7

Risk Analysis

7.1 Introduction

In previous chapters two assumptions were made about data needed for successful simulation runs. It was first proposed that necessary data is completely available and second that it is good quality. So it was implicitly concluded that each model is unique. In practice, this is usually not the case. Data sets have gaps and the data values often have wide error bars. These uncertainties lead to the following three types of questions:

1. What is the impact of uncertainties in the input data on the model?

What is the chance or the probability of having special scenarios? How large is the risk or the probability of failure? Is the simulation result stable or does a slight variation of some input parameters cause a completely different result? How sensitive is the relationship between a given parameter variation and the resulting model variation or how do the error bars of the input data map to the error bars of the results?

2. What are the important dependencies in our model?

Not every uncertainty of an input parameter has an impact on each uncertainty of the simulation result values. Which parameter influences which result? How strong are the different influences? Do they have a special form? Studying these questions is especially necessary for the understanding of the model and the processes it contains. Understanding is again necessary if conclusions are to be drawn, which go beyond a plain collection of results.

3. Which set of input data leads to agreement when considering additional comparison data?

Very often additional calibration data are available which cannot be used directly for the modeling but can be compared to simulation results. Is it possible to reduce the uncertainty in the input data by excluding models related to simulation results which are not matching the calibration data? In the literature, procedures treating this problem are often listed under the keywords “inversion” or “calibration”.

T. Hantschel, A.I. Kauerauf, *Fundamentals of Basin and Petroleum Systems Modeling*, DOI 10.1007/978-3-540-72318-9_7,

This chapter deals with these three topics under the headings “Risking”, “Understanding”, and “Calibration”.

The classical approach of tackling such problems would be to perform scenario runs. Starting with a first best guess model, which is commonly named “reference model” or “master run” (Fig. 7.1), model parameters are modified manually and scenario runs are performed according to the knowledge, speculations, expectations, and understanding of the modeler (Fig. 7.2).

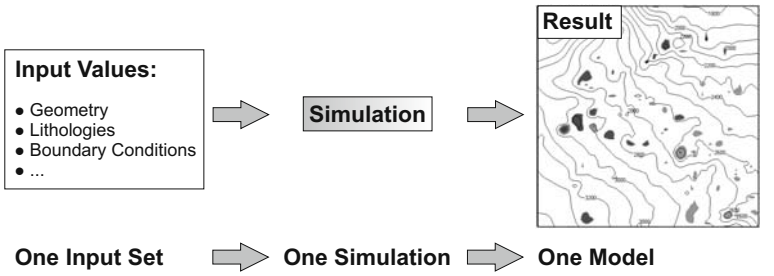

Fig. 7.1. Result of one – deterministic – 3D simulation. It is implicitly assumed, that uncertainties in input data do not exist

An example, with an uncertain temperature history caused by unknowns in heat flow and thermal conductivity could have this typical form: A high heat flow scenario and a low heat flow scenario are simulated, whereas other uncertain input parameters such as thermal conductivities are held at fixed values. The high heat flow scenario is found to be realistic by looking for example at the calibration data or through the experience of the modeler. Next, a high thermal conductivity scenario and a low conductivity scenario are modeled with a fixed high heat flow. High thermal conductivity matches the calibration data best. So, a high heat flow combined with high thermal conductivity is found to be the most realistic scenario.

Two main problems arise with such an approach:

The procedure of variation and selection of the input parameters is not systematic. There is a possibility of overseeing other realistic scenarios e.g.: in this example, low heat flow with low thermal conductivity. With a higher amount of uncertain parameters such mistakes can become normal.

The choice of “probable” scenarios is not very well quantified: A sensitivity analysis of how precise the heat flow has to be known to match the result is not performed, a quantification of the reduction of uncertainty is missing and the repercussions of the variation of other parameters, simultaneously with the heat flow, is omitted completely.

The reliability of risk results gained by scenario runs is primarily dependent on the knowledge of the involved modelers. Scenarios are usually not

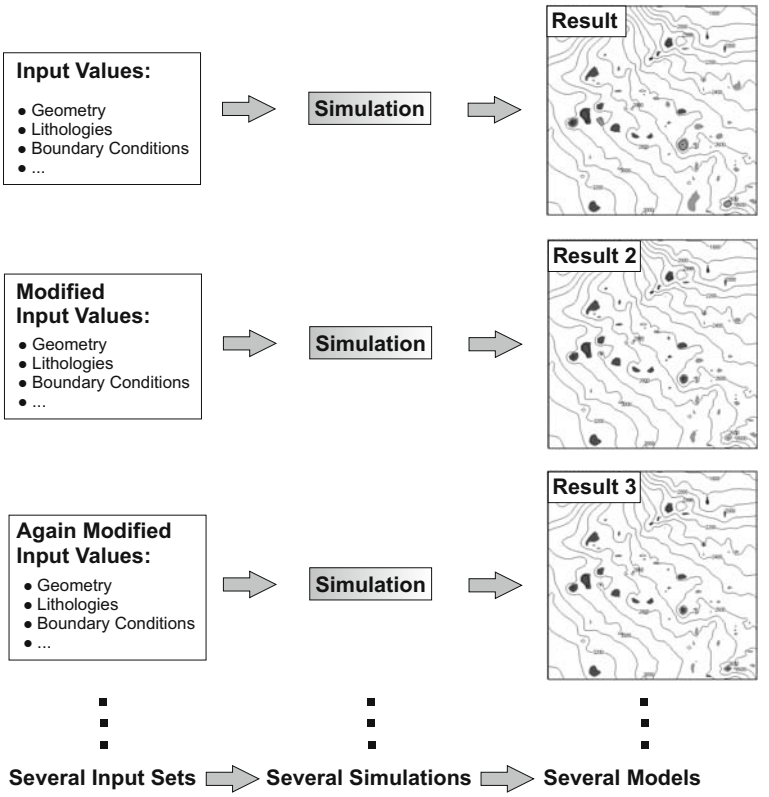

Fig. 7.2. Approach with "Scenario Runs". The first run is usually called the "Master Run" or "Reference Model"

performed systematically and the discussion of the results is qualitative but not quantitative.

The main goal of this chapter is to describe a systematic approach to deal with these issues. A more concrete formulation of the tasks involved with the three topics are:

- **Risking:** Calculation of probabilities, confidence intervals and error bars.
- **Understanding:** Calculation and analysis of correlations.
- **Calibration:** Calculation of the probability of how good a model fits calibration data and search for the best fitting model.

All topics contain words such as "probability" or "correlation" which are related to the language of stochastics and statistics. It is possible to treat all three topics simultaneously with a stochastic method such as a "Monte Carlo Simulation". This has the big advantage that expensive and time consuming simulation runs can be reused for the analysis of three distinct topics.

An introduction into probabilistic methods of applied basin modeling can be found in Thomsen (1998). Other approaches are usually less general and restricted in applicability or assumptions. Nevertheless, the efficiency can be significantly raised by studying limited problems or tasks with different methods. Highly specialized methods of inversion, for interpolation and extrapolation of simulation results are discussed in later sections.

7.2 Monte Carlo Simulation

The starting point for a Monte Carlo simulation is a reference model and a list of uncertainties belonging to the data. The reference model is based on a parameter set within the limits of these uncertainties, which typically represents a best first guess. Additionally, a quantification of the uncertainties must be known. The most precise quantification is a probability distribution (Fig. 7.3) which defines the probability of a data value to be exact.[1]

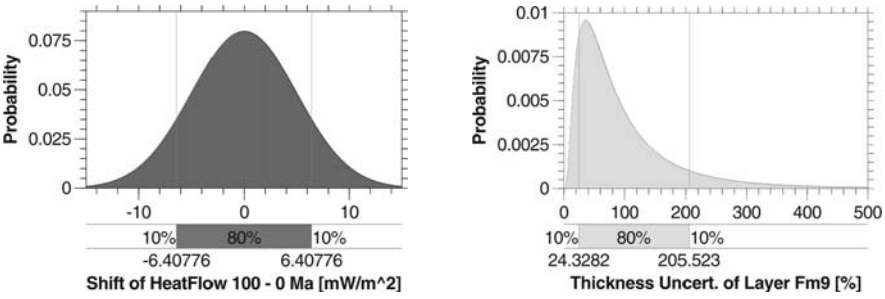

Fig. 7.3. Examples of normally and log. normally distributed uncertainties

Very often the distribution is not known but only some more general statements about the type and size of the uncertainties. It is usually not difficult and also not critical to construct a distribution from this knowledge. This is discussed in Sec. 7.2.1 and typical examples are demonstrated.

One important point, which must be mentioned, is that the uncertain model parameters should be independent. In Sec. 7.2.3 this is discussed in more detail.

With this setup the "Monte Carlo Workflow" is straightforward: A set of random numbers according to the distributions is drawn and a simulation run with this parameter set is performed. This procedure is repeated while the results are collected (Fig. 7.4). Output parameters are collected, visualized, and analyzed with statistical tools such as histograms.

[1] It more precisely defines a probability density with probabilities of values to be within certain intervals.

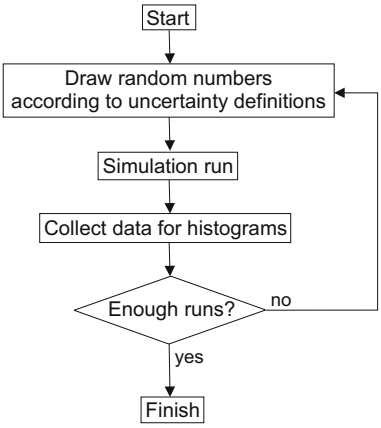

Fig. 7.4. Flow chart for Monte Carlo simulation runs

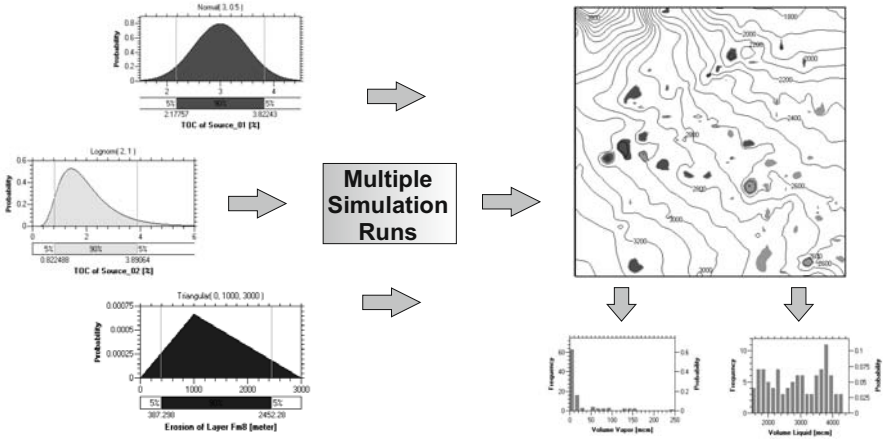

Fig. 7.5. Monte Carlo simulation with histograms of accumulated petroleum

It will now be shown that the main topics "risking", "understanding", and "calibration" can be solved with the Monte Carlo simulation approach:

Risking

Confidence intervals related to risking can directly be read off from result histograms. They define the probability to find a result within a given interval. E.g. it is possible to formulate statements such as *"With* 80% *probability the accumulated liquid petroleum amount is between* 1623 *and* 1628 *million cubic meters"* (Fig. 7.6).

