Uncertainty Visualization Using Copula-Based Analysis in Mixed Distribution Models

Subhashis Hazarika, Ayan Biswas and Han-Wei Shen, Member, IEEE

Abstract—Distributions are often used to model uncertainty in many scientific datasets. To preserve the correlation among the spatially sampled grid locations in the dataset, various standard multivariate distribution models have been proposed in visualization literature. These models treat each grid location as a univariate random variable which models the uncertainty at that location. Standard multivariate distributions (both parametric and nonparametric) assume that all the univariate marginals are of the same type/family of distribution. But in reality, different grid locations show different statistical behavior which may not be modeled best by the same type of distribution. In this paper, we propose a new multivariate uncertainty modeling strategy to address the needs of uncertainty modeling in scientific datasets. Our proposed method is based on a statistically sound multivariate technique called *Copula*, which makes it possible to separate the process of estimating the univariate marginals and the process of modeling dependency, unlike the standard multivariate distributions. The modeling flexibility offered by our proposed method makes it possible to design distribution fields which can have different types of distribution (Gaussian, Histogram, KDE etc.) at the grid locations, while maintaining the correlation structure at the same time. Depending on the results of various standard statistical tests, we can choose an optimal distribution representation at each location, resulting in a more cost efficient modeling without significantly sacrificing on the analysis quality. To demonstrate the efficacy of our proposed modeling strategy, we extract and visualize uncertain features like isocontours and vortices in various real world datasets. We also study various modeling criterion to help users in the task of univariate model selection.

Index Terms—Uncertainty visualization, probability distribution, probabilistic feature, statistical modeling, copula

1 Introduction

Most of the numerical simulations which are used to model complex real world physical phenomenon generate uncertain data. The lack of a proper ground truth and/or simulation parameter knowledge are some of the common causes of uncertainty. In order to avoid making erroneous decisions using such data, it is important to incorporate the uncertainty into the analysis process itself. For example, tasks like feature extraction and visualization should reflect the effect of uncertainty in the data. With recent advances in computing power and resources, scientists are able to model the uncertainty by running ensemble of simulations with varying experiment parameters, thus, generating multiple realizations of the same physical phenomenon. These multiple realizations/values at each of the spatially sampled points (grid locations) represent the uncertainty in that location and are often modeled as stochastic random variables. Various approaches have been proposed [2,33,38] to extract probabilistic/uncertain features from such a field of random variables using standard statistical tools.

An important property to be taken into account while modeling uncertainty in spatially sampled scientific datasets is the correlation among the grid locations due to the inherent local data continuity [36,37,39,41]. Therefore, various multivariate distribution models have been proposed to model the uncertainty in the data which can preserve the dependency/correlation among the random variables at each grid locations. The choice of the statistical model plays an important role in any distribution driven uncertainty analysis. Among the parametric models, the multivariate Gaussian distribution is the most popular choice [40,41], while, most common nonparametric models are histograms, empirical distributions and kernel density estimates (KDE) [2,39]. Multivariate Gaussian distributions are useful to model the multivariate dependency but it has the basic assumption that the univariate marginal distributions (at the grid locations) are all Gaussians. This can lead to misleading results if the underlying distribution at a location does not follow a normal distribution. Pöthkow et al. [39] highlighted this problem in parametric models and extended their work to consider nonparametric multivariate models which fits the data better. However, there are two possible challenges with such nonparametric models. First, the estimation and subsequent analysis of multivariate nonparametric distribution models is computationally intensive (both in terms of time and memory footprint). Second, they are susceptible to generate biased results if an over-fitted nonparametric model is chosen for a sample which shows high confidence of following a particular parametric model. A general problem with all standard multivariate models (both parametric and nonparametric) is that they consider all the univariate marginals to follow the same family/class of distribution. But in reality, not all the locations in the data show uncertainty trends which can be best modeled by the same type of distribution. For example, a simple statistical normality test can reveal the fact that not all the grid locations show equal confidence of following a normal distribution. Some locations show high certainty, whereas, others show very low certainty. Recently, Bensema et al. [4] in their modality driven analysis, have shown that the ensemble distributions at different locations can vary significantly. Therefore, there is a need to adopt a different multivariate strategy to model uncertainty in scientific datasets, which is flexible enough to model the univariate marginals by different types of distributions as well as be able to model the multivariate dependency among the random variables.

In this paper, we propose a new uncertainty modeling and analysis technique for scientific datasets which can separate the estimation of multivariate dependency structures from the process of estimating univariate marginal distributions at each grid location. Our technique is based on a statistically sound multivariate modeling scheme called *Copula*, which has been widely used in the field of financial modeling and machine learning [6,11,13,30,44]. The proposed technique makes it possible to choose the best possible univariate distribution to model the uncertainty at each grid location. The resulting distribution field, which can have different types of distribution (Gaussian, Histogram, KDE, GMM etc.) at different grid locations is henceforth referred to as a *mixed distribution field* in our work. In fact, copula-based techniques can accommodate any univariate distribution type, as long as it is a continuous distribution with a valid cumulative density function (CDF). A major advantage, for example, of using such a flexible strategy is that

- *Subhashis Hazarika and Han-Wei Shen are with the GRAVITY research group, Department of Computer Science and Engineering, The Ohio State University. E-mail: hazarika.3@osu.edu, hwshen@cse.ohio-state.edu.*
- *Ayan Biswas is with the Los Alamos National Laboratory. E-mail: ayan@lanl.gov.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

we can choose the aforementioned computationally intensive nonparametric models only for those grid locations where parametric models fail to model with sufficient confidence. This can significantly reduce the computational cost without compromising the quality of the analysis. To demonstrate the effectiveness of our flexible scheme, we propose copula-based techniques to visualize uncertain/probabilistic features in the resulting mixed distribution fields. We introduce methods to compute the level-crossing probability values to extract probabilistic isocontours in mixed scalar distribution fields as well as vortex-core probabilities to determine uncertain vortex features in mixed vector distribution fields. We compare the results of our probabilistic features against the results generated by existing methods which use standard multivariate models and evaluate them based on correctness and computational complexity. The selection of the optimal univariate model at each grid location is an important task. Since there are multiple statistical tests available at our disposal, we guide the users by identifying the ones that are useful for scientific data modeling and highlight their advantages. To summarize, the contribution of our work is threefold:

- We propose a statistically sound copula-based uncertainty modeling technique which makes it possible to model the uncertainty at each grid location independently with different types of univariate distribution while preserving the spatial correlation at the same time.
- We offer guidelines for the task of univariate distribution model selection based on criteria like goodness-of-fit, model complexity and eventual goal of analysis.
- We propose specific copula-based methods to compute level-crossing probability and vortex core probability in a mixed distribution based representation of uncertain data.

2 Related Work

Distribution Driven Uncertainty Analysis and Visualization: Uncertainty analysis and visualization of scientific datasets is considered as one of the top few challenges in our field [20, 21, 55]. Over the past few years, there have been significant research contributions towards visualizing and modeling uncertain data [5, 34, 42]. Here, we specifically discuss only the works that use statistical distributions to model uncertainty and are related to our proposed technique. However, besides the use of standard probability distributions, other statistical tools like data-depth have also been used in the field to perform quantile-level analysis of uncertain features [17, 54], but are not directly related to our proposed strategy.

Distributions have been used to address different aspects of uncertainty analysis. Techniques were proposed to visualize datasets where each grid locations have data distributions rather than single data point [25, 26, 31]. The use of distributions to model the uncertainty in data and its subsequent analysis to extract probabilistic features have gained popularity in the recent past. Pöthkow et al. [38] proposed the concept of level-crossing probability (LCP) to compute probabilistic isocontours in uncertain data. LCP computes the probability of an isocontour passing through a cell of the data. It assumed that the data at each grid location follows a Gaussian distribution and there is no correlation among the grid location. This approach was later extended to introduce the local spatial correlation [41]. An alternative method of computing first-crossing probability integrated with ray-casting algorithm to visualize probabilistic isocontours with spatial correlation was proposed by Pfaffelmoser et al. [36]. Expensive Monte Carlo computations were replaced by fast techniques such as maximum edge crossing probability and linked pairs to speed up computation for interactive visualization of probabilistic isocontours [40]. Apart from uncertain isocontours, methods have also been proposed to extract and analyze features like vortex and critical points using statistical distributions as uncertainty modeling tools [23, 29, 32, 33, 35]. All these works assumed that the data at each grid location follow a Gaussian distribution. Pöthkow et al. [39] later extended uncertainty analysis to include nonparametric models. Athawale et al. [1, 2] proposed closed-form analytic solution to compute uncertain isocontours in nonparametric distribution models. Pfaffelmoser et al. [37] performed detailed study on the properties of global and local correlation in uncertain data. Schlegel et al. [47] proposed Gaussian process regression based interpolation scheme and investigated the influence of correlation functions on the level-crossing probabilities in Gaussian random field. Another class of distributions widely used are the Gaussian Mixture Models (GMM), a type of semi-parametric model. Liu et al. [24] used GMMs to approximate large ensemble datasets for volume visualization. Dutta et al. [9], on the other hand, used GMMs to perform feature tracking in time varying data. Despite its flexibility, it is computationally very expensive to estimate GMMs for multivariate models where correlation has to be accounted for.

