Carlos Alberola, S.; Sánchez Galdón, AI.; Ginestar Peiro, D.; Martorell Alsina, SS. (2013). Using finite mixture models in thermal-hydraulics system code uncertainty analysis. Nuclear Engineering and Design. 262:306-318. doi:10.1016/j.nucengdes.2013.04.030.

Abstract

Nuclear power plant safety analysis is mainly based on models that predict the plant behavior under normal or accidental conditions. As the models introduce approximations, it is necessary to perform an uncertainty analysis of the results obtained. The most popular approach, based on Wilks' method, obtains a tolerance/confidence interval, but it does not completely characterize the output variable behavior. In order to obtain more information about the output variable there exist different techniques to reconstruct the probability distribution using the information provided by a sample of values as, for example, the finite mixture models. In this paper, the Expectation Maximization and the k-means algorithms are used to obtain a finite mixture model that reconstructs the output variable probability distribution from data obtained with RELAP-5 simulations. Both methodologies were applied to a separated effects experiment, and an integral effects simulation.

Keywords: thermal-hydraulic codes, uncertainty, finite mixture models, Expectation Maximization algorithm, k-means algorithm.

1. Introduction

Nuclear power plant safety analysis is mainly based on neutronic and thermal hydraulic models that predict the plant behavior under normal or accidental conditions (Guba et al., 2003). Thermal hydraulic calculations can be performed either using conservative or best estimate codes, providing this latter option more realistic results. Nevertheless, as the model introduces approximations, it is necessary to perform an uncertainty analysis of the results obtained. In this way, the general modeling methodology process comprises different steps, from capturing reality to conceptual models to convert those models to computerized codes. In this process there are always numerous simplifications, approximations, round-off errors, numerical techniques, and so on, which cause uncertainties in the calculation (Pourgol-Mohammad, 2009). Thus, the uncertainty of the results obtained with the code has to be quantified in order to give credit to the predictions obtained (Pourgol-Mohammad et al., 2011).

From the last decade, the regulatory bodies allow the use of thermal hydraulic simulation codes to guarantee the safe operation of nuclear installations, but only if the uncertainty associated with the simulation is properly quantified (Boyack et al.,1990; Wilson et al., 1990; Wulf et al., 1990). The plant simulations undertaken using best

estimate codes combined with uncertainty analysis is known as Best Estimate Plus Uncertainty (BEPU) approach (Crécy et al. 2008).

In the literature there exist different approaches to quantify uncertainty in best estimate codes. For example, in Cacuci and Ionescu-Bujor (2000a) and in Cacuci and Ionescu-Bujor (2000b) a deterministic approach is followed using the adjoint sensitivity analysis method for RELAP code. This approach needs to implement the uncertainty quantification method in a new code, which has to be coupled with the thermal-hydraulic code. This is known as an intrusive method as the original thermal hydraulic code has to be modified. But the most popular approaches to quantify code uncertainty are the non intrusive methods. Such approaches use the thermal-hydraulic code as a black box model, in which the output variables are linked to the input variables (Guba et al., 2003). That is, given and input variable vector, $\vec{x}$, the computer code transforms it into a vector $\vec{y}$ of output variables,

$$\vec{y}\,(t) = f(\vec{x}, t\,). \tag{1}$$

In practical, this link is very complex but it is assumed to be deterministic, that is, once the input variables are fixed the same output is obtained from the code within the computation accuracy of each run. In this situation, the uncertainty associated with the input variables will be propagated to the output variables, and should be quantified. Some of the non intrusive methodologies developed to quantify best estimate codes uncertainty are the CSAU (Boyeck et al.,1990; Wilson et al., 1990; Wulf et al., 1990), the GRS methodologies (Glaeser et al., 1994) ASTRUM and IMTHUA (Pourgol-Mohammad, 2009). Reference Pourgol-Mohammad (2009) provides a detailed comparison of the uncertainty methodologies developed, and applied to the thermal hydraulic calculations.
In these methodologies, it is assumed that the input variables are uncertain, and follow a statistical distribution. In this way, fixing the time of the transient, after N runs, N randomly samples of a varying output variable are obtained, which carry information on the fluctuating input and the code properties.

The uncertainty in the output variables can be quantified by obtaining a tolerance/confidence interval, making use of the advantage of order statistics (Guba et al., 2003). Thus, assuming there is one output variable, *y* , with a probability distribution *g(y)*. If we carry N runs with fluctuating inputs, we obtain a sample $\{y_1, y_2, \dots, y_N\}$ of the random variable y . The usual approach is to construct two random functions $L = L(y_1, y_2, \dots, y_N)$ and $U = U(y_1, y_2, \dots, y_N)$, called tolerance limits, such that

$$P\left\{\int_L^U g(y)dy > \gamma\right\} = \beta, \tag{2}$$

where

$$\int_L^U g(y)dy = A(y_1, y_2, \dots, y_N), \quad (3)$$

is a random variable, called probability content, which measures the portion of the distribution included in the random interval $[L, U]$. Probability β is the confidence level, and γ is a non-negative real number not grater that 1. It is desirable to have values of β and γ as large as possible inside the interval [0, 1]. Having fixed β and γ, it becomes possible to determine the number of runs N necessary to determine an appropriate interval [L,U]. The first works that discussed the problem of setting tolerance/confidence intervals based on samples where developed by Wilks (1941), and they are the basis of uncertainty methodologies for quantifying best estimate codes uncertainty such as Glaesser (1994) and Guba et al. (2003). This approach has the advantage that the number of runs, N, necessary to determine the tolerance limits, is much lower than the runs necessary in a Monte Carlo approach. This is of great interest in the study of the uncertainty in best estimate codes, since, in most of situations, the code execution requires a high computational cost. However, the information provided by the tolerance limits methodology does not completely characterize the output variable behavior.

In order to obtain more information about the output variable, there exist different techniques to reconstruct its probability distribution using the information provided by a sample of values. For example, recently, polynomial chaos expansion methods have been used to reconstruct the probability distribution and to estimate its parameters, as a lower number of runs are needed compared with Monte Carlo approaches (Sundret, 2008; Eaton and Williams, 2010; Gili et al., 2012). This methodology provides very good results for unimodal distributions, but for multimodal probability distributions the order of the expansions needed to reconstruct the probability function increases and the computational cost becomes very large, especially in the multidimensional case (Nouy, 2010).

In the process of thermal hydraulic modelling using best estimate (BE) codes the initial plant state is represented by the initial and boundary conditions of the plant model, which are the input variables in the BE simulation. In many cases their values are unkown or uncertain, an such uncertainty is transmitted through the code to the output variable. So, assuming that the input variables are random and follow a certain probability distribution, if its variance or range of variation are small, generally, the probability distribution of the output variables can be reconstructed using polynomial chaos expansion and described by its first moment (mean) and its second moment

(variance), as the output variable follows an unimodal distribution. But if the variance or range of variation of the input variables is larger, the output variable distribution can become multimodal and the mean and the variance of its probability distribution are not enough to describe it.

The analysis of multimodal distributions has been successfully carried out using finite mixture models. Finite mixtures of distributions have provided a mathematical-based approach to the statistical modeling of a wide variety of random phenomena, because they are able to represent arbitrarily complex probability density functions (McLachlan, 2000). Because of their usefulness as an extremely flexible method of modeling, finite mixture models have continued to receive increasing attention over the years, from both a practical and theoretical point of view. Fields in which mixture models have been successfully applied include astronomy, biology, genetics, medicine, psychiatry, economics, engineering, and marketing, among many other fields in the biological, physical, and social sciences.

So, the use of finite mixture models can be of interest in the uncertainty quantification of thermal-hydraulic simulations if the output variables to be studied follow multimodal probability distributions. Two approaches can be used: In the first one, a Gaussian mixture model is built to reconstruct the output variable probability distribution and then classify the results into different clusters that can be characterized by its mean and standard deviation. The second approach consists of performing a previous classification of the output variable to assure that the clusters obtained follow, approximately, a Gaussian distribution. From this previous classification, a Gaussian mixture model is built to reconstruct the complete probability distribution of the output variable.

The rest of the paper is organized as follows: In section 2 finite mixture models, the Expectation Maximization based algorithm, section 3 reviews the k-means algorithm, section 4 presents two applications of both algorithms to a separated test effects, considering the experiments undertaken in the RIT facility and an integral test effects, considering a large break LOCA in a pressurized water reactor. Finally, section 5 presents the conclusions of the work.

2. Finite Mixture models

Finite Mixture Modeling was introduced by Karl Pearson to study a population of crabs using Gaussian distributions, and a moments based approach to determine the mixture parameters. The interest on Finite Mixture Models has increased since Dempster et al. (1977) published the Expectation Maximization (EM) algorithm, since this algorithm

has simplified the process of computing the mixture parameters. In the following we review the finite mixture models and the estimation of the parameters from observed data, using the EM algorithm.

