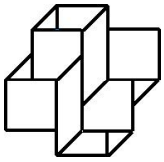

Laboratório Nacional de Computação Científica

Programa de Pós-Graduação em Modelagem Computacional

Generalized Lambda Distribution for Uncertainty Quantification of Large-scale Spatio-temporal Models

Noel Moreno Lemus

Petrópolis, RJ - Brasil

Abril de 2018

Noel Moreno Lemus

Generalized Lambda Distribution for Uncertainty Quantification of Large-scale Spatio-temporal Models

Thesis submitted to the examining committee in partial fulfillment of the requirements for the degree of Doctor of Sciences in Computational Modeling.

Laboratório Nacional de Computação Científica

Programa de Pós-Graduação em Modelagem Computacional

Supervisor: Fábio André Machado Porto

Petrópolis, RJ - Brasil

Abril de 2018

XXXX

Moreno Lemus, Noel

Generalized Lambda Distribution for Uncertainty Quantification of Large-scale Spatio-temporal Models / Noel Moreno Lemus. – Petrópolis, RJ - Brasil, Abril de 2018-

109 p. : il. ; 30 cm.

Orientador(es): Fábio André Machado Porto e

Thesis (D.Sc.) – Laboratório Nacional de Computação Científica
Programa de Pós-Graduação em Modelagem Computacional, Abril de 2018.

1. Uncertainty Quantification. 2. Big Data. 3. Information Entropy. I. Machado Porto, Fábio André. II. LNCC/MCTI. III. Title

CDD: XXX.XXX

Noel Moreno Lemus

Generalized Lambda Distribution for Uncertainty Quantification of Large-scale Spatio-temporal Models

Thesis submitted to the examining committee in partial fulfillment of the requirements for the degree of Doctor of Sciences in Computational Modeling.

Approved by:

Prof. Fábio André Machado Porto, D.Sc.
(Presidente)

Prof. Fernando Alves Rochinha, D.Sc.

Prof. Hugo de La Cruz, Ph.D.

Prof. Artur Ziviani, Ph.D.

Petrópolis, RJ - Brasil
Abril de 2018

Dedication

To my little and special family.

Acknowledgements

O autor manifesta reconhecimentos às pessoas e instituições que colaboraram para a execução de seu trabalho.

"Essentially, all models are wrong, but some are useful."
(George Edward Pelham)

Abstract

Segundo a **??**, 3.1-3.2), o resumo deve ressaltar o objetivo, o método, os resultados e as conclusões do documento. A ordem e a extensão destes itens dependem do tipo de resumo (informativo ou indicativo) e do tratamento que cada item recebe no documento original. O resumo deve ser precedido da referência do documento, com exceção do resumo inserido no próprio documento. (. . .) As palavras-chave devem figurar logo abaixo do resumo, antecedidas da expressão Palavras-chave:, separadas entre si por ponto e finalizadas também por ponto.

Keywords: latex. abntex. editoração de texto.

Abstract

Large-scale spatio-temporal simulations with quantified uncertainty enable scientists/decision-makers to precisely assess the degree of confidence of their simulation-based predictions. This uncertainty could be qu antified or characterized in different ways, from the use of low order statistical moments (the most commonly used), to the evaluation of a complete PDF (a most complete approach). The latter provides a more comprehensive description of the uncertainty leading to aware decisions. However, fitting PDFs to the data is computational intesive. Moreover, due to heterogeneity the uncertainty computed in regions of the dataset is hampered by the existence of different PDF types.

In this thesis, we propose a new method to quantify the uncertainty in large-scale spatio-temporal models based on the Generalized Lambda Distribution (GLD). GLD is a family of PDFs that nicely models the heterogeneity of uncertainty as discussed above. It is specified by 4 parameters that simplifies PDFs comparisons easing analytical processing, such as clustering. We show how the dataset modeled through GLDs can be used to answer queries, such as: (*i*) how to group the output of the UQ process based on the simillarity of the uncertainty?, (*ii*) what is the uncertainty in some spatio-temporal locations not previously analysed?, (*iii*) what is the uncertainty of an specific spatio-temporal region?, (*iv*) how to compare two regions as a function of its uncertainty?, and (*v*) what is the less uncertain model from a set of models? The proposed method has been tested in realistic use cases from various scientific areas. Additionally, an R package has been implemented with all the functionalities discussed in the thesis.

Keywords: Uncertainty Quantification, Large-scale spatio-temporal models, Big Data, Generalized Lambda Distribution

List of Figures

List of Tables

List of abbreviations and acronyms

UQ	Uncertainty Quantification
FP	Forward Problem
QoI	Quantity of Interest
GLD	Generalized Lambda Distribution
p.f.	percentile function
r.v.s	random variables
GLDEX	r package to compute the GLD
M&S	Modeling and Simulation
LSSTM	Large-scale spatio-temporal model
KDE	Kernel Density Estimation

List of symbols

σ	random variable
Σ	random space
ζ	Letra grega minúscula zeta
$\in$	Pertence

Contents

1 Introduction

"A measurement result is complete only when accompanied by a quantitative statement of its uncertainty. The uncertainty is required in order to decide if the result is adequate for its intended purpose and to ascertain if it is consistent with other similar results"

The rapid growth of high-performance computing and the advances in numerical techniques in the last two decades have provided an unprecedented opportunity to explore complex physical phenomena using large-scale spatio-temporal modeling and simulation. At the same time, scientific community is leaving behind the traditional deterministic approach, which offers point predictions with no associated uncertainty (JOHNSTONE et al., 2016); to include Uncertainty Quantification (*UQ*) as a common practice in their researches.

Large-scale spatio-temporal simulations with quantified uncertainty enable scientists to make precise statements about the degree of confidence they have in their simulation-based predictions. These approaches find practical applicability in models for predicting the behavior of weather, hurricane forecasts (TOBERGTE; CURTIS, 2013), subsurface hydrology (BARONI; TARANTOLA, 2014), geology (GUERRA et al., 2016), nuclear reactor design, financial portfolios (CHEN; FLOOD; SOWERS, 2008), and biological phenomena, just to name a few. They also allow to study physical phenomena that are impossible to assess experimentally, for example: simulate nuclear accidents, or the conditions that some spatial vehicle will find at landing in Mars, and so on. The success of these techniques has made them increasingly important tools for high impact predictions and decision making.

UQ includes different aspects that warranty the predictive fidelity of a numerical simulation, such as the uncertainty in the experimental data, which is used for defining the parameter values of a model; the propagation of uncertain parameters through the model; and the choice of the model itself. *UQ* is a complex process that covers the following main tasks: (i) uncertainty characterization (CRESPO; KENNY; GIESY, 2014), also called model calibration (FARRELL, 2015) or statistical inverse problem (ESTACIO-HIROMS; PRUDENCIO, 2012); (ii) sensitivity analysis; (iii) forward problem or uncertainty propagation; and (iv) model selection.

This paper is focused on *forward propagation*, whose objective is to quantify the uncertainties in model output(s) propagated from uncertain inputs. The targets of *forward propagation* analysis can be: (i) evaluate low-order moments (i.e. mean and variance) of

the outputs, (ii) evaluate the reliability of the outputs, and/or (iii) assess the complete probability distribution (*PDF*) of the outputs.

When dealing with large-scale spatio-temporal models, a huge among of data is generated as a result of the simulation process. Indeed, on each spatio-temporal location $(s_i, t_j) \in \mathcal{S} \times \mathcal{T} \subseteq \mathbb{R}^3 \times \mathbb{R}$, usually more than 10^4 simulations are performed. Then, the size of the output dataset is in the order of $N_s \times N_t \times N_{sim}$, where: N_s is the number of spatial locations, N_t is the number of time steps, and N_{sim} is the number of simulations. An example of the volume of data generated by these simulations is given in the experimental section **??** of this paper, where the output dataset is about 2.4 TB. This turn *forward propagation* in a data intensive problem.

Another important aspect, which is often not taken into account, is that the uncertainty need to be quantified in some way that can be used after, to answer questions that arise in the *UQ* context. In that sense, assess the complete *PDF* could be the best way to quantify uncertainty, because if you can find the *PDF* that best fit the dataset with reasonably accurately, you can get all the statistical properties under one roof. At the same time, we can substitute the original data by the *PDFs*, which represents a huge reduction in the volume of data to manipulate.

Contradictorily, statistical moments (e.g. mean and standard deviation) are possibly the most used ways to quantify the uncertainty, despite the fact that they doesn't have information about the manner in which the data are distributed (LAMPASI; Di Nicola; PODESTA, 2006). This is because of the difficulty to find the *PDF* that best fit a dataset (KARIAN; DUDEWICZ, 2011), even more, when dealing with large-scale spatio-temporal models where the *PDF* needs to be derived on each spatio-temporal location, and therefore the *forward propagation* problem becomes time consuming and computationally intensive too.

However, the use of low order moments alone prevents us from making accurate analysis with respect to the uncertainty. They are not enough neither for the characterization nor for the quantification of the uncertainty, and questions such as:

- What is the uncertainty in the spatio-temporal region $\mathcal{S}_i \times \mathcal{T}_j$ associated to the *QoI* q_k and a computational model $\mathcal{M}_m$?
- How to compare different spatio-temporal regions $\mathcal{S}_i \times \mathcal{T}_j$ with respect to the uncertainty?
- What is the less uncertain model from the set of models $\mathcal{M} = \mathcal{M}_1, \mathcal{M}_2, ...\mathcal{M}_m$, to predict the value of a QoI q_k, over a spatio-temporal region $\mathcal{S}_i \times \mathcal{T}_j$?

can be poorly answered. So, we emphasize that only the characterization of the uncertainty by using the *PDF* allows aware decisions.

A first effort to try to estimate the *PDFs* on large-scale spatio-temporal simulations was done by (LIU; NIU; LIAO, 2018) Ji et. al. in ***Parallel Computation of PDFs on Big Spatial Data Using Spark***. They propose a new solution to efficiently compute the *PDFs* in parallel using Spark, through three methods: data grouping, machine learning prediction and sampling. The main drawback of the proposed approach is that you should try many different distributions, to find the PDF that best fits the dataset on each specific spatio-temporal location. Another drawback is that, as we mentioned above, the uncertainty needs to be quantified in the way that facilitates its further use; and the heterogeneity of the functions used in the approach doesn't facilitate it.

To face these challenges, in this paper we propose a general framework to quantify the uncertainty in large-scale spatio-temporal models. It uses a data-driven approach and combines the generalized lambda distribution (*GLD*), clusters algorithms and information entropy, for helping researchers to answer the above questions and many others that arise in *UQ* context. Our proposal provides a generally applicable and easy-to-use tool that supports the representation and analysis of uncertainty, as was suggested in the "*Workshop on Quantification, Communication, and Interpretation of Uncertainty in Simulation and Data Science*" (TOBERGTE; CURTIS, 2013).

In order to illustrate the use of the proposed framework, a case study is discussed. The main results obtained are: (i) the *GLD* good fits for more than the 80 % of the dataset, (ii) the use of the *GLD* allows to include clustering algorithms to group the spatio-temporal locations with similar uncertainty, (iii) the centroids of the clusters can be used as a faithful representation of the rest of the spatio-temporal locations, which significantly reduces the data corresponding to the simulation outputs, (iv) with the use of these centroids we can characterize the uncertainty in any spatio-temporal region as a mixture of *GLDs*.

The rest of the paper is organized as follows: Section **??** gives the theoretical foundations of *UQ* and highlights some interesting aspects included in our proposal. Section **??** describes the principal characteristics of the *GLD* that make it suitable for this proposal. Section **??** presents the proposed approach, the workflow we implement and some considerations of the implementation. Section **??** presents a use case and discusses the results. This use case allows us to explain our approach in the context of a real problem, which facilitates its understanding. Section **??** covers the related works and finally, section **??** concludes the paper and proposes some future works.

1.1 Research Objectives

The main objective of this thesis is a new method to quantify the uncertainty in large-scale spatio-temporal models based on the Generalized Lambda Distribution (GLD).

To achieve that goal the following research questions need to be answered:

RQ1. how to group the output of the UQ process based on the similarity of the uncertainty?

RQ2. what is the uncertainty in some spatio-temporal locations not previously analyzed?

RQ3. what is the uncertainty of an specific spatio-temporal region?

RQ4. how to compare two regions as a function of its uncertainty?

RQ5. what is the less uncertain model from a set of models?

RQ1 and **RQ2** are answered in chapter 4, while **RQ3**, **RQ4** and **RQ5** are answered in chapter 5. In chapter 6 all the questions are answered again for all the use cases.

1.2 Highlights of the Dissertation

1.3 Organization of the Dissertation

The structure of the remainder of this thesis is outlined for reference.

Chapter 2 background of UQ.

Chapter 3 Ji paper.

Chapter 4 GLD.

Chapter 5 GLD clustering and kriging.

Chapter 6 Workflow.

Chapter 7 Use cases.

Chapter 8 Conclusions and future works.

2 Uncertainty Quantification Background

"UQ cannot tell you that your model is 'right' or 'true', but only that, if you accept the validity of the model (to some quantified degree), then you must logically accept the validity of certain conclusions (to some quantified degree)"
(SULLIVAN, 2015)

Uncertainty Quantification (UQ) is a topic of great importance and hence widespread interest in computational analyses that are used to support important societal decisions on issues related to climate change (ALLEN et al., 2000; PATT; KLEIN; VEGA-LEINERT, 2005), hurricane forecasts (TOBERGTE; CURTIS, 2013), subsurface hydrology (BARONI; TARANTOLA, 2014), geology (GUERRA et al., 2016), reactor safety [20-26], radioactive waste disposal [27-34], nuclear weapon safety [35-38], financial portfolios (CHEN; FLOOD; SOWERS, 2008), environmental degradation [44- 47], and many additional areas of concern and challenge.

In this Chapter, we review the UQ specialized literature, comment some interesting results and highlight the remainder challenges we are interested in to solving in the present thesis. The rest of the Chapter is organized as follow: in Section 2.1 we define what we understand as uncertainty, the differences between uncertainties and errors, some classifications of the uncertainties, and finally what is uncertainty quantification. In Section 2.2 we introduce the mathematical formalisms used to represent the uncertainty. Section 2.3 present the main problems that UQ covers. Next, in Section 2.4 the two principal *forward propagation* methods are referenced. Section 2.5 review the UQ challenges when we are in presence of large-scale spatio-temporal models; and finally Section 2.6 summarize the Chapter.

2.1 Definitions

2.1.1 Errors vs Uncertainties

The mismatch between the true physical phenomena and the prediction obtained by modeling and simulation (M&S) process can arise from the mathematical representation of a real problem, a physical problem (are the values of the parameters a good representation of the reality?), a computational problem (translation of a mathematical formulation into a numerical algorithm and a computational code) (MELOROSE et al., 2015). Uncertainty and error can be considered as the broad categories that are normally associated to

this mismatch. Until recently terms uncertainty and error have commonly been used interchangeably. It is believed, however, that failure to distinguish between these terms is detrimental to the quantification of credibility in M&S. According to (ALVIN et al., 1998) we can classify errors and uncertainties as follow:

- **errors:** recognizable deficiencies of the model or the algorithms employed. Errors are associated to: physical approximations to simplify the modeling of a physical process, translation of the mathematical to computational model, numerical approximations (truncation or roundoff), etc. When the errors are knowns there are reasonable means of estimating the magnitude of the error introduced.
- **uncertainties:** potential deficiency that is due to lack of knowledge. The different sources of uncertainty can be:
  - **parameter uncertainty**, which comes from the model parameters that are inputs to the computer model (mathematical model) but whose exact values are unknown to experimentalists and cannot be controlled in physical experiments, or whose values cannot be exactly inferred by statistical methods. Examples are the local free-fall acceleration in a falling object experiment, various material properties in a finite element analysis for engineering, and multiplier uncertainty in the context of macroeconomic policy optimization (KENNEDY; O'HAGAN, 2001).
  - **model inadequacy**, no model is perfect. Even if there is no parameter uncertainty, so that we know the true values of all the inputs required to make a particular prediction of the process being modeled, the predicted value will not equal the true value of the process. The discrepancy is model inadequacy. Since the real process may itself exhibit random variability, we define model inadequacy to be the difference between the true mean value of the real world process and the code output at the true values of the inputs.
  - **parametric variability**, which comes from the variability of input variables of the model. For example, the dimensions of a work piece in a process of manufacture may not be exactly as designed and instructed, which would cause variability in its performance.
  - **structural uncertainty**, aka model inadequacy, model bias, or model discrepancy, which comes from the lack of knowledge of the underlying true physics. It depends on how accurately a mathematical model describes the true system for a real-life situation, considering the fact that models are almost always only approximations to reality. One example is when modeling the process of a falling object using the free-fall model; the model itself is inaccurate since there

always exists air friction. In this case, even if there is no unknown parameter in the model, a discrepancy is still expected between the model and true physics.

- **algorithmic uncertainty**, aka numerical uncertainty, which comes from numerical errors and numerical approximations per implementation of the computer model. Most models are too complicated to solve exactly. For example, the finite element method or finite difference method may be used to approximate the solution of a partial differential equation, which, however, introduces numerical errors. Other examples are numerical integration and infinite sum truncation that are necessary approximations in numerical implementation.
- **experimental uncertainty**, aka observation error, which comes from the variability of experimental measurements. The experimental uncertainty is inevitable and can be noticed by repeating a measurement for many times using exactly the same settings for all inputs/variables.
- **interpolation uncertainty**, which comes from a lack of available data collected from computer model simulations and/or experimental measurements. For other input settings that don't have simulation data or experimental measurements, one must interpolate or extrapolate in order to predict the corresponding responses.

A more elegant definition of what uncertainty is, is enunciated in (HELTON, 2009) as:

Definition 2.1. Uncertainty is a best estimate of the range of a particular metric which may derive from one or two broad sources. Uncertainties that reflect a lack of knowledge about the appropriate value to use for a quantity that is assumed to have (missing modifier: a fixed?) value in the context of a particular analysis are termed ***epistemic***. Uncertainties that arise from an inherent randomness in the behavior of the system under study are termed ***aleatoric***.

2.1.2 Aleatoric vs Epistemic Uncertainty

It is sometimes assumed that uncertainty can be classified into those two categories, ***aleatoric*** and ***epistemic***, although the validity of this categorization is open to debate (KIUREGHIAN; DITLEVSEN, 2009).

Aleatoric uncertainty arises from an inherent randomness in the properties or behavior of the system under study. For example, the weather conditions at the time of a reactor accident are inherently random with respect to our ability to predict the future. Other examples include the variability in the properties of a population of weapon components and the variability in the possible future environmental conditions

that a weapon component could be exposed to. Alternative designations for **aleatory uncertainty** include variability, stochastic, irreducible and type A. (HELTON, 2009)

Epistemic uncertainty derives from a lack of knowledge about the appropriate value to use for a quantity that is assumed to have a fixed value in the context of a particular analysis. For example, the pressure at which a given reactor containment would fail for a specified set of pressurization conditions is fixed but not amenable to being unambiguously defined. Other examples include minimum voltage required for the operation of a system and the maximum temperature that a system can withstand before failing. Alternative designations for **epistemic uncertainty** include state of knowledge, subjective, reducible and type B. (HELTON, 2009)

While **epistemic uncertainty** can be reduced through experiments, improvement of the numerical methods and so on, **aleatory uncertainty** can not be reduced.

2.1.3 Uncertainty Quantification

UQ is not a mature field like linear algebra or single-variable complex analysis, with stately textbooks containing well-polished presentations of classical theorems. Both because of its youth as a field and its very close engagement with applications, UQ is much more about problems, methods and 'good enough for the job'. There are some very elegant approaches within UQ, but as yet no single, general, overarching theory (SULLIVAN, 2015).

UQ neither have a unique and globally accepted definition. In the reviewed literature we find some definitions that, from our point of view, are those that better describe what UQ is.

In the Wikipedia we find the following definition:

Definition 2.2. Uncertainty Quantification is the science of quantitative characterization and reduction of uncertainties in applications. It tries to determine how likely certain outcomes are if some aspects of the system are not exactly known.

This definition is very general and may be ignore some important aspects, by as a first approach is a good one.

In October of 2009 the U.S. Department of Energy organize a comission to study the impact of Extreme Scale computing in its National Security. One of the aspects analysed by the comision was UQ. In the report *"Scientific Grand Challenges in National Security: The Role of Computing at the Extreme Scale"* (U.S. Department of Energy, 2009), the authors define UQ as:

Definition 2.3. UQ studies all sources of error and uncertainty, including the following: systematic and stochastic measurement error; ignorance; limitations of theoretical models;

limitations of numerical representations of those models; limitations of the accuracy and reliability of computations, approximations, and algorithms; and human error. A more precise definition is UQ is the end-to-end study of the reliability of scientific inferences. (U.S. Department of Energy, 2009)

A more recent definition was introduced by Higdon et al. (HIGDON, 2017) in the *"Handbook of Uncertainty Quantification"*:

Definition 2.4. UQ is the rational process by which proximity between predictions and observations is characterized. It can be thought of as the task of determining appropriate uncertainties associated with model-based predictions. More broadly, it is a field that combines concepts from applied mathematics, engineering, computational science, and statistics, producing methodology, tools, and research to connect computational models to the actual physical systems they simulate. In this broader interpretation, UQ is relevant to a wide span of investigations. These range from seeking detailed quantitative predictions for a well-understood and accurately modeled engineering systems to exploratory investigations focused on understanding trade-offs in a new or even hypothetical physical system. (HIGDON, 2017)

Just to remark, in this definition the sentence: **a field that combines concepts from applied mathematics, engineering, computational science, and statistics, producing methodology, tools, and research to connect computational models to the actual physical systems they simulate**, illustrate the multidisciplinary nature of UQ and the main objectives of this research field.

2.2 Uncertainty Representation

An immediate challenge in the development of an appropriate treatment of uncertainty is the selection of a mathematical structure to be used in its representation (HELTON et al., 2010). Traditionally, probability theory has provided this structure [48-55]. However, in the last several decades, additional mathematical structures for the representation of uncertainty such as evidence theory [56-63], possibility theory [64- 70], fuzzy set theory [71-75], and interval analysis [76-81] have been introduced. This introduction has been accompanied by a lively discussion of the strengths and weaknesses of the various mathematical structures for the representation of uncertainty [82-90]. For perspective, several comparative discussions of these different approaches to the representation of uncertainty are available [72; 91-98].

This section briefly summarizes some of this approaches, and discuss in more details probability theory as this is the main one used in the rest of the thesis.

2.2.1 Interval Analysis

Interval analysis is the simplest way to represent the uncertainty used when nothing more can be said about some unknown quantity than a range of it possible values. All the values had the same probability. For example, in a case of some unknown variable x, such information may expressed as: $x \in [a, b]$, where a and b represent the left and right limit of the interval. This is a very basic form of uncertainty (SULLIVAN, 2015).

2.2.2 Variance

Suppose that, for a random variable $X \in \mathcal{X}$, the knowledge is summarized by a probability function $p \in P(\mathcal{X})$. The probability measure p is a very rich and high-dimensional object, but some times we need to summarize the uncertainty in p, with just one number. Variance is an obvious statistic to summarize this. The formal definition of Variance is:

Definition 2.5. Variance is the expectation of the squared deviation of a random variable from its mean. The variance measures how far each number in the set is from the mean.

If we know the mean m of p, then the Variance can be computed as:

$$\mathcal{V}(p) = \int_{\mathcal{X}} \|x - m\|^2 dp(x) = E_{X\ p}[\|X - m\|^2] \tag{2.1}$$

If $\mathcal{V}(p)$ is small, then we are **relatively certain** that the values of X are quite close to the mean m, and if $\mathcal{V}(p)$ is large, then we are more uncertain about X.

Variance and standard deviation $\sigma = \sqrt{\mathcal{V}(p)}$ are the standard way of quantify the uncertainty of a random variable. This is due to its simplicity of interpretation and easy to compute. But, there are some interesting questions that arise in UQ context that can't be answered correctly by the use of low order statistical moments, as Lampasi et al. (LAMPASI; Di Nicola; PODESTA, 2006) say, only assess the complete PDF allow aware decisions.

Is clear that, assess the complete PDF of each random variables on each spatio-temporal location, and worse yet, when we are in presence of a large-scale models, is a computational intensive task. But, this is exactly the main objective of this thesis, present a new workflow to compute the PDF on each spatio-temporal location, and demonstrate its practical use answering the research questions enunciated in Chapter 1.

2.2.3 Information Entropy

The concept of information entropy was first defined by Shannon (1948) in a study performed to identify the amount of information required to transmit English text.

The underlying idea was that, given the probabilities of letters occurring in the English alphabet, it is possible to derive a measure describing the missing information to determine the full text of a partially transmitted message, where information is understood as the information required to identify the message, not the information of the message itself. Based on several theoretical considerations, Shannon derived the following equation to classify a measure of the missing information, often referred to as information entropy:

$$H = -\sum_{i}^{N} p_i \log p_i \tag{2.2}$$

The information entropy H is defined as the sum of the product of the probability p for each possible outcome i of N, total possible outcomes, with its logarithm. The minimum value is 0, because $\log 1 = 0$.

2.2.3.1 Information entropy in a spatio-temporal context

For each spatio-temporal region, the information entropy can be described as:

$$H(s,t) = -\sum_{m=1}^{M} p_m(s,t) \log p_m(s,t) \tag{2.3}$$

where s denotes the location of the subregion, M represents the number of possible (exclusive) members the subregion may contain, and t is the physical time.

2.2.3.2 Information entropy as a meause of uncertainty

Based on 2.2.3 and 2.2.3.1, if the possible outcomes of the model and the probability of each outcome on each (s,t), are known, then the information entropy could be used as a qualitative measure of the uncertainty of the model output (WELLMANN; REGENAUER-LIEB, 2012). For example, in a spatio-temporal region (s,t) where the outcome is always the same, the information entropy is 0, because the outcome is known. On the other hand, in the worse case where all the outcomes have the same probability in (s,t), the entropy is maximum and the uncertainty too.

2.2.4 Probability Theory

Probability theory is based on the specification of a triple $(\Omega, \mathcal{F}, P)$, where Ω is the set of all possible outcomes, $\mathcal{F}$ is a suitably restricted collection of subsets of Ω, and P defines the probability of the elements of Ω.

The probability measure P is a function returning an event's probability, with the properties that $0 \leq P \leq 1$ and $P(\Omega) = 1$

One way to characterize the probability is through the probability density function (*PDF*). It is a mathematical function that, stated in simple terms, can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment.

