model. Second, the non-equidistant bin grid requires careful treatment in both visualization and likelihood computation. In particular, the first bin $[0, 0.06]$ is notably wider than the others, which may mask finer structure in the small-particle regime. Third, the present analysis is restricted to a two-component mixture; we did not systematically explore whether a three-component mixture or alternative flexible distributions (e.g., gamma or Weibull mixtures) might provide a better fit. Finally, the EM estimates depend on the specific jittering realization, introducing a source of variability not present in the direct binned MLE. This dependence could be quantified through repeated jittering with different random seeds, though the close agreement between EM and MLE estimates suggests it is small.

7.4 Future Work

A natural extension would be to fit a three-component mixture and compare it to the two-component model via BIC, in order to determine whether additional components are statistically warranted. It would also be valuable to explore alternative distributional families (e.g., Weibull or gamma mixtures) and assess whether they provide a better fit to the data. From a scientific perspective, investigating the physical processes that generate the observed bimodal size distribution could connect the two mixture components to distinct particle formation mechanisms, such as precipitation versus wind erosion. Finally, the fitted Monte Carlo simulator could be integrated into a full snow transport model, following the framework of Melo et al. (2022), to quantify the impact of the particle size distribution on saltation dynamics and drifting snow flux.

References

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1): 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Melo, Daniela B., Varun Sharma, Francesco Comola, Anna Sigmund, and Michael Lehning. 2022. "Modeling Snow Saltation: The Effect of Grain Size and Interparticle Cohesion." *Journal of Geophysical Research: Atmospheres* 127 (3): e2021JD036137. https://doi.org/10.1029/2021JD036137.

Roberts, Gareth O., Andrew Gelman, and Walter R. Gilks. 1997. "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms." *The Annals of Applied Probability* 7 (1): 110–20. https://doi.org/10.1214/aoap/1034625254.