Probabilistic computation techniques

Appendix 10 to the Annex 1 to the
Draft Stage 1 Evaluation Report

July 2009

Contents

1 Introduction 1

2 Principles of probabilistic flood computations 2

2.1 Introduction 2

2.2 Example 1: the univariate case 2

2.3 Example 2: the multivariate case 4

2.3.1 Introduction 4

2.3.2 The need for classification 4

2.3.3 Derivation of damage functions 6

2.3.4 Identification of the *T*-year flood event 7

2.3.5 Probability of exceedance for an individual threshold value 8

2.4 Flood hazard versus flood risk 17

1 Introduction

In this Appendix 10 to Annex 1 background to the probabilistic computations is given. It serves as an introduction for those who are not familiar with probabilistic computation methods. For actual applications of these techniques in the focal areas the reader is referred to:

Annex 1: Nam Mae Kok
Annex 3: Se Bangfai
Annex 6: Mekong Delta

2 Principles of probabilistic flood computations

2.1 Introduction

The magnitude of floods in the focal areas is related to a combination of influencing factors. The influencing factors can be different for different locations. For the focal areas downstream of Phnom Penh these factors are peak flow and flow volume of the Mekong River. For areas along Mekong tributaries the influencing factors are the flow in the Mekong and the flow in the tributaries.

Since these influencing factors vary from year to year, i.e have a random nature, we will refer to them as *random variables*. Because of the random nature, flood probabilities in the focal areas can only be quantified through the application of a *multivariate probabilistic* modelling approach. "Multivariate" refers to the fact that more than one random variable is involved. "Probabilistic" refers to the fact that these factors have a random nature.

This chapter serves as an introduction for those who are not familiar with probabilistic computation methods. The purpose of these methods is to derive flood probabilities or, reversely, conditions related to the *T*-year flood, i.e. the flood conditions which have a probability of 1/T per year to be exceeded in a single year. For T, generally a range of values like 5, 10, 20, 25, 50 or 100 years are evaluated.

For the sake of clarity we use simplified examples in this chapter to demonstrate the principles of probabilistic computation. We emphasise that the numbers used in these examples are hypothetic and used with the sole purpose to demonstrate probabilistic computation methods.

2.2 Example 1: the univariate case

The univariate case is the situation where the severity of the flood is determined by only one random variable. This may be for instance the (annual) peak river discharge or the flow volume in the flood season. If we take the example of the peak discharge, it means an increase in the peak discharge automatically implies an increase in the severity of the flood event. The derivation of the 100-year flood event then "only" comes down to the derivation of the 100-year peak river discharge, $Q_P(100)$. The 100-year discharge is defined as follows:

$Q_P(100)$ is the discharge level which has a probability of being exceeded equal to 1/100 each year.

In general, the T-year discharge is defined as follows:

$Q_P(T)$ is the discharge level which has a probability of being exceeded equal to 1/T each year.

High values of T refer to extreme events. So, the higher the value of T, the higher the value of the peak discharge, $Q_P(T)$. In this example, where peak discharge Q_P is the only variable involved, we need to derive the function $Q_P(T)$. Since the return period, T, is directly related to the probability, function Q(T) is referred to as a *probability distribution function*. Alternatively the name *frequency distribution function* is often used as well, since the probabilities and frequencies are strongly related.

Generally these type of distribution functions are derived from *extreme value analysis*. Most commonly, strictly statistical methods are used in extreme value analysis and a parameterised probability distribution function is fitted to a set of observed extremes of the target variable. Figure 2.1 shows an example of the fit of the so-called "Generalised Extreme Value (GEV) distribution function" on observed peak discharges of the Mekong river at location Stung Treng. The vertical axis shows possible values of the annual peak discharge at Stung Treng, the horizontal axis shows its probability of exceedance. The dots represent observed annual maximum flows at Stung Treng, the line is the fitted distribution function.

Once this distribution is available it is very straightforward to derive exceedance probabilities for any threshold value of the peak river discharge and, vice versa, river discharges that correspond to a certain probability of exceedance. So for instance the 100-year peak river discharge can be derived directly from this Figure by verifying where the distribution function crosses the line x = 1/100 (i.e. $x = 10^{-2}$). In this example the 100-year peak discharge is approximately equal to 75,000 m^3/s. Similarly it can be derived that the 1/10 year (10^{-1}) discharge is approximately equal to 67,000 m^3/s.

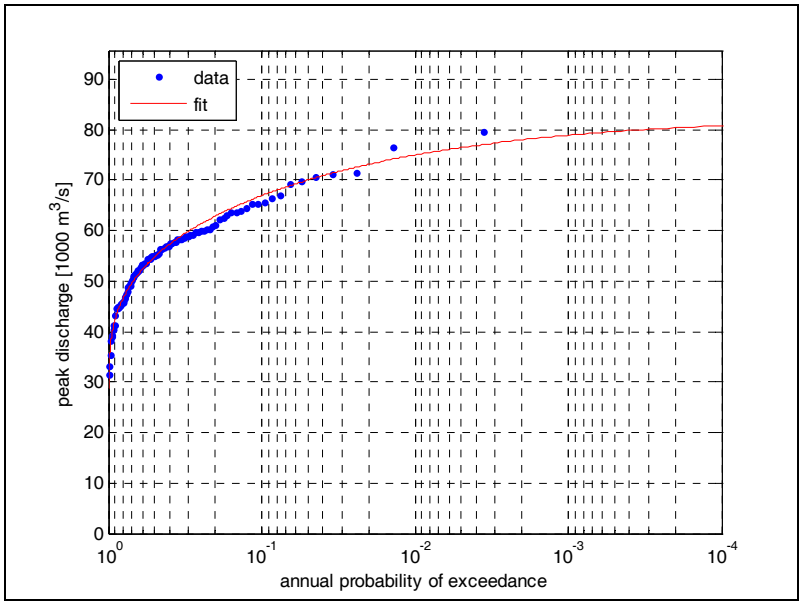

Figure 2.1 Example of a frequency distribution function of river discharge along the Mekong

2.3 Example 2: the multivariate case

2.3.1 Introduction

If more than one random variable is involved the problem is characterised as "multivariate". As we will demonstrate this introduces some complexities that are not present in the univariate case. For the sake of clarity the example in this section will be bivariate, i.e. 2 random variables are involved. However, the problems and solutions addressed also apply to situations with more than two variables involved.