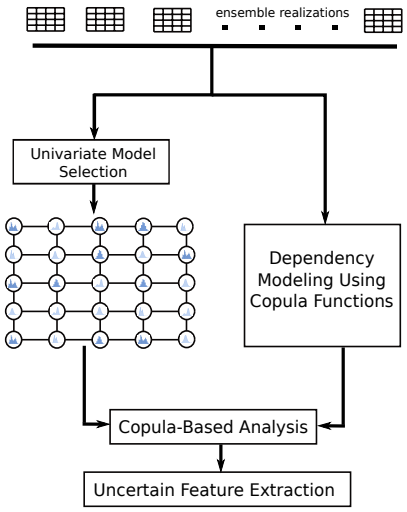


Fig. 1: Overview

In general, depending on the modeling scenario, all distribution types have their own advantages and disadvantages. To the best of our knowledge, none of the current distribution-driven feature analysis works try to utilize the benefits of using different distribution types to model the uncertainty at different locations. Our proposed copula-based multivariate modeling strategy facilitates such flexibility in use of distribution models while preserving the local spatial dependency at the same time.

Copula-Based Methods: The term *Copula* was derived from the latin word *copulare*, which means to connect or to join. The relationship between copula and generic multivariate functions was first postulated by Sklar in 1959 [50]. This led to its wide-spread popularity as a flexible multivariate dependency modeling framework, specially, in the field of financial modeling and risk management [6, 13, 30]. Schmidt [48] have provided detailed explanation of the working of copula and its possible applications in the article titled, *Coping with Copula.* Of late, copula-based methods (specially Gaussian copula) are increasingly used to address many machine learning problems [11, 44]. Copula-based methods have been used to perform independent component analysis [22], component analysis [27], mixture models [15, 52] and dependency seeking clustering [45]. Besides, copula related analysis have been used in the field of Uncertainty Quantification as well [3]. The field of scientific visualization, especially any multivariate distribution based analysis can greatly benefit from the flexibility offered by copula functions.

3 Motivation and Overview

Motivation: Our proposed technique is specifically tailored to meet the needs of uncertainty modeling in scientific datasets. Some of the key aspects of distribution-based uncertainty modeling in scientific datasets that need to be taken into account are as follows:

1. **Univariate Model Selection:** A single class/type of distribution may not be the best way to model the underlying uncertainty at each spatially sampled grid location because different locations show different statistical properties. Therefore, to identify the optimal distribution model, some form of statistical test must be performed at each grid location.

2. **Dependency Modeling:** Since scientific datasets have the inherent property of local data continuity, there exists a spatial correlation among the values at the grid locations. Therefore, any uncertainty modeling technique must consider the spatial correlation/dependency among the univariate distributions at each grid location.

3. **Computational Complexity:** The size of the distribution field (indicated by the number of univariate distribution models) is dictated by the resolution of the dataset. Performing tasks like feature extraction and/or query on large distribution fields can become computationally expensive (in terms of time and memory footprint) if complex models are used. There needs to be a balance between the degree of correctness and the overall computational expense of generating the results.

Standard multivariate distribution models are not flexible enough to meet all of these requirements at the same time because the multivariate dependency structure is strongly coupled with the type of univariate marginal distribution. Our proposed copula based technique provides a framework to separate the two tasks i.e, univariate model selection and dependency modeling. As a result, users have the flexibility to choose an optimal distribution at each location to model uncertainty and still preserve the correlation among the locations. This independent execution of the two tasks and the fact that we can now use the computationally expensive models only when it is absolutely necessary, in turn, helps us to achieve faster analysis time while generating similar, if not more statistically reliable results.

Overview: A high-level overview of our proposed idea is shown in Figure 1. Ensemble realizations are used to select an optimal distribution model at each grid location by performing standard statistical tests. The resulting *mixed distribution field* can have different types of distribution models at each grid location as illustrated in Figure 1. On the other hand, the spatial correlation among the neighboring grid locations is modeled from the ensemble realizations separately using Coupla functions (Gaussian copula). Finally, using a sequence of quantile transformations we model the desired multivariate distribution, comprising of mixed marginal distributions and the preserved dependency structure. We use this technique to extract uncertain features like isocontours and vortices in various mixed distribution fields. Our proposed method uses a Monte-Carlo based integration technique to model the multivariate distribution. We compare our results with other Monte-Carlo based techniques that use multivariate Gaussian models [41] and other nonparametric models (histograms, KDE) [39]. In this paper, we first explain the concept of Copula along with our proposed dependency modeling strategy. We follow it up with a discussion about the task of univariate model selection and subsequent implementation to extract uncertain features.

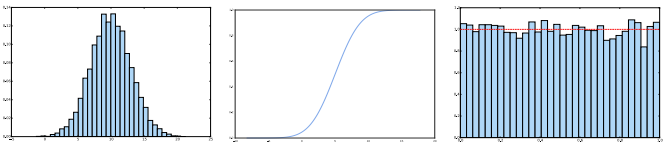
4 Concept of Copula

Copula functions are used as tools for modeling dependence/interrelation of several random variables. The idea of copula is closely tied with the definition of multivariate distribution. In the subsequent paragraphs of this section we first revisit some of the important definitions and properties of multivariate distributions which are relevant to understand the concept of copula and then formally introduce the copula functions along with an example to show dependency modeling using copula.

4.1 Multivariate Probability Distribution

Consider a set of d real valued random variables, $X_1, X_2, ..., X_d$. The joint cumulative distribution function (CDF) is defined as,

$$F(x_1, x_2, ..., x_d) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_d \leq x_d) \quad (1)$$



(a) (b) (c)

Fig. 2: Quantile transformation property of univariate distributions: (a) distribution of Gaussian samples, (b) CDF of the corresponding Gaussian samples, (c) output of the CDF for the given samples, which is uniformly distributed.

where, x_i is a realization of the random variable X_i and $P(.)$ is the joint probability density function. The joint CDF, $F : \mathbb{R}^d \rightarrow [0,1]$, maps the multivariate random variable to a scalar value in between 0 and 1. A marginal distribution $F_i(x_i)$ can be obtained from the joint distribution by marginalizing out the other dimensions (setting probability to 1) except for the i^{th} dimension. A marginal CDF, which is essentially a univariate distribution, has two interesting transformation properties [28], which are used in our proposed copula-based technique.

- ***Property 4.1*:** If U is a uniform random variable (i.e, $U \sim U[0,1]$) and F_X is a univariate CDF then its inverse function, $F_X^{-1}(U)$, corresponds to the random variable X, (i.e, $F_X^{-1}(U) \sim X$)

$$P(F_X^{-1}(U) \leq x) = F_X(x) \quad (2)$$

- ***Property 4.2*:** If a real valued random variable X has a continuous cumulative distribution function F_X then

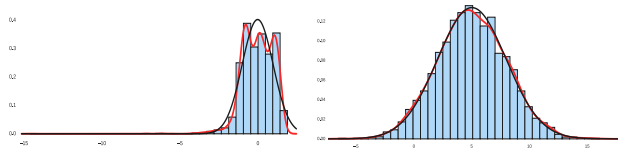
$$F_X(X) \sim U[0,1] \quad (3)$$

Of the two stated properties, the former is very well known and has been extensively used to simulate random variables with arbitrary CDF from uniform distributions. But its reverse scenario, the second property, which states that we can transform any continuous CDF to a uniform distribution is not frequently used. Proofs for these properties can be found in [44]. Figure 2 shows the results of these transformation properties for a Gaussian distribution. Figure 2(a) shows the distribution of 1000 samples drawn from a Gaussian distribution. Figure 2(b) shows the CDF of the Gaussian distribution, say Φ. Property 4.2 states that the output of $\Phi(x)$ for all samples in (a) follows a uniform distribution as shown in Figure 2(c). Conversely, by property 4.1, when we feed a uniform distribution to the inverse CDF i.e, $\Phi^{-1}(u)$, we get back the samples drawn from the original distribution. In our proposed method, we use Property 4.2 to model the copula functions from the initial samples, whereas, Property 4.1 is used to transform the copula models to the desired distribution form.