Probability density function: for a continuous random variable X, we can define the probability that X is in $[a, b]$ as:

$$P(a <= X <= b) = \int_a^b f(x)dx \tag{2.4}$$

where $f(x)$ is a probability density function, which satisfies two properties:

$$f(x) >= 0$$
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

a, b are real numbers. The *PDF* defines the probability that $X <= a$ as $P(X <= a) = \int_{-\infty}^{a} f(x)dx$

2.3 Some Typical UQ Problems

Many typical UQ problems can be illustrated in the context of a system F, that maps input X in some space $\mathcal{X}$ to outputs $Y = \mathcal{M}(X)$ in some space $\mathcal{Y}$, through a mathematical/computational model $\mathcal{M}$. Some common UQ objectives include: *forward propagation or push-forward problem*, Section 2.3.1; *reliability or certification problem*, Section 2.3.2; *prediction problem*, Section 2.3.3; *inverse problem*, Section 2.3.4; *sensitivity analysis*, Section 2.3.5; and *model reduction or model calibration problem*, Section 2.3.6. More-less all of this objectives are summarized in Figure 1.

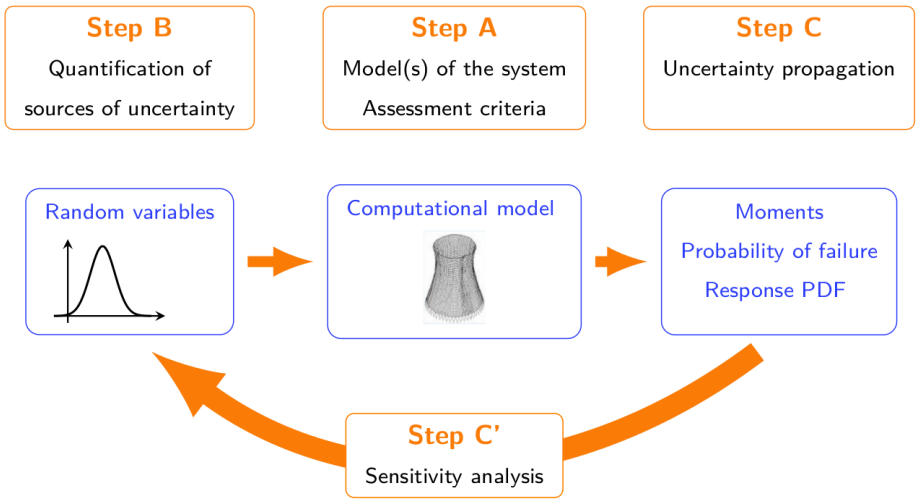

Figure 1 – Uncertainty Quantification workflow. Taken from UQLab.

2.3.1 Forward propagation or push-forward problem

Given the equation $\boldsymbol{Y} = \mathcal{M}(\boldsymbol{X})$ where:

- $\boldsymbol{X} \in \mathcal{X}$ is a vector of input parameters of the model,
- $\mathcal{M}$ is a computational model, and
- $\boldsymbol{Y} \in \mathcal{Y}$ is a vector that represents quantities of interest (QoI).

Suppose that the uncertainty about the inputs of $\mathcal{M}$ can be summarized in a probability distribution P on $\mathcal{X}$. Then in a *forward propagation*, the objective is to quantify the uncertainty of $\boldsymbol{Y}$, induced by $\boldsymbol{X}$ through $\mathcal{M}$.

The main objective of this thesis is a new method to quantify the uncertainty of the output of large-scale spatio-temporal models, that is a *forward propagation problem*. In Section 2.4 some methods for *forward propagation* are exposed; while in Section 2.5 we discuse about *forward propagation* in large-scale spatio-temporal models.

2.3.2 Reliability or certification problem

Suppose that some set $\mathcal{Y}_{fail} \subseteq \mathcal{Y}$ is identified as a 'failure set'. Then in a reliability analysis we are interesting in, given appropriate information about the input X and a process F, determine the failure probability

$$\mathcal{P}[\mathcal{M}(X) \in \mathcal{Y}_{fail}] \tag{2.5}$$

Furthermore, how large will the deviation from acceptable performance be, and what are the consequences? (SULLIVAN, 2015)

2.3.3 Prediction problem

Similar to the reliability problem, given a maximum acceptable probability error $\epsilon > 0$, find a set $\mathcal{Y}_\epsilon \subseteq \mathcal{Y}$ such that

$$\mathcal{P}[\mathcal{M}(X) \in \mathcal{Y}_\epsilon] \geq 1 - \epsilon \tag{2.6}$$

in other works, the prediction $\mathcal{M}(X) \in \mathcal{Y}_\epsilon$ is wrong with probability at most ϵ.

2.3.4 Inverse problem or parameter estimation

Given some experimental measurements of the output Y of the system and some computer simulation results from its mathematical model $\mathcal{M}$, inverse uncertainty quantification estimates the discrepancy between the experiment and the mathematical model

(which is called *bias correction*), and estimates the values of unknown parameters in the model if there are any (which is called *parameter calibration*) (Gharib Shirangi, 2014). Generally this is a much more difficult problem than forward uncertainty propagation; however it is of great importance since it is typically implemented in a model updating process.

2.3.5 Sensitivity Analysis

Sensitivity analysis refers to the determination of the contributions of individual uncertain analysis inputs to the uncertainty in analysis results. The goal in sensitivity analysis is to apportion the uncertainty in Y to the uncertainty in inputs X, (SANKARARAMAN, 2012).

2.3.6 Model reduction or model calibration problem

Construct another model $\mathcal{M}_h$ such that $\mathcal{M}_h \approx \mathcal{M}$ in an appropriate sense.

2.3.7 Model selection

If, for the system F we have a set of models $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_n\}$, then a model selection problem consist in the selection of the most plausible model $\mathcal{M}_i$ that best fit the experimental data.

Sometimes a UQ problem consists of several of these problems coupled together, for example, one might have to solve an ***inverse problem*** to produce or improve some model parameters, and then use those parameters to propagate some other uncertainties ***forwards***, and hence produce a ***prediction*** that can be used for decision support in some ***certification problem*** (SULLIVAN, 2015).

In this thesis we focus in ***forward propagation problem*** although in chapter 6 we introduce some queries that help to solve ***reliability or certification*** and ***prediction problem***.

2.4 Methods for Forward Propagation

Two different methods are used to study how the uncertainty is propagated through a computational model, *intrusive* and *non-intrusive*. *Intrusive methods* require the modification of the mathematical/computational model. On the other hand, *non-intrusive methods* consider the mathematical/computational model as a black-box, and therefore the simulation codes don't need to be rewritten. For this reason, non-intrusive methods are attractive (KAWAI; SHIMOYAMA, 2014). The most popular methods for non-intrusive *UQ* are sampling methods, such as Monte Carlo (*MC*) and Latin hypercube sampling

(*LHS*). To date, the MC simulation is the most powerful method for the uncertainty evaluation, (**??**).

In this thesis we are not interesting into the methods itself because our principal objective is to process the uncertainty data generated as an output of a forward propagation process. To illustrate in a better way the problem, lets consider a *non-intrusive method* as an example.

To estimate a stochastic behavior of the output solution $\boldsymbol{q}$ in terms of input uncertainties $\boldsymbol{\theta}$, the sampling methods analyze the values of $\mathcal{M}(\boldsymbol{\theta})$ at multiple sampled conditions in the Θ space (called stochastic space or parameters space) directly from numerical simulations. Basically, *MC* and *LHS* methods randomly sample in the stochastic space, and hence both require many sample calculations to achieve a convergence of stochastic estimations (although the *LHS* method is more efficient than the *MC* method). As a result, the method returns multiple realizations of $\boldsymbol{q}$, and then other methods to measure the uncertainty, as those proposed in Section 2.2, need to be applied, (BAXTER; COOL, 2016; ESTACIO-HIROMS; PRUDENCIO, 2012; FARRELL; ODEN; FAGHIHI, 2015).

The selection of one UQ representation or other depend of the characteristics of each problems and the accuracy we are interesting in. But, as all results of interest can be derived from the **PDF** (COX et al., 2012), our objective in this thesis is to use **PDFs** as the representation of the uncertainty.

Derive a PDF from uncertain data is a fitting process. Given a random sample $q_1, q_2, q_3, ..., q_n$, the basic problem in fitting a statistical distribution to these data is that of approximating the distribution from which the sample was obtained.

Until now we review what uncertainty is, some definitions of uncertainty quantification, how to quantify the uncertainty, and some typical UQ problems. In the next section we explore what happens when we are in presence of a large-scale spatio-temporal model, which is the problem that really motivates this thesis.

2.5 UQ in Large-scale Spatio-temporal models

First to all lets define what is a large-scale spatio-temporal model (**LSSTM**). Although it is an extremely used term in the area, we don't find any exact definition in the reviewed literature. According to the Cambridge Dictionary, the mean of large-scale is:

Definition 2.6. Large-scale: involving many people or things, or happening over a large area.

In the simulation context **happening over a large area** is the most appropriated

term. The spatio-temporal part of the concept is straightforward. Join both ideas the definition of a large-scale spatio-temporal model could be something like:

Definition 2.7. Large-scale spatio-temporal model: a mathematical/computational model that study the spatio-temporal evolution of a physical system over a large area.

In this context, a computational model of the form: $\boldsymbol{q} = \mathcal{M}(\boldsymbol{\theta})$ represents the spatio-temporal evolution of a complex systems, and the *QoI* $\boldsymbol{q}$ can be represented as:

$$\mathbf{Q} = (\mathbf{q}(s_1, t_1), \mathbf{q}(s_2, t_2), ..., \mathbf{q}(s_n, t_n)) \tag{2.7}$$

where:

- $(s_1, t_1), (s_2, t_2), ..., (s_n, t_n) \in \mathcal{S} \times \mathcal{T} \subseteq \mathbb{R}^3 \times \mathbb{R}$ represent a set of distinct spatio-temporal locations, and
- $\mathbf{q}(s_i, t_j)$ represents a value of the *QoI* at the spatio-temporal location (s_i, t_j)

Many *QoI* can be analyzed, but for simplicity in this thesis we consider only one.

When dealing with **LSSTMs**, a huge amount of data is generated as a result of the simulation process. Indeed, on each spatio-temporal location $(s_i, t_j) \in \mathcal{S} \times \mathcal{T} \subseteq \mathbb{R}^3 \times \mathbb{R}$, usually more than 10^4 simulations are performed. Then, the size of the output dataset is in the order of $N_s \times N_t \times N_{sim}$, where: N_s is the number of spatial locations, N_t is the number of time steps, and N_{sim} is the number of simulations. An example of the volume of data generated by these simulations is given in the experimental Section **??**, where the output dataset is about 2.4 TB. From a computational point of view, this classifies *forward propagation* as a data intensive problem.

This information can be modeled as:

$$S(s_i, t_j, simId, q(s_i, t_j)) \tag{2.8}$$

where $simId$ represents the *id* of one simulation (realization).

The emerging field of data science, nevertheless, is largely lacking in generalizable methods for quantifying the uncertainty in the output of analyzed systems. As a result, a major new research initiative needs to be initiated in this area (TOBERGTE; CURTIS, 2013).

Remembered, we are interested in characterizing the uncertainty on each spatio-temporal location through a PDF. If we know, by theoretical considerations, that the distribution on each location is of certain type (e.g. normal, gamma, beta), then through moment matching, or some other means, one can determine a specific distribution that fits the data, (KARIAN; DUDEWICZ, 2011; Mustafa Inchasi, 2016). This is usually not the

case, even worse in **LSSTM**, because is impossible to know what could be a distribution type on each location. In those cases it makes sense to use a flexible family of distributions and choose a specific member of that family.

By a **flexible family** we mean one whose members can:

(i) assume a large variety of shapes: skewness in either direction, tails that are truncated or extend to infinity on either or both sides, bell-shaped distributions as well as inverted bell-shaped ones,

(ii) be able to represent a wide range of distributional characteristics such as moments (or combination of moments) or percentiles (or combinations of percentiles),

(iii) to have convenient expressions for at least one of their p.d.f., c.d.f., and quantile function,

(iv) no prior knowledge was needed to fit the distribution to a dataset, which is practical and suitable for automatic and software procedures.

Another challenge that emerge in **LSSTM** is related to the amount of data. It is not always convenient to retain the $N_s \times N_t \times N_{sim}$, (vector) values produced by MC and use them subsequently (COX et al., 2012). This fact introduce another desirable characteristic of the distribution family.

(v) reduce the amount of data to use in subsequent UQ analysis, after the **PDF** is obtained.

In Chapter 3 we present one of this distribution families an discus why this family is suitable to our proposes.

2.6 Summary

Summarizing, in this Chapter we present a review about UQ, showing what is uncertainty quantification, what are the principal mathematical tools used to represent it, and what are the typical UQ problems. Immediately we talk about some methods for uncertainty propagation and what are the problems we are interesting in. Basically, we are interesting into propose a new Big Data approach that allow to compute a **PDF** on each spatio-temporal location of the output of a forward propagation problem.

3 The Generalized Lambda Distribution

"There are good reasons for using the GLD distribution methods... GLD fits have been used successfully in many fields ... Try the GLD first and stop there if the results are acceptable."
(Karian and Dudewicz, 2011)

Fitting statistical distribution to data (real or simulated), is an important task in uncertainty quantification forward problem. When fitting data, one typically first selects a general class, or family, of distributions and then finds values for the distributional parameters that best match the observed data (LAKHANY; MAUSSER, 2000). One of this families is the Generalized Lambda Distribution (*GLD*), originally proposed by Ramberg and Schmeiser in 1974, as a generalization of the Tukey's distribution (1947). The *GLD* has been used in for years to fit data in the most diverse areas such us: analysis of brain MRI scan data, human twin data for quantifying genetic (vs. environmental) variance, rainfall distributions, radiation in soil, velocities within galaxies, exchange rate data for Japanese yen, and so on (KARIAN; DUDEWICZ, 2011). This is due to its flexibility to describe a variate of shapes and to provide good fits to many well know distributions.

Contradictorily, the *GLD* is not extensively used in UQ, despite the features that make it attractive for these purposes. In the present Chapter we show the features that motivated us to select the *GLD* as a distribution to characterize the uncertainty in our proposed approach.

The rest of the Chapter is organized as follow: Section 3.1 present the different parameterizations of the *GLD*, highlight the advantages and drawbacks of those parameterizations and justify the selected one. In Section 3.2 the shapes of the *FMKL* parametrization of the *GLD* are investigated. Next, in Section 3.3 the principal references to the numerical method to estimate the parameters of the *GLD* are shown, with some considerations and comments about future works in this area. Section 3.4 show how the *GLD* fit many well know distributions. Section 3.5 describe the flexibility of the *GLD* to fit mixture of distributions (bimodal and multimodals). The ability of the *GLD* as a random variate generator is presented in Section 3.6, that reinforces the idea of using it in UQ context. In Section 3.7 we resume all the characteristics of the *GLD* that make it suitable to UQ. In Section 3.8 we present **GLDEX** an R package that represent the state-of-the-art of the algorithms to work with the *GLD*. Finally in Section 3.9 we summarize the principal results of the Chapter.

3.1 The Generalized Lambda Distribution

The generalized lambda distribution is a continuous distribution defined in terms of it quantile function. Various parameterizations exist (see Section 3.1.3), but the most popular are the proposed by Ramberg and Schmeiser in 1974, Section 3.1.1; and the proposed by Freimer et al. in 1988, Section 3.1.2.

3.1.1 The Ramberg and Schmeiser Parameterization

The Generalized Lambda Distribution (GLD) was proposed by Ramberg and Schmeiser in 1974 as an extention of the Tukey's distribution. It is represented by the quantil function:

$$Q_{RS}(y|\lambda) = Q_{RS}(y|\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{y^{\lambda_3} - (1-y)^{\lambda_4}}{\lambda_2} \tag{3.1}$$

where $Q_{RS} = F^{-1}$ is the quantile function for probabilities y, wiht $y \in [0, 1]$; λ_1 and λ_2 are the location and scale parameteres, and λ_3 and λ_4 determine the skewness and kurtosis of the $GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$.

The probability density function of the GLD at the point $x = Q_{RS}(y)$ is given by:

$$f(x) = f(Q_{RS}(y)) = \frac{\lambda_2}{\lambda_3 y^{\lambda_3 - 1} + \lambda_4 (1-y)^{\lambda_4 - 1}} \tag{3.2}$$

In order to have a valid distribution function, the probability density function $f(x)$ need to be positive for all x and integrates to one over the allowed domain:

$$f(x) \geqslant 0 \tag{3.3}$$

$$\int f(x)dx = 1 \tag{3.4}$$

This impose complex constraints on the parameters and support regions of the RS parameterization, as summarized in table 1 and figure 2.

Table 1 – Support regions of the GLD and conditions on the parameters given by the RS parameterization to define a valid distribution function (KARIAN; DUDEWICZ, 2011). The support regions are displayed in Fig. 2. Note that there are no conditions on λ_1 to obtain a valid distribution.

Region	λ_2	λ_3	λ_4	Q(0)	Q(1)
1	< 0	$<= -1$ $-1 < \lambda_3 < 0$ $\frac{(1-\lambda_3)^{1-\lambda_3}(\lambda_4-1)^{\lambda_4-1}}{(\lambda_4-\lambda_3)^{\lambda_4-\lambda_3}} = \frac{-\lambda_3}{\lambda_3}$	$>= 1$ > 1	$-\infty$	$\lambda_1 + \frac{1}{\lambda_2}$
2	< 0	$>= 1$ > 1 $\frac{(1-\lambda_4)^{1-\lambda_4}(\lambda_3-1)^{\lambda_3-1}}{(\lambda_3-\lambda_4)^{\lambda_3-\lambda_4}} = \frac{-\lambda_4}{\lambda_3}$	$<= -1$ $-1 < \lambda_4 < 0$	$\lambda_1 - \frac{1}{\lambda_2}$	∞
3	> 0	> 0 $= 0$ > 0	> 0 > 0 $= 0$	$\lambda_1 - \frac{1}{\lambda_2}$ λ_1 $\lambda_1 - \frac{1}{\lambda_2}$	$\lambda_1 + \frac{1}{\lambda_2}$ $\lambda_1 + \frac{1}{\lambda_2}$ λ_1
4	< 0	< 0 $= 0$ < 0	< 0 < 0 $= 0$	$-\infty$ λ_1 $-\infty$	∞ ∞ λ_1

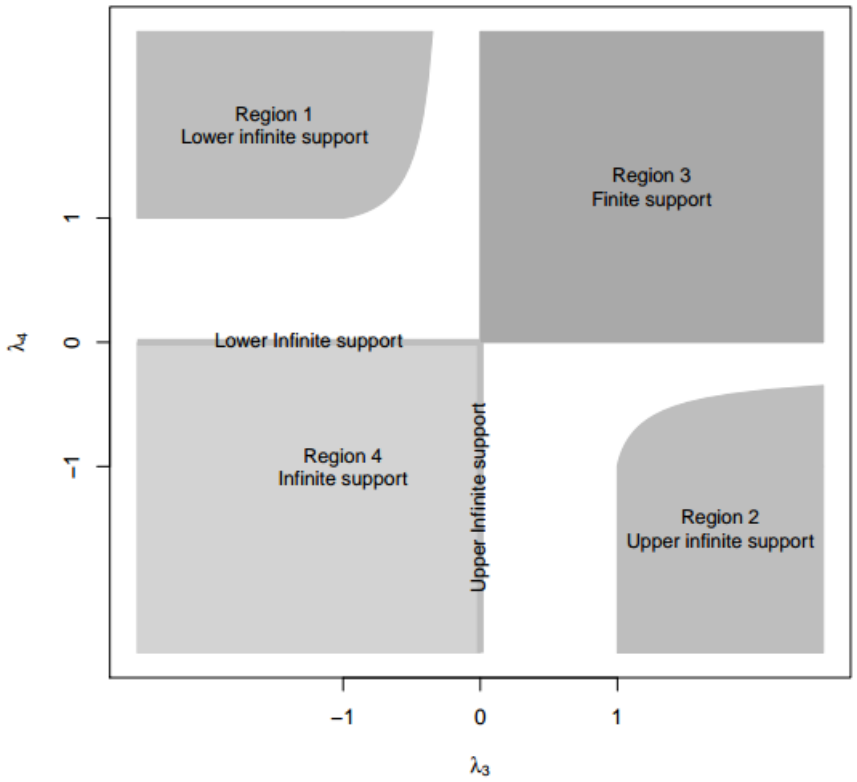

Figure 2 – Support regions of the GLD in the RS parameterization that produce valid statistical distributions.

Table 2 – Support regions of the *GLD* given by the *FMKL* parameterization (MARCONDES; PEIXOTO; MAIA, 2017).

λ_3	λ_4	Q(0)	Q(1)
> 0	> 0	$\lambda_1 - \frac{1}{\lambda_2\lambda_3}$	$\lambda_1 + \frac{1}{\lambda_2\lambda_4}$
> 0	≤ 0	$\lambda_1 - \frac{1}{\lambda_2\lambda_3}$	∞
≤ 0	> 0	$-\infty$	$\lambda_1 + \frac{1}{\lambda_2\lambda_4}$
≤ 0	≤ 0	$-\infty$	∞

3.1.2 The FMKL Parameterization

To circumvent the constraints on the *RS* parameter values, Freimer et al. (FREIMER; LIN; MUDHOLKAR, 1988) introduced a new parameterization called *FKML*, equation 3.5.

$$Q_{FMKL}(y|\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{1}{\lambda_2}\left[\frac{y^{\lambda_3} - 1}{\lambda_3} - \frac{(1-y)^{\lambda_4} - 1}{\lambda_4}\right] \tag{3.5}$$

As in the previous parameterization, λ_1 and λ_2 are the location and scale parameters, but in this one λ_3 and λ_4 are the tail index parameters. The advantage over the previous parameterization is that the only constraint on the parameters is that λ_2 must be positive. Figure 3 displays the support regions of the *GLD* in the *FKML* parameterization, table 2.

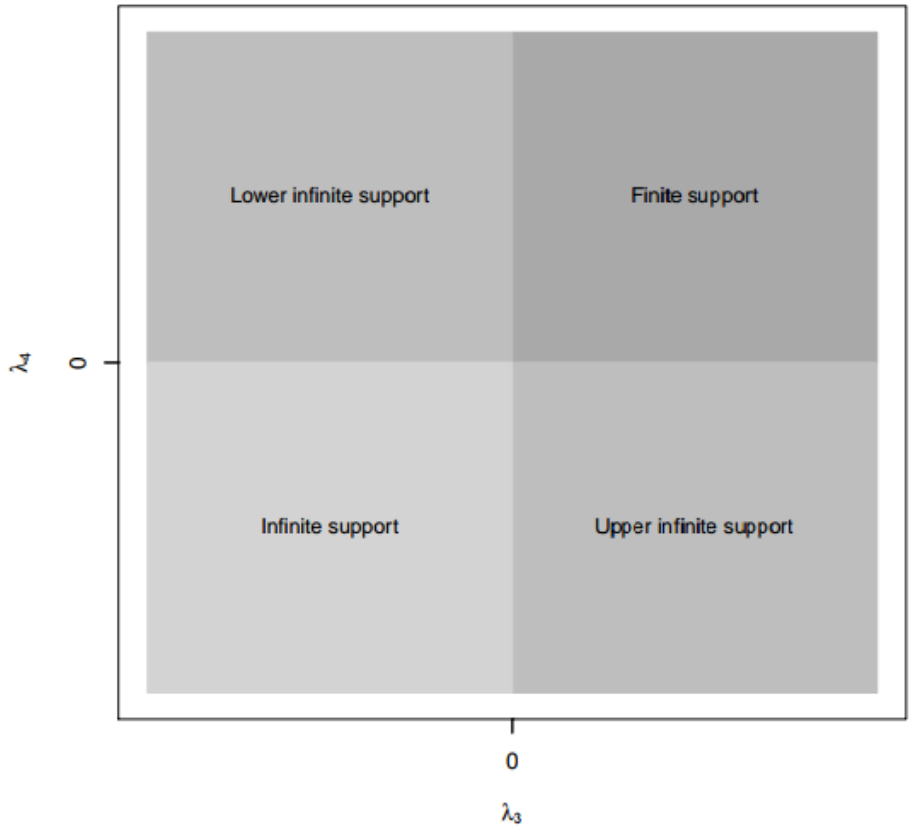

Figure 3 – Support regions of the *GLD* in the *FMKL* parameterization that produce valid statistical distributions.

The probability density function of the *FMKL-GLD* at the point $x = Q_{FMKL}(y)$ is given by (SU, 2015):

$$f(x) = f(Q_{FMKL}(y)) = \frac{\lambda_2}{y^{\lambda_3 - 1} + (1-y)^{\lambda_4 - 1}} \tag{3.6}$$

Although both the *RS* and *FMKL GLD* are generalizations of Tuckey's Lambda Distribution, they are not equivalent, so that the distribution fitted by one parametrization to a dataset differs in general from the one fitted by the other (MARCONDES; PEIXOTO; MAIA, 2017). Both of these representations can present a wide variety of shapes and therefore are utilized in practice; however, it is generally the *FMKL GLD* preferred due to the ease in its use (CORLU; METERELLIYOZ, 2016). In this thesis, we also use the *FMKL GLD* representation.

3.1.3 Other Parameterizations

One of the criticisms of the *GLD* is that its skewness is expressed in terms of both tail indices λ_3 and λ_4. In one approach addressing this concern, a five-parameter *GLD* was introduced by Joiner et al. (JOINER; ROSENBLATT, 1971), which, expressed in the

FKML parameterization, can be written as,

$$Q_{JR}(y|\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{1}{2\lambda_2}\left[(1-\lambda_5)\frac{y^{\lambda_3}-1}{\lambda_3} - (1+\lambda_5)\frac{(1-y)^{\lambda_4}-1}{\lambda_4}\right] \tag{3.7}$$

It has λ_1 and λ_2 as the location and scale parameters, and an asymmetry parameter, λ_5, which weights each side of the distribution and the two tail indices, λ_3 and λ_4. The conditions on the parameters are $\lambda_2 > 0$ and $-1 < \lambda_5 < 1$. The drawback of this parameterization is that the additional parameter can make the estimation of the parameter values even more difficult.

In (CHALABI; DIETHELM; SCOTT, 2012) the authors introduce a new parameterization of the *GLD* that transform the *FMKL* parameterization, equation 3.5 in terms of an asymmetry and steepness parameter without adding a new variable. Its major advantage is that provides an intuitive interpretation of its parameters. A new **R** package called **gldist** was implemented with the new *GLD* parameterization, along with the parameter estimation methods they present in his work. The problem with this parameterization is that the **R** package was removed from the official repository because of the code is out of date.

3.2 FMKL GLD Shapes

Both *RS* and *FMKL GLD* can describe a variety of shapes, examples: U-shaped, bell shaped, triangular, and exponentially (SU, 2007). At the same time they also provide good fits to many well know distributions. In the case of *RS GLD* and extensive study can be found in (KARIAN; DUDEWICZ, 2011), for the *FMKL GLD* see (FREIMER; LIN; MUDHOLKAR, 1988).

Those properties of the *GLD* are important to our purpose for two reasons: first we don't need previous knowledge to use the *GLD* to fit any dataset, that is exactly the case in large-scale spatio-temporal models, and second the *GLD* can be compared and grouped based on its shapes, that allow us to answer the **RQ1** as we show in Chapter 4.

