G. Gori, A. Raimondi, A. Guardone
Snowflakes Shape Characterization via Bayesian Inference: Exploring the Challenges
in: AIAA Aviation 2021 Forum, AIAA, 2021, ISBN: 9781624106101, p. 1-16, AIAA 2021-2683
[AIAA Aviation 2021 Forum, Virtual Event, 2-6 Aug. 2021]
doi:10.2514/6.2021-2683

Snowflakes shape characterization via Bayesian inference: exploring the challenges

Giulio Gori* and Alessio Raimondi† and Alberto Guardone‡
Department of Aerospace Science and Technology, via La Masa 34, 20156, Milano, Italy

This work explores the challenges underlying the inference of characteristic parameters describing the aerodynamics of blowing snowflakes, with application to in-flight snow accretion and engine ingestion. Due to the shortage of experimental observations, the classical Bayesian problem is formulated with respect to synthetic data generated from statistics of falling snow measurements. The goal is to expose issues possibly hindering the aerodynamic shape inference process, in order to anticipate barriers and envisage solutions to apply when a comprehensive experimental data set will be available. This paper provides guidelines for implementing novel experiments, including specifications concerning the desirable accuracy and precision of measurement systems.

I. Nomenclature

ρ	=	Density of air
Φ	=	Particle sphericity
$\Phi_\perp$	=	Particle crosswise sphericity
$\Phi_{//}$	=	Particle lengthwise sphericity
A_p	=	Particle cross-sectional area
$A_{\mathrm{p},\perp}$	=	Particle cross-sectional area projected on a plane perpendicular to the velocity vector
C_D	=	Aerodynamic drag coefficient of a snowflake
d_v	=	Diameter of the sphere with equivalent volume
F_b	=	Buoyancy force
Re	=	Reynolds number
v_lim	=	Falling limit velocity
W	=	Weight force

II. Introduction

SNOWFLAKES are mould into complex beautiful shapes under the action of peculiar physical mechanisms occurring in cold moisture-saturated regions within the Earth atmosphere. Typically, snowflakes nucleate around mineral or organic particles suspended in subfreezing air masses. Driven by electrostatic forces, single ice crystals amalgamate with others, thus steadily growing in size and weight. Once the snowflake reaches a sufficient weight, it starts precipitating through regions of the atmosphere characterized by different temperature and humidity which affect the growth (or reduction) process. During the falling, the shape of the crystal is also modeled by the action of the aerodynamic forces now acting on the particle. Complex unique shapes emerge as each flake precipitates to the ground.

Being able to accurately model the snowflakes shape is of the utmost relevance for many diverse applications. Here, we are concerned with in-flight icing, a quite frequent phenomenon posing serious risks to safety and operation. Generally, the severity of icing conditions varies enormously depending on the weather conditions. Compared to supercooled droplets or freezing rain, dry snow encounters represent a less significant hazard as usually crystals bounce against the airframe without sticking. However, snow is likely to accumulate on the engine air intakes, on the leading edges of aerodynamic surfaces and on windshields, thus limiting the pilots visibility. In the context of in-flight icing, an improved knowledge of the snowflakes geometry would as well allow an improved estimation of the aerodynamic forces

*Post-doc, , Department of Aerospace Science and Technology, via La Masa 34, 20156, Milano, Italy. Email: giulio.gori@polimi.it
†Master Student, Department of Aerospace Science and Technology, via La Masa 34, 20156, Milano, Italy. Email: alessio.raimondi@mail.polimi.it
‡Full Professor, Department of Aerospace Science and Technology, via La Masa 34, 20156, Milano, Italy. Email: alberto.guardone@polimi.it

acting on the snowflake, which is fundamental for accurately reconstructing the trajectory of these particles w.r.t. the aircraft, to evaluate which regions of the airframe are prone to snow accumulation during flight. The identification of such critical zones and their extension is an essential process for enhancing flights safety in adverse weather conditions.

In the meteorological field, the aerodynamics of snowflakes has been studied for decades. In 1965, Magono and Nakamura published a study titled 'Aerodynamic Studies of Falling Snowflakes' [1]. In 1971, Jiusto and Bosworth published 'Fall velocity of Snowflakes' [2]. Other notable examples are the studies of [3–8]. All these articles are focused on the experimental characterization of snowflakes in free-fall. The quantities most commonly measured are dimensions, the density, the temperature and the velocity. In early experiments, measurement techniques were quite rudimentary e.g, dimensions were measured once the snowflake settled on the ground or velocity was measured manually with a stopwatch. Recent studies rely on more advanced and precise techniques involving the use of thermometers, hydrometers, anemometers, multiple cameras and image processing codes. These set-ups provide a larger amount of information, including for instance the volume, the surface and the density of the particle. However, the characterization of the aerodynamics of a snowflake for aeronautic applications is not straightforward. Indeed, experiments typically deal with snowflakes falling in still air. Due to the wide range of sizes, snow particles may display various falling behaviors that encompass stable and unstable falling trajectories. For instance, unstable trajectories include periodic, with particle tumbling and oscillations, or chaotic motion. Typically, it is possible to discern between the stable and unstable regimes by considering a threshold Reynolds number of about 100 [9]. Note that the Reynolds number of a free-falling particle is typically defined as

$$Re = \frac{\rho \mathrm{vD}}{\mu}, \tag{1}$$

where D is the particle reference dimension and μ is the fluid viscosity. The v term is the module of the velocity of the particle relative to the fluid.

Differently, when a snowflake is lifted from the ground by the wind, we talk about blowing snow regime or blizzard. This differentiate from the falling snow regime since the snowflake is advected by the wind through the generation of aerodynamic forces that overwhelm gravity. To a great extent, the blowing regime also occurs when a snowflake suspended in the atmosphere is approached by an object moving with a velocity in the order of that of a flying aircraft. Namely, the aerodynamic field developing around the close proximity of the object will affect the particle motion.

The types of motions that a particle exhibits at a given regime is fairly poorly understood for particles with complex irregular shapes. It is thus difficult to achieve high accuracy in the drag prediction, especially at moderately high Reynolds numbers. Ideally, the blowing and falling regimes imply two physically different mechanisms requiring dedicated investigation approaches. For instance, blowing snowflakes characteristics should be measured in a wind tunnel. Nevertheless, such experiments are complex and, therefore, a comprehensive data base is still missing. Numerics helps shedding lights on the physics of snow motion. Simulations targeting the flow developing around conical graupel falling at various inclination angles have been presented in [10]. In [11] and in [12], the authors are concerned with investigating falling hail with spherical and lobed shape. However, the snow crystals present more complex shapes and, more importantly, they are known to be very different one from another.

In order to assess the snow accumulation risk for an aircraft flying through a snowfall, we need to provide a description of the composition of that snowfall in terms of snowflakes shape families. The aim of this work is to explore the challenges underlying the process of inferring the shape of snowflakes given an experimental data set reporting the falling velocity of crystals of different dimensions. The goal is to expose issues possibly hindering the shape inference process, in order to anticipate barriers and envisage solutions to exploit the moment a comprehensive experimental data set will be available. The expected results should lead to establishing more realistic snowfall models, reducing the uncertainty on our knowledge about the snowflakes shape, therefore improving the accuracy of numerical simulations for in-flight icing investigations. Ultimately, meteorologic stations at ground level could take advantage of these models to gather real-time observations about the composition of the snowfall, to complement the information dispatched in weather advisories to air traffic.

This paper is structured as follows. In Sec. III we present the mathematical model linking the snowflakes shape to some physical observables. The parametrization of the crystal geometry is also described. Section IV introduces the formulation of the Bayesian approach, the hypotheses underlying the choice of the likelihood function and of the prior probability distribution. In Sec. V we present preliminary results, whereas in Sec. VI we summarize the findings and discuss the continuation of the research.

III. The Mathematical Model

We are targeting the inference of shape parameters from the observation of the snowflake falling velocity at the ground level. Therefore, we need to rely on a mathematical model to establish a physical relation between the observable and the unknown parameters to be inferred. We employ a simple falling model based on the Newton's Law

$$\frac{1}{2}\rho v_{lim}^2 A_p C_D\,(\mathrm{Re}(v_{lim}), \mathrm{model}, \mathrm{shape}) = W - F_b, \tag{2}$$

where the particle is assumed to have reached its maximum falling velocity v_{lim} (therefore the inertial term is null).

The main challenge lies in estimating the drag coefficient of non-spherical particles in incompressible viscous fluids. A critical evaluation of the available models was presented in [13]. According to [13], it follows that a suitable model for determining the drag coefficient (C_D) of a particle must take into account both its shape and orientation. Among the available options, we select the simplified form of the Höltzer and Sommerfield model [14], hereinafter referred to as HS. In the following, we present a brief description of the HS models, referring the reader to the nomenclature for a comprehensive list of the names of the variables and the terms appearing in the equations.