4.2 Copula

A d-dimensional copula $C : [0,1]^d \rightarrow [0,1]$ is a CDF with uniform marginals. For d-uniform random variables, it can be denoted as $C(u_1, u_2, ..., u_d)$.

Every joint CDF in $\mathbb{R}^d$ inherently embodies a copula function. If we choose a copula and some marginal distributions and entangle them in the right way, we will end up with the proper multivariate distribution function. Therefore, Copula functions allow splitting the problem of joint distribution estimation into two parts, 1) estimate of the marginal distributions and 2) estimation of dependencies between random variables. Due to this decoupling, it is possible for a joint distribution to have marginal distributions from different distribution families. This powerful connection between copula functions and general multivariate distribution functions was formalized by *Sklar's theorem* [50] and stated as follows:



(a) (b)

Fig. 3: Bivariate Example: The univariate distributions of the initial 1000 samples. (a) for the first random variable X, (b) for the second random variable Y. The black plot shows the estimated Gaussian distribution for the given samples while the red plot shows the estimated KDEs. For X, KDE is a better fit than the Gaussian, whereas, for Y, Gaussian is a better choice of model than KDE.

Theorem 1. ***(Sklar's Theorem)***

1. *Let F be a joint CDF with marginals $F_1,...F_d$. Then, there exists a copula $C : [0,1]^d \rightarrow [0,1]$ such that*

$$F(x_1,...,x_d) = C(F_1(x_1),...,F_d(x_d)), \forall x_i \in [-\infty,+\infty] \quad (4)$$

   Furthermore, if the marginals are continuous, then the copula is unique.

2. *Conversely, if C is a copula and $F_1,...,F_d$ are univariate CDFs, then F defined as in equation 4 is a multivariate CDF with margins $F_1,...,F_d$ and copula C.*

Equation 4 can be also rewritten to compute the copula C itself. Using $F_i \cdot F_i^{-1}(x) \geq x$, we obtain

$$C(u_1,...,u_d) = F(F_1^{-1}(u_1),...,F_d^{-1}(u_d)) \quad (5)$$

Equation 4 states that, given the univariate marginal CDFs and a copula function, we can derive the original multivariate CDF. This is the major idea behind many copula-based modeling techniques (including ours) when the copula function C is known. But the obvious question is what copula function to use? Equation 5 gives the formula to compute the copula function from existing multivariate distributions. Copulas extracted using equation 5 are called *implicit copulas*. The most popular implicit copula is the *Gaussian* copula, which is computed from multivariate Gaussian distribution and is based on the correlation matrix which can capture the multivariate dependency among the random variables. Therefore, it is a good choice for uncertainty modeling in scientific data. In recent years, Gaussian copula have found wide-spread usage in the field of machine learning to perform dependency based clustering and classification in multivariate models [44]. We discuss more in details about the properties of Gaussian copula in the next section along with an example to show its practical usage in dependency modeling. However, it is important to know that there is a good number of other predefined copulas as well, referred to as *explicit copulas*. Explicit copulas (like *Gumbel*, *Clayton* and *Frank* copula) are designed keeping in mind special statistical tasks at hand [13,48]. They have gained sufficient popularity in the field of financial modeling. Unlike Gaussian copula, usage of explicit copulas to perform uncertainty analysis will not be straightforward because of the predefined objective of their usage. However, it is an interesting field of future research to see how the explicit copulas can be put to use to solve scientific visualization problems.

4.3 Gaussian Copula-based Dependency Modeling

The copula extracted from a multivariate Gaussian distribution is called a *Gaussian* copula. More formally, for a d-dimensional multivariate Gaussian distribution $\mathcal{N}_d(\mu,\Sigma)$ with mean vector μ, covariance matrix Σ and correlation matrix ρ, the corresponding copula derived using equation 5 is called a Gaussian copula with parameter ρ. We denote a Gaussian copula by C_ρ^G and can be written as,

$$C_\rho^G(u_1,...u_d) = \Phi_\Sigma(\Phi^{-1}(u_1),...,\Phi^{-1}(u_d)) \quad (6)$$

where, Φ represents the CDF of the normal distribution and Φ^{-1} its inverse function. An important property to note here is that C_ρ^G depends on the correlation but not on the mean or the variance of the marginals. This means that the class of all Gaussian variables having correlation matrix ρ share the same copula irrespective of means and variances C_ρ^G [28]. In fact, the real benefit of copula lies in the fact that besides retaining the multivariate dependency (via, ρ, for Gaussian copula), the corresponding marginals are uniformly distributed in the interval $[0,1]$. Because of this property, the uniform marginals of the copula can be fed into the inverse CDFs of the desired form (using Property 4.1) to successfully estimate any multivariate distribution with arbitrary marginals while preserving the dependence structure. Though a Gaussian assumption is made to preserve the dependency at the beginning, the subsequent transformations lead to a form where the final model no longer holds the Gaussian properties, in fact, such models are also popularly termed as *meta-Gaussian* models [46].

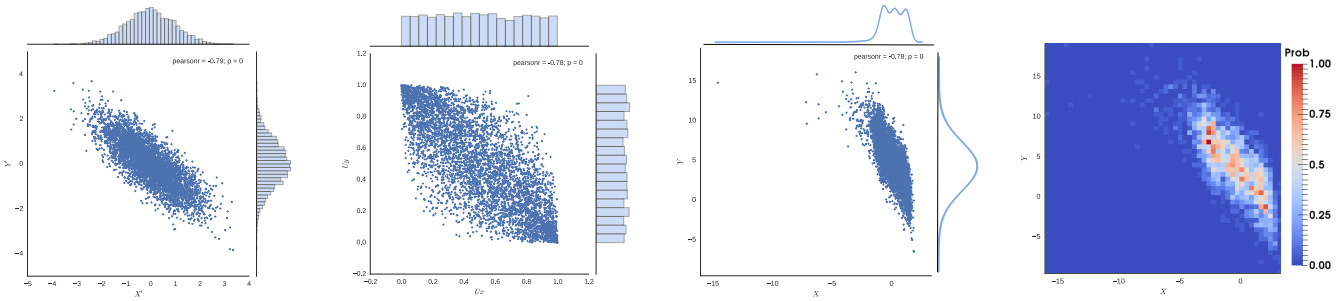
Next, we illustrate the process of modeling a Gaussian copula for a simple bivariate dataset (i.e, the numerical estimation of Equation 6). Consider a bivariate distribution of two negatively correlated random variable X and Y, where X follows a mixture of three non-Gaussian distributions and Y follows a Gaussian distribution. We generated 1000 bivariate samples with correlation coefficient -0.8. Figure 3(a) and (b) shows the univariate marginal distributions of X and Y respectively. As shown by the black plot in Figure 3(a), the data does not fit a Gaussian representation, while a nonparametric KDE (red plot) can capture the distribution better. On the other hand, for Y, as shown by the black plot in Figure 3(b) a Gaussian model is a good assumption for the data. While the KDE for Y (red plot) also fits the data very well, we choose the Gaussian model because it is a much more generic representation of the underlying data. We discuss more about the benefits of selecting a parametric model over a nonparametric model in the next section while explaining the model selection goals. Let, F_X be the CDF of the estimated KDE of X and F_Y be the Gaussian CDF of Y. The multivariate relationship between X and Y will be modeled using a Gaussian copula C_ρ^G where $\rho = -0.8$. The respective CDFs and the correlation matrix are sufficient to successfully model the bivariate distribution using Gaussian copula. We now generate new samples to prove that we have successfully modeled the bivariate distribution with different marginal distribution representations. It involves a three step process as shown below and are illustrated in Figure 4.

1. Generate N (5000) new bivariate samples from a bivariate standard normal distribution with mean vector $[0,0]$ and correlation matrix corresponding to $\rho(= -0.8)$. As shown in Figure 4(a) the samples currently preserve only the correlations (sample correlation = -0.79).

2. Using Property 4.2, transform the samples to uniform distributions by applying the inverse standard normal distribution function on the respective variables. This step models the corresponding Gaussian copula for this example case. The correlation of the samples are still preserved (sample correlation = -0.78) but their marginals have been transformed to uniform distributions, as is shown in Figure 4(b).

3. Final step is to the transform uniform samples to a form which respects the previously estimated marginals, F_X and F_Y respectively. Using property 4.1, the uniform marginals are fed into the respective inverse CDF's (F_X^{-1} and F_Y^{-1}) to transform the samples into the desired form. The final bivariate samples respect the correlation among the variables (sample correlation = -0.78) as shown in Figure 4(c).