As the *FMKL GLD* parameterization is the one selected to be used in this thesis, in the next sub-sections we present a brief review of its shapes and how this parameterization fit some well-known distributions.

The shape of the *GLD* are dependent of its λ values. In the case of the *FMKL GLD* parameterization, Freimer et al. (FREIMER; LIN; MUDHOLKAR, 1988) classify the shapes into five categories depending on the variety of distributions that can be represented by the several combinations of the shape parameters λ_3 and λ_4. In particular, Class-I ($\lambda_3 < 1, \lambda_4 < 1$) represents unimodal densities with continuous tails. This class is subdivided in I_a ($\lambda_3, \lambda_4 \leq 1/2$), I_b ($1/2 < \lambda_3 < 1, \lambda_4 \leq 1/2$) and I_c ($1/2 < \lambda_3 < 1, 1/2 < \lambda_4 < 1$).

In I_a we find distributions such us $Gaussian(Normal)$, $Beta(2.3)$ and $\Gamma(\alpha = 5)$; in I_b $\Gamma(\alpha = 3)$ and $Lognormal(\sigma = 0.5)$; and in I_c distributions as the example of Class-I in Figure 4.

Class-II ($\lambda_3 > 1$, $\lambda_4 < 1$) represents monotone pdfs similar to the *Exponential* distribution, $Beta(1.2)$ or $Lognormal(\sigma = 1.0)$. Class-III ($1 < \lambda_3 < 2$, $1 < \lambda_4 < 2$) represents U-shaped densities with truncated tails, Class-IV ($\lambda_3 > 2$, $1 < \lambda_4 < 2$) represents S-shaped densities, and Class-V ($\lambda_3 > 2$, $\lambda_4 > 2$) represents unimodal densities with truncated tails. Figure 4 provides the shapes that are represented by the parameters indicated in Table 3.

Table 3 – Examples of the five categories of distributions the *FMKL GLD* can represent.

	λ_1	λ_2	λ_3	λ_4
Class-I	0	1	0.5	0.6
Class-II	0	1	2	0.5
Class-III	0	1	1.5	1.5
Class-IV	0	1	2.5	1.5
Class-V	0	1	3	3

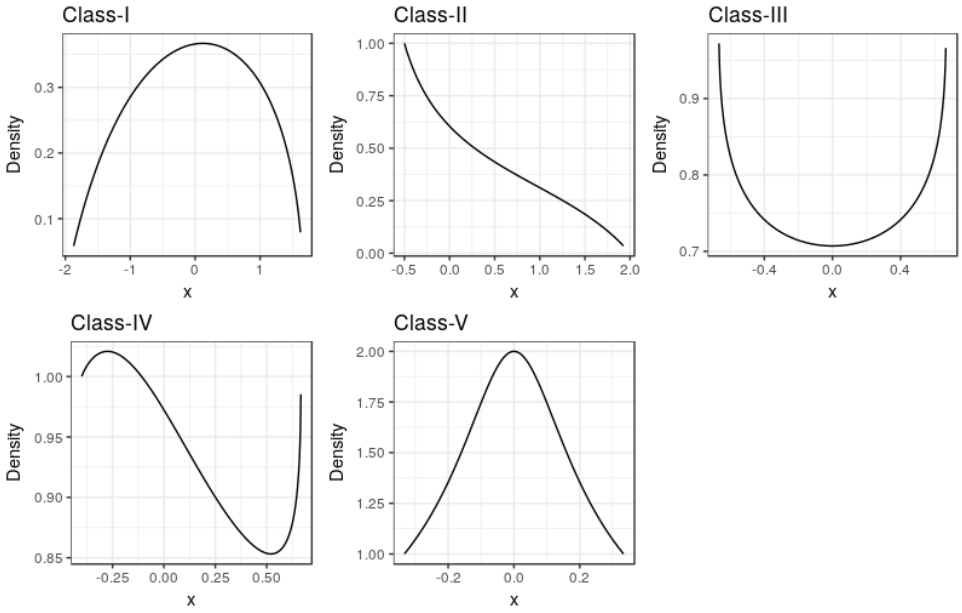

Figure 4 – Examples of the five categories of shapes the *FMKL GLD* can represent.

Figure 5 show the five categories of the *FMKL GLD* shapes in (λ_3, λ_4) space. There are two regions in this figure that were left out of the analysis, the regions with ($\lambda_3 < 1$, $\lambda_4 > 1$) and ($1 < \lambda_3 < 2$, $\lambda_4 > 2$). Those regions are symmetric to region II and IV respectively, see Figure 6.

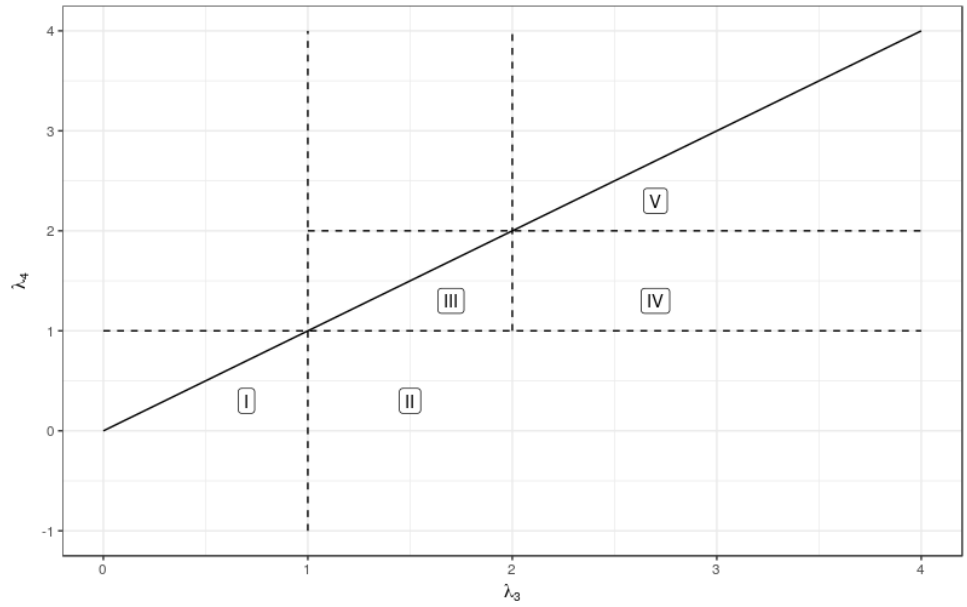

Figure 5 – The five categories of shapes of the *FMKL GLD* in the (λ_3, λ_4) space.

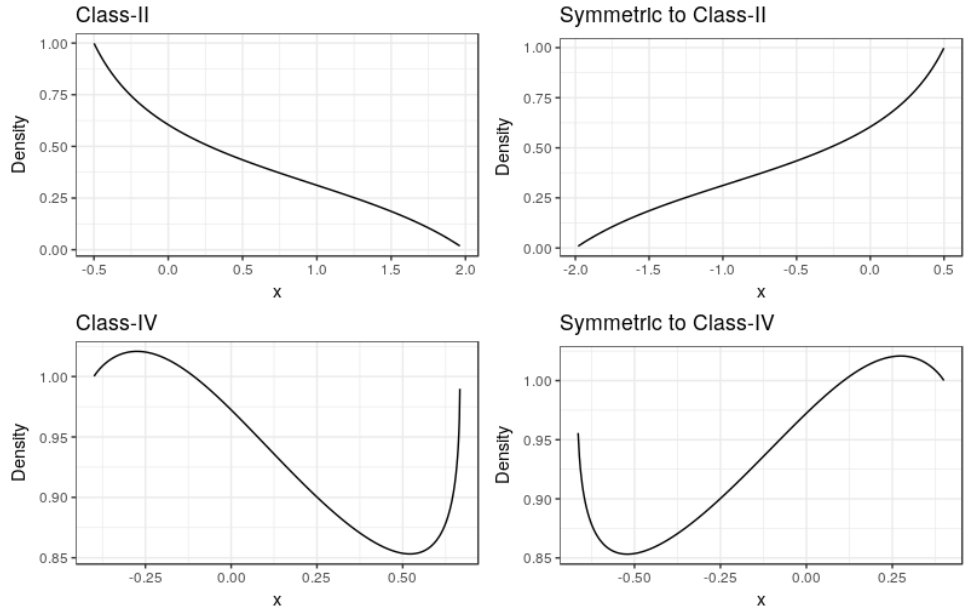

Figure 6 – Symmetry of the regions ($\lambda_3 < 1$, $\lambda_4 > 1$) and ($1 < \lambda_3 < 2$, $\lambda_4 > 2$) with respect to region II and IV.

The results presented in this section are extremely important to our approach because they are the bases of the clustering algorithms we propose in Chapter 4.

3.3 Numerical Methods to Fit the GLD to Data

Given a random sample $x_1, x_2, x_3, ...x_n$, the basic problem in fitting a statistical distribution to this data is that of approximating the distribution from which the sample

was obtained. If it is known, because of theoretical considerations, that the distribution is of a certain type (e.g., a gamma distribution with unknown parameters), then through moment matching, or some other means, one can determine a specific distribution that fits the data. This, however, is generally not the case and, in the absence of any knowledge regarding the distribution, it makes sense to appeal to a flexible family of distributions and choose a specific member of that family (KARIAN; DUDEWICZ, 2011).

There are two different parameter estimation philosophies, **direct estimation methods**, such as least-squares estimation with order statistics and with percentiles (FOURNIER et al., 2007; KARIAN; DUDEWICZ, 2011); the methods of moments (LODZIENSIS, 2013), L-moments (KARVANEN; NUUTINEN, 2008), and trimmed L-moments (FOURNIER et al., 2007); and the goodness-of-fit method with histograms (**??**) and with maximum likelihood estimation (SU, 2007). On the other side, **stochastic methods** have been introduced with various estimators such as goodness-of-fit (LAKHANY; MAUSSER, 2000) or the starship method [King and MacGillivray, 1999].

Without doubts the major contributions in the implementation of parameter estimation algorithms are due to Steve Su (SU, 2007; SU, 2011; SU, 2015; SU, 2016), that besides the theoretical contributions is the author of the state-of-the-art R package to work with the *GLD*. A brief review of this package is presented in Section 3.8, as this package is the one we use in this thesis to solve many task related to the *GLD*.

Out of the two estimation philosophies presented above, (CORLU; METEREL-LIYOZ, 2016) a genetic algorithms approach to estimate the parameters of the *GLD* was introduced.

In the current year Marcondes at al. (MARCONDES; PEIXOTO; MAIA, 2017) present a new parameterization of the *GLD* with its respective numerical methods. The main contribution of this paper is that the new parameterization allow to fit the *GLD* to highly skewed data, with a great number of zeros and heavy tails.

The methods to fit the *GLD* to data are out of the scope of this thesis, as we are interesting into demonstrate its usability in UQ. Nevertheless is important to remark that, fit the *GLD* to data is computationally intensive but suitable to parallelize, but we don't find any work in the literature to explore the possibility of increase the performance of the fitting process by mean of parallelization. This is an open problem we are interesting into explore in the future.

3.4 GLD Approximations of Some Well-Known Distributions

For the $GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ to be useful for fitting distributions to data, it should be able to provide good fits to many of the distributions the data may come from. In

(KARIAN; DUDEWICZ, 2011), the authors explore how the *GLD* fit sixteen well-known distributions using the *RS GLD* parameterization. Here we explore how the *GLD* fit eight distributions, but using the *FMKL* parameterization.

In Table 4 we show the original distribution, the four λ values of the fit and the result of apply a Kolmogorov-Smirnov test to validate the good of the fit.

Table 4 – GLD Approximations of 8 Well-Known Distributions

Distribution	Parameters	λ_1	λ_2	λ_3	λ_4	KS-test
Normal	$N(0,1)$	-0.04263	1.49039	0.13787	0.12571	951
Uniform	$U(0,1)$	0.46250	2.16223	1.00008	0.8614	912
Exponential	$\theta = 1$	0.49150	1.40546	1.44813	-0.10419	923
Chi-Square	$nu = 5$	4.141853	0.486702	0.508298	-0.045440	911
Gamma	$(\alpha = 5, \theta = 3)$	1.535078	1.846312	0.410183	0.027492	885
Weibull	$(\alpha = 1, \beta = 5)$	2.684657	0.263865	1.413826	-0.067658	940
Lognormal	$(\mu = 0, \sigma = 1/3)$	0.984696	4.516254	0.324879	-0.074348	903
Beta	$(\beta_3 = \beta_4 = 1)$	0.50092	2.00702	0.99505	1.00060	906

As was expected the results are slightly different to those presented by Karian et al., as we use other parameterization. The KS-test value was over 900 in seven cases and near 900 in the case of the Gamma distribution. This result suggests us that the fit was good. All the fit provides (λ_3, λ_4) values that match the shape region each distribution belongs to.

3.5 Fitting Mixture Distributions Using a Mixture of Generalized Lambda Distributions

In general, a *mixture distribution* is the probability distribution of a random variable that is derived from a collection of other random variables. Mathematically, given a finite set of *PDFs* $p_1(x), p_2(x), \ldots, p_n(x)$, and weights $w_1, w_2, \ldots, w_n$ such that $w_i \geq 0$ and $\sum w_i = 1$, the mixture distribution can be represented by writing the density $f(x)$ as a sum (which is a convex combination):

$$f(x) = \sum_{i}^{n} w_i p_i(x) \tag{3.8}$$

Since its introduction by Karl Pearson in 1894 *mixture distribution* are extensively used, but in the majority of the cases using normal mixtures. The advantage of using the GLD family is that the GLD can fit the normal well, hence whenever a mixture of normals will fit data well, so will a mixture of at most the same number of GLDs. Meanwhile, the GLD family is a much broader family, and can do well in cases where the normal cannot (NING; GAO; DUDEWICZ, 2008). Due to the versatile and rich shapes of the *GLD*, they

are particularly suited for mixture modeling as they eliminate the need to choose between a wide range of different distributions on the same data set (SU, 2007).

The *GLD* mixture distribution can be represented as:

$$f(x) = \sum_{i}^{n} w_i GLD_i(\lambda_1, \lambda_2, \lambda_3, \lambda_4) \tag{3.9}$$

Some studies that compare the use of *GLD* mixtures with normal mixtures are presented by Ning et al. (NING; GAO; DUDEWICZ, 2008), where the author show that the mixture of *GLDs* perform as well as the mixture of normal distributions, and sometimes even better. In (SU, 2011) the author present examples of fitting bimodal and trimodal data with mixture of *GLDs*, again with excellent results. Numerical methods to fit mixture of *GLDs* to data are discus in (SU, 2007; SU, 2011), while the implementations are part of the **GLDEX** R package (SU, 2007).

We go back to *GLD* mixtures in Chapter 5 where present how to answer the **RQ.3** by the use of a mixture of *GLDs*.

3.6 GLD Random Variate Generation

An important thing to take into account when we substitute the raw data produced as an output of a simulation process, by its *PDF* is that those *PDFs* need to allow us to reproduce the original data as close as possible. The outcome produced by a particular *PDF* is known as random variate, its definition is:

Definition 3.1. A **random variate** is a particular outcome of a random variable. The random variates which are other outcomes of the same random variable might have different values.

Random variates are used when simulating processes driven by random influences. One of the important applications of the *GLD* has been the generation of random variables for Monte Carlo studies (Mustafa Inchasi, 2016).

This fact is justified by the following theorem, enunciated by Karian and Dudewicz (2010).

Theorem 3.2. *If $Q_X(y)$ is the percentile function of a random variable X, and U is a uniform random variable on $(0, 1)$ then $Q_X(U)$ has the same PDF as does X.*

For a proof, also see p. 156 of Karian and Dudewicz (1999). The percentile function is not available in a closed (or easy-to-work-with) form for many of the most important distributions, such as the normal distribution. However, the GLD is (see sections **??** and **??**) defined by its p.f., which is a simple-to-calculate expression.

Thus, **r.v.s for a simulation study can easily be generated from any distribution that can be modeled by a GLD.**

Example 3.3. Suppose we have modeled an important **r.v.** by an approximate standard normal distribution X. We show in Section 3.4 that a close fit to the standard normal is available via the *RS-GLD* with

$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0, 0.1975, 0.1349, 0.1349) \tag{3.10}$$

and this *GLD* has **p.f.**

$$Q(y) = \frac{y^{0.1349} - (1-y)^{0.1349}}{0.1975} \tag{3.11}$$

Thus, if $U_1, U_2, ...$ are independent uniform **r.v.s** on $(0, 1)$, then

$$Q(U_1), Q(U_2), ... \tag{3.12}$$

are independent and (approximately) $N(0, 1)$ **r.v.s** for the simulation study at hand.

This theorem means that, independently of the nature of the dataset (Normal, Exponential, etc.), when we fit a GLD to it, we can proceed similarly to the example above. That is, we just need to generate a stream of independent uniform **r.v.s** on $(0, 1)$, and then evaluate the equation 3.12. There are a number of good sources of independent uniform r.v.s on $(0, 1)$ (KARIAN; DUDEWICZ, 2011).

This is an important property of the *GLD* that allow us to substitute the raw data by the four lambdas of the *GLD* that best fit it (if the fit is a good one), with the warranty that if we need to go back, the *GLD* could generate a good representation of the original data.

It is evident that the *GLD* allows easy generation of random variables from every kind of distribution, because featuring an explicit and accessible $Q_X(y)$ reduces it to a uniform generation in $[0, 1]$, (LAMPASI; Di Nicola; PODESTA, 2006).

3.7 GLD and Uncertainty Quantification

From the best of our knowledge, the first effor to use the *GLD* in UQ is due to Lampasi et al. (LAMPASI; Di Nicola; PODESTA, 2006) in the paper *"Generalized Lambda Distribution for the Expression of Measurement Uncertainty"*. In this paper the authors argue why the *GLD* is suitable for UQ, (i) how the use of the *GLD* to represent the uncertainty both at the input and the output of the models helps to homogenize the information we are processing in UQ worflows, (ii) the *GLD* allows easy generation of random variables from every kind of distribution (see Section 3.6), (iii) the amount of data we need to storage is extremely smaller than required by histograms or by the approximation

presented in (JCGM, 2008), as a *GLD* is full described by its four parameters, and (iv) how their introduction is practical and suitable for automatic and software procedures, as required by the industrial standards.

Later Cox et al. (COX et al., 2012) comment that as a *GLD* is defined with respect to its quantile function, drawing random samples from the resulting model approximation is straightforward, and then its use is suitable because is not always convenient to retain the among of values produced by MC simulations and use them subsequently. Then we can substitute the raw data generated in MCS by the *GLDs* that represent this raw data.

A more general reference of the use of the *GLD* in UQ is due to Hack et al. in (**??**). They do a literature review about UQ and include the *GLD* as a very interesting option to characterize the uncertainty.

In (MOVAHEDI; LOTFI; NAYYERI, 2013), a solution to determining the reliability of products using the *GLD* is presented. The novelty here is because of the variability of distributions of the products they need a flexible distribution family that allow to fit those distributions without previous knowledge.

Finally, in (RAJAN et al., 2016) the authors present a benchmark test distributions for expanded uncertainty evaluation algorithms. Between many other distributions, the *GLD* is included because of its potential use in UQ.

3.7.1 Relevance of GLD in Uncertainty Quantification

The use of the *GLD* to quantify the uncertainty is justified because:

- the *GLD* fits the *PDF* of a wide variety of datasets, including those that follow distributions such as normal, uniform, Student's t, U-shaped, exponential, etc;
- no prior knowledge is needed to fit the *GLD* to a dataset, which is practical and suitable for automatic and software procedures;
- the *PDF* is completely characterized by the four parameters of the *GLD*, which represents a reduction in the amount of data that must be stored for post-processing;
- the shape of the *GLD* is governed by its parameters, so the *GLDs* can be grouped based on their shapes, which is especially useful for further queries;
- in cases where mixture of distributions are needed, *GLD* mixtures could be a very good option; and
- the *GLD* allows easy generation of random variables from every kind of distribution.

3.8 The GLDEX R package

In the implementation of our approach, we use the ***GLDEX***[1] ***R*** package (SU, 2007). The GLDEX R package provides fitting algorithms with two objectives: (i) to provide a smoothing device to fit distributions to data using the weight and unweighted discretised approach based on the bin width of the histogram; (ii) to provide a definitive fit to the data set using the maximum likelihood estimation.

The GLDEX package also provides diagnostic tests to examine the quality of fit through the resample Kolmogorov-Smirnoff test, quantile plots and comparison of the mean, variance, skewness and kurtosis between the empirical data and the fitted distribution.

The GLDEX package is used in this thesis to: (i) fit the *GLD* distribution to a dataset on each spatio-temporal location; (ii) examine the quality of the fit; (iii) sampling any spatio-temporal location based on its *GLD*.

3.9 Summary

The main contribution of this Chapter is to summarize the relevance of *GLD* in UQ. Besides that, we select the *FMKL* parameterization as the one we use in this thesis, because it is more general than the *RS-GLD*. The shapes of the *FMKL-GLD*, the capacity to describe many well know distributions and the fact that *GLD* mixtures outperform normal mixtures, allow us to conclude that the *GLD* is a good candidate to be used in our approach.

The *GLD* also fit the five requirements enunciated in Section 2.5 of Chapter 2, as the desirable characteristics of a flexible family, to by used in the quantification of uncertainty in **LSSTM**.

[1] https://cran.r-project.org/web/packages/GLDEX/index.html

4 Clustering Uncertain Data Based on GLD Similarity

In Chapter 3 we exposed the two most important parameterizations of the *GLD* and select the *FMKL* as the one to be used for the rest of the thesis. In this parameterization λ_1 represent the location of the *GLD* and is directly related to the mean of the distribution. λ_2 is the scale, directly related to the standard deviation; and λ_3 and λ_4 represent the left and right tails of the distribution. Combinations of λ_3 and λ_4 can be used to estimate the skewness and kurtosis of the distribution.

As λ_2 define the dispersion, and λ_3 and λ_4 the shape of a *GLD*, so the combination of those parameters are the responsible of the quantification of the uncertainty, from the *GLD* point of view.

The **RQ.1** we formulate in the Introduction (see Chapter 1) is:

> **RQ1.** how to group the output of the UQ process based on the similarity of the uncertainty?

If the uncertainty in a *GLD* is characterized by λ_2, λ_3 and λ_4 and we are interesting into group the uncertainty based on the similarity, is clear that we need to explore how to group the *GLDs* based on its λ values. This is the main objective of this Chapter, that is organized as follow: in Section 4.1 a brief review of some related works is performed, and the advantage and drawbacks of those works are highlighted. Some considerations about the possibilities of the use of the *GLD* to solve some of the drawbacks are commented. Next, in section 4.2 our hypothesis about the use of the *GLD* to clustering uncertain data, is presented. Sections 4.3 and 4.4 present two synthetics datasets and the results of the clustering technique. Those results help us to validate our hypothesis. Finally, section 4.5 summarize and discuss the main results of the Chapter.

4.1 Related Works

Clustering uncertain data is recognized as an essential tasks in mining uncertain data (JIANG et al., 2011). It impose significant challenges in both modeling similarities between uncertain objects and developing efficient computational methods. Many methods are implemented as extensions of k-means and density-based clustering methods like DBSCAN, but using geometric distances between objects. Such methods cannot handle uncertain objects that are geometrically indistinguishable, such as products with the same

mean but very different variances, as the uniform and normal distributions in Figure 7.

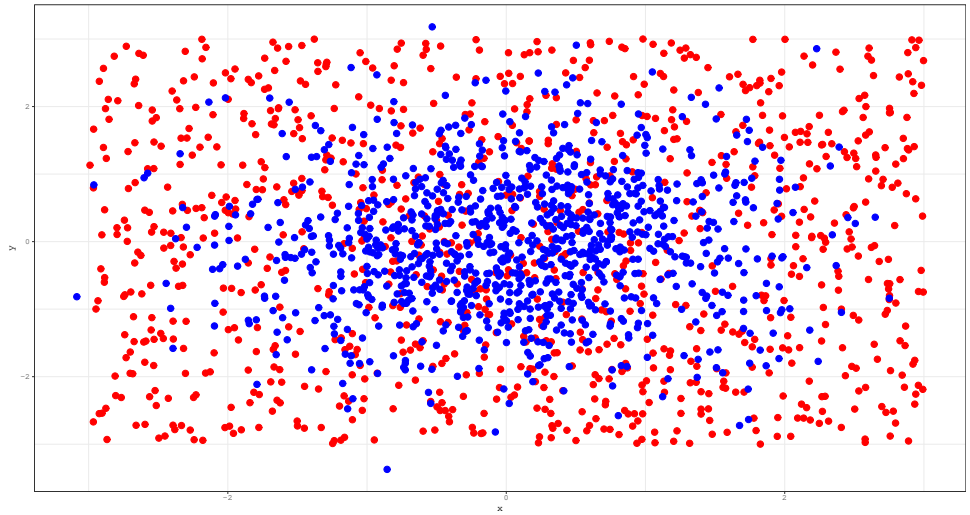

Figure 7 – Two geometrically indistinguishable distributions. Both distributions have the same mean but different variances. In color blue a bivariate Gaussian distribution, in red a bivariate uniform distribution.

Until recently, probability distributions, which are essential characteristics of uncertain objects, had not been considered in measuring the similarity between uncertain objects. In the last years, a plethora of new methods to clustering uncertain data emerge, based in the use of statistical metrics like Kullback-Leibler divergence (KL-divergence). From the best of our knowledge, the first effort in this direction was the work of Jiang et al. (JIANG et al., 2011), where they propose a new clustering method based on the use of Kernel Density Estimation (KDE) to fit a *PDF* to the uncertain data, and modifications to the k-means algorithm to use the KL-divergence as a distance function. Some improvements to the computational cost of the previous method were introduced by the same authors in (JIANG et al., 2013).

More recently Lui et al. (LIU; NIU; LIAO, 2018) present some contributions to the k-means algorithm to clustering uncertain objects but with unsatisfactory results. The proposed approach could be useful just to find the optimal number of clusters.

Basically, those methods are two steps methods: (i) first a method to find the *PDF* that best fit the uncertain objects, and (ii) the clustering algorithm to group the uncertain objects using an statistical metric as a distance function. Two major drawbacks can be remarked here: first, as we mention in Chapter 3, find the *PDF* that best fit an uncertain object is not so easy. And second the computation of any statistical distance between many objects is a time consuming and computationally intensive task. Other restrictions are also imposed to some methods, for example in (JIANG et al., 2011) all the *PDF* need to be defined in the same domain.

Some of those drawbacks can be solved by the use of the *GLD*. For example, the *GLD* is defined in the same domain, we don't need any previous knowledge to fit the *GLD* to an uncertain object. In the next section we formulate an hypothesis about how the *GLD* can be used to clustering uncertain data.