The original HS model [14] depends upon three different shape parameters, namely the sphericity (Φ), the crosswise sphericity ($\Phi_\perp$) and the lengthwise sphericity ($\Phi_{//}$). In the same paper, the authors also derive a simplified model of similar performance, in terms of accuracy w.r.t. all the available data, depending on two parameters only, namely Φ and $\Phi_\perp$. Here, we chose to take advantage of this simplified model since it relies on a reduced number of parameters. According to [14], the drag coefficient can be expressed as

$$C_D = \frac{8}{Re}\frac{1}{\sqrt{\Phi_\perp}} + \frac{16}{Re}\frac{1}{\sqrt{\Phi}} + \frac{3}{\sqrt{Re}}\frac{1}{\Phi^{\frac{3}{4}}} + 0.4210^{0.4(-\log\Phi)^{0.2}}\frac{1}{\Phi_\perp}, \tag{3}$$

where

$$\Phi = \frac{\pi\, d_v^2}{A_p} \quad \text{and} \quad \Phi_\perp = \frac{\frac{\pi}{4}\, d_v^2}{A_{p,\perp}}. \tag{4}$$

To provide a physical interpretation of the shape parameters, Φ can be seen as the ratio between the surface area of the volume-equivalent sphere (with diameter d_v) and the area of the actual particle (A_p). Instead, $\Phi_\perp$ can be interpreted as the ratio between the cross-sectional area of the volume-equivalent sphere w.r.t. the cross-sectional area of the actual particle projected on a plane perpendicular to the velocity vector ($A_{p,\perp}$). Note that the definition of both Φ and $\Phi_\perp$ depend on the scale of the volume-equivalent sphere through d_v. In the following, we will assume d_v known and provided together with synthetic observations of the particle terminal falling speed. Therefore, inferring Φ and $\Phi_\perp$ is equivalent to inferring A_p and $A_{p,\perp}$. In future works, we plan to relax this assumption and consider d_v as an additional unknown parameter.

Once the drag model (3) is substituted into the force balance expression (2), we are left with an implicit equation for computing the particle terminal velocity. Indeed, the estimation of the drag coefficient requires the evaluation of the Reynolds number which in turn includes the v_{lim} term. Depending on the particle shape, through the pair Φ and $\Phi_\perp$, and dimension d_v, the terminal velocity is determined by means of an iterative numerical procedure.

IV. Methodology

The Bayes rule is employed to infer the geometry of snowflakes by taking advantage of falling velocity measurements (in fact, synthetic data). In the following, we denote $\mathbf{q}$ the vector of all quantities involved in the inference, the shape parameters described in Sec. III, and $\mathbf{o}$ the vector of the observations. Having a priori knowledge of the values that shape parameters can take, in the form of the prior probability distribution $\mathcal{P}(\mathbf{q})$, the objective is to derive the posterior distribution of $\mathbf{q}$ using the available observations $\mathbf{o}$. To this end, the Bayes formula reads

$$\mathcal{P}\left(\mathbf{q} \mid \mathbf{o}\right) = \frac{\mathcal{P}\left(\mathbf{o} \mid \mathbf{q}\right)\mathcal{P}\left(\mathbf{q}\right)}{\mathcal{P}\left(\mathbf{o}\right)}. \tag{5}$$

In (5), we have denoted $\mathcal{P}\left(\mathbf{q} \mid \mathbf{o}\right)$ the posterior density of the parameters vector $\mathbf{q}$, $\mathcal{P}(\mathbf{o})$ the probability density of the data (also called evidence), and $\mathcal{P}\left(\mathbf{o} \mid \mathbf{q}\right)$ the so-called likelihood of the data. Note that the explicit computation of the evidence is not necessary since it acts as a scaling constant normalizing the integral of the posterior distribution to 1. Here, we are interested in inferring the shape parameters only i.e., in finding peaks in the posterior distributions, therefore we can spare the effort of computing the evidence. The definition of the posterior distribution in (5) requires

the prior and the likelihood to be specified. We assume uniform prior distributions only, thus fixing a (finite) range Q with $\mathbf{q} \in Q$ of possible values for each of the unknowns, disregarding any correlation or dependence between the parameters in $\mathbf{q}$. In practice, the prior ranges are selected in relation to some geometrical considerations. For instance, the sphericity Φ is a value ranging from 1 (a perfect sphere) and 0 (a degenerate flat geometry). Note that the goal of this paper is to assess whether it is possible to infer the geometry of falling snowflakes based on the velocity measurement at the ground level: the point is more to show the information brought by the data, rather than inferring an exact shape. In other words, the use of a synthetic data set ensures that the truth is included within the selected bounds. The uniform prior assumption is a convenient choice since it results into the definition of a posterior which is proportional to the likelihood for $\mathbf{q} \in Q$ and vanishing for $\mathbf{q} \notin Q$:

$$\mathcal{P}\left(\mathbf{q} \mid \mathbf{o}\right) \propto \begin{cases} \mathcal{P}(\mathbf{o} \mid \mathbf{q}) & \mathbf{q} \in Q, \\ 0 & \mathbf{q} \notin Q. \end{cases} \tag{6}$$

The likelihood is the probability of observing the synthetic falling velocity $\mathbf{o}$, according to the model and given a particular choice of the shape parameters $\mathbf{q}$. For the definition of the likelihood, we follow the suggestions of [15], claiming that, if we are in possession of some testable information about our experimental data, we should make the assignment for the likelihood following the principle of maximum entropy. If the testable information consists in the knowledge of just the individual variances (as we will assume hereinafter), the maximum entropy principle translates straightforwardly to assuming a product of Gaussian distributions, each centered on a specific single observation [16]. Namely, we penalize values of $\mathbf{q}$ that yield larger discrepancies between the synthetic data $\mathbf{o}$ and their prediction $\tilde{\mathbf{o}}(\mathbf{q})$ obtained by evaluating the mathematical model presented in Sec. III. The primary assumption underlying the product form of the likelihood in (7) concerns the independence of the error on each of the components of $\mathbf{o}$. Although synthetic data are expected to involve correlated discrepancies, the use of independent error model is justified in the case of experimental measurement noise and is therefore appropriate for our numerical experiments. According to the above assumptions, the likelihood reads

$$\mathcal{P}(\mathbf{o} \mid \mathbf{q}) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{|o_i - \tilde{o}(\mathbf{q})|^2}{2\sigma_i^2}\right]. \tag{7}$$

In (7), we have denoted o_i the single observation i.e., the i-th component of the vector $\mathbf{q}$ and σ_i the typical error (standard deviation) associated to o_i. $\tilde{o}$ is the prediction from the HS model, for a specific set off parameters $\mathbf{q}$.

The likelihood described in (7) suits the inference problem considering only one set of unknown parameters $\mathbf{q}$ characterizing the population i.e., all snowflakes belong to the same family and differences among crystals are seen as the variance of the shape parameters w.r.t. a unique mean value. Unfortunately, the complexity of the physical mechanisms that mould the snowflakes during the fall give birth to a wide variety of shape families, making the above choice limiting for real case applications. By acknowledging the presence of different crystals type subpopulations in a single snowfall, we extend the formulation of the likelihood function (7) to a Gaussian Mixture Model (GMM). Namely, we assume the snowfall to be characterized by M snowflake families, and we define the unknown probability π_m for a snowflake to belong to subpopulation m. The likelihood then becomes

$$\mathcal{P}(\mathbf{o} \mid \mathbf{q}) = \prod_i \sum_m^M \pi_m \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{|o_i - \tilde{o}(\mathbf{q}_m)|^2}{2\sigma_i^2}\right], \tag{8}$$

where the π_m are unknowns additional parameters to be inferred through the Bayesian approach. Clearly, the above procedure is valid under the constraint of that $\sum_m^M \pi_m = 1$. $\mathbf{q}^m$ is the vector containing the mean shape parameter values associated to the snowflake class m.

According to our approach, in Eq. (8) the number M of sub-populations characterizing the snowfall is a finite arbitrary quantity. There is always uncertainty concerning the selection of the number of the M mixture components to be included in the statistical model. Since the π_m are to be inferred, a large M can quickly lead to an explosion of the computational cost of the Bayesian procedure. At the same time, one should choose M such that all the shape families are represented, to ensure the adequacy of the model fit. A standard procedure is to begin with a small number of mixture components and then to increase it until sub-populations are not reflected by the model anymore.

Having specified all the ingredients defining the posterior distribution of the unknown parameters, it is now possible to infer the shape of crystals characterizing a snowfall by taking advantage of fall velocity observations.