2.3.2 The need for classification

In the univariate case the classification of the "severity" of the flood was straightforward. In the example of the previous section the severity was directly related to the peak river discharge, i.e. an increase in the peak discharge automatically implies an increase in the severity of the flood event. For multivariate problems this is not the case anymore as will be demonstrated in the following example with two random variables: peak discharge and total flow volume (above a certain discharge threshold). Figure 2.2 shows 4 hypothetic realisations of these two variables. In comparing these 4 events (years) the following observations can be made:

- event 4 is more severe than event 3, because the peak discharge in event 4 is higher
- event 2 is more severe than event 1, because the peak discharge in event 2 is higher
- event 4 is more severe than event 2, because the flood volume in event 4 is larger
- event 3 is more severe than event 1, because the flood volume in event 3 is larger

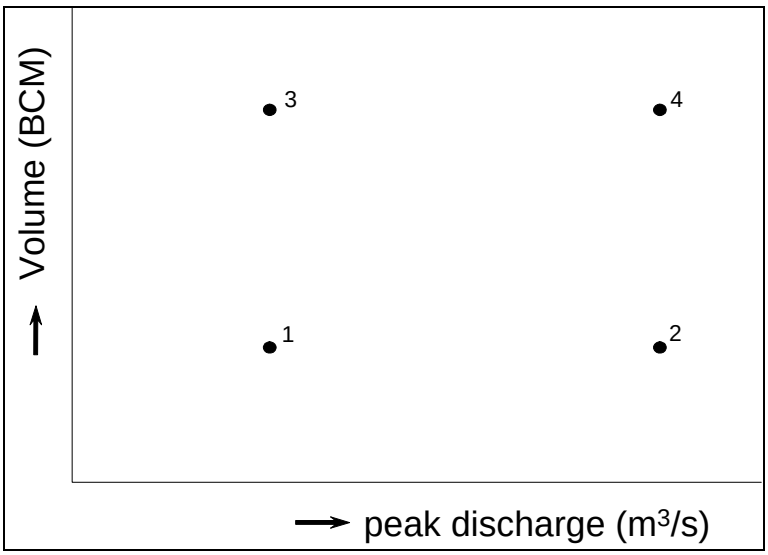

Figure 2.2 Four hypothetic events (years), characterised by realisations of the peak river discharge and the flow volume in the flood season.

The main question that remains is which of the following two events is more severe: event 2 or event 3? Event 2 has a higher peak discharge, while event 3 has a larger flow volume. This means we need additional information and criteria to be able to rank these two events

and subsequently be able to derive the T-year flood event (e.g. T equal tot 5, 20 or 100 years). Generally a "damage" function is used for this purpose. This damage function may represent a combination of loss of crops, property, infrastructure and economic productivity. Even casualties may be taken into account, although it is always hard to quantify the economic value of casualties. In the FMMP we will also consider a much simpler "damage" function: the maximum water depth in the flooded areas. If the maximum water level is used as criterion we speak of "flood hazard analysis", if economic damages are considered we speak of "flood risk analysis".

Figure 2.3 shows a hypothetic example of a damage function. The horizontal axis shows the peak discharge and the vertical axis the flow volume. The contour lines show the damage as a function of these two variables. If the peak river discharge is moderate or low (left side of Figure 2.3) no flooding occurs and consequently the damage is equal to zero. On the other hand for large peak discharges in combination with large flow volumes, there will be damage. Clearly the damage increases with increasing values of volume and peak discharge.

Figure 2.4 shows a similar plot, but in this case the "damage"-function is simply the maximum water depth at a certain location in a focal area. Again, for low peak discharges no flooding occurs, which means the water depth at this location is equal to zero. For larger peak discharges the water depth increases with increasing values of volume and peak discharge. Note: this is a simplified example, and that is why the coloured contour lines in both figures are linear. In "reality" this is generally not the case.