Let y be a random variable, it is said that this variable follows a k-component finite mixture distribution if its probability density function can be written as:

$$f(y) = \sum_{i=1}^{k} \pi_i \, f_i(y), \tag{4}$$

where $f_i(y)$ are the different probability density functions and $\pi_1, \pi_2 \ldots, \pi_n$ are the mixing probabilities satisfying $0 < \pi_i < 1$, $i = 1 \ldots, k$ and

$$\sum_{i=1}^{k} \pi_i = 1 . \tag{5}$$

For the particular case of a mixture of Gaussian distributions, the expression of $f_i(y)$ is given by:

$$f_i(y) = \frac{1}{\sigma_i \sqrt{2\,\pi}} \, e^{-\frac{(y-\mu_i)^2}{2{\sigma_i}^2}}, \tag{6}$$

where μ_i and σ_i are the mean and the standard deviation of the distribution, respectively. A finite mixture of Gaussian distributions is considered in this work, although more general mixtures can be defined.

The standard method to fit finite mixture models to observed data is the Expectation Maximization (EM) algorithm (Dempster et al., 1997; McLachlan, 2000). In order to review how the EM algorithm works, we consider the particular case of a mixture of two Gaussian distributions given by

$$f(y) = \pi_1 f_1(y) + (1 - \pi_1) f_2(y), \tag{7}$$

where the parameters of this distributions have to be estimated from the samples $\{y_1, y_2, \ldots, y_n\}$ of the variable *y*. The generalization of the method for a larger number of distributions is straight forward.

The estimation of the mixture parameters using the maximum likelihood method implies to solve a system of nonlinear equations (McLachlan, 2000). Instead of using this method, the problem is interpreted as an incomplete data problem. In this way classification variables are introduced, $z_{1i}, z_{2i}, \; i = 1, \ldots, n$; being

$$z_{1i} = \begin{cases} 1 & if\ y_i\ follows\ f_1(y) \\ 0 & if\ y_i\ follows\ f_2(y) \end{cases}; \qquad z_{2i} = \begin{cases} 0 & if\ y_i\ follows\ f_1(y) \\ 1 & if\ y_i\ follows\ f_2(y) \end{cases}, \quad (8)$$

and satisfying $z_{1i} + z_{2i} = 1$.

For a given observation of the variable, y_j, calling $\theta=(\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ the vector of the mixture parameters, we have the conditional probabilities

$$P(y_j, z_{1j}, z_{2j}|\theta) = P(y_j|z_{1j}, z_{2j}, \theta)P(z_{1j}, z_{2j}|\theta), \quad (9)$$

where

$$P(y_j|z_{1j}, z_{2j}, \theta) = f_1^{z_{1j}}(y_j)\, f_2^{z_{2j}}(y_j), \quad (10)$$

and

$$P(z_{1j}, z_{2j}|\theta) = \pi_1^{z_{1j}}\,(1-\pi_1)^{z_{2j}}. \quad (11)$$

Thus, for a sample y_j , we can write

$$P(y_j, z_{1j}, z_{2j}|\theta) = f_1^{z_{1j}}(y_j)\, f_2^{z_{2j}}(y_j)\, \pi_1^{z_{1j}}\,(1-\pi_1)^{z_{2j}}, \quad (12)$$

and the logarithm

$$\log\left(P(y_j, z_{1j}, z_{2j}|\theta)\right) = z_{1j}\log\left(f_1(y_j)\right) + z_{1j}\log(\pi_1) + z_{2j}\log\left(f_2(y_j)\right) + z_{2j}\log(1-\pi_1)\,. \quad (13)$$

For the n samples $y = (y_1, y_2, \dots, y_n)$ we can write

$$L(y, z|\theta) = log(P(y, z|\theta)) = \sum_{j=1}^{n}\left(z_{1j}\log\left(f_1(y_j)\right) + z_{1j}\log(\pi_1) + z_{2j}\log\left(f_2(y_j)\right) + z_{2j}\log(1-\pi_1)\right). \quad (14)$$

To apply the EM algorithm we need an initial estimation of the parameters, which is taken for the z variables, $z^{(1)}$. With this estimation, the means and the standard deviations are initialized as

$$\mu_1^{(1)} = \frac{\sum_{j=1}^{n} z_{1j}^{(1)} y_j}{\sum_{j=1}^{n} z_{1j}^{(1)}}; \qquad \mu_2^{(1)} = \frac{\sum_{j=1}^{n} z_{2j}^{(1)} y_j}{\sum_{j=1}^{n} z_{2j}^{(1)}}, \quad (15)$$

$$\sigma_1^{2(1)} = \frac{\sum_{j=1}^{n} z_{1j}^{(1)} (y_j-\mu_1^{(1)})^2}{\sum_{j=1}^{n} z_{1j}^{(1)}}; \qquad \sigma_2^{2(1)} = \frac{\sum_{j=1}^{n} z_{2j}^{(1)} (y_j-\mu_2^{(1)})^2}{\sum_{j=1}^{n} z_{2j}^{(1)}}, \quad (16)$$

and, finally, the probability

$$\pi_1^{(1)} = \frac{\sum_{j=1}^{n} z_{1j}^{(1)}}{n}. \tag{17}$$

With this initialization the algorithm follows with the expectation step,

$$z_{1j}^{(i)} = P\left(z_{1j} = 1 \middle| \theta^{(i-1)}, y_j\right) = \frac{\pi_1^{(i-1)} f_1(y_j, \theta^{(i-1)})}{\pi_1^{(i-1)} f_1(y_i, \theta^{(i-1)}) + \left(1 - \pi_1^{(i-1)}\right) f_2(y_i, \theta^{(i-1)})}, \tag{18}$$

$$z_{2j}^{(i)} = P\left(z_{2j} = 1 \middle| \theta^{(i-1)}, y_j\right) = \frac{\left(1 - \pi_1^{(i-1)}\right) f_2(y_j, \theta^{(i-1)})}{\pi_1^{(i-1)} f_1(y_i, \theta^{(i-1)}) + \left(1 - \pi_1^{(i-1)}\right) f_2(y_i, \theta^{(i-1)})}, \tag{19}$$

which is a result of applying the Bayes' rule (McLachlan and Krishnan, 2008).

After the expectation step, the maximization step follows updating the means, standard deviations and the mixing probabilities by means of:

$$\mu_1^{(i)} = \frac{\sum_{j=1}^{n} z_{1j}^{(i)} y_j}{\sum_{j=1}^{n} z_{1j}^{(i)}}; \qquad \mu_2^{(i)} = \frac{\sum_{j=1}^{n} z_{2j}^{(i)} y_j}{\sum_{j=1}^{n} z_{2j}^{(i)}}, \tag{20}$$

$$\sigma_1^{2(i)} = \frac{\sum_{j=1}^{n} z_{1j}^{(i)} (y_j - \mu_1^{(i)})^2}{\sum_{j=1}^{n} z_{1j}^{(i)}}; \qquad \sigma_2^{2(i)} = \frac{\sum_{j=1}^{n} z_{2j}^{(i)} (y_j - \mu_2^{(i)})^2}{\sum_{j=1}^{n} z_{2j}^{(i)}}, \tag{21}$$

$$\pi_1^{(i)} = \frac{\sum_{j=1}^{n} z_{1j}^{(i)}}{n}. \tag{22}$$

The EM algorithm has several drawbacks (Figuereido and Jain, 2002), it is a local method, that is, it obtains local maxima of the likelihood function, and it is sensitive to the initialization used because the likelihood function of a mixture model can have several maxima. In Figuereido and Jain (2002) some modifications to EM algorithm are done to mitigate these problems. Such modifications include a criterion to select the optimal number of Gaussian distributions in the finite mixture model based on the Minimum Message Length criterion (MML) (Oliver et al. 1996). The implementation proposed by Figuereido and Jain (2002) is the one used in this paper.

Once the finite mixture model is obtained from the samples of a given variable, different clusters can be defined associated with each Gaussian distribution. It is assumed that a sample y_j belongs to the ith-cluster if it satisfies

$$\left|y_j - \mu_i\right| \leq 3\sigma_i \,. \tag{23}$$

The objective to construct the clusters is that inside each one of the clusters the output variable approximately follows a Gaussian distribution and can be characterized using its mean and the standard deviation, which is the usual approach followed in uncertainty quantification.

3.- K-means algorithm

An alternative approach to study a multimodal distribution from a population is to perform a previous classification of the population into clusters that approximately follow a Gaussian distribution. A possible classification is obtained from the k-means algorithm. There are many interpretations for this method, but all of them follow similar guidelines (Jain and Dubes, 1998).