This shows that using the transformation properties of univariate CDFs and the Gaussian copula-based dependency modeling we are able to successfully estimate the bivariate distribution where both the marginals are represented by unique distribution types. The final dependency structure is not strictly Gaussian (meta-Gaussian) but still preserves the correlation structure of the original 1000 samples, as shown by the joint distribution in Figure 4(d). Moreover, the value



(a) Step 1 (b) Step 2 (c) Step 3 (d) Original joint distribution

Fig. 4: Bivariate Example: The three steps of performing a copula-based sampling. (a) Step 1: Generate N (=5000) samples from a bivariate standard normal distribution. (b) Step 2: Model the corresponding Gaussian copula by transforming the points to uniform marginals using property 4.2. (c) Step 3: Finally, applying property 4.1, transform the uniform samples to desired marginal forms. (d) the joint distribution of the original 1000 samples.

of joint entropy [7] (a popular information-theoretic measure of the uncertainty associated with a set of variables) for the new and the original sample is very similar (5.73 for the new samples and 5.65 for the original samples).

5 Univariate Model Selection

The task of selecting a proper model to represent the uncertainty at each grid location is vital for subsequent uncertainty analysis. The choice of model often varies on the types of data and the eventual goal that needs to be achieved via modeling. Statistical models are broadly classified into two categories, *parametric* and *nonparametric*. Parametric models are based on assumptions about the distribution of the underlying population from which the sample was taken. Nonparametric models do not rely on any assumptions about the shape or parameters of the underlying population distribution. Depending on factors like initial sample size, type of post-hoc analysis to be performed and computational complexity both forms of models have distinct advantages and disadvantages. Therefore, it is highly advisable to judge the pros and cons before deciding to select specific models.

A popular statistical maxim is that nonparametric models generally have less *statistical power*[1] for the same sample size than the corresponding parametric model which shows high certainty [12]. Therefore, any statistical test showing high certainty for a parametric model should be preferred over nonparametric models. However, if an appropriate parametric model cannot be ascertained with sufficient confidence, often a nonparametric form of model is recommended. For small sample sizes, statistical tests for parametric models often fail, in which case, nonparametric models are the only option. But one must be aware of the potential side-effects of an over-fitted nonparametric model. When the models are used for some Monte-Carlo sampling based post-hoc analysis, the risk of over-fitted nonparametric models generating biased results increases. Cost of working with the eventually selected model is also another important factor. While parametric forms have a fixed (usually low) number of parameters, the nonparamteric models, despite the name, have to store parameters in proportion with sample data size. As a result, working with the nonparametric models may require more computational time compared to the parametric models. For large sample sizes, working with nonparametric models become computationally very expensive (both in terms of time and memory).

Many standard statistical tests currently exist to decide which model can best represent a given sample. However, often a single test cannot inspect all the aforementioned concerns. Therefore, users have to carefully design their tests based on their requirements. Depending on the application and scale of operation, model selection tasks can be as simple as graphical validation of the shapes of distributions to as complex as solving an optimization function with desired requirements as the function variables. In this section, we put forward some of the most commonly used model testing practices prevalent in the field and the ones that are specifically useful for scientific datasets.

Normality Test: Checking for a Gaussian behavior is the most common and useful test that can be made before testing any other distribution because most data simulating a natural phenomenon is Gaussian in nature [18]. However, if the underlying distribution is not normal, making a normality assumption to represent the data can lead to erroneous results. In statistics, normality tests are used to compute how likely it is for a random variable of a dataset to be normally distributed. Studies have shown that for the same sample size Shapiro-Wilk test [49] is the most powerful (i.e, *statistical power*) normality test [43]. The Shapiro-Wilk test returns a likelihood value, commonly referred to as pValue, which lies between 0 and 1. Small pValues lead to the rejection of normality whereas a value of 1 ascertains normality with high confidence. A pValue in the range of $[0.05, 0.1]$ is often considered as a good threshold value to make a call to decide normality. Data showing pValues less than the threshold normality value can be checked for further parametric distributional properties or select a suitable nonparametric model.

Generalized Goodness-of-fit Test: While the normality test tells us whether a dataset follows a normal distribution or not it does not offer a means to check the sample for multiple distribution types. Kolmogorov-Smirnov goodness-of-fit test (KS test) [51] is a more generic platform for such comparative validation. It compares the CDF of the distribution we want to test for against the empirical CDF (ECDF). Goodness-of-fit is decided by how close the CDF of a distribution is to the ECDF. If $F(x)$ represents the CDF of the hypothesized distribution and $F_e(x)$ represents the ECDF, then the KS test measure is given as,

$$K = sup_x |F(x) - F_e(x)| \tag{7}$$

where *sup* stands for supremum, which means the greatest. This is a more generalized statistical test which lets us test for any continuous distribution as along as it has a valid CDF (i.e $F(x)$). Several goodness-of-fit test are in fact refinement of the KS test. One big advantage of the KS test is the ability to compare a parametric model versus a nonparametric model which is not provided by many other complex statistical test.

Bayesian Information Criterion Bayesian Information Criterion (BIC) [14] is a popular model selection tool for selecting among a finite set of parametric models. It is based on the log likelihood of a given model on the sample data. It is defined as,

$$BIC = -2L_p + p\log(n) \tag{8}$$

where n is the sample size, L_p is the maximized log-likelihood of the

[1] *Statistical power* of any test is defined as the probability that it will reject a false null hypothesis.

chosen model and p is the number of parameters in the model. A low BIC value indicates a better model. BIC attempts to address the risk of over-fitting by introducing a penalty term $p\log(n)$, which grows with the number of parameters. This eliminates overly complicated models with large number of parameters. BIC severs as a good tool for our model selection task when the desired distributions are all parametric.

6 Uncertain Feature Extraction

In this section, we put together the modeling techniques to extract and visualize uncertain features. Our copula-based method allows us to separate the process of estimating the univariate model at each grid location from the dependency modeling of the locations. Using the model selection guidelines proposed in Section 5, we create a mixed distribution field by independently modeling the univariate distributions at each grid locations with the desired distribution type (we store the CDF representations rather than the PDF form). On the other hand, dependency is modeled by first computing the correlation matrix of a local neighborhood from the initial sample values at the corresponding grid locations and then transforming to the more general Gaussian Copula representation as discussed in Section 4.3. The spatial correlation among the locations diminishes with Euclidean distance, therefore it is sufficient to consider the correlation within localized spatial regions [36, 37, 39, 41, 47]. Using this proposed strategy, we focus on the extraction of two specific types of features, uncertain isocontours in scalar distribution fields and uncertain vortices in vector distribution fields.

6.1 Copula-based Uncertain Isocontour Extraction

Level-crossing probability (LCP) is a popular uncertain isocontour detection measure for distribution based data [36, 38–41]. It involves computing the probability of the level-set/isocontour of a given isovalue passing through the cells of the dataset. Similar to the method proposed by Pöthkow et al. [39,41], we adopt a Monte-Carlo based sampling strategy with the difference that we use a copula-based sampling method to handle the mixed distributional representations.

Consider a 2D cell with four neighboring vertices V_1, V_2, V_3, V_4. Let, $F_1(x), F_2(x), F_3(x), F_4(x)$ be the respective CDFs representing the uncertainty at the four vertices (as mentioned before the CDFs can be in the form of any continuous family of distribution). Let $\rho_{4\times4}$ be the correlation matrix which captures the multivariate dependency in the cell. Using the three step sampling method as discussed in the example in Section 4.3, we draw multivariate samples for the considered cell. Using the cases of the traditional marching cube algorithm, each of the final multivariate samples, representing the cell configuration, are checked to see if a level-set of the given isovalues passes through it or not. The number of times we find such a cell (multivariate sample) out of the total number of samples drawn, determine the level-crossing probability of the cell. Computing this for all the cells give us a density field that highlights the regions through which the level-set of an isovalue is most likely to pass through. As can be seen in this approach, it is really important to model the uncertainty at each grid location in an optimal way to get a result which is neither biased or erroneous and also easy to compute for a dataset of high resolution. The procedure to compute the LCP of a single 2D cell is formalized in the pseudocode in Algorithm 1

Algorithm 1 LCP computation for a single 2D cell

1: $S[numSamples] \leftarrow getSamples(\mathcal{N}(\mathbf{0}, \rho_{4\times4}))$ ▷ $\mathbf{0} =< 0,0,0,0 >$
2: $numCrossing \leftarrow 0$
3: **for all** $\mathbf{s}$ in $S[.]$ **do** ▷ $\mathbf{s} =< s_1, s_2, s_3, s_4 >$
4: $\quad \mathbf{u} \leftarrow \Phi(\mathbf{s})$ ▷ uniform samples $\mathbf{u} =< u_1, u_2, u_3, u_4 >$
5: $\quad s_i = F_i^{-1}(u_i) \quad \forall i \in \{1, ...4\}$
6: $\quad$ **if** isCrossing($\mathbf{s}, isoValue$) **then**
7: $\quad\quad numCrossing \leftarrow numCrossing + 1$
8: $LCP \leftarrow numCrossing / numSamples$

First step in the algorithm is to generate *numSample* multivariate samples from a standard normal distribution with correlation matrix $\rho_{4\times4}$. These samples currently only preserves the multivariate dependency among the four cell vertices. Second step is to transform the samples to uniform marginals using the Property 4.2. This is shown in step 4 of the pseudocode 1. The third and the final step involves transforming the uniform marginals $u =< u_1, u_2, u_3, u_4 >$ to their correct forms using the inverse of the predetermined CDF functions (i.e, $F_1^{-1}, F_2^{-1}, F_3^{-1}, F_4^{-1}$). The corresponding 3D version of this implementation will have 8 neighboring random variables for each voxel.