4.2 Clustering Based on GLD

According to Lampasi et al. (LAMPASI; Di Nicola; PODESTA, 2006), a particular $GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ can be rewrite as:

$$GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{1}{\lambda_2} GLD(0, 1, \lambda_3, \lambda_4) \tag{4.1}$$

Based on the second term of Equation 4.1, our hypothesis is that we can group the uncertainty using clustering algorithms above λ_2, λ_3 and λ_4, looking to λ_2 first and then refine the groups with the values of λ_3 and λ_4. As λ_2 characterize the dispersion (variance or standard deviation) of the *GLD*, in the first step of the algorithm we group all the *GLDs* with similar dispersion. Dispersion alone don't tell us how the data is distributed, then to refine the clusters we look at λ_3 and λ_4, that are the parameters that define the shape of the distribution.

Intuitively, suppose we have two distributions, one normal and one exponential, and both distributions have the same standard deviation. The value of λ_2 of both distributions is similar. In the first step of the algorithm both distributions are grouped together. But, as the normal distribution is symmetric it has left and right tails, differently of the exponential that only have tail in one direction. The values of λ_3 and λ_4 of both distributions are dissimilar, then in the second step of the algorithm both distributions are separated in different clusters.

To test our hypothesis, we generate two synthetic datasets using 4 different probability density functions: Gaussian, Exponential, Uniform and Gamma. The structure of the datasets is represented in 4.2.

$$S(x_i, < v_j >) \quad i = 1.....n, j = 1.....m \tag{4.2}$$

where:

- n represent the number of objects of the dataset and,
- m represent the size of each object.

The datasets are described in details in Sections 4.3 and 4.4.

4.2.1 Fit the GLD to a dataset

When we generate a synthetic dataset, the next step is to find the *GLD* that best fit $< v_j >$ on each x_i. As the fitting process is computationally intensive we implement a parallel algorithm using **R**. The pseudo-code is shown in Algorithm 1.

Algorithm 1 Fitting the GLD to a synthetic dataset

1: **function** GLDFIT($S(x_i, < v_1, v_2, ..., v_n >)$)
2: $< \lambda_1, \lambda_2, \lambda_3, \lambda_4 > \leftarrow$ FIT.GLD.LM($< v_1, v_2, ..., v_n >$)
3: $isValid_{(x_i)} \leftarrow$ VALIDITYCHECK($< \lambda_3, \lambda_4 >$)
4: **if** $isValid_{(x_i)}$ **then**
5: $[pvalue, D]_{(x_i)} \leftarrow$ KS($< \lambda_1, \lambda_2, \lambda_3, \lambda_4 >_{(x_i)}$)
6: **if** $pvalue_{(x_i)} > 0.05$ **then**
7: STORELAMBDAS($< \lambda_1, \lambda_2, \lambda_3, \lambda_4 >, x_i$)

The algorithm receive a dataset represented by 4.2 and, for each position x_i, call a function ***fit.gld.lm*** from the **R** package **GLDEX** presented in section 3.8, line 2 of Algorithm 1. In line 3 we check the validity of the *GLD* returned by the function (remember from Chapter 3 that the *GLD* is not always valid). In line 5 a good-of-fit test is perform to be sure that each *GLD* is a good representation for the dataset in x_i. Finally all the *GLD* with $pvalue > 0.05$ are stored to be used in the next section.

The final result of this process is a new dataset with the form:

$$S(x_i, < \lambda_1, \lambda_2, \lambda_3, \lambda_4 >) \quad i = 1.....n \tag{4.3}$$

4.2.2 Clustering the GLD

The clustering algorithm follow the hypothesis mentioned above, see Algorithm 2. The dataset 4.3 is modified to remove λ_1 as we don't use it in the clustering process.

Algorithm 2 Clustering the GLD based on its $\lambda_{(2,3,4)}$ values.

1: **function** GLDCLUSTERING($S(x_i, < 0, \lambda_2, \lambda_3, \lambda_4 >)$)
2: $S(x_i, clusterID_I) \leftarrow$ FIRSTCLUSTERSTEP($S(x_i, \lambda_2)$)
3: **for each** $clId_I$ **do**
4: $S(x_i, clusterID_{II}) \leftarrow$ SECONDCLUSTERSTEP($S(x_i, < \lambda_3, \lambda_4 >), S(x_i, clId_I)$)

To test the influence of the three parameters in the clustering process, we first run the Algorithm 2 as it, over $< \lambda_2, \lambda_3, \lambda_4 >$. But in the second experiment we run it over $< \lambda_3, \lambda_4 >$. The results are discuses in sections 4.3 and 4.4.

4.3 Synthetic Data I

To generate the first synthetic data set we use 11 probability density functions, where 5 are Gaussian, 5 Exponential, and one Uniform, figures 8, 9 and 10. The standard deviation of the 5 Gaussian distributions is $0.05 * i$, with $i = 1, 2, 3, 4, 5$, and we generate 90 samples of each distribution. This is, the first 90 objects where generated from a Gaussian distribution with standard deviation 0.05, and so on. Similarly, the rate of the 5 Exponential distributions is i, with $i = 1, 2, 3, 4, 5$, and again we generate 90 samples of each one. Finally, 100 samples of a Uniform distribution between $[0, 1]$ were generated. In resume, we have 1000 objects, where the first 450 were sampled from a Gaussian distributions, the next 450 from an Exponential and the last 100 from a Uniform distribution. As we generate a synthetic dataset in this way, we have the ground truth of the clustering in the dataset. This ground truth is used to evaluate the clustering quality of our algorithms.

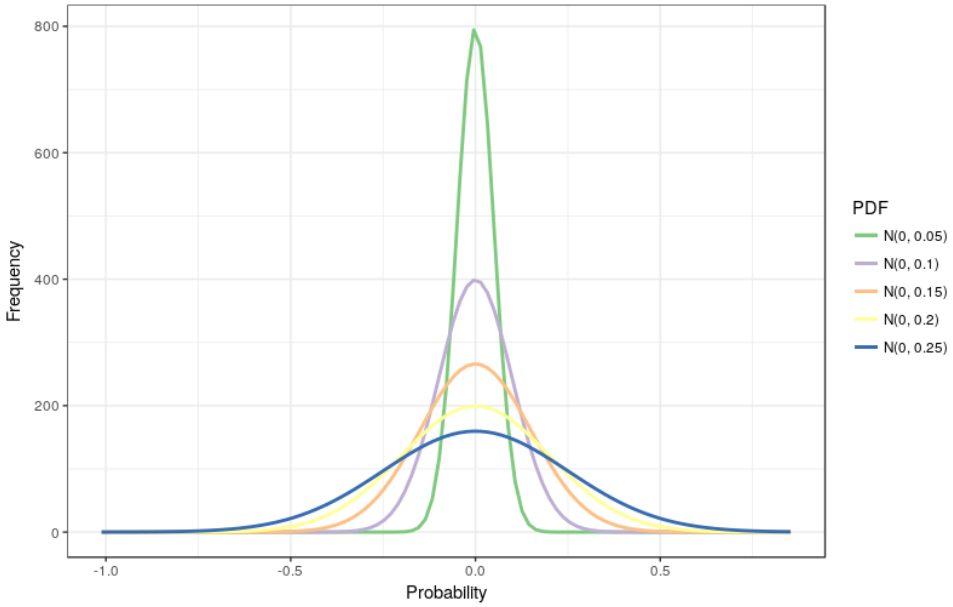

Figure 8 – Gaussian (Normal) distributions used to generate the synthetic dataset.

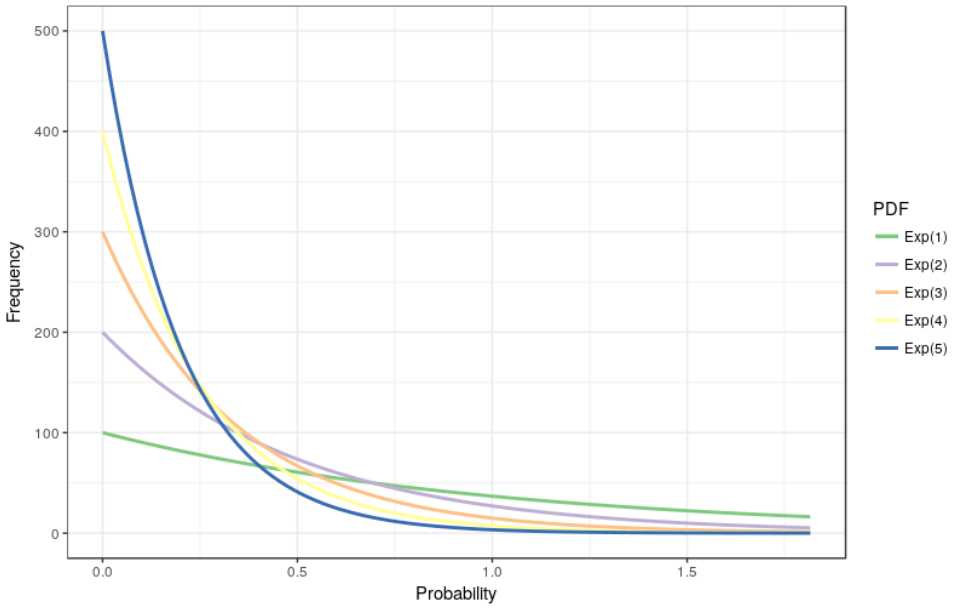

Figure 9 – Exponential distributions used to generate the synthetic dataset.

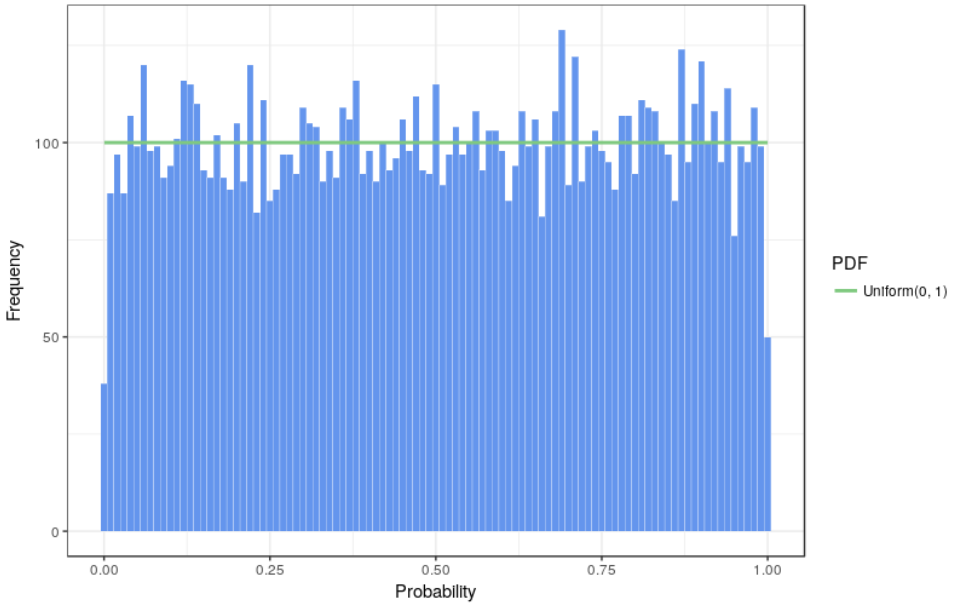

Figure 10 – Uniform distribution used to generate the synthetic dataset.

This dataset could be represented as a multidimensional array where for each position x_i, we have 1000 values v_j, equation 4.4. In this case i and j vary from 1 to 1000 casually.

$$S(x_i, < v_j >) \quad i, j = 1, 2.....1000 \tag{4.4}$$

The fitting algorithm proposed in subsection 4.2.1 is applied over 4.4. The good-of-fit test return that all the *GLDs* are good fit for its corresponding distribution. As a result

the dataset 4.5 is genarated.

$$S(x_i, < \lambda_1, \lambda_2, \lambda_3, \lambda_4 >) \quad i = 1.....1000 \tag{4.5}$$

4.3.1 Clustering using λ_2, λ_3 and λ_4

As we mention above, our idea is to test what happen if we use a simple k-means with euclidean distance over the λ_2, λ_3 and λ_4 values of the *GLDs*. Similar to the paper (JIANG et al., 2011), as we use 11 *PDFs* to generate the synthetic dataset I, we expect that the k-mean algorithm will return 11 clusters as well (one for each distribution). Then 11 is the number we use with the k-means algorithm.

In figure 11 and table 5 the distribution of the clusters returned by the k-means is shown.

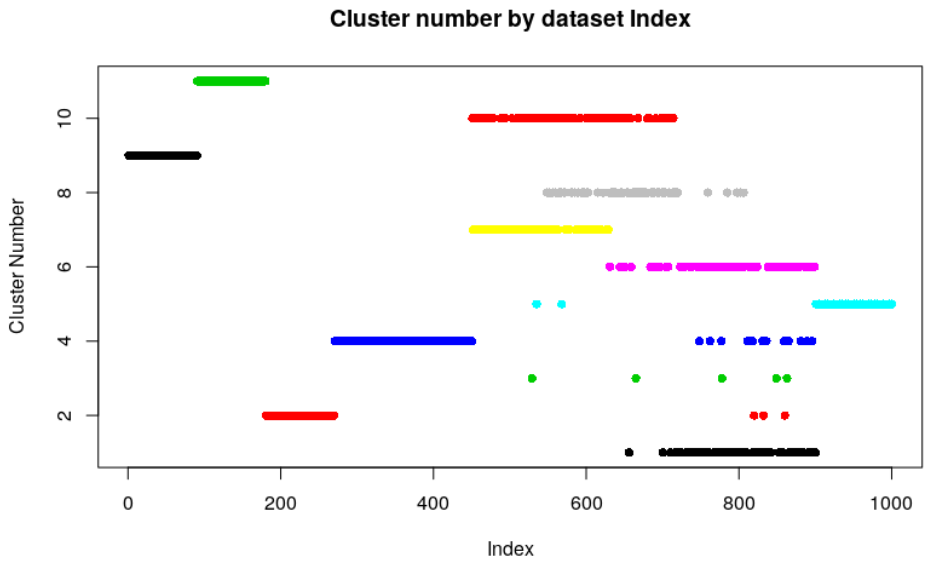

Figure 11 – Distribution of the clusters using k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*.

Remembered, the first 450 elements are normal distributions, the next 450 are exponential and the last 100 are uniform. Looking to those regions in general, the first observation is that we have 23 false positives, three in cluster 2, 18 in cluster 4 and two in cluster 5. The second observation is that, the normal distributions were grouped in 4 clusters (2, 4, 9 and 11), cluster 2 group perfectly it 90 elements with 2 false positives, clusters 9 and 11 group exactly its 90 elements each. The cluster 4 group the last 180 elements of the Normal distribution, with 18 false positives

The Uniform distribution was grouped totally in cluster 5, with two false positives as was mention above. The last observation is that the algorithm can't separate the 5

Table 5 – Distribution of the clusters using k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*.

Cluster	Type of Distribution	No. of Elements
1	Exponential	82
2	Normal	93
3	Exponential	5
4	Normal	198
5	Uniform	102
6	Exponential	83
7	Exponential	91
8	Exponential	82
9	Normal	90
10	Exponential	84
11	Normal	90

Exponential distributions, but this is not a bad result as we will show soon.

In figures between 12 and 22 we show the *PDFs* of all the distributions that belongs to the same cluster. If we take a look at figures 12, 13, 14, 20, 21 and 22 we see that the exponential distribution was well grouped. Really the problem is that the rate value of $0.05 * i$ used to generate the exponential distribution does not have such a big difference between one and another.

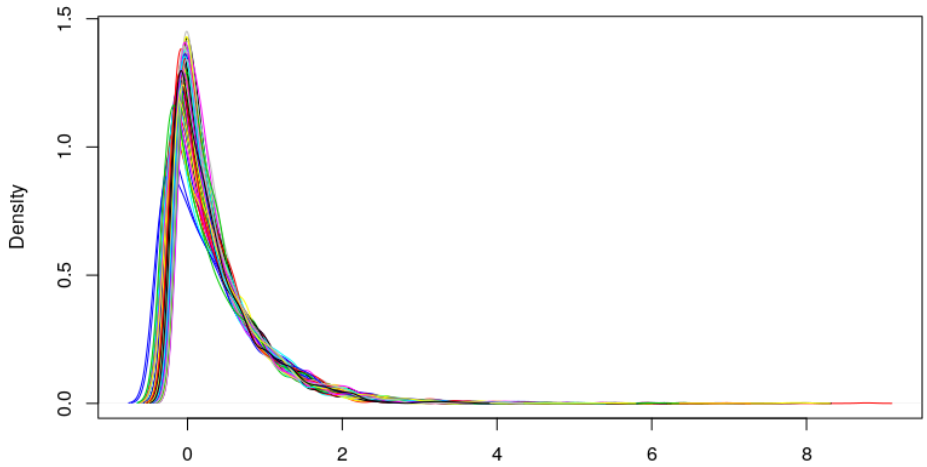

Figure 12 – Cluster 1 returned by the k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*, synthetic dataset I.

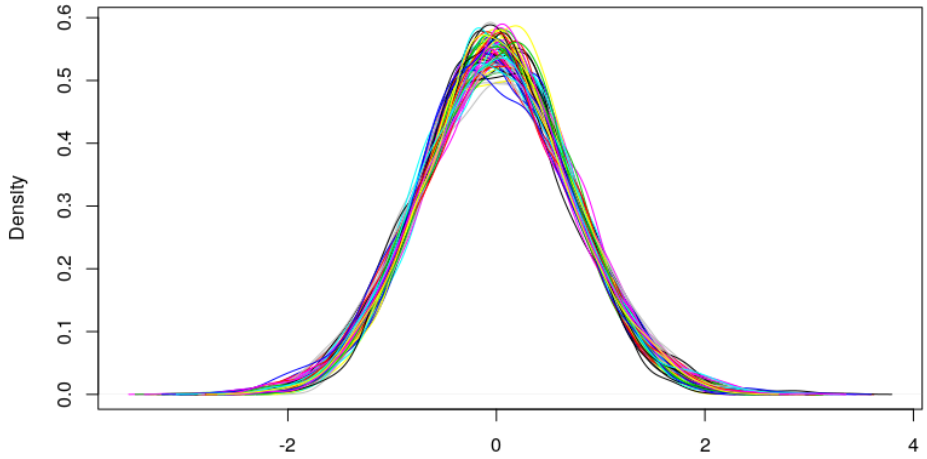

Figure 13 – Cluster 2 returned by the k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*, synthetic dataset I.

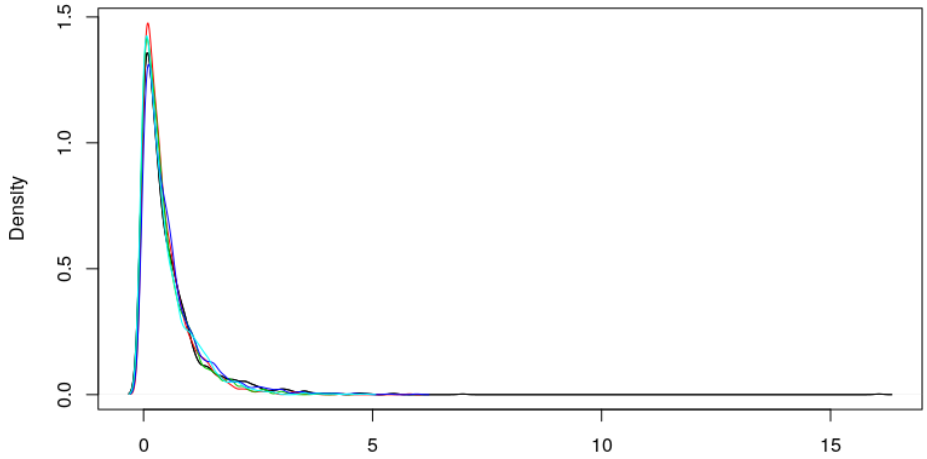

Figure 14 – Cluster 3 returned by the k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*, synthetic dataset I.

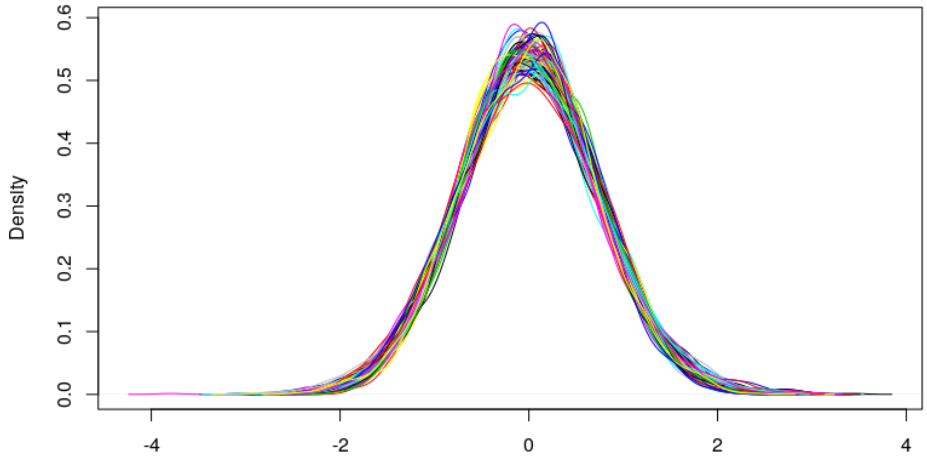

Figure 15 – Cluster 4 returned by the k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*, synthetic dataset I.

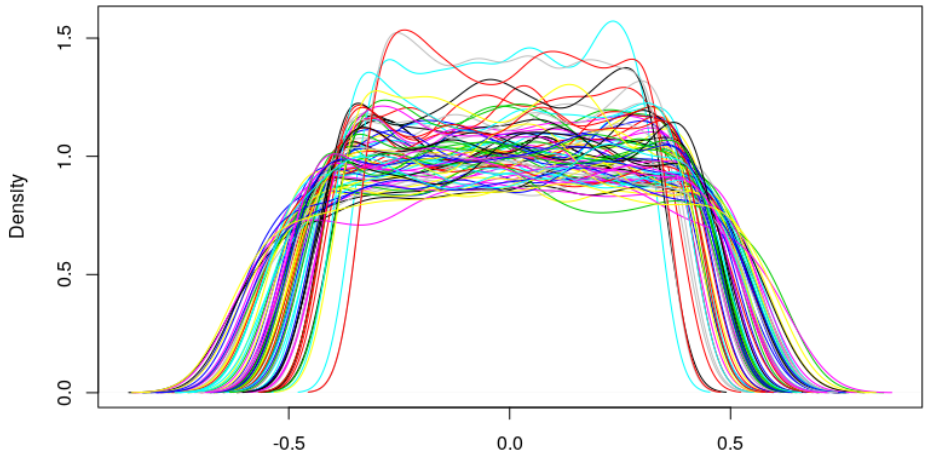

Figure 16 – Cluster 5 returned by the k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*, synthetic dataset I.

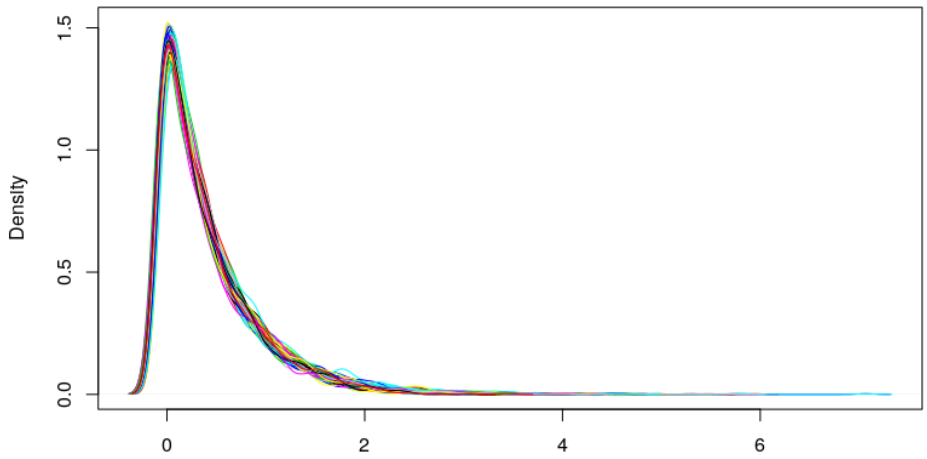

Figure 17 – Cluster 6 returned by the k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*, synthetic dataset I.

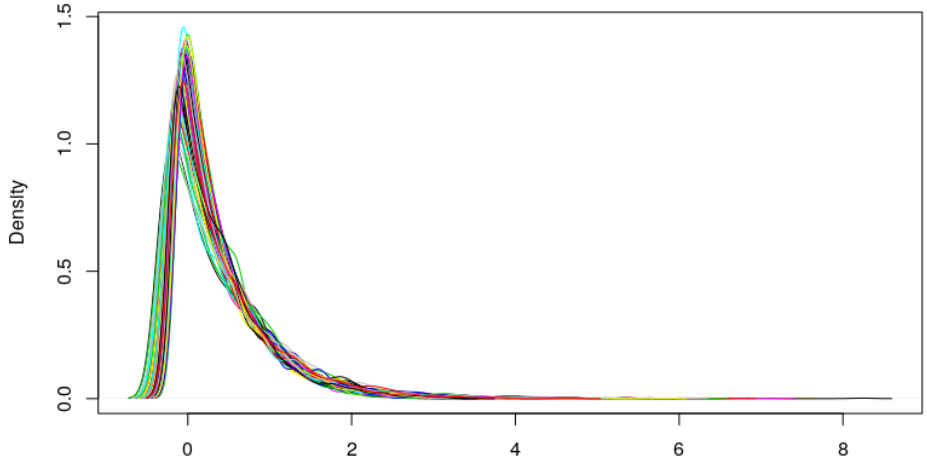

Figure 18 – Cluster 7 returned by the k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*, synthetic dataset I.

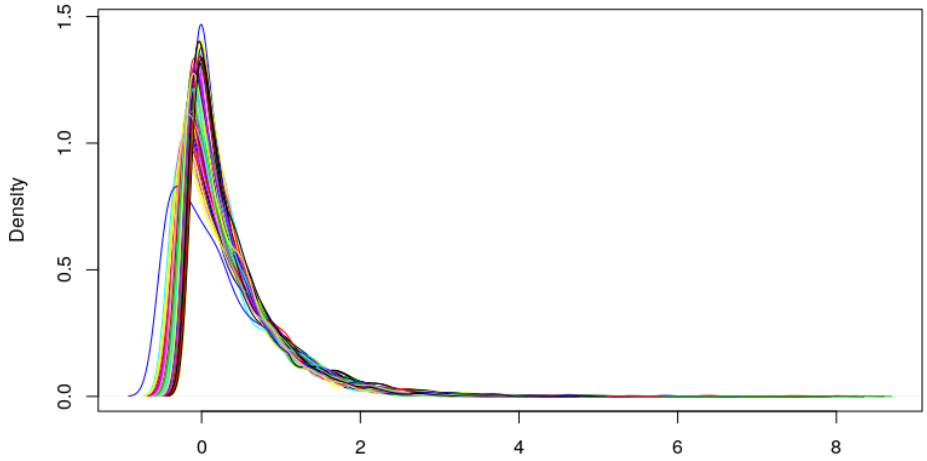

Figure 19 – Cluster 8 returned by the k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*, synthetic dataset I.