Unfortunately, the likelihood (7) has no closed-form expression since model predictions $\tilde{o}(\mathbf{q})$ are obtained solving the implicit expression (2). This limitation prevents the analytical evaluation of the posterior. Therefore, brute force approaches or sampling methods are required for its estimation. When the dimensionality of the inference problem allows for it, namely when the number of unknowns (M and $\ell(q)$) is limited, brute force approaches may be employed i.e., the posterior can be numerically evaluated on a regularly spaced grid spanning the full prior probability space. Naturally, a brute force approach becomes quickly unfeasible since the number of likelihood evaluations (implying model evaluations) explodes exponentially, with the dimensionality of the prior probability space. Alternatively, Markov Chain Monte Carlo (MCMC) methods are often used to draw samples from complex multi-dimensional distributions with a large number of dimensions. Among the different MCMC alternatives, in this work we take advantage of the popular Metropolis-Hastings (MH) algorithm [17] to draw a sequence of samples from the posterior distribution. The core idea of the MH algorithm is to construct a random Markovian sequence of values for $\mathbf{q}$. A considerable number of steps, typically tens to hundreds of thousands, is necessary to produce sufficiently many samples representative of the whole distribution. This of course implies evaluating $\tilde{o}(\mathbf{q})$ multiple times. Luckily, the evaluation of the mathematical model considered here is computationally immediate, thus making the implementation of direct Monte Carlo methods feasible i.e., no surrogate modelling is required.

Note that in the formulation adopted in this paper, we do not enforce the equality constraint $\sum_m^M \pi_m = 1$ explicitly in the generation of samples from the proposal distribution. Instead, we leave the inferential procedure free to explore the uncertainty space with no limit if not just the prior bounds on each π_m. Besides the standard rejection rule from the Metropolis-Hasting algorithm, we also reject proposed samples based on the inequality constraint $\sum_m^M \pi_m \leq 1$. Results reveal that this naïve approach is still capable of retrieving acceptable solutions for which $\sum_m^M \pi_m = 1$. Nevertheless, we do also acknowledge that more advanced approaches can be deployed to handle this constraint. We leave the evaluation of refined methodologies for future works, as the goal of this contribution is a preliminary exploration of the challenges implied by the process of inferring the shape of snowflakes from experimental observations. Nevertheless, we indicate that one possible way to proceed could be to formulate a Hierarchical models, perhaps taking advantage of Dirichlet processes [18].

A. The Generation of the Synthetic Data Set

The observation of the velocity of hydrometeors requires the deployment of complex experimental rigs based on optical devices. In [6], the authors take advantage of a disdrometer to investigate the aggregate terminal velocity (the velocity at ground level) of falling snowflakes and demonstrate the capability of obtaining detailed information, including the shape and the size of the particles. Nevertheless, even though hexagonal forms, capped columns, and needles are in principle detectable, the component forms of aggregates are not easily identified and therefore not reported.

In addition to the work of Brandes [6], we mention Refs. [2, 7] concerning experimental campaigns aimed at investigating the snowflakes shape and their terminal velocity. Besides these works, literature is in short supply of experimental data concerning snowflakes. Moreover, authors do not always provide a comprehensive or complete description of the data set e.g., Ref. [6] does not report the component forms of aggregates. For these reasons, and because we are aiming at exposing possible challenges underlying the inference of characteristic parameters describing the shape of falling snowflakes, we decide to rely on a synthetically generated data set. This choice gives us full control over the observations and their precision, making it possible to investigate the inference problem from different perspectives. For each component, we generate observations by evaluating the final velocity model w.r.t. samples from a multivariate normal distribution of Φ and Φ $\perp$, for which we arbitrarily impose the mean vector and the covariance matrix. In multi-component tests, observations are generated considering different components mean and mixing lengths. The numerical values employed to build the synthetic data sets will be given later for each test case. Here, we describe the general procedure.

The vector of target shape parameters for each component $\mathbf{q}_m = (\Phi^m, \Phi^m_\perp)^T$ with $m = (1, \ldots, M)$ is generated arbitrarily. Note that we can define a set of classes identifying the various snowflake families $(C1, C2, \ldots, CM)$. The mixing length of each components π_m are also arbitrarily chosen under the constraint of that $\sum_1^M \pi_m = 1$. Having decided these figures, we first sample a multinomial (categorical) distribution, having the π_m parameters specifying the probabilities of the snowflake to belong to one of the classes. After, we sample the dimension of the snowflake, the equivalent diameter d_v, from a uniform distribution bounded in between 1 and 16 mm. Once the class and the scale are defined, we finally sample the shape of the snowflake from the multivariate normal distribution corresponding to the

parent class $\mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m)$. The mean vector $\boldsymbol{\mu}^m$ and the covariance matrix read

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_\Phi \\ \mu_{\Phi_\perp} \end{bmatrix}, \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\Phi & 0 \\ 0 & \sigma_{\Phi_\perp} \end{bmatrix}, \tag{9}$$

where implicitly we assume that Φ^m and $\Phi^m_\perp$ are uncorrelated parameters (null off-diagonal covariance terms).

Once the parameters set is generated, we evaluate the terminal velocity o_i of the snowflake by evaluating Eq. (2). To each synthetic observation, we associate an error-bar (equal for all points) σ_i emulating some sort of gaussian measurement noise affecting the precision of the measurement system. Note that we assume that the information about the snowflake scale i.e., the value of d_v, is a complementary information provided together with the terminal velocity.

V. Results

In Section V.A, we consider a snowfall composed by a single class of snowflakes, to explore the challenges underlying the inference of the snowflakes' shape. Namely, we first verify the correctness of the implementation of our Bayesian framework by assessing whether the computerized program is capable of retrieving the synthetic solution. We also take advantage of the single-component snowfall problem in order to explore how the inferential procedure varies w.r.t. the observation alleged precision. Secondly, we consider a two and three-component snowfall, in Sec. V.B, and we rely on the Gaussian mixture formulation with the goal of inferring the shape parameters and the mixing length of each component.

In all the test cases presented, the MCMC chain is left exploring the probability space for an initial burn-in phase of 35000 samples. The burn-in phase is required to ensure that the sampling procedure becomes independent from the (random) initial point, and to adapt the proposal distribution according to the empirical covariance revealed by the collected points. Note that we initialize the chain assuming that the unknowns are uncorrelated i.e., the initial proposal distribution is spherical. After the burn-in, the actual MCMC chain is then built considering 80000 samples. Empirical trials showed that this setting is sufficient for retrieving statistically meaningful sample sets.

A. Single component snowfall

This first test case considers three different snowfalls. Each snowfall is characterized by the presence of a unique family (or class) of snowflakes. Therefore, we generate three different synthetic data sets corresponding to different families of snowflakes $C1$-$C2$-$C3$. Namely, observations are generated by evaluating the falling velocity equation (2) (and drag model HS) w.r.t. samples from three different multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m)$) reading

$$\boldsymbol{\mu}^1 = \begin{bmatrix} 0.2 \\ 0.2 \end{bmatrix}, \quad \boldsymbol{\mu}^2 = \begin{bmatrix} 0.5 \\ 0.9 \end{bmatrix}, \quad \boldsymbol{\mu}^3 = \begin{bmatrix} 0.9 \\ 0.3 \end{bmatrix}, \quad \text{with} \quad \boldsymbol{\Sigma}^1 = \boldsymbol{\Sigma}^2 = \boldsymbol{\Sigma}^3 = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}. \tag{10}$$

Note that we explicitly assume that the two shape parameters are not correlated since the off diagonal terms in Σ are null. Each of the three data sets includes 100 observation points. The inferential procedure is carried out (independently) on each of the three data sets, assuming a single component mixture model for each snowfall. Figure 1, 2 and 3 report, respectively, the shape parameters sampled for the targeted snowflake family (left-hand side) and the corresponding synthetic terminal velocity observations (right-hand side). The synthetic velocity observations are quite insensitive to the snowflakes shape (and family) when the equivalent diameter d_v is small. Indeed, all the three plots report comparable values. Differences in particles terminal velocity become relevant at the large scale. Namely, $C1$ snowflakes reach a maximum falling speed of about 0.35 m/s, whereas $C2$ reaches 0.9 m/s and $C3$ a value larger than 1 m/s.

