International Journal of Approximate Reasoning 122 (2020) 24–46

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

On the quantification and efficient propagation of imprecise probabilities with copula dependence

Jiaxin Zhang [a,b], Michael Shields [b,*]

[a] Oak Ridge National Laboratory, Oak Ridge, TN 37831, United States of America
[b] Department of Civil and Systems Engineering, Johns Hopkins University Baltimore, MD 21218, United States of America

ARTICLE INFO

Article history:
Received 5 June 2019
Received in revised form 9 April 2020
Accepted 9 April 2020
Available online 30 April 2020

Keywords:
Imprecise probability
Uncertainty quantification
Copula
Bayesian inference
Small data
Multimodel inference

ABSTRACT

This paper addresses the problem of quantification and propagation of uncertainties associated with dependence modeling when data for characterizing probability models are limited. Practically, the system inputs are often assumed to be mutually independent or correlated by a multivariate Gaussian distribution. However, this subjective assumption may introduce bias in the response estimate if the real dependence structure deviates from this assumption. In this work, we overcome this limitation by introducing a flexible copula dependence model to capture complex dependencies. A hierarchical Bayesian multimodel approach is proposed to quantify uncertainty in dependence model-form and model parameters that result from small data sets. This approach begins by identifying, through Bayesian multimodel inference, a set of candidate marginal models and their corresponding model probabilities, and then estimating the uncertainty in the copula-based dependence structure, which is conditional on the marginals and their parameters. The overall uncertainties integrating marginals and copulas are probabilistically represented by an ensemble of multivariate candidate densities. A novel importance sampling reweighting approach is proposed to efficiently propagate the overall uncertainties through a computational model. Through an example studying the influence of constituent properties on the out-of-plane properties of transversely isotropic E-glass fiber composites, we show that the composite property with copula-based dependence model converges to the true estimate as data set size increases, while an independence or arbitrary Gaussian correlation assumption leads to a biased estimate.

1. Introduction

Uncertainty Quantification (UQ) is widely applied to better understand complex stochastic physical and mathematical systems. Typically, computational simulations aim to estimate statistics of the response of a system subject to random inputs. These inputs are commonly modeled by a random vector $\boldsymbol{X}$ with their joint probability density $f_{\boldsymbol{X}}(\boldsymbol{x})$. The uncertainty associated with the inputs are quantified probabilistically and propagated through a computational model $\mathcal{M}$. The corresponding output $Y = \mathcal{M}(\boldsymbol{X})$ is the quantity of interest (QoI), which is uncertain. If the computational model is deterministic, all uncertainties in Y result from the uncertainty in $\boldsymbol{X}$.

* Corresponding author.
E-mail addresses: zhangj@ornl.gov (J. Zhang), michael.shields@jhu.edu (M. Shields).

https://doi.org/10.1016/j.ijar.2020.04.002
0888-613X/

Practically, the inputs are often assumed to be mutually independent or to possess a multivariate Gaussian dependence structure because it is simple to model and to fit from data. Some conventional UQ approaches, for example, importance sampling [1] and polynomial chaos expansions [2], take advantage of mutually independent inputs. If the inputs are dependent, a number of UQ approaches require to map the model inputs $\boldsymbol{X}$ onto an input $\boldsymbol{X}^*$ with independent components. When $f_{\boldsymbol{X}}(\boldsymbol{x})$ has multivariate Gaussian dependence structure, the map corresponds to the Nataf transformation [3,4]. A more general way that maps the input $\boldsymbol{X}$ onto $\boldsymbol{X}^*$ is the Rosenblatt transformation [5], which needs to know the conditional probability distribution functions (pdfs) that are often infeasible in practice. For this reason, the Gaussian dependence assumption is widely applied in the context of UQ. The Gaussian assumption and the associated dependence provides a convenient representation of the input dependencies, but it may introduce a bias in the response estimate if the real dependence structure deviates from this assumption.

Dependence modeling has recently received widespread attention in the engineering and mathematics communities. This is mainly due to the significant development of copula models [6–8], and vine copulas [9–13] in particular. Copula theory is used to separately model the dependence and the marginal distribution, but it is often limited to low-dimensional problems, typically bivariate or simple copula families, such as Gaussian or Archimedean families [6]. Copula-based approaches have been recently used in various dependence modeling studies, for example in reliability and risk analysis [14–19], sensitivity analysis [20,21], and prognostics and health management (PHM) [22,23]. Copulas also have widespread applications in engineering practice, such as ocean and offshore [24,25], wind [26], and earthquake [27] engineering. To overcome the limitation of copula theory in high dimensions, the vine copula theory was first proposed by Joe [28,29] by formulating multivariate copulas as a product of bivariate copulas among pairs of random variables. Bedford and Cooke [30] introduced a graphical model for describing multivariate copulas using pair-copulas, which provides a flexible and easy interpretation. Czado presented a series of productive studies in the context of vine copulas [31,32] and successfully applied them to financial modeling [33,34]. Recently, vine copula approaches have become increasingly attractive in engineering applications [16,35–38].

Conventionally, the dependence structure of multivariate inputs is built probabilistically through a known joint probability measure. Therefore, the first step of copula-based dependence modeling is to identify or assume a reasonable copula or vine copula model for the input variables. However, it may not be straightforward to identify the appropriate copula model when data characterizing the input parameters are sparse. This process may therefore give rise to a form of *epistemic uncertainty* [39] - which is due to a lack of knowledge or data. Epistemic uncertainty plays an essential role in UQ and must be considered, particularly when it arises from a lack of data.

Many theories have been developed to address the various forms of epistemic uncertainty. It has been argued that epistemic uncertainty needs a different mathematical treatment than *aleatory uncertainty* [40] that are naturally stochastic and treated probabilistically. It remains an open debate as to what that mathematical treatment should be. This desire also has given rise to the field of so-called *imprecise probabilities* wherein epistemic uncertainty contributes a level of "imprecision" and aleatory uncertainty are quantified by classical probability theory. There are numerous approaches to model this imprecision that includes the use of fuzzy sets [41,42] and measures [43], random sets [44–47], intervals and probability boxes [48,40,49,45] and Dempster-Shafer theory [50,51]. Efforts from Walley [52,53] have worked to unify these theories under an over-arching theory of imprecise probabilities. An extensive review of many of these imprecise probabilities approaches for engineering applications can be found in [54].

To the author's knowledge, relatively few studies have accounted for the problem of *imprecise dependence modeling* in UQ. Some recent studies focus on the investigations of Sklar's theorem for imprecise copulas using fuzzy theory [55,56]. Coolen-Maturi et al. [57] combine nonparametric predictive inference that quantifies the uncertainties through imprecise probability with a parametric copula to model and estimate the dependence structure. Among the most comprehensive studies of UQ with dependence modeling is that conducted by Kurowicka and Cooke [58], who discussed UQ in bivariate as well as high dimensional dependence modeling. More recent works include those of Schefzik et al. [59], who propose a general multistage procedure called ensemble copula coupling to quantify the uncertainty in complex simulation models, particularly in weather and climate predictions, and Emiliano et al. [35] who use vine copulas to develop a general data-driven UQ framework for dependence modeling of complex input.

In this paper, we investigate copula-based dependence modeling in the context of imprecise probability that specifically results from a lack of data. This is motivated by the difficulty of data collection under complex conditions, for example, long-time cycle and expensive experiments, in engineering practice. When only scarce data is available, it is a challenging task to assign an objective and accurate probability distribution for the random inputs and precisely estimate their dependence relationship. The developed method builds on the previous work of the authors who proposed information-theoretic [60] and Bayesian [61] multimodel probabilistic methodologies to quantify and efficiently propagate combined aleatory and epistemic uncertainty given small data sets. This work introduces a copula-based dependence modeling framework, which is flexible enough to capture complex dependence structures. To fully quantify the uncertainty in dependence modeling, we propose a hierarchical Bayesian multimodel approach that allows to first identify a set of candidate marginal models and their associated model probabilities, and then estimate the copula model-form and model parameter uncertainties, which are conditioned on the uncertain marginals and their parameters. Using the proposed method, an ensemble of candidate multivariate densities are identified as random inputs that need to be propagated through a complex model to estimate the response of an engineering system. Propagation of these families of densities is particularly difficult because it requires nested Monte Carlo calculations, which are often computationally infeasible even for simple models. This paper proposes a

novel efficient importance sampling reweighting algorithm that allows simultaneous propagation of the multiple densities through one Monte Carlo simulation. The proposed method can further achieve an adaptive updating as additional data are collected but without requiring additional computational evaluation.

This paper is structured as follows. Section 2 provides a brief review of copula-based dependence modeling, particularly bivariate copula theory and vine copula theory. Section 3 presents the uncertainty analysis for copula-based multivariate dependence modeling, including copula uncertainty and marginal uncertainty. An efficient uncertainty propagation with imprecise copula dependence modeling is proposed in Section 4. Section 5 shows an application of the proposed method to the probabilistic prediction of unidirectional composite lamina properties. Some discussions and concluding remarks are given in Section 6.

2. Copula-based modeling of dependence structure

2.1. Measures of statistical dependence

The most well-known measure of dependence between random variables is the Pearson's correlation coefficient, commonly named simply the correlation coefficient, which measures linear dependence. Considering two random variables X and Y with mean values μ_X and μ_Y and standard deviations σ_X and σ_Y, the correlation coefficient $\rho_{X,Y}$ is defined as

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

where $E[\cdot]$ is the expectation and cov is the covariance. All correlation coefficient values are bounded in the interval $[-1, 1]$, indicating the degree of linear dependence between two variables. The closer the coefficient is to either 1 or -1, the stronger the correlation between the variables. If the variables are linearly independent, the correlation coefficient is 0.

Another common measure of dependence is Kendall's τ, or Kendall's rank correlation coefficient, which measures the difference between the concordance and discordance probability and can be used to detect some nonlinear dependence. Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed random vectors, then Kendall's tau is defined as

$$\tau_{X,Y} = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]. \tag{2}$$

Rank correlation can also be expressed using Spearman's ρ (defined as the correlation coefficient – Eq. (1) – between the ranks of the variables) and both Kendall's τ and Spearman's ρ can be shown to be special cases of a generalized rank correlation [62].

However, the information given by a correlation coefficient (Pearson's ρ, Kendall's τ, or Spearman's ρ) is only enough to define the dependence structure between random variables in special cases, e.g. Gaussian random variables. In general, the complete dependence structure requires knowledge of the full joint distribution. One method to capture the complete dependence structure is to model the joint distribution using a copula. In practice, many data structures exhibit different marginal distributions, nonsymmetric/nonlinear dependencies, and/or tail dependencies between variables. These variables cannot be modeled by a Gaussian or multivariate t distribution. This challenge is overcome by the copula approach, which models the dependencies and marginal distributions separately.

2.2. Copula theory

Consider $F_{\boldsymbol{X}}(\boldsymbol{x})$ as the d-dimensional joint distribution function of the random vector $\boldsymbol{X} = (X_1, ..., X_d)^T$ with marginal distributions $F_1(x_1), ..., F_d(x_d)$. According to Sklar's theorem [63], there exists a copula C such that for all $\boldsymbol{x} = (x_1, ..., x_d)^T \in [-\infty, \infty]^d$,

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = C(F_1(x_1), ..., F_d(x_d)) \tag{3}$$

If $F_1(x_1), ..., F_d(x_d)$ are continuous, the copula C is unique. The copula C can be interpreted as the joint distribution function of a d-dimensional random vector on $[0, 1]^d$ with uniform marginals.

Sklar's theorem can also be restated with respect to probability densities. The corresponding copula density can be expressed as:

$$c(F_1(x_1), ..., F_d(x_d)) = \frac{\partial C(F_1(x_1), ..., F_d(x_d))}{\partial F_1(x_1), .., \partial F_d(x_d)} \tag{4}$$

which implies the joint multivariate pdf can be formulated by

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = c(F_1(x_1), ..., F_d(x_d)) \cdot f_1(x_1) \cdots f_d(x_d) \tag{5}$$

where $f_k(x_k), 1 \le k \le d$ are the marginal pdfs. For the bivariate case, Joe [29] and Nelsen [6] provided a rich variety of copula families from the two major classes of *Elliptical* and *Archimedean* copulas. Elliptical copulas are directly derived by

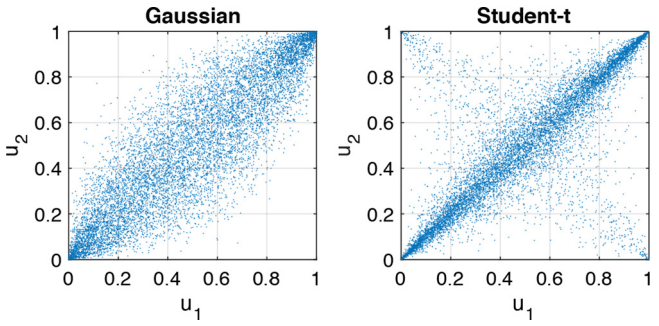

Fig. 1. Elliptical copula family. Samples drawn from (left) Gaussian copula and (right) Student-t copula.

Table 1
Properties and definition of elliptical copula families.

Elliptical family	Parameter range	Kendall's τ	Tail dependence
Gaussian	$\rho \in (-1, 1)$	$\frac{2}{\pi}\arcsin(\rho)$	0
Student-t	$\rho \in (-1, 1), \nu > 2$	$\frac{2}{\pi}\arcsin(\rho)$	$2t_{\nu+1}(-\sqrt{\nu+1}\sqrt{\frac{1-\rho}{1+\rho}})$

inverting Sklar's theorem, shown in Eq. (3). Given a bivariate cumulative distribution function $F_{\boldsymbol{X}}(\boldsymbol{x})$ with marginals $F_1(x_1)$ and $F_2(x_2)$, then

$$C(u_1, u_2) = F(F_1^{-1}(u_1), F_2^{-1}(u_2)) \tag{6}$$

is a bivariate copula for $u_1, u_2 \in [0, 1]$. One of the most commonly used bivariate elliptical copula is the bivariate Gaussian copula

$$C(u_1, u_2) = \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \tag{7}$$

where Φ_ρ is the joint cumulative distribution of bivariate standard normal distribution with correlation coefficient ρ and Φ^{-1} is the inverse standard normal cdf.