6.2 Copula-based Uncertain Vortex Detection

The concept of copula-based analysis in mixed distribution datasets can also be extended to uncertain vector datasets. We use it to extract vortex probabilities along the lines of what Otto et al. [33] proposed. It involves a Monte-Carlo based algorithm of sampling vector fields and using a known vortex detection method (like λ_2-criterion or $Q-$criterion) to compute the probability of observing a vortex core at each grid location.

Unlike the uncertain isocontour extraction approach, vortex detection is a per grid location based computation. Therefore we have to generate samples for each grid location rather than a cell. Sampling at each grid location also involves considering the correlation among the neighboring locations. For a regular 2D dataset, there are at most 4 connected neighbors and each consists of a vector with two components. Therefore, there are 10 (5×2) random variables to be taken into account. Consider a grid location V_0 in a 2D dataset with neighbors V_1, V_2, V_3, V_4. Let, $F_{u_0}(x), F_{u_1}(x), ..., F_{u_4}(x)$ be the marginal CDFs of the respective u-velocities while $F_{v_0}(x), F_{v_1}(x), ..., F_{v_4}(x)$ be the corresponding v-velocity CDFs. Let, $\rho_{10\times10}$ be the correlation matrix for this neighborhood. We then apply a similar copula-based Monte-Carlo sampling to draw sample vectors for location V_0. We used the λ_2 criterion to estimate vortex core probability. Regions with λ_2-criterion below a threshold value (generally 0) is considered highly likely to have a vortex core. Therefore, for all the sampled vector fields we compute the probability of a location having λ_2-criterion less that 0. The resulting density field serves as a visualization of uncertain vortex cores in a mixed vector distribution data. The corresponding procedure to compute the vortex core probability (VCP) for a single grid location is formalized in the pseudocode in Algorithm 2

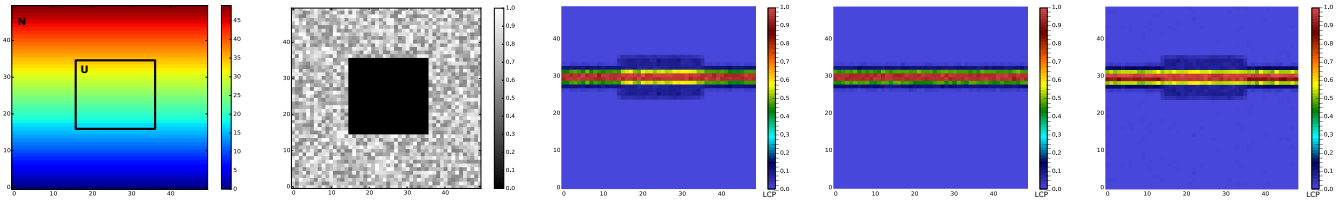
Algorithm 2 Vortex core probability for a grid location

1: $S[numSamples] \leftarrow getSamples(\mathcal{N}(\mathbf{0}, \rho_{10\times10}))$
2: $numVortex \leftarrow 0$
3: **for all** $\mathbf{s}$ in $S[.]$ **do** ▷ $\mathbf{s} =< s_0, s_1, ..s_9 >$
4: $\quad \mathbf{u} \leftarrow \Phi(\mathbf{s})$ ▷ $\mathbf{u} =< u_0, u_1, ..., u_9 >$
5: $\quad s_i = F_i^{-1}(u_i) \quad \forall i \in \{0, 1, ...9\}$
6: $\quad$ **if** $\lambda_2(\mathbf{s}) < 0$ **then**
7: $\quad\quad numVortex \leftarrow numVortex + 1$
8: $VortexProb. \leftarrow numVortex / numSamples$

7 Results and Discussion

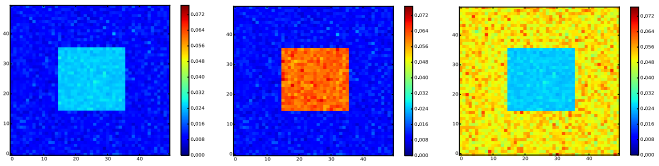
To illustrate the effectiveness of our copula-based analysis on a mixed distribution field, we first show the results on a synthetic dataset. We then apply our method on three different real world ensemble simulation datasets. All computations were performed on a standard workstation PC (Intel i7 at 3.40GHz and 16GB RAM) and implemented using C++.

Synthetic Data: We created a synthetic ensemble dataset of size 50×50 by generating 1000 samples from either a normal distribution or a uniform distribution at each grid location. The mean of the distribution at a location is taken to be the corresponding y-coordinate value at that location and a fixed standard deviation value (1.15) is used across all locations. Figure 5(a), colored by the y-coordinate values, illustrates our data creation process. The black rectangle in the image encloses the locations where we used a uniform distribution to generate the samples, while a normal distribution was used for rest of the locations . Also, a high correlation value of $\rho = 0.8$ among the neighboring locations was used to generate the samples. The result of the Shapiro-Wilk normality



(a) (b) (c) (d) (e)

Fig. 5: Synthetic Data: (a) Illustrates the synthetic data creation process. Samples are drawn from a uniform distribution for the locations inside the rectangle and from a normal distribution for outside. (b) shows the result of the Shapiro-Wilk normality test on the initial samples at each grid location. We compute the level-crossing probability (LCP) for isovalue 30 (c) using our proposed method on a mixed distribution field, (d) using the multivariate Gaussian distributions [41] and (e) using multivariate histograms.



(a) (b) (c)

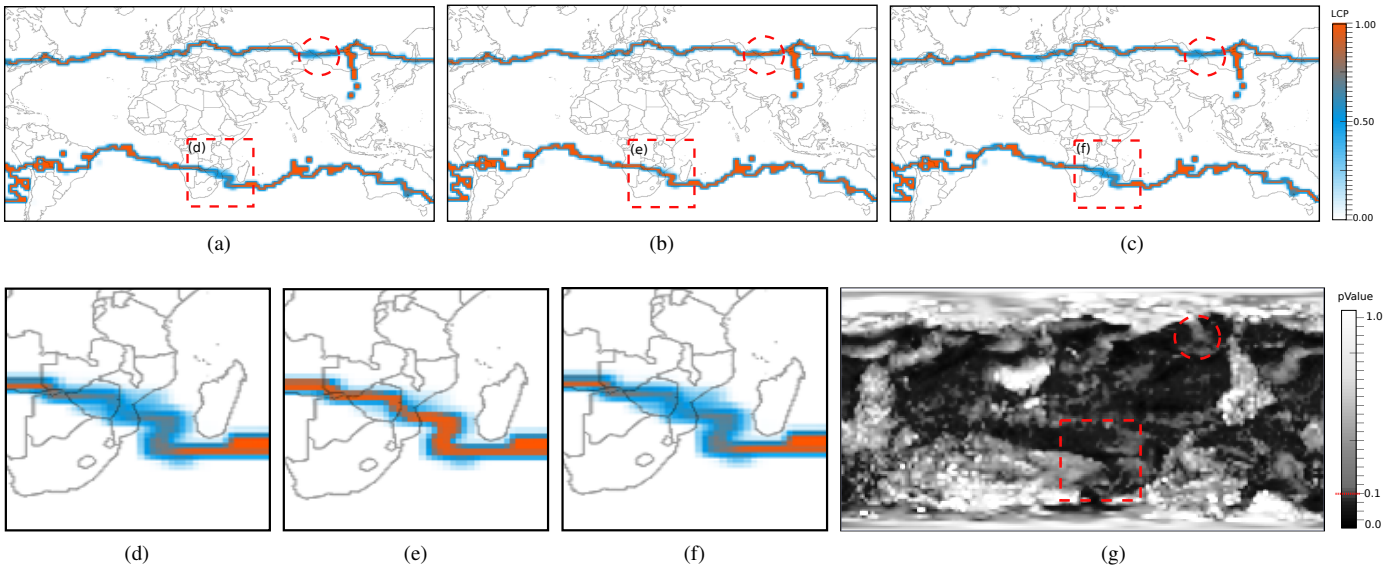
Fig. 6: Synthetic Data: The KS test for goodness-of-fit, reflects how good are the selected models at each location. The lower the KS test value, the better. The KS test values at each location for (a) mixed distribution field (Gaussian and histogram), (b) only Gaussian at all location and (c) only histogram at all locations is shown in this figure.

test is shown in Figure 5(b). A low p-Values in the center block indicates that a Gaussian distribution is not a good choice for modeling the uncertainty at those locations. In this example, we decided to model such locations (i.e, p-Value < 0.1) with histograms. The rest of the locations are modeled using Gaussian distributions.