In the inferential procedure, we consider four different levels of measurement precision for the synthetically generated data (assumed equal for all the data points) namely, the standard deviation is $\sigma_o = (0.6, 0.3, 0.1, 0.05)^T$ m/s. These values are coherent with real instrumentation characteristics, see [6]. Namely, in general it is difficult to estimate the error associated to terminal velocity observations. Indeed, the error depends on the density, the shape, the size, and the orientation of the snowflake. Nevertheless, it is possible to take advantage of the alleged dispersion in terminal velocities for raindrops [5] (for which the standard error varies from 0.4 m/s, for drops with d_v of about 0.5 mm, to 0.2 m/s, for drops larger than 2 mm). Errors for snowflakes should be less because of the finer vertical resolution. Therefore, the inferential procedure is carried out on each of three data sets, considering each of the four observation confidence levels. Figure 4(a-d) report the results of the inferential procedure. Namely, each row (a-d) corresponds to a certain measurement precision, whereas each column is related to a snowflake family (from left to right, C1-C3). In

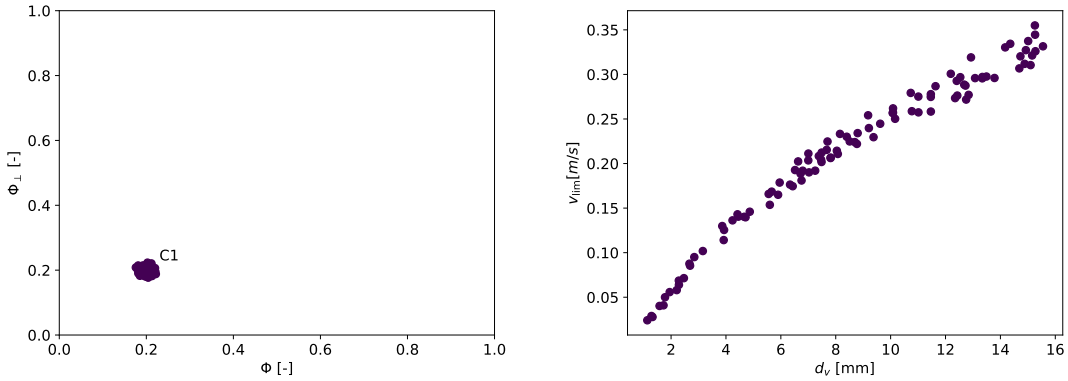


Fig. 1 Test case A. The shape parameters sampled for snowflakes belonging to the $C1$ family (left-hand side) and corresponding synthetic terminal velocity observations (right-hand side).

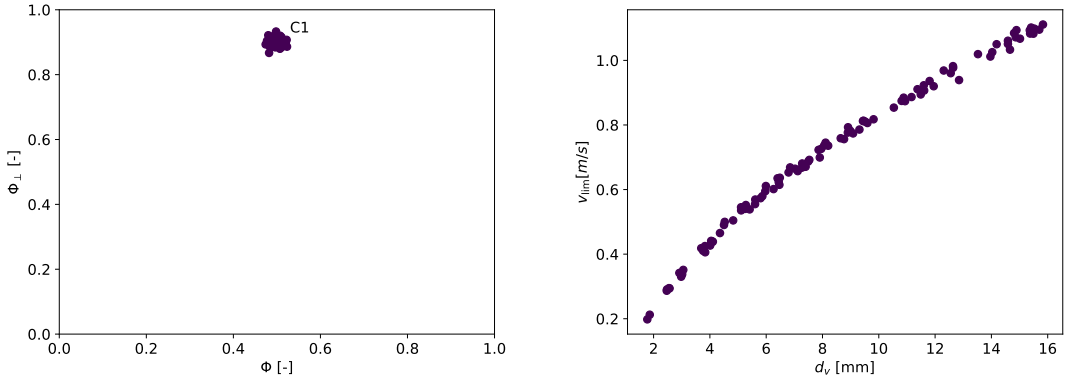


Fig. 2 Test case A. The shape parameters sampled for snowflakes belonging to the $C2$ family (left-hand side) and corresponding synthetic terminal velocity observations (right-hand side).

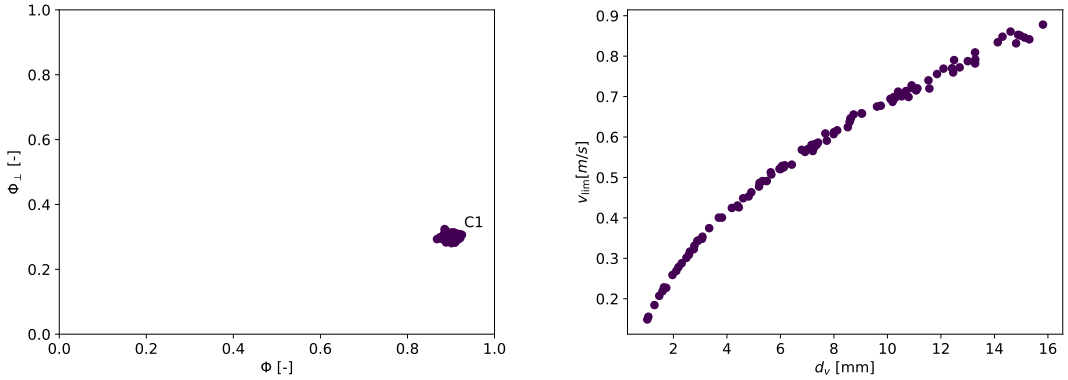
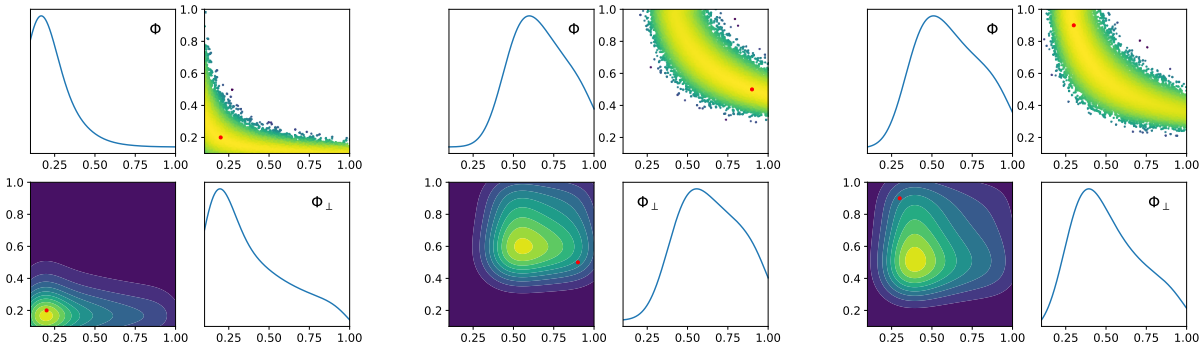
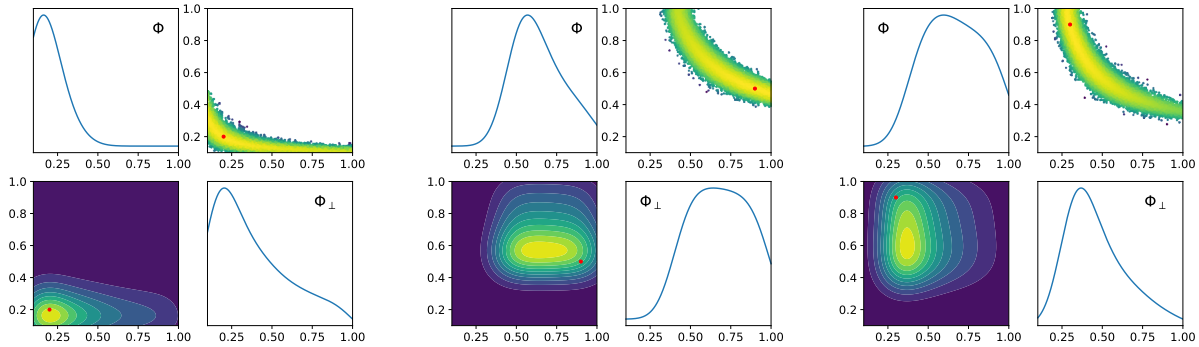


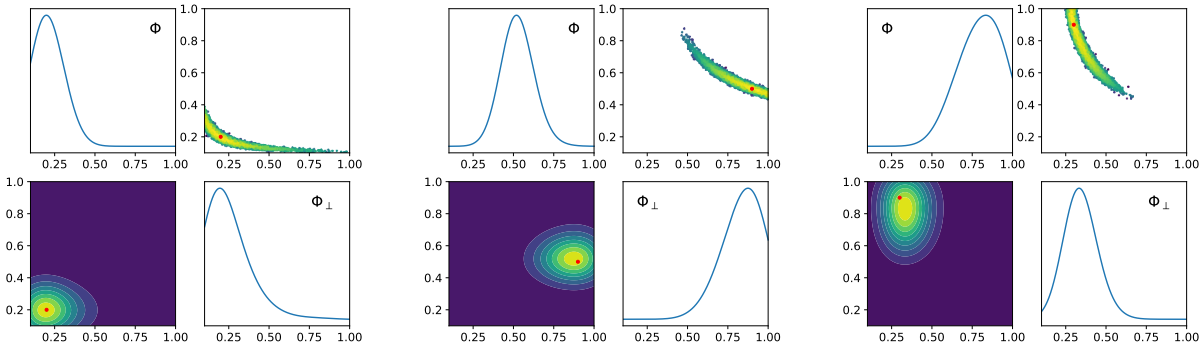
Fig. 3 Test case A. The shape parameters sampled for snowflakes belonging to the $C3$ family (left-hand side) and corresponding synthetic terminal velocity observations (right-hand side).