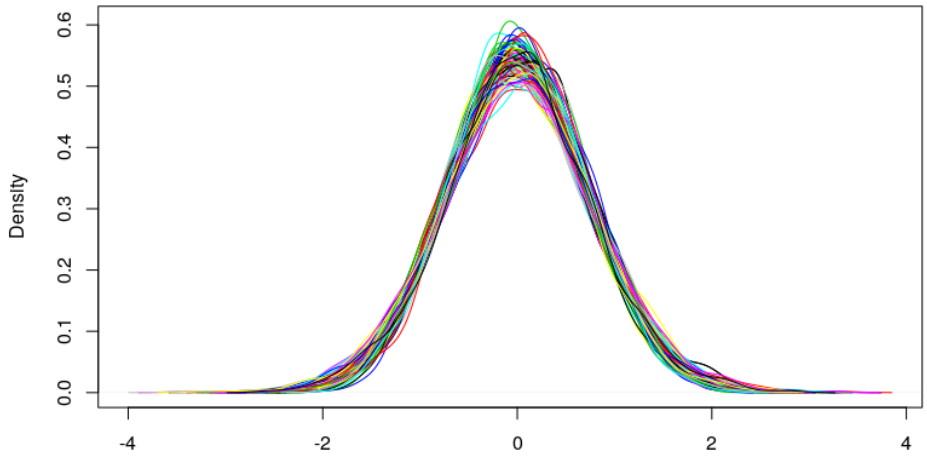

Figure 20 – Cluster 9 returned by the k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*, synthetic dataset I.

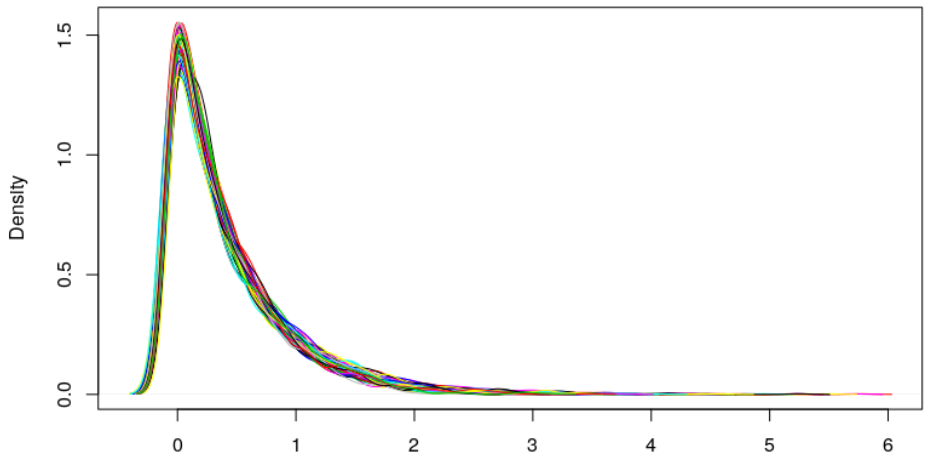

Figure 21 – Cluster 10 returned by the k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*, synthetic dataset I.

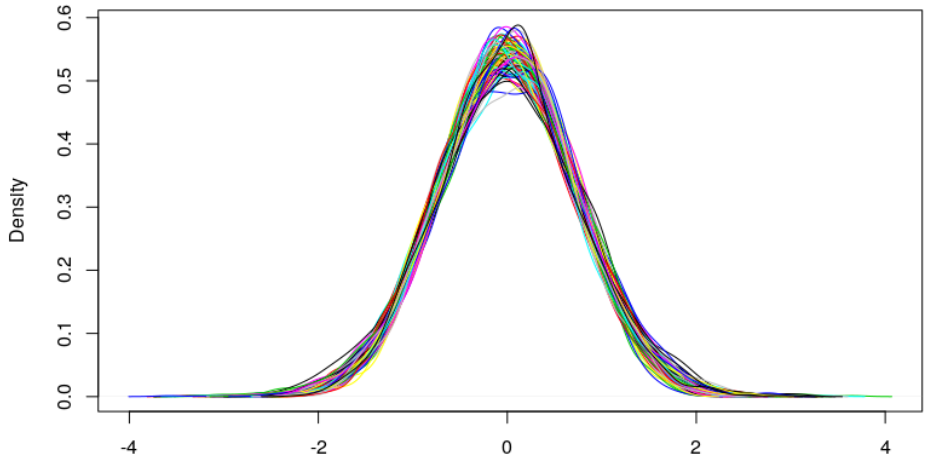

Figure 22 – Cluster 11 returned by the k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*, synthetic dataset I.

Another interesting result is show in figures 23 and 24. As we can see, clusters 2, 4, 9 and 11 that represent the Normal distribution are all at the same region over the λ_3 and λ_4 space, near $\lambda_3 = 0$ and $\lambda_4 \in [0, 0.3]$. Similarly cluster 5, that represent the Uniform distribution is on the top left of the λ_3 and λ_4 space, $\lambda_3 \in [0, 0.3]$ and $\lambda_4 \in [0.7, 1.5]$. And finally the rest of the clusters that represent the Exponential distribution are distributed

in the bottom of the λ_3 and λ_4 space, $\lambda_3 \in [0.2, 7]$ and $\lambda_4 \in [-0.1, 0.1]$. As we see in the rest of the thesis, this result is repeated in all the use cases.

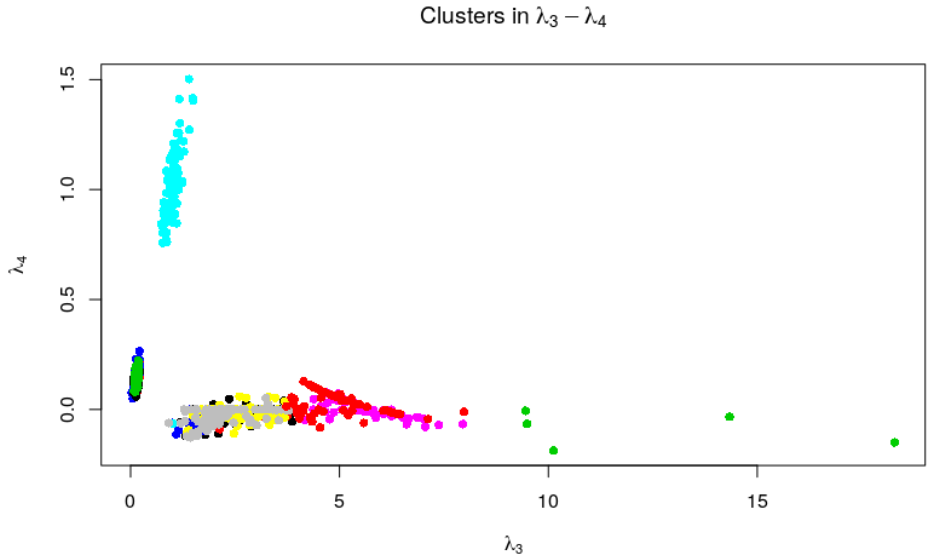

Figure 23 – Distribution of the clusters over the λ_3 and λ_4 space.

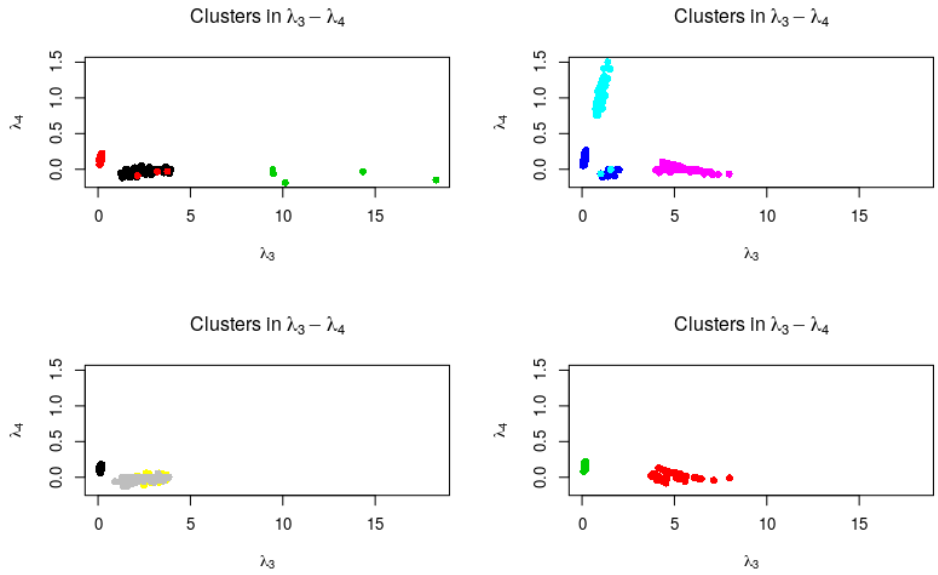

Figure 24 – Distribution of the clusters over the λ_3 and λ_4 space. In the top left corner: clusters 1, 2 and 3. Top right corner: clusters 4, 5 and 6. Bottom left: clusters 7, 8 and 9. Bottom right: clusters 10 and 11.

4.3.2 Clustering using λ_3 and λ_4

In this section we proceed similar to section 4.3.1, but the k-means algorithm run over λ_3 and λ_4. The distribution of the clusters is shown in figure 25 and table 6.

Table 6 – Distribution of the clusters using k-means over the λ_3 and λ_4 values of the *GLDs*.

Cluster	Type of Distribution	No. of Elements
1	Normal	197
2	Exponential	118
3	Exponential	110
4	Uniform	35
5	Exponential	41
6	Uniform	65
7	Exponential	2
8	Normal	131
9	Exponential	74
10	Exponential	105
11	Normal	122

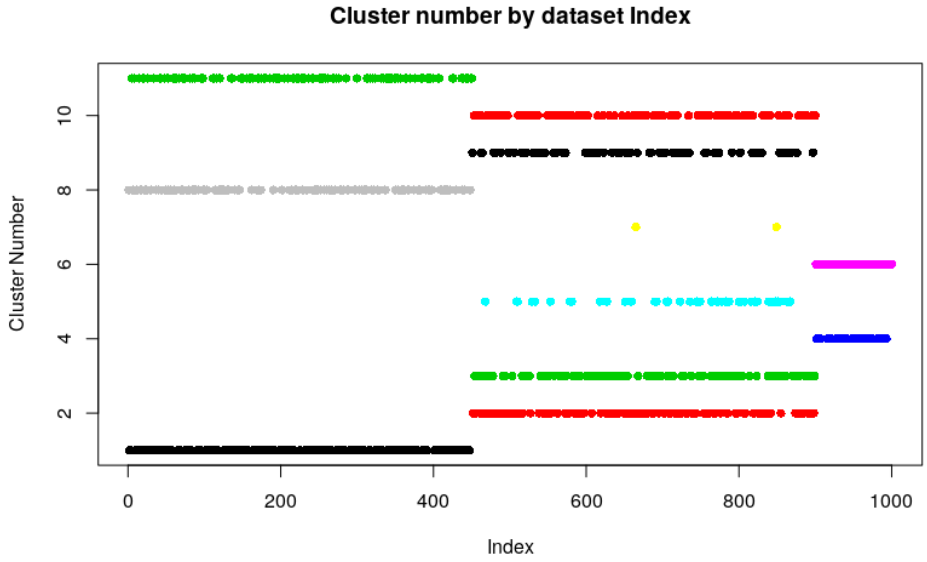

Figure 25 – Distribution of the clusters using k-means over the λ_3 and λ_4 values of the *GLDs*.

As we don't use λ_2 here, is clear that the algorithm can't distinguish the distributions by its standard deviation. But, as the shape of the *GLD* is defined by λ_3 and λ_4, what we expect is that the algorithm can separate the objects by type of distribution. As we see in figure 25 this is exactly what we get, there is no any false positive in this case, the three regions (Normal, Exponential and Uniform) are identified by the k-means.

Clusters 1, 8 and 11 group all the Normal distributions, clusters 4 and 6 group the Uniform and the rest group the Exponential.

In the λ_3 and λ_4 space the behavior is very similar at the one we get in subsection 4.3.1, figures 26 and 27. Again the distributions are concentrated near the same (λ_3, λ_4)

values.

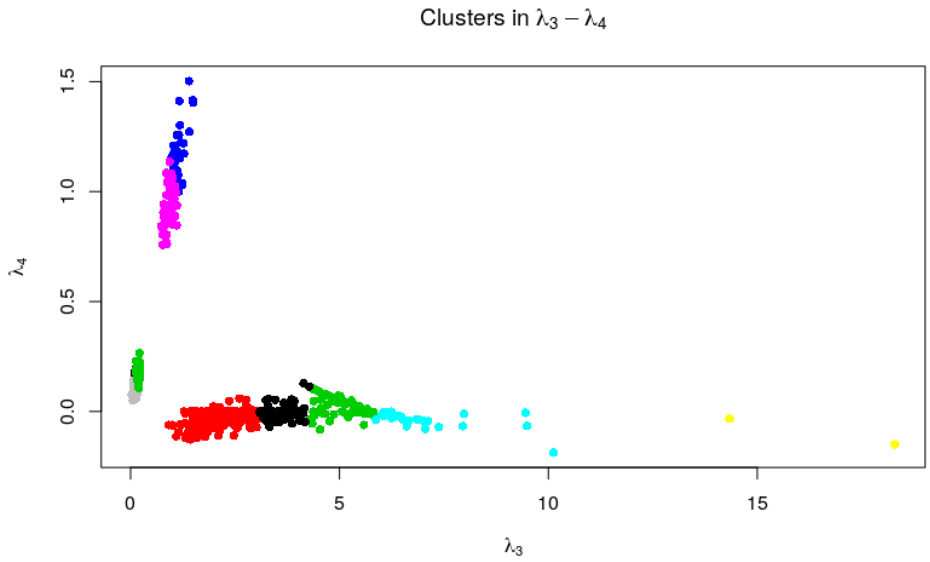

Figure 26 – Distribution of the clusters over the λ_3 and λ_4 space.

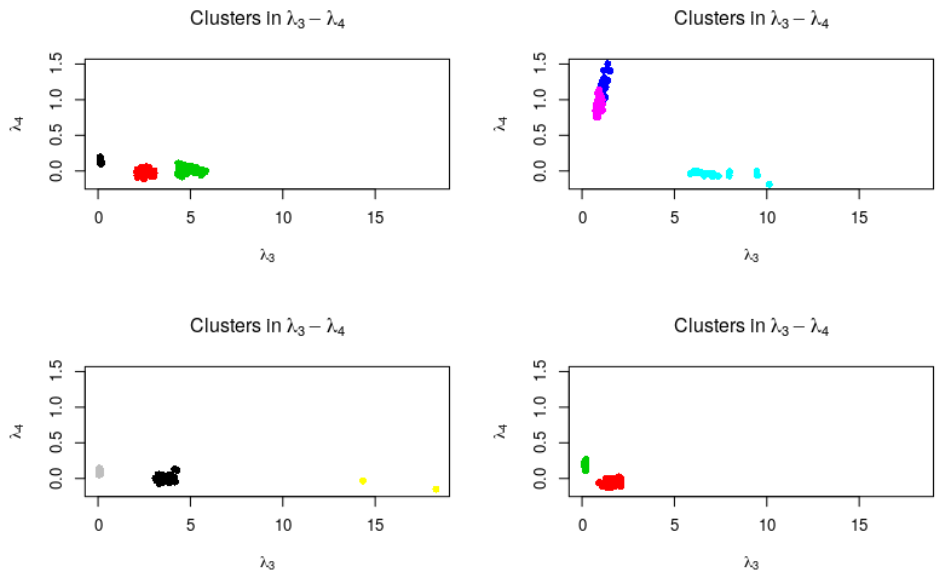

Figure 27 – Distribution of the clusters over the λ_3 and λ_4 space. In the top left corner: clusters 1, 2 and 3. Top right corner: clusters 4, 5 and 6. Bottom left: clusters 7, 8 and 9. Bottom right: clusters 10 and 11.

4.4 Synthetic Data II

The second synthetic dataset is similar to the first one, here we include 5 Gamma distributions, between the Exponential and the Uniform, figure 28. The shape of the

Gamma distribution is i, with $i = 1, 2, 3, 4, 5$. This dataset have 1450 objects, where the first 450 were sampled from a Gaussian distributions, the next 450 from an Exponential, the next 450 are Gamma, and the last 100 from a Uniform distribution. As we use 16 different distributions, this is the number of clusters to be used with the k-means algorithm.

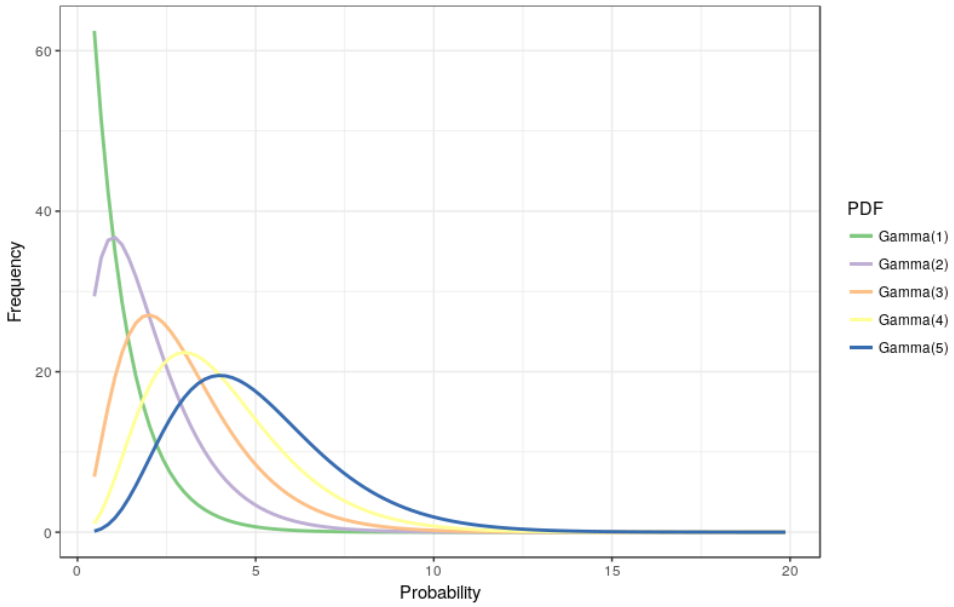

Figure 28 – Gamma distributions used to generate the synthetic dataset.

Similar to the dataset I, the fitting algorithm proposed in subsection 4.2.1 is applied over dataset II. The good-of-fit test return that all the *GLDs* are good fit for its corresponding distribution.

4.4.1 Clustering using λ_2, λ_3 and λ_4

The distribution of the clusters returned by the k-means algorithm is shown in figure 29 and table 7.

Table 7 – Distribution of the clusters using k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*.

Cluster	Type of Distribution	No. of Elements
1	Normal	90
2	Exponential	44
3	Uniform	45
4	Normal	179
5	Gamma	60
6	Uniform	55
7	Exponential	87
8	Exponential	58
9	Exponential	67
10	Exponential	74
11	Exponential	25
12	Exponential	90
13	Normal	96
14	Normal	90
15	Gamma	30
16	Gamma	360

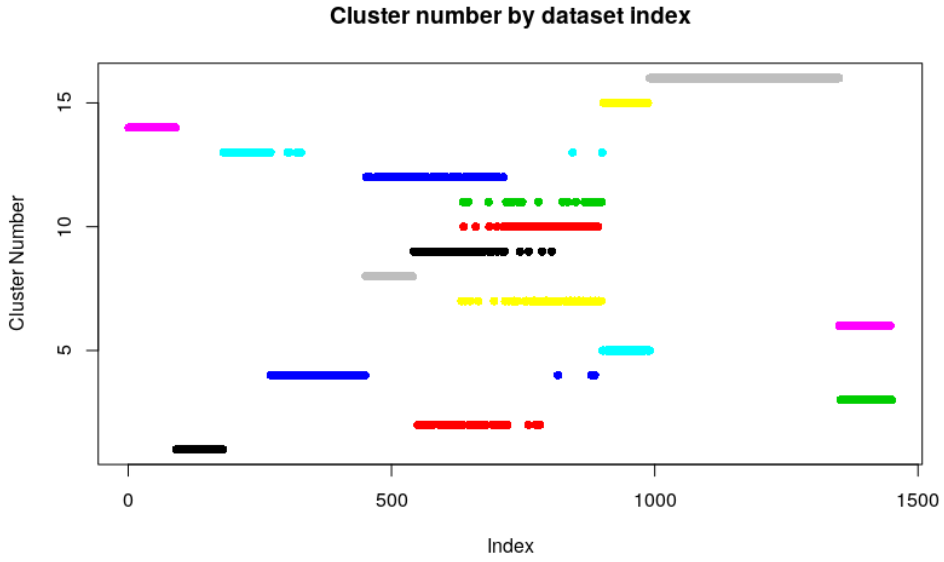

Figure 29 – Distribution of the clusters using k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*.

In general the results are very similar to the results of the section 4.3, but we get less false positives, 5 in total. 3 false positives in cluster 4 and 2 false positves in cluster 13. The normal distribution was groping again in for clusters: 1, 4, 13 and 14. The uniform distribution was grouping in clusters 3 and 6 without false positives. The gamma distribution introduced here was grouped in clusters 5, 15 and 16, without false positives. And finally the rest of the clusters are for the exponential distribution.

The projection of the clusters over the λ_3 and λ_4 space is show in figure 30. The two clusters of the uniform distribution are located again in the top-left region of the figure. The normal distribution is located in the same place, near $\lambda_3 = 0$ and $\lambda_4 \in [0, 0.3]$. The exponential distribution is distributed in the bottom of the λ_3 and λ_4 space. The gamma distribution is overlapped together with the normal distribution.

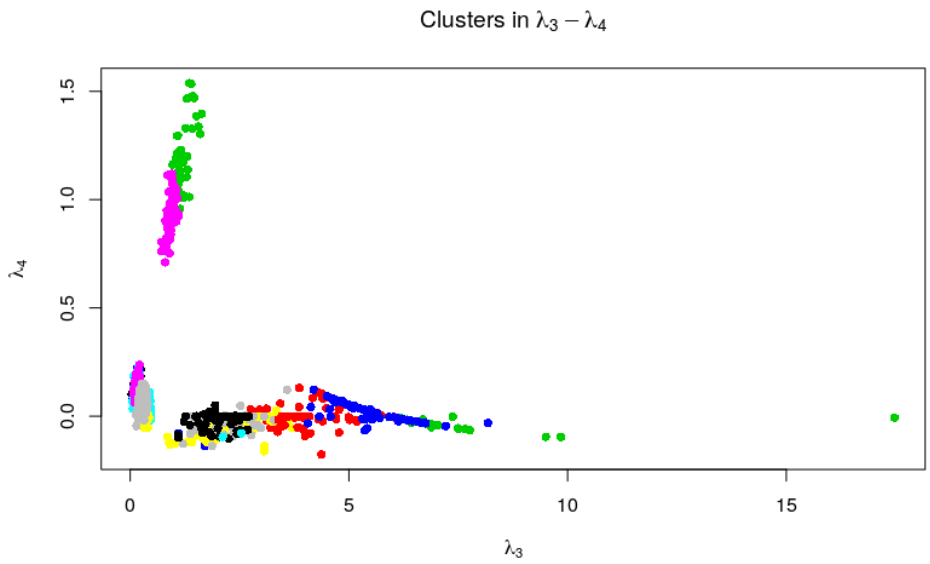

Figure 30 – Distribution of the clusters over the λ_3 and λ_4 space.

4.4.2 Clustering using λ_3 and λ_4

The distribution of the clusters returned by the k-means when using the values of λ_3 and λ_4 to group the second synthetic dataset are shown in figure 31 and table 8.

A few false positives are observed in clusters 5, 6 and 12, but nothing to worry about. Again the regions of the four distribution families are perfectly separated by the algorithm.

Table 8 – Distribution of the clusters using k-means over the λ_3 and λ_4 values of the *GLDs*.

Cluster	Type of Distribution	No. of Elements
1	Exponential	64
2	Exponential	126
3	Exponential	1
4	Exponential	57
5	Gamma	83
6	Normal	139
7	Uniform	67
8	Gamma	148
9	Gamma	108
10	Exponential	75
11	Exponential	80
12	Normal	112
13	Exponential	44
14	Normal	201
15	Gamma	112
16	Uniform	33

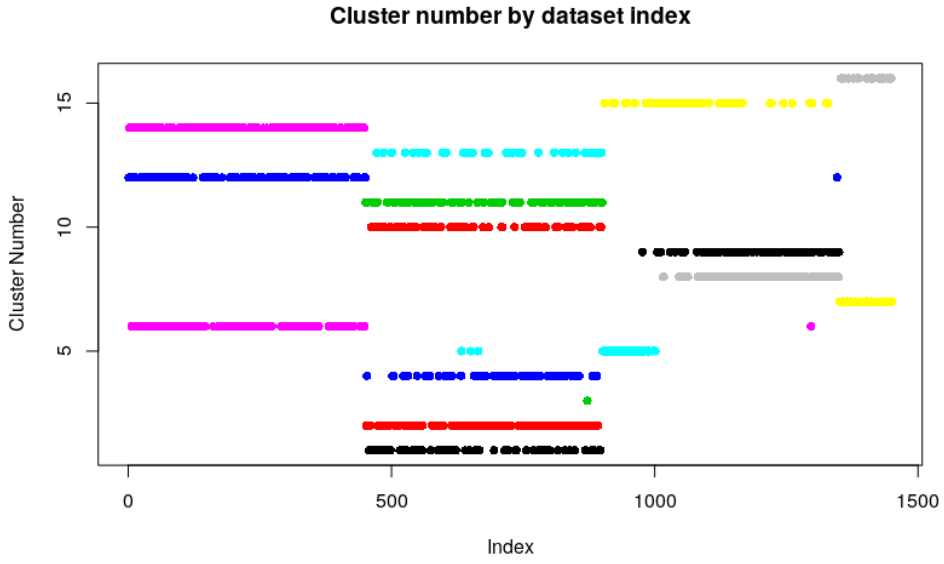

Figure 31 – Distribution of the clusters using k-means over the λ_2, λ_3 and λ_4 values of the *GLDs*.

The projection of the clusters over the λ_3 and λ_4 space is show in figure 32.