Another common copula is the Student-t copula, whose bivariate density is given by

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\Gamma(\frac{\nu+2}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\pi\nu)^2|\boldsymbol{\Sigma}|}}\left(1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+2}{2}} \tag{8}$$

where ν is the number of degrees of freedom, $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is a positive-definite matrix. Since the copula remains invariant under a standardization of the marginal distributions, the copula of a $t(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is identical to that of a $t(\nu, 0, \boldsymbol{P})$ distribution where $\boldsymbol{P}$ is the correlation matrix implied by the dispersion matrix $\boldsymbol{\Sigma}$ [64]. Thus, the corresponding Student-t copula is given by

$$C(u_1, u_2) = \int_{-\infty}^{t_\nu^{-1}(u_1)} \int_{-\infty}^{t_\nu^{-1}(u_2)} \frac{\Gamma(\frac{\nu+2}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\pi\nu)^2|\boldsymbol{P}|}}\left(1 + \frac{\boldsymbol{x}'\boldsymbol{P}^{-1}\boldsymbol{x}}{\nu}\right)^{-\frac{\nu+2}{2}} d\boldsymbol{x}. \tag{9}$$

For bivariate case, we simplify the notation to

$$C(u_1, u_2) = t_{\rho,\nu}(t_\nu^{-1}(u_1), t_\nu^{-1}(u_2)) \tag{10}$$

where ρ is the off-diagonal element of $\boldsymbol{P}$ [64], t_ν^{-1} is defined as the inverse Student-t marginal distribution function with ν degrees of freedom. Fig. 1 shows samples from the elliptical copula family with Gaussian and Student-t copulas. Table 1 provides the basic properties of the Gaussian and Student-t copulas.

Another important copula family, Archimedean copulas are defined as

$$C(u_1, u_2) = \psi^{[-1]}(\psi(u_1) + \psi(u_2)) \tag{11}$$

where ψ is the generator function of the copula C, which is a continuous strictly decreasing convex function which satisfies $\psi(1) = 0$ and $\psi^{[-1]}$ is defined as

$$\psi^{[-1]}(t) = \begin{cases} \psi^{-1}(t), & 0 \le t \le \psi(0) \\ 0, & \psi(0) \le t \le \infty \end{cases} \tag{12}$$

Table 2
Definitions of Archimedean copula families.

Name of Copula	Bivariate copula $C_\theta(u_1, u_2)$	Parameter θ
Clayton	$\left[\max\left\{u_1^{-\theta} + u_2^{-\theta} - 1, 0\right\}\right]^{-1/\theta}$	$\theta \in [-1, \infty) \setminus \{0\}$
Frank	$-\frac{1}{\theta}\log\left[1 + \frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right]$	$\theta \in \mathbb{R} \setminus \{0\}$
Gumbel	$e^{-\left((-\log(u_1))^\theta + (-\log(u_2))^\theta\right)^{1/\theta}}$	$\theta \in [1, \infty)$

Table 3
Properties of Archimedean copula families.

Name of Copula	Generator	Kendall's τ
Clayton	$\frac{1}{\theta}(t^{-\theta} - 1)$	$\frac{\theta}{\theta+2}$
Frank	$-\log[\frac{e^{-\theta t}-1}{e^{-\theta}-1}]$	$1 - \frac{4}{\theta} + 4\frac{D_1(\theta)}{\theta}$
Gumbel	$(-\log t)^\theta$	$1 - \frac{1}{\theta}$

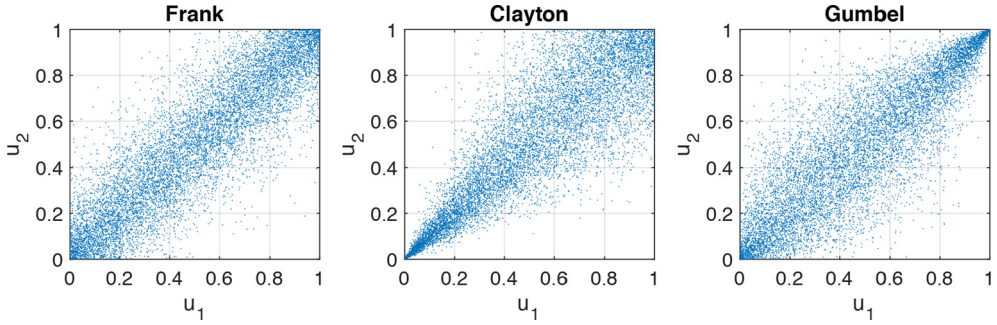

Fig. 2. Archimedean copula family. Samples drawn from (left) Frank copula, (middle) Clayton copula and (right) Gumbel copula.

The most common single parameter Archimedean copulas are the Clayton, Gumbel and Frank [6]. Their bivariate copula formulations are shown in Table 2, with their corresponding properties (generator and Kendall's τ) shown in Table 3 where $D_1(\theta) = \frac{1}{\theta}\int_0^\theta \frac{t}{e^t-1}dt$ is the Debye function [29,6]. Fig. 2 shows examples of samples drawn from these copulas for two random variables u_1 and u_2.

2.3. Vine copulas

Copula families perform well in the bivariate case, but in arbitrarily high dimension the choice of adequate copula families is very limited. Elliptical families and Archimedean copulas lack the flexibility to accurately model the dependence structure of high dimensional variables. Simple extensions of these bivariate families offer some improvement, but typically become intricate and introduce additional limitations that, for example, they can not be applied to establish a distribution consistent with arbitrary correlation [65].

Vine copulas (also called tree structures) do not suffer from these issues and have been widely used in many fields of application. Bedford and Cooke [30] introduced a graphical model for describing multivariate copulas using a cascade of bivariate copulas, denoted by *pair-copulas*. This pair-copula construction provides a flexible way to decompose a multivariate probability density into bivariate copulas such that each pair-copula is independent of the others.

Consider a d-dimensional joint density function $f_{\boldsymbol{X}}(x_1, ..., x_d)$ for a random vector $\boldsymbol{X} = (X_1, ..., X_d)$. This density can be decomposed based on the law of total probability

$$f(x_1, ..., x_d) = f_n(x_d) \cdot f(x_{d-1}|x_d) \cdot f(x_{d-2}|x_{d-1}, x_d) \cdots f(x_1|x_2, ..., x_d). \tag{13}$$

From Sklar's theorem, we also know the joint probability density can be formulated as shown in Eq. (5). In the bivariate case, Eq. (5) simplifies to

$$f(x_1, x_2) = c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \cdot f_2(x_2) \tag{14}$$

where c_{12} is the appropriate *pair-copula density* for the pair of transformed variables $F_1(x_1)$ and $F_2(x_2)$. It is straightforward to write a conditional density

$$f(x_1|x_2) = c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \tag{15}$$

in terms of the pair-copula. Similarly, it easily follows for three random variables X_1, X_2 and X_3 as follows

$$f(x_1|x_2,x_3) = c_{12|3}(F(x_1|x_3), F(x_2|x_3)) \cdot f(x_1|x_3) \tag{16}$$

for the appropriate pair-copula $c_{12|3}$ which is used for the transformed variables $F(x_1|x_3)$ and $F(x_2|x_3)$. An alternative decomposition is

$$f(x_1|x_2,x_3) = c_{13|2}(F(x_1|x_2), F(x_3|x_2)) \cdot f(x_1|x_2) \tag{17}$$

where $c_{13|2}$ differs from the pair-copula in Eq. (16). We can further decompose $f(x_1|x_2)$ in Eq. (17) based on Eq. (15)

$$f(x_1|x_2,x_3) = c_{13|2}(F(x_1|x_2), F(x_3|x_2)) \cdot c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1). \tag{18}$$

By extension, the conditional marginal can be decomposed into the appropriate pair-copula using the general form given by [10,31]

$$f(x|\boldsymbol{v}) = c_{xv_j|\boldsymbol{v}_{-j}}(F(x|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j}))f(x|\boldsymbol{v}_{-j}) \tag{19}$$

where v_j is an arbitrarily excluded element from vector $\boldsymbol{v}$ and $\boldsymbol{v}_{-j}$ denotes the vector $\boldsymbol{v}$ after excluding v_j. Hence, a multivariate density $f_X(\boldsymbol{x})$ can be expressed as a product of bivariate copula density functions with marginal conditional CDFs in the form of $F(x|\boldsymbol{v})$ that can be formulated recursively as follows [29]

$$F(x|\boldsymbol{v}) = \frac{\partial C_{x,v_j|\boldsymbol{v}_{-j}}(F(x|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j}))}{\partial F(v_j|\boldsymbol{v}_{-j})} \tag{20}$$

where $C_{x,v_j|\boldsymbol{v}_{-j}}$ is a bivariate copula distribution function.

Note that a d-dimensional multivariable density can be factorized into a number of different conditional pair-copulas based on the vine copula construction proposed by Bedford and Cooke [30]. Except regular vine structure (R-vine), there are two special types of regular vines: canonical vine (C-vine) and drawable vine (D-vine). For the C-vine, each tree has a unique node that is connected to all other nodes, and the corresponding joint pdf $f_X(\boldsymbol{x})$ is

$$f_X(\boldsymbol{x}) = \prod_{k=1}^{d} f_k(x_k) \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c(F(x_j|x_1,\ldots,x_{j-1}), F(x_{j+i}|x_1,\ldots,x_{j-1})). \tag{21}$$

In contrast, each tree in a D-vine is a path and the corresponding joint pdf $f_X(\boldsymbol{x})$ is

$$f_X(\boldsymbol{x}) = \prod_{k=1}^{d} f_k(x_k) \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c(F(x_i|x_{i+1},\ldots,x_{i+j-1}), F(x_{i+j}|x_{i+1},\ldots,x_{i+j-1})) \tag{22}$$

where the subscript indices indicate the conditional random variables to be drawn.

Copula theory and vine copulas are an important tool for modeling the dependence of multivariate densities in either low or high dimension. A following critical question is how to select and estimate all components of a bivariate copula model or tree structure model from limited data. The paper mainly focuses on the bivariate copula model to show how to efficiently quantify the uncertainties associated with copula model selection and the corresponding parameters. The proposed method can be extended to high dimensional problems with dependence given a specified vine copula structure. The next sections discuss this issue in detail.

3. Statistical inference of copula dependence models

Given a d-dimensional probability density, we can decompose it into products of marginal densities and bivariate copula densities and represent this decomposition with a nested set of trees that fulfill a proximity condition. However, it is often difficult to directly identify a d-dimensional probability density. Instead, more commonly, only data are provided and statistical inference is necessary for model selection and parameter estimation. Small data sets create additional uncertainties which pose a significant challenge to the inference of the copula dependence model.

Assuming known marginal distributions, copula dependence modeling consists of three principal components: tree structure, copula form and copula parameters. However, for small data sets, uncertainty in the marginals cannot be ignored. Consequently, the marginal form and marginal distribution parameters must also be included in the inference process. As a result, the total uncertainty when inferring joint probability model form, U_{all}, includes the following five components:

$$U_{all} = \left\{U_t, U_{cf}, U_{cp}, U_{mf}, U_{mp}\right\} \tag{23}$$

where U_t is uncertainty in the tree structure, U_{cf} and U_{cp} are the uncertainty in copula families and parameters respectively, and U_{mf} and U_{mp} represent the uncertainty in marginal distribution families and parameters. To quantify these uncertainties, statistical methods are adopted for model selection and parameter estimation.

The model uncertainty in tree structure is particularly challenging to address. This is mainly because the possible decomposition of pair-copulas is potentially large, especially in high dimensions. Typically, the tree structure is assumed to follow a specified model based on the analyst's knowledge or experience. There are several model selection approaches for the specification of tree structures, including optimal C-vine structure selection [66], Bayesian approaches for D-vine selection [67] and maximum spanning trees for R-vines [33]. Here, the tree model selection is not our first priority, so we do not elaborate on these methods. Instead, our emphasis is on how to efficiently quantify the uncertainties associated with copula form selection and the corresponding parameters given a specified vine copula structure.

3.1. Copula form selection and parameter estimation

When a specific vine copula structure is determined, classical statistical approaches, including goodness-of-fit tests [68], independence test [69] and AIC/BIC [70] are capable of handling copula form selection when data sets are large. When both tree structure and copula form are known and the data set is large, the copula parameters can be estimated using sequential estimation [10,66], maximum likelihood estimation [71], or Bayesian parameter estimation [67,72]. However, these classical approaches fall short when inferring from small data sets.

Traditionally, statistical inference is applied to select a single "best" model given a set of candidate models and available data, and the model is the sole model used for probabilistic modeling. Any uncertainty associated with model selection is simply ignored. However, it is often difficult (even impossible) to identify a unique best model without significant (and potentially problematic) assumptions. Consequently, it is necessary to consider model uncertainty and compare the validity of multiple candidate models – a process referred to as multimodel inference, as introduced by Burnham and Anderson [70]. In this study, we generalize the Bayesian multimodel inference developed previously by the authors [60,61] to include uncertainty in the form and parameters of the copula dependence model.

Given a data set $\boldsymbol{d}$, the model selection problem is to identify the model M_i that "best" fits the data from a collection of N candidate models $\mathbb{M} = \{M_j\},\ \ j = 1, \ldots, N$. The notion of best fit varies depending on the selected metric. In the Bayesian setting used here, initial model prior probabilities $\tilde{\pi}_j = p(M_j)$ with $\sum_{j=1}^{N} \tilde{\pi}_j = 1$ are assigned to each model $M_j \in \mathbb{M}$. According to Bayes' rule, the posterior model probability, given the data $\boldsymbol{d}$ can be calculated by

$$\pi_j = p(M_j|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|M_j)p(M_j)}{\sum_{k=1}^{N} p(\boldsymbol{d}|M_k)p(M_k)}, \quad j = 1, \ldots, N \tag{24}$$

having $\sum_{j=1}^{N} \pi_j = 1$ and where

$$p(\boldsymbol{d}|M_j) = \int_{\boldsymbol{\theta}_j} p(\boldsymbol{d}|\boldsymbol{\theta}_j, M_j)p(\boldsymbol{\theta}_j|M_j)d\boldsymbol{\theta}_j, \quad j = 1, \ldots, N \tag{25}$$

is referred as to the marginal likelihood or evidence of model M_j.

