Ch 3. Model and data limitations: the sources and implications of epistemic uncertainty

Jonathan Rougier
Department of Mathematics
University of Bristol

Keith Beven
The Lancaster Environment Centre
University of Lancaster

Late draft prepared for *Risk and Uncertainty Assessment for Natural Hazards*, J.C. Rougier, R.S.J. Sparks and L.J. Hill eds, 2013, Cambridge, UK: Cambridge University Press.

1. Introduction

Chapter 2 focused entirely on aleatory uncertainty. This is the uncertainty that arises out of the randomness of the hazard itself, and also, possibly, out of the responses to the hazard outcome. That chapter chased this uncertainty through the footprint function and a loss operator to arrive at an Exceedance Probability (EP) curve. Such a structured approach (e.g. as opposed to a purely statistical approach) was motivated by the need to evaluate different interventions, for choosing between different actions; and by the possibility of non-stationarity in the boundary conditions, on policy-relevant time-scales measured in decades.

Different Risk Managers will have different loss operators, and hence different EP curves. Likewise, the same Risk Manager will have different EP curves for different actions. A very simple summary statistic of an EP curve is the area underneath it, which corresponds to the expected loss ('expectation' taken in the mathematical sense), which is defined to be the *risk*. This presupposes a well-defined EP curve, i.e. a probability function F such that, if X is the loss, then

$$F_X(x) = \Pr(X \le x),$$

adopting the standard convention that capitals represent uncertain quantities and lower case represent ordinates or specified values. Given F_X, the EP curve is defined as $EP(x) = 1 - F_X(x)$.

This chapter considers the conditions under which the EP curve is itself uncertain, and examines the consequences. This is the *epistemic uncertainty* that arises in both model representations and the data used to drive and evaluate models. The source of epistemic uncertainty is limitations in the analysis—it could, in principle, be reduced with additional time and resources. Thus epistemic uncertainty arises partly out of the need for the Risk Manager to take actions on the basis of what she currently knows, rather than being able to wait. For simplicity, the presentation in this chapter is in terms of computing an EP curve for a specific action; i.e. the analysis will need to be replicated across actions if the task is to choose between actions. Section 2 presents a general framework for thinking about epistemic uncertainty, and examines its main limitations. Sections 3, 4, and 5 consider particular sources of epistemic uncertainty, and how they might be incorporated into the EP curve, and into the decision analysis. Section 6 concludes with a summary. The interested reader will also want to consult Spiegelhalter and Riesch (2011), a wide-ranging review of current practice in uncertainty assessment, with recommendations for good practice.

As in Chapter 2, the emphasis here is on the needs of the Risk Manager, whose role involves managing the decision process and recommending actions. She is answerable to several different stakeholders, each of whom is entitled (and advised) to hire an Auditor. The Risk Manager must be able to demonstrate to the Auditor(s) that her analysis is valid, and that it is suitable. Validity can be ensured by operating within a probabilistic framework, and this also promotes transparency, making it easier to demonstrate suitability. The same comments on "Why probability (as opposed to some other formal method for uncertainty quantification)?" made in Chapter 2 apply here as well. Statistics is a framework for assessing the impact of judgements on choices. It provides a set of rules, demonstrably valid, by which the mechanics of turning judgements into choices is to a large extent automated. This allows us to focus on the judgements, and, with efficient mechanics, to subject them to stress testing in the form of sensitivity analysis, which will reveal the extent to which different defensible judgements would have yielded the same choice of action.

It is understood, however, that probability, although powerful, also makes very strong demands that are hard to meet in complex situations such as natural hazards, and which entail many simplifications. This is especially true of epistemic uncertainty. Therefore this chapter suggests a somewhat pragmatic approach. It is crucial that the various sources of epistemic uncertainty are clearly identified. But the degree to which each one can be fully incorporated into the EP curve will vary. Some are straightforward (uncertainty about the hazard process, for example), some moderately challenging (uncertainty about the footprint function and loss operators), and some extremely demanding (uncertainty about external factors, such as greenhouse gas emissions in the current century).