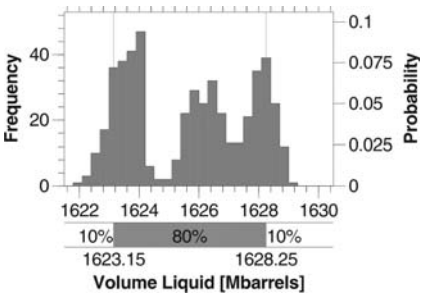

Fig. 7.6. Histogram of liquid accumulations: "With 80% probability the accumulated liquid petroleum amount is between 1623 and 1628 million cubic meters"

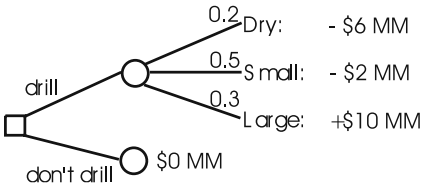

Fig. 7.7. Decision analysis with tree: The expected value EV for gains or losses of the "drill branch" is the average $EV(\text{drill}) = 0.2 \times (-6) + 0.5 \times (-2) + 0.3 \times 10 = \$\,0.8\,\text{MM}$

Other valuable characteristics of a histogram are the modus, which defines the location of the most probable result, or the average. The concept of calculating expectation values, such as the average, is extremely important in economics: For example complex decision procedures in companies are often analyzed with decision trees (Fig. 7.7). These trees are based on the evident statistical law, that the optimal decision strategy is found by following the branches with the highest expectation values, which can be calculated from averages.

A measure for the width of a histogram is the standard deviation. This quantity can be set in relation to the less precisely defined error bar. Together, average and standard deviation are often used as "value with error bar" (Fig. 7.8). Big standard deviations of resulting histograms indicate high uncertainties and give rise to the conclusion that the master run is not representative and therefore not probable.

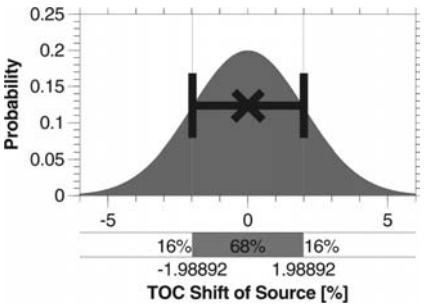

Fig. 7.8. Gauss distribution with mean $\mu = 0$ and standard deviation $\sigma = 2$. The standard deviation can be interpreted as the size of an error bar. About 68% of numbers drawn from this distribution will be inside the range of the error bar

The analysis of the result widths as functions of the uncertainty widths is called "sensitivity analysis". More precisely, it can be represented by the relation between standard deviations of uncertainty and result parameters. A result is highly sensitive/unsensitive to an uncertainty if its error bar is large/small compared to the error bar of the uncertainty parameter. Therefore, sensitivity analysis can be a guiding tool for the understanding of a system.

Understanding

A very important problem arises with the question of where future efforts concerning the reduction of uncertainties should be spent. A reduction of uncertainties can be achieved by further data acquisition, which can be very costly. Sensitivity analysis directly leads to the parameters, which are of importance (Fig. 7.9). So, an expensive collection of unnecessary data could be avoided.

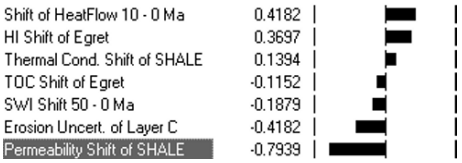

Fig. 7.9. Tornado diagram depicting the influence of some uncertainties on the porosity at a defined location in a well. Spearman rank order correlation coefficients (Press et al., 2002) are plotted as bars. As expected, the permeability shift (highlighted) (anti–)correlates mostly with the porosity

Understanding can be improved by searching for dependencies, e.g. via cross plots (Fig. 7.10). Correlations can be visualized and with the help of correlation coefficients quantified. In case of strong correlations, it is possible to interpolate the results and for forecasting purposes state formulas of dependency. For expensive simulation runs this is very valuable. Generalizations of such techniques are discussed in Sec. 7.5.

Calibration

It is obvious that calibration could be performed with Monte Carlo simulations in a simple way by just looking at the model that best fits the calibration data. The investigation of uncertainty space is performed by sampling the uncertainties according to their probability distributions. Random combinations of parameters are used for the Monte Carlo models. This method ensures a global sampling of the space of uncertainty. The risk of missing regions with good calibration, becomes small with a high number of Monte Carlo runs.

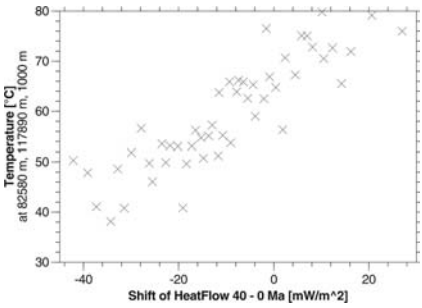

Fig. 7.10. Cross plot of temperature against heat flow shift. A correlation is visible and a linear interpolation might be performed

The Monte Carlo method is not intelligent in a way that it searches for models with good calibration. Search algorithms would be more efficient but obviously cannot be performed simultaneously in combination with risking and understanding. Additionally, these algorithms often search in local regions of space so it could be that they end up with an erroneous calibration. Therefore, a global investigation of the uncertainty space such as with a Monte Carlo analysis has to be performed as a first step before beginning with a search algorithm. Global stability and the prevention of extra simulation runs are often of higher importance compared to high–quality calibration. However, in Sec. 7.5 more sophisticated calibration methods, which combine the advantages of both approaches are discussed.

7.2.1 Uncertainty Distributions

Uncertainty distributions must be specified for Monte Carlo simulations. The properties of some well known distributions are now discussed with regard to their usage in Monte Carlo simulations.

Normal Distribution

The normal or Gauss distribution

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \tag{7.1}$$

with mean μ and standard deviation σ is the most widely used distribution in science (Fig. 7.3). Assume that a quantity X is measured independently N times with the values $x_1 \ldots x_N$. Following the central limit theorem of statistics, the average

$$x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{7.2}$$

is Gauss distributed for $N \to \infty$ with[2]

[2] In practice $N > 7$ is enough for high numerical accuracy.

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\mu_i \qquad \text{and} \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2 \,. \tag{7.3}$$

Here, μ_i and σ_i are the means and the standard deviations of the probability distributions for each measurement i. They are often the same for all i. Parameters which are used in large scale basin models are often provided as upscaled averages of higher resolution data or averages of multiple measurements. Assuming independency, it is often possible to assign a Gauss distributed uncertainty to such a parameter.

Logarithmic Normal Distribution

This distribution is also called lognormal or lognorm distribution and has the form

$$p(x) = \frac{1}{\sigma x\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \quad \text{for} \quad x > 0 \qquad \text{and} \qquad p(x) = 0 \quad \text{else} \,, \tag{7.4}$$

compare with Fig. 7.3. It has similar properties to the normal distribution. If a quantity Y is normally distributed then $X = \exp Y$ is lognormally distributed. The central limit theorem for the arithmetic average of some Y_i becomes a geometric average for the related X_i namely

$$x = \prod_{i=1}^{N} x_i^{1/N} \,. \tag{7.5}$$

The equations for μ and σ stay the same as in (7.3) but it should be remembered that μ and σ are not the mean and standard deviation of the lognormal distribution, they are only the mean and standard deviation of the related normal distribution.

A lognormal distribution is of special interest to "scale quantities", which by definition cannot be negative. The logarithm of a scale quantity can be calculated every time and the distribution is zero for negative values. Many physical quantities especially material properties such as thermal conductivities are limited to positive values. And the calculation of averages e.g. upscaling, is often performed with geometrical averaging (Chap. 8). The lognorm distribution can be a proper choice for the description of uncertainties related to such quantities.

Uniform Distribution

The uniform distribution (Fig. 7.11) is defined by

$$p(x) = \frac{1}{b-a} \quad \text{for} \quad a \le x \le b \qquad \text{and} \qquad p(x) = 0 \quad \text{else} \tag{7.6}$$

with $a < b$. The uniform distribution is a good choice if nothing except some limiting statements can be made about the form of the uncertainty. The two discontinuities of the distribution are often the subject of criticism: They have sharp edges and are therefore argued to be in contradiction to the assumption of ignorance about the "tails" of the distribution. Additionally, it often seems unreasonable that the central inner parts of an uncertainty have the same probabilistic weight as the more outer parts.

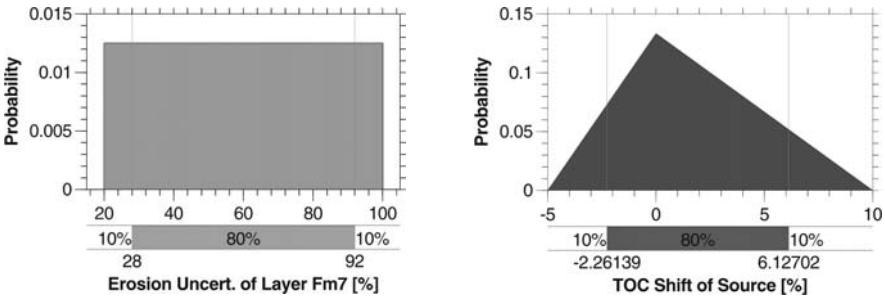

Fig. 7.11. Examples of uniform and triangular distributions

Triangular Distribution

The triangular distribution (Fig. 7.11) does not have the principal problems, which come with the uniform distribution. It is given by

$$p(x) = \begin{cases} \dfrac{2(x-a)}{(c-a)(b-a)} & \text{for} \quad a < x \leq b, \\ \dfrac{2(c-x)}{(c-a)(c-b)} & \text{for} \quad b < x < c, \\ 0 & \text{else} \end{cases} \tag{7.7}$$

with $a < b < c$. The distribution is zero at a and c and its median is located at b. A triangular uncertainty distribution can be directly constructed when the uncertainty limits and also the most probable value are known. It can also be used as an "easy to use" approximation to normal and lognorm distributions (Lerche, 1997; Thomsen, 1998).

Other distributions, such as exponential or beta distributions, are more sophisticated alternatives to uncertainty descriptions (Figs. 7.12, 4.5, Rinne 1997). They are only used under special circumstances.

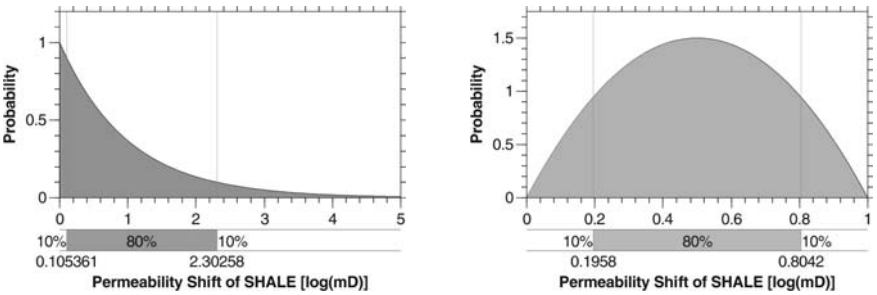

Fig. 7.12. Examples of exponential and beta distributed uncertainties

Nominal Distributions

Sometimes it is necessary to assign uncertainties to discrete parameters (Fig. 7.13). In the most general case these parameters are without order relation, which signifies that there are no "less than" or "bigger than" defined between them. Then they are called nominal parameters. Typical examples are lithologies or kinetic type assignments as lithologies and kinetics are usually specified by a large number of parameters. Therefore uncertain nominal parameters often imply strong result variations.