Let us assume that we start with a sample of n values of a given variable $\{y_1, \dots, y_n\}$ that has to be grouped in k clusters $\{c_1, \dots, c_k\}$ with a n_g data per cluster, in such a way that

$$\sum_{g=1}^{k} n_g = n \quad . \tag{23}$$

The centroids of each cluster are

$$\bar{y}^g = \frac{1}{n_g} \sum_{i=1}^{n_g} y_i, \tag{24}$$

the square error for each cluster is

$$e_g^2 = \sum_{i=1}^{n_g} (y_i - \bar{y}^{(g)})^T (y_i - \bar{y}^{(g)}), \tag{25}$$

and the total square error is

$$E_K^2 = \sum_{g=1}^{k} e_g^2 \, . \tag{26}$$

The k-means method to obtain the k-clusters has the following steps:

Step 1.- Select an initial partion with k-clusters and compute the centroid of each cluster.

Step 2.- Generate a new partition by assigning each pattern to its clostest cluster centroid.

Step 3.- Compute new cluster centers as the centroids of the each cluster.

Step 4.- Repeat steps 2 and 3 until an optimum value of the square error function is found.

Step 5.- Adjust the number of clusters by merging and splitting existing clusters or by removing small or outlier clusters.

This algorithm presents some weakness. For example, when the number of data is small the initial grouping will determine the final clustering. Moreover, the number of clusters has to be defined beforehand. And also it is not robust to outliers, that is, very far data from the centroids may pull the centroid away from the real one. Many implementations of the different steps of the method have been proposed to minimize these drawbacks. The particular implementation used here is the one included in the kmeans() function of MatLab package (Sober, 1984; Spath, 1985).

Once the k clusters $\{c_1, \dots, c_k\}$ are obtained with a n_i data per cluster, the mean, μ_i, and the standard deviation, σ_i, are computed for each one of the clusters. A finite mixture model can be built as

$$f(y) = \sum_{i=1}^{k} \pi_i \, f_i(y), \tag{27}$$

being $f_i(y)$

$$f_i(y) = \frac{1}{\sigma_i \sqrt{2\,\pi}} \, e^{-\frac{(y-\mu_i)^2}{2{\sigma_i}^2}}, \tag{28}$$

and the probability mixtures are

$$\pi_i = \frac{n_i}{n}. \tag{29}$$

4.- Thermal-hydraulic simulations

In the following we will study the uncertainty propagation for typical transients using the best estimate code RELAP5 (Information Systems Laboratories, 2001) where the output variables follow multimodal distributions showing the usefulness of finite mixture technique to describe them. Also a clustering technique will be investigated to classify the output variables into several groups following a Gaussian distribution. In particular the probability distribution function of the maximum wall temperature is reconstructed for a dry-out transient associated with an experiment in the Royal Institute of technology facility. Also, the peak cladding temperature (PCT) distribution obtained

from a loss of coolant accident in a typical PWR is reconstructed, using both the EM and k-means algorithms.

4.1. Royal Institute of technology transients

In the scope of the International Thermal-Hydraulic Code Assessment and Application Program (ICAP), a series of experiments were performed on the RIT facility to study the accuracy of the thermal-hydraulic codes to simulate the post-dry-out heat transfer process. ICAP program is oriented to assess the codes capabilities in order to improve future code versions.

Fig. 1 shows a simplified flow diagram of the RIT facility (Nilsson, 1993). In this Figure we can observe the test section, which is constituted by a vertical pipe. This pipe is electrically heated and the wall temperatures are measured by different thermocouples distributed along its length, as shown in Fig 2.

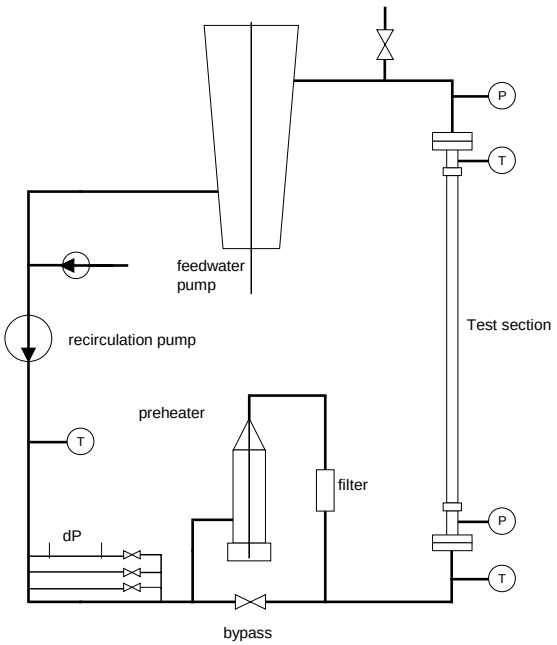

Fig. 1. Diagram of the RIT experimental facility.

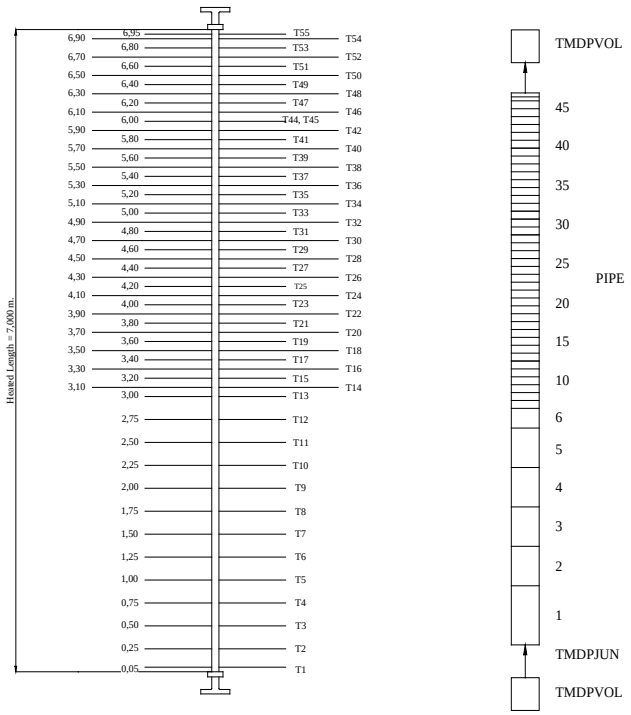

Fig. 2. Placement of the thermocouples in the heated pipe and nodalization for the RELAP5 model.

As the important phenomena take place inside the vertical pipe, the RELAP5 plant model only considers this zone, and the rest of the installation is substituted by boundary conditions. As can be observed in Fig 2., the length of the volumes that constitute the PIPE element has been chosen according to the location of the thermocouples, to guarantee the temperature calculated by RELAP5 corresponds to an experimental measure.

The boundary conditions are defined at the bottom of the PIPE by a TMDPVOL connected to the PIPE by a TMDPJUN, to assure that the pressure and the inlet mass flow will remain constant along the simulation. At the top of the PIPE a TMDPVOL defines a constant pressure during the transient. Using this plant model several transients were performed with different conditions. In this paper experiment 136 has been selected to perform the uncertainty analysis. The parameters which define the transient are shown in Table 1 (Nilsson, 1993).

Table 1.

Experiment RIT 136 parameters

Pressure (MPa)	Mass flux (kg/s)	Inlet temp. (^oK)	Heat flux (kW/m^2)	Measured CHF localization (m)	Oulet steam quality (x)

13.99	0.34456	599.6	509	5.55	0.384

Taking data from experiment RIT-136 as a base case, different examples have been analyzed changing the range and number of the uncertain input variables and reconstructing the probability distribution of the maximum wall temperature.

4.1.1. RIT-C1: Uncertain mass flux. Unimodal output.

First, we study the uncertainty propagation when only one input variable is considered as uncertain. In particular, we assume that the inlet mass flux can be described as a Gaussian distribution with mean 0.34456 kg/s and a standard deviation of 0.025 kg/s. As output variable to be studied we use the maximum wall temperature reached in the pipe. This example will be referred as RIT-C1.

In order to reconstruct the output variable probability distribution we obtained three populations of size 100, 500 and 40000 for the inlet mass flux and by executing RELAP5 we obtained their corresponding populations for the maximum wall temperature. For all of these populations the output variable follows an unimodal distribution, shown in Fig 3., that can be fitted using a Gaussian distribution, with the parameters shown in Table 2. As the values of the parameters exposed in Table 2 are quite close, a sample of size 100 is sufficient to approximately describe the maximum wall temperature distribution.