We tested our proposed copula-based analysis technique on the resulting mixed distribution field to compute the level-crossing probability (LCP) for isovalue 30 using the technique outlined in Algorithm 1. Figure 5(c), (d) and (e) show the LCP fields generated by our method on a mixed distribution field, by using standard multivariate Gaussian distribution [41] and by using multivariate histograms [39] respectively. Since the mean and the standard deviations for the normal and the uniform distributions are the same, Figure 5(d) shows a much smoother result and does not reflect the high uncertainty corresponding to uniform distribution in the middle block. Whereas, when a histogram representation is used for those locations, we are able to see the underlying high uncertainty in those locations (Figure 5(c) and (d)). But using the multivariate histogram for all the location is computationally expensive. By selectively choosing histograms only for regions where a Gaussian test fails, we are able to significantly reduce the modeling effort in our proposed method and still generate results similar to the nonparametric version. Our proposed method took 3.4 minutes (including 12 seconds for normality test), whereas, using multivariate histogram took 6.5 minutes. Though the multivariate Gaussian model took 1.10 minutes, it was not able to reflect the true underlying uncertainty in regions where uniform distributions were used to sample. Moreover, the overall memory requirement for the multivariate Gaussian model, proposed copula-based model and multivariate histogram was 0.77MB, 0.89MB and 314.7MB respectively. Besides, the results generated by our method is more reliable compared to the other two approaches because we performed a statistical verification (test) of model at each grid before deciding on a model to pick from. A goodness-of-fit test like the KS test discussed in Section 5 can be used to quantify how good are the selected models at each location. Figure 6 shows the results of KS-test performed at each grid location for the three cases. Lower the value of the KS test, the better the model represents the data. As can be seen in Figure 6(a) the KS test values of our mixed distribution field is lower for all the locations compared to the KS test values for using only Gaussian models (Figure 6(b)) and Histograms (Figure 6(c)). Besides, the visual validation of the results we quantified the difference in the three LCP results by computing their Root-Mean Square Deviation (RMSD). The LCP field of the multivariate Gaussian has a 12.7% and 13.4% deviation from the corresponding results of mixed distribution field and multivariate histogram respectively. Whereas, our proposed method produces 1.7% deviation from the multivariate histogram result and still takes 48% less time.

Global Ensemble Weather Forecast: We applied our copula-based technique on a 144×73 resolution real world weather forecast ensemble dataset with 21 members, generated by the Global Ensemble Forecast System (GEFS) [16] to compute the probabilistic isocontours of global temperature values. In this example, we used KDE to model the uncertainty at grid locations where normality test fails to show sufficient confidence. Figure 7(a) shows the result of our copula-based method on the mixed distribution field (Gaussian and KDE), Figure 7(b) shows the result of using multivariate Gaussian models and Figure 7(c) shows the result of using nonparametric KDE to model all the grid locations. The result of the Gaussian model is more smoothed out and can be misleading in many regions where a Gaussian distribution is a bad fit as shown in the highlighted regions in the results. The KDE model is able to highlight those uncertain regions which the Gaussian model fails but at the cost of high computational time and memory usage. Our flexible modeling strategy allows us to use KDE only where Gaussian fails, as a result, we are able to generate similar results as multivariate KDE models but with much less modeling effort. We have marked some of the regions where there are differences in probability values between the three techniques with dotted rectangles and circles in red. The zoomed-in views (Figure 7(d),(e),(f)) of the selected rectangular region (southeastern coast of Africa) clearly show the variation of probability values across the three approaches. This is validated by the p-Value results in Figure 7(g). The root mean square deviation (RMSD) of the LCP field of the multivariate Gaussian model is 5.9% and 4.1% of the results generated by the KDE model and the mixed model respectively. While the result of the mixed model has only 0.6% deviation from the KDE model. The computation time of the multivariate Gaussian, KDE and mixed models are 3.13, 19.06 and 7.54 minutes respectively. Besides the computational time, the overall memory usage for the multivariate Gaussian, KDE and mixed models are 0.331MB, 1.535MB and 0.414MB respectively.

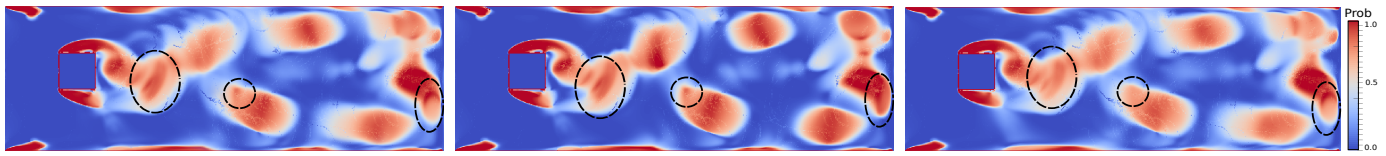
Square Cylinder Vortex Ensemble: We tested our coupla-based technique on vector distribution fields to detect vortex core probabilities. The dataset represents the flow field behavior around a square shaped cylinder in a 600×200 resolution 2D grid. A set of 10 simulations were generated with slightly different parameters to model the uncertainty in flow structures. Figure 8 shows the results of vortex probabilities (λ_2-criterion values) as computed by the method outlined in Section 6.2. Figure 8(a) shows the result generated by our copula-based method where we used either a Gaussian or a KDE to model uncertainty of the vector components at each grid location. Figure 8(b) and (c) show the results generated by assuming only Gaussian and only KDE re-



(a) (b) (c)

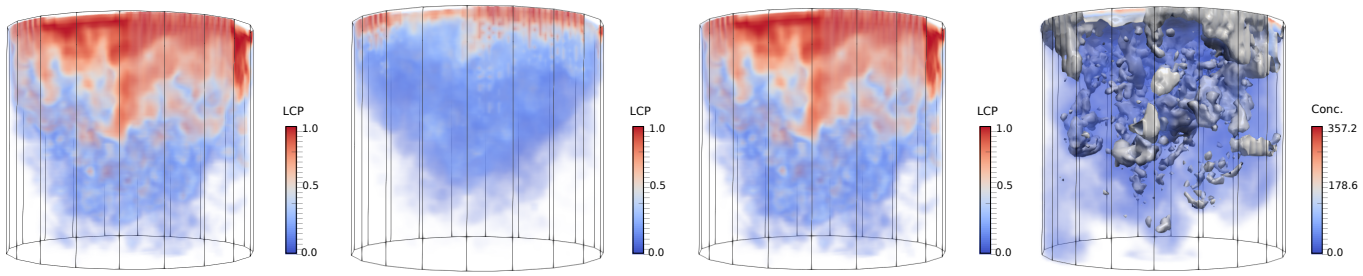
(d) (e) (f) (g)

Fig. 7: Global Ensemble Weather Forecast: The level-crossing probability for isovalues 280K. (a) The result of using copula-based method on the distribution field with Gaussian and KDE models. (b) The result of using only Gaussian models. (c) The result of using only KDEs. (d) zoomed in view of the selected region in figure (a). (e) zoomed in view of the selected region in figure (b). (f) zoomed in view of the selected region in figure (c). (g) The result of Shapiro-Wilk normality test, a low p-Value indicates the underlying data is less likely to follow normal distribution



(a) (b) (c)

Fig. 8: Square Cylinder Vector Ensemble: The results of vortex core probability using (a) copula-based method on mixed distribution field of Gaussian and KDE models, (b) only Gaussian models and (c) only KDE models. The marked regions highlights the difference in vortex structures detected by the three modeling strategies.