One useful way of picturing the treatment of epistemic uncertainty advocated here is in terms of 'internal' and 'external' uncertainty, as discussed by Goldstein (2011). He states

> "Internal uncertainties are those which arise directly from the problem description. Many analyses in practice are carried out purely on the basis of assessing all of the internal uncertainties, as these are an unavoidable component of any treatment of the problem. External uncertainties are all of the additional uncertainties which arise when we consider whether the treatment of the internal uncertainties indeed provides us with a satisfactory uncertainty description … Most of the conceptual challenges ... arise in the appropriate treatment of the external uncertainties" (section 3.1).

In the standard risk assessment for natural hazards, the EP curve captures aleatory uncertainty, and everything else is external. Of course, external uncertainty is not forgotten. Instead, it is accounted for using a lumped adjustment, such as a margin for error. In reinsurance, for example, it appears as a loading which increases the premium above that of the expected loss. In flood defence design it appears as an "freeboard" allowance. In other areas of engineering it appears as an acceptable factor of safety. The message of this chapter is that it is feasible and beneficial to move some aspects of epistemic uncertainty into the EP curve, i.e. to transfer them from external to internal uncertainty. In this way, the uncertainty accounted for by the margin for error diminishes, and the difficulties associated with specifying its size become less important in the overall risk assessment, and in the choice between different actions.

2. Epistemic uncertainty in the EP curve

2.1. Recapitulation

We briefly recapitulate the main points from Chapter 2. Loss, denoted X, is expressed as a random quantity dependent on the hazard outcome, denoted ω. Probabilities are attached to each

outcome, and this uncertainty 'cascades' down to the loss. This calculation requires the enumeration of all possible hazard outcomes, and their probabilities, and for this reason is often not very practical. Instead, a random simulation approach can be used to approximate the probability distribution F_X. In this chapter, random simulation is termed 'sampling' according to standard statistical usage, and one such random sample is termed a 'realisation'. This is because the word 'simulator' has a different technical meaning, as explained in section 4.1 below.

In the random sampling approach, n realisations are independently sampled from the hazard process, and for each realisation a loss is sampled. The resulting set of sampled losses approximates the loss distribution, and summaries of this distribution such as the expectation (the *risk*) or the 99.5th percentile can be estimated, including confidence intervals to quantify variability. This was discussed in Section 5 of Chapter 2 of this volume. In this chapter we proceed as though n can be set large enough that variability can be ignored, purely to simplify the presentation.

2.2. Relaxing the suppositions

Now consider this calculation in more detail. While providing a general framework for estimating the EP curve, it involves a series of suppositions. The simplest way to conceptualise epistemic uncertainty is that it allows the Risk Manager to relax the stringent condition that the suppositions have to be true. As already discussed, she will relax this condition anyhow, in her use of the EP curve for decision analysis. For example, by including a lumped margin for error which will be larger the more unlikely the suppositions underlying the EP curve are judged to be. But by thinking probabilistically the Risk Manager can relax some of the suppositions within a formal quantitative framework, which enables her to guarantee formal validity, and also enhances transparency.

In principle, epistemic uncertainty can be fully incorporated into the EP curve, as demonstrated by the following thought experiment. Note that this thought experiment is designed to point out the challenges for quantifying epistemic uncertainty: it is *not* a template for how to proceed. First, group all of the suppositions underlying a particular EP curve under the label 'θ'. In this chapter θ will be used wherever there is a source of epistemic uncertainty. Any particular distribution function for loss should be written not as $F_X(x)$ but as $F_X(x \mid \theta)$, that is, the distribution function for loss conditional on θ. We refer to $1 - F_X(x \mid \theta)$ as a *Conditional Exceedance Probability* curve, or CEP.

Now for the thought experiment. Imagine that is it possible to enumerate all the different sets of suppositions under which a CEP curve might be constructed, denoted by the set Θ. Attach a probability to each θ in Θ, where this probability represents the Risk Manager's judgement that this θ is the appropriate θ ('right θ' is too strong here, even in a thought experiment), and where the probabilities sum to one over Θ. Then the probability calculus asserts that a collection of CEP curves should be combined into a EP curve by a vertically weighted average, where the weights are given by the probabilities over the θ values in Θ.