Commonly, the model $M^* \in \mathbb{M}$ with the highest posterior model probability $p(M^*|\boldsymbol{d})$ is selected as the single "best" model. By contrast, Bayesian multimodel inference ranks the candidate models by their posterior model probabilities calculated by Eq. (24) and retains all plausible models with non-negligible probability. Once the plausible models and their associated model probabilities have been identified, model parameter uncertainties are assessed by applying Bayesian parameter estimation. For each model in the set of plausible models, $M_i,\ \ i = 1, \ldots, N_d\ \ (N_d \leq N)$, we begin by assigning a prior (often a noninformative prior) to the model parameters $\boldsymbol{\theta}_i$, denoted $p(\boldsymbol{\theta}_i|M_i)$. We then estimate the posterior parameter distribution using Bayes' rule:

$$p(\boldsymbol{\theta}_i|\boldsymbol{d}, M_i) = \frac{p(\boldsymbol{d}|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i|M_i)}{p(\boldsymbol{d}|M_i)} \propto p(\boldsymbol{d}|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i|M_i), \quad i = 1, \ldots, m \tag{26}$$

where $p(\boldsymbol{d}|\boldsymbol{\theta}_i, M_i)$ is the likelihood function. The posterior $p(\boldsymbol{\theta}_i|\boldsymbol{d}, M_i)$ is identified implicitly through Markov Chain Monte Carlo (MCMC) without requiring the calculation of model evidence $p(\boldsymbol{d}|M_i)$. However, the evidence, as evident from Eq. (25) is critical in Bayesian multimodel inference and needs to be calculated with caution. A detailed discussion of the evidence calculation can be found in [61].

In the classical setting, a unique set of model parameters $\boldsymbol{\theta}_i$ is identified from the posterior samples using, for example, the maximum a posterior (MAP) estimator,

$$\tilde{\boldsymbol{\theta}}_j^{\text{MAP}}(\boldsymbol{d}, M_j) = \arg\max_{\boldsymbol{\theta}_j} p(\boldsymbol{\theta}_j|\boldsymbol{d}, M_j) = \arg\max_{\boldsymbol{\theta}_j} p(\boldsymbol{d}|\boldsymbol{\theta}_j, M_j)p(\boldsymbol{\theta}_j|M_j). \tag{27}$$

When $p(\boldsymbol{\theta}_j|M_j)$ is a noninformative prior, the MAP estimator is equivalent to the maximum likelihood estimate (MLE). Due to a lack of data, the posterior parameter probability will likely possess a large variance. Rather than discarding the full uncertainty by selecting a single set of MLE or MAP parameters or integrating out its variability using Bayesian model averaging [73], we retain the full posterior densities for each plausible model.

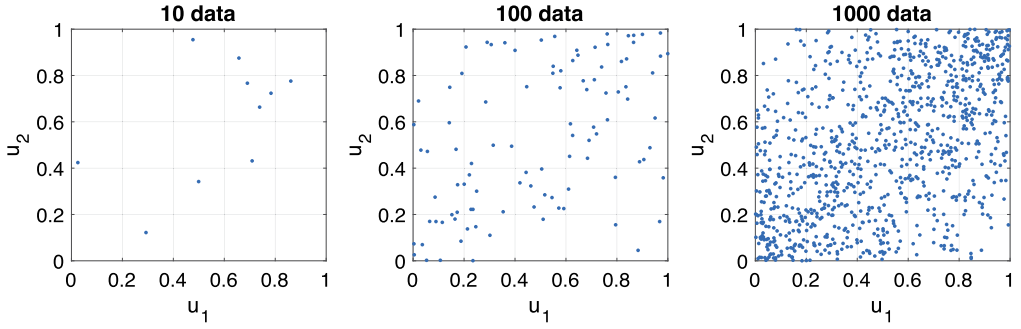

Fig. 3. Bivariate correlated data drawn from `Frank`(3) copula model, showing 10 data, 100 data and 1000 data.

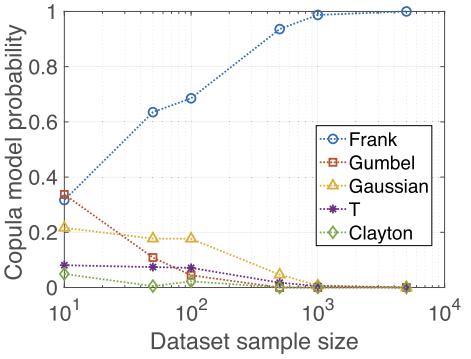

Fig. 4. Posterior copula model probability as a function of dataset size.

In this work, the Bayesian multimodel inference method is generalized to address copula dependence model selection and parameter estimation. A simple bivariate example is used to illustrate the process and its performance. Consider a bivariate random vector $\boldsymbol{u} = [u_1, u_2]$ whose dependence follows the Frank copula model with parameter $\theta = 3$ (denoted `Frank`(3)). Fig. 3 shows data sets of varying size drawn from the `Frank`(3) copula. Notice that, given only 10 data, one cannot decipher a clear dependence relation. Only after 100 data are drawn does the dependence begin to emerge and it finally becomes clear when 1000 points are drawn.

From these data, Bayesian multimodel inference is first used to quantify the copula form uncertainty. Five copula models – the Gaussian, Student-t, Clayton, Gumbel and Frank copulas – are selected as the candidate copula forms. Without an informative prior, all candidate copula models are assumed to have equal probability. The Monte Carlo method is adopted to compute the evidence from Eq. (25). Then the posterior copula model probabilities are obtained using Eq. (24). Fig. 4 shows the posterior probabilities for each candidate copula model as a function of dataset size. Notice that the model probability for the Frank copula becomes gradually larger as the data set size increases but the Bayesian multimodel inference does not select the correct Frank copula model conclusively until 1000 correlated data are collected.

Next, Bayesian inference is employed to estimate the copula parameter for each plausible candidate model. Fig. 5 shows the posterior probability distribution for the Frank copula parameter θ for increasing data set size. Note that the posterior variance is large when the data set size is small and the estimate gradually narrows with increasing data set size. Finally, the posterior density with 1000 data converges towards a narrow distribution that includes the true value ($\theta = 3$).

This simple example illustrates the Bayesian multimodel inference process for model selection and parameter estimation of copula dependence modeling. More specifically, it illustrates the fact that inference is inherently imprecise from small data sets. When data sets are small, it is impossible to uniquely identify the copula form (and the associated copula model parameters) from which the data are drawn. In the following section, we turn our attention to uncertainty in the marginal distributions.

3.2. Uncertainty in marginal distributions

As observed in authors' previous studies [60,61,74], uncertainty in the marginal distributions play a critical role in uncertainty quantification from small datasets. Consider again for simplicity, the bivariate case where the joint pdf can be expressed as:

$$f_X(x_1, x_2) = c_{12}(F_1(x_1, \theta_1), F_2(x_2, \theta_2), \theta_c) \cdot f_1(x_1, \theta_1) \cdot f_2(x_2, \theta_2) \tag{28}$$

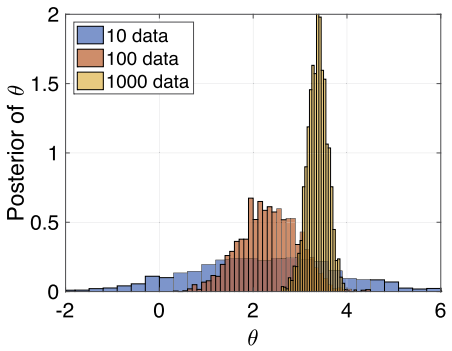

Fig. 5. Posterior histogram of the Frank copula model parameter given different data set sizes.

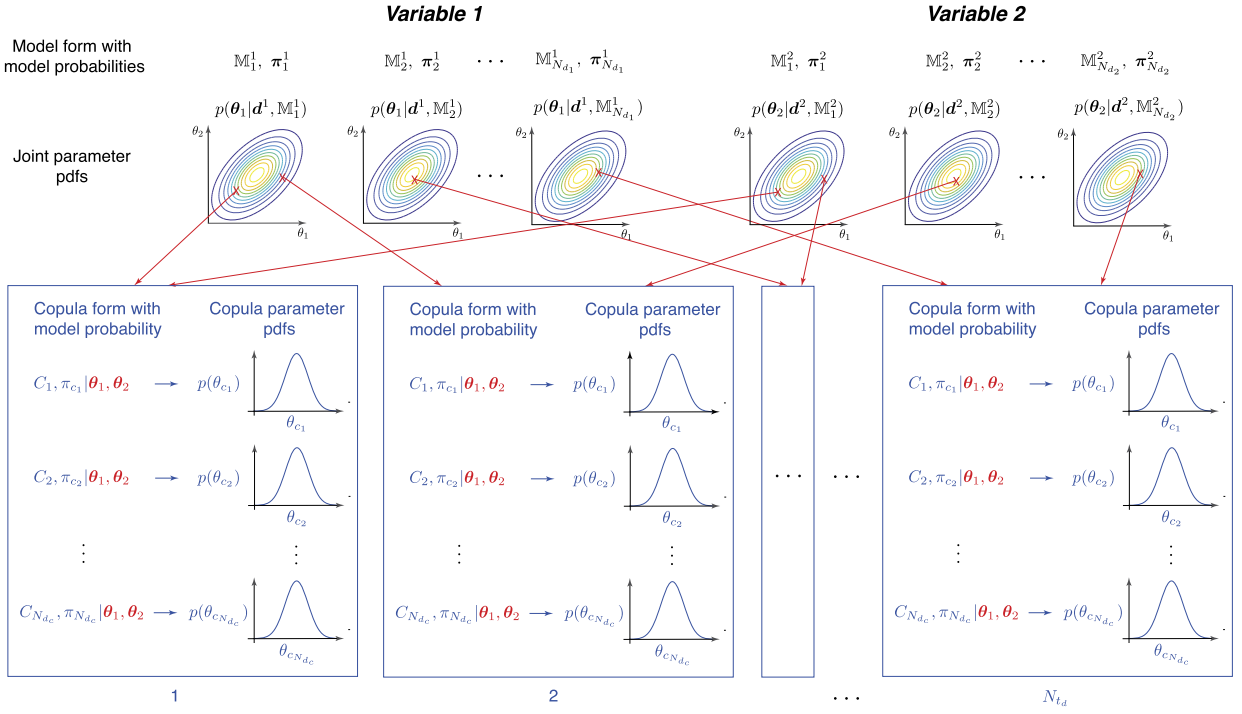

Fig. 6. Hierarchy of Bayesian multimodel inference for copulas and marginals.

where θ_c are the copula parameters. Given this expression of the joint density, it is clear that the copula model is conditional on the marginals and their parameters, which the previous studies have shown to have very large uncertainties when data sets are small. Consequently, it is necessary to identify copula model probabilities and copula parameter probabilities for each set of inferred candidate marginals. This induces a hierarchy of probabilities that includes both the copula model and the marginal model. We therefore propose a hierarchical Bayesian multimodel inference method, as illustrated in Fig. 6. The procedure is summarized for each pair of variables as follows:

- Step 1: Marginal multimodel inference – First identify the candidate marginal model sets $\mathbb{M}^1 = \{M_j^1\}$, $j = 1, \ldots, N_{d_1}$ and $\mathbb{M}^2 = \{M_j^2\}$, $j = 1, \ldots, N_{d_2}$ for each variable and compute the marginal model probabilities $\boldsymbol{\pi}^1 = \{\pi_1^1, \pi_2^1, \ldots, \pi_{N_{d_1}}^1\}$ and $\boldsymbol{\pi}^2 = \{\pi_1^2, \pi_2^2, \ldots, \pi_{N_{d_2}}^2\}$ using Eq. (24). Notice that this induces a set of $N_{d_1} \times N_{d_2}$ possible marginal pairs. Then estimate the posterior joint pdf for the marginal parameters for all plausible models, $p(\boldsymbol{\theta}_j^1 | \boldsymbol{d}^1, M_j^1)$, $j = 1, \cdots, N_{d_1}$ and $p(\boldsymbol{\theta}_j^2 | \boldsymbol{d}^2, M_j^2)$, $j = 1, \cdots, N_{d_2}$ using Eq. (26).
- Step 2: Define a finite set of marginal distributions – Theoretically, the above process yields an infinite set of parameterized probability models. Practically, it is necessary to reduce this to a finite but statistically representative set of N_{t_d} marginal probability model pairs. This is achieved by randomly selecting a model family for each variable from $\mathbb{M}^1$

and $\mathbb{M}^2$ with probabilities $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$ respectively, and randomly selecting the parameters of each model from the appropriate posterior joint pdf $p(\boldsymbol{\theta}_1|\boldsymbol{d}^1, M_j^1)$ and $p(\boldsymbol{\theta}_2|\boldsymbol{d}^2, M_k^2)$.

- Step 3: Copula multimodel inference – For each pair of marginal distributions $f_1(x_1|\boldsymbol{\theta}_1, M_j^1)$ and $f_2(x_2|\boldsymbol{\theta}_2, M_k^2)$, standardize the data using $F_1(\boldsymbol{d}^1)$ and $F_2(\boldsymbol{d}^2)$. Compute the posterior copula model probabilities $\boldsymbol{\pi}_c = \left\{\pi_{c_1}, \cdots, \pi_{c_{N_{d_c}}}\right\}$ for each candidate copula model $\left\{C_1, \cdots, C_{N_{d_c}}\right\}$ using Eq. (24) where N_{d_c} is the number of plausible copula models for the specified marginal pair. Next, estimate the posterior pdf for the copula parameters for each plausible copula model, $p(\boldsymbol{\theta}_{c_k}|\boldsymbol{d}, C_k)$, $k = 1, \ldots, N_{d_c}$ using Eq. (26). As in step 2, a finite set of N_{t_c} (N_{t_c} can be arbitrarily large) copulas (copula models and parameters) are determined for each marginal pair $\left\{f_1(x_1|\boldsymbol{\theta}_1, M_j^1), f_2(x_2|\boldsymbol{\theta}_2, M_k^2)\right\}$.
- Step 4: Identify bivariate joint densities – Combine the set of marginal densities and copula densities to define the full set of candidate joint densities $f_X(x_1, x_2)$, as in Eq. (28). This, however, may lead to a prohibitively large number, $N_{t_d} \times N_{t_c}$, of candidate bivariate densities. In the following section, we discuss a strategy to keep this number tractable.

The result is a set of $N_{t_d} \times N_{t_c}$ joint distributions that are representative of the uncertainty in marginal model form, marginal parameters, copula model form, and copula parameters. We now consider how to propagate this set of joint distributions through a computational model. Note that the cost of propagation depends only weakly on $N_{t_d} \times N_{t_c}$, the number of joint densities in the set. That is, increasing $N_{t_d} \times N_{t_c}$ does not increase the number of model evaluations necessary for uncertainty propagation. Therefore, it is advantageous to make $N_{t_d} \times N_{t_c}$ as large as possible, as undersampling it will result in artificially narrow uncertainty bounds.

4. Uncertainty propagation with copula dependence modeling

In the previous study [60], we proposed an efficient algorithm for propagation of the imprecise probabilities characterized by a multimodel set with independent marginals. Here, we extend this algorithm to the propagation of imprecise probabilities with copula dependence modeling. For illustration, and without loss of generality, we derive here the propagation method for bivariate random variables. Its extension to higher-dimensional vectors with copula dependence, particularly vine copulas that rely on a series of bivariate copulas, follows naturally.