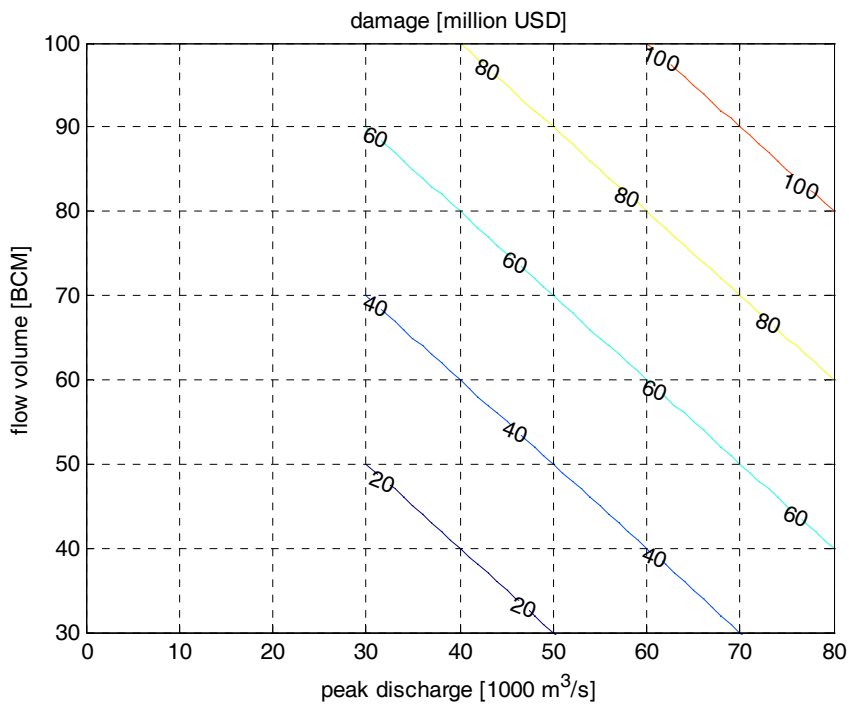

Figure 2.3 Hypothetic example of damage (contour lines) due to flooding as a function of the peak discharge and flow volume.

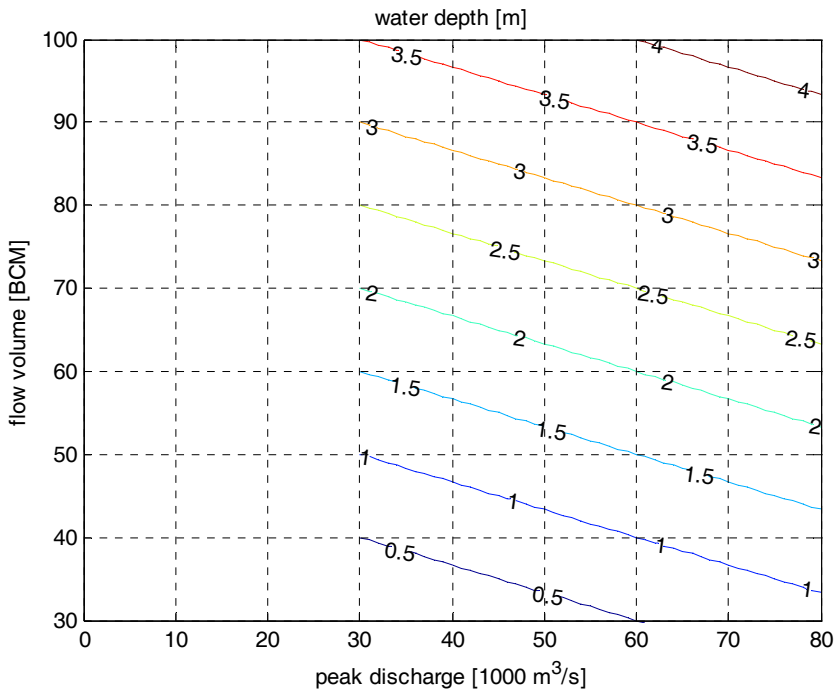

Figure 2.4 Hypothetic example of maximum water depth at a certain location in focal area (contour lines) as a function of the peak discharge and flow volume.

It is very important to note that plots like Figure 2.3 and Figure 2.4 will be different for different locations or areas. In areas where the water depth mainly depends on the flow volume the contour lines will be nearly horizontal, whereas in areas where the water depth mainly depends on the peak discharge the contour lines will be nearly vertical. The ranking of events will therefore differ form location to location and the same holds for the T-year flood event! Probabilistic flood hazard computations therefore generally need to be executed separately for different locations or areas.

2.3.3 Derivation of damage functions

In the previous section it was demonstrated for the multivariate case that:

- a criterion is required to be able to rank events and to identify the T-year flood event;
- the *T*-year flood event can be expected to differ for different locations.

With respect to the first bullet, a damage function often is used to assess the severity of flood events. This damage function translates the random variables like peak discharge and flow volume into damages. In order to derive such a function two types of analysis are required:

1. hydraulic analysis (and simulations) to translate flow characteristics into flood depth;
2. economic analysis to translate flood depth into damages.

These analyses are depicted in the flow chart of Figure 2.5. In this example, the starting point of the hydraulic analysis is a flow hydrograph. If more than one river is involved, multiple hydrographs are required. Based on hydraulic analysis of the river system and potential

flooded areas a simulation model is produced. The simulation model uses the flow hydrographs as input and computes the flood depth (as a function of time) at various locations. Subsequently economic damage as a result of the flood is based on an ecomic survey of, among others, crops, property, infrastructure and economic productivity.

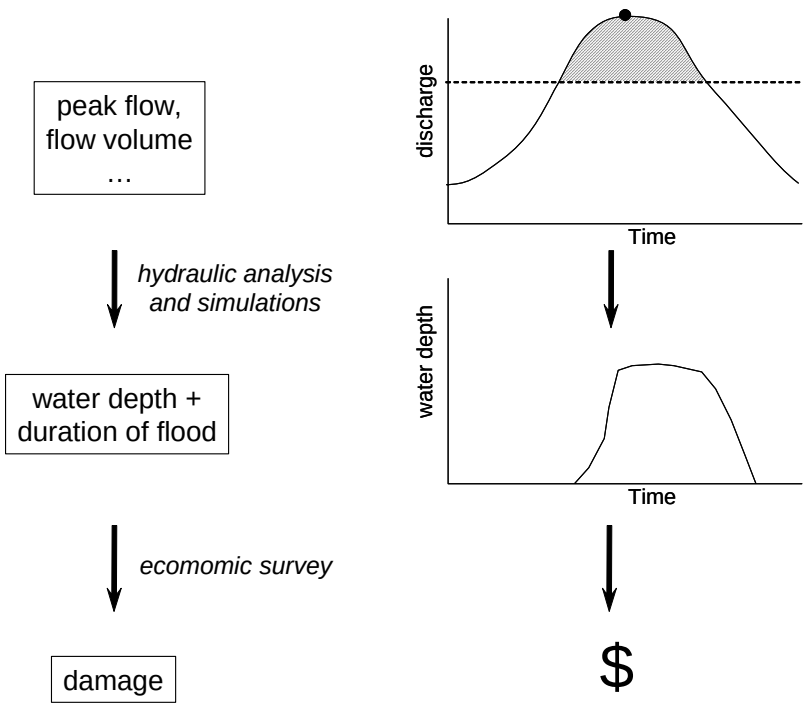

Figure 2.5 Schematic view of the derivation of damage as a function of flow characteristics.