An interpretation of results derived from nominal uncertainties can be difficult especially if nominal and continuous uncertainties are mixed in the same sequence of risk runs. This should therefore be avoided.

Substit. of Lithology SANDSTONE_04

ID	Name	Prob. [%]	0 % 100 %
108	SANDSTONE_04	70.00	
42	SAND&SILT	15.00	
43	SAND&SHALE	15.00	

Fig. 7.13. Discrete distributed lithologies

7.2.2 Derived Uncertainty Parameters

An uncertainty is described as a distribution of one number, e.g. the thermal conductivity of shale. But very often it is associated with more than only one number. For example a heat flow uncertainty is related to the complete basal heat flow, which is space and time dependent and cannot be described with one number only.

However, the basal heat flow of the master run can be shifted, tilted, twisted, etc.. A restriction to special forms of variation which can be described with one number only allows the assignment of an uncertainty to this "derived parameter".

It is well known from mathematics that arbitrary variations can often be decomposed into infinite series of orthogonal functions which would yield infinite uncertainty parameters. In practice, one is thus restricted to the most important variations, which are often defined by the first terms of such series.

The simplest form of basal heat flow variations is a value shift in a defined time interval (Fig. 7.14). Therefore, it is possible to assign an uncertainty distribution to a shift of the whole basal heat flow.

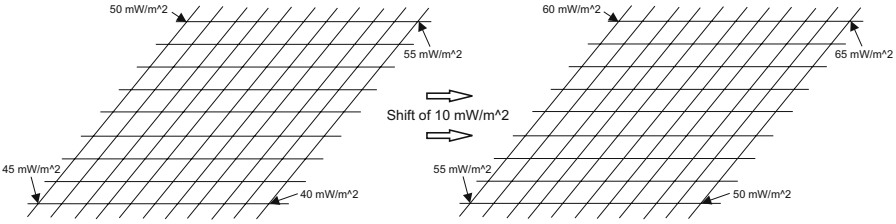

Fig. 7.14. Shift of (gridded) basal heat flow map

Complex structural uncertainties can easily be risked with the prize of restriction to special forms of variation (Fig. 3.36).

7.2.3 Latin Hypercube Sampling (LHC)

Arbitrary random sampling of the uncertainty distributions has some drawbacks. Clusters of drawn numbers can occur (Fig. 7.15) and low probability tails of distributions are often not sampled, although they might contribute significantly to the analysis (e.g. calculation of moments) especially if they have a wide range. The statistics (e.g. estimating a mean with the average over a set of random numbers) becomes increasingly better, as less clusters exist and the smoother the sampling is.

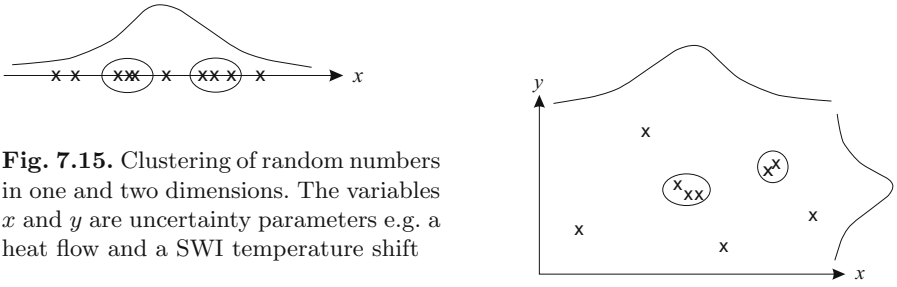

Fig. 7.15. Clustering of random numbers in one and two dimensions. The variables x and y are uncertainty parameters e.g. a heat flow and a SWI temperature shift

Latin hypercube sampling is a technique which helps to avoid clustering and samples low probability tails without affecting basic statistics. It consists

of primarily two parts: The first part is an improved drawing algorithm in the series drawn from one distribution. The second part refers to the "hypercube" and deals with multiple drawings in the multi–dimensional uncertainty "hyperspace".

Latin Hypercube Sampling in One Dimension

The interval within which the uncertainty parameter is defined can be divided into intervals of the same cumulative probability which are called "strips" (Fig. 7.16). Drawing a random number is now performed in two steps: First, a strip is selected. Then, a random number is drawn according to the probability distribution in this strip (Fig. 7.17). It is not allowed to use a strip again until all others have been selected for drawing. The best efficiency is obviously achieved if the number of drawings equals the number of strips or is a small integer multiple.

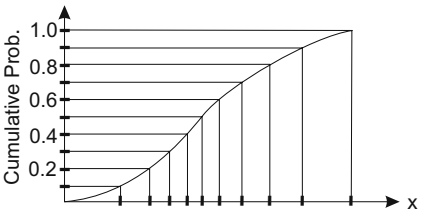

Fig. 7.16. Segmentation into ten equal probable intervals

LHC sampling is three times more efficient for the calculation of basic statistical quantities such as means or confidence intervals (Newendorp and Schuyler, 2000). When considering the huge size of simulation efforts for big basin models, this is a good deal for the price.

On the other hand it is easy to see, that this method does not reproduce auto–correlations between successive drawings. In practical implementations, the additional effort for the calculation of the strips and for the bookkeeping of used and unused strips has also to be taken into account.

Latin Hypercube Sampling in Multiple Dimensions

In more than one dimension, each distribution is segmented into the same number of strips. When drawing the random numbers, it is necessary to avoid correlations between the selection of the strips. Therefore the strips must be selected randomly too.

The final result is a subdivision of uncertainty space into equal probable hypercubes (Fig. 7.17). In the case that the number of drawings equals the number of strips, each cube contains only one drawn number at maximum.

In two dimensions each column or row contains one drawn number and in N dimensions each $N-1$ subspace contains exactly one number.

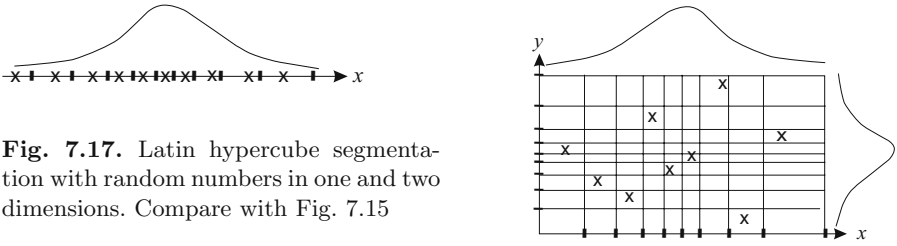

Fig. 7.17. Latin hypercube segmentation with random numbers in one and two dimensions. Compare with Fig. 7.15

Again, bookkeeping of strips has to be performed but the advantages are the same as in the one–dimensional case. LHC is a very efficient method for global sampling of the uncertainty space.

7.2.4 Uncertainty Correlations

Up to now, it was assumed that the uncertainty parameters were independent. The opposite of independence is dependency. This does not need to be further discussed, as a dependent uncertainty parameter can obviously be eliminated from the list of uncertainties and treated like a simulation result. Besides these two extremes, the region of correlation exists where specified combinations of the drawn numbers are favored above others.

An example could be the thermal conductivity of two layers which are known to have similar lithologies but it is not known what they are. So, for heat flow analysis a modeler would prefer to study combinations of similar conductivities.

A complete joint probability distribution, which defines the probability for all combinations of all values of the uncertainty parameters, would be the most thorough description. However, data and theoretical foundations of certain joint probability distribution forms usually do not exist. In practice it is sufficient to deal with correlation coefficients which are used to link marginal distributions. The rest of the joint probability distribution remains unspecified.

Nevertheless, drawing random numbers of correlated distributions is problematic enough. Explicit formulas exist for correlated Gauss distributions. The simplest case are two correlated Gauss distributions which have the following form (Beyer et al., 1999)

$$p(\mathbf{x}) = \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-\frac{1}{2}\mathbf{x}^T \Sigma^{-1}\mathbf{x}} \,. \tag{7.8}$$

with two variables $\mathbf{x}^T = (x_1, x_2)$. The correlation is defined by the covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{7.9}$$

which is symmetric and positive definite (Fahrmeir and Hamerle, 1984). Here $\sigma_i =< x_i^2 >$ are the variances and $\rho =< x_1 x_2 > /\sigma_1\sigma_2$ is the correlation coefficient with $-1 \leq \rho \leq 1$. Without loss of generality $\sigma_1 = \sigma_2 = 1$ is further assumed. The correlation matrix can be Cholesky decomposed (Beyer et al., 1999; Press et al., 2002) into

$$\Sigma = A^T A \qquad \text{with} \qquad A = \begin{pmatrix} 1 & \rho \\ 0 & \sqrt{1-\rho^2} \end{pmatrix} . \tag{7.10}$$

Thus $\mathbf{x}^* = A\mathbf{x} = (x_1 + \rho x_2, \sqrt{1-\rho^2}x_2)^T$ is Gauss distributed without correlation. Obviously, this can easily be generalized to higher dimensions.

Correlations of arbitrary marginal distributions can be forced with numerical methods. At least three different algorithms are known to exist (Miller, 1998). In the case of many distributions with many correlations, these algorithms become computationally very expensive. Especially, if one parameter is correlated multiple times with other parameters, these methods are not affordable anymore. Another disadvantage of these algorithms is their incompatibility with latin hypercube sampling. Abdication of LHC sampling reduces the performance significantly.

A new approximative method to get correlated random numbers is now described: All random numbers can be drawn before performing the risk runs if the total number of simulation runs is known at the beginning and correlations are ignored. These random numbers can be sorted afterwards with the following algorithm: An uncertainty parameter is randomly selected and after that two random numbers out of its sequence are randomly selected again. These numbers are swapped, if the resulting covariance matrix approximates the target covariance matrix more closely than before. This procedure can be repeated until a high degree of accuracy is reached. The sum of the squared deviations of all correlation coefficients can be taken as a measure for the total deviation.

It is clear that this permutation procedure certainly does not lead to a sufficient approximation and never to the exact reproduction of all correlation coefficients if the number of simulations is small.[3] But experience has shown that an almost exact numerical agreement can be reached very fast on modern computers in practical relevant examples. Even for about twenty runs with a few correlated parameters a good numerical agreement could be achieved.

[3] An extension of this method with some acceptance/rejection probability of a number swap would transfer it into a "Markov chain Monte Carlo" (MCMC) algorithm. It can be proved that such an algorithm finds the optimal approximation over longer time intervals. Due to rejection the MCMC algorithm shows generally a poorer performance. By experience, the authors found the MCMC algorithm here not necessary. MCMC algorithms in general will be discussed in more detail in Sec. 7.5.