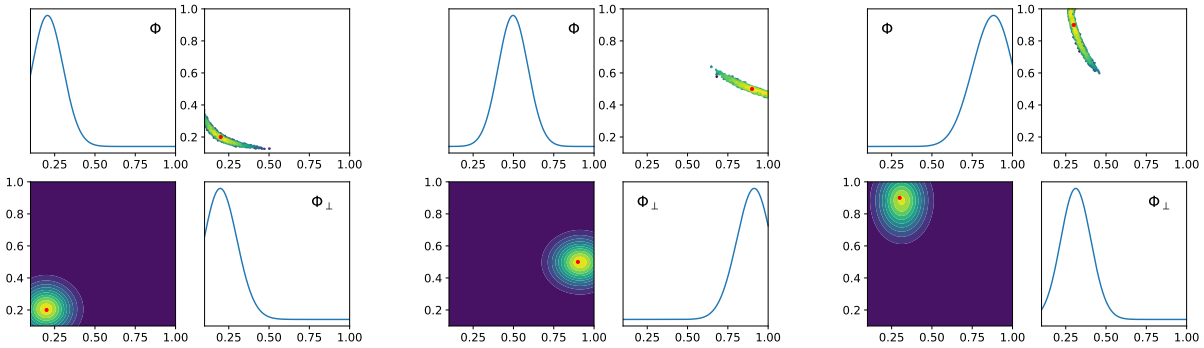
(a) σ_o = 0.6 m/s



(b) σ_o = 0.3 m/s



(c) σ_o = 0.1 m/s



(d) σ_o = 0.05 m/s

Fig. 4 Test case A. The shape parameters inferred for snowflakes belonging to the $C1$ (left), $C2$ (middle) and $C3$ (right) families.

each sub-figure, we report the Φ and $\Phi_\perp$ posterior marginal distributions inferred from the data. On the off diagonal term, in position top-right, we report the MCMC chain (neglecting the burn-in phase) employed to explore the prior space. In position low-left we plot instead the Φ and $\Phi_\perp$ joint distribution (computed under the assumption of statistical independence). Red dots in the off diagonal plots indicate the targeted values i.e., the μ^m parameters employed to generate the data. Clearly, the inferential procedure finds it difficult to retrieve the target values of $C2$ and $C3$ if the precision of the measurement system has standard deviation of 0.6 and 0.3 m/s. This indicates that, according to the physical model employed, the difficulty of inferring the shape of a falling snowflake depends on the shape itself. For each of the considered classes, the MCMC chain keeps sampling an arc-shaped region of highly probable pairs of parameter shape values, including the target coordinates. At the same time, the posterior marginal distributions clearly indicate a maximum a posteriori value for the each class shape parameters. Nevertheless, for $C1$ both posterior maxima are correctly located in the close proximity of $(\Phi, \Phi_\perp) = (0.2, 0.2)$. Moreover, the joint probability surface clearly shows a peak at (0.2, 0.2). Instead, for $C2$ the maximum for Φ is approximately 0.5, whereas for $\Phi_\perp$ the mass of the posterior is smeared over a region centered on 0.6 and it fails to retrieve the target value of 0.9. Moreover, the maximum of the joint probability surface does not include the target value, if not just marginally. Same considerations apply to $C3$. These results indicate that the inference process, or rather the HS model, suffers from a certain dependency w.r.t. the coordinates of the unknown parameters. Namely, the inference of the shape of certain snowflakes appears to be easier ($C1$ w.r.t. $C2$ and $C3$). As the precision is increased to 0.1 or 0.05 m/s, we can appreciate that the procedure is capable of retrieving the targeted parameters also for the $C2$ and $C3$ families. For $\sigma_o \leq 0.1$, a clear indication about the true value of the $C2$-$C3$ unknowns is given by sharp marginal posterior distributions. The joint posterior also reveals a sharp peak in correspondence of the target values for Φ, $\Phi_\perp$. In all cases, the MCMC chain is associated to a satisfactory acceptance rate. For $C1$, we record an acceptance rate of about 11.45%, 17.86%, 12.34% and 11.76%, w.r.t. the increasing observation precision. At the same time, the inferential process of $C2$ show an acceptance rate of about 32.26%, 25.83%, 19.20% and 11.72%. Similar acceptance rate values are associated to $C3$ (28.90%, 22.34%, 19.02% and 11.05%). Interestingly, even though we employ the very same numerical set up, the inference of $C1$ is endowed with a slightly lower acceptance rate. That is, depending on the shape to be inferred, the chain may be left exploring the probability space for a larger number of steps in order to obtain a converged posterior prediction.

This test case serves a twofold objective. On one hand, it verifies the correctness of the implementation of our Bayesian framework into a computerized program. Indeed, the code is capable of retrieving the synthetically generated data. On the other hand, we obtain an indication concerning the minimum precision requirement for the measurement system employed to observe the snowflake terminal velocity on the ground. Moreover, results also show that, depending on the snowflake family, it may be easier to infer the shape. Indeed, the family characterized by $(\Phi, \Phi_\perp) = (0.2, 0.2)$ is correctly inferred even if observations are associated to a low precision. Of course, this result is strongly related to the computational model employed to relate the shape of the crystal to its terminal velocity i.e., the HS drag model. Employing a different model may lead to a different outcome.

B. Multiple components snowfall

This second example considers multiple snowflakes families characterizing the snowfall at once. The number of families effectively considered to generate the data is again three. Class parameters $\mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m)$ resemble the ones employed in Sec. V.A (reported here for clarity)

$$\boldsymbol{\mu}^1 = \begin{bmatrix} 0.2 \\ 0.2 \end{bmatrix}, \quad \boldsymbol{\mu}^2 = \begin{bmatrix} 0.5 \\ 0.9 \end{bmatrix}, \quad \boldsymbol{\mu}^3 = \begin{bmatrix} 0.9 \\ 0.3 \end{bmatrix}, \quad \text{with} \quad \boldsymbol{\Sigma}^1 = \boldsymbol{\Sigma}^2 = \boldsymbol{\Sigma}^3 = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}, \tag{11}$$

The dimension of the data set is again 100, but, differently than the previous case, it now includes observations related to different families of snowflakes at once. In addition to the previous case, we also consider additional unknown parameter i.e., the mixing length π^m of each component. In the following, we consider different mixing length combinations.

1. Two classes

We assume that $\boldsymbol{\pi} = (0.3, 0.7, 0.0)$. Therefore, only two families are effectively present in the data set. Figure 5 reports the sampled parameters and their corresponding synthetic observations. Clearly, different snowflakes families give birth to different observation trends. Again, we consider four different level of precision for the synthetically generated data, $\sigma_o = (0.6, 0.3, 0.1, 0.05)^T$ m/s. We unroll our Bayesian sampling procedure considering a two-components (M=2)

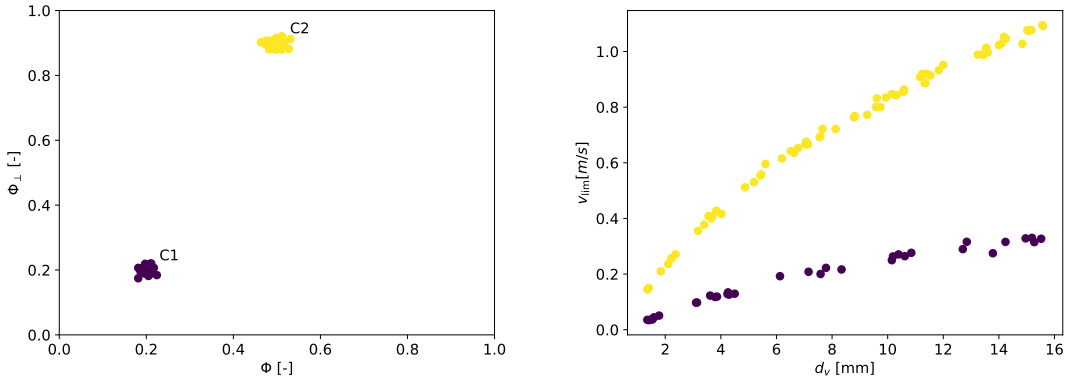


Fig. 5 Test case B, two classes. The shape parameters sampled for the different snowflakes categories (left-hand side) and corresponding synthetic terminal velocity observations (right-hand side).