4.1. Importance sampling for bivariate joint probability density

Consider the performance function $g(\boldsymbol{X}_1, \boldsymbol{X}_2)$ defining the response quantity of interest for a mathematical or physical system. The aim of uncertainty propagation is to evaluate the expectation $E(g(\boldsymbol{X}_1, \boldsymbol{X}_2))$ where $(\boldsymbol{X}_1, \boldsymbol{X}_2) \in \Omega$ is a random vector having bivariate joint probability density $p(\boldsymbol{x}_1, \boldsymbol{x}_2)$. The classical Monte Carlo estimator is computed as follows:

$$\mu = E_p[g(\boldsymbol{X}_1, \boldsymbol{X}_2)] = \int_\Omega g(\boldsymbol{x}_1, \boldsymbol{x}_2) p(\boldsymbol{x}_1, \boldsymbol{x}_2) d\mathbf{x} \approx \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{x}_1^i, \boldsymbol{x}_2^i) \tag{29}$$

where $E_p[\cdot]$ is the expectation with respect to $p(\cdot)$ and $(\boldsymbol{x}_1^i, \boldsymbol{x}_2^i)$ are bivariate random samples drawn from $p(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Importance sampling allows samples to be drawn from an alternate density $q(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and then reweights the samples to obtain the estimator. The Monte Carlo estimator in Eq. (29) is modified as:

$$\begin{aligned} \mu = E_q\left[g(\boldsymbol{X}_1, \boldsymbol{X}_2)\frac{p(\boldsymbol{X}_1, \boldsymbol{X}_2)}{q(\boldsymbol{X}_1, \boldsymbol{X}_2)}\right] &= \int_\Omega g(\boldsymbol{x}_1, \boldsymbol{x}_2)\frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2)}{q(\boldsymbol{x}_1, \boldsymbol{x}_2)} q(\boldsymbol{x}_1, \boldsymbol{x}_2) d\mathbf{x} \\ &\approx \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_1^i, \boldsymbol{x}_2^i) w(\boldsymbol{x}_1^i, \boldsymbol{x}_2^i) \end{aligned} \tag{30}$$

where $E_q[\cdot]$ denotes expectation for $(\boldsymbol{X}_1, \boldsymbol{X}_2) \sim q(\cdot)$ and the importance weights are defined as:

$$w(\boldsymbol{x}_1^i, \boldsymbol{x}_2^i) = \frac{p(\boldsymbol{x}_1^i, \boldsymbol{x}_2^i)}{q(\boldsymbol{x}_1^i, \boldsymbol{x}_2^i)}. \tag{31}$$

4.2. Optimal important density for bivariate joint probability density with copula dependence: derivation

The efficient propagation of multimodel imprecise probabilities is performed by identifying an "optimal" importance sampling density, propagating this optimal density, and reweighting the samples according to each distribution in the multimodel set. The optimal sampling density is derived as the distribution that "best" matches the multimodel distribution set according to some metric. In the prior work, the authors [60] derive an explicit analytical optimal importance sampling

density given an ensemble of target marginal probability densities that minimizes the total expected mean square difference, $\mathcal{M}(\mathbb{M} \parallel Q)$, between the model set $\mathbb{M} = \{M_j\}, j = 1, \ldots, N_d$ and the importance sampling density $Q = q(\boldsymbol{x})$ given by:

$$\mathcal{E} = \sum_{j=1}^{N_d} E_\theta \left[\mathcal{M}(M_j \parallel Q)\right] = E_\theta \left[\sum_{j=1}^{N_d} \frac{1}{2} \int \left(p_j(\boldsymbol{x}|\boldsymbol{\theta}) - q(\boldsymbol{x})\right)^2 d\boldsymbol{x}\right]. \tag{32}$$

In other words, the following optimization problem is solved:

$$\begin{aligned} &\underset{q}{\text{minimize}} \quad \mathcal{L}(q) = E_\theta \left[\int \mathcal{F}(\boldsymbol{x}, \boldsymbol{\theta}, q(\boldsymbol{x})) d\boldsymbol{x}\right] \\ &\text{subject to} \quad \mathcal{I}(q) = \int q(\boldsymbol{x}) d\boldsymbol{x} - 1 = 0 \end{aligned} \tag{33}$$

where the action functional $\mathcal{F}(\cdot)$ is the total square differences:

$$\mathcal{F}(\boldsymbol{x}, \boldsymbol{\theta}, q(\boldsymbol{x})) = \frac{1}{2} \sum_{j=1}^{N_d} \left(p_j(\boldsymbol{x}|\boldsymbol{\theta}) - q(\boldsymbol{x})\right)^2 \tag{34}$$

and E_θ is the expectation with respect to the posterior probability of the model parameters $\boldsymbol{\theta}$. $\mathcal{I}(q)$ ensures that $q(\boldsymbol{x})$ is a valid pdf. Solving this optimization problem yields a closed-form solution given by the convex mixture model [60]

$$q^*(\boldsymbol{x}) = \frac{1}{N_d} \sum_{j=1}^{N_d} E_\theta \left[p_j(\boldsymbol{x}|\boldsymbol{\theta})\right] \tag{35}$$

When the posterior model probabilities are not equal, this solution generalizes as

$$q^*(\boldsymbol{x}) = \sum_{j=1}^{N_d} \pi_j E_\theta \left[p_j(\boldsymbol{x}|\boldsymbol{\theta})\right] \tag{36}$$

where each term is weighted by the corresponding posterior model probabilities π_j computed by Eq. (24). The interested reader can find more details in [60].

It is straightforward to generalize this solution from the one-dimensional probability density to multivariate joint probability densities. If the bivariate joint probability density has independent marginals, the optimal sampling density is expressed as:

$$q^*(\boldsymbol{x}) = \frac{1}{N_{d_1} N_{d_2}} \sum_{i=1}^{N_{d_1}} \sum_{j=1}^{N_{d_2}} E_\theta \left[p_{ij}(\boldsymbol{x}|\boldsymbol{\theta})\right] \tag{37}$$

and the bivariate joint probability density $p_{ij}(\boldsymbol{x}|\boldsymbol{\theta})$ can be decomposed by marginal distribution $f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1)$ and $f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)$ as follows:

$$p_{ij}(\boldsymbol{x}|\boldsymbol{\theta}) = f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1) \cdot f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2) \tag{38}$$

where N_{d_1} and N_{d_2} are the number of candidate probability models for the marginal densities respectively and $N_d = N_{d_1} \cdot N_{d_2}$ is the total number of candidate probability models for the bivariate joint probability density. Thus, the optimal sampling density for independent bivariate joint density can be expanded in terms of the marginals as:

$$\begin{aligned} q^*(\boldsymbol{x}) &= \frac{1}{N_{d_1} N_{d_2}} \sum_{i=1}^{N_{d_1}} \sum_{j=1}^{N_{d_2}} E_\theta \left[f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1) f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)\right] \\ &= \frac{1}{N_{d_1} N_{d_2}} \sum_{i=1}^{N_{d_1}} \sum_{j=1}^{N_{d_2}} E_{\theta_1} \left[f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1)\right] E_{\theta_2} \left[f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)\right] \\ &= \frac{1}{N_{d_1} N_{d_2}} \sum_{i=1}^{N_{d_1}} E_{\theta_1} \left[f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1)\right] \sum_{j=1}^{N_{d_2}} E_{\theta_2} \left[f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)\right] \end{aligned} \tag{39}$$

Again, it is straightforward to show that this solution generalizes for unequal model probabilities as:

$$q^*(\boldsymbol{x}) = \sum_{i=1}^{N_{d_1}} \pi_i^1 E_{\theta_1}\left[f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1)\right] \sum_{j=1}^{N_{d_2}} \pi_j^2 E_{\theta_2}\left[f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)\right] \tag{40}$$

where π_i^1 associated with marginal density $f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1)$ is the posterior model probability for model M_i satisfying $\sum_{i=1}^{N_{d_1}} \pi_i^1 = 1$ and π_j^2 associated with marginal density $f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)$ is the posterior model probability for model M_j satisfying $\sum_{j=1}^{N_{d_2}} \pi_j^2 = 1$.

If the bivariate joint probability density has copula dependence, with copula density $c_{12}^k(F_1(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2(\boldsymbol{x}_2|\boldsymbol{\theta}_2)|\boldsymbol{\theta}_c)$, we can express the bivariate joint probability density as:

$$p_{ij}^k(\boldsymbol{x}|\boldsymbol{\theta}) = c_{12}^k(F_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)|\boldsymbol{\theta}_c) \cdot f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1) \cdot f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2) \tag{41}$$

where $k = 1, \ldots, N_{d_c}$ indexes the candidate copula models. Similarly, we can derive the optimal sampling density for dependent bivariate joint probability density with copula dependence as follows. We start by applying the joint density in Eq. (41) to the optimal density in Eq. (37) where we require an additional summation over all N_{d_c} candidate copula models:

$$q_c^*(\boldsymbol{x}) = \frac{1}{N_{d_1} N_{d_2} N_{d_c}} \sum_{i=1}^{N_{d_1}} \sum_{j=1}^{N_{d_2}} \sum_{k=1}^{N_{d_c}} E_\theta\left[c_{12}^k(F_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)|\boldsymbol{\theta}_c) \cdot f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1) \cdot f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)\right]. \tag{42}$$

Next, let us apply the law of total expectation as:

$$E[X] = E[E[X|Y]] = \int_Y E[X|Y = y]p(y)dy \tag{43}$$

where

$$X = c_{12}^k(F_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)|\boldsymbol{\theta}_c) \cdot f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1) \cdot f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2) \tag{44}$$

and $Y = y$ is the condition that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ take specific values, i.e.

$$\boldsymbol{\theta}_1 = \theta_n, \text{ and } \boldsymbol{\theta}_2 = \theta_m. \tag{45}$$

Applying the law of total expectation, the summand in Eq. (42) can be expressed as

$$\int_{\boldsymbol{\theta}_1}\int_{\boldsymbol{\theta}_2} E_\theta\left[c_{12}^k(F_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)|\boldsymbol{\theta}_c, \boldsymbol{\theta}_1 = \theta_n, \boldsymbol{\theta}_2 = \theta_m) \cdot f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1 = \theta_n) \cdot f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2 = \theta_m)\right] \cdot$$
$$p(\boldsymbol{\theta}_1 = \theta_n, \boldsymbol{\theta}_m = \theta_2)d\theta_n d\theta_m. \tag{46}$$

Recognizing that the first term is conditioned on $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ taking specific values, the expectation can be written entirely with respect to $\boldsymbol{\theta}_c$ and the marginal densities can be taken outside the expectation. We further recognize that $p(\boldsymbol{\theta}_1 = \theta_n, \boldsymbol{\theta}_2 = \theta_m) = p(\boldsymbol{\theta}_1 = \theta_n|\boldsymbol{d}, M_i) \cdot p(\boldsymbol{\theta}_2 = \theta_m|\boldsymbol{d}, M_j)$ because $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are independent and inferred from the data for each variable. Hence, Eq. (46) becomes:

$$\int_{\boldsymbol{\theta}_1}\int_{\boldsymbol{\theta}_2} E_{\theta_c}\left[c_{12}^k(F_1(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2(\boldsymbol{x}_2|\boldsymbol{\theta}_2)|\boldsymbol{\theta}_c, \boldsymbol{\theta}_1 = \theta_n, \boldsymbol{\theta}_2 = \theta_m)\right] \cdot f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1 = \theta_n) \cdot f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2 = \theta_m) \cdot$$
$$p(\boldsymbol{\theta}_1 = \theta_n|\boldsymbol{d}, M_i) \cdot p(\boldsymbol{\theta}_2 = \theta_m|\boldsymbol{d}, M_j)d\theta_n d\theta_m. \tag{47}$$

Plugging this into Eq. (42) and letting

$$\hat{c}_{12}^{mn}(F_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)) = \frac{1}{N_{d_c}} \sum_{k=1}^{N_{d_c}} E_{\theta_c}\left[c_{12}^k(F_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)|\boldsymbol{\theta}_c, \boldsymbol{\theta}_1 = \theta_n, \boldsymbol{\theta}_2 = \theta_m)\right] \tag{48}$$

be the expected conditional copula for marginal parameter pair $(\boldsymbol{\theta}_1 = \theta_n, \boldsymbol{\theta}_2 = \theta_m)$ gives:

$$q_c^*(\boldsymbol{x}) = \frac{1}{N_{d_1} N_{d_2}} \sum_{i=1}^{N_{d_1}} \sum_{j=1}^{N_{d_2}} \int_{\boldsymbol{\theta}_1}\int_{\boldsymbol{\theta}_2} \hat{c}_{12}^{mn}(F_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)) \cdot f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1 = \theta_n) \cdot f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2 = \theta_m) \cdot$$
$$p(\boldsymbol{\theta}_1 = \theta_n|\boldsymbol{d}, M_i) \cdot p(\boldsymbol{\theta}_2 = \theta_m|\boldsymbol{d}, M_j)d\theta_n d\theta_m. \tag{49}$$

Next, recognizing that we likely cannot know $p(\boldsymbol{\theta}_1 = \theta_n|\boldsymbol{d}, M_i)$ and $p(\boldsymbol{\theta}_2 = \theta_m|\boldsymbol{d}, M_j)$ explicitly because we do not have the parameter posterior density in closed form (instead, we have sampled it from MCMC), we will rely on Monte Carlo estimation of the integrals over θ_n, θ_m with $N_n \times N_m \to \infty$ samples such that θ_n and θ_m are drawn randomly from the posterior parameter density (i.e. from MCMC samples) and allowing us to express the optimal density as

$$q_c^*(\boldsymbol{x}) = \frac{1}{N_{d_1} N_{d_2} N_n N_m} \sum_{i=1}^{N_{d_1}} \sum_{j=1}^{N_{d_2}} \sum_{n=1}^{N_n} \sum_{m=1}^{N_m} \hat{c}_{12}^{mn}(F_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)) \cdot f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1 = \theta_n) \cdot f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2 = \theta_m). \quad (50)$$

The optimal sampling density in Eq. (50) can be generalized to account for the posterior model probabilities as follows:

$$q_c^*(\boldsymbol{x}) = \frac{1}{N_n N_m} \sum_{i=1}^{N_{d_1}} \sum_{j=1}^{N_{d_2}} \sum_{n=1}^{N_n} \sum_{m=1}^{N_m} \hat{c}_{12}^{mn}(F_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)) \cdot \pi_i^1 f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1 = \theta_n) \cdot \pi_j^2 f_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2 = \theta_m) \quad (51)$$

where the expected conditional copula $\hat{c}_{12}^{mn}(F_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2))$ in Eq. (48) is replaced by:

$$\hat{c}_{12}^{mn}(F_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)) = \sum_{k=1}^{N_{d_c}} \pi_c^{k,mn} E_{\theta_c}\left[c_{12}^k(F_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^j(\boldsymbol{x}_2|\boldsymbol{\theta}_2)|\boldsymbol{\theta}_c, \boldsymbol{\theta}_1 = \theta_n, \boldsymbol{\theta}_2 = \theta_m)\right] \quad (52)$$

where $\pi_c^{k,mn}$ is the posterior copula model probability conditioned on $\boldsymbol{\theta}_1 = \theta_n$ and $\boldsymbol{\theta}_2 = \theta_m$.