2.3.4 Identification of the *T*-year flood event

In this section it is assumed that there is a damage function, *D*, available for the location(s) of interest. The *T*-year flood event is then related to:

"the threshold damage D_T which has a probability of exceedance of p=1/T per year"

In order to derive D_T a relation between damage, *D*, and probability of exceedance, *p*, is required. The common approach is to iteratively derive the probability of exceedance of a number of well-chosen threshold values of damage function *D*: D_1, D_2 … D_k. For example in case of Figure 2.3 these threshold values may be 5, 10, 15, …,100 (million USD). Figure 2.6 shows the possible outcome of such an analysis (section 2.3.5 explains how these numbers can be obtained). The squares show the computed probabilities for the individual threshold values of *D*. The dotted line is a fit through the squares and with this line the relation can be reversed: for any probability of exceedance the corresponding damage can be obtained. So for example the 100-year damage D_{100} can simply be derived by checking were the dotted line crosses the horizontal threshold of 1/100. In the example of Figure 2.6 this occurs at around 70 million USD, so D_{100} =70 million USD.

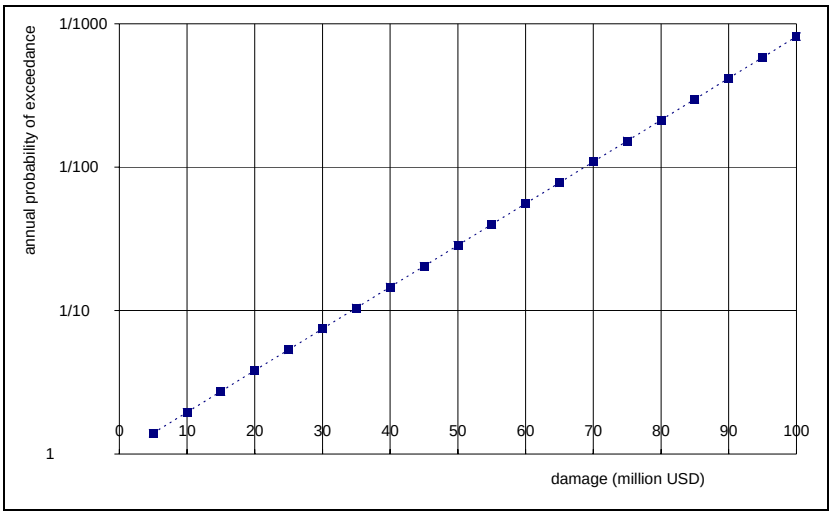

Figure 2.6 Hypothetic relation between damage and annual probability of exceedance

2.3.5 Probability of exceedance for an individual threshold value

We will now describe how the frequency of exceedance of a single threshold value, *D**, of damage function *D* is determined (i.e an individual dot in Figure 2.6). Suppose the threshold value *D** equals 80 (million USD). This amount is represented by the yellow line in Figure 2.3. The area to the upper right of this yellow line consists of all combinations of peak discharges and volumes for which the resulting damage exceeds *D**. This is also schematically depicted in Figure 2.7.

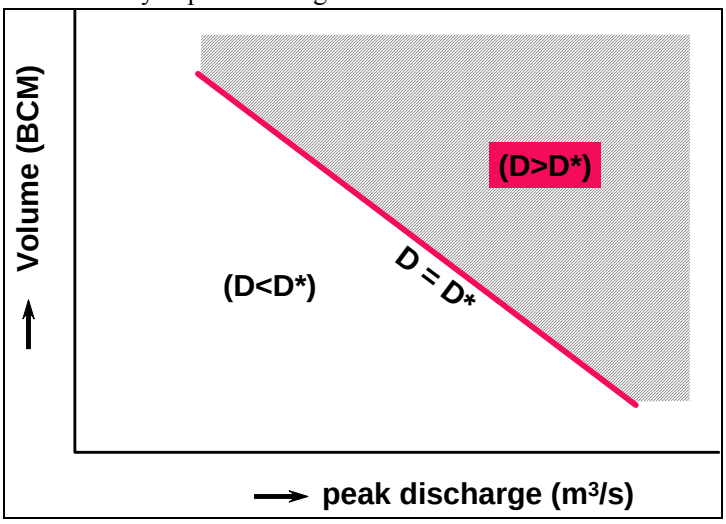

Figure 2.7 Schematic representation of the total damage flooding as a function of the peak discharge and flow volume. The shaded area represents combinations for which the damage exceeds the threshold value of D* million USD.

The red line in Figure 2.7 connects all combinations of peak discharge and volume that will lead to a damage of exactly *D**. The shaded area shows all combinations that lead to a damage that exceeds *D**. This is the area of interest since the probability of exceedance of *D** needs to be derived. It is therefore relevant to derive the location of this area and/or the red line that divides the area where D> *D** from the area where D≤ *D**. This may seem

straightforward at first, but generally that is not the case. In order to derive Figure 2.7 a large number of hydrodynamic model simulations need to be executed for different combinations of peak discharges and volumes. Furthermore, an economic survey of the area of interest is required (see section 2.3.3).