Another big advantage of this procedure is its compatibility with latin hypercube sampling. Hypercubes and therefore higher performance coming from a lower number of necessary risk runs can be conserved.

An example of a correlation matrix constructed with this permutation method is given here: the matrix which links some marginal probability distributions of uniform, triangular, normal, and lognorm form is defined as

$$\begin{pmatrix} 1 & & & & & & & \\ 0 & 1 & & & & & & \\ 0 & 0.2 & 1 & & & & & \\ 0 & 0.3 & 0 & 1 & & & & \\ -0.1 & 0 & 0 & 0 & 1 & & & \\ -0.7 & 0 & -0.1 & 0.1 & 0 & 1 & & \\ 0 & -0.2 & 0 & 0 & 0 & 0 & 1 & \\ -0.6 & 0.2 & 0 & 0 & 0 & 0 & -0.5 & 1 \end{pmatrix}.$$

Because of symmetry, only the lower triangular part of the matrix is shown here. For 20 random numbers, which correspond to 20 simulation runs, the following approximation could be achieved with a maximum deviation of 0.0321 of any correlation value

$$\begin{pmatrix} 1 & & & & & & & \\ -0.0085 & 1 & & & & & & \\ -0.0092 & 0.2088 & 1 & & & & & \\ 0.0131 & 0.3004 & -0.0008 & 1 & & & & \\ -0.0978 & -0.0060 & -0.0005 & 0.0057 & 1 & & & \\ -0.6705 & -0.0063 & -0.0973 & 0.1091 & 0.0006 & 1 & & \\ 0.0178 & -0.1956 & 0.0008 & -0.0033 & -0.0015 & 0.0172 & 1 & \\ -0.5679 & 0.1957 & -0.0102 & -0.0078 & 0.0024 & 0.0310 & -0.4865 & 1 \end{pmatrix}$$

and for 100 runs

$$\begin{pmatrix} 1 & & & & & & & \\ -0.0009 & 1 & & & & & & \\ 0.0004 & 0.2000 & 1 & & & & & \\ 0.0000 & 0.2998 & 0.0002 & 1 & & & & \\ -0.1000 & -0.0002 & 0.0002 & 0.0005 & 1 & & & \\ -0.6938 & 0.0000 & -0.0998 & 0.0999 & -0.0002 & 1 & & \\ 0.0031 & -0.1997 & 0.0001 & 0.0000 & 0.0016 & -0.0023 & 1 & \\ -0.5928 & 0.1992 & 0.0005 & 0.0000 & 0.0011 & 0.0051 & -0.4965 & 1 \end{pmatrix}$$

with a maximum deviation of 0.0072. The maximum relative error of any correlation coefficient with an absolute value bigger than 0.1 is then less than 2%. This error is usually far beyond the accuracy of the knowledge of the correlation coefficients.

7.2.5 Analysis of Results

Many good textbooks are available about probability theory and statistics, e.g. Beyer et al. (1999) or Spiegel and Stephens (1999). Due to some specific problems associated with basin modeling some subjects are reviewed here.

An introduction to statistical analysis has already been given in Sec. 7.2. Histograms and cross plots are used to visualize the data while numbers, such as average values, can be calculated for further analysis. Statements can be quantified, e.g. with percentiles. A histogram is a binned approximation of a probability distribution. The width of a bin should neither be too narrow nor too wide because the visualization would become meaningless or grouping errors would become an issue (Spiegel and Stephens, 1999). Especially the calculation of percentiles for risking directly from histograms yields only "gridded values" (Fig. 7.18). In this case it is often better to use a linear interpolated form of cumulative probability. It is more precise for the calculation (e.g. percentiles) and has a smoother visualization (Fig. 7.18) but the data have to be available in raw form. This corresponds to a higher allocation of resources in a computer implementation.

Nevertheless, the binning width of a well sampled histogram indicates the statistical error of extracted quantities such as mean values or percentiles. Analysis (Fig. 7.18) shows that at least about 100 data values and therefore 100 risk runs are necessary for statistics with an acceptable relative error of around a few percent.

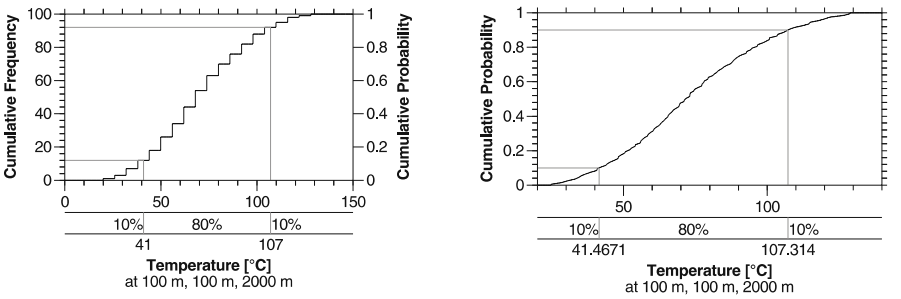

Fig. 7.18. Histogram with cumulative frequency of 100 drawn temperature values on the left and linear interpolated cumulative probability based on the same data on the right

Correlation coefficients are usually calculated for the analysis of possible dependencies. One should not forget that correlation is necessary but not sufficient for dependency. So, finding dependencies is not only part of analysis but also interpretation. Unfortunately, dependencies can have a variety of forms. Standard approaches of statistics test only for special forms.

Most commonly used is the Pearson correlation coefficient. It is a measure for the deviation of a cross plot from a straight line. If its correlation value is 1

then the cross plot fits perfectly to a straight line, which is positively inclined, and if it is −1 to a line, which is negatively inclined. If the correlation is 0 then there is no similarity to a straight line at all. Intermediate values indicate an approach to a straight line which becomes better with increasing absolute values (Fig. 7.19).

A straight line is the most important form of dependency but it is also a strong restriction to a special form. The Spearman rank order correlation coefficient is more general. It models the order of data points and is a measure of the deviation to an arbitrary monotonic increasing or decreasing correlation. Even a little more general but nearly the same is "Kendall's tau". It relies more on relative ordering and less on ranks. Some example cross plots with different correlation coefficients are shown in Fig. 7.19.

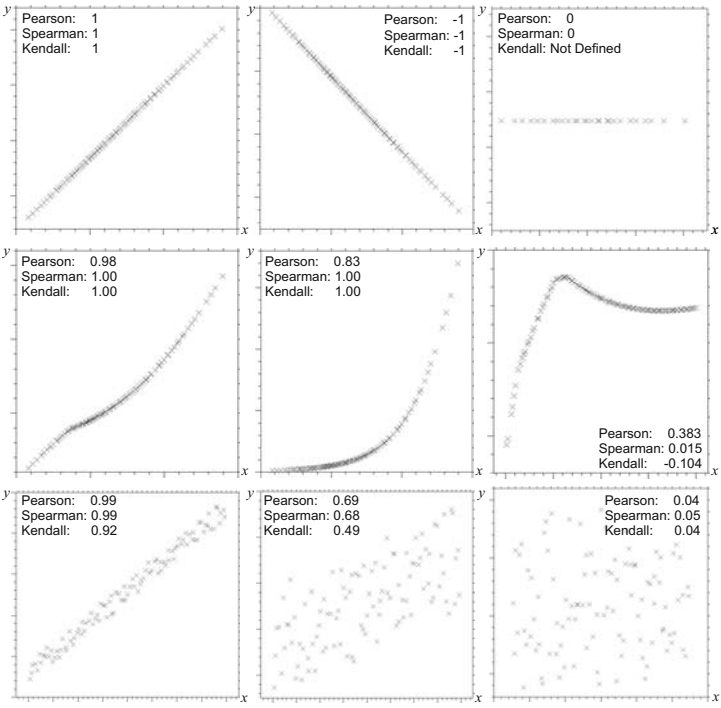

Fig. 7.19. Some examples of cross plots and their correlation coefficients

Spearman's rank order coefficient ranges also from −1 to 1 but it is overall known to be more robust than Pearson's correlation coefficient (Press et al., 2002). Commonly, it is used for tornado diagrams where lists of correlations are ranked and visualized (Fig. 7.9).

The existence of correlation can generally not be determined with a correlation coefficient alone. It only describes the strength of the correlation of a

specific data set. For example a small dataset can be randomly correlated. The most extreme case are two points which fall on a straight line every time. But fortunately it is often possible to estimate significance levels for the existence of correlation. For a more detailed discussion see Press et al. (2002).

Nominal distributions must be treated differently. Because of missing order relations a correlation cannot be defined properly anymore. Instead, associations are calculated. Typical association values are Cramer's V or the contingency coefficient C. Their interpretation can be complicated. Alternative measures of association exist and are based on entropy (Press et al., 2002).

The values of association reach from 0 to 1 from no– to full– association. They cannot be negative just like their continuous counterparts, which must be kept in mind when plotted e.g. in tornado diagrams. Similar to continuous correlations, significance levels can also be estimated for associations.

If a uncertain nominal parameter is associated with a result distribution of the corresponding model, the result distribution must obviously be available in discretized, e.g. binned, form.

7.2.6 Model Data

Basin models are usually very large in size and contain a vast amount of data. It is not possible to store all the data of each Monte Carlo run completely. Only selected and restricted amounts of data can be handled and therefore not all statistical methods can be used for analysis every time.

The calculation of the average is exceptional. It can be obtained just by adding the results of each run and finally, after the last run, by dividing through the number of runs. Hence a storage of all results of all runs can be avoided. Variance can be treated in a similar way. More precisely, average and variance of a quantity need only resources of the same size necessary for the storage of the corresponding result values of two simulation runs. It is therefore possible to store them for all quantities of interest even for grid based spatial overlays on huge three dimensional models.

Sophisticated statistical analysis can only be performed if the complete data sets or at least histograms are available. They are usually collected only at some special points or for quantities of special interest. In basin modeling it is common to collect all the results of the different simulation runs at all well locations with logging information because calibration against these measurements might be performed. Additionally, the sizes of petroleum accumulations and column heights as primary targets of petroleum systems modeling are tracked over all risk runs.

It is common to define "risk points" which are special points of interest, where additionally all the data of all runs is collected. These are usually points in source rocks which are of interest for maturation and expulsion timing or points located at faults which can e.g. be important for petroleum migration.

If the spatial density of these points, with full risk data, is high and the intermediate behavior of the fields belonging to the stored values smooth, then

it is possible to interpolate the data through space for full reconstruction of each risk run. Especially, in combination with predictive methods for risking, which are discussed later in Sec. 7.5, these methods can be used to estimate full modeling results without performing the accordant expensive risk run. Such forecasted data sets can be generated very fast and used for further more sophisticated statistical analyses.

Besides the risk points it is also common to explicitly track hydrocarbon mass amounts related to layers, facies, faults or individual structures such as introduced in Sec. 6.10. Especially the characteristic HC masses of a petroleum system are subjected to statistical analysis (Sec. 6.10.2). Analyses of individual reservoir structures and accumulations raise similar problems for identification and tracking of a structure in different risk runs as in different events (Sec. 6.10.3). It can be solved in the same manner as in Sec. 6.10.3 just by treating a risk run similarly as a paleo event. Again, the same structure and accumulation tracking problems arise but are assumed to decrease with an increasing number of grouped drainage areas.