One difficulty with this thought experiment is that it is generally not possible to enumerate the set Θ. By the very nature of epistemic uncertainty, the enumeration of the possible θs might be incomplete (an example is the small number of assumed scenarios for future emissions that underlie the IPCC and UKCP09 projects of climate change in this century). We may appreciate quite clearly in assessing CEP curves that there are other possibilities (as in the case of emissions scenarios) but accept that, for practical limitations of computer resources or availability of measurement techniques, not all possible θs can be incorporated in the analysis.

A second difficulty is that, even were we able to enumerate a representative set Θ, it is extremely challenging to assign a probability to each member, such that the probabilities sum to one. This is

even more challenging if the set Θ is clearly incomplete, for then there will be some probability left over with nowhere to go.

But we reiterate the point made at the beginning of this chapter: the objective is to transfer some of the uncertainty in the risk from 'external' to 'internal'. Not being able to enumerate all possible members of Θ, or not being able to assign probabilities, should not prevent the Risk Manager from benefiting from those situations where she can identify some of the members of Θ, and/or provide approximate probabilities for some of the members. Therefore we describe the benefits that follow from being able to enumerate Θ, and from being able to sample from Θ. The Risk Manager must decide whether these benefits are worth the extra effort, in terms of providing a clearer justification for her choice of action. And then she must justify her decision to the Auditor.

Benefits of enumeration

Suppose that it is possible to enumerate Θ, or, more likely, to identify a representative collection of values for θ which roughly span the possible outcomes for epistemic uncertainty. In this case quite a lot can be achieved without having to assign probabilities to the individual values, if it is possible to compute the CEP curve corresponding to $F_X(x \mid \theta)$ for each θ. In particular, it is possible to compute lower and upper envelopes for the EP curve, and thus lower and upper bounds on the risk. It is a basic property of probabilities that the unconditional probability of an event always lies within the range of the conditional probabilities. Expressed in terms of EP curves,

The EP curve always lies within the lower and upper envelope created by the CEP curves.

This is illustrated in Figure 1.

[FIGURE 1 ABOUT HERE]

This result allows the EP curve to be bounded without having to make any reference to the probabilities of each of the possible outcomes for epistemic uncertainty. Therefore it may be possible for the Risk Manager to choose between actions entirely in terms of the bounds. If it is clear that the lower bound of the EP curve for action A is above the upper bound of the EP curve for action B then the probabilities that are attached to the epistemic uncertainties are immaterial as far as the risk ordering is concerned: the risk of A will always be greater than that of B. Technically to be sure of this all possible values of θ must be enumerated; but a high level of confidence will follow from a reasonable attempt to span the range of possible values for θ.

It is also sometimes possible to rule out certain choices, as being *inadmissible*. This will be discussed further below, in section 5. But, briefly, if the CEP for action A is above that of the CEP curve for action B for each θ, then action A would never be preferred to B, and A can be removed from further consideration.

Benefits of sampling

Suppose now that it is possible to sample values of θ from the distribution Pr(θ). In this case it is also possible to proceed by enumeration, as above. The sampled values are amalgamated into a representative set, and then a CEP curve is computed for each θ in the representative set. Thus lower and upper bounds on the EP curve are available. Alternatively, it is possible to combine the CEP curves into a single EP curve, by vertical averaging, using probabilities approximated by the relative frequencies in the sample.

A more standard and less laborious calculation would simply be to attach the sampling of θ to the front end of the sampling procedure for the loss X. First, sample θ, then sample the hazard outcome ω conditional on θ, then sample the loss X conditional on θ and ω. Repeating this

process many times generates a distribution for losses that incorporates epistemic uncertainty about θ, and aleatory uncertainty about ω. Hidden inside this calculation is a collection of CEP curves, but they are never explicitly constructed. Sometimes this is the natural way to proceed, if there is a ready mechanism for generating probabilistic samples for θ. In the next section, for example, sampling θ and then ω given θ is a very natural extension to just sampling ω given θ.