Figure 2.8 shows an example of a 2D-grid of combinations which may be selected to execute hydrodynamic model simulations and thus explore the damage function as a function of peak discharges and volumes. After executing the hydrodynamic simulations and economic analysis, the damages at the grid points is are known. In order to estimate the position of the red line these results of the grid points need to be interpolated on a 2D-grid.

In order to derive the probability of exceedance of D^* we need a procedure that basically performs the following two tasks:

1. find the "failure domain", i.e. the area where D>D*;
2. determine the cumulative probability of occurrence of all combinations in the failure domain.

This shows that the probabilistic approach consist of a combination of deterministic computations (step 1) and statistic computations (step 2). We demonstrate the concept by applying it to a very simple problem: a game of two dices. Although there seems to be no relation between floods and dices, the basic probabilistic approach is actually very similar. The reason to show the example of two dices is that it makes the basic concept of probabilistic computation easy to understand.

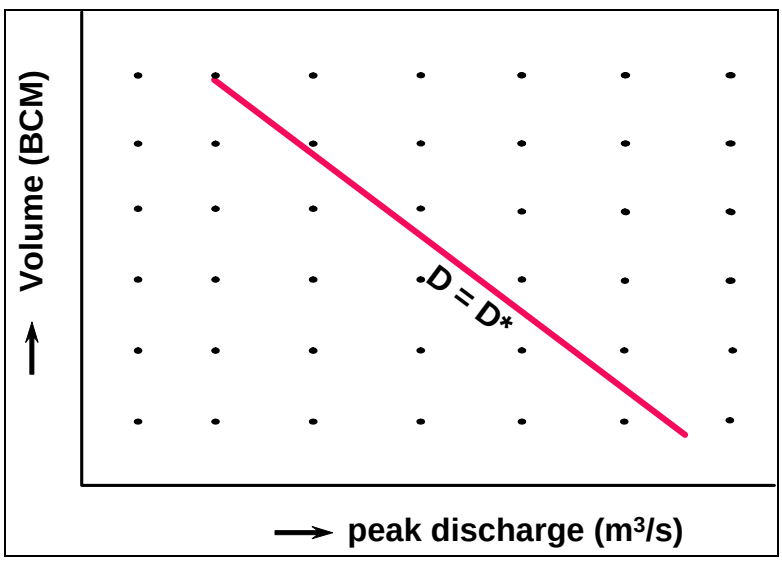

Figure 2.8 Schematic representation of the 2D grid of hydrodynamic computations.

In the example we need to derive the probability that the sum of the outcome of two dices equals or exceeds a total of 10. So the sum of the two dices can be considered as the "damage function" D, and the objective is to derive $P(D \geq 10)$. Figure 2.9 shows the possible outcomes of the two dices. The horizontal axis shows the set of possible outcomes of the first dice (1, 2, 3, 4, 5 or 6), the second axis shows the same for the second dice. Notice the similarity between Figure 2.9 and Figure 2.8!

The total number of combinations shown is 6×6=36, which means each combination has a probability of occurrence of 1/36. The red triangle shows all 6 combinations for which ($D \geq 10$). Since each combination has a probability of occurrence of 1/36 it can be directly seen that $P(D \geq 10) = 6/36 = 1/6$.

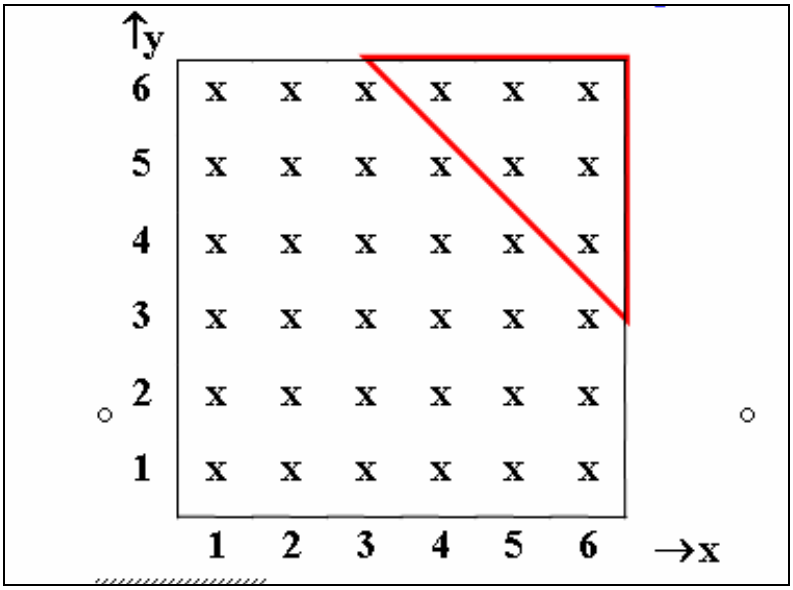

Figure 2.9 Possible outcomes of two dices x and y

Now for the random variables in the Mekong basin basically the same procedure is applied as in the example of the two dices. However, the actual implementation of the procedure is more complex and time consuming because of the following reasons:

- in some cases more than 2 random variables are involved;
- probability distribution functions are generally continuous and more complex than for dice;
- random variables in nature are often correlated;
- time-consuming hydraulic model runs need to be executed.

For these reasons it is generally not possible to find an exact analytical solution for these type of problems. Therefore, the probability of exceedance has to be based on approximation techniques. There are several methods available for this purpose, such as:

1. *numerical integration*
2. *crude Monte Carlo*;
3. *FORM (First Order Reliability Method)*..

Below these methods are briefly described.