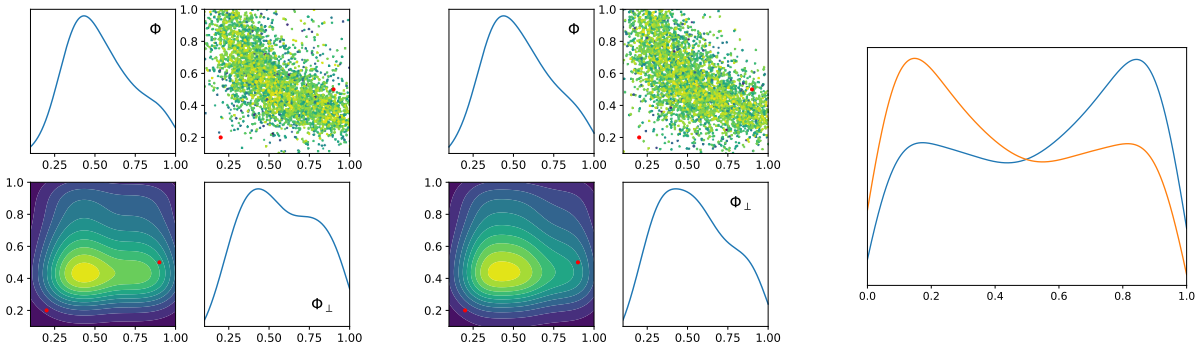
GMM model and we obtain the results reported in Fig. 6. Again, the inference procedure leads to inferring the target values only if the measurement system is endowed with a precision of at least 0.1 m/s. In such cases, the framework is capable of identifying the two classes with a fair confidence. This is evident from the joint probability plots, that clearly highlight the region marked with the red reference dot. In the other cases, the framework instead fails to identify the correct region.

Figure 6 reports also the mixing length components inferred from the data sets, on the right hand-side column. The same conclusions about the experimental rig requirements can be drawn. Indeed, results show that the precision of the measurement system greatly affects the inference of the mixing lengths and that a precision of at least 0.1 m/s is required to properly infer the correct mixing length values of 0.3 and 0.7.

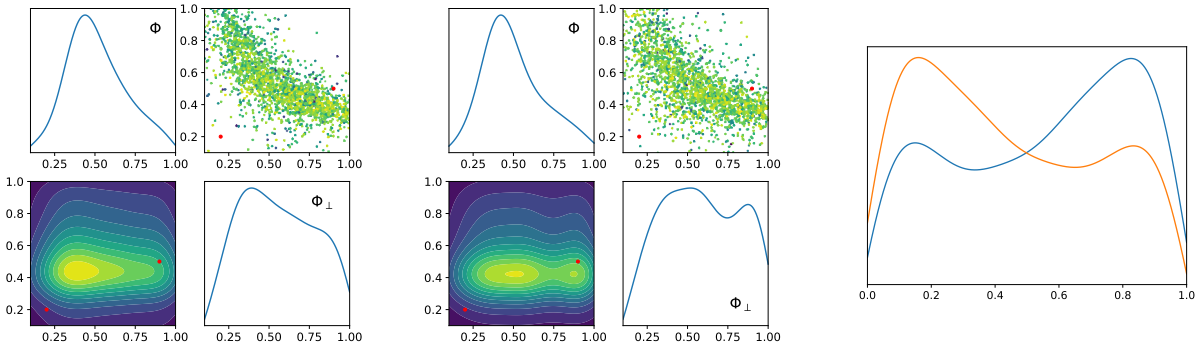
For the two-class inference problem, we observe an abrupt decrease of the MCMC acceptance rate. Indeed, the values obtained are about 4.68%, 2.60%, 2.86% and 0.95% for, respectively, $\sigma_o = (0.6, 0.3, 0.1, 0.05)^T$. These values reveal a very inefficient sampling process characterized by a high rejection rate. Namely, a large number of samples (from 90% to 60%) is rejected due to failing in fulfilling the constraint of that the mixture lengths sum must be smaller than 1. Even though the naïve implementation of our Bayesian framework is capable of retrieving the targeted solution, the low efficiency of the inferential procedure points out the need for a more advanced approach for handling constraints.

2. Three classes

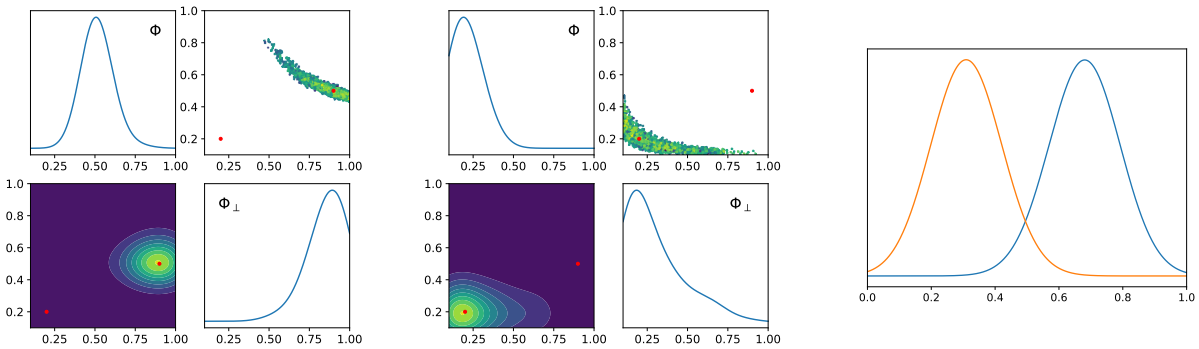
We assume that $\boldsymbol{\pi} = (0.3, 0.2, 0.5)$. Therefore, three snowflakes families are considered. Figure 7 reports the sampled parameters and their corresponding synthetic observations. Despite a quite significant difference in between $C2$ and $C3$, synthetic observation does not depart much from each other, especially for a value of $d_v \lesssim 6$ mm. On the contrary, the trend of the synthetic observations related to $C1$ is well distinguishable within the whole span of dimension range. As shown in previous test cases, this possibly explains why the $C1$ class is easier to infer than $C2$ and $C3$. We now unroll our Bayesian sampling procedure considering a three-components (M=3) GMM model and, again, assuming $\sigma_o = (0.6, 0.3, 0.1, 0.05)^T$ m/s. Results from the inference process are reported in Fig. 8. The inference procedure now reveals that a more strict measurement precision requirement must be enforced to obtain informative posterior distributions for all classes. Indeed, if $\sigma_o = 0.1$ m/s, the inferential process fails in retrieving two out of three classes ($C2$ and $C3$). Namely, just the class associated to the (Φ, $\Phi_\perp$) = (0.2, 0.2) pair is clearly inferred. This is possibly related to the fact that the synthetic observation trends related to $C2$ and $C3$ are superimposed for a certain range of d_v values, while they just slightly depart in the remaining portion of the particle scale domain, see 7. Therefore, a higher precision (in the order of $\sigma_o = 0.05$ m/s) is needed to handle the three class problem properly (at least considering the figures selected for this test case). Under this condition, the framework is capable of identifying the three classes with a fair confidence. Figure 9 reports the mixing length components inferred from the data sets w.r.t. the different σ_o. With a strict precision requirement of $\sigma_o = 0.05$ m/s, the targeted mixing length are fairly inferred. Namely, the peaks are found at π of about 0.15, 0.3 and 0.55. The slight overestimation of the denser $C3$ class, and the slight underestimation of the emptier $C2$ class, are possibly due to the model scarce sensitivity w.r.t. data points related to small scale particles. The model indeed returns similar terminal velocity values for particles of $d_v \lesssim 4$ mm which make



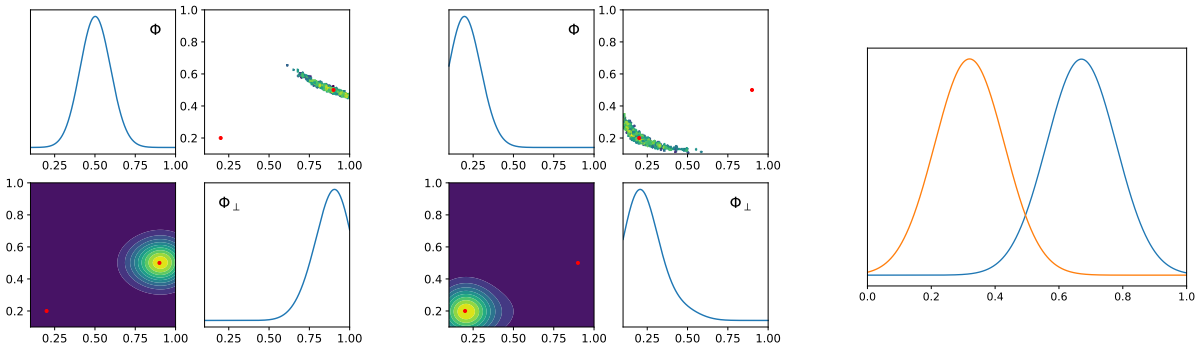
(a) σ_o = 0.6 m/s



(b) σ_o = 0.3 m/s



(c) σ_o = 0.1 m/s



(d) σ_o = 0.05 m/s

Fig. 6 Test case B, two classes. The shape parameters inferred for snowflakes belonging to the $C1$-$C2$ families (left and middle columns) and the inferred mixing lengths (right column).