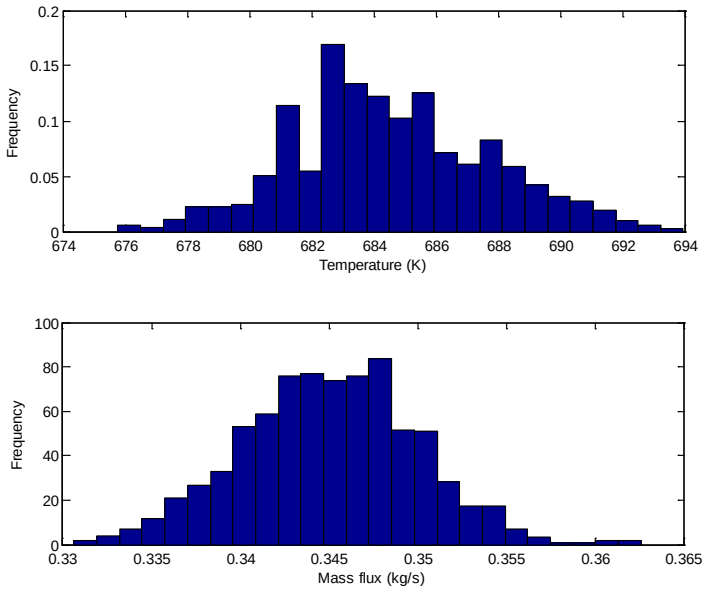

Fig 3. RIT-C1 maximum wall temperature and Gaussian mass flux histograms.

Table 2.

RIT-C1 parameters of the Gaussian maximum temperature distribution for different sample sizes.

Sample size	Mean	Standard Deviation
100	684.42	3.41
1000	684.53	3.19
4000	684.73	3.22

4.1.2. RIT-C2 Case. Uncertain mass flux. Multimodal output.

If the variance of the inlet mass flux distribution is increased, the maximum temperature distribution becomes multimodal. In this example, referred as RIT-C2, we have considered that the input variable follows a Gaussian distribution N(0.34456, 0.05) kg/s. The histograms associated with the output and the input variables are shown in Fig 4.

This is due to the fact that there are some values of the mass flux that do not lead to a dry-out in the heated zone (see Fig.1 and Fig. 2), the heat is properly removed and the pipe is always full of water, so the temperature remains at the initial temperature during all the transient, which correspond to the values near 620 K in Fig. 4. If the mass flux value leads to a dry-out in the heated zone the wall temperature rises, obtaining values of temperature between 660 K and 740 K, as it is also observed in Fig 4.

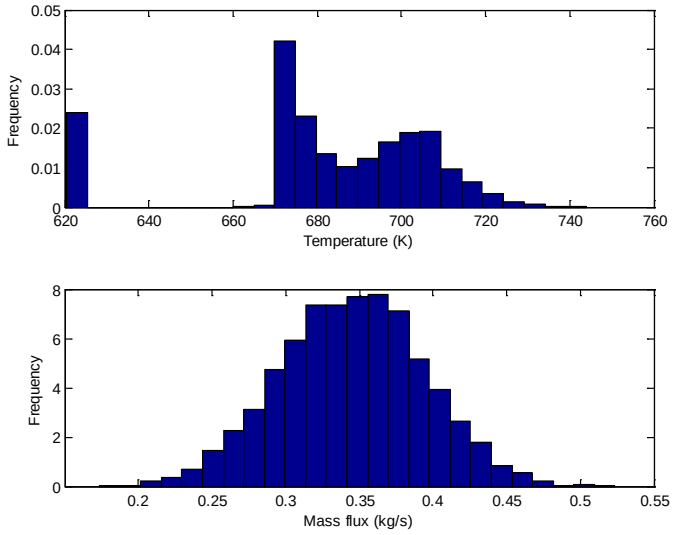

Fig 4. RIT-C2 maximum wall temperature and Gaussian mass flux histograms.

As can be observed in Fig 4, the histogram obtained for the maximum temperature presents several maxima, and the probability distribution followed by this variable is

multimodal. To describe this distribution, a mixture of three Gaussians is used, which is the optimal number of Gaussian distributions provided by the EM algorithm.

$$f_i(y) = \pi_1 \frac{1}{\sigma_1\sqrt{2\,\pi}}\, e^{-\frac{(y-\mu_1)^2}{2{\sigma_1}^2}} + \pi_2 \frac{1}{\sigma_2\sqrt{2\,\pi}}\, e^{-\frac{(y-\mu_2)^2}{2{\sigma_2}^2}} + \pi_3 \frac{1}{\sigma_3\sqrt{2\,\pi}}\, e^{-\frac{(y-\mu_3)^2}{2{\sigma_3}^2}} .$$

To reconstruct the probability distribution two methods will be used. In the first one, the finite mixture parameters will be obtained by means of the EM algorithm, explained in section 2. To use this method three population sample sizes of 100, 1000 and 4000, respectively, have been considered. The parameters of the reconstructed mixture with the EM method are shown in Table 3.

Table 3.
RIT-C2 parameters of the finite mixture of Gaussian distributions obtained with the EM algorithm considering different sample sizes.

Sample size	Probabilities (π_1, π_2, π_3)	Means (μ_1, μ_2, μ_3)	Standard Deviations (σ_1, σ_2, σ_3)
100	(0.13, 0.37, 0.50,)	(620.65, 674.50, 700.33,)	(0.06, 1.94,11.52)
1000	(0.11, 0.32, 0.57)	(620.68,673.78, 697.91)	(0.03,1.78,12.51)
4000	(0.12, 0.27, 0.61)	(620.68, 673.92,697.92)	(0.01, 1.73, 13.51)

Fig. 5. shows the histogram of the maximum wall temperature distribution and its reconstruction using the finite mixture model obtained by the EM algorithm when a sample of 4000 runs are considered.

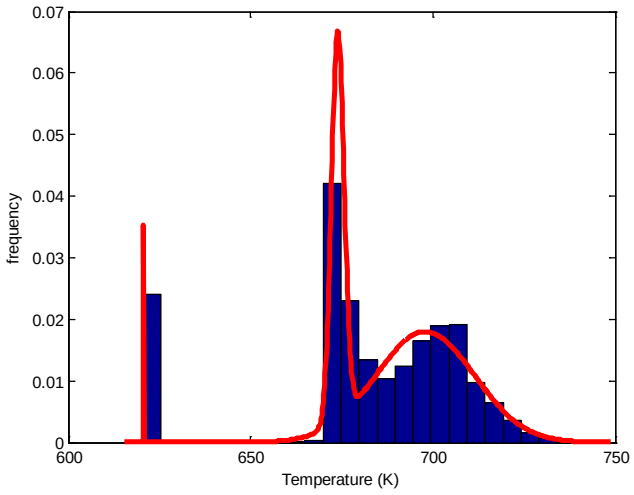

Fig. 5. RIT-C2 case histogram and EM reconstruction of the maximum temperature distribution.

From the parameters found for the mean and standard deviation in the finite mixture model, it is possible to divide the population of the output variable into three different clusters. Considering into a cluster those values of temperature whose distance to a given mean is less than three times its corresponding standard deviation. That is, cluster c_i satisfies,

$$c_i = \{y : |y - \mu_i| \leq 3\sigma_i\}.$$

This clustering in the output variable induces the corresponding clustering in the input variable, which is the mass flux measured in kg/s, shown in Table 4.

Table. 4. RIT-C2 clusters for mass flux, using the finite mixtures obtained with the EM algorithm.

Sample size	Cluster1	Cluster 2	Cluster 3
100	[0.24, 0.40]	[0.35, 0.40]	[0.40, 0.52]
1000	[0.20, 0.40]	[0.35, 0.40]	[0.40, 0.49]
4000	[0.20, 0.40]	[0.35, 0.40]	[0.40, 0.47]

The second method used to reconstruct the probability distribution of the output variable is based on the use of the k-means algorithm, described in section 3. Once the clusters, c_i, are obtained the mean, standard deviation and the probability weights of the finite mixture model are computed using expressions (20), (21) and (22), respectively.

In order to compare the results obtained with both methods, three clusters are preselected for the k-means algorithm, as this is the optimal number of Gaussians provided by the EM algorithm.

The different clusters performed by the k-means algorithm in the maximum temperature induce a clustering in the input variable (mass flux in this case of application) presented in Table 5. Comparing Table 4 and Table 5 we can observe that both methodologies provide similar results for case RIT-C2.

Table 5. RIT-C2 clusters for mass flux obtained with the k-means algorithm for different sample sizes.

Sample size	Cluster 1	Cluster 2	Cluster 3
100	[0.24, 0.33]	[0.34,0.40]	[0.40, 0.52]
1000	[0.20, 0.34]	[0.34, 0.40]	[0.40, 0.52]
4000	[0.17, 0.33]	[0.33, 0.40]	[0.40, 0.52]

To reconstruct the maximum wall temperature distribution, the parameters of the finite mixture model are obtained using expressions (27), (28) and (29) for each one of the

clusters predicted by the k-means algorithm. Table 6 presents the parameters obtained with the k-means algorithm for this case of application. Fig. 6 shows the histogram and the reconstructed distribution obtained with the parameters of Table 6.

Table 6. RIT-C2 parameters of the finite mixture of Gaussian distributions obtained from the k-means algorithm using different number of samples.