Numerical integration

This method explores the *n*-dimensional space of possible realisations (where *n* equals the number of random variables) by dividing it into a number of discrete *n*-dimensional

"cubicals" and related grid points. In two dimensions the grid will look somewhat like the dots in Figure 2.10. Each gridpoint represents a combination of realizations of the *n* random variables. For each gridpoint the following is computed:

a) the probability of occurrence;
b) whether or not the grid point is in the "failure domain" i.e. leads to an exceedance of the threshold damage *D**.

The probability of exceedance of D* is equal to the sum of probabilities of all grids in the failure domain. In Figure 2.10 the gridpoints in the failure domain are circled in red. For each combination in the failure domain the probability of occurrence is computed, based on the combined distribution function of the underlying random variables. The sum of these probabilities is the estimated probability of exceedance of D*..

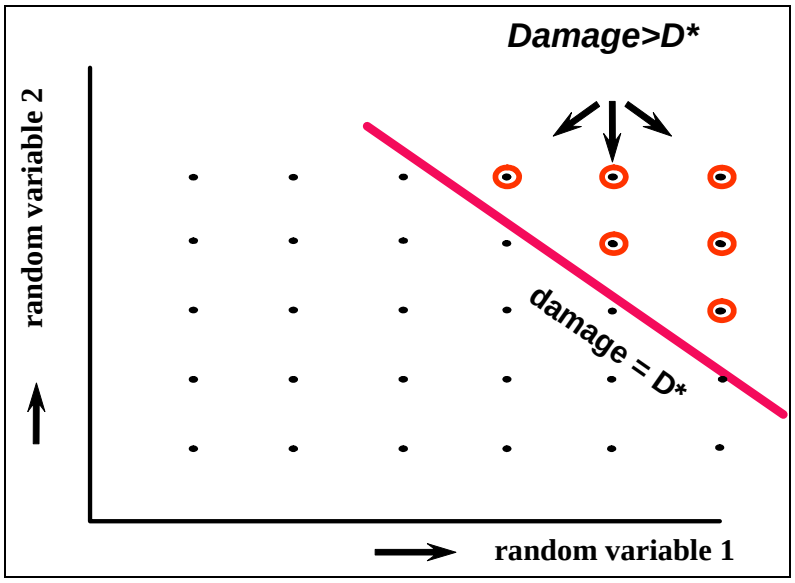

Figure 2.10 Schematic view of numerical integration

Crude Monte Carlo

Crude Monte Carlo (MC) is relatively straightforward to apply and still, if enough samples are taken, very accurate. In essence, MC involves a large number (*N*) of deterministic model simulations, where for each individual simulation the random parameters are sampled from their respective probability distribution functions. For each of the *N* simulations the resulting damage is derived. The probability that the damage exceeds *D** can simply be estimated by counting the total number of simulations for which *D>D** (*M*) and divide them by the total number of samples (*N*) (see Figure 2.11). The name "Monte Carlo" refers to the famous casino in Monaco. The use of randomness and the repetitive nature of this method are analogous to the activities conducted at a casino, i.e. repetitive activities with a rondom outcome like the throw of the dice.

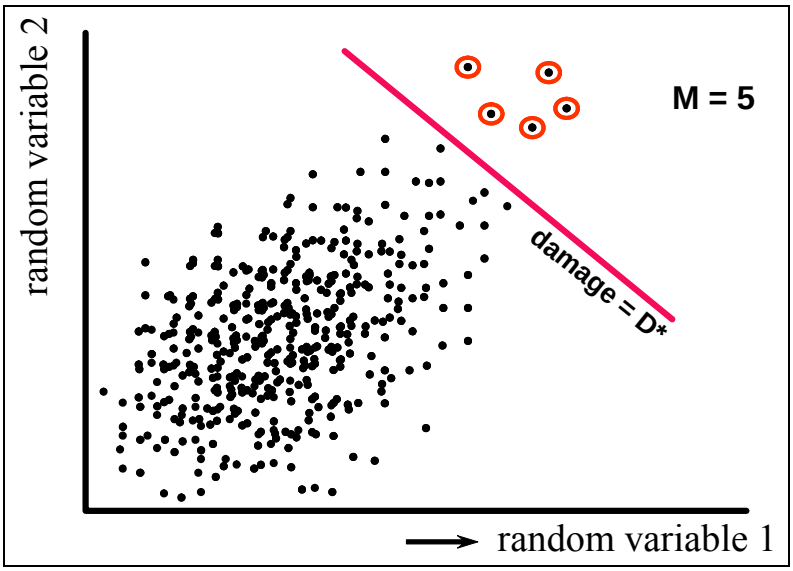

Figure 2.11 Schematic view of a Monte Carlo simulation

Example 1: an example of a Crude Monte Carlo (MC) simulation will clarify the method. Table 2.1 shows 20 combined MC-realisations for two random variables, peak discharge (1000 m^3/s) and flow volume (BCM), sampled from their combined distribution function. We emphasise that these numbers are hypothetic and used with the sole purpose to demonstrate the Monte Carlo procedure.

In 3 out of 20 cases the peak discharge and flow volume are such that the total damage exceeds the threshold value D^* = 80 million USD. This means the estimated probability of exceedance equals 3/20 each year. Note that in this example a relatively small amount of (20) samples is used. Generally, much more samples are used in order to reduce the uncertainty in the resulting estimate of the exceedance probability.