Sampling the epistemic uncertainty as well as the aleatory uncertainty is the incremental update to a procedure that was already sampling aleatory uncertainty. From this point of view, it is a cheap upgrade, and unlikely to induce complications such as coding errors. Uncovering the underlying CEP curves, though, provides additional information which can be used to assess robustness and sensitivity.

2.3. Epistemic uncertainty and risk

The previous discussion has been in terms of the relationship between the CEP curves and the EP curve. This has obvious implications for the relationship between the conditional risk and the (unconditional) risk. The risk of the EP curve could be lower or higher than the conditional risk of a specified CEP curve. So generalising a CEP curve to account for epistemic uncertainty does not necessarily increase the risk. The Risk Manager could discover, for example, that other possible choices for θ implied much lower losses.

As well as a risk, each EP or CEP curve also has a variance, which is a summary of the uncertainty of the loss. The variance of a specified CEP curve is a conditional variance. It is a standard probabilistic result that the overall variance is equal to the variance of the conditional risk, plus the mean of the conditional variance (see, e.g., Grimmett and Stirzaker, 2001, chapter 3, section 7). A single CEP curve only contributes to the second term, and therefore generalising a CEP curve to account for epistemic uncertainty will tend to increase uncertainty about the loss, exactly as might be expected. As has already been stressed, though, this is not ‘new’ uncertainty. This is uncertainty that has been ‘internalised’, i.e. moved from a margin for error into the EP curve for a more transparent analysis. Consequently the margin for error ought to be reduced, to compensate. To pose the very difficult question “How much should the margin for error be reduced?” is simply to emphasise the importance of putting as much epistemic uncertainty as possible into the EP curve.

3. Uncertainty about the hazard process

The hazard process is characterised by a set of outcomes ω in Ω, and a probability for each outcome. Typically, these probabilities are derived from a statistical model fitted to a catalogue of previous events. Given the scarcity of large hazard events, and their lack of homogeneity, such statistical models will be somewhat crude, and the estimated parameters ill-constrained. The need to treat parametric uncertainty carefully, as possibly constituting a major source of epistemic uncertainty about the EP curve, suggests operating within a Bayesian statistical framework for estimating the parameters, and simulating the hazard process.

Consider earthquakes, by way of illustration. First, suppose that it is possible to specify a stochastic process for earthquakes in a specified region. The nature of this process is not important; what is important is that the stochastic process is represented as a member of a well-specified ensemble of models indexed by a parameter θ. This θ is uncertain, but the catalogue of earthquakes in the region is informative about it. A common approach is to derive a point-estimate for θ by finding that value which best explained the observations; maximum likelihood methods for parameter estimation are discussed in Davison (2003), chapters 4 and 7, and stochastic

processes, including for earthquakes, in chapter 6. Then it is common to use the point estimate for θ as though it were the true θ. In other words, to sample the hazard process using its estimated parameter values. This does not allow for the epistemic uncertainty associated with our inability to fully constrain θ using the catalogue. A two-pronged approach can be taken to rectify this. The first is to perform the sampling within a Bayesian statistical approach, explicitly allowing for uncertainty about θ, and the second is to extend the process of learning about θ to other regions, also following a Bayesian approach.

In the Bayesian approach θ is treated as epistemically uncertain, and given its own initial distribution, the 'prior' distribution, which incorporates general judgements about θ formed 'prior' to consulting the catalogue. Learning about θ takes the form of conditioning θ on the catalogue. Sampling a realisation for future earthquakes is a two-stage process

1. Sample one realisation of θ from the distribution of θ conditional on the catalogue;

2. Sample one realisation of the earthquake process for this realisation of θ.

By allowing θ to remain uncertain, Bayesian simulation of the earthquake process incorporates both aleatory and epistemic uncertainty. If the EP curve was already being computed by simulation then this is a 'free upgrade'. Note that if the catalogue is very detailed and we can treat the underlying process as stationary, then there will be very little epistemic uncertainty about θ after conditioning, in which case the Bayesian approach reduces to the simple simulation method. Bayesian estimation and simulation is discussed in Davison (2003), chapter 11.