7.3 Bayesian Approach

Calibration can be non–unique or numerically unstable dependent on the available data. A bayesian approach for generalized calibration is presented in this section. It can be read almost independently from the rest of this chapter and can also be skipped if calibration topics are not of special interest.

In the following, it is assumed that N calibration data values $\mathbf{d}^T = (d_1, \ldots, d_N)$ are available. They are measured values and have an error, so they can be described by $d_i \pm \Delta_i$. Further, it is assumed that M uncertainties x_k exist. Performing a simulation run with fixed values $\mathbf{x}^T = (x_1, \ldots, x_M)$ yields a model with simulation results $f_i(\mathbf{x})$ as model data, which can be compared to the calibration data.

In arbitrary calibrations it is possible to calculate the probability of how calibration data fit a given model. Under the assumption of small error bars and a statistical independency of measurements belonging to the data points, it is postulated that the measurement values are normally distributed. So, the probability of how well a model fits the calibration data is given by

$$p(\mathbf{d}|\mathbf{x}) \propto \prod_{i=1}^{N} \exp\left[-\frac{1}{2}\left(\frac{d_i - f_i(\mathbf{x})}{\Delta_i}\right)^2\right] . \tag{7.11}$$

Calibration does now imply a search of the values $\mathbf{x}$, which fit the calibration data best. An obviously good criterion for the best fit is looking for the highest probability, which is called the “Maximum Likelihood” method in statistics (Beyer et al., 1999). Caused by the minus sign in the exponent of (7.11) it is equivalent to the search for the minimum of

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{d_i - f_i(\mathbf{x})}{\Delta_i} \right)^2 \tag{7.12}$$

which is the classic chi–square formula for fitting models to data. Its interpretation is easy and can be visualized for simple cases: Equation (7.12) becomes a linear regression of a straight line with the assumption of $M = 2$, a simple "simulator" $f_i = x_1 r_i + x_2$ and data values measured at some locations r namely $d_i = d_{r_i}$ (Fig. 7.20).

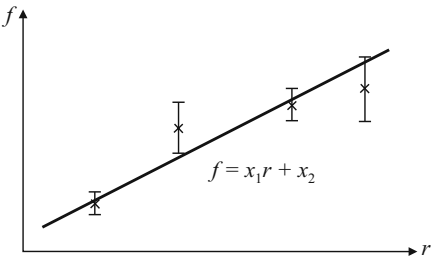

Fig. 7.20. Regression for the fit of a straight line through a "cloud" of measurement data. The inverse of the error bar size determines the weight of each point

Some problems can arise with (7.12) in practice. Often it occurs that a calibration is not unique (Fig. 7.21). In such a case the practitioner would choose a value somewhere out of the middle or at the highest probability of the according uncertainty distribution of the heat flow.

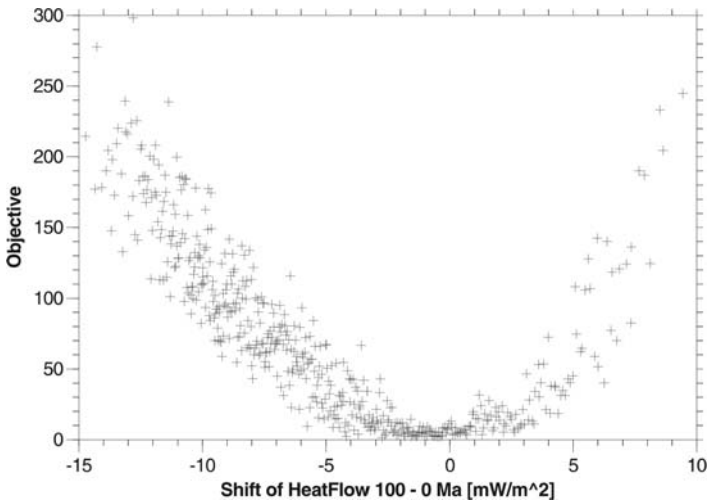

Fig. 7.21. Example with $M = 5$, $N = 19$, and objective χ^2 plotted against heat flow shift, which is known to be the most sensitive parameter. The calibration of heat flow is not unique in the range of $-5 \ldots 3\,\mathrm{mW/m^2}$

Even worse, occasionally a calibration is numerically unstable or yields completely unrealistic results caused by insufficient data points combined with some outliers. Such awkward effects can already be found in mathematically very simple situations: Without loss of generality $\sigma_i = 1$ is assumed for the following discussion. Further a simple "linear simulator" with

$$f_i = \sum_{j=1}^{M} R_{ij} x_j \qquad \text{or} \qquad \mathbf{f} = \mathbf{R} \cdot \mathbf{x} \tag{7.13}$$

is studied. The matrix $\mathbf{R}$ describes this simple "linear simulator". It is easy to show that minimizing (7.12) directly leads to

$$\mathbf{R}\mathbf{x} = \mathbf{d} \tag{7.14}$$

which is a set of linear algebraic equations with $\mathbf{x}$ as unknowns. This leads directly to the following statements:

1. The inverse R^{-1} does not exist in general, especially if $M > N$ which denotes that a unique calibration is not possible.
2. If R^{-1} exists it could be numerically unstable (Press et al., 2002).
3. If a solution is found, it is not ensured to be physically or geologically meaningful.

The first statement means that calibration data could be insufficient for calibration. For example present day temperature data alone is never sufficient for paleo–heat flow calibration. The second statement expresses that calibration data might be inconsistent leading to possibly different calibration scenarios and the third states that calibration might be optimal outside of the allowed parameter range of the model, e.g. a negative thermal conductivity.

The problem now, is how to get rid of these possibly awkward calibration behaviors and introduce a method which is at least as good as the workflow of the practitioner.

A possible solution could be a so called "Singular Value Decomposition" of the matrix R (Press et al., 2002). This is a projection on parameters which can be calibrated numerically stable with the available data. The rest of the uncertainty parameters are ignored. This method has two drawbacks: First of all it is only well defined for linear problems such as the "linear simulator". A generalization to non–linear problems would be very complicated if possible at all. Second, there is still a problem with the parameters which cannot be calibrated. Which value should they have?

Regularization is another attempt which can be tried. Instead of minimizing χ^2 it is proposed to minimize

$$\chi^2 + \lambda \left(\mathbf{x}^T \mathbf{x} \right) \tag{7.15}$$

with a number λ which has to be selected properly. It is easy to see that at least the first and second statement are solved with this method because (7.14) changes to

$$\left(\mathbf{R}^T\mathbf{R} + \lambda\mathbf{1}\right)\mathbf{x} = \mathbf{R}^T\mathbf{d}$$

which are regularized normal equations of (7.14).[4] But which value should be taken for λ? A few ideas can be found in (Press et al., 2002) but in general the problem remains unsolved.

All three problems are of principle nature and it is necessary to go back to the basics of the probability definition (7.11). It is written in conditional form stating a probability of calibration data fitting a given model. Instead it is possible to evaluate the probability of models fitting given calibration data following Bayes law

$$p(\mathbf{x}|\mathbf{d}) = p(\mathbf{x})\frac{p(\mathbf{d}|\mathbf{x})}{p(\mathbf{d})} \tag{7.16}$$

which leads down to the roots of probability theory and logic (Jaynes, 2003; Robert, 2001). The term on the left side is called the "posterior", the classical probability (7.11) is situated in the nominator and called the "likelihood" and the first term $p(\mathbf{x})$ on the right side the "prior". The term in the denominator does not play a central role, it is for normalization only.

It is possible to evaluate (7.16) similar to (7.11) under the assumption that all distributions have Gaussian form. One yields a minimization rule for the objective function Φ with

$$\Phi = \sum_{i=1}^{N}\left(\frac{d_i - f_i(\mathbf{x})}{\Delta_i}\right)^2 + \sum_{i=1}^{M}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2 \tag{7.17}$$

where μ_i and σ_i are means and standard deviations of the uncertainty distributions. The first term on the right side is the "classical" χ^2 followed by an additional term. It is derived from the uncertainty distributions and implies that the knowledge for the definition of their shape has the same value as the knowledge about error bars of calibration data and should be taken into account with the same weight for calibration. The knowledge entering the definition of the uncertainty distributions is therefore called "prior information" and the distributions often just "priors".

In case of the linear simulator, the objective

$$\Phi = \sum_{i=1}^{N}\left(\frac{d_i - \sum_{j=1}^{M} R_{ij}x_j}{\Delta_i}\right)^2 + \sum_{i=1}^{M}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$$

must be minimized. This formula has basically the same form as (7.15). In case of $\mu_i = 0$ and $\sigma_i = \sigma$ for all i they are the same with $\lambda = 1/\sigma^2$.[5] It is

[4] This is only half of the truth because it is known that normal equations usually have worse numerical properties than their "non–normal" counterparts. Depending on the numeric value of λ, the stability can still be a problem.

[5] This relation yields some additional hints to how parameters such as λ should be chosen in regularization problems.

easy to see that the first and second statements concerning the existence and numerical stability of R^{-1} vanish with the usage of (7.17).

The term associated with the prior tries to move the calibration in the direction of the μ_i where the center of the distribution is located. The most extreme case would be if there were no given calibration data. Then (7.17) would lead to $x_i = \mu_i$. In general, one can assume that the priors are defined for physically and geologically meaningful parameter ranges. The attraction of the parameters into this region by the prior therefore ensures meaningful solutions and solves for the problem of the third statement. This behavior automates the procedure of the practitioner.

On the other hand if either a huge amount of data or qualitatively very good calibration data is available the prior term can be neglected and calibration approaches the classical χ^2 method. In the intermediate region both terms balance Φ in the same way as different data points balance pure χ^2 calibration.

The discussion is the same for the nonlinear case (7.17) and therefore it is expected that the prior term removes the problems associated with all three statements in almost all cases.

The formula (7.17) provides the very simple interpretation that an uncertainty parameter is used for calibration in exactly the same manner as a calibration data point. For example, if definitions of uncertainty distributions are deduced from measurements, there is no reason why they should not be used for calibration in the same way as calibration data.

A calibration with the classical χ^2 takes only calibration data into account, whereas calibration with the objective Φ calibrates the whole model including parameter uncertainties as well as calibration data uncertainties.

The important point about a Bayesian approach for calibration is the definition of the prior distribution. If it is derived from measurements with error bars, everything is o.k. But very often priors are defined just through the experience of the modeler. So, e.g. the basement heat flow is simply known not to be below $20\,\mathrm{mW/m^2}$ and never to be above $140\,\mathrm{mW/m^2}$. With the definition of a prior such knowledge is taken quantitatively into account and must now withstand critical considerations.

An iterative refinement of uncertainties as feedback of risk results is not allowed in the Bayesian approach because independency of all calibration data values must be ensured. This is completely different to the classic approach where calibration error bars are usually mapped to uncertainty parameter ranges. These ranges are afterwards often taken as “obvious” limits for uncertainty distribution definitions. Distributions constructed in such a way are not allowed to be used as priors in objective functions. Nevertheless, they are often a good choice for Monte Carlo simulations in general.