Number of samples	Probabilities (π_1, π_2, π_3)	Means (μ_1, μ_2, μ_3)	Standard Deviations (σ_1, σ_2, σ_3)
100	(0.14, 0.46, 0.40)	(620.65, 676.40, 704.60)	(0.06, 4.46, 8.53)
1000	(0.12, 0.47, 0.41)	(620.67, 676.69, 704.18)	(0.03, 5.14, 8.50)
4000	(0.12, 0.46, 0.42)	(620.68, 677.49, 704.98)	(0.02, 5.61, 8.50)

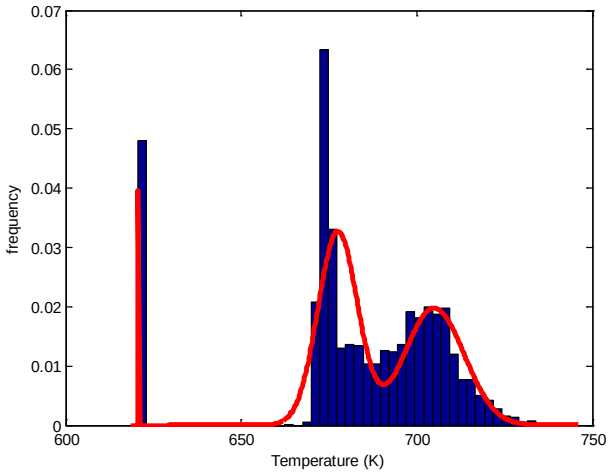

Fig. 6. RIT-C2 case histogram and k-means reconstruction of the maximum temperature distribution.

It can be observed that the reconstructions provided by EM algorithm and k-means algorithm are slightly different, see Fig.5 and Fig. 6. In particular, for the second Gaussian distribution obtained the k-means algorithm predicts a larger standard deviation, and the mean is shifted to a higher value, as can be observed comparing Tables 3 and 6.

4.1.3. RIT-C3 Case. Uncertain mass flux and power.

A similar study has been developed assuming that there are two uncertain input variables, mass flux (kg/s) and power (W), which are considered to follow Gaussian distributions N(0.344, 0.025) and N(23826.20, 1191.31), respectively. This example is referred as RIT-C3 case. The histograms associated with the output variable, which is the maximum wall temperature, and the input variables, mass flux and power, are shown in Fig 7. It is observed that, in this case the probability distribution of the output

variable is also a multimodal distribution. To reconstruct this distribution a finite mixture of three Gaussian distributions is used.

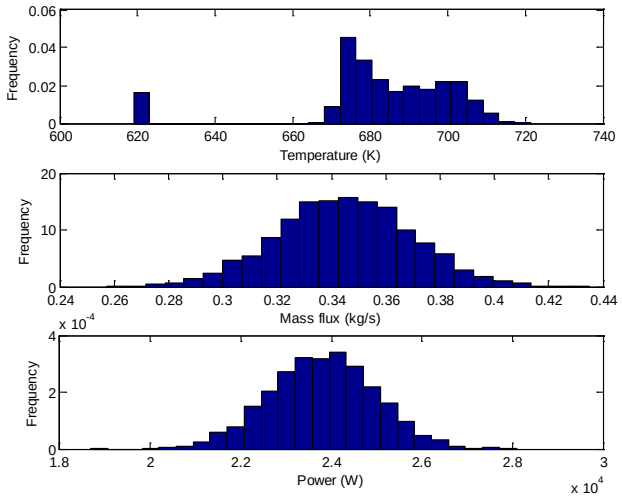

Fig. 7. RIT-C3 maximum wall temperature distribution, mass flux and power histograms.

The parameters of the reconstructed finite mixture using the EM method are shown in Table 7, for different sample sizes.

Table 7.
RIT-C3 parameters of the finite mixture of Gaussian distributions computed with the EM algorithm for different samples sizes.

Sample size	Probabilities (π_1, π_2, π_3)	Means (μ_1, μ_2, μ_3)	Standard Deviations ($\sigma_1, \sigma_2, \sigma_3$)
100	(0.04, 0.53, 0.43)	(620.23, 677.97, 697.41)	(0.18, 3.63, 7.33)
1000	(0.06, 0.42,0.52)	(620.23, 676.02, 695.49)	(0.19, 3.19, 8.63)
4000	(0.07, 0.42, 0.51)	(620.17, 676.65, 696.29)	(0.24, 3.60, 8.17)

Fig. 8 shows the histogram and its reconstruction using the finite mixture model obtained using a sample size of 4000.

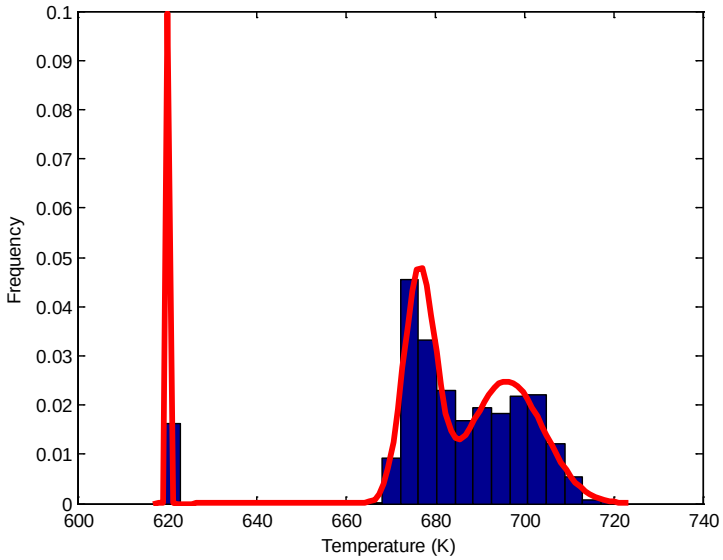

Fig. 8. RIT-C3 maximum temperature histogram and finite mixture reconstruction.

From the finite mixture model, we can obtain three clusters or intervals in the maximum temperature, corresponding with the three Gaussian distributions of the mixture. These intervals divide the plane mass flux-power in three regions where the maximum temperature follows a Gaussian distribution. Fig. 9 shows the different clusters in mass flux and power induced by the finite mixture obtained using a sample size of 4000. This previous classification of the regions of the possible values of power and mass flux is necessary if a classical uncertainty quantification for the maximum wall temperature is carried out.

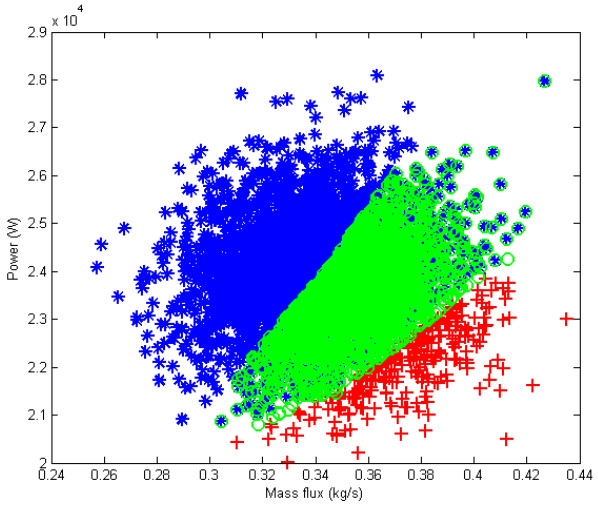

Fig. 9. Clustering in mass flux and power induced by the Gaussian finite mixture model in RIT-C3.

For the RIT-C3 case, the clustering in mass flux and power induced by the k-means algorithm for a sample size of 4000 is presented in Fig 10. This method provides more

independent clusters than the finite Gaussian mixture, as can be observed when comparing Fig. 9 and Fig 10.

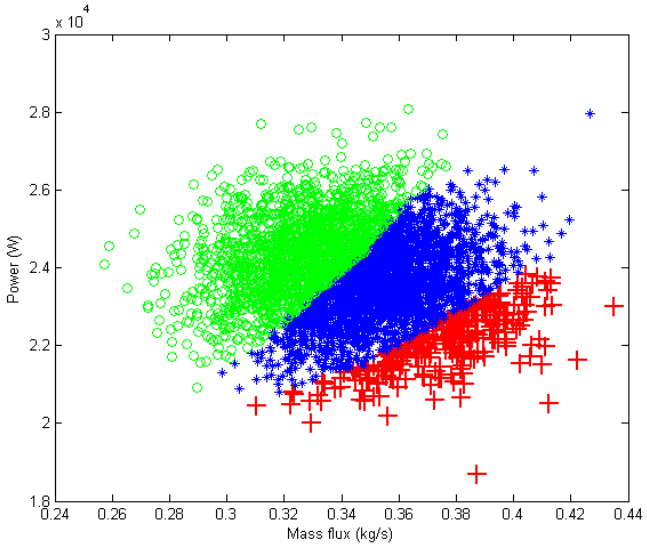

Fig. 10. Clustering in mass flux and power induced by the k-means in RIT-C3 case.