One limitation of this approach is that it only uses the catalogue from one site or region to learn about the θ for that region. It is tempting to augment the catalogue with events from other regions, because increasing the size of the catalogue is likely to reduce epistemic uncertainty about θ. But, at the same time, it does not seem defensible to insist that θ is the same across all regions, especially where the observational data might be of varying quality and subject to epistemic errors. Bayesian hierarchical modelling provides a statistical solution. Each region is allowed to have its own value of θ, say θ_r for region r. The θ_r are assumed to be similar, in the sense that they are treated as being drawn independently from the same underlying distribution. By allowing this underlying distribution itself to be uncertain, information from the different regions can be pooled. This allows us to learn about the θ_r of a region for which we have observations, which will be strongly affected by the catalogue from that region, but also affected by the catalogues from the other regions. It also allows us to learn about the θ_r of a region for which we do not have observations.

This same framework extends to considering the role of historical catalogues in describing the future hazard process, when boundary conditions may have changed. More sophisticated approaches replace the simple exchangeability across regions described here with a more detailed treatment that can include regional covariates, and explicit representations of space and time. Again, as in the special case of just a single region, the output from this type of modelling is a simulation of the hazard process, which accounts for both aleatory and epistemic uncertainty. Therefore it can be plugged in to the front of a calculation that estimates the EP curve by sampling.

Bayesian hierarchical models are discussed in many textbooks. See, e.g., Davison (2003) section 11.4 for an introduction, or Gelman *et al.* (2004) for a more detailed treatment, including computation and diagnostics. Banerjee *et al.* (2004) covers hierarchical modelling of spatial data. Sahu and Mardia (2005) review spatial-temporal modelling.

For simplicity, the above explanation was for an uncertain parameter within a specified statistical model. In a more general treatment, both the model and the parameter (which is then expressed conditionally on the model) are allowed to be uncertain. This is an important generalisation, given the typical absence of really large events in a typical catalogue, and the convexity of the loss operator (itself very uncertain for large hazard events). So, for example, switching from a Frechet to a Gumbell distribution for extreme values in flood assessment may well make a large difference to the risk, even though both curves fit equally well over the catalogue. This case could be handled by embedding both alternatives within the Generalised Extreme Value (GEV) family of distributions, but only if parametric uncertainty is explicitly included.

Other situations are more complicated. Draper (1995) provides an example in which two different statistical models fitted on the same data (the Challenger O-ring data) give non-overlapping predictions. More recently, the ongoing debate about climate reconstruction for the last millennium has been largely about the effect of using different statistical models on the same database of historical temperature proxies; see, e.g., McShane and Wyner (2011), and the discussion that follows. Issues of model choice and model criticism are a core part of statistical theory and practice.

Uncertainty about the loss operator. Exactly the same comments also apply to the loss operator, if this is represented as a stochastic relationship between the hazard outcome and the loss that follows: (i) simulate within a Bayesian framework to allow for epistemic uncertainty in the relationship between hazard and outcome; (ii) use hierarchical models to extend the calibration inference to other datasets which, while not identical, are sufficiently similar to be informative; (iii) allow for alternative statistical models.

4. Uncertainty about the footprint function

The footprint function maps a hazard outcome into a trajectory in space and time, from which loss can be computed. The footprint function will often be influenced by the Risk Manager's choice of action. For example, the decision to increase the height of seawalls will restrict the footprint of large storm surges.

4.1. Three sources of epistemic uncertainty

In the simplest case, which is also the most common, the footprint function is treated as a deterministic function of the hazard outcome, usually expressed in terms of a particular hazard event. Thus, the hazard event might be a storm, described in terms of a time-series of precipitation over a catchment, and the footprint function might be a hydrological model describing how that storm affects river-flow in the catchment and consequent inundation of the flood plain, using a hydrodynamic model based on the shallow water equations. Or the hazard event might be a volcanic eruption, described in terms of location, orientation, and magnitude, and the footprint function might be a dynamical model describing the evolution of pyroclastic, lahar, or lava flows, e.g. using multi-phase versions of the shallow water equations. We refer to such functions as (deterministic) *simulators*. This is to avoid another instantiation of the word 'model', which is already heavily overloaded.