4.3. Optimal important density for bivariate joint probability density with copula dependence: implementation

In the derived form, the optimal sampling density in Eqs. (51) and (52) is difficult to implement, involving several nested loops. For every pair of marginals $\{f_1^i(\cdot), f_2^j(\cdot)\}$, we need to randomly sample N_n and N_m samples respectively from the parameter densities using MCMC. Then, for each pair of the $N_n \times N_m$ model parameters, we need N_{θ_c} samples of the copula parameters for each of the N_{d_c} candidate copula models for a total computational complexity of $N_{d_1} \times N_{d_2} \times N_n \times N_m \times N_{d_c} \times N_{\theta_c}$. Here, we propose a Monte Carlo sampling approach to reduce the complexity of this calculation.

This is performed by first populating the marginal sets. That is, we perform the multimodel selection process for the marginal distributions to obtain $\mathbb{M}^1$ and $\mathbb{M}^2$ and the model probabilities $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$. Next we, perform Bayesian parameter estimation for each of the marginals in $\mathbb{M}^1$ and $\mathbb{M}^2$, which provides a set of N_m and N_n parameter values following the joint parameter distributions of each model M_i^1 and M_i^2, respectively. Next, instead of combining all combinations of marginals and parameters ($N_{d_1} \times N_{d_2} \times N_n \times N_m$), we set a feasible value N_{t_d} of total marginal combinations to be considered. Note that while the total number of combinations is likely to be in the millions, e.g. $4 \times 4 \times 1000 \times 1000 = 16,000,000$, we generally select $N_{t_d} \approx 1,000$. This set of N_{t_d} probability models is selected by randomly drawing marginals from $\mathbb{M}^1$ and $\mathbb{M}^2$ with probabilities $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$ and then randomly drawing their respective parameters from the MCMC samples for each marginal.

This first simplification reduces the estimator in Eq. (51) to the following form:

$$q_c^*(\boldsymbol{x}) = \frac{1}{N_{t_d}} \sum_{l=1}^{N_{t_d}} \hat{c}_{12}^l(F_1^l(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^l(\boldsymbol{x}_2|\boldsymbol{\theta}_2)) \cdot f_1^l(\boldsymbol{x}_1|\boldsymbol{\theta}_1 = \theta_1^l) \cdot f_2^l(\boldsymbol{x}_2|\boldsymbol{\theta}_2 = \theta_2^l) \quad (53)$$

where l is a single index associated with a pair of marginals randomly selected according to their model probabilities as well as random parameters for each of these marginals selected from their joint posterior pdf.

For each of the N_{t_d} marginal pairs, we perform copula model selection to obtain the copula model probabilities π_c^l and then, again perform MCMC to obtain samples of the copula parameters following their posterior distribution. To estimate the expected conditional copula, we again reduce the samples from $N_{d_c} \times N_{\theta_c}$, which might be on the order of 10,000, to a smaller number N_{t_c} (≈ 500). We estimate Eq. (52) by randomly drawing N_{t_c} copula models according to π_c^l and randomly drawing the parameter values from the MCMC samples for that model obtained during Bayesian inference. Procedurally, Eq. (52) is re-expressed in the following form for use in Eq. (53):

$$\hat{c}_{12}^l(F_1^l(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^l(\boldsymbol{x}_2|\boldsymbol{\theta}_2)) = \frac{1}{N_{t_c}} \sum_{k=1}^{N_{t_c}} c_{12}^k(F_1^l(\boldsymbol{x}_1|\boldsymbol{\theta}_1), F_2^l(\boldsymbol{x}_2|\boldsymbol{\theta}_2)|\boldsymbol{\theta}_c^k, \boldsymbol{\theta}_1 = \theta_1^l, \boldsymbol{\theta}_2 = \theta_2^l) \quad (54)$$

where the superscript k in c_{12}^k denotes that the form of the model for the kth copula is random and follows the model probabilities π_c^l, while superscript k in $\boldsymbol{\theta}_c^k$ denotes that the copula parameters are randomly drawn from the posterior parameter density associated with copula model $c_{12}^k(\cdot)$.

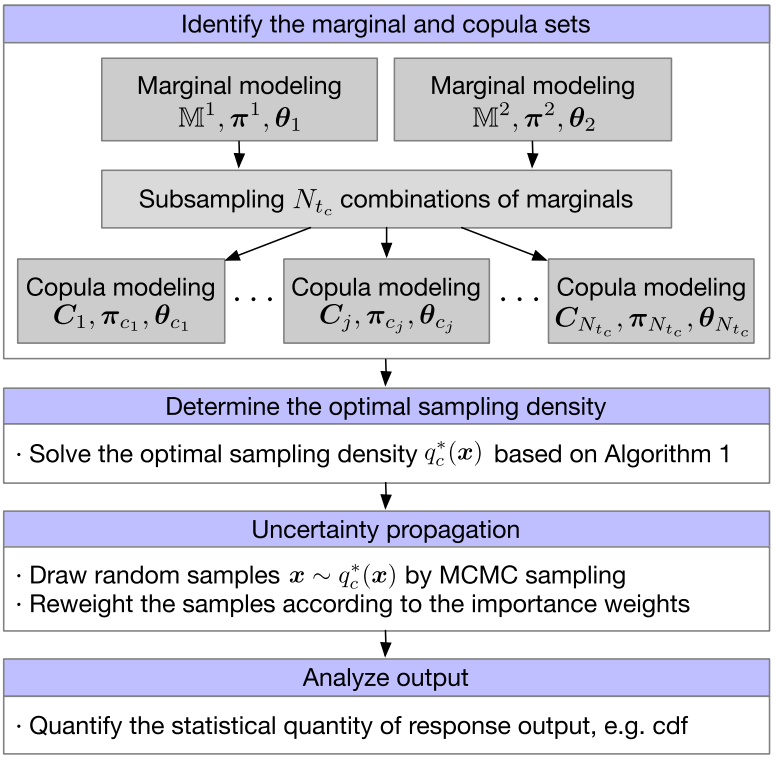

Fig. 7. Flowchart for propagation of imprecise probabilities with copula-based dependence modeling.

Eqs. (53) and (54) are then actually used for optimal sampling density estimation. Overall, this reduces the complexity of the optimal sampling density estimation from $N_{d_1} \times N_{d_2} \times N_n \times N_m \times N_{d_c} \times N_{\theta_c} \sim \mathcal{O}(10^{11} - 10^{12})$ to $N_{t_d} \times N_{t_c} \sim \mathcal{O}(10^5 - 10^6)$, while retaining a statistically representative set of joint probability models from which to estimate the optimal.

We further emphasize here that calculation of the optimal sampling density is generally much less expensive than evaluation of the computational model through which uncertainties are being propagated. Nonetheless, the optimal sampling density must be called for every sample re-weighting, which can lead to additional computational burden. One simple way to alleviate this burden is to compute the optimal joint density once via the approach described above and develop an inexpensive surrogate or lookup table to call it rapidly.

The implementation procedure for copula-based optimal sampling density estimation is summarized as Algorithm 1.

Algorithm 1 Copula-based optimal sampling density.

1: Identify the marginal models, $\mathbb{M}^1$ and $\mathbb{M}^2$, and their model probabilities, $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$, using Bayesian multimodel inference.
2: Perform Bayesian parameter estimation using MCMC to obtain sample parameters following the posterior parameter density, $p(\boldsymbol{\theta}_i|\boldsymbol{d}, M_i)$ for each marginal model.
3: Randomly select a pair of marginals $\left\{f_1^i(\boldsymbol{x}_1|\boldsymbol{\theta}_1 = \theta_n), f_2^i(\boldsymbol{x}_2|\boldsymbol{\theta}_2 = \theta_m)\right\}$ by drawing the marginal models with probabilities $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$ and randomly drawing the parameters from the MCMC samples of the posterior parameter density.
4: Identify the candidate copula models and their model probabilities π_c for the specific marginal pair using Bayesian multimodel inference.
5: Perform Bayesian parameter estimation using MCMC to obtain sample parameters following the posterior parameter density for each copula model.
6: Randomly draw N_{t_c} copula models according to their model probabilities π_c and their associated parameters from the MCMC samples for the posterior parameter density.
7: Estimate the expected conditional copula $\hat{c}_{12}^I$ according to Eq. (54).
8: Determine the expected joint density by multiplying the marginals and copula.
9: Repeat Step 3 - 8 for a large number, N_{t_d}, of marginal pairs.
10: Determine the copula-based optimal sampling density $q_c^*(\boldsymbol{x})$ by averaging the N_{t_d} joint densities as shown in Eq.. (53).
11: (Optional) Create a surrogate optimal sampling density or lookup table to expedite sample re-weighting.

4.4. Propagation of imprecise probabilities with copula dependence modeling

With the constituents outlined in the previous section, the importance sampling reweighting approach for imprecise uncertainty propagation with copula dependence is summarized here and a flowchart is shown in Fig. 7.

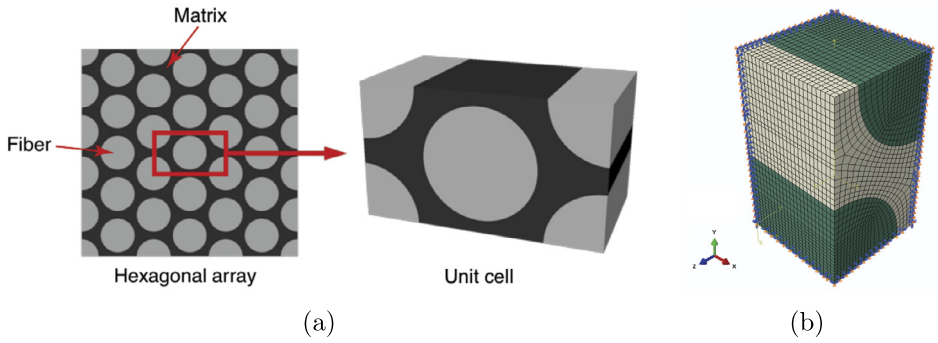

Fig. 8. Unidirectional fiber reinforced composite (a) Hexagonal RVE unit and (b) RVE FEM model.

- Step 1: Identify the marginal and copula sets – Given a small data set, the hierarchical Bayesian multimodel inference outlined in Section 3.2 is used to identify the candidate sets of marginal distributions and copulas. We first identify candidate marginal forms and associated model probabilities, and construct combinations of marginals by randomly drawing N_{t_d} marginal pairs. For each pair of marginals, identify copula forms and estimate the copula model probabilities and copula parameters.
- Step 2: Determine the optimal sampling density – Combine all the candidate marginals and associated copulas modeling from Step 1. Solving the optimization problem yields the optimal sampling density $q_c^*(\boldsymbol{x})$, shown in Eq. (51), which is practically solved as described in Sec. 4.3 (Eqs. (53) and (54)), i.e. according to the Algorithm 1.
- Step 3: Uncertainty propagation – Uncertainty associated with copula-based dependence modeling is propagated using importance sampling with optimal sampling density $q_c^*(\boldsymbol{x})$. Samples are drawn from $q_c^*(\boldsymbol{x})$ using MCMC sampling and are reweighted for each model according to the importance weights $w(\boldsymbol{x}) = p(\boldsymbol{x})/q_c^*(\boldsymbol{x})$
- Step 4: Analyze output – Quantify the distribution of the statistical response quantity of interest.

5. Application to probabilistic prediction of unidirectional composite lamina properties

This section applies the proposed methodology to understand the influence of the constituent material properties on the out-of-plane elastic properties (Young's modulus) of a unidirectional composite lamina.

5.1. Problem description

Fiber reinforced composite materials are popular and widely used in many engineering fields because of their attractive properties, for example, high stiffness and strength combined with low weight. In order to evaluate the performance of a composite part, the accurate prediction of its mechanical properties in the layup is important [75]. Several numerical and experimental methods have been proposed to determine the mechanical properties of unidirectional lamina based on the elastic properties of the constituent materials (fibers and matrix) [76,77]. In this work, the finite element method (FEM) with a representative volume element (RVE) is used to predict the out-of-plane elastic properties of a unidirectional composite lamina given the constituent (fiber and matrix) material properties.

Typically, unidirectional composites are considered as transversely isotropic materials composed of two phases: a fiber reinforcement phase and a matrix phase, as shown in Fig. 8 (a) for a hexagonal packing configuration. Commonly, the reinforced-fiber phase for traditional materials is modeled as isotropic (e.g. glass fibers) or orthotropic (e.g. carbon fiber) and the matrix phase is typically composed of an isotropic material (e.g. epoxy). The overall mechanical properties of transversely isotropic unidirectional fiber reinforced lamina with a hexagonal packing geometry are determined by five independent engineering constants which are given by the following compliance matrix:

$$C = \begin{bmatrix} 1/E_{11} & -\nu_{12}/E_{11} & -\nu_{12}/E_{11} & 0 & 0 & 0 \\ -\nu_{12}/E_{11} & 1/E_{22} & -\nu_{23}/E_{22} & 0 & 0 & 0 \\ -\nu_{12}/E_{11} & -\nu_{23}/E_{22} & 1/E_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/G_{23} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/G_{12} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/G_{12} \end{bmatrix} \tag{55}$$

where E_{11} and E_{22} are the longitudinal and transverse Young's moduli respectively, G_{12} and G_{23} are the longitudinal and transverse shear moduli, ν_{12} is the major Poisson's ratio and ν_{23} is the minor Poisson's ratio. The transverse shear modulus is determined from the minor Poisson's ratio ν_{23} and elastic modulus E_{22} as [78]:

$$G_{23} = \frac{E_{22}}{2(1+\nu_{23})} \tag{56}$$

Table 4
Constituent material properties of E-Glass fiber/LY556 Polyester Resin composites.