7.3.1 Prior Information of Derived Parameters

It is sometimes problematic to use the Bayesian approach with uncertain derived parameters. For example the shift of a whole basement heat flow in a huge basin model is taken into account with the same weight as e.g. one measured bottom hole temperature value by (7.17). The weights are only given by the size of the uncertainty but one shift of a basement heat flow, shifts many grid values. Its prior knowledge is not based on experience alone but form a variety of argumentations, consistency arguments, indirect observations, etc. So the prior information which enters the calibration is obviously larger than assumed by (7.17). For that reason it should be possible to increase the weight of such a parameter.

There is no fixed rule for how much the weight should be increased. If the relative uncertainties of the calibration data and the model parameter are of the same size, then the prior term should be multiplied by approximately the number of correlated calibration data points. If it would be much less, the prior term would not affect the calibration and if it would be much bigger, the calibration data would not show significant contributions. In balance, the prior information is believed to be as important as the calibration data itself, which is a reasonable starting point in many cases.

In the example which is shown in Figs. 7.21 and 7.22 the uncertainty in the heat flow shift could be reduced by about two thirds by using the Bayesian approach.

7.3.2 Correlations of Priors

Correlations of priors can be directly taken into account in the Bayesian approach. Instead of using only the diagonal elements of the covariance matrix in (7.17) the whole matrix Σ which is explicitly written down in the two dimensional case in (7.9) must be used:

$$\Phi = (\mathbf{d} - \mathbf{f}(\mathbf{x}))^T \mathbf{C}^{-1} (\mathbf{d} - \mathbf{f}(\mathbf{x})) + (\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \; . \tag{7.18}$$

Here a matrix notation with $C_{ij} = \Delta_i \delta_{ik}$ with $\delta_{ik} = 1$ for $i = j$ and $\delta_{ik} = 0$ else was chosen.

7.3.3 Prior Information of Nominal Uncertainties

Nominal distributions (Fig. 7.13) need a special treatment in the Bayesian framework. For example, one continuous uncertainty of the objective is

$$\Phi = \chi^2 + \Phi_{pc} \tag{7.19}$$

with the continuous prior term

$$\Phi_{pc} = \left(\frac{x - \mu}{\sigma} \right)^2 \tag{7.20}$$

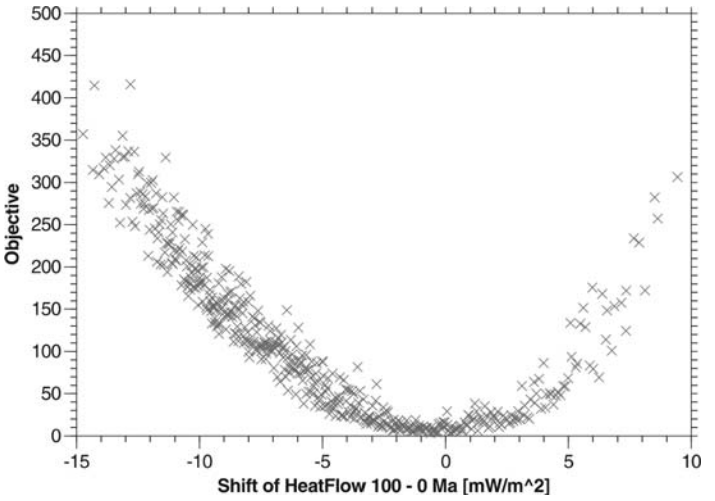

Fig. 7.22. The same example as Fig. 7.21 but with Bayesian objective Φ plotted against the heat flow shift. The calibration is almost unique with a shift in the range of $-2 \ldots 1\,\mathrm{mW/m^2}$. An extra prior weight of 19 was assumed for the heat flow shift and further rise narrows the range and moves it continuously into the direction of the master run without shift

but for nominal distributions a mean μ or a variance σ does not exist by definition.

In Sec. 7.2.5 association was used instead of correlation. Variance is an "auto–correlation" so it is obvious to try

$$\Phi_{pn} = \sum_{i=1}^{n} \frac{(N_i - n_i)^2}{n_i} \tag{7.21}$$

with n defined as the number of bins of the distribution, $n_i = Np_i$ with N as the number of samples, p_k the probability of the bin k and N_i the number of samples in bin i. It is $\sum_{i=1}^{N} p_i = 1$. Equation (7.21) is known to follow χ^2 statistics as well as its continuous counterpart Φ_{pc} (Press et al., 2002).

The prior must be calculated for one run so $N = 1$ and $N_i = \delta_{ik}$ with k as the bin of the drawn sample. Evaluation of (7.21) yields

$$\Phi_{pn} = \frac{1}{p_k} - 1\,. \tag{7.22}$$

This is a reasonable choice because the objective Φ_{pn} is decreasing with increasing p_k similar as Φ_{pc} with σ^2. With rising uncertainty the prior becomes less important.

In practice, it is possible to add a constant which does not influence the minimization procedure of the objective and use

$$\Phi_{pn} = \frac{1}{p_k} - \frac{1}{p_m} \tag{7.23}$$

with m as the index of the bin of the master run instead of (7.22). This has the advantage that $\Phi_{pn} = 0$ for the master run. If $x = \mu$ is chosen in the continuous case then it is analogously $\Phi_{pc} = 0$ and $\Phi = \chi^2$ for the master run.

7.4 Deterministic Sampling

In the previous section it was shown that the Monte Carlo method is very general. Many topics such as risking, calibration and understanding could be treated simultaneously by just analyzing the results of one Monte Carlo simulation. Occasionally, one is interested only in special questions which are not related to the topic of general risking. In these cases a random and global sampling of the space of uncertainty is often not necessary anymore and it is possible to avoid expensive simulation runs.

The most extreme cases are special algorithms for highly specialized questions, e.g. one is only interested in classical calibration. This can be seen as a minimization problem of one χ^2 function. Special algorithms exist to optimize such a minimization (Press et al., 2002). Expensive simulations are avoided and high numerical accuracy is achieved. However, such algorithms have some serious drawbacks. First, in basin modeling high numerical accuracy is usually not needed because of many uncertainties. Second, these algorithms are so specialized that expensive simulations performed for a minimization of χ^2 are not reusable for a minimization of Φ, which is also often an issue.

Other disadvantages are technical in nature (e.g. bad parallelization properties) because many sophisticated algorithms are of primarily sequential nature, e.g. following a gradient downhill to the minimum.

Thus one is looking for methods, which are more efficient than arbitrary Monte Carlo simulations, for the price of losing generality and which are less special than sophisticated algorithms with high numerical accuracy. Obviously, the targets of interest must be specified exactly before starting to search for appropriate methods.

Risking is the part of Monte Carlo simulations which is most dependent on the random structure of sampling caused by the nature of probabilities. Thus one has to dispense with risking in its general form. On the other hand, one does not need to dispense with "simple" risking such as the calculation of minimum and maximum scenarios. Meaningful targets are hence "simple" risking, calibration, and "simple" understanding as far as understanding can be found without the calculation of statistical quantities.

Other targets, which will be treated more explicitly in Sec. 7.5, are inter– and extrapolation techniques between different simulations for forecasting results. Abdication of risking is not necessary anymore. It can be studied with forecasted models.

7.4.1 Cubical Design

The most simple uncertainty sampling design, which fulfills the conditions of the previous discussion, is simply sampling all combinations of minimum and maximum choices of all uncertainty parameters. In uncertainty space this has the form of a (hyper)cube (Fig. 7.23) and is therefore called cubical design.[6]

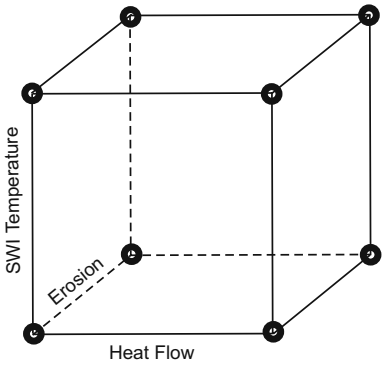

Fig. 7.23. Example of cubical design in uncertainty space with three parameters. The bold circles depict the parameter choices for the simulation runs

Cubical sampling can be used for "simple" risking, calibration, "simple" understanding, and forecasting (Sec. 7.5): the topic of "simple" risking is solved under the assumption of "non–pathological" behavior of the simulator. In such a case, minimum and maximum values of uncertainty parameters would map to minimum and maximum simulation results and thus to result ranges similar to error bars. It is clear that complicated processes such as migration cannot be treated this way.

Calibration is performed analogously. Cubical design samples the uncertainty space regularly and so simulation results are easy to interpolate for good calibration.

Understanding is improved because at least all extreme combinations are simulated. Again under the assumption of smoothness it is possible to calculate interaction effects out of the results (Montgomery, 2001). In general, all results can be inter– and extrapolated, e.g. linearly, which is forecasting. More about this in Sec. 7.5.1.

Nevertheless, cubical design has the serious drawback that the number of simulation runs to be performed increases exponentially with the number of uncertain parameters. For that reason one is often forced to omit certain combinations. There exists a whole theory of "Design of Experiments (DOE)" treating problems such as this (Montgomery, 2001). Keywords are "fractional factorial design" for omitting special combinations of uncertainty parameters or "screening". Screening is important because it tries to find the important

[6] It is exactly cubical if the units of the uncertainties are chosen so that the minimum and maximum values have the same numerical value for all parameters.

and sensitive parameters. The other parameters could be omitted, which drastically reduces the number of combinations. "Blocking" is another important method, which omits sensitive parameters for the better recognition of the effects of the less important parameters. Other designs such as pure cubical designs are proposed, too, e.g. cubical face centered, which obviously can be very valuable. But the theory was created for engineering needs based on real experiments and not on deterministic reproducible computer simulations. Thus only basic ideas such as screening or blocking can be transferred.

Simulation runs for the cubical design can be performed in parallel. The number of parallel runs is restricted to one cube. Multiple cubes themselves are evaluated sequentially if their design is iteratively refined as proposed in Montgomery (2001). Thus simulation runs can to some degree be performed in parallel but not in such a general way as for arbitrary Monte Carlo simulations.

Cubical designs are very valuable for fast uncertainty analyses especially if the number of uncertainties is small.

7.4.2 Other Deterministic Designs

A method that is similar to latin hypercube sampling is Sobol' sequence of "quasi–random" numbers (Press et al., 2002). It guarantees a smoother and more homogeneous sampling than pure "pseudo–random" sampling.[7] The sampling is smoothly refined by increasing numbers in the sequence and it is not influenced by extra parameters such as strip widths.

Sobol' sequence generates quasi–random numbers, which are calculated in any case with a deterministic algorithm, whereas LHC sampling is based on pseudo–random numbers.[8] It can therefore be expected, that LHC sampling generates random numbers with better statistical properties.[9] Additionally, an implementation of Sobol' sequence, such as in Press et al. (2002), does not allow an arbitrary number of independent and different sampling realizations, which can be achieved easily for LHC sampling by different initialization of the random number generator. Independently created samplings can thus usually not be merged to one large with better statistical properties. In practical work this is e.g. a drawback for parallel processing or merging of different risk scenarios. Workarounds, such as sequential precalculation of the random numbers, must be performed (Bücker et al., 2008).