Simulators are limited in their ability to represent the impact of the hazard. In general, the 'gap' between the simulator output and the system behaviour can be thought of as the accumulation of three sources of epistemic uncertainty. First, simulators usually require the input of initial and boundary conditions in addition to the description of the hazard event, and these inputs may be imperfectly known, giving rise to **input uncertainty**. For example: a compartmental hydrological

simulator requires an initial specification of the saturation of each of the boxes; a fire simulator requires the dryness of the vegetation. In many cases, initial condition uncertainty can be somewhat finessed by running the simulator through a spin-up period prior to the hazard event, during which the effect of uncertainty about the condition at the start of the spin-up period is reduced through the assimilation of observational data. This is the basis of weather forecasting (see, e.g., Kalnay, 2003). Boundary condition uncertainties, such as the pattern of rainfall over a catchment in an extreme flood event, are more difficult to finesse and, for a given measurement system, might vary significantly in their characteristics from event to event. As in weather forecasting, data assimilation can sometimes help compensate for such epistemic uncertainties for real-time forecasting purposes (e.g. Romanowicz et al., 2008) but it can often be difficult to construct realisations of epistemic boundary condition uncertainties in simulation.

The second source of epistemic uncertainty is **parametric uncertainty**. Within the simulator there are typically parameters that do not have well-defined analogues in the system, or whose analogues are not measurable. This is an inevitable feature of simulators, which, through their abstraction, introduce ambiguity into the relationship between the simulator state vector and the system. So, for example, in a compartmental hydrological simulator the rates connecting the different compartments will be uncertain, and usually catchment-dependent. Parameters are often tuned or calibrated in an exercise designed to reduce the misfit between simulator output and system behaviour, for events in which the inputs were approximately known (e.g. previous storms on the same catchment). Here we distinguish between 'tuning', which describes a more-or-less *ad hoc* procedure, and 'calibration', which is explicitly statistical and aims to produce some assessment of parametric uncertainty.

The third source of epistemic uncertainty is **structural uncertainty**. This accounts, crudely, for all of the limitations of the simulator that cannot be eliminated by calibrating the parameters of the simulator. Formally, and this is another unrealistic thought experiment, we suppose that the inputs are known exactly, and that the 'best' values of the parameters have been identified in some way. The structural uncertainty is the uncertainty about the system that remains after running the simulator with the correct inputs and best parameter values. In the simplest situation, structural uncertainty is quantified without reference to the precise value of the inputs and parameters, as a lumped value. Often it is given in very crude terms. An avalanche simulator, for example, might predict steady state velocity as a function of slope and snow density; it might be considered to have a structural uncertainty of $\pm$1m/s, where this is stated without reference to any particular slope or density. Note that this kind of assessment can be improved with more resources but initially a crude estimate will often suffice, in terms of identifying where the major sources of uncertainty lie. Structural uncertainty is discussed further in section 4.3.

The loss operator again. In situations where the loss operator is a deterministic function of the hazard outcome, then exactly the same concerns apply.

4.2. Parametric uncertainty

It is a cannon of uncertainty assessment that one must start, if possible, with quantities that are operationally defined. The 'best' value of the parameters of a simulator, however, are not operationally defined and, as a consequence, nor is the discrepancy between the system and the simulator output at its best parameterisation. Of course science is beset by such difficulties, and we do the best we can. In the case of uncertainty about the simulator, *it is surely far better to include a crude assessment of parametric and structural uncertainties than simply to ignore them.* To ignore them is effectively to set these uncertainties to zero, and relegate them to the external

uncertainty and the margin for error. On this basis, we proceed to outline a simple assessment, advocating that such an assessment is much better than nothing.