Material property	Physical meaning	Mean value	Coefficient of variation
V_f	Fiber volume fraction	0.6	0.05
E_m	Matrix's Young's modules	3.375	0.05
ν_m	Matrix Poisson's ratio	0.35	0.05
E_{1f}	Fiber Young's modules along 1 direction	73.01	0.05
ν_{12f}	Fiber Poisson's ratio along 1-2 direction	0.228	0.05

Experimental determination of the in-plane lamina properties are typically straightforward and generally provide accurate values for these properties. However, the out-of-plane lamina properties are difficult to obtain experimentally [79–81], and consequently numerical prediction becomes an attractive alternative to predict these lamina properties. In this example, we focus on the determination of the elastic modulus E_{22} which is an independent out-of-plane lamina property.

The overall mechanical properties in Eq. (55) depend on the constituent properties (fibers and matrix). Table 4 shows the four independent constituent material properties and the fiber volume fraction, which are needed to define the lamina properties for the isotropic resin and fiber materials.

In this work, we study a common composite lamina fabricated from E-glass fibers and LY556 polyester resin matrix. The finite element method is employed to construct a three-dimensional RVE with two symmetry planes in the $x-y$ and $x-z$ directions and periodic boundary conditions, as shown in Fig. 8 (b). The model has a total 22750 nodes and 20448 C3D8R solid elements and is solved using the commercial solver Abaqus.

5.2. Identification of probabilistic input model

From engineering experience, the five inputs in Table 4 may be correlated or dependent and thus one task is to identify the dependence relationship among these five random variables from data. Commonly, the matrix properties E_m and ν_m are considered to be dependent and the fiber properties E_{1f} and ν_{12f} are dependent. However, fiber and matrix properties are independent of one another and the fiber volume fraction is often assumed independent of constituent properties. Therefore, the five probability inputs are composed of two bivariate dependent models and one independent variable: $\{E_m, \nu_m\}$, $\left\{E_{1f}, \nu_{12f}\right\}$ and $\left\{V_f\right\}$.

Although this type of composite materials has been used extensively in many engineering applications, statistical data for its constituent properties are very limited. Typically, only nominal design values are provided without adequate guidance on their variability. The nominal values in Table 4 were compiled from the literature for each constituent property and candidate probability distributions were identified for each property. The interested readers can find an extensive list of references for the relevant data and literature in the authors' recent work [82].

Due to a lack of statistical data for characterizing the constituent material properties, it is difficult to assign accurate and objective probabilistic models for the properties, specifically the dependence model for the constituent properties. For reference purposes, we assume normal distributions with nominal mean value in Table 4 and 5% coefficient of variation (COV) as the "true" marginal distributions for each fiber and matrix property. The matrix properties $\{E_m, \nu_m\}$ and the fiber properties $\left\{E_{1f}, \nu_{12f}\right\}$ are assumed to be strongly correlated with a "true" Frank copula with parameter $\theta = -10$. Fig. 9 shows the "true" probabilistic input model, which includes the marginal histogram and dependence relationship between each of these input variables. It can be observed that $\{E_m, \nu_m\}$ and $\left\{E_{1f}, \nu_{12f}\right\}$ have a strong dependence that follows the true Frank(-10) copula model. We assume this probabilistic model to be the truth and generate 20 random data, as shown in Fig. 10 for the joint matrix and fiber properties. These serve as the initial data from which uncertainty needs to be quantified and propagated. Clearly, a single bivariate dependence model cannot be precisely identified from these data – although it is clear that the properties are dependent.

5.3. Probabilistic prediction of composite properties

The multimodel inference approach proposed herein is applied to this problem, given the limited data characterizing the constituent material properties and their clear dependencies. We first identify a set of candidate marginal probability models, which include the Gaussian, Gamma, Lognormal and Weibull distributions. The Bayesian multimodel approach in Eq. (24) is used to estimate the posterior model probabilities and the corresponding model parameter uncertainties are estimated by Bayesian inference using MCMC sampling. Combining these model-form and model parameter uncertainties, we therefore obtain an ensemble of plausible probability densities for the five input variables shown in Fig. 11.

In this example, we identify 500 candidate densities for each marginal such that the total number of combinations of these marginal distributions is $500^5 = 3.125^{13}$, which is computationally prohibitive. Instead, a representative 1000 marginal pairs are compiled by Latin hypercube sampling. To evaluate the elastic modulus E_{22}, 5,000 random samples are drawn from the optimal sampling density, shown in the thick black thick curves in Fig. 11, for each material property and computational model evaluations are performed using FEM. Hence, the computational advantage of the approach lies in the vastly reduced number of model evaluations needed to propagate the full model set. In this case, we need only 5,000 simulations where conventional multi-loop Monte Carlo approaches require on the order of $5,000^3$ simulations to cover the full set of copulas,

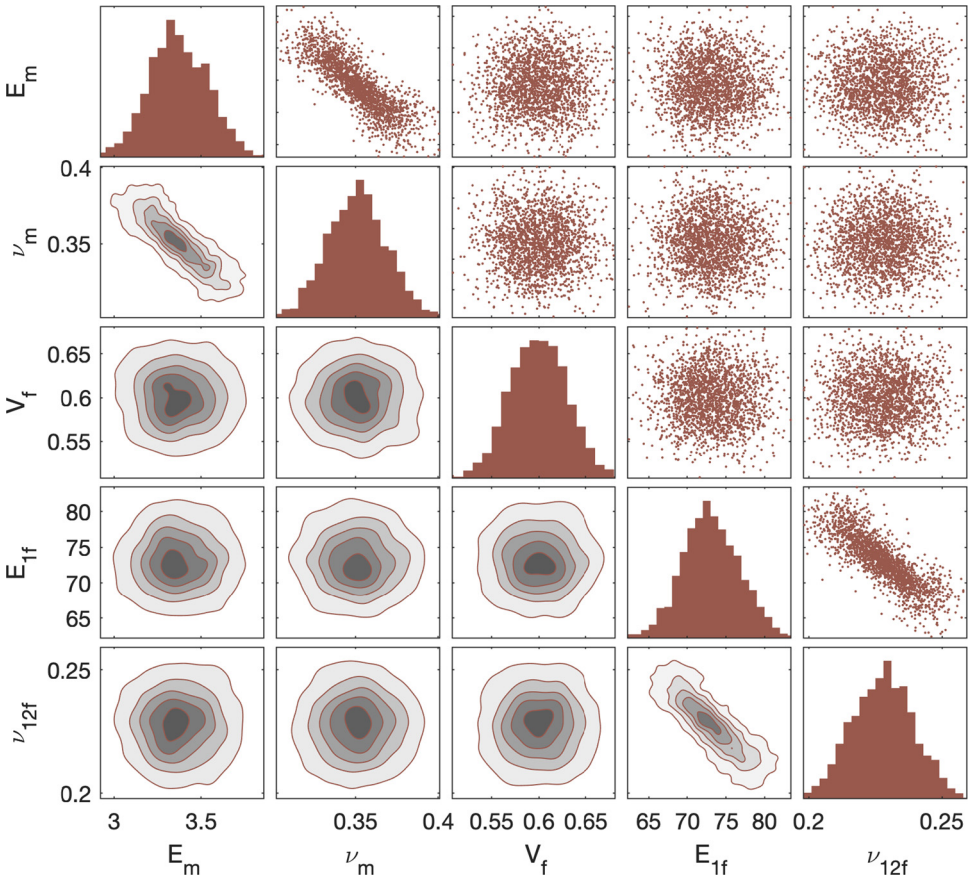

Fig. 9. Dependent probabilistic input model.

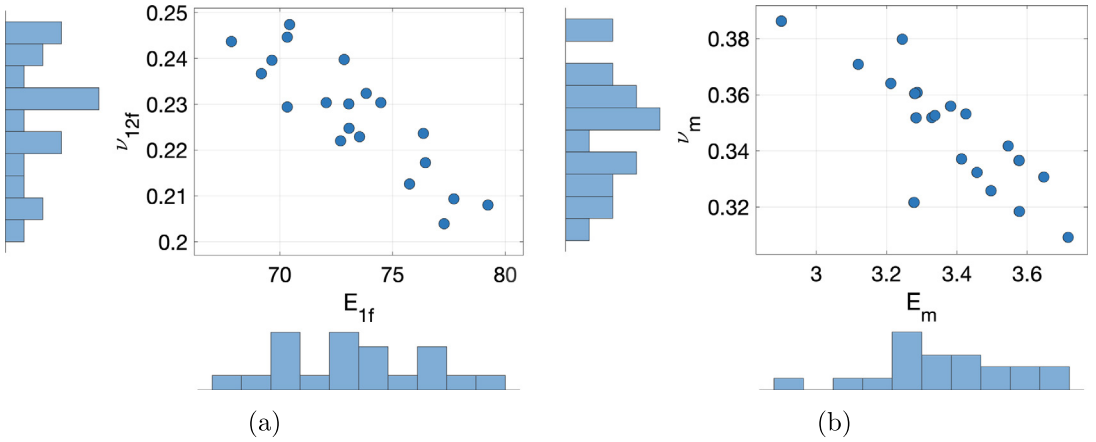

Fig. 10. 20 randomly generated constituent material properties that serve as the initial dataset (a) fiber property and (b) matrix property.

marginals, and marginal parameters. For the composite model used herein, the 5,000 simulations take approximately 28 cpu-hours to complete, making the conventional strategy infeasible.

If the multivariate input is assumed independent, we can easily achieve the probabilistic prediction of overall material property E_{22} by multiplying each marginal. Fig. 12 shows the cloud of candidate empirical CDFs for E_{22} based on multi-model inference from the 20 data assuming the marginals are independent and Gaussian correlated with $\rho = 0.8$. The "true" CDF in Fig. 12 (with variable dependence) is shown in black. Note that the collection of CDFs compiled under the independence assumption (blue) and Gaussian correlation (green) as well as the true estimate with dependence (black) seem to

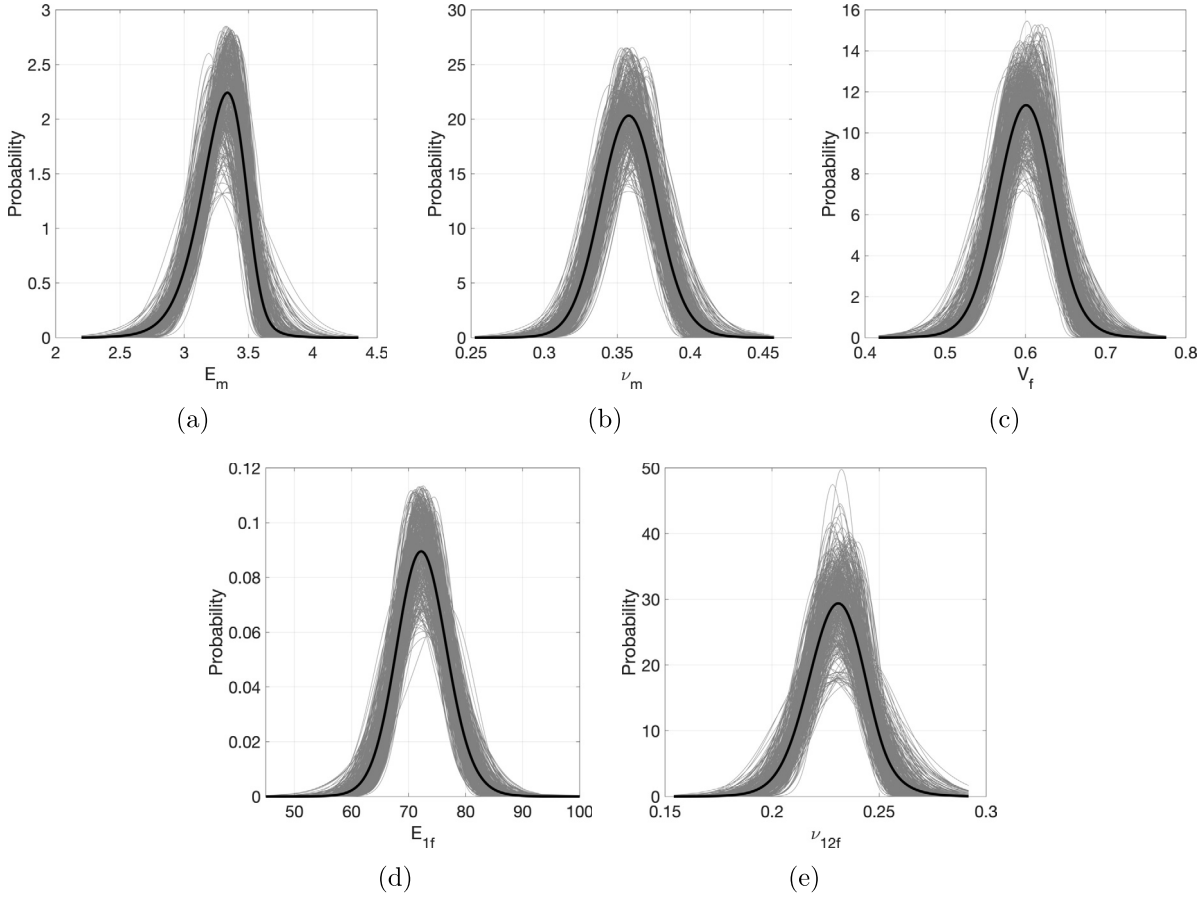

Fig. 11. Multiple candidate probability densities for marginals (a) E_m, (b) ν_m, (c) V_f, (d) E_{1f} and (e) ν_{12f}.

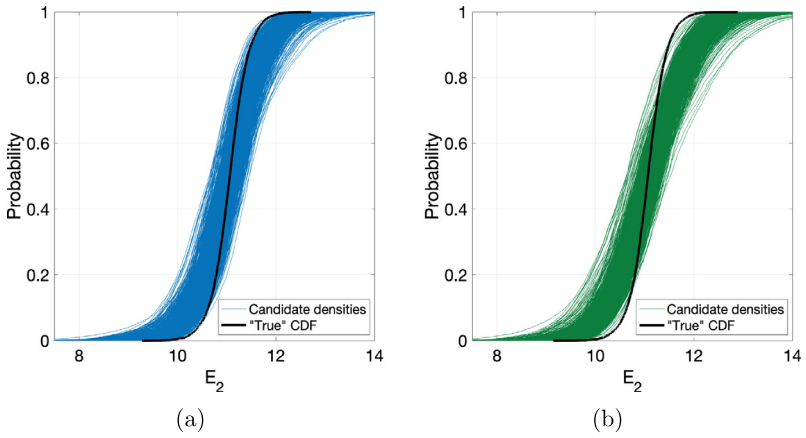

Fig. 12. Collection of candidate empirical CDFs for Young's modulus E_{22} given the initial 20 data, assuming (a) independent marginal distributions and (b) Gaussian correlation. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

overlap – suggesting that perhaps the independence assumption is sufficient to bound the elastic properties. However, as we show next, this result underestimates the uncertainty in E_{22}.