Table 2.1 Monte Carlo realisations with hypothetic numbers

nr.	peak discharge 1000 m^3/s	volume BCM	damage million USD	damage>80? yes/no
1	34.8	12.0	0.0	no
2	69.2	122.7	131.9	**yes**
3	32.7	20.8	0.0	no
4	58.3	44.8	43.1	no
5	36.1	10.8	0.0	no
6	25.7	2.4	0.0	no
7	32.2	5.5	0.0	no
8	53.6	89.9	83.5	**yes**
9	59.9	51.3	51.2	no
10	48.7	64.0	52.7	no
11	36.2	12.4	0.0	no
12	41.5	20.0	1.5	no
13	44.4	37.8	22.2	no
14	35.4	19.8	0.0	no
15	25.8	3.2	0.0	no
16	52.6	57.2	49.8	no

nr.	peak discharge	volume	damage	damage>80?
17	51.6	24.4	16.0	no
18	71.7	74.0	85.7	**yes**
19	40.1	14.1	0.0	no
20	30.3	5.7	0.0	no

Example 2

Since the sampling in Monte Carlo is random, the outcome is also random. This may be hard to understand, or, accept. However if a large enough number of samples are taken, the random nature of the outcome decreases. In statistics, this is known as the "law of large numbers". This is demonstrated by the example below where the two dice are thrown repetitively.

The possible outcome for the sum of two dice ranges from 2 (two times 1) to 12 (two times 6). The average outcome is 7, which can be verified by throwing the two dice a large (N) number of times and subsequently taking the average of the outcome. Figure 2.12 shows the results of such an experiment. The dice where thrown 10,000 times and the average as a function of the number of throws is shown (blue line). For the purpose of validation the experiment has been executed a second time (pink line).

Figure 2.12 shows that for a small number of throws (=10), the outcome varies. The first experiment (blue line) started with a couple of higher throws, which caused the average to significantly exceed the value of 7. Experiment 2, on the other hand, started of with a couple of low throws, which caused the average to stay below 7. However, for an increasing number of throws the differences with 7 clearly diminish. In other words: the law of large numbers causes the outcome to get increasingly close to the real value.

This example shows that as long as a "large enough" number of samples are taken the outcome of the procedure is reliable. At the same time it shows one should be careful not to take too less samples. While executing a Monte carlo analysis it is therefore worthwhile to examine the outcome by making plots similar to Figure 2.12. A (hypothetic) example is shown in Figure 2.13, where the probability of exceedannce of a flood level is estimated by a large number of Monte Carlo simulations. If the result still shows a relatively large variation over the whole range of the figure, additional samples should be taken. In this case, however, the results on the right side of the figure converges to the probability of approximately 0.05 (1/20) per year.

Figure 2.12 Monte Carlo sampling procedure for the sum of two dice. The average outcome of the sum of two dice is depicted as a function of the number of throws.

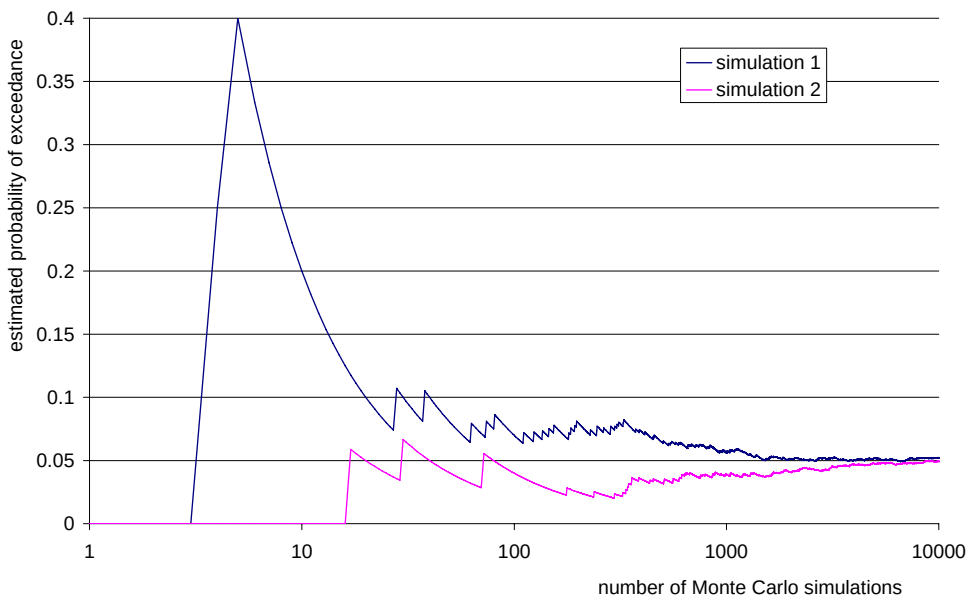

Figure 2.13 Monte Carlo sampling procedure for the derivation of the probability of exceedance of a certain flood level.

FORM

FORM: FORM is short for First Order Reliability Method. This method derives the design point with an iterative procedure and subsequently linearises the limit state function, $D=D^*$,

in order to derive the probability of exceedance of D^*. The design point is the combination of realisations of the involved random variables for which:

1. $D=D^*$, i.e it is part of the limit state;
2. of all combinations that form the limit state it has the highest probability of occurrence.