The parameters obtained for the finite mixture of Gaussian distributions computed form the clusters in the maximum temperature obtained with the k-means algorithm are shown in Table 8. In this Table, it can be observed that the parameters are quite similar for any sample size, so using the smallest sample size it would be sufficient to reconstruct the temperature probability distribution.

Table 8.

RIT-C3 parameters of the finite mixture of Gaussian distributions obtained from the k-means classification for different sample sizes.

Number of samples	Probabilities (π_1, π_2, π_3)	Means (μ_1, μ_2, μ_3)	Standard Deviations (σ_1, σ_2, σ_3)
100	(0.05, 0.56, 0.39)	(620.23, 678.46, 698.86)	(0.20, 4.00, 6.26)
1000	(0.06, 0.54, 0.40)	(620.23, 677.75, 699.01)	(0.19, 4.62, 6.31)
4000	(0.07, 0.53, 0.41)	(620.17, 678.28, 699.25)	(0.24, 4.81, 6.17)

The histogram of the maximum temperature and its reconstruction with the finite mixture of Gaussian distributions computed from the k-means classification when 4000 samples are used is shown in Fig. 11.

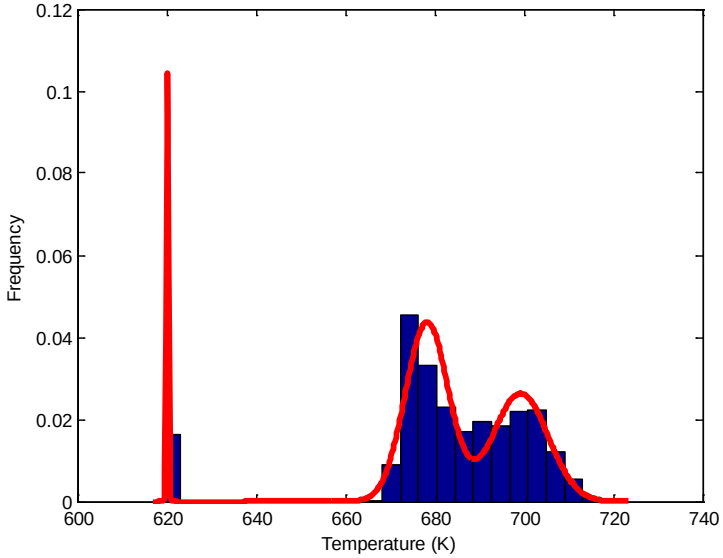

Fig. 11. RIT-C3 maximum temperature histogram and finite mixture reconstruction using k-means algorithm.

4. Pressurized Water Reactor 4” LOCA.

The model used in this analysis corresponds to a four loop Pressurized Water Reactor of 3600 MW of nominal power. The model constitutes one of the RELAP5 test cases recognized by typpwr.i. This input file is used to simulate a small break in the cold leg of a four loop PWR. In this particular model, three of the loops have been lumped, so the plant model used in the calculations constitutes a two asymmetric PWR. The loop containing the break is modeled as a single loop, and is presented in Fig. 12, but the other three loops are coalesced into one loop. The model also contains the steam generators secondary side of each of the two steam generators and the safety injections are modeled as boundary conditions. The transient is initiated by the break opening, after that a scram signal is activated, scram is produced and the primary pumps stop, and then safety injections actuate.

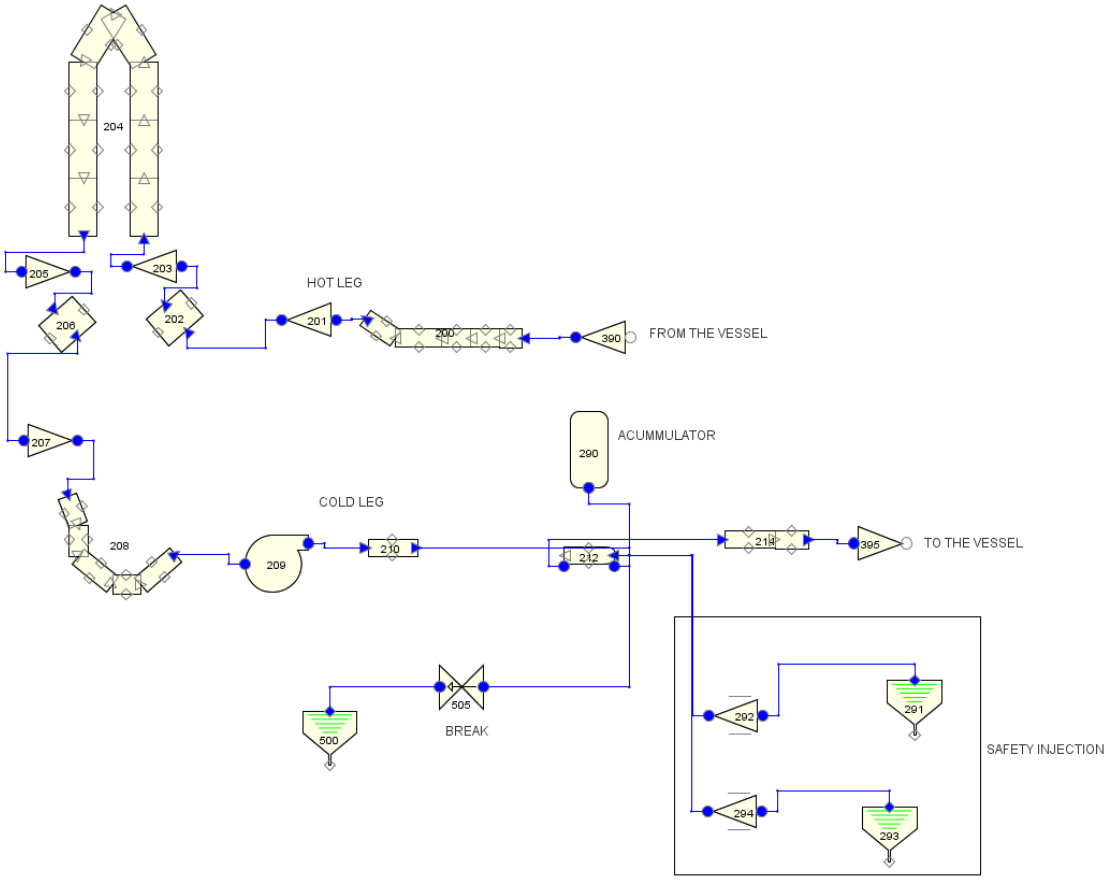

Fig. 12. Broken leg nodalization of the typical PWR model.

4.1. PWR-C1 Case: Uncertain break area.

In this case of application the input parameter considered uncertain is the break area, and it is supposed to follow a Gaussian distribution N(0.012,0.002) m^2, and the output variable to be studied is the Peak Cladding Temperature (PCT). This case of application is called PWR-C1. Fig. 13 shows the histograms associated with the maximum PCT and the break area considering a sample size of 4000.

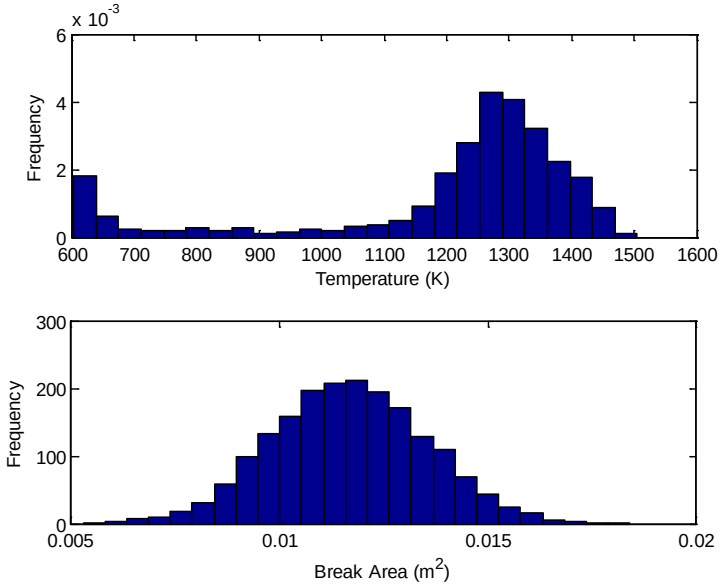

Fig. 13. PWR-C1 maximum wall temperature and break area probability distributions.

In Fig 13, it can be observed that the histogram of the temperature presents two clear maxima, what coincides with the optimal number of Gaussians predicted by the EM algorithm. The results obtained using the EM algorithm to fit the parameters of the mixture model, are presented in Table 9. The highest mean temperature value remains quite stable for all the sample sizes as well as its standard deviation. However, to get a stable value of the lowers mean more than 100 runs are needed. The Gaussian distribution associated with lower temperature values has a large standard deviation.

Table 9.
PWR-C1 parameters of the finite mixture of Gaussian distributions computed with the EM algorithm for different samples sizes.