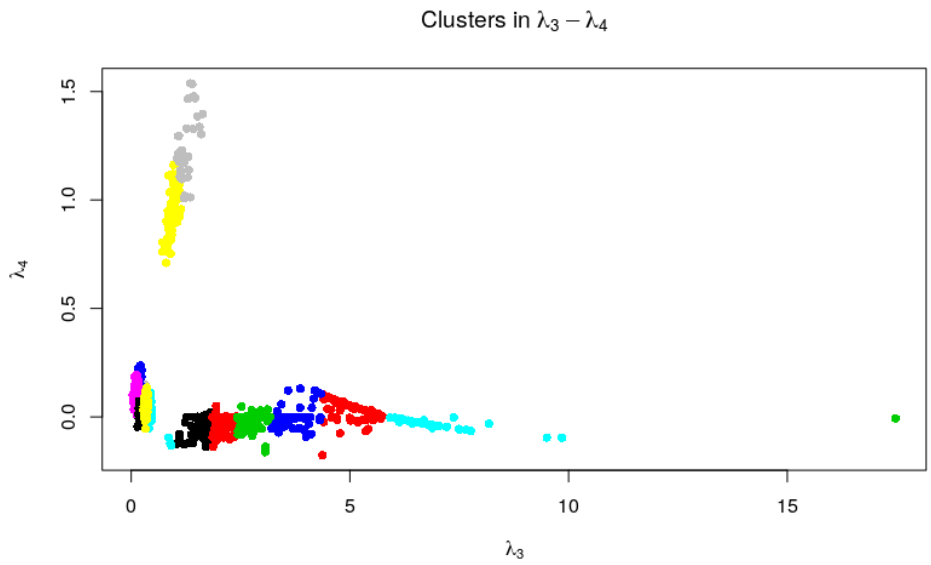

Figure 32 – Distribution of the clusters over the λ_3 and λ_4 space.

4.5 Summary

In this Chapter, we explore clustering uncertain data based on the similarity between their distributions. The idea is to answer the **RQ.1** *"how to group the output of the UQ process based on the similarity of the uncertainty?"*. The hypothesis enunciated at the beginning of the Chapter was tested against two synthetic datasets, and the results corroborate that we can group uncertain data based in the λ_i values of the *GLDs* that describe the data.

As a second result we test that, when using λ_2, λ_3 and λ_4 the separations of the different distributions of the dataset was almost perfect with a few false positives; and when using λ_3 and λ_4 all the elements of the same family are grouping together without consider the difference in the standard deviation. Both results are exactly what we expect.

Another important result of this Chapter, is that if we look at the clustering technique proposed here and compare it with the state-of-the-art, our approach is a competitive one. For example, in the approach proposed by (JIANG et al.,) the computational cost of its algorithm depend of two factors, the fit of the distributions using Kernel Density Estimation (KDE) and the computation of the KL-divergence (the distance measure used in its approach). Both factors are computationally intensive. In our approach we substitute KDE by *GLD* fit, that is most costly form the computational point of view; but at the same time we substitute the KL-divergence by simple distance comparisons in $\mathbf{R}$ and $\mathbf{R^2}$. On the other hand, some limitations of the KL-divergence approach as: (i) the *PDF* of every uncertain object to be clustered need to be defined in the same domain and (ii) the needs to select an appropriate kernel to fit the data using KDE, are solved with the use of

the *GLD*.

All of this observation join with the rest of the advantage of the use of the *GLD* to quantify the uncertainty, make our approach far superior to those reported in the literature.

5 Our Approach

In Chapter 3 we present the *GLD* and argue why this distribution family is suitable to be used in UQ, while in Chapter 4 we explore the possibility to use the λ values of the *GLD* to clustering uncertain data. Now in this Chapter we present a workflow to quantify the uncertainty in large-scale spatio-temporal models using the *GLD*. In Chapter 4 we present a solution to the **RQ.1**, in this Chapter we present a solution to the other four research questions.

The rest of the Chapter is organized as follow: Section 5.1 describes the proposed workflow and comment briefly the motivations to propose it and its steps. Section 5.2 presents the fitting step, that is divided in three sub-steps: the fitting process, *GLD* validity check and quality of the fit. Section 5.3 presents how to use spatio-temporal interpolation over the λ values of the *GLD* to estimate the uncertainty in spatio-temporal locations not previously analyzed (**RQ.2**). Section 5.4 discuses the integration of the clustering algorithm proposed in Chapter 4 into the workflow. Section 5.5 presents how to use the previous results to answer queries that arise in the UQ context, such us those queries we formulate in Chapter 1 (**RQ.3, RQ.4 and RQ.3**). Section 5.6 presents an implementation of the proposed workflow in an R package named Simulation Uncertainty Quantification Querying (SUQ^2). Finally Section 5.7 summarize the Chapter.

5.1 UQ Proposed Dataflow

The proposed workflow to quantify the uncertainty in **LSSTM** based in the *GLD*, is divided in four steps, Figure 33. The first step is the **fitting process**, where we implement algorithms to estimate the parameters of the *GLD* that best fit the dataset on each spatio-temporal location. The second step is the **spatio-temporal interpolation (kriging)**. This step is included because in UQ is common to estimate the uncertainty (e.g. low order statistical moments as a standard deviation) in some points and the interpolate the uncertainty to other points. Analogously we propose to do the same but using the λ values of the *GLD*. The third step is c**lustering uncertain data**, using the algorithm proposed and tested in Chapter 4. And finally the four step is the **queries** step, where we expose how the results we get in the previous steps help us to answer queries that arise in UQ.

In the next sections we detail every step of the process.

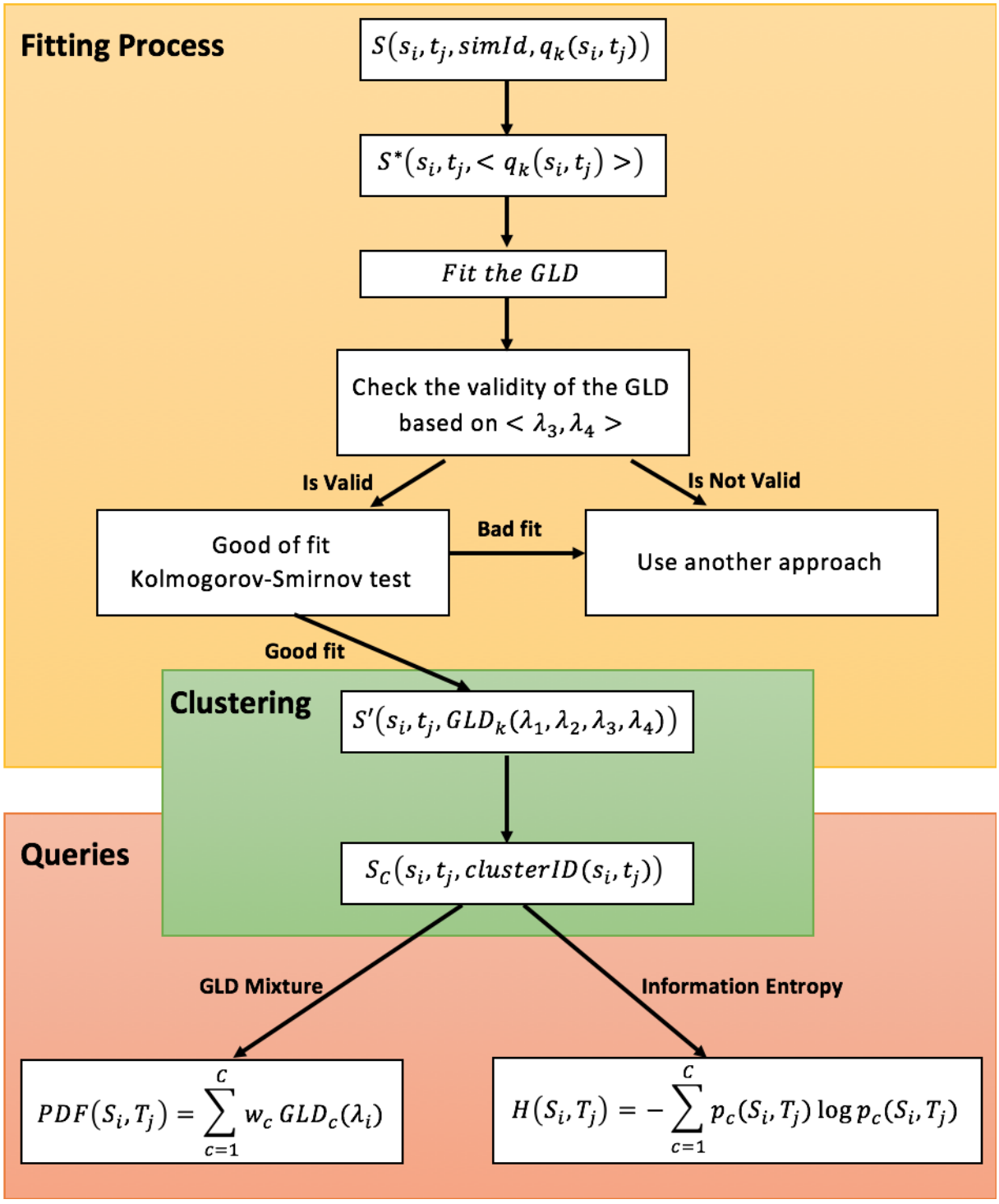

Figure 33 – Proposed workflow. The workflow was divided in three steps, (a) the fitting process, (b) the clustering of the GLDs and, (c) the queries over the results of the clustering process.

5.2 Fitting a GLD to a spatio-temporal dataset

In the more general case the computational model $\boldsymbol{q} = \mathcal{M}(\boldsymbol{\theta})$ represents the spatio-temporal evolution of a complex systems, and the *QoI* $\boldsymbol{q}$ could be represented as:

$$\mathbf{Q} = (\mathbf{q}(s_1, t_1), \mathbf{q}(s_2, t_2), ..., \mathbf{q}(s_n, t_n)) \tag{5.1}$$

where:

- $(s_1, t_1), (s_2, t_2), ..., (s_n, t_n) \in \mathcal{S} \times \mathcal{T} \subseteq \mathbb{R}^3 \times \mathbb{R}$ represent a set of distinct spatio-temporal locations, and
- $\mathbf{q}(s_i, t_j)$ represents a value of the *QoI* at the spatio-temporal location (s_i, t_j)

In the presence of a stochastic problem, on each spatio-temporal location (s_i, t_j) we have many realizations of $q(s_i, t_j)$. A structure of a database to store this information can be modeled as:

$$S(s_i, t_j, simId, q(s_i, t_j)) \tag{5.2}$$

where $simId$ represents the id of one simulation (realization).

The first step of our approach consists in find the *GLD* that best fits our simulations on each spatio-temporal location. This step is divided in three minor tasks:

- Fit the *GLD* to the data.
- Evaluate the validity of the resulting *GLD* on each spatio-temporal location.
- Perform a ks-test to evaluate the quality of the fit on each spatio-temporal location.

The fitting process has been implemented following the Algorithm 3. Before starting the fitting process, we group all the simulations that correspond to the same spatio-temporal location (s_i, t_j). As a result a new dataset with the following structure is created$S^*(s_i, t_j, < q_1, q_2, .., q_n >)$, where $q_i, 1 \leq i \leq n$, represents a vector of all the values of q at point (s_i, t_j).

5.2.1 Fitting process

Now, for each spatio-temporal location $(s_i, t_j) \in \mathcal{S} \times \mathcal{T}$ we use a function of the GLDEX R package described in section 3.8, to fit the *GLD* to a vector $< q_1, q_2, .., q_n >$, line 2 of Algorithm 3. As a result of this task we get the λ values of the *GLD* that best fit the dataset at each spatio-temporal location, Equation 5.3.

$$S'(s_i, t_j, GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)) \tag{5.3}$$

5.2.2 GLD validity check

As we mention in section 3.1 the *GLD* is not always valid, it depends of the λ_3 and λ_4 values. The evaluation of the validity of the *GLD* is straightforward, if λ_3 and λ_4 are in the gray regions of Figure 2 the *GLD* is not valid, on the other case is valid.

The validity check is performed in line 3 of the Algorithm 3, and as a result we get:

$$S_{validity}(s_i, t_j, valid(s_i, t_j)), \tag{5.4}$$

where:

$$valid(s_i, t_j) = \begin{cases} 1 & \text{if GLD is valid in } (s_i, t_j) \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

5.2.3 Quality of the fit

Now at the remaining points, where the *GLD* is valid, we need to evaluate how good is the fit. That is, we evaluate whether the *GLD* (PDF) correctly describes the dataset. We use here the Kolmogorov-Smirnov test (KS-test). The KS-test determines if two datasets differ significantly. In this case, these datasets are: the original dataset and a second one generated using the fitted *GLD*. As a result, this test returns two values: a Kolmogorov-Smirnoff Distance (D); and a p-value, line 5 of Algorithm 3. The distance D is the maximum distance between both cumulative density functions (CDF), as shown in Figure 34. A small distance means that both, the dataset and the fitted PDF, are similar.

The second value, the p-value, is a more robust test, as it helps us to determine the significance of our results. Suppose we have two hypotheses, the null hypothesis is that our PDF is a good fit to our dataset, and the alternative hypothesis is that it is not. Then, a small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis. A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis. p-values very close to the cutoff (0.05) are considered to be marginal (could go either way).

At the end of this task we have two new multidimensional arrays with the values of $\mathcal{D}$ and p-value on each spatio-temporal locations.

$$S_{\mathcal{D}}(s_i, t_j, \mathcal{D}(s_i, t_j)) \tag{5.6}$$

$$S_{p_{value}}(s_i, t_j, p_{value}(s_i, t_j)) \tag{5.7}$$

Finally, in line 7 of Algorithm 3 we store the λ values of those *GLDs* that are valid and return p-values greater than 0.05.

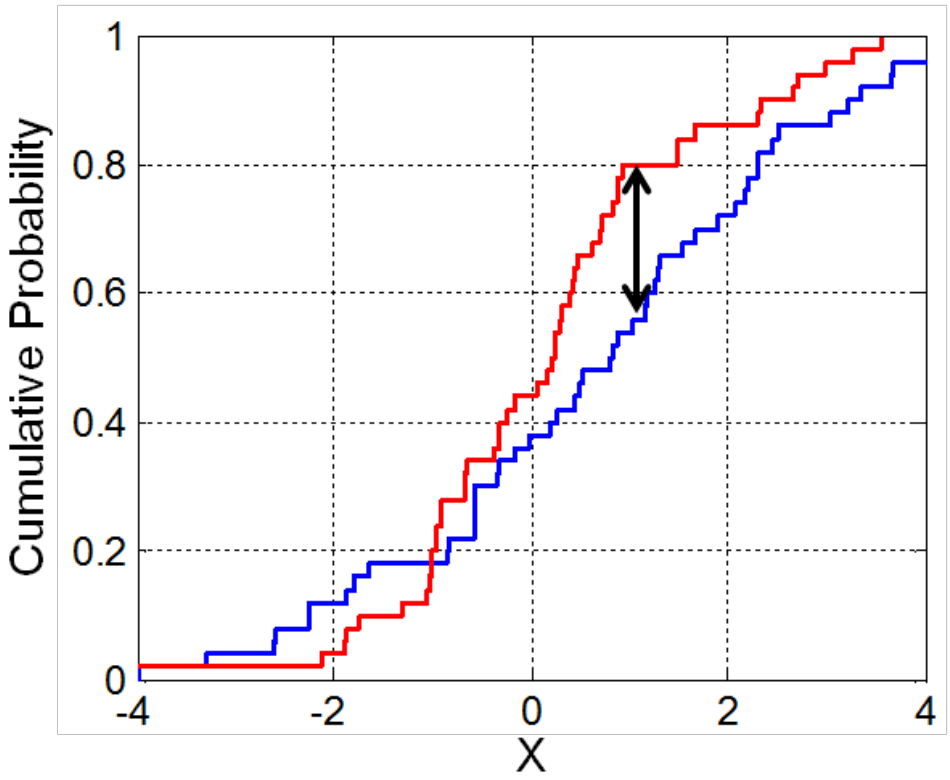

Figure 34 – Illustration of the two-sample Kolmogorov–Smirnov statistic. Red and blue lines each correspond to an empirical distribution function, and the black arrow is the two-sample KS statistic.

Algorithm 3 Fitting the GLD to a spatio-temporal dataset

function GLDFIT($S(s_i, t_j, < q_1, q_2, ..., q_n >)$)
  $< \lambda_1, \lambda_2, \lambda_3, \lambda_4 > \leftarrow$ FIT.GLD.LM($< q_1, q_2, ..., q_n >$)
  $isValid_{(s_i,t_j)} \leftarrow$ VALIDITYCHECK($< \lambda_3, \lambda_4 >$)
  if $isValid_{(s_i,t_j)}$ **then**
    $[pvalue, D]_{(s_i,t_j)} \leftarrow$ KS($< \lambda_1, \lambda_2, \lambda_3, \lambda_4 >_{(s_i,t_j)}$)
  if $pvalue_{(s_i,t_j)} > 0.05$ **then**
    STORELAMBDAS($< \lambda_1, \lambda_2, \lambda_3, \lambda_4 >, s_i, t_j$)

5.3 Spatio-Temporal Interpolation

Although our interest is to quantify the uncertainty on each spatio-temporal locations, this is a timely and computationally prohibitive task. Neither the **GLD** or simples approach, as the evaluation of low order statistical moments, can be computed over the whole output space. Usually researchers put control points at particular points of interest, and then interpolate the uncertainty to other points as needed.

Spatial or temporal interpolation independently, are very well studies, and dozens of algorithms exist to compute new values by mean of those methods. However, working with spatio-temporal domain implies that variability in space and time must be modelled,

which is more complicated than modelling purely spatial or purely temporal variability (GRALER; PEBESMA; HEUVELINK, 2016).

Our purpose here is not to provide a new interpolation method over the **GLDs** λ values, but show how spatio-temporal interpolation can be used to answer the **RQ.2**:

> **RQ2.** what is the uncertainty in some spatio-temporal locations not previously analyzed?

For this reason we just select the, from the best of our knowledge, state-of-the-art spatio-temporal interpolation method proposed by Graler et al. in (GRALER; PEBESMA; HEUVELINK, 2016), and implemented in the R package **gstat**. The selection obey the fact that this implementation include time as another dimension, and allow us to interpolate in space and time together.

5.3.1 Kriging over GLD

The workflow of spatio-temporal interpolation using **gstat** have three steps: (i) variogram, (ii) fitting, and (iii) interpolation. The implementation of this workflow over the **GLD** using **gstat** is shown in Algorithm 4.

Algorithm 4 Spatio-temporal interpolation over the $\lambda_{(2,3,4)}$ values of the GLD.

1: **function** GLDKRIGING($S(s_i, t_j, < 0, \lambda_2, \lambda_3, \lambda_4 >)$)
2: $gldModel \leftarrow$ VGMST($S(s_i, t_j, < 0, \lambda_2, \lambda_3, \lambda_4 >)$)
3: $fitGldVariogram \leftarrow$ FIT.STVARIOGRAM($gldModel$)
4: $S''(s_i, t_j) \leftarrow$ KRIGEST($fitGldVariogram, (\mathcal{S}_i, \mathcal{T}_j)$)

As a result of this algorithm a new dataset with the interpolated values of the **GLD** in the spatio-temporal locations not previously analyzed is returned, Equation 5.8.

$$S''(s_i, t_j, GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)) \tag{5.8}$$

5.4 Clustering

In Chapter 4 we discussed how to use a *GLD* to group uncertain data. The Algorithm 5 proposed there is using here to clustering the *GLDs*.

Algorithm 5 receive as an input the result of the previous steps, could be a result of the fitting process, Equation 5.3 or the result of the kriging process Equation 5.8. The output of the algorithm is:

$$S_{\mathcal{C}}(s_i, t_j, GLD_k, clusterID) \tag{5.9}$$

Algorithm 5 Clustering the GLD based on its $\lambda_{(2,3,4)}$ values.

1: **function** GLDCLUSTERING($S(s_i, t_j, < 0, \lambda_2, \lambda_3, \lambda_4 >)$)
2: $S(s_i, t_j, clusterID_I) \leftarrow$ FIRSTCLUSTERSTEP($S(s_i, t_j, \lambda_2)$)
3: **for each** $clId_I$ **do**
4: $S(s_i, t_j, clusterID_{II}) \leftarrow$ SECONDCLUSTERSTEP($S(s_i, t_j, < \lambda_3, \lambda_4 >), S(s_i, t_j, clusterID_I)$)

where: $clusterID$ represents the ID of the cluster to which the *GLD* at the spatio-temporal location (s_i, t_j) belongs.

With the *GLD* clusterized, we can use this result to characterize the uncertainty in a particular spatio-temporal region, or to measure numerically the corresponding uncertainty. In subsections 5.5.1 and 5.5.2, we describe how those approaches are implemented (see Figure 33).

5.5 Queries

In this section we present how questions as:

RQ3. what is the uncertainty of an specific spatio-temporal region?
RQ4. how to compare two regions as a function of its uncertainty?
RQ5. what is the less uncertain model from a set of models?

can be answered with the use of the *GLD* and the results of the previous steps of the workflow.

5.5.1 Use of GLD mixture to characterize the uncertainty in an spatio-temporal region

Lets start with **RQ.3**. The naive algorithm to answer this question could be something as follow: given a spatio-temporal region $(\mathcal{S}_i \times \mathcal{T}_j)$, read and analyze all the data generated during the simulation process on that region. This is a process current used by many researchers and proposed for example in provenance softwares. Is clear that this approach is not so efficient. Now, in our approach we first find the *GLD* that best fit the dataset on each (s_i, t_j), then we test if the fit is a good one. As a *GLD* is proved as a good random variate generator (see Section 3.6), then we can substitute the raw data by the *GLD*. In the clustering step of our approach we group all the *GLDs* by its similarities, and then we test if the centroid of each cluster represent statistically the other members of each cluster. If this condition is met, then is possible to substitute each *GLD* by the centroid of the cluster is belongs to. At the end of this process we substitute the raw data by a few centroids of the clusters.

Now, coming back to the question: *"what is the uncertainty of an specific spatio-temporal region* $(\mathcal{S}_i \times \mathcal{T}_j)$*?"*, we can answer it looking to the centroids of the clusters.

In $(\mathcal{S}_i \times \mathcal{T}_j)$ each cluster may be qualified with a weight given by:

$$w_k = \frac{1}{N} \sum_{i=1}^{S} \sum_{j=1}^{T} w(s_i, t_j) \tag{5.10}$$

where:

$$w(s_i, t_j) = \begin{cases} 1 & \text{if } clusterID(s_i, t_j) = k \\ 0 & \text{otherwise} \end{cases} \tag{5.11}$$

and N is the number of points in the region $(\mathcal{S}_i \times \mathcal{T}_j)$.

The weight w_k is the frequentist probability of occurrence of the cluster k in the region, and complies with the conditions outlined in section 3.5 that $w_k \geq 0$ and $\sum w_k = 1$.

Remember that the mixture of the *GLDs* can be written as:

$$f(x) = \sum_{k=1}^{K} w_k GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4) \tag{5.12}$$

So, if we have the weights and a representative *GLD* for each cluster, we have the mixture of *GLD* that characterizes the uncertainty in the spatio-temporal region $(\mathcal{S}_i \times \mathcal{T}_j)$.

The process to generate the mixture of *GLDs* that characterize the uncertainty on a region $(\mathcal{S}_i \times \mathcal{T}_j)$ is presented in Algorithm 6.

Algorithm 6 GLD mixture in a region $(\mathcal{S}_i \times \mathcal{T}_j)$

```
1: function GLDMIXTURE(S_i × T_j, C_{S_i × T_j})
2:     for each p_i in (S_i × T_j) do
3:         c ← cluster(p_i)
4:         w_c = w_c + 1
5:         N = N + 1
6:     end for
7:     return (1/N) Σ_c^{C_{(S_i × T_j)}} w_c * c.getGLD()
```

5.5.2 Information Entropy as a measure of the uncertainty in an spatio-temporal region

Another way to answer question **RQ.3** is using the Information Entropy (see Section 2.2.3.2). In that section we highlight a limitation related to the fact that we need to know the possible outcomes of the system to use Information Entropy. In cases where we can use the clusters we get in 5.4 as the different outcomes of the system, Information Entropy could be used as a measure of the uncertainty in an spatio-temporal region.

The equation 2.3 can be rewrite as follow:

$$H(s,t) = -\sum_{c=1}^{C} p_c(s,t) \log p_c(s,t) \tag{5.13}$$

where c represent a particular cluster of the total number of clusters C, and $p_c(s,t)$ represent the probability of occurence of the cluster c in the spatio-temporal region (s,t).

Algorithm 7 computes the Information Entropy in a region $C_{(\mathcal{S}_i \times \mathcal{T}_j)}$. In lines 2 to 7, we compute the probability of each cluster in the region, similar to section 5.5.1. Using this probability we compute the Information Entropy $H(s,t)$, line 8, and finally we return the result in line 9.

Algorithm 7 Information Entropy in a region $(\mathcal{S}_i \times \mathcal{T}_j)$

```
1: function GLDMIXTURE(S_i × T_j, C_{S_i × T_j})
2:     for each p_i in (S_i × T_j) do
3:         c ← cluster(p_i)
4:         w_c = w_c + 1
5:         N = N + 1
6:     end for
7:     p_c(s,t) = w_c / N
8:     H(s,t) ← − Σ_{c=1}^{C} p_c(s,t) log p_c(s,t)
9:     return H(s,t)
```

The advantage of the Information Entropy is that synthesize the uncertainty in a single number.

5.5.3 Information Entropy and regions comparison

Information Entropy can be used to answer other questions as: *"**RQ4.** how to compare two regions as a function of its uncertainty?"*. Its application here is straightforward, as was mentioned in Section 2.2.3, the Information Entropy is zero when we are certain and maximum when the uncertainty is maxims too. Then, if we want to compare two regions we compute the $H(s,t)$ of each one, and that region with the smaller value of $H(s,t)$ is the less uncertain one.

5.5.4 Information Entropy and model selection

Similar to the previous section, we can use Information Entropy to compare two models. Suppose we have a set of models $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_n\}$ and we want to know what is the less uncertain in $(\mathcal{S}_i \times \mathcal{T}_j)$. We proceed to compute $H(s,t,\mathcal{M}_i)$ for each model, and that one with less value of $H(s,t,\mathcal{M}_i)$ is the less uncertain in that region.

5.5.5 Other queries

Although in this section we present four different queries to answer three different questions (**RQ.3**, **RQ.4** and **RQ.5**), our approach leave open the option to develop new queries to solve questions that arise in UQ context. In Section xxx we show how new queries can be answered quickly, thanks to the flexibility of this approach.

5.6 SUQ^2 R package

The proposed approach was implemented as an R package named SUQ^2, an acronym of **S**imulation **U**ncertainty **Q**uantification **Q**uerying, freely available at SUQ^2. As this package is in a development stage, to install it you need to install the R **devtools** package first and then use a function **install_github** to install SUQ^2.

```
> install.packages("devtools")
> install_github("nmlemus/suq2")
```

The package was divide into five sub-package, one for each step or the workflow and other one to show the results. The name of the sub-package are: **fit**, **kriging**, **clustering**, **queries** and **plot**.