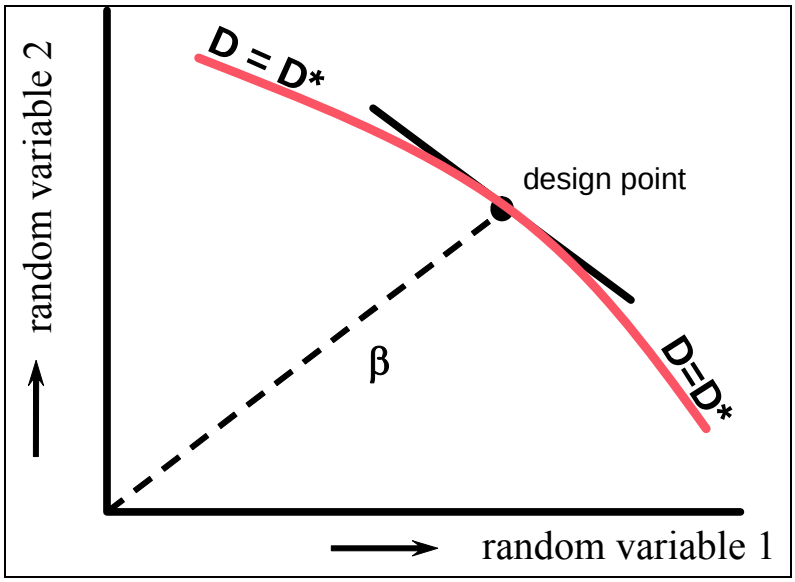

Figure 2.14 Schematic view of a FORM computation

Figure 2.14 shows a schematic view of a FORM computation. It shows the limit state (red curve), the design point (black dot) and the linearised limit state (straight black line). The advantage of the linearization is that the probability of exceedance of D^* can be estimated straight away, without further exploration of the area where $D>D^*$. The disadvantage is that this straight line is not the true limit state, i.e. an error is introduced in the estimated probability of exceedance. In most situations this error is relatively small though. This is because in the neighbourhood of the design point the linearised limit state function is very close to the actual limit state function. Since the design point has the highest probability of occurrence of all points on the limit state, the maximum contribution to the probability of exceedance of D^* is generally in the neighbourhood of the design point. Therefore the estimate based on the linearised function is generally very accurate.

The design point can be considered as the most likely combination of realizations of random variables that will lead to an exceedance of D^*. Generally FORM is applied with a set of standard normally distributed random variables $u_1, \ldots u_n$ that correspond with the "real" random variables $X_1, ..X_n$. The transformation from u to X or vice versa is done by setting equal their respective probability of non-exceedance:

$$F_i(x_i) = \Phi(u_i)$$

Where F_i is the distribution function of variable x_i and Φ is the standard normal distribution function. The u-variables have a mean of 0 and a standard deviation of 1. Furthermore they are mutually independent. The advantage of the transformation to u-variables is that is gives a clear interpretation of the design point (see Figure 2.15). In the u-space, it is the point on the limit state that is closest to the origin.

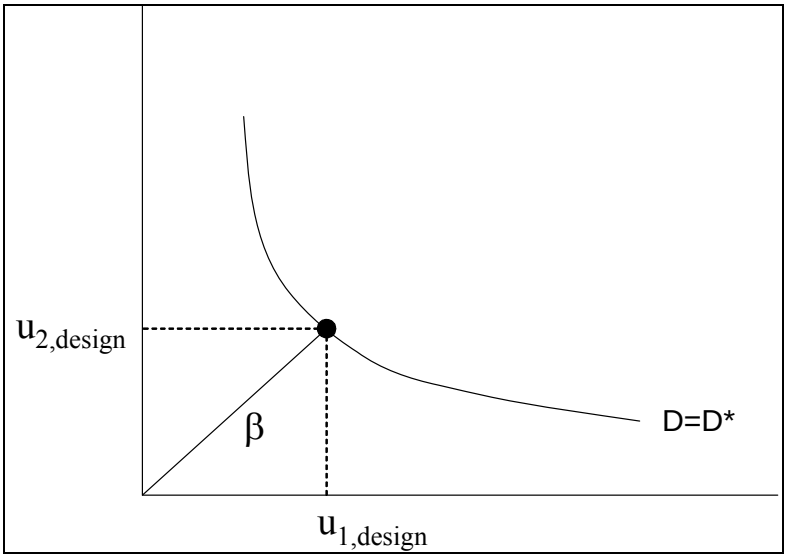

Figure 2.15 Design point in u-space

The distance, β, from the design point to the origin is the *reliability index*. Large values of β, indicate large deviations of u-values from their mean (0) and therefore extreme events. The probability of failure is computed directly from β:

$$P(D>D^*) \approx P(D_L > D^*) = \Phi(-\beta)$$

Where D_L is the linearised version of damage function D in the design point:

$$D_L = B + A_1 u_1 + \ldots A_n u_n$$

Rationale

The above described methods are all approximation techniques, necessary because an exact analytical solution cannot be derived. Naturally, the error introduced by the approximation should be as small as possible. Generally, these errors are relatively small compared to errors in the measurements, hydraulic modelling or estimates of (economic) damage.

The error in numerical integration depends on the size of the intervals between the grid points: the smaller the intervals, the smaller the error. For Monte Carlo the error decreases with increasing number of (N) samples. So, in theory the approximations of these two techniques are exact if the grid intervals are infinitely small (numerical integration) or N is infinitely large (Monte Carlo). However, this is always at the expense of the computation time, since this requires a large amount of hydraulic model simulations. And since the available computation time is always limited there will always be some error.

FORM has the advantage that it requires relatively little computation time. The disadvantage of this method is that the iterative algorithm to find the design point sometimes does not converge and results are not always reliable. This is especially the case if the relation between the input and output of the simulation model is highly non-linear.

The selection of the “best” approximation technique, i.e. the technique which introduces the smallest error within reasonable computation time depends on the problem under consideration.

2.4 Flood hazard versus flood risk

In flood *hazard* analysis the focus is on hydraulic features. With respect to the previous examples this means only probabilities of flood water levels are derived and no damages are considered. In flood *risk* analysis the combination of water levels and damages are considered. The previous sections showed examples of probabilities of exceedance of flood both water levels and or damages. It shows that the basic principles of probabilistic modelling are the same for flood hazard analysis as for flood risk analysis.