[7] "Pseudo–random" numbers are deterministic numbers which are generated in a way that they pass statistical tests for random numbers. Therefore they are random in practice. "Quasi–random" numbers only appear to be "random".

[8] Obviously, it is even possible to combine LHC sampling with real random numbers.

[9] Better statistical properties are here defined as a larger number of passed statistical tests.

7.5 Metamodels

Interpolation and extrapolation of results between different simulations is forecasting. A method which forecasts all important results is called a "metamodel" or a "surrogate" (Simpson et al., 1997). A metamodel can be used for everything that can be done with the corresponding "real model", e.g. risking.

Metamodels are very important in basin modeling because of the high simulation effort, especially the long simulation times of 3D basin models. In contrast, forecasting is usually very fast, often by a factor of more than a million.

Forecasting does usually not produce the exact results but only approximations. Thus metamodels are often restricted in their applicability. Highly non–linear effects such as hydrocarbon spilling can usually not be forecasted with metamodels.

An overview of common metamodeling methods is given in the following. After this the usage of metamodels for calibration is discussed. Based on the high performance of response surface based metamodels, it is shown that Markov chain Monte Carlo algorithms can be applied.

7.5.1 Response Surfaces

The usage of response surfaces for interpolation and extrapolation is very popular in many fields of science. Very good textbooks are available (Myers and Montgomery, 2002; Box and Draper, 1987; Khuri and Cronell, 1996; Montgomery, 2001). The method is also becoming popular in basin modeling (Wendebourg, 2003).

Response surfaces are low order multivariate polygons which are fitted with least squares regression techniques to the simulation results. Thus, for example a model $f(x_1, x_2)$ with two uncertainty parameters, is typically approximated by

$$f \approx b_0 + b_1 x_1 + b_2 x_2 + b_{11} x_1^2 + b_{22} x_2^2 + b_{12} x_1 x_2 \tag{7.24}$$

with b_i and b_{ik} calculated from a least squares fit (Fig. 7.24).

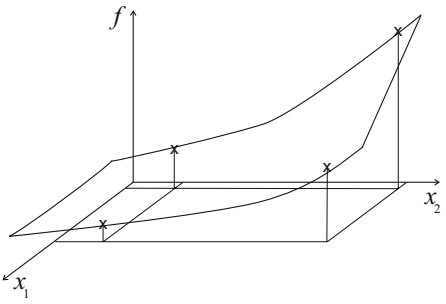

Fig. 7.24. Illustration of a response surface with two uncertainty parameters x_1 and x_2. Crosses indicate simulation results for given x_i e.g. temperatures for given heat flow and bulk conductivity values. Generally, they do not match the response surface exactly

Therefore response surfaces are ideal for approximating smooth and continuous dependencies. Discontinuities and oscillations cannot be reproduced.

It is common to introduce a short hand notation for quadratic terms so that (7.24) becomes

$$f \approx b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_5 x_5 \quad \text{with} \quad x_3 = x_2 x_2,\ b_3 = b_{22},\ \ldots\ .$$

In general there are k parameters x_j with $j = 1, \ldots, k$. The number k is determined by the number of uncertainties M and it is $k = M$ for linear response surfaces or $k = M(M+3)/2$ for approximations including quadratic terms.

For the fit, some data values $y_i = f_i(x_{i1}, \ldots, x_{ik})$ of already performed simulations are needed. So finally a vector $\mathbf{b}^T = (b_0, \ldots, b_k)$ for optimization of approximation

$$y_i \approx b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$$

or in vector notation $\mathbf{y} \approx \mathbf{Xb}$ with $X_{ij} = x_{i,j-1}$ for $j > 1$ and $X_{i1} = 1$ is searched. A least squares fit results in minimization of $(\mathbf{y} - \mathbf{Xb})^2$ and evaluation yields [10]

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\ . \tag{7.25}$$

A measure of goodness σ_g of this approximation can be evaluated by

$$\sigma_g^2 = \frac{(\mathbf{y} - \mathbf{Xb})^2}{N}\ . \tag{7.26}$$

This simply denotes the quality of a fit by summing up the quadratic deviations and dividing through the number of points.[11] A safer alternative, which takes outliers into account, can be defined by the maximal deviation

$$\sigma_g = \max_i \left| y_i - \sum_k X_{ik} b_k \right|\ . \tag{7.27}$$

Design forms of cubical type are very often used as the "natural" sampling procedure for the creation of response surfaces (Myers and Montgomery, 2002; Montgomery, 2001). Under the assumption of smooth behavior of the approximated model, it is obvious that cubical design is an effective sampling

[10] Equation (7.25) is known to be highly unstable and badly conditioned in many practical examples. More robust for the solution of $\mathbf{b}$ is a decomposition of $\mathbf{X}$ into singular values and direct solution of $\mathbf{y} = \mathbf{Xb}$ in appropriate subspaces (Press et al., 2002).

[11] At first glance this seems to be in contradiction to unbiased estimators of variance such as $\sigma^2 = (\mathbf{y}-\mathbf{Xb})^2/(N-k-1)$ and defined in Myers and Montgomery (2002). But this formula is linked tightly to the variation of $\mathbf{b}$ under random variation of $\mathbf{X}$ due to measurement uncertainties. This is a completely different objective.

strategy as all minimum–maximum combinations are studied. Additionally, the number of unknowns to be determined for a quadratic response surface $k+1 = M(M+3)/2+1$ almost matches, in cases of small numbers of uncertainties, M the number of simulations 2^M+1 which have to be performed:[12]

M	1	2	3	4	5	...
$k+1$	3	6	10	15	21	...
2^M+1	3	5	9	17	33	...

Thus only an optimal small number of simulations have to be performed and good matches at the points of simulation itself, leading to small σ_g, are enforced.

In Figs. 7.25 and 7.26 two typical diagrams of response surface models are shown. The formulas describing these isolines can easily be extracted and used for further studies. A high value of the coefficient of a cross term, e.g. x_1x_2, indicates an interaction between the impact of the corresponding uncertainty parameters. This is information which may help to better understand a model.

In Fig. 7.26 negative values for the transformation ratio appear. This is due to the polygonal form of the method. It can therefore not be used in these regions. Generally, it often occurs that simulation results vary faster than can be approximated with simple polygons. In such cases response surface modeling is often performed only in limited regions with adapted sampling of the uncertainty space (Montgomery, 2001).

Another example, which could not be treated in general by response surfaces, are pressure calibrations via variation of permeabilities, especially if permeability is expressed in logarithmic units. Pore pressure is restricted to the lower limit by hydrostatic pressure and to lithostatic pressure at the upper limit but a polynomial fit of response surface type is generally unlimited for infinitely increasing or decreasing uncertainty parameters. However, in a limited region of permeability variations a smooth behavior of pore pressure with good fitting response surfaces can be obtained.

The creation of a metamodel is mainly determined by the solution of a linear set of equations of dimension $(k+1)\times(k+1)$ for each result point which is modeled. In case of calibration purposes this number is given by the number of calibration points N. Thus, the creation of a response surface metamodel is performed within seconds on modern computers because in practice mostly $M < 10$ and $N < 1000$.

The calculation of response surface metamodel results is an evaluation of simple polygons and this is extremely fast. In most computer applications this appears to be almost instantaneous.

7.5.2 Fast Thermal Simulation

Fast thermal simulation is a special method of fast heat flow analysis (Nielsen, 2001). It is based on the approximative linear form of the partial differential

[12] All combinations plus master run.

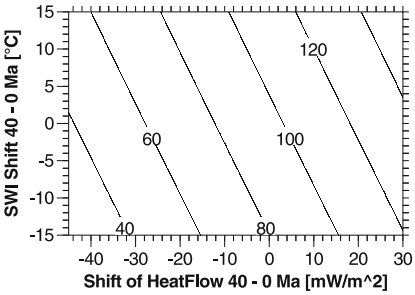

Fig. 7.25. Response surface isolines for temperature in Celsius in a source rock. Dependent on heat flow and SWI temperature variations the isolines are linear as expected

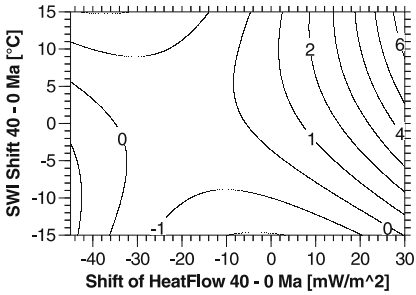

Fig. 7.26. Response surface isolines for transformation ratio in [%] at same point as in Fig. 7.25. Negative values indicate a region where the response surface method cannot be used

equation

$$\rho c \frac{\partial T}{\partial t} - \boldsymbol{\nabla} \cdot (\boldsymbol{\lambda} \cdot \boldsymbol{\nabla} T) = Q \tag{7.28}$$

of heat flow. Here T is the temperature, $\boldsymbol{\lambda}$ are the thermal conductivities, t is the time, ρ the density, c the specific heat capacity, and Q are external heat sources. For a unique compilation of heat flow analysis, boundary and initial conditions must be specified. At the top of the basin usually the temperature is given, at the sides, a condition of prohibited horizontal heat flow is applied and at the bottom, basement heat flows are specified:

$$\begin{aligned} T &= T_{\mathrm{SWI}} && \text{on top,} \\ \boldsymbol{\nabla} T &= 0 && \text{at the sides,} \\ -\boldsymbol{\lambda} \cdot \boldsymbol{\nabla} T &= \mathbf{q} && \text{at the bottom,} \\ T|_{t=\mathrm{initialtime}} &= T_0 && \text{at initial time} \end{aligned} \tag{7.29}$$

with T_{SWI} the "Sediment Water Interface" temperature at top of the basin, $\mathbf{q}$ the basement heat flow, and T_0 the temperature profile at initial time.

If $\boldsymbol{\lambda}$, ρ, and c are assumed to be smooth and weakly temperature–dependent then (7.28) is almost linear. This property can be utilized:

Firstly, one should take a look at the following boundary value problem:

$$\begin{aligned} \rho c \frac{\partial \tilde{T}}{\partial t} - \boldsymbol{\nabla} \cdot (\boldsymbol{\lambda} \cdot \boldsymbol{\nabla} \tilde{T}) &= 0 \text{ with} \\ \tilde{T} &= 0 && \text{on top,} \\ \boldsymbol{\nabla} \tilde{T} &= 0 && \text{at the sides} \\ -\boldsymbol{\lambda} \cdot \boldsymbol{\nabla} \tilde{T} &= \tilde{\mathbf{q}} && \text{at the bottom and} \\ T|_{t=\mathrm{initialtime}} &= 0 && \text{at initial time.} \end{aligned} \tag{7.30}$$

A temperature profile $T + x_q \tilde{T}$ with T as the solution of (7.28) with boundary and initial conditions (7.29) and $\tilde{T}$, a solution of (7.30) is a solution of (7.28) with boundary and initial conditions such as (7.29), which must only be modified at the bottom by $-\boldsymbol{\lambda} \cdot \boldsymbol{\nabla} T = \mathbf{q} + x_q \tilde{\mathbf{q}}$. Herein, x_q is just an arbitrary number, which can be interpreted as a derived uncertainty parameter for a variation of form $\tilde{\mathbf{q}}$.