To account for variable dependence, for each pair of marginals we must identify a set of candidate copulas. For this we perform the hierarchical Bayesian multimodel selection for the Gaussian, Clayton, Frank and Gumbel copulas. We first compute the posterior copula model probabilities and then compute the associated joint parameter densities. For each pair of marginals, we then construct an ensemble of copula model sets by randomly selecting the copula models and copula parameters. Finally, the optimal sampling density in Eq. (53) is determined and employed for propagation of the multiple

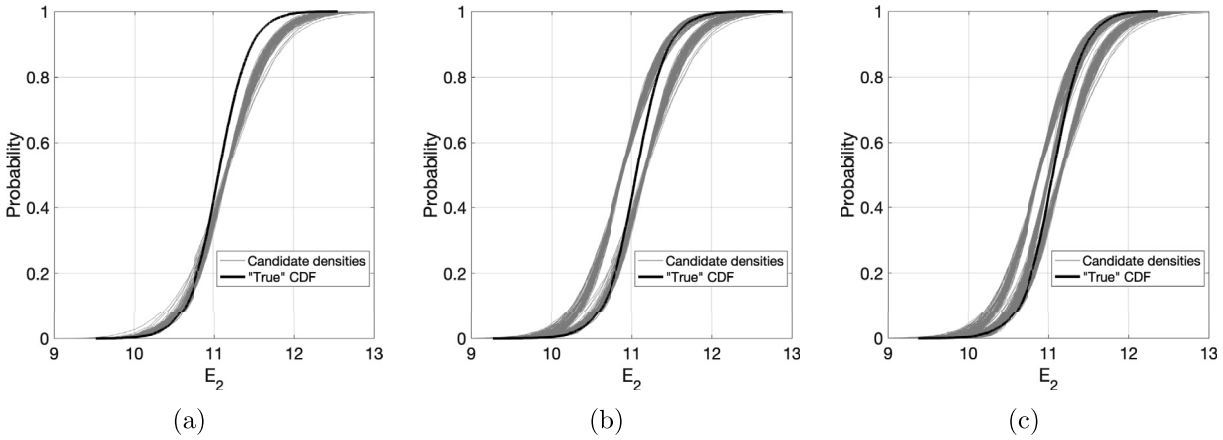

Fig. 13. Collection of candidate empirical CDFs for Young's modulus E_{22} with only copula uncertainty given (a) one pair of marginals, (b) two pairs of marginals and (c) three pairs of marginals.

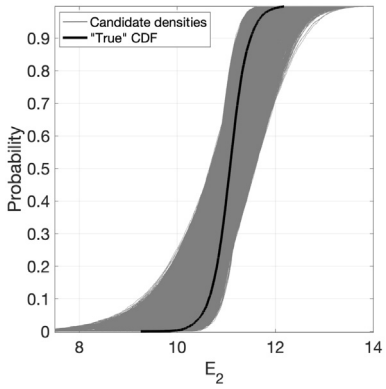

Fig. 14. Total collection of candidate empirical CDFs for Young's modulus E_{22} with uncertainty in dependence modeling given 20 data.

candidate densities with copula dependence. Fig. 13 shows three examples illustrating the influence of copula dependence uncertainty for specific marginal density pairs. Notice that, when the marginals are assumed to be independent a single cdf for E_{22} is generated. However, with uncertainty in the copula dependence, there are several candidate pdfs for each pair of marginal densities. In other words, the uncertainty associated with the spread in the sets of cdfs in Fig. 13 is ignored if we assume independent marginals.

When we combine the uncertainties from the copula model and marginal model together in Fig. 14, we see that the overall uncertainty is considerably wider than it was when assuming the marginals to be independent or Gaussian correlated (Fig. 12). That is, the candidate densities with dependence modeling show a much wider band than the densities with independent or Gaussian correlated assumption.

5.4. Influence of dataset size

In this section, we investigate the convergence of the composite material properties as a function of dataset size. As discussed in the previous section, small datasets led to large uncertainties including the copula model and marginal model in the composite material properties. This raises a critical question: "How much data is necessary to gain adequate confidence in the probabilistic prediction of composite material properties?"

Here, additional data are generated from the true joint probability density. We begin with the initial 20 data and increase to 50 data, 500 data and 5000 data, as shown in Fig. 15. As the data set size increases, we more clearly see the true dependence emerge. Both the normal marginals become increasingly pronounced and the nature of the underlying copula dependence becomes clear.

Fig. 16 shows the results of the multimodel uncertainty propagation to estimate the cdf of the transverse modulus E_{22} for increasing data set size. The figure shows the convergence of the approach under assumptions of independent marginals (Fig. 16a-c), Gaussian correlation (Fig. 16d-f) and with dependence included (Fig. 16g-i). The true cdf (with the known joint probability densities) is shown for reference. As expected, in all three cases the band of cdfs narrow as additional data are collected – i.e. uncertainty in the prediction of E_{22} is reduced. However, we notice that under the assumption

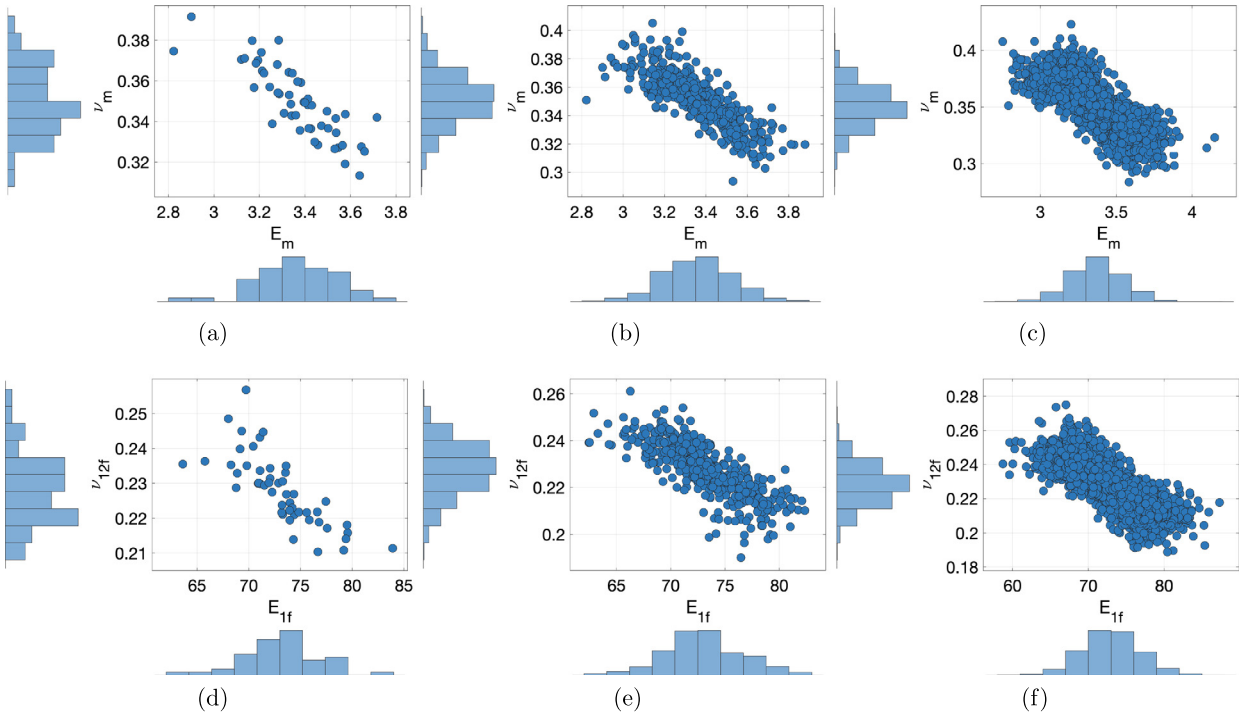

Fig. 15. Increasing data set size for dependent matrix and fiber properties: (a,d) 50 data, (b,e) 500 data and (c,f) 5000 data.

of independent marginals and Gaussian correlation, the band of cdfs do not converge to the true cdf. Instead, there is a bias introduced by the assumption of independent and Gaussian correlated marginals. Only when we account for the variable dependence in the multimodel UQ approach are we able to converge to the true cdf of the modulus. This is an important conclusion because it shows that, although uncertainty bands generated under the incorrect assumption of independence **may** initially bound the true probability distribution, they (i) are likely to underestimate the uncertainty in the estimated distribution as shown in Section 5.3, and (ii) provide biased bounds on the true probability distribution that will not converge as the data set size increases.

6. Conclusion

In this work, we propose a hierarchical multimodel approach to investigates the effect of uncertainties associated with small data sets for quantifying and propagating probabilistic model inputs with dependencies. The joint CDF of the probabilistic model inputs is composed of marginal distributions and copulas, which are modeled separately. The proposed approach is set in a hierarchical Bayesian multimodel inference framework, where the model-form and model parameter uncertainties associated with marginals are first quantified, and uncertainties associated with the copula are conditioned on specified marginal pairs. This results in an ensemble of joint probability densities that represent the imprecise probabilities in the assignment of probability model inputs with statistical dependence. A novel importance sampling reweighting algorithm is derived to efficiently propagate the imprecise probabilities through a mathematical or physical model, which is often computationally intensive. The proposed approach therefore estimates the uncertainty in the quantity of interest given multiple candidate model input distributions at a low computational cost when compared with the typical nested Monte Carlo simulations.

The methodology is demonstrated on an engineering application which aims to understand the influence of constituent properties on the overall out-of-plane properties of a transversely isotropic E-Glass fiber/LY556 Polyester Resign composites. A strong correlation between the constituent properties (fibers and matrix) is assumed and described using a Frank copula model. The results show that the assumption of independent and arbitrary Gaussian correlated marginals in the imprecise UQ modeling both underestimates the uncertainty in predictions of the modulus and yields biased statistical estimates. When copula-based dependence is integrated into the multimodel UQ framework, the model achieves more realistic bounds on the uncertainty and more accurate probabilistic predictions.

Declaration of competing interest

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

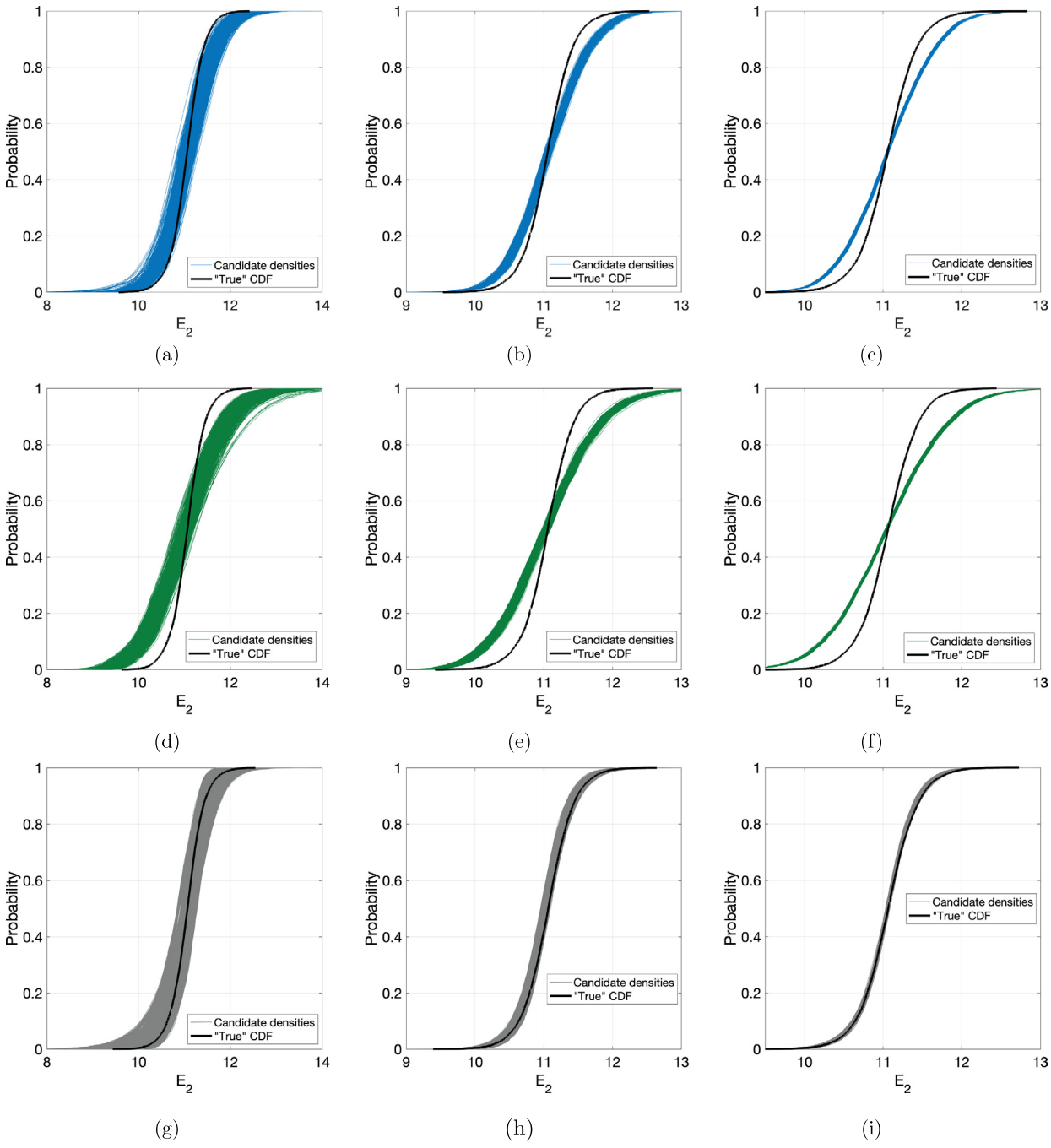

Fig. 16. Uncertain CDFs for transverse elastic modulus E_{22} with increasing data set size under the assumption of independent marginals (a-c), Gaussian correlation (d-f) and accounting for copula dependence (g-i): (a,d,g) 50 data, (b,e,h) 500 data and (c,f,i) 5000 data.

Acknowledgements

The work presented herein has been supported by the Office of Naval Research under Award Number N00014-16-1-2582 with Dr. Paul Hess as the program officer. The work of J. Zhang was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under contract ERKJ352; and by the Artificial Intelligence Initiative at the Oak Ridge National Laboratory (ORNL). ORNL is operated by UT-Battelle, LLC., for the U.S. Department of Energy under Contract DEAC05-00OR22725. The authors are grateful to Prof. Stephanie Termaath for providing models and support for materials applications.