Sample size	Probabilities (π_1, π_2)	Means (μ_1, μ_2)	Standard Deviations (σ_1, σ_2)
100	(0.20, 0.80)	(805.90, 1304.50)	(169.78, 78.51)
1000	(0.17, 0.83)	(748.50, 1299.30)	(163.63, 80.96)
4000	(0.17, 0.83)	(745.70, 1295.20)	(155.14, 81.90)

Fig. 14 shows the histogram of the PCT and its reconstruction for a sample size of 4000 and using a finite mixture of two Gaussians, with the parameters presented in Table 9. It can be observed that the model reconstructs quite accurately the peak of high values of temperature, but it fails to predict the distribution for low temperature values. This must be due to the asymmetry of the probability distribution of the temperature in the region of low temperatures. Such behavior can be explained by the value of the initial temperature value, which takes a value of 600 K, so if there is any run in which the

safety system injection could decrease this value it is not taken into consideration in this application as the value recorded is always the maximum one.

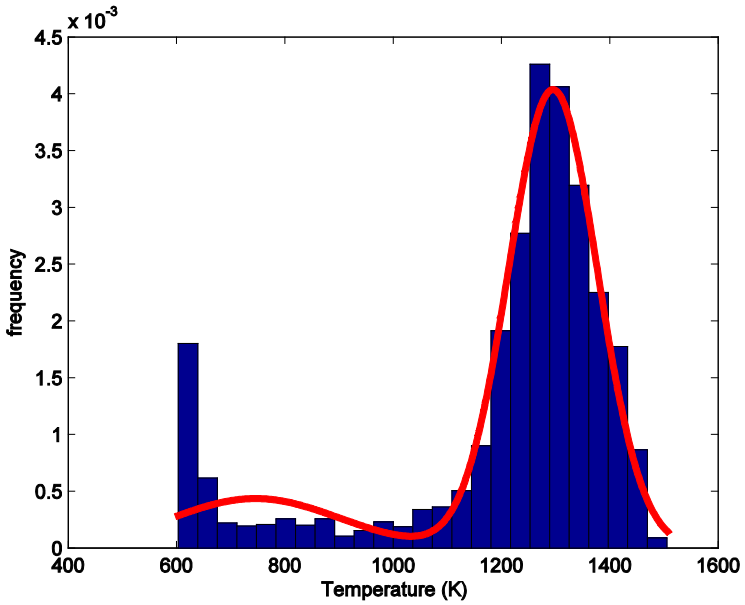

Fig. 14. PWR-C1 PCT reconstruction using the EM algorithm.

The clustering induced by the finite mixture model in the break area is presented in Table 10. In this case, both clusters are overlapped, since the standard deviation of the Gaussian obtained to predict the range of temperatures [600, 1000] is quite large (see Table 9) and this produces an overlapping of both clusters.

Table 10.
PWR-C1 break area clusters obtained using the EM algorithm.

Sample size	Cluster1	Cluster 2
100	[0.10, 0.17]	[0.07, 0.17]
1000	[0.06, 0.17]	[0.09, 0.20]
4000	[0.09, 0.20]	[0.05, 0.20]

The same problem is studied using the k-means method and assuming that the distribution can be reconstructed using two Gaussian distributions. The k-means algorithm provides the clusters in the break area values, presented in Table 11. It can be observed that also in this case the clusters predicted are not independent.

Table 11.
PWR-C1 break area clusters obtained using k-means algorithm.

Sample size	Cluster1	Cluster 2
100	[0.07, 0.15]	[0.10, 0.17]
1000	[0.06, 0.15]	[0.09, 0.20]
4000	[0.06, 0.12]	[0.09, 0.20]

The parameters obtained to reconstruct the PCT distribution using the k-means method are exposed in Table 12. And Fig. 15 shows the histogram and the reconstruction of the PCT distribution predicted by the k-means algorithm.

Table 12.
PWR-C1 parameters of the finite mixture of Gaussian distributions computed with the k-means algorithm for different samples.

Sample size	Probabilities (π_1, π_2)	Means (μ_1, μ_2)	Standard Deviations (σ_1, σ_2)
100	(0.17, 0.83)	(750.25, 1295.40)	(127.62, 90.03)
1000	(0.16, 0.84)	(701.47, 1292.80)	(116.56, 89.07)
4000	(0.15, 0.85)	(711.88, 1290.50)	(121.69, 87.86)

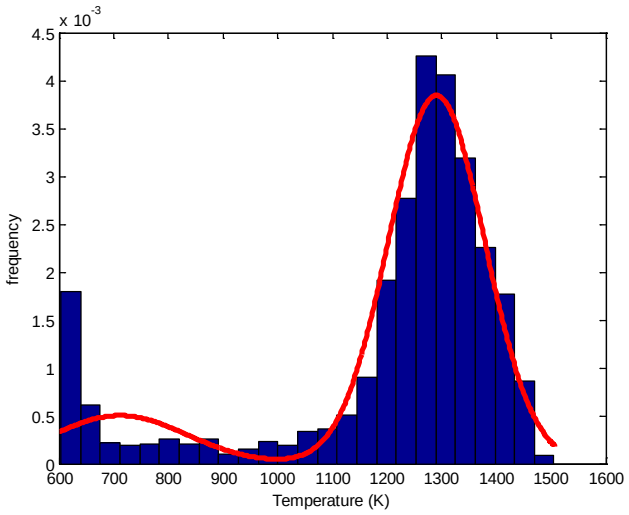

Fig. 15. PWR-C1 PCT reconstruction using k-means method for two Gaussians.

4.2. PWR-C2 Case: Uncertain break area and nominal power.

This second case of application considers the break area and nominal power as uncertain parameters for the LOCA transient. Assuming that the break area follows a distribution N(0.012,0.002) m^2, and the power is modeled as N(3600, 36) MW, the objective is to reconstruct the PCT distribution. This case is referred as PWR-C2. Fig. 16 shows the histogram of the PCT distribution together with the histograms of the

break area and the power values. All these histograms have been obtained with a sample size of 4000.

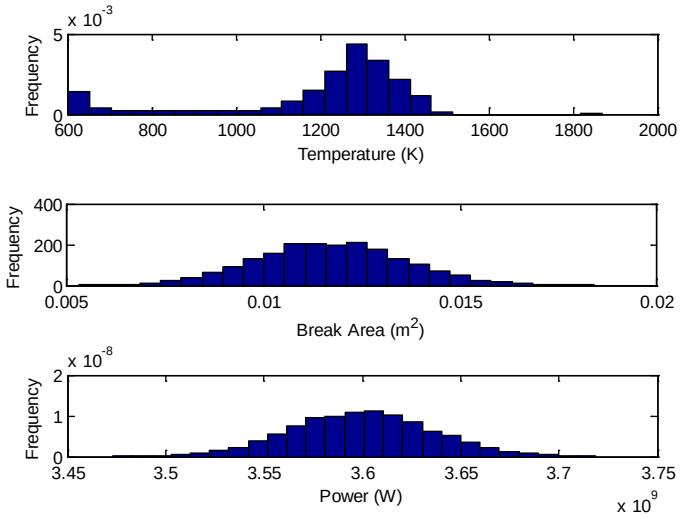

Fig. 16. PWR-C2. PCT, break area and power histograms.

The parameters of the finite mixture obtained with the EM algorithm are presented in Table 13 for different number of sample sizes, and Fig. 17 shows the histogram associated with of the PCT and its reconstruction using the finite mixture model, with the parameters of Table 13 for a sample size of 4000.

Table 13. PWR-C2 parameters of the Gaussian mixture for the PCT distribution for different number of sample sizes.

Sample size	Probabilities (π_1, π_2)	Means (μ_1, μ_2)	Standard Deviations (σ_1, σ_2)
100	(0.13, 0.87)	(780.10, 1307.50)	(167.35, 69.77)
1000	(0.17, 0.83)	(785.20, 1289.20)	(171.51, 79.98)
4000	(0.16, 0.84)	(736.40, 1291.10)	(139.78, 86.51)

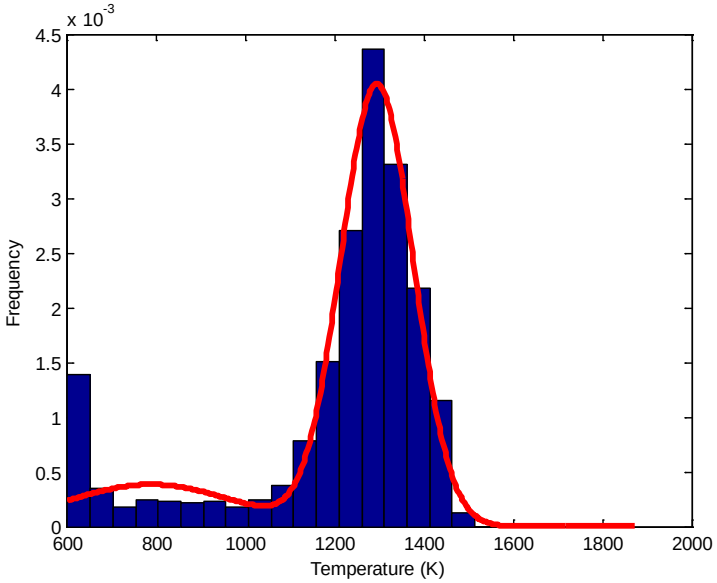

Fig. 17. PWR-C2 PCT probability distribution reconstruction using EM algorithm.