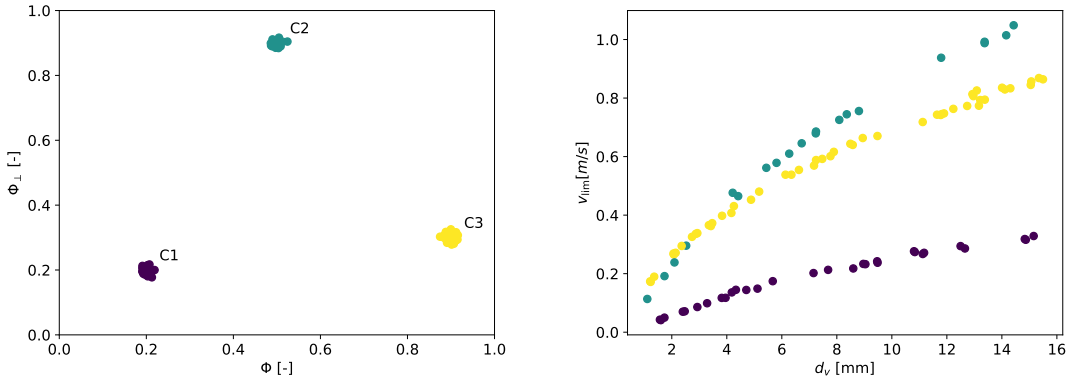


Fig. 7 Test case B, three classes. The shape parameters sampled for the different snowflakes categories (left-hand side) and corresponding synthetic terminal velocity observations (right-hand side).

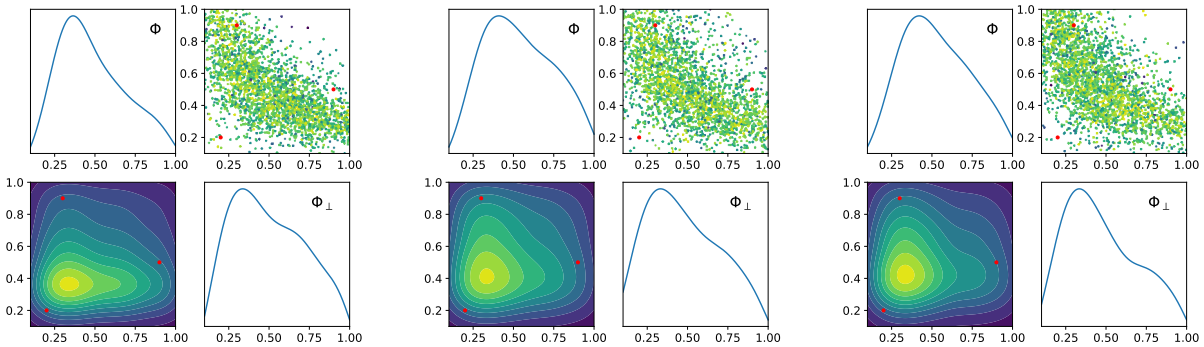
the two classes indistinguishable. Nevertheless, the inferential procedure is capable of retrieving a solution consistent with the requirement of that $\sum_1^M \pi_m = 1$. That is, the information brought by the data is sufficient to infer the unknown mixing lengths. The fact that they sum up to one is just a consequence of the fact that data were generated accordingly.

For the three-class inference problem, we observe an even (slight) further decrease of the MCMC acceptance rate. Indeed, the values obtained are about 3.14%, 1.92%, 1.86% and 0.90% for, respectively, $\sigma_o = (0.6, 0.3, 0.1, 0.05)^T$. Our numerical experiments suggest that the sampling process becomes more inefficient as the number of classes to be inferred increases. Again, the high rejection rate is due to a large number of samples (from 90% to 60%) rejected due to failing in fulfilling the applied constraints. Because of the low acceptance rate, this test case was analyzed considering also a burn-in chain of 135000 steps and a MCMC chain of 380000 samples. This analysis confirm the observations arising from the shorter chain.

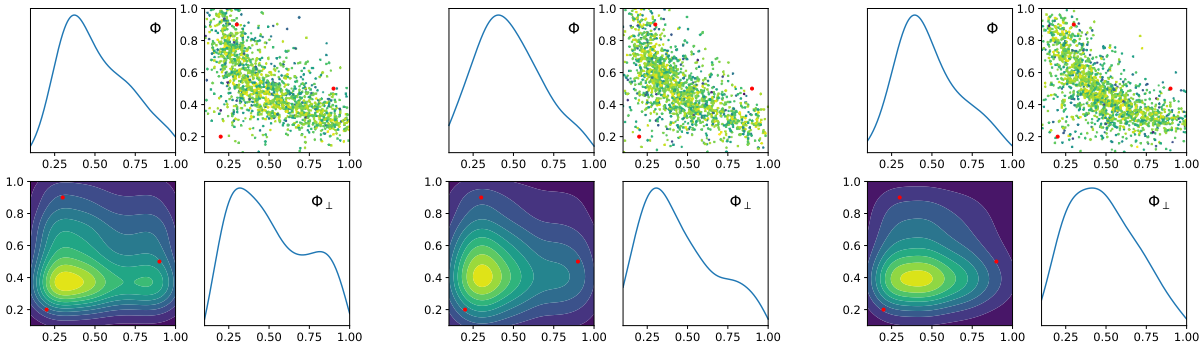
VI. Conclusions

In this paper, we explored the challenges underlying the inference of characteristic parameters describing the shape of snowflakes. The inference relies on synthetic observations emulating experimental measurements targeting the falling speed of snowflakes at ground level. The computational framework implements a Bayesian approach capable of correctly inferring the shape parameters of snowflakes based on the provided observations. Despite the simplicity of the proposed formulation, results confirm its soundness and its ability to correctly retrieve an artificial solution. Nevertheless, numerical experiments suggest that snowflakes endowed with a flat shape (a small value of Φ) and having their major dimension oriented in the direction perpendicular to the fall (a small value of $\Phi_\perp$) are more easily identifiable. At the same time, numerical experiments show that it is possible to discern snowflakes belonging to up to three different families within the same snowfall. The achieving of this goal requires a measurement system of the highest precision (better than 0.1 m/s). However, we should emphasize that the conclusions of a Bayesian analysis are always conditional on the data, our prior knowledge and the mathematical model used to describe the phenomena.

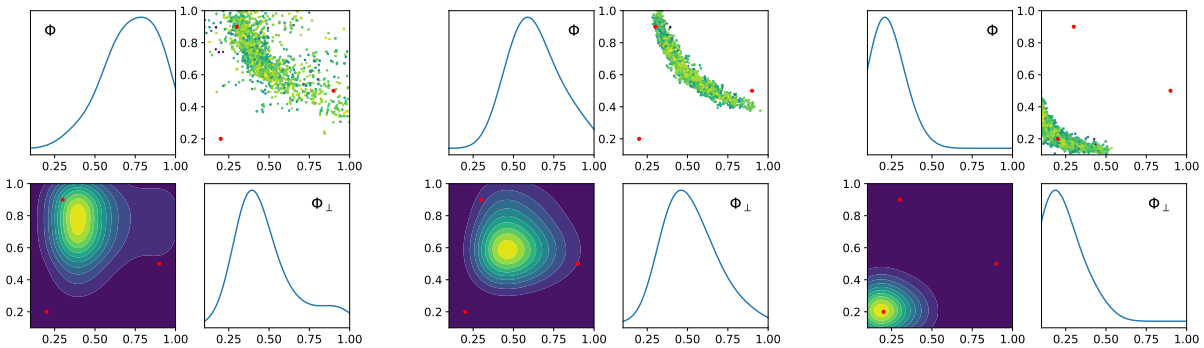
Nevertheless, moving from the results shown here, we anticipate barriers and envisage solutions to exploit the moment a comprehensive experimental data set will be available. Increasing the number of classes to be included in the mixture model naturally increases the complexity of the inference procedure. Though Monte Carlo-based methods surely represent the way to proceed, approaches more advanced than the one employed in this paper must be implemented. Namely, our methodology is endowed with a scarce efficiency worsening with an increasing number of snowflake classes to be inferred. That is, a large number of samples is rejected because of the failing in fulfilling the constraint on the mixture lengths sum. Possibly, Hierarchical models, perhaps taking advantage of Dirichlet processes, may help improving the overall efficiency of the inference. At the same time, regularization strategies may help introducing some parsimony in establishing the number of relevant classes of snowflakes characterizing the snowfall. An open concern remains about considering different drag models for producing the very same analysis reported in this paper. Indeed, drag models other than the HS may produce significantly different indications.