References

[1] R.E. Melchers, A.T. Beck, Structural Reliability Analysis and Prediction, John Wiley & Sons, 2018.
[2] R. Li, R. Ghanem, Adaptive polynomial chaos expansions applied to statistics of extremes in nonlinear random vibration, Probab. Eng. Mech. 13 (2) (1998) 125–136.
[3] A. Nataf, Determination des distribution don t les marges sont donnees, C. R. Acad. Sci. 225 (1962) 42–43.
[4] R. Lebrun, A. Dutfoy, An innovating analysis of the nataf transformation from the copula viewpoint, Probab. Eng. Mech. 24 (3) (2009) 312–320.
[5] M. Rosenblatt, Remarks on a multivariate transformation, Ann. Math. Stat. 23 (3) (1952) 470–472.
[6] R.B. Nelsen, An Introduction to Copulas, Springer Science & Business Media, 2007.
[7] H. Joe, Dependence Modeling with Copulas, CRC Press, 2014.
[8] S. Wisadwongsa, S. Tasena, Bivariate quadratic copula constructions, Int. J. Approx. Reason. 92 (2018) 1–19.
[9] H. Joe, D. Kurowicka, Dependence Modeling: Vine Copula Handbook, World Scientific, 2011.
[10] K. Aas, C. Czado, A. Frigessi, H. Bakken, Pair-copula constructions of multiple dependence, Insur. Math. Econ. 44 (2) (2009) 182–198.
[11] H. Joe, H. Li, A.K. Nikoloulopoulos, Tail dependence functions and vine copulas, J. Multivar. Anal. 101 (1) (2010) 252–270.
[12] T. Nagler, C. Bumann, C. Czado, Model selection in sparse high-dimensional vine copula models with an application to portfolio risk, J. Multivar. Anal. 172 (2019) 180–192, https://doi.org/10.1016/j.jmva.2019.03.004.
[13] D. Müller, C. Czado, Dependence modelling in ultra high dimensions with vine copulas and the Graphical Lasso, Comput. Stat. Data Anal. 137 (2019) 211–232, https://doi.org/10.1016/j.csda.2019.02.007.
[14] Á. Rózsás, Z. Mogyorósi, The effect of copulas on time-variant reliability involving time-continuous stochastic processes, Struct. Saf. 66 (2017) 94–105.
[15] C. Wang, H. Zhang, Roles of load temporal correlation and deterioration-load dependency in structural time-dependent reliability, Comput. Struct. 194 (2018) 48–59.
[16] D. Xu, M. Xing, Q. Wei, Y. Qin, J. Xu, Y. Chen, R. Kang, Failure behavior modeling and reliability estimation of product based on vine-copula and accelerated degradation data, Mech. Syst. Signal Process. 113 (2018) 50–64.
[17] L. He, Z. Lu, X. Li, Failure-mode importance measures in structural system with multiple failure modes and its estimation using copula, Reliab. Eng. Syst. Saf. 174 (2018) 53–59.
[18] F. Wang, H. Li, The role of copulas in random fields: characterization and application, Struct. Saf. 75 (2018) 75–88.
[19] Y. Pan, S. Ou, L. Zhang, W. Zhang, X. Wu, H. Li, Modeling risks in dependent systems: a Copula-Bayesian approach, Reliab. Eng. Syst. Saf. 188 (2019) 416–431, https://doi.org/10.1016/j.ress.2019.03.048.
[20] P. Wang, Z. Lu, K. Zhang, S. Xiao, Z. Yue, Copula-based decomposition approach for the derivative-based sensitivity of variance contributions with dependent variables, Reliab. Eng. Syst. Saf. 169 (2018) 437–450.
[21] Z. Hu, S. Mahadevan, Probability models for data-driven global sensitivity analysis, Reliab. Eng. Syst. Saf. 187 (2019) 40–57, https://doi.org/10.1016/j.ress.2018.12.003.
[22] Z. Xi, R. Jing, P. Wang, C. Hu, A copula-based sampling method for data-driven prognostics, Reliab. Eng. Syst. Saf. 132 (2014) 72–82.
[23] Z. Xi, X. Zhao, An enhanced copula-based method for data-driven prognostics considering insufficient training units, Reliab. Eng. Syst. Saf. 188 (2019) 181–194, https://doi.org/10.1016/j.ress.2019.03.015.
[24] Y. Zhang, M. Beer, S.T. Quek, Long-term performance assessment and design of offshore structures, Comput. Struct. 154 (2015) 101–115.
[25] M. Masina, A. Lamberti, R. Archetti, Coastal flooding: a copula based approach for estimating the joint probability of water levels and waves, Coast. Eng. 97 (2015) 37–52.
[26] W.P. Warsido, G.T. Bitsuamlak, Synthesis of wind tunnel and climatological data for estimating design wind effects: a copula based approach, Struct. Saf. 57 (2015) 8–17.
[27] K. Goda, S. Tesfamariam, Multi-variate seismic demand modelling using copulas: application to non-ductile reinforced concrete frame in Victoria, Canada, Struct. Saf. 56 (2015) 39–51.
[28] H. Joe, Multivariate extreme-value distributions with applications to environmental data, Can. J. Stat. 22 (1) (1994) 47–64.
[29] H. Joe, Multivariate Models and Multivariate Dependence Concepts, CRC Press, 1997.
[30] T. Bedford, R.M. Cooke, Vines: a new graphical model for dependent random variables, Ann. Stat. (2002) 1031–1068.
[31] C. Czado, Pair-copula constructions of multivariate copulas, in: Copula Theory and Its Applications, Springer, 2010, pp. 93–109.
[32] C. Czado, E.C. Brechmann, L. Gruber, Selection of vine copulas, in: Copulae in Mathematical and Quantitative Finance, Springer, 2013, pp. 17–37.
[33] J. Dissmann, E.C. Brechmann, C. Czado, D. Kurowicka, Selecting and estimating regular vine copulae and application to financial returns, Comput. Stat. Data Anal. 59 (2013) 52–69.
[34] E.C. Brechmann, C. Czado, K. Aas, Truncated regular vines in high dimensions with application to financial data, Can. J. Stat. 40 (1) (2012) 68–85.
[35] E. Torre, S. Marelli, P. Embrechts, B. Sudret, A general framework for data-driven uncertainty quantification under complex input dependencies using vine copulas, Probab. Eng. Mech. 55 (2019) 1–16.
[36] Y. Qiu, Q. Li, Y. Pan, H. Yang, W. Chen, A scenario generation method based on the mixture vine copula and its application in the power system with wind/hydrogen production, Int. J. Hydrog. Energy 44 (11) (2019) 5162–5170.
[37] Q. Li, Y. Cai, H. Wang, Z. Lv, E. Li, An efficient D-vine copula-based coupling uncertainty analysis for variable-stiffness composites, Compos. Struct. 219 (2019) 221–241, https://doi.org/10.1016/j.compstruct.2019.03.067.
[38] R. Niemierko, J. Töppel, T. Tränkler, A d-vine copula quantile regression approach for the prediction of residential heating energy consumption based on historical data, Appl. Energy 233 (2019) 691–708.
[39] A. Der Kiureghian, O. Ditlevsen, Aleatory or epistemic? Does it matter?, Struct. Saf. 31 (2) (2009) 105–112.
[40] S. Ferson, L.R. Ginzburg, Different methods are needed to propagate ignorance and variability, Reliab. Eng. Syst. Saf. 54 (2–3) (1996) 133–144.
[41] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338–353.
[42] D. Dubois, H. Prade, Fundamentals of Fuzzy Sets, vol. 7, Springer Science & Business Media, 2012.
[43] Z. Wang, G.J. Klir, Fuzzy Measure Theory, Springer Science & Business Media, 2013.
[44] J.O. Berger, E. Moreno, L.R. Pericchi, M.J. Bayarri, J.M. Bernardo, J.A. Cano, J. De la Horra, J. Martín, D. Ríos-Insúa, B. Betrò, et al., An overview of robust bayesian analysis, Test 3 (1) (1994) 5–124.
[45] T. Fetz, M. Oberguggenberger, Propagation of uncertainty through multivariate functions in the framework of sets of probability measures, Reliab. Eng. Syst. Saf. 85 (1–3) (2004) 73–87.
[46] I. Molchanov, I.S. Molchanov, Theory of Random Sets, vol. 19, Springer, 2005.
[47] T. Fetz, M. Oberguggenberger, Imprecise random variables, random sets, and Monte Carlo simulation, Int. J. Approx. Reason. 78 (2016) 252–264.
[48] R.E. Moore, Methods and Applications of Interval Analysis, vol. 2, SIAM, 1979.
[49] S. Ferson, V. Kreinovich, L. Ginzburg, D.S. Myers, K. Sentz, Constructing Probability Boxes and Dempster-Shafer Structures, vol. 835, Sandia National Laboratories Albuquerque, 2002.
[50] A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Stat. (1967) 325–339.
[51] G. Shafer, A Mathematical Theory of Evidence, vol. 1, Princeton University Press, Princeton, 1976.

[52] P. Walley, Statistical Reasoning with Imprecise Probabilities, Monographs on Statistics and Applied Probability, vol. 42, Chapman and Hall, London, 1991.
[53] P. Walley, Towards a unified theory of imprecise probability, Int. J. Approx. Reason. 24 (2–3) (2000) 125–148.
[54] M. Beer, S. Ferson, V. Kreinovich, Imprecise probabilities in engineering analyses, Mech. Syst. Signal Process. 37 (1–2) (2013) 4–29.
[55] I. Montes, E. Miranda, R. Pelessoni, P. Vicig, Sklar's theorem in an imprecise setting, Fuzzy Sets Syst. 278 (2015) 48–66.
[56] R. Pelessoni, P. Vicig, I. Montes, E. Miranda, Imprecise copulas and bivariate stochastic orders, in: Proc. EUROFUSE 2013, Oviedo, 2013, pp. 217–224.
[57] T. Coolen-Maturi, F.P. Coolen, N. Muhammad, Predictive inference for bivariate data: combining nonparametric predictive inference for marginals with an estimated copula, J. Stat. Theory Pract. 10 (3) (2016) 515–538.
[58] D. Kurowicka, R.M. Cooke, Uncertainty Analysis with High Dimensional Dependence Modelling, John Wiley & Sons, 2006.
[59] R. Schefzik, T.L. Thorarinsdottir, T. Gneiting, et al., Uncertainty quantification in complex simulation models using ensemble copula coupling, Stat. Sci. 28 (4) (2013) 616–640.
[60] J. Zhang, M.D. Shields, On the quantification and efficient propagation of imprecise probabilities resulting from small datasets, Mech. Syst. Signal Process. 98 (2018) 465–483.
[61] J. Zhang, M.D. Shields, The effect of prior probabilities on quantification and propagation of imprecise probabilities resulting from small datasets, Comput. Methods Appl. Mech. Eng. 334 (2018) 483–506.
[62] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93.
[63] M. Sklar, Fonctions de repartition an dimensions et leurs marges, Publ. Inst. Stat. Univ. Paris 8 (1959) 229–231.
[64] S. Demarta, A.J. McNeil, The t copula and related copulas, Int. Stat. Rev. 73 (1) (2005) 111–129.
[65] E. Brechmann, U. Schepsmeier, Cdvine: modeling dependence with c-and d-vine copulas in r, J. Stat. Softw. 52 (3) (2013) 1–27.
[66] C. Czado, U. Schepsmeier, A. Min, Maximum likelihood estimation of mixed c-vines with application to exchange rates, Stat. Model. 12 (3) (2012) 229–255.
[67] A. Min, C. Czado, Bayesian model selection for d-vine pair-copula constructions, Can. J. Stat. 39 (2) (2011) 239–258.
[68] F.J. Massey Jr, The Kolmogorov-Smirnov test for goodness of fit, J. Am. Stat. Assoc. 46 (253) (1951) 68–78.
[69] J.H. McDonald, Handbook of Biological Statistics, vol. 2, Sparky House Publishing, Baltimore, MD, 2009.
[70] K.P. Burnham, D.R. Anderson, Multimodel inference understanding aic and bic in model selection, Sociol. Methods Res. 33 (2) (2004) 261–304.
[71] J. Stöber, U. Schepsmeier, Is there significant time-variation in multivariate copulas?, arXiv preprint, arXiv:1205.4841.
[72] L. Gruber, C. Czado, et al., Sequential bayesian model selection of regular vine copulas, Bayesian Anal. 10 (4) (2015) 937–963.
[73] S. Sankararaman, S. Mahadevan, Likelihood-based representation of epistemic uncertainty due to sparse point data and/or interval data, Reliab. Eng. Syst. Saf. 96 (7) (2011) 814–824.
[74] J. Zhang, M.D. Shields, Efficient Monte Carlo resampling for probability measure changes from bayesian updating, Probab. Eng. Mech. 55 (2019) 54–66.
[75] J. Zhang, M. Shields, S. TerMaath, Probabilistic modeling and prediction of out-of-plane unidirectional composite lamina properties, Mech. Adv. Mat. Struct. (2020) 1–17.
[76] I.M. Daniel, O. Ishai, I.M. Daniel, I. Daniel, Engineering Mechanics of Composite Materials, vol. 3, Oxford University Press, New York, 1994.
[77] R. Younes, A. Hallal, F. Fardoun, F.H. Chehade, Comparative review study on elastic properties modeling for unidirectional composite materials, in: Composites and Their Properties, Intech, 2012.
[78] Z. Hashin, Analysis of composite materials—a survey, J. Appl. Mech. 50 (3) (1983) 481–505.
[79] T. King, D. Blackketter, D. Walrath, D. Adams, Micromechanics prediction of the shear strength of carbon fiber/epoxy matrix composites: the influence of the matrix and interface strengths, J. Compos. Mater. 26 (4) (1992) 558–573.
[80] K. Gipple, D. Hoyns, Measurement of the out-of-plane shear response of thick section composite materials using the v-notched beam specimen, J. Compos. Mater. 28 (6) (1994) 543–572.
[81] P. Soden, M. Hinton, A. Kaddour, Lamina properties, lay-up configurations and loading conditions for a range of fibre reinforced composite laminates, in: Failure Criteria in Fibre-Reinforced-Polymer Composites, Elsevier, 2004, pp. 30–51.
[82] J. Zhang, S. TerMaath, M. Shields, Imprecise global sensitivity analysis using Bayesian multimodel inference and importance sampling, Mech. Syst. Signal Process. (2020) (in review).