(a) (b) (c) (d)

Fig. 9: Salt Concentration Ensemble: Results of uncertain isocontours representing the viscous fingers for salt concentration level of 50 and generated by (a) copula-based technique on a mixed distribution field, (b) assuming Gaussian model at each grid location, (c) assuming trimodal GMM at each grid location. (d) shows the shape of an isosurface from a randomly chosen ensemble member.

spectively across all the locations. The probability values are different when using only Gaussian as compared to using only KDE. The region marked by the dotted black circles in Figure 8(b) and (c) highlights some of those differences. Result generated by our proposed method as shown by Figure 8(a) is a mixed representation of both the type of distributions, therefore, the regions with high certainty of following a Gaussian distribution are able to show the probable vortex structures that Figure 8(b) reflects. Whereas, regions where we used KDE to model the data where able to show results similar to Figure 8(c). For example, the region marked by the right-most black circle highlights a feature which was missed out by assuming Gaussian distribution, but was captured by both KDE and our mixed representation. The computation time of using multivariate Gaussian distribution was 10.4 minutes , while, the time for using only KDE was 52 minutes. On the other hand, our copula-based method took 28.5 minutes (including the time to perform normality test for both the vector components at each location), but generated results which are less prone to be erroneous compared to the other two approaches because of the prior statistical

test for model selection. The RMSD of the result generated by the Gaussian model is 6.5% of the result generated by the mixed model and 8.2% of the KDE model. On the other hand, result of the mixed model has a deviation of 2% compared to the KDE model results. While, the overall memory usage for the multivariate Gaussian, mixed model and multivariate KDE are 23.23MB, 27.56MB and 453.3MB respectively.

Salt Concentration Ensemble Our final dataset is a 3D dataset ($64 \times 64 \times 64$) from the field of fluid dynamics [19]. It represents the density field of concentration of salt particles dissolving in a fluid contained in a cylindrical container. Various boundary conditions were used to study an interesting fluid property called viscous fingers, the shape of which is represented by the isosurfaces of salt concentration values. The data comprises of 50 ensemble members and we selected a single time-step to perform our study. We computed the LCP for isosurface of concentration value 50. Figure 9 shows the results of the uncertain viscous fingers (isosurfaces). To model uncertainty at each grid location, we decided to use only parametric models because a nonparametric model like KDE will be computationally very expensive both in terms of time and storage because the number of ensemble members are relatively high for this dataset(50). Instead, we decided to use Gaussian Mixture Models (GMM) of different modes to model the uncertainty. Using the Bayesian Information Criterion (BIC) test explained in Section 5, we decide on a optimal number of modes for GMMs (out of 1, 2 and 3) at each grid location. Figure 9(a) shows the LCP results of using a copula-based technique on a mixed distribution field of unimodal, bimodal and trimodal GMMs. We compared this against distribution fields where we used only Gaussian distribution (equivalent to unimodal GMMs) and only multivariate GMMs with 3 modes, the results of which are shown in Figure 9(b) and (c) respectively. One interesting property of the isosurfaces in this dataset is that there are many disjoint components across the space as shown in Figure 9(d) (it shows the isosurface for isovalue 50 in a randomly chosen ensemble member). Because of this structure, Gaussian model assumption produces an overly smoothed-out result as shown in Figure 9(b) and no clear finger structures are visible. Whereas, the results produced by using a multivariate GMM with 3 modes is able to reveal those finger-like structures as shown in Figure 9(c). However, the overall estimation of multivariate GMM with 3 modes for each voxel in the dataset was very expensive and took 1 hour 10 minutes to generate the LCP field. On the other hand, our proposed copula-based method on a mixed distribution field took only 35.6 minutes (including the 1.5 minutes for the BIC test) and produced similar results as the multivariate GMM approach (RMSD value of 1.5%). Using only Gaussian models it took just 7.2 minutes but the result was not trustworthy (RMSD value was 22.3 % of the GMM result). The overall memory usage for multivariate Gaussian model, proposed mixed GMM model and multivariate GMM with 3 modes are 30.09MB, 34.07MB and 291.05MB respectively.

8 Discussion

Performance Study: Table 1 summarizes the overall performance (time and memory) and quantitative comparison values (RMSD) for the various example case studies. One important thing to note here is that the performance gain is attributed to the fact that our proposed uncertainty modeling technique allows us to model the univariate distribution at each grid location and the multivariate dependency separately. Because of this flexibility we can use the computational expensive models only when it is really necessary, thus, bringing down the cost without sacrificing on quality. Based on the modeling requirements and the degree of correctness required in the task, the performance can vary and is strictly controlled by the complexity of the univariate models involved. Even the quality of the result is limited only by the quality that the individual models offer. In this paper, we compared our results with the other Monte-Carlo based uncertain feature extraction methods [39, 41] because a copula-based method is inherently a Monte-Carlo process. Closed-form solution in copula-based models is not so straight-forward to derive for mixed univariate marginals [44], therefore, we did not compare our approach with the work proposed by Athawale et al. [2].

The effectiveness of the result generated by the proposed method is highly dependent on the effectiveness of the statistical test performed to decide the distribution type. As was mentioned in Section 5, one important factor which determines the effectiveness of a statistical test is the sample size. Small sample sizes are susceptible to generate unreliable test results. A general rule of thumb in statistic is to fall back upon a nonparametric model when the parametric tests do not give reliable results for smaller sample sizes. Therefore, one must be careful and critical while performing such statistical test, especially on small sample sizes, which is common for many real world ensemble experiments.

Table 1: Performance Summary of the Example Case Studies

	Complexity			RMSD
	Model	Time (mins)	Memory (MB)	
Synthetic Data	Mixed	3.4	0.89	1.7% (Mixed vs Hist)
	Hist	6.5	314.7	12.7% (Mixed vs Gaussian)
	Gaussian	1.1	0.77	13.4% (Gaussian vs Hist)
GEFS	Mixed	7.54	0.41	0.6% (Mixed vs KDE)
	KDE	19.06	1.53	4.1% (Mixed vs Gaussian)
	Gaussian	3.13	0.33	5.9% (Gaussian vs KDE)
Sq. Cylinder	Mixed	28.5	27.56	2% (Mixed vs KDE)
	KDE	52	453.3	6.5% (Mixed vs Gaussian)
	Gaussian	10.4	23.23	8.2% (Gaussian vs KDE)
Salt Conc.	Mixed	35.6	34.07	1.5% (Mixed vs GMM_3)
	GMM_3	70	291.05	21.6 % (Mixed vs Gaussian)
	Gaussian	7.2	30.09	22.3 % (Gaussian vs GMM_3)

Nonlinear Correlation: Gaussian copula, used in this work is a form of elliptic copula, which models the multivariate dependency based on the correlation values. Correlation is good only to capture the linear relationships. Therefore, Gaussian copula alone, cannot capture any nonlinear correlation with good accuracy. Tewari et al. [52] have recently proposed Gaussian Mixture Copula Models (GMCM) to capture nonlinear relationships. However, for the purpose of our work, which is to model the dependency among the grid locations (the spatial neighbors) we have found that there is no significant improvement of the results using GMCM over Gaussian copula.

Scope of Application: In this paper we have used only the popular distribution models, but our proposed method facilitates incorporating other distribution models into the analysis framework as well. Also, the scope of our proposed copula-based modeling strategy is not limited to only extracting probabilistic features. Since, distributions are not only used for modeling uncertainty but also to perform data reduction in large-scale simulations [8, 10, 53], our flexible strategy can allow users to perform adaptive data reduction by using complex models for sensitive regions and simpler models for the other regions. Many other possible applications can benefit from our flexible modeling strategy.

9 Conclusion and Future Work

In this paper, we have proposed a flexible distribution based uncertainty modeling strategy based on a statistically sound multivariate technique called *Copula*. Using Gaussian copula, we model multivariate distributions where the marginal distributions are represented by distinct class/family of distributions. Our proposed technique is specifically tailored to meet the needs of modeling uncertainty in scientific datasets. Using this flexible strategy we have proposed ways to extract uncertain/probabilistic features both in scalar and vector fields. We have also offered guidelines to select the univariate distributions to model the uncertainty at each grid location. In future, we would like to explore more of such scenarios where modeling flexibility is required, which standard models fail to deliver. As stated in the previous section, data reduction is one such interesting field to explore. Also, multivariate visualization and analysis techniques can greatly benefit from the dependency modeling framework that copula offers. This is something that we are interested to look into in the near future.

Acknowledgments

This work was supported in part by NSF grants IIS- 1250752, IIS-1065025, and US Department of Energy grants DE- SC0007444, DE-DC0012495, program manager Lucy Nowell.