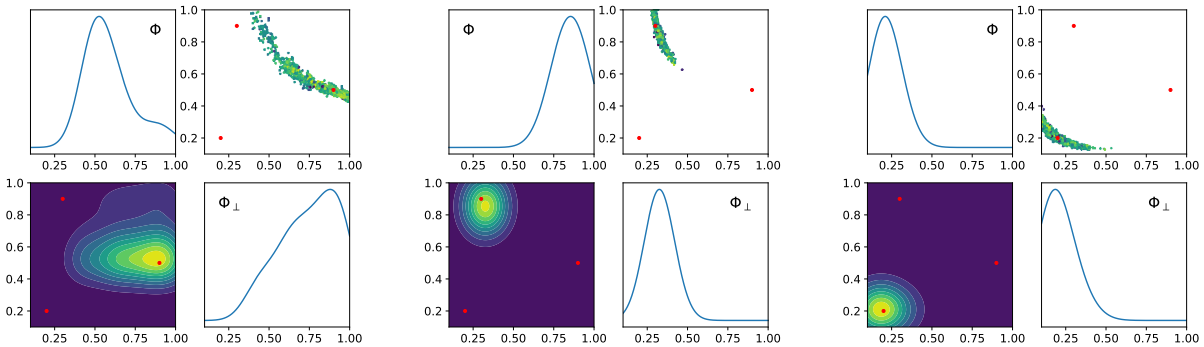
(a) σ_o = 0.6 m/s



(b) σ_o = 0.3 m/s

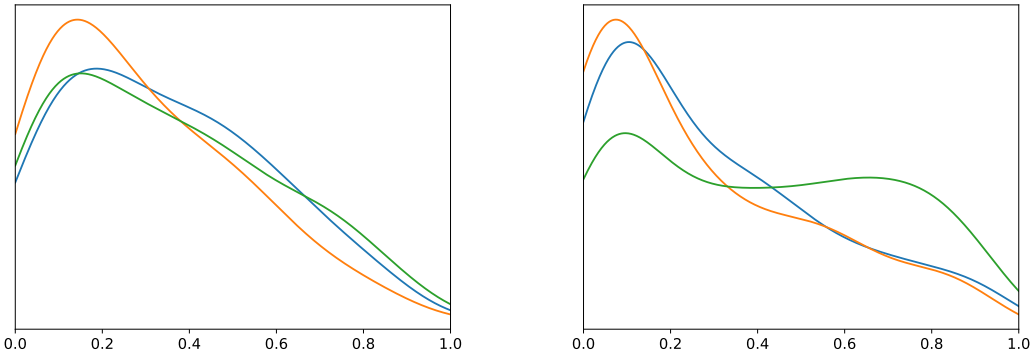


(c) σ_o = 0.1 m/s



(d) σ_o = 0.05 m/s

Fig. 8 Test case B, three classes. The shape parameters inferred for snowflakes belonging to the $C1$-$C3$-$C3$ families (note that columns are not necessarily in this order).



(a) σ_o = 0.6 m/s

(b) σ_o = 0.3 m/s

(c) σ_o = 0.1 m/s

(d) σ_o = 0.05 m/s

Fig. 9 Test case B, three classes. The mixing length inferred for the three considered families, assuming different measurement precision.

Acknowledgments

A. Guardone acknowledges funding from the European Union's Horizon 2020 under grant agreement No. 824310 ICE GENESIS.

References

[1] Magono, C., and Nakamura, T., "Aerodynamic Studies of Falling Snowflakes," *Journal of the Meteorological Society of Japan. Ser. II*, Vol. 43, No. 3, 1965, pp. 139–147. https://doi.org/10.2151/jmsj1965.43.3_139.

[2] Jiusto, J. E., and Bosworth, G. E., "Fall Velocity of Snowflakes," *Journal of Applied Meteorology and Climatology*, Vol. 10, No. 6, 1971, pp. 1352 – 1354. https://doi.org/10.1175/1520-0450(1971)010<1352:FVOS>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/apme/10/6/1520-0450_1971_010_1352_fvos_2_0_co_2.xml.

[3] List, R., and Schemenauer, R. S., "Free-Fall Behavior of Planar Snow Crystals, Conical Graupel and Small Hail," *Journal of Atmospheric Sciences*, Vol. 28, No. 1, 1971, pp. 110 – 115. https://doi.org/10.1175/1520-0469(1971)028<0110:FFBOPS>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/atsc/28/1/1520-0469_1971_028_0110_ffbops_2_0_co_2.xml.

[4] Kajikawa, M., "Observation of the Falling Motion of Early Snow Flakes Part I. Relationship between the Free-fall Pattern and the Number and Shape of Component Snow Crystals," *Journal of the Meteorological Society of Japan*, Vol. 60, No. 2, 1982. URL https://www.jstage.jst.go.jp/article/jmsj1965/60/2/60_2_797/_pdf.

[5] Brandes, E., Ikeda, K., Zhang, G., Schoenhuber, M., and Rasmussen, R., "A Statistical and Physical Description of Hydrometeor Distributions in Colorado Snowstorms Using a Video Disdrometer," *Journal of Applied Meteorology and Climatology - J APPL METEOROL CLIMATOL*, Vol. 46, 2007, pp. 634–650. https://doi.org/10.1175/JAM2489.1.

[6] Brandes, E., Ikeda, K., Thompson, G., and Schoenhuber, M., "Aggregate Terminal Velocity/Temperature Relations," *Journal of Applied Meteorology and Climatology - J APPL METEOROL CLIMATOL*, Vol. 47, 2008, pp. 2729–2736. https://doi.org/10.1175/2008JAMC1869.1.

[7] Zawadzki, I., Jung, E., and Lee, G., "Snow Studies. Part I: A Study of Natural Variability of Snow Terminal Velocity," *Journal of The Atmospheric Sciences - J ATMOS SCI*, Vol. 67, 2010, pp. 1591–1604. https://doi.org/10.1175/2010JAS3342.1.

[8] Garrett, T. J., Yuter, S. E., Fallgatter, C., Shkurko, K., Rhodes, S. R., and Endries, J. L., "Orientations and aspect ratios of falling snow," *Geophysical Research Letters*, Vol. 42, No. 11, 2015, pp. 4617–4622. https://doi.org/https://doi.org/10.1002/2015GL064040, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GL064040.

[9] Tagliavini, G., McCorquodale, M., Westbrook, C., Corso, P., Krol, Q., and Holzner, M., "Drag coefficient prediction of complex-shaped snow particles falling in air beyond the Stokes regime," *International Journal of Multiphase Flow*, Vol. 140, 2021, p. 103652. https://doi.org/https://doi.org/10.1016/j.ijmultiphaseflow.2021.103652, URL https://www.sciencedirect.com/science/article/pii/S0301932221001002.

[10] Kubicek, A., and Wang, P. K., "A numerical study of the flow fields around a typical conical graupel falling at various inclination angles," *Atmospheric Research*, Vol. 118, 2012, pp. 15–26. https://doi.org/https://doi.org/10.1016/j.atmosres.2012.06.001, URL https://www.sciencedirect.com/science/article/pii/S0169809512001664.

[11] Cheng, K.-Y., and Wang, P. K., "A numerical study of the flow fields around falling hails," *Atmospheric Research*, Vol. 132, 2013, pp. 253–263. https://doi.org/10.1016/j.atmosres.2013.05.016.

[12] Wang, P. K., Chueh, C.-C., and Wang, C.-K., "A numerical study of flow fields of lobed hailstones falling in air," *Atmospheric Research*, Vol. 160, 2015, pp. 1–14. https://doi.org/10.1016/j.atmosres.2015.02.013.

[13] Chhabra, R., Agarwal, L., and Sinha, N., "Drag on non-spherical particles: an evaluation of available methods," *Powder Technology*, Vol. 101, No. 3, 1999, pp. 288 – 295. https://doi.org/https://doi.org/10.1016/S0032-5910(98)00178-8, URL http://www.sciencedirect.com/science/article/pii/S0032591098001788.

[14] Hölzer, A., and Sommerfeld, M., "New simple correlation formula for the drag coefficient of non-spherical particles," *Powder Technology - POWDER TECHNOL*, Vol. 184, 2008, pp. 361–365. https://doi.org/10.1016/j.powtec.2007.08.021.

[15] Jaynes, E. T., "Information Theory and Statistical Mechanics," *Phys. Rev.*, Vol. 106, 1957, pp. 620–630. https://doi.org/10.1103/PhysRev.106.620, URL https://link.aps.org/doi/10.1103/PhysRev.106.620.

[16] Sivia, D. S., and Skilling, J., *Data Analysis - A Bayesian Tutorial*, 2nd ed., Oxford Science Publications, Oxford University Press, 2006.

[17] Hastings, W. K., "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, Vol. 57, No. 1, 1970, pp. 97–109. https://doi.org/10.1093/biomet/57.1.97, URL http://biomet.oxfordjournals.org/cgi/content/abstract/57/1/97.

[18] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M., "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, Vol. 101, No. 476, 2006, pp. 1566–1581. URL http://www.jstor.org/stable/27639773.