The important point is that with only two solutions T and $\tilde{T}$ one can construct multiple solutions $T + x_q \tilde{T}$ for heat flow variations $\mathbf{q} + x_q \tilde{\mathbf{q}}$ for any value x_q just by linear combination. The form of $x_q \tilde{\mathbf{q}}$ defines the space of possible heat flow variations. Obviously, multiple variations $x_{q,i} \tilde{\mathbf{q}}_i$ can be combined just by summing up the solutions $x_{q,i} \tilde{T}_i$. Hence it is possible to quickly create flexible variations of the original heat flow and temperature pattern.

It has been proposed by Nielsen (2001) to vary the heat flow below each of the four model corners.[13] Each paleo–heat flow map is calculated by interpolation with two dimensional form functions analogously to (8.22). A corner variation with its shape function states one variation $\tilde{\mathbf{q}}_i$ with $i = 1, \ldots, 4$ for all four corners. The sum of all four corner variations describes tilting and twisting variations of the original heat flow distribution. The method should not be applied in cases when mismatches to the measured data require heat flow shifts very differently in many locations, e.g. from well to well. For such cases methods as described in Sec. 3.9 are more advantageous.

Additionally it must be noted, that $\tilde{\mathbf{q}}$ can vary independently to $\mathbf{q}$ through time. Hence heat flow variations in time can be incorporated. For example, it is possible to shift the heat flow of each of the four corner points linearely with time.

The set T and $\tilde{T}$ can be interpreted as a metamodel for forecasting manifold heat flow histories.

A response surface which is created as an interpolation of two different solutions T_1 and T_2 of (7.28) with the boundary conditions (7.29) once with bottom heat flow $\mathbf{q}$ and the other with $\mathbf{q} + \tilde{\mathbf{q}}$, yields almost the same results as the fast thermal simulation caused by the linearity of the differential equation system. The main difference between both methods comes from the fact, that the differential equation is usually not exactly linear. Parameters such as the thermal conductivity are typically weak but non–linearly temperature dependent. The response surface can then be interpreted rather as a "secant approach", whereas the fast thermal simulation is following more a "tangent" (Fig. 7.27). Response surfaces can incorporate smooth non–linearities to some degree of accuracy with their quadratic terms (7.24). On the other hand, as they are caused by the linear regression, they do not need to match the simulation results from which they were created exactly, whereas fast thermal

[13] It is assumed here that the model has a rectangular base area. Generally, it does not matter if the corner points are not inside of the model.

simulations reproduce, at least in a region of small variation, the original models exactly.

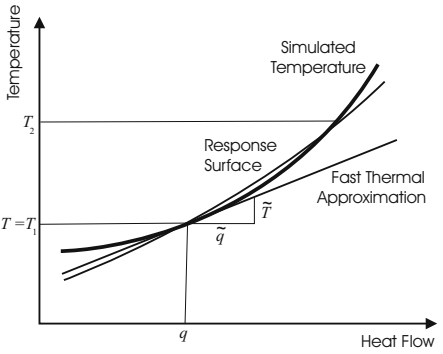

Fig. 7.27. Illustration of the differences between response surface and fast thermal simulations. The curvature of temperature, which comes from the nonlinearities is exaggerated for demonstration. Due to quadratic terms the response surface is able to approximate non–linearities whereas the fast thermal simulation approximation is restricted to a straight "tangential" line. The response surface does not in general need to match the temperature exactly at any point whereas the fast thermal simulation matches at $T = T_1$

The calculation of T and $\tilde{T}$ needs about the same effort. The evaluation of a forecast is just the evaluation of $T + x_q \tilde{T}$ and thus can be performed almost instantaneously. Thus the effort and the needed resources for the creation of the metamodel as well as the evaluation performance of the fast thermal simulation and the response surface are almost the same.

Fast thermal simulations are not limited to heat flow variations only but can also be applied to thermal conductivity variations. This is achieved by introducing a derived uncertainty parameter x_λ for the variation of thermal conductivity according to $\boldsymbol{\lambda} + x_\lambda \hat{\boldsymbol{\lambda}}$ with a $\hat{\boldsymbol{\lambda}}$ describing the form of the variation. The solution $\hat{T}$ of the differential equation

$$\boldsymbol{\nabla} \cdot (\boldsymbol{\lambda} \cdot \boldsymbol{\nabla} \hat{T}) = -\boldsymbol{\nabla} \cdot (\hat{\boldsymbol{\lambda}} \cdot \boldsymbol{\nabla} T) \tag{7.31}$$

with boundary conditions such as (7.29) but with $\mathbf{q} = 0$ can be added to the solution T in the same manner as $\tilde{T}$ to construct valid heat flow histories for thermal conductivity variations. Here, the fast thermal simulation is restricted to "small" variations in conductivity because quadratic terms are neglected in the deviation of (7.31).

7.5.3 Kriging

Kriging is another method for interpolation and extrapolation in multi dimensional spaces. It is based on the minimization of statistical correlations and derived as the best linear unbiased estimator. Originally it was developed for spatial inter– and extrapolation only but it can also be applied to abstract uncertainty spaces. Various different methods of kriging exist. To the authors

knowledge it has not been applied up to now in any case as a metamodel in basin modeling and thus we refer only to the literature of geostatistics (Davis, 2002). However, it can be expected that kriging might yield good results in many cases especially when simple functions such as (7.24) are not appropriate at all for the description of the model or process of interest.

7.5.4 Neural Networks

Neural Networks can be interpreted as metamodels.

Neural Networks must be trained. They learn. Three classes of learning are usually distinguished (Zell, 1997):

- supervised learning
- reinforcement learning
- unsupervised learning

Supervised learning is based on comparison with correct results. These results are simulation results in basin modeling. Supervised learning is usually the fastest way of learning. Nevertheless many expensive simulation runs must be performed for this way of learning.

Reinforcement learning is based on reduced feedback. The network is taught only with information about the correctness of its output but not the correct result itself. Therefore reinforced learning networks need even more training than supervised learning networks. Although the amount of feedback data is small it, too, must be available. This means that many expensive simulation runs must be performed for this method.

Unsupervised learning is performed without feedback. The network should learn by classifications in its own right e.g. by "self organization". Caused by the complexity of a typical basin model, it is expected that unsupervised learning neural networks will be improper for result predictions.

The high effort for learning leads to the conjecture that neural networks are not the best alternative for basin metamodeling.

7.5.5 Other Methods for Metamodeling

Methods such as rule based expert systems or decision trees are obviously limited in their applicability for forecasting. Under special circumstances they can be interpreted as metamodels but not in general.

Other special techniques are based on analysis in frequency space. Due to the complex geometry in geology, these methods cannot usually be applied to basin modeling.

7.5.6 Calibration with Markov Chain Monte Carlo Series

The Markov chain Monte Carlo (MCMC) method is designed for the sampling of multi–variate probability distributions (Neal, 1993; Besag, 2000).

Calibration is a search of high probability regions where data fits the model which is not trivial in high dimensional spaces. So MCMC can be misused to find the regions of calibration. Additionally, the sampling allows the error bars of measurement values to be mapped to uncertainty distributions.[14] The whole subject of search and mapping is called "inversion" and thus MCMC is also a method for inversion.

Several different algorithms for MCMC exist, which can be shown to be directly related (Neal, 1993; Besag, 2000). This section is restricted to the classical Metropolis algorithm. It basically works as follows:

According to a distribution, with some special properties which are of no interest here, random jumps are performed in uncertainty space. If the probability density of the distribution, which should be sampled increases, the jump is accepted. If the density decreases it can be rejected or accepted by a special criterion with a random level of acceptance. This ensures that MCMC also samples low probability regions but it focuses primarily on the highly probable regions. In Fig. 7.28 such a MCMC "random walk" is illustrated. Sampling can only be performed on a small fraction of jumps typically every 100th or less to ensure independency of the samples.

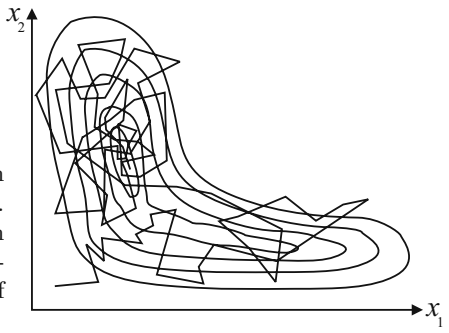

Fig. 7.28. Illustration of Markov chain Monte Carlo sampling with a random walk. Isolines indicate the probability density in the $x_1 - x_2$ uncertainty diagram. The "random walk" is attracted by the region of high probability

Obviously, it is clear that MCMC is not very efficient in sampling because most of the jumps which are model results are ignored for sampling, some of the jumps are rejected and additionally, the random walks can become very long before reaching their high probability calibration targets. Proper MCMC sampling can become a delicate choice of the start point and jump width. Therefore, in basin modeling MCMC is only usable with fast metamodels such as response surfaces or fast thermal simulations. On the other hand, in theory MCMC guarantees to find the regions of interest.

[14] These distributions are not allowed to be used in a Bayesian approach, see Sec. 7.3.

Summary: Models are usually constructed on the basis of uncertain data. These uncertainties cause additional tasks during comprehensive model analysis. Firstly, modeled results must be classified according to their probability. For example, confidence intervals of special output scenarios should be specified or even more concrete a risk of failure must be quantified. Secondly, the behavior of model results with the variation of uncertain parameters should be understood. Which parameter affects which part of the result? Finally, uncertainties should be reduced by comparison with additional calibration data. The three tasks are "risking", "understanding", and "calibration".

Obviously, all three tasks can be studied with multiple simulation runs. Uncertain parameters must therefore be varied according to their range of uncertainty. Monte Carlo simulations are an effective method of treating all three tasks simultaneously. Multiple simulation runs with randomly drawn uncertainty parameters, according to their probability of occurrence, are performed. Another advantage of the approach is the possibility of unrestricted parallel processing. This is especially valuable because simulation runs are often very time consuming and therefore expensive. The method can further be optimized with latin hypercube sampling, which avoids clustering of parameter combinations.

A model can be calibrated in two different ways, with and without consideration of information which describes data uncertainties of the model, e.g. limits or ranges of an uncertain input parameter. This "prior" information is taken into account in the Bayesian approach. Ambiguous and geologically meaningless calibrations can be avoided with this approach.

Simulation runs are very time consuming. Response surfaces are a method for fast interpolation between simulation results. Other methods for rapid result prediction, such as the fast thermal simulation, are also discussed. Particularly with regard to heat flow problems, response surfaces and fast thermal simulations can be used efficiently for calibration. A very robust algorithm concerning inversion is the Markov Chain Monte Carlo (MCMC) sampling, which in principal guarantees the best possible calibration due to random jumps in the uncertainty space.

References

J. Besag. Markov Chain Monte Carlo for Statistical Inference. Working paper, Center for Statistics and the Social Sciences, University of Washington, 2000.

O. Beyer, H. Hackel, V. Pieper, and J. Tiedge. *Wahrscheilichkeitsrechnung und Statistik*. B. G. Teubner Stuttgart Leipzig, 8th edition, 1999.

G. E. P. Box and N. R. Draper. *Empirical Model–Building and Response Surfaces*. John Wiley & Sons, Inc., 1987.