As R is an interactive language, to warranty the easy to use of the functions inside the package we use the following pattern: (i) the name of all the function of the package start with **suq2.**; (ii) after **suq2.** we add a name that identify a sub-package, for example the name of all the functions of the sub-package **plot** start with **suq2.plot.**; (iii) and finally a name that identify the functionality of the function. For example a function to plot a gld based on its λ values is named **suq2.plot.gld()**, while a function to clustering the *GLDs* based on λ_2 is **suq2.clustering.lambda2()**.

The full documentation of the package is available inside the installation and in Anexo A.

5.7 Summary

In this Chapter we present the workflow of the proposed approach. The workflow is divided into four steps, fitting, kriging, clustering and queries. The algorithms of all the steps are presented and commented. In Section 5.5 the queries to answer the research questions raised in the introduction were discuss. Finally, in Section 5.6 we present SUQ^2, an R package that implements our approach.

6 Use Cases

In the present chapter we are going to test the UQMS in three different scenarios, spatial only domain, section 6.1 , spatio-temporal domain, section 6.3, and finally a multidisciplinary system, section 6.4.

6.1 Case Study: Wave Propagation Problem

6.1.1 The Dataset

In the HPC4e benchmark, the models have been designed as a set of 16 layers with constant physical properties. The top layer delineates the topography and the other 15 delineate different layer interface surfaces or horizons. To generate a single cube with dimensions $250 \times 501 \times 501$ we can use the values provided in the benchmark. For example, to generate a cube in the $v_p(m/s)$ variable we can use the fixed values of Table 9.

The first slice of this cube is shown in Figure 35.

Layer	$v_p(m/s)$
1	1618.92
2	1684.08
3	1994.35
4	2209.71
5	2305.55
6	2360.95
7	2381.95
8	2223.41
9	2712.06
10	2532.22
11	2841.03
12	3169.31
13	3252.35
14	3642.28
15	3659.22
16	4000.00

Table 9 – Values of v_p used in the generation of a single velocity field cube.

Layer	**PDF Family**	**Parameters**
1	Gaussian	[1619, 711.2]
2	Gaussian	[3368, 711.2]
3	Gaussian	[8839, 711.2]
4	Gaussian	[7698, 301.5]
5	Lognormal	[7723, 294.7]
6	Lognormal	[7733, 292.2]
7	Lognormal	[7658, 312.1]
8	Lognormal	[3687, 368.7]
9	Exponential	[3949, 394.9]
10	Exponential	[5983, 711.2]
11	Exponential	[3520, 352.0]
12	Exponential	[3155, 315.5]
13	Uniform	[2541, 396.4]
14	Uniform	[2931, 435.3]
15	Uniform	[2948, 437.0]
16	Uniform	[3289, 471.1]

Table 10 – PDFs and its parameteres used to sampling the v_p, to generate n velocity models.

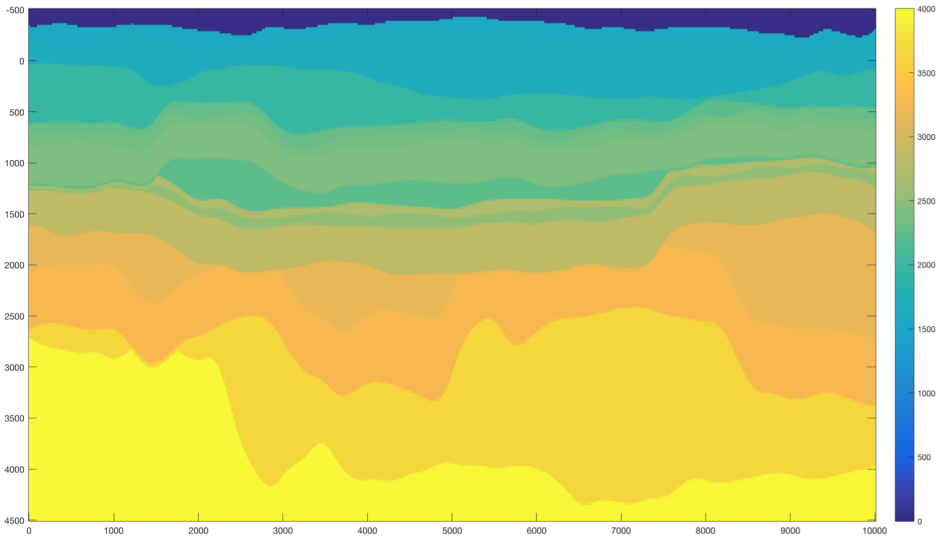

Figure 35 – One slice of the $250 \times 501 \times 501$ cube. In the slice we can distinguish between the different layers.

Now as our purpose is to study the uncertainty in the output as a result of the propagation of the input uncertainty throughout the model, we cannot use this benchmark as it is. We need the input, $v_p(m/s)$ in this case, to be uncertain. In order to achieve so, we compute $v_p(m/s)$ as a random variable with the *PDFs* shown in Table 10.

Then, using a Monte Carlo method we generate a sampling of 1000 realizations of the $v_p(m/s)$ variable, Figure 36; and using a Matlab script provided by the HPC4e

benchmark we simulate 1000 times, one for each realization, and generate 1000 cubes (230 GB) as an output. The resulting cubes are $250 \times 501 \times 501$ multi-dimensional arrays.

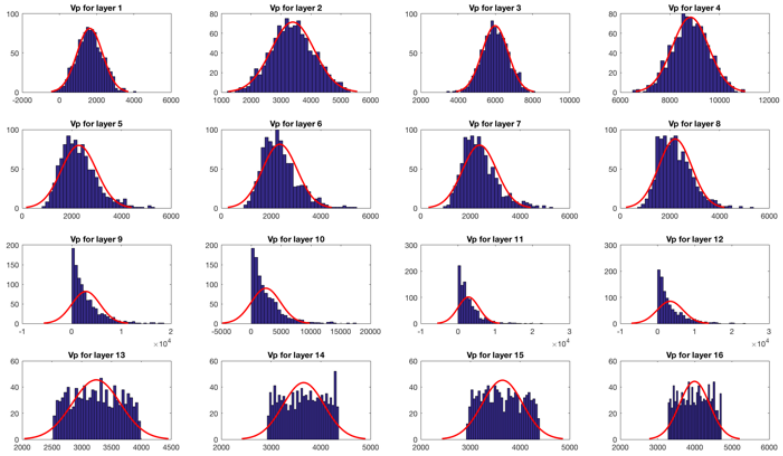

Figure 36 – Histograms of the 1000 samplings generated using Monte Carlo method and the PDFs reported in Table 10.

To simplify the computational process and visualize the results, we select the slice 200 to be used here, then we have 1000 realizations of a slice with size of 250×501. The equation 5.2 can be simplified because we have two dimensions in space and don't have time domain, then our dataset can be represented as $S(x_i, y_j, simId, v_p(x_i, y_j))$. In this new representation (x_i, y_j) are the 2D coordinates and $v_p(x_i, y_j)$ is the velocity value at point (x_i, y_j). *simId* still represents the Id of the simulation and its range here is between 1 and 1000.

Now that we have an experimental dataset we can start to apply our workflow, step by step.

6.1.2 Fitting the GLD

The first step is to find the *GLD* that best fits the dataset at each spatial location. Running the algorithm proposed in Section 5.2.1 we get as a result a new 2D array:

$$S'(x_i, y_j, GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)) \tag{6.1}$$

The raw data is reduced and our dataset is characterized by four lambda values at each spatial location. Now we need to assess the validity of the *GLDs* and how well they fit the dataset. Those analyses are described in sections 6.1.3 and 6.1.4.

6.1.3 GLD validity check

Once the algorithm to check the validity of the *GLD* is run on the experimental dataset, we obtain as a result that the *GLD* is valid in all the (x_i, y_j) space.

6.1.4 Quality of the fit

The next step is to check how good is the fit. To do this we use an algorithm that returns the D and *p-value* for the KS-test at each spatial location. As we show in figure 37, and remember that with a *p-value* > 0.05 we cannot reject the null hypothesis, we conclude that the fit of the GLD is acceptable in most cases. To be more exact, the p-value was greater than 0.05 in 82 % of the spatial locations, figure 38.

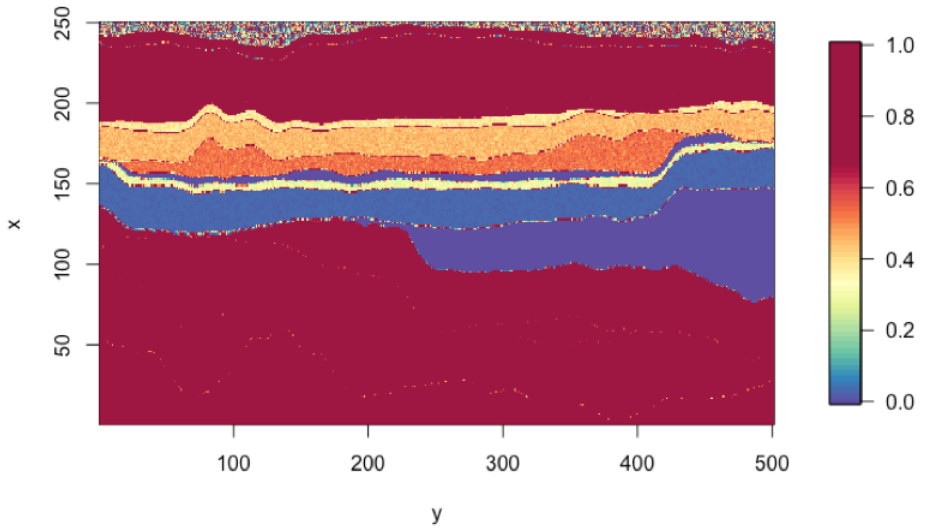

Figure 37 – Goodness of the fit based on the p-value returning by the KS-test. p-value $>$ 0.05 represent a good fit of the GLD to the dataset at (x_i, y_j).

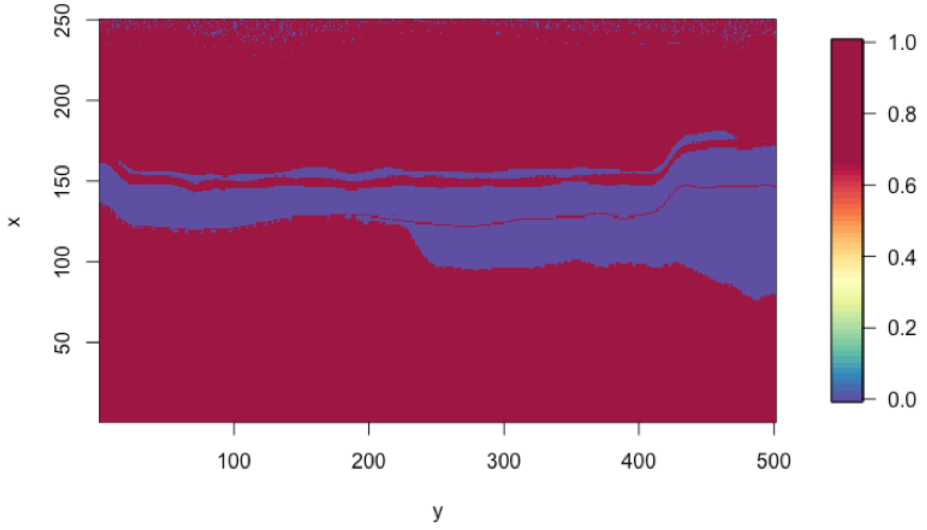

Figure 38 – The red color shows where the p-value was greater than 0.05.

If we consider the distance D, returned by the KS-test, the result is similar, figure 39. We can see a blue region that is common in figures 38 and 39. This region is where

the quality of the GLD fit is below a threshold. On those cases, some *GLD* extensions proposed in (KARIAN; DUDEWICZ, 2011) could be used.

As the main purpose of this paper is to demonstrate the utility of the use of the *GLD* in *UQ*, then we are not going to deep in other algorithms to solve this particular problem.

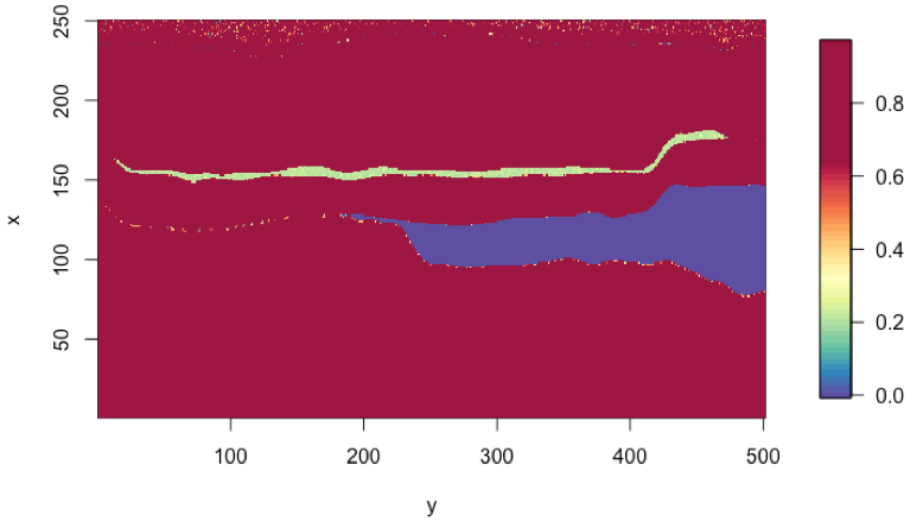

Figure 39 – Kolmogorov-Smirnoff Distance (D). The red regions represent where the GLD fits well.

6.1.5 Clustering

At this point we have our dataset characterized by the schema depicted by Equation 6.1, then using a clustering algorithm, such as k-means, we are going to group the GLDs based on its $(\lambda_2, \lambda_3, \lambda_4)$ values, as those are the values that describe the shape of the distribution at each point of the dataset.

In this paper we use k-means algorithm with $n = 10$, where n is the number of clusters to be made. This is an arbitrary value, we are investigating other algorithms as DBSCAN and what are the ϵ of this algorithm that warranty a good clusterization, but discussing alternative GLDs clustering algorithms is beyond the scope of this paper.

Once the clustering algorithm has been applied, a new dataset is produced, where for each spatial location we have a label that indicates the cluster the GLD at each position belongs to (see the schema at Equation 6.2), Figure 40. Note that, in Figure 40, the blue region corresponding to cluster 11 is not a cluster itself. It is rather the region where the *GLD* is not valid, see section 6.1.4.

$$S_{\mathcal{C}}(x_i, y_j, clusterID, GLD_{x_i,y_j}) \tag{6.2}$$

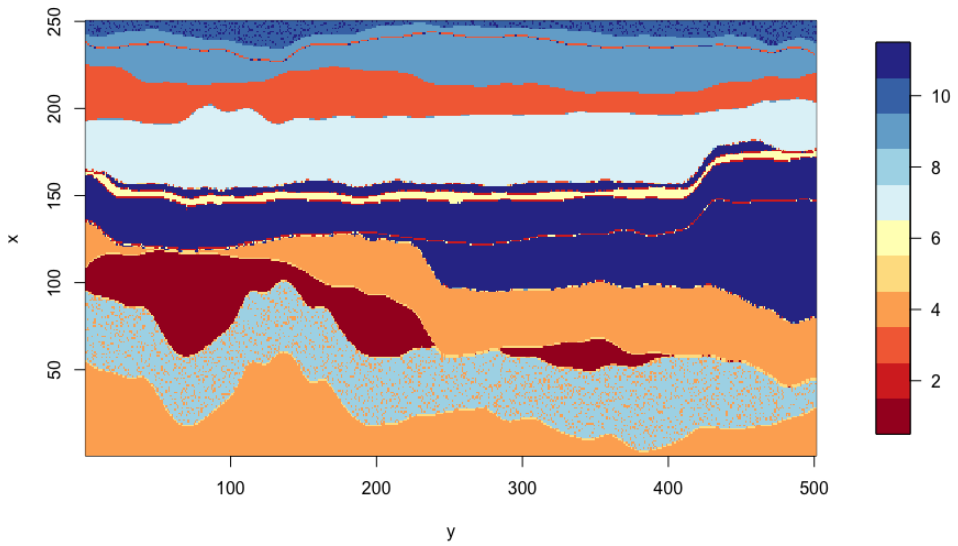

Figure 40 – Result of the clusterization using k-means with $n = 10$.

If we visually compare Figure 35 with Figure 40, we observe a close similarity between the two. It is clear that they can not be equal because we are talking about a slice of a deterministic model, and the result of making clusters on 1000 realizations of a stochastic model, but as the model used here is very linear, this is the result we expect.

Another interesting result is shown in Figure 41, where we plot the clusters in (λ_3, λ_4) space. As we mention in section **??**, the shape of the *GLD* depends on the values of λ_3 and λ_4. In this scenario, the expected result is that the members of the same cluster share similar values of λ_3 and λ_4. This is exactly the result we can observe in Figure 41.

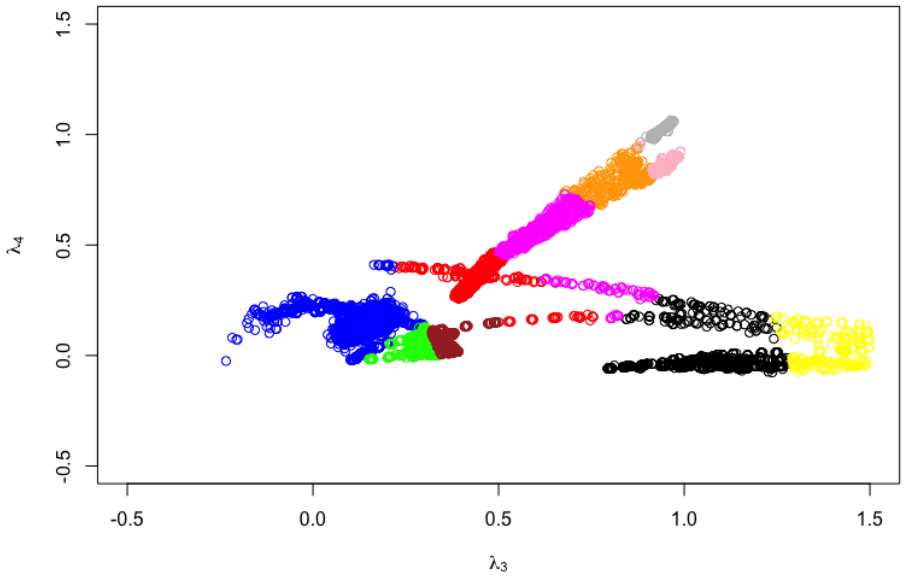

Figure 41 – Distribution of the clusters in the (λ_3, λ_4) space. The points that belongs to a same cluster are one near the others, as was expected.

To further corroborate this fact, in Figure 42 we show the *PDFs* of 60 members of the 10 clusters. Visually assessing the figures we have an idea of how similar are the shapes of the members of a same cluster and how dissimilar are the shapes of the members of different clusters. This suggests that our approach is valid. A product of these observations is that we can pick one member of each cluster (the centroid) as a representative of all the members of this cluster, Table 11. The selected member is going to be used to answer the queries in the next sections.

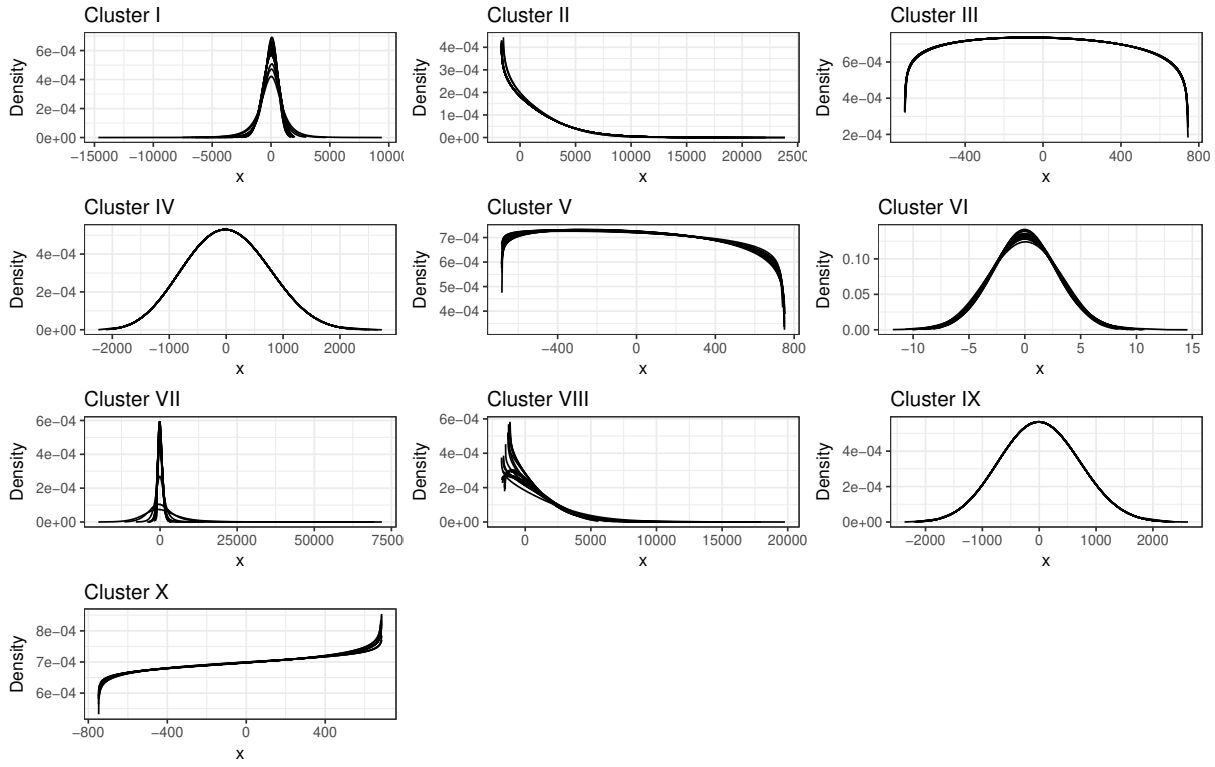

Figure 42 – *PDFs* of 60 members of the 10 clusters obtained using k-means over the $(\lambda_2, \lambda_3, \lambda_4)$ values.

The 125250 points of the slice are distributed through the clusters following the histogram of the figure 43 and Table 12.

Cluster	λ_2	λ_3	λ_4
1	0.0013937313	0.9585829	1.04696461
2	0.0005291388	1.1633978	-0.07162550
3	0.0020630696	0.1349486	0.17305941
4	0.0016238358	0.8653824	0.83857646
5	0.0027346929	0.5084664	0.39199164
6	0.0003894541	1.4076354	-0.01925743
7	0.0021972784	0.3253562	0.01493809
8	0.0015421749	0.9491101	0.86699555
9	0.0018672401	0.2176002	0.17862024
10	0.4856397733	0.1404140	0.14011298

Table 11 – Centers of the clusters.

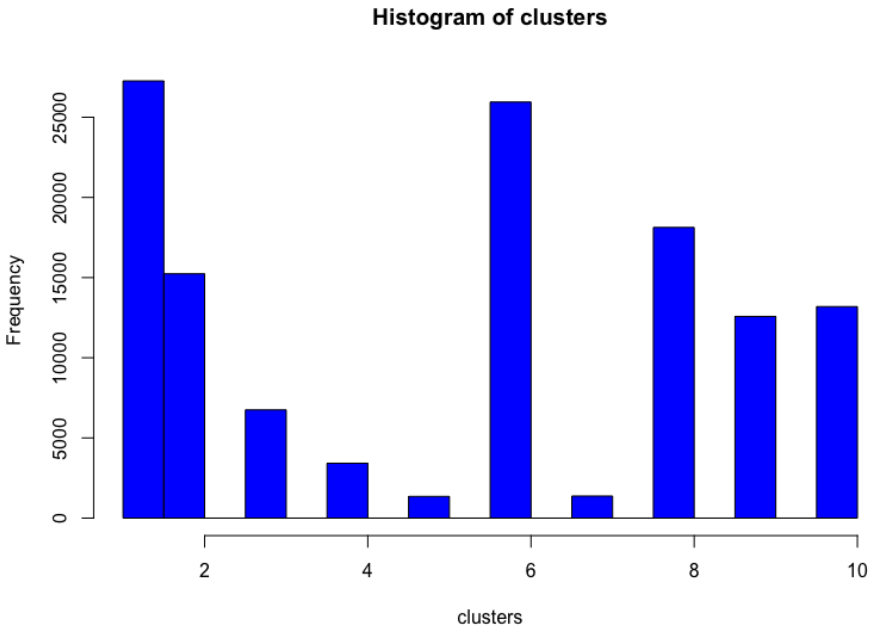

Figure 43 – Distribution of the clusters.

6.1.6 Spatio-temporal queries

At this point, the initial dataset is summarized as depicted by the schema in equation 6.2. It can be used to answer queries and to validate our approach, comparing the results with the raw data.

First of all we select four spatio-temporal regions of the dataset where the clusters suggest us different behaviors. The regions are shown in Figure 44 and the values of $[x_1, x_2], [y_1, y_2]$ that define the regions are shown in Table 13.

Cluster	**No. of members**
1	27217
2	15223
3	6749
4	3421
5	1353
6	25853
7	1374
8	18103
9	12051
10	13156

Table 12 – Distribution of the clusters.

Region	x_1	x_2	y_1	y_2
Region 1	210	250	0	40
Region 2	150	250	50	150
Region 3	0	75	100	200
Region 4	0	250	300	400

Table 13 – Analysis Regions.

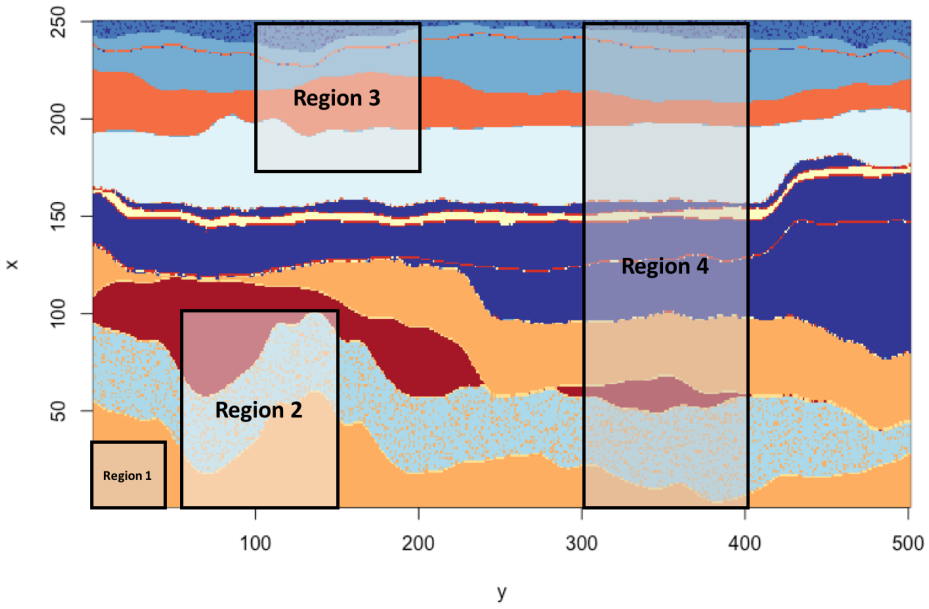

Figure 44 – Analysis Regions.