REFERENCES

[1] T. Athawale and A. Entezari. Uncertainty quantification in linear interpolation for isosurface extraction. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2723–2732, Dec 2013.

[2] T. Athawale, E. Sakhaee, and A. Entezari. Isosurface visualization of data with nonparametric models for uncertainty. *IEEE Trans. Vis. Comput. Graph.*, 22(1):777–786, 2016.

[3] M. Baudin, A. Dutfoy, B. Iooss, and A.-L. Popelin. *OpenTURNS: An Industrial Software for Uncertainty Quantification in Simulation*, pages 1–38. Springer International Publishing, Cham, 2016.

[4] K. Bensema, L. Gosink, H. Obermaier, and K. Joy. Modality-driven classification and visualization of ensemble variance. *IEEE transactions on visualization and computer graphics*, (5), 2015-12-10 00:00:00.0.

[5] K. Brodlie, R. Allendes Osorio, and A. Lopes. *A Review of Uncertainty in Data Visualization*, pages 81–109. Springer London, London, 2012.

[6] U. Cherubini and E. Luciano. Bivariate option pricing with copulas. *Applied Mathematical Finance*, 9:69–85, 2002.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.

[8] S. Dutta, C. M. Chen, G. Heinlein, H. W. Shen, and J. P. Chen. In situ distribution guided analysis and visualization of transonic jet engine simulations. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):811–820, Jan 2017.

[9] S. Dutta and H. Shen. Distribution driven extraction and tracking of features for time-varying data analysis. *IEEE Trans. Vis. Comput. Graph.*, 22(1):837–846, 2016.

[10] S. Dutta, J. Woodring, H.-W. Shen, J.-P. Chen, and J. Ahrens. Homogeneity guided probabilistic data summaries for analysis and visualization of large-scale data sets. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*, pages 1–1, 2017 (accepted).

[11] G. Elidan. *Copulas in Machine Learning*, pages 39–60. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[12] P. D. Ellis. *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press, Cambridge ; New York, 2010.

[13] P. Embrechts, F. Lindskog, and A. McNeil. Modelling dependence with copulas and applications to risk management. *Handbook of heavy tailed distributions in finance*, 8(1):329–384, 2003.

[14] D. F. Findley. Counterexamples to parsimony and bic. *Annals of the Institute of Statistical Mathematics*, 43(3):505–514, 1991.

[15] R. Fujimaki, Y. Sogawa, and S. Morinaga. Online heterogeneous mixture modeling with marginal and copula selection. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 645–653, New York, NY, USA, 2011. ACM.

[16] GEFS. Global ensemble forecast system. `https://www.ncdc.noaa.gov/`.

[17] M. G. Genton, C. Johnson, K. Potter, G. Stenchikov, and Y. Sun. Surface boxplots. *Stat*, 3(1):1–11, 2014.

[18] A. Ghasemi and S. Zahediasl. Normality tests for statistical analysis: A guide for non-statisticians. *Int J Endocrinol Metab*, 10(2):486–489, 2012.

[19] IEEE. Scivis contest 2016. `http://www.uni-kl.de/sciviscontest/`.

[20] C. Johnson. Top scientific visualization research problems. *IEEE Computer Graphics and Applications*, 24(4):13–17, July 2004.

[21] C. R. Johnson and A. R. Sanderson. A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications*, 23(5):6–10, Sept 2003.

[22] S. Kirshner and B. Póczos. Ica and isa using schweizer-wolff measure of dependence. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 464–471, New York, NY, USA, 2008. ACM.

[23] T. Liebmann and G. Scheuermann. Critical points of gaussian-distributed scalar fields on simplicial grids. *Computer Graphics Forum*, 35(3):361–370, 2016.

[24] S. Liu, J. A. Levine, P. T. Bremer, and V. Pascucci. Gaussian mixture model based volume visualization. In *IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, pages 73–77, Oct 2012.

[25] A. L. Love, A. Pang, and D. L. Kao. Visualizing spatial multivalue data. *IEEE Computer Graphics and Applications*, 25(3):69–79, May 2005.

[26] A. Luo, D. Kao, and A. Pang. Visualizing spatial distribution data sets. In *Proceedings of the Symposium on Data Visualisation 2003*, VISSYM '03, pages 29–38, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.

[27] J. Ma and Z. Sun. Copula component analysis. *CoRR*, abs/cs/0703095, 2007.

[28] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management: concepts, techniques and tools*. Princeton series in finance. Princeton University Press, Princeton (N.J.), 2005.

[29] M. Mihai and R. Westermann. Visualizing the stability of critical points in uncertain scalar fields. *Computer Graphics Forum*, 41(0):13–25, 2014.

[30] R. B. Nelsen, J. J. Quesada-Molina, J. A. Rodriguez-Lallena, and M. Úbeda-Flores. On the construction of copulas and quasi-copulas with given diagonal sections. *Insurance: Mathematics and Economics*, 42:473–483, 2008.

[31] R. A. Osorio and K. Brodlie. Contouring with uncertainty. In *6th Theory and Practice of Computer Graphics Conference*, pages 59–66, 2008.

[32] M. Otto, T. Germer, H.-C. Hege, and H. Theisel. Uncertain 2d vector field topology. *Computer Graphics Forum*, 29(2):347–356, 2010.

[33] M. Otto and H. Theisel. Vortex analysis in uncertain vector fields. In *Computer Graphics Forum*, volume 31, pages 1035–1044. Blackwell Publishing Ltd, 2012.

[34] A. Pang, C. Wittenbrink, and S. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, Nov 1997.

[35] C. Petz, K. Pthkow, and H.-C. Hege. Probabilistic local features in uncertain vector fields with spatial correlation. *Computer Graphics Forum*, 31(3pt2):1045–1054, 2012.

[36] T. Pfaffelmoser, M. Reitinger, and R. Westermann. Visualizing the positional and geometrical variability of isosurfaces in uncertain scalar fields. In *Computer Graphics Forum*, volume 30, pages 951–960. Wiley Online Library, 2011.

[37] T. Pfaffelmoser and R. Westermann. Visualization of Global Correlation Structures in Uncertain 2D Scalar Fields. *Computer Graphics Forum*, 2012.

[38] K. Pöthkow and H. C. Hege. Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1393–1406, Oct 2011.

[39] K. Pöthkow and H.-C. Hege. Nonparametric models for uncertainty visualization. *Computer Graphics Forum*, 32(3pt2):131–140, 2013.

[40] K. Pöthkow, C. Petz, and H.-C. Hege. Approximate level-crossing probabilities for interactive visualization of uncertain isocontours. *International Journal for Uncertainty Quantification*, 3(2), 2013.

[41] K. Pöthkow, B. Weber, and H.-C. Hege. Probabilistic marching cubes. In *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis'11, pages 931–940, Chichester, UK, 2011. The Eurographs Association ; John Wiley ; Sons, Ltd.

[42] K. Potter, P. Rosen, and C. R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *Uncertainty Quantification in Scientific Computing*, pages 226–249. Springer, 2012.

[43] N. M. Razali, Y. B. Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.

[44] M. Rey. Copula models in machine learning. 2015.

[45] M. Rey and V. Roth. Copula Mixture Model for Dependency-seeking Clustering. *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 927–934, 2012.

[46] M. Rey and V. Roth. Meta-gaussian information bottleneck. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1925–1933, 2012.

[47] S. Schlegel, N. Korn, and G. Scheuermann. On the interpolation of data with normally distributed uncertainty for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2305–2314, Dec 2012.

[48] T. Schmidt. Coping with Copulas. *Copulas - From Theory to Applications in Finance*, (15):1–23, 2006.

[49] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

[50] A. Sklar. *Fonctions de repartition a n dimensions et leurs marges*. 1959.

[51] M. A. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, 1974.

[52] A. Tewari, M. J. Giering, and A. Raghunathan. Parametric characterization of multimodal distributions with non-gaussian modes. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 286–292,

Dec 2011.

[53] D. Thompson, J. A. Levine, J. C. Bennett, P. T. Bremer, A. Gyulassy, V. Pascucci, and P. P. Pébay. Analysis of large-scale scalar data using hixels. In *1st IEEE Symposium on Large-Scale Data Analysis and Visualization 2011, LDAV 2011 - Proceedings*, pages 23–30, 2011.

[54] R. T. Whitaker, M. Mirzargar, and R. M. Kirby. Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2713–2722, 2013.

[55] P. C. Wong, H. W. Shen, C. R. Johnson, C. Chen, and R. B. Ross. The top 10 challenges in extreme-scale visual analytics. *IEEE Computer Graphics and Applications*, 32(4):63–67, July 2012.