The finite mixture model obtained with the EM algorithm induces the clustering in the plane break area-power presented in Fig. 18, for a sample size of 4000. In the Figure an overlapping of both clusters is observed, due to the large standard deviation predicted for the first Gaussian distribution (see table 13).

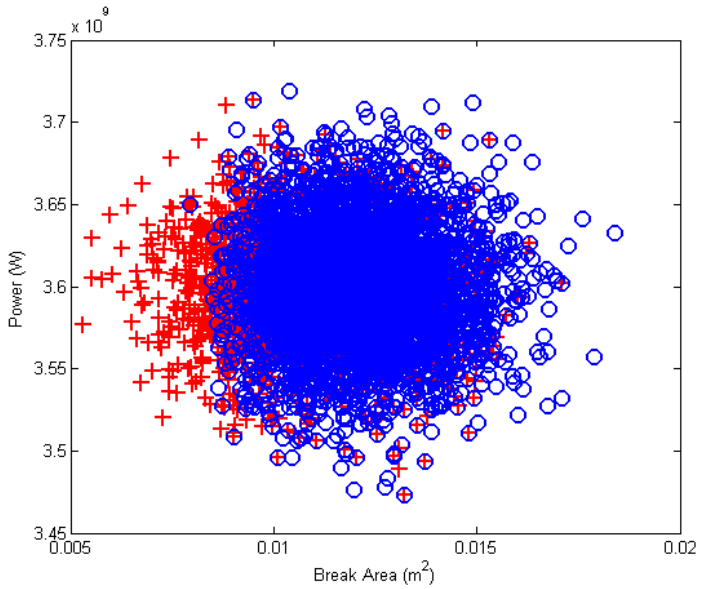

Fig. 18. PWR-C2. Clusters in break area and power induced by the EM algorithm.

Finally, the same problem was solved using the k-means method to reconstruct the PCT distribution considering the break area and the power as uncertain inputs. The clusters induced by the k-means algorithm for break area and power values are shown in Fig. 19. Also, with this method it can be observed an overlap of the two clusters associated with the two Gaussian distributions. Such situation is due again to the large standard deviation predicted for the Gaussian distribution obtained for low values of the

temperatures. Nevertheless, the values obtained using the k-means method are not so large as the ones predicted by the finite mixture model, as it can be observed by comparing the values presented in Table 13 and Table14.

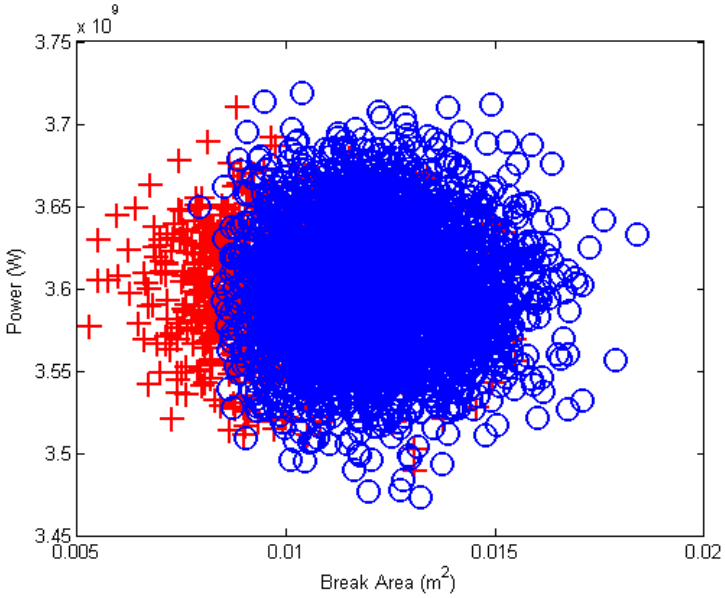

Fig. 19. PWR-C2. Clustering in break area and power induced by the k-means algorithm.

The parameters for the Gaussian mixture model obtained from the clusters computed with the k-means algorithm using different number of sample sizes are exposed in Table 14. Once again the Gaussian for high PCT values is quite stable, but this is not the case for low temperature values, and also the standard deviations in this cluster are quiet large.

Table 14. Parameters of the Gaussian mixture model obtained from the k-means method using different number of runs for case PWR-C2 case.

Number of runs	Probabilities (π_1, π_2)	Means (μ_1, μ_2)	Standard Deviations (σ_1, σ_2)
100	(0.12, 0.88)	(746.54, 1304.10)	(145.65, 75.58)
1000	(0.15, 0.85)	(739.89, 1284.20)	(131.98, 85.78)
4000	(0.15, 0.85)	(722.17, 1288.80)	(125.69, 89.34)

Fig. 20 shows the histogram for the PCT and its reconstruction with the k-means method using a sample size of 4000 for the PWR-C2 case.

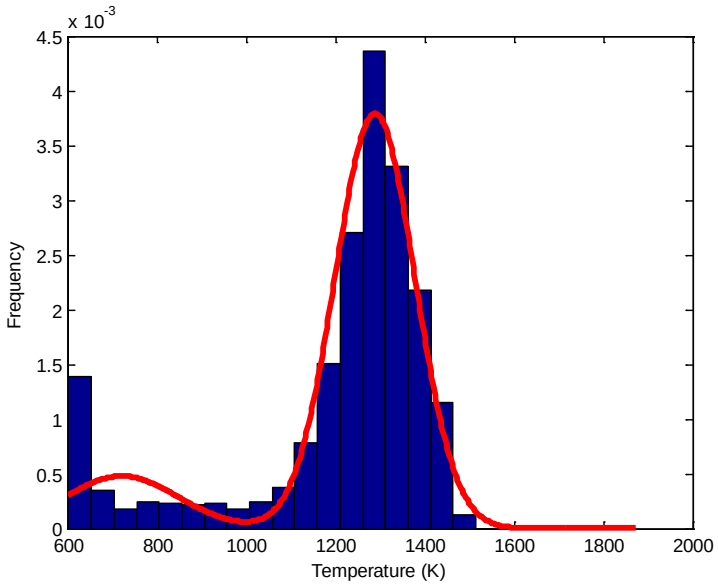

Fig. 20. PWR-C2. PCT probability distribution reconstruction using k-means algorithm.

We observe that the k-means method also fails to reconstruct the probability distribution for the region of low temperatures, where the probability distribution is asymmetric. In fact the PCT distribution reconstructed with the k-means is practically identical to the one reconstructed using the EM algorithm, as can be observed comparing Fig. 17 and Fig. 20.

5. Conclusions

As thermal-hydraulic simulations using best estimate codes play an important role in Nuclear Power Plants safety analysis, it is also important to analyze the uncertainty associated with such simulations, and its propagation to the code results. Great advances have been made during the last years using statistical techniques as the order statistics, which provides a tolerance/confidence interval of the output variable. However, there is not a complete knowledge of the output variable probability distribution. Finite mixture models have been successfully used to reconstruct the probability distribution of random variables, even if such probability distribution presents a multimodal behavior.

This paper presents two different methodologies to reconstruct the output variable probability distribution of a best estimate code using a Gaussian mixture model: The EM and the k- means algorithms. Both techniques have been applied in the study of the uncertainty of two typical applications. A separated effects problem has been analyzed using both methods, considering the maximum wall temperature as output variable, obtaining good agreement between the reconstructed distribution and the histogram. Moreover, using the finite mixture techniques it has been possible to cluster the input

variable domains induced by the value of the output variable, which is of great interest to estimate the value of the output depending on the value of the input parameters selected. The definition of the different clusters in the input variables, also allows assuming an approximate Gaussian behavior of the output variable inside each cluster, where the classical statistical analysis can be applied.

The EM and k-means methods have also be applied to an integral effects problem. In this case the reconstructed distribution agrees with the histogram for high temperature values but they fail to reproduce the low range of temperatures. This situation is due to the asymmetric behavior of the PCT given by boundary conditions. That is, there exist many situations in which safety systems are able to maintain the PCT at the initial value. In any case, the reconstructed PCT can give us an estimation of the probability of exceeding its limit value. In this application the clusters induced by the PCT are overlapped, so a deeper study on how to tackle this situation is needed. The results in this case of application could be improved if more general finite mixture models where asymmetrical distributions are considered.

Acknowledgements

This work has been partially supported by the Consejo de Seguridad Nuclear under the contract with reference STN/2369/08/640.