With these four regions we assess the adoption of the *GLD* mixture to obtain the *PDF* that characterizes the uncertainty in an specific region, section 6.1.6.1; and in section 6.1.6.2. We use the Information Entropy to assign a value that measures the uncertainty at each region. In section 6.1.6.1, we expect the GLD mixture to characterize well the raw

Cluster	Region 1	Region 2	Region 3	Region 4
1	0	2250	0	979
2	0	0	0	268
3	0	0	2596	1468
4	1640	4467	0	5173
5	0	149	0	269
6	0	0	0	416
7	0	0	1967	3920
8	0	3335	0	3432
9	0	0	1918	3280
10	0	0	901	583

Table 14 – Distribution of the clusters by regions.

data; and in 6.1.6.2 we hope that the information entropy is zero in region 1 and increases between regions 2, 3 and 4.

6.1.6.1 GLD mixture

The experiment here is to use the representative *GLDs* at each cluster and the weight associated to it in the region. Using these parameters we can build a *GLD mixture* that characterizes the uncertainty on that region. Here we use the algorithm described in section **??**.

First of all we query the region to find the clusters represented inside it, and how are they distributed. Below we show the R codes to query the four regions. The retrieved results are shown in Table 14.

```
> clRegion1 = clByRegion(210, 250, 0, 40)
> clRegion2 = clByRegion(150, 250, 50, 150)
> clRegion3 = clByRegion(0, 75, 100, 200)
> clRegion4 = clByRegion(0, 250, 300, 400)
```

If we divide the columns of Table 14 by the sum of the elements of each column we get the weight needed to formulate the *mixed GLDs*. It is clear that the *GLD* in region 1 is represented by the *GLD* of cluster 4. On the other 3 cases we get:

Metrics	**Region 1**	**Region 2**	**Region 3**	**Region 4**
p-value	0.73	0.56	0.34	0.08

Table 15 – p-values by regions.

entropy	**Region 1**	**Region 2**	**Region 3**	**Region 4**
value	0	1.122243	1.41166	2.024246

Table 16 – Information Entropy by regions.

$$
\begin{aligned}
&GLD_{region1} = GLD_{c4} \\
&GLD_{region2} = 0.22GLD_{c1} + 0.44GLD_{c4} + 0.014GLD_{c5} \\
&+ 0.33GLD_{c8} \\
&GLD_{region3} = 0.34GLD_{c3} + 0.26GLD_{c7} + 0.25GLD_{c9} \\
&+ 0.12GLD_{c10} \\
&GLD_{region4} = 0.22GLD_{c1} + 0.44GLD_{c4} + 0.014GLD_{c5} \\
&+ 0.33GLD_{c8}
\end{aligned}
$$

Now we need to evaluate if the *mixture of GLDs* describes well the uncertainty in the regions. To do this we perform the same *ks-test* used to evaluate the goodness of the fit and described in Section 5.2.3.

Based on the *p-value*, Table 15, we can conclude that in all 4 regions the *mixture of GLDs* is a good fit to the raw data.

6.1.6.2 Information Entropy

Now we are going to evaluate what happens with the information entropy. Based on the distribution of clusters inside the regions, table 14; we can compute the entropy. In this case we use an R function called *entropy*, implemented in the r-package of the same name (HAUSSER; STRIMMER, 2008).

As we expect, Table 16, the entropy in region 1 is zero, because the region contains only members of the cluster 4. On the other regions the entropy increases from region 2 to region 4, as we expected.

It is clear that the information entropy is a very good and simple measure of the uncertainty, and here it is demonstrated its utility combined with the *GLD*.

The first one is a geophysical tests for wave propagation problems

As a first case study we use the “HPC4E Seismic Test Suite”, a collection of four 3D models and sixteen associated tests that can be downloaded freely at the project’s

website (https://hpc4e.eu/downloads/datasets-and-software). The models include simple cases that can be used in the development stage of any geophysical imaging practitioner (developer, tester ...) as well as extremely large cases that can only be solved in a reasonable time using ExaFLOPS supercomputers. The models are generated to the required size by means of a Matlab/Octave script and hence can be used by users of any OS or computing platform. The tests can be used to benchmark and compare the capabilities of different and innovative seismic modelling approaches, hence simplifying the task of assessing the algorithmic and computational advantages that they pose.

In our case, we are going to use the "HPC4E Seismic Test Suite" as a case study of the porposed UQMS. As we mention in the introduction of this chapter this model is a spatial only domain problem, because we are going to consider a multidimentional array as an Input and a multidimentional array as an output, but of them time independet.

6.1.7 Mathematical Formulation

6.1.8 Model and Dataset Description

The models have been designed as a set of 16 layers with constant physical properties. The top layer delineates the topography and the other 15 different layer interface surfaces or horizons. In the following, an interface horizon is associated with properties that apply to the layer that exists between itself and the immediately next layer horizon. The model covers an area of 10 x 10 x 5 km, with maximum topography at about 500 m and maximum depth at about 4500 m. The layer horizons have been sampled very finely with 1.6667 m spacing so that a highly accurate representation can be honored at high frequencies. For simulation schemes based on unstructured grids, the layer horizons can be used easily to constrain model blocks. For simulation schemes based upon Cartesian grids, a simple script is provided that can generate 3D grids for any desired spatial sampling. Table 17 shows the properties of each of the layers included in the models.

6.1.9 Adding uncertainty into the model

The "HPC4E Seismic Test Suite" does not provide uncertainty sources, because all the input parameters of the model have fixed values. Then, to the purpose of our work we need to add some uncertainties into the inputs. Let's suppose the variable V_p is uncertain. As this variable have 16 different values, one for each layer, we can consider it as a random vector, equation 6.3. We associate to each of the V_{p_i} a Normal distribution with μ_i equal to the value reported in Table 17 and $\sigma = 2$.

$$V_p =< V_{p_i}, \mathcal{N}(\mu_i, \sigma_i) > \tag{6.3}$$

Table 17 – Layer constant properties and their depth range. “Star” layers are only used in the flat case, in substitution of their non-star equivalents

Layer Id	**Vp (m/s)**	**Vs (m/s)**	**Density (Kg/m3)**	**Max. depth (m)**	**Min. depth (m)**
1	1618.92	500.00	1966.38	-135.55	-476.35
2	1684.08	765.49	1985.88	41.50	-394.90
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
2*					
3*					

6.2 Case study with cross-correlated variables

Este caso de estudio es el primero del paquete R spup. Nuestros resultados son los mismos que ellos muestran haciendo uso de la GLD. Una query interesante es encontrar donde determinado valor es mayor que 24. En los codigos del capitulo esta la respuesta super facil haciendo uso de la funcion qgl del paquete GLDEX. Esta funcion nos devuelve el valor del quantil 90 y de ahi buscamos donde este valor es mayor que 30, el resultado es el mismo que ellos muestran en el ejemplo.

6.3 Case Study: Austin, queso library

6.4 Case Study: Multidisciplinary System (NASA)

6.5 Case Study: Spatio-temporal Nicholson-Bailey model

Este esta en el software uqlab, en la carpeta Doc Manuals

7 Conclusions and Future Works

Large-scale spatio-temporal simulations produce a huge amount of data that need to be interpreted in order to assess the simulation quality in different regions of space-time. Querying these data poses a great challenge due to their volume and different data distributions. In order to solve this problem, in this paper we propose SUQ^2, a general approach to answer uncertainty quantification queries.

The approach uses *GLD* that enables the representation of a spatio-temporal simulation output using a single functional formalism. By modeling each spatio-temporal point by a GLD instance, we can synthesize the region in a number of clusters, represented by their centroid GLD function. From this basis, queries can be answered by combining the centroids in a spatio-time region into GLD-mixture functions. Moreover, by using information entropy techniques, a value can be assigned that represents the uncertainty in a region. The proposed approach is implemented in a workflow that can be extended to solve new UQ queries.

We ran extensive experiments using a seismic use case. The results showed that GLD representation of the data is valid on 85% of the dataset. Other extensions of the GLD formalism, such as EGLD (KARIAN; DUDEWICZ, 2011), can be evaluated to improve the GLD dataset coverage. Moreover, we showed that the computed centroid function is a good representation of the function instances in its cluster. Additionally, we use the Kolmogorov-Smirnov test to evaluate the quality of the GLD mixture. The p-value, larger than 0.05, assures that the results of the mixture is a good representation of the raw data in the region. Finally, the adoption of the Information Entropy technique was validated by showing the correspondence of the computed values with the uncertainty in the spatio-temporal regions.

To the best of our knowledge, this is the first work to use GLD as the basis for answering UQ queries in spatio-temporal regions and to compile a series of techniques to produce a query answering workflow.

7.1 Revisiting the Research Questions

7.2 Significance and Limitations

7.3 Open Problems and Future Work

Some of the future directions we are interested in pursuing were mentioned above. For example, in Section 6.1.5 we mention that for the purpose of this paper we select *k-means* as the clustering algorithm to be used. This arbitrary selection needs to be studied, and some algorithms implemented to provide an automatic way to cluster the *GLDs*, based on the shapes described in Section **??**.

In Section 6.1.4, there is a region where the *GLD* does not fit well the dataset. If we want to provide a general purpose computational approach for *forward propagation* we need to further investigate this issue.

The use of Information Entropy to quantify the uncertainty is very powerful. However, when applied on clusters of PDFs, such as the GLD, it observes the information variation as a function of the PDF definition, in the case of GLD this is given by its for λ parameters. In this context, a complete region modeled by a single GLD function would have a very low information entropy value. This, however would not express the uncertainty modeled by the GLD function, which could be very high. The outcome of the information entropy evaluation must be interpreted by the user.

7.4 Final Considerations

Acknowledgments

This work has been funded by CNPq, CAPES, FAPERJ, Inria (SciDISC project) and the European Commission (HPC4E H2020 project) and performed (for E. Pacitti and P. Valduriez) in the context of the Computational Biology Institute (www.ibc-montpellier.fr) and for (F. Porto, H. Lustosa and N. Lemus) in the context of the DEXL Laboratory (dexl.lncc.br) at LNCC.

Bibliography

ALLEN, M. R. et al. Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature*, Nature Publishing Group, v. 407, n. 6804, p. 617, 2000. Citado na página 23.

ALVIN, K. F. et al. Uncertainty quantification in computational structural dynamics: a new paradigm for model validation. *Society for Experimental Mechanics, Inc, 16 th International Modal Analysis Conference.*, v. 2, p. 1191–1198, 1998. Citado na página 24.

BARONI, G.; TARANTOLA, S. A General Probabilistic Framework for uncertainty and global sensitivity analysis of deterministic models: A hydrological case study. *Environmental Modelling and Software*, Elsevier Ltd, v. 51, p. 26–34, 2014. ISSN 13648152. Disponível em: <http://dx.doi.org/10.1016/j.envsoft.2013.09.022>. Citado 2 vezes nas páginas 19 and 23.

BAXTER, M. J.; COOL, H. E. M. Reinventing the wheel? Modelling temporal uncertainty with applications to brooch distributions in Roman Britain. *Journal of Archaeological Science*, Elsevier Ltd, v. 66, p. 120–127, 2016. ISSN 10959238. Citado na página 33.

CHALABI, Y.; DIETHELM, W.; SCOTT, D. J. Flexible Distribution Modeling with the Generalized Lambda Distribution. 2012. Disponível em: <https://pdfs.semanticscholar.org/6b34/5bfa8ca3e73fadc11359155c2c5f33e63a7b.pdf>. Citado na página 41.

CHEN, J.; FLOOD, M. D.; SOWERS, R. B. Measuring the Unmeasurable: An Application of Uncertainty Quantification to Financial Portfolios Measuring the Unmeasurable An application of uncertainty quantification to financial portfolios. *Quantitative Finance*, v. 7688, n. January, p. 1–18, 2008. ISSN 14697696. Disponível em: <http://dx.doi.org/10.1080/14697688.2017.1296176>. Citado 2 vezes nas páginas 19 and 23.

CORLU, C. G.; METERELLIYOZ, M. Estimating the Parameters of the Generalized Lambda Distribution: Which Method Performs Best? *Communications in Statistics: Simulation and Computation*, v. 45, n. 7, p. 2276–2296, 2016. ISSN 15324141. Citado 2 vezes nas páginas 40 and 44.

COX, M. et al. Numerical aspects in the evaluation of measurement uncertainty. *IFIP Advances in Information and Communication Technology*, v. 377 AICT, p. 180–192, 2012. ISSN 18684238. Citado 3 vezes nas páginas 33, 35, and 48.

CRESPO, L. G.; KENNY, S. P.; GIESY, D. P. The NASA Langley Multidisciplinary Uncertainty Quantification Challenge. *16th AIAA Non-Deterministic Approaches Conference*, n. January, p. 1–9, 2014. Disponível em: <http://arc.aiaa.org/doi/abs/10.2514/6.2014-1347>. Citado na página 19.

ESTACIO-HIROMS, K. C.; PRUDENCIO, E. E. User's Manual: Quantification of Uncertainty for Estimation, Simulation, and Optimization (QUESO). 2012. Citado 2 vezes nas páginas 19 and 33.

FARRELL, K.; ODEN, J. T.; FAGHIHI, D. A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems. *Journal of Computational Physics*, Elsevier Inc., v. 295, p. 189–208, 2015. ISSN 10902716. Disponível em: <http://dx.doi.org/10.1016/j.jcp.2015.03.071>. Citado na página 33.

FARRELL, K. A. Selection , Calibration , and Validation of Coarse-Grained Models of Atomistic Systems. 2015. Citado na página 19.

FOURNIER, B. et al. Estimating the parameters of a generalized lambda distribution. *Computational Statistics and Data Analysis*, v. 51, n. 6, p. 2813–2835, 2007. ISSN 01679473. Citado na página 44.

FREIMER, M.; LIN, C. T.; MUDHOLKAR, G. S. A Study Of The Generalized Tukey Lambda Family. *Communications in Statistics - Theory and Methods*, v. 17, n. 10, p. 3547–3567, 1988. ISSN 1532415X. Citado 2 vezes nas páginas 39 and 41.

Gharib Shirangi, M. History matching production data and uncertainty assessment with an efficient TSVD parameterization algorithm. *Journal of Petroleum Science and Engineering*, v. 113, p. 54–71, 2014. ISSN 09204105. Citado na página 32.

GRALER, B.; PEBESMA, E.; HEUVELINK, G. Spatio-Temporal Interpolation using gstat. *Wp*, v. 8, p. 1–20, 2016. ISSN 20734859. Citado na página 77.

GUERRA, G. M. et al. Uncertainty quantification in numerical simulation of particle-laden flows. *Computational Geosciences*, v. 20, n. 1, p. 265–281, 2016. ISSN 1420-0597. Disponível em: <http://link.springer.com/10.1007/s10596-016-9563-6>. Citado 2 vezes nas páginas 19 and 23.

HAUSSER, J.; STRIMMER, K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. n. October 2008, p. 1–18, 2008. ISSN <null>. Disponível em: <http://arxiv.org/abs/0811.3579>. Citado na página 92.

HELTON, J. Conceptual and computational basis for the quantification of margins and uncertainty. n. June, 2009. Disponível em: <http://www.osti.gov/energycitations/product.biblio.jsp?osti{_}id=958>. Citado 2 vezes nas páginas 25 and 26.

HELTON, J. C. et al. Representation of analysis results involving aleatory and epistemic uncertainty. *International Journal of General Systems*, v. 39, n. 6, p. 605–646, 2010. ISSN 0308-1079. Citado na página 27.

HIGDON, D. *Handbook of Uncertainty Quantification*. [s.n.], 2017. ISBN 978-3-319-12384-4. Disponível em: <http://link.springer.com/10.1007/978-3-319-12385-1>. Citado na página 27.

JCGM, J. C. F. G. I. M. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. *International Organization for Standardization Geneva ISBN*, v. 50, n. September, p. 134, 2008. ISSN 00099147. Disponível em: <http://www.bipm.org/en/publications/guides/gum.html>. Citado na página 48.

JIANG, B. et al. CLUSTERING UNCERTAIN DATA BASED ON PROBABILITY DISTRIBUTION SIMILARITY. Disponível em: <https://pdfs.semanticscholar.org/e172/2c8911b7db1a3114fbd38b3ea5a9e93d1290.pdf>. Citado na página 70.

JIANG, B. et al. Clustering Uncertain Data Based on Probability Distribution Similarity. *IEEE Transactions on Knowledge and Data Engineering*, p. 1–14, 2011. ISSN 1041-4347. Disponível em: <https://pdfs.semanticscholar.org/e172/2c8911b7db1a3114fbd38b3ea5a9e93d1290.pdf>. Citado 3 vezes nas páginas 50, 51, and 56.

JIANG, B. et al. Clustering Uncertain Data Based on Probability Distribution Similarity. *IEEE Transactions on Knowledge and Data Engineering*, v. 25, n. 4, p. 751–763, apr 2013. ISSN 1041-4347. Disponível em: <http://ieeexplore.ieee.org/document/6051435/>. Citado na página 51.

JOHNSTONE, R. H. et al. Uncertainty and variability in models of the cardiac action potential: Can we build trustworthy models? *Journal of Molecular and Cellular Cardiology*, The Authors, v. 96, p. 49–62, 2016. ISSN 10958584. Disponível em: <http://dx.doi.org/10.1016/j.yjmcc.2015.11.018>. Citado na página 19.

JOINER, B. L.; ROSENBLATT, J. R. Some Properties of the Range in Samples from Tukey's Symmetric Lambda Distributions. *Journal of the American Statistical Association*, v. 66, n. 334, p. 394–399, jun 1971. ISSN 0162-1459. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482275>. Citado na página 40.

KARIAN, Z. A.; DUDEWICZ, E. J. *Handbook of fitting statistical distributions with R*. [S.l.: s.n.], 2011. ISSN 1098-6596. ISBN 9788578110796. Citado 11 vezes nas páginas 12, 20, 34, 36, 38, 41, 44, 45, 47, 86, and 95.

KARVANEN, J.; NUUTINEN, A. Characterizing the generalized lambda distribution by L-moments. *Computational Statistics and Data Analysis*, v. 52, n. 4, p. 1971–1983, 2008. ISSN 01679473. Disponível em: <https://ia801001.us.archive.org/27/items/arxiv-math0701405/math0701405.pdf>. Citado na página 44.

KAWAI, S.; SHIMOYAMA, K. Kriging-model-based uncertainty quantification in computational fluid dynamics. *32nd AIAA Applied Aerodynamics Conference*, n. June, p. 1–16, 2014. Disponível em: <http://arc.aiaa.org/doi/10.2514/6.2014-2737>. Citado na página 32.

KENNEDY, M. C.; O'HAGAN, A. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Blackwell Publishers Ltd., v. 63, n. 3, p. 425–464, aug 2001. ISSN 1369-7412. Disponível em: <http://doi.wiley.com/10.1111/1467-9868.00294>. Citado na página 24.

KIUREGHIAN, A. D.; DITLEVSEN, O. Aleatory or epistemic? Does it matter? *Structural Safety*, v. 31, n. 2, p. 105–112, mar 2009. ISSN 01674730. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0167473008000556>. Citado na página 25.

LAKHANY, A.; MAUSSER, H. Estimating the Parameters of the Generalized Lambda Distribution. *ALGO RESEARCH QUARTERLY*, v. 3, n. 3, 2000. Disponível em: <https://pdfs.semanticscholar.org/0f9d/1848671969232d58cb7cf3d2d06d9c4c347e.pdf?{_}ga=2.56724463.474793085.1524894682-1121088995.1524894>. Citado 2 vezes nas páginas 36 and 44.

LAMPASI, D. A.; Di Nicola, F.; PODESTA, L. Generalized lambda distribution for the expression of measurement uncertainty. *IEEE Transactions on Instrumentation and Measurement*, v. 55, n. 4, p. 1281–1287, 2006. ISSN 00189456. Citado 4 vezes nas páginas 20, 28, 47, and 52.

LIU, C. M.; NIU, Z.; LIAO, K. T. *Mechanisms to improve clustering uncertain data with UKmeans*. North-Holland, 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169023X16303287>. Citado 2 vezes nas páginas 21 and 51.

LODZIENSIS, A. U. Generalizations of tukey-lambda distributions. 2013. Citado na página 44.

MARCONDES, D.; PEIXOTO, C.; MAIA, A. C. FITTING A HURDLE GENERALIZED LAMBDA DISTRIBUTION TO HEALTHCARE EXPENSES. *Annals of Applied Statistics*, 2017. Disponível em: <https://arxiv.org/pdf/1712.02183.pdf>. Citado 4 vezes nas páginas 12, 39, 40, and 44.

MELOROSE, J. et al. *A PROBABILISTIC FRAMEWORK FOR UNCERTAINTY QUANTIFICATION IN LARGE-SCALE SIMULATIONS: APPLICATION IN SEISMIC IMAGING*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2015. Citado na página 23.

MOVAHEDI, M. M.; LOTFI, M. R.; NAYYERI, M. A solution to determining the reliability of products Using Generalized Lambda Distribution. *Research Journal of Recent Sciences Res.J.Recent Sci*, v. 2, n. 10, p. 41–47, 2013. Disponível em: <http://www.isca.in/rjrs/archive/v2/i10/7.ISCA-RJRS-2013-227.pdf>. Citado na página 48.

Mustafa Inchasi, E. A. *The Generalized Lambda Distribution and Its Use in Fitting Distributions to Data*. Tese (Doutorado), 2016. Citado 2 vezes nas páginas 34 and 46.

NING, W.; GAO, Y.; DUDEWICZ, E. J. Fitting mixture distributions using generalized lambda distributions and comparison with normal mixtures. *American Journal of Mathematical and Management Sciences*, v. 28, n. 1-2, p. 81–99, 2008. ISSN 01966324. Citado 2 vezes nas páginas 45 and 46.

PATT, A.; KLEIN, R. J. T.; VEGA-LEINERT, A. de la. Taking the uncertainty in climate-change vulnerability assessment seriously. *Comptes Rendus Geoscience*, Elsevier, v. 337, n. 4, p. 411–424, 2005. Citado na página 23.

RAJAN, A. et al. Benchmark Test Distributions for Expanded Uncertainty Evaluation Algorithms. *IEEE Transactions on Instrumentation and Measurement*, v. 65, n. 5, p. 1022–1034, 2016. ISSN 00189456. Citado na página 48.

SANKARARAMAN, S. *Uncertainty Quantification and Integration*. Tese (Doutorado) — Vanderbilt University, 2012. Citado na página 32.

SU, S. Fitting Single and Mixture of Generalized Lambda Distributions to Data via Discretized and Maximum Likelihood Methods: GLDEX in R. *Journal of Statistical Software*, v. 21, n. 9, 2007. Citado 4 vezes nas páginas 41, 44, 46, and 49.

SU, S. Maximum Log Likelihood Estimation using EM Algorithm and Partition Maximum Log Likelihood Estimation for Mixtures of Generalized Lambda Distributions. *Journal of Modern Applied Statistical Methods*, v. 10, n. 2, p. 599–606, 2011. ISSN 1538-9472. Disponível em: <http://digitalcommons.wayne.edu/jmasm/vol10/iss2/17>. Citado 2 vezes nas páginas 44 and 46.

SU, S. Flexible parametric quantile regression model. *Statistics and Computing*, v. 25, n. 3, p. 635–650, 2015. ISSN 15731375. Citado 2 vezes nas páginas 40 and 44.

SU, S. Fitting Flexible Parametric Regression Models with GLDreg in R. *Journal of Modern Applied Statistical Methods*, v. 15, n. 2, p. 768–787, 2016. ISSN 1538-9472. Disponível em: <http://digitalcommons.wayne.edu/jmasmhttp://digitalcommons.wayne.edu/jmasm/vol15/iss2/46>. Citado na página 44.

SULLIVAN, T. J. *Introduction to Uncertainty Quantification*. Springer, 2015. ISBN 9783319233949. Disponível em: <http://www.springer.com/series/1214>. Citado 5 vezes nas páginas 23, 26, 28, 31, and 32.

TOBERGTE, D. R.; CURTIS, S. Workshop on Quantification, Communication, and Interpretation of Uncertainty in Simulation and Data Science. *Journal of Chemical Information and Modeling*, v. 53, n. 9, p. 1689–1699, 2013. ISSN 1098-6596. Citado 4 vezes nas páginas 19, 21, 23, and 34.

U.S. Department of Energy. *Scientific Grand Challenges in National Security: The Role of Computing at the Extreme Scale*. [S.l.], 2009. 255 p. Citado 3 vezes nas páginas 26, 27, and 104.

WELLMANN, J. F.; REGENAUER-LIEB, K. Uncertainties have a meaning: Information entropy as a quality measure for 3-D geological models. *Tectonophysics*, Elsevier B.V., v. 526-529, p. 207–216, 2012. ISSN 00401951. Disponível em: <http://dx.doi.org/10.1016/j.tecto.2011.05.001>. Citado na página 29.

Appendix

APPENDIX A – uqms R package

A.1 Título da seção

Aqui temos uma seção dentro do Apêndice.

APPENDIX B – Ideas

B.0.1 Variance, Information and Entropy

Variance.

Information and Entropy.

B.0.2 Information Gain, Distances and Divergences

B.1 Sensitivity Analysis

Sensitivity analysis is the systematic study of how model inputs—parameters, initial and boundary conditions—affect key model outputs. Depending on the application, one might use local derivatives or global descriptors such as Sobol's functional decomposition or variance decomposition. Also, the needs of the application may range from simple ranking of the importance of inputs to a response surface model that predicts the output given the input settings. Such sensitivity studies are complicated by a number of factors, including the dimensionality of the input space, the complexity of the computational model, limited forward model runs due to the computational demands of the model, the availability of adjoint solvers or derivative information, stochastic simulation output, and high-dimensional output. Challenges in sensitivity analysis include dealing with these factors while addressing the needs of the application. (U.S. Department of Energy, 2009)

$$E = mc^2 \tag{B.1}$$

APPENDIX C – Título do apêndice C

Annex

ANNEX A – Título do anexo A

A.1 Título da seção

Aqui temos uma seção dentro do Anexo.

ANNEX B – Título do anexo B

ANNEX C – Título do anexo C