The first step in addressing parametric uncertainty is to identify hard-to-quantify parameters, and to assess plausible ranges for each of them. Notionally, the parameter ranges represent the possible values that the 'best' parameterisation might take. These ranges could be assessed on the original scale, or on a log-scale for parameters that are strictly positive; in the latter case they might often be assessed proportionately, for example ±10% around the standard value. Experience in a wide range of environmental modelling applications suggests that the effective value of a parameter, in terms of generating acceptable simulator outputs, can be well away from the system equivalent. Sometimes this is for well-understood reasons, as with viscosity in ocean simulators, where it is common to distinguish molecular viscosity, which is a property of water, from eddy viscosity, which is determined partly by the solver resolution. Generally, however, a good strategy is to be prepared to revise the parameter ranges after a pilot study of simulator runs have been compared to system measurements. Proceeding sequentially is absolutely crucial, even though it might be cheaper and more convenient to submit the entire experiment as one batch.

Once ranges have been assigned (and perhaps revised), the shape of the parametric uncertainty distribution can be specified, to construct simple probability representations. This is a more subtle issue than might first appear. It is natural (although by no means necessary) to treat uncertainty about the parameters as probabilistically independent, in the absence of strong information to the contrary. However, the joint distribution of large numbers of probabilistically independent random variables has some counterintuitive properties. It is also natural to use a peaked distribution rather than a uniform, to reflect the common judgement that the best value of each parameter is more likely to be found in the centre of its range than on the edge. However, if there are more than a handful of parameters, these two judgements made together will cause the joint probability over all the parameters to be highly concentrated at the centre of the parameter space.

For example, if two parameters are each given independent symmetric triangular distributions on the unit interval, the inscribed sphere (i.e., in two dimensions, the circle with radius 0.5) occupies 79% of the volume, but contains 97% of the probability. For five parameters, this becomes 16% of the volume and 68% of the probability; for ten parameters, 0.2% of the volume and 14% of the probability. On the other hand, if independent uniform distributions are used, then for ten parameters 0.2% of the volume will contain 0.2% of the probability; hence with ten parameters switching from uniform to triangular causes a substantial increase in the probability near the centre of the parameter space, and corresponding substantial decrease in the probability in the corners and along the edges. The moral of this story is to think carefully about using marginal distributions that have sharp peaks if there are lots of parameters. It might be better, in this situation, to focus initially on just a few of the key uncertain parameters and treat the others as fixed; or to use a much less-peaked marginal distribution.

Once a probability distribution for the parameters has been quantified, parametric uncertainty in the simulator can be incorporated into the EP curve by sampling: sample the hazard outcome ω, sample the simulator parameters θ, run the simulator with arguments ω and θ, compute the loss; and then repeat. If a statistical model of input uncertainty is postulated, then input uncertainty can also be introduced by sampling. This simple Monte Carlo approach can be substantially improved by statistical techniques for variance reduction, which will be particularly important when the simulator is slow to evaluate. For many natural hazards simulators, however, for which the evaluation time tends to be measured in seconds rather than hours (compare to weeks for a large climate simulator) the balance currently favours simple Monte Carlo simulation, when taking into

account the possibility of coding and computation errors. This will shift, though, as simulators become more complex (e.g. 3D multi-phase flow simulations for volcanic pyroclastic flows, taking hours), which seems to be inevitable. Variance reduction methods to improve on simple Monte Carlo simulation are discussed in Davison (2003), chapter 3, section 3.

The choices for parametric uncertainty, and structural uncertainty discussed next, can both be assessed and possibly revised when there are system measurements that correspond, either directly or indirectly, to simulator outputs. This is discussed in section 4.4.

4.3. Structural uncertainty

It is usually not hard to construct a list of the main sources of structural uncertainty in a simulator; indeed, the chapters on individual hazards in this volume do exactly that. But it is very difficult to quantify these, either jointly or individually (as the chapters illustrate). Mostly, structural uncertainty is ignored at the sampling stage (to be relegated to the margin for error), or, in the case of simulator calibration, to be discussed in section 4.4, treated as small relative to measurement error.

Where structural uncertainty is incorporated, the method that is favoured in statistics is to use a lumped additive 'discrepancy' (possibly additive on a log-scale for strictly positive values). An example for a simulator with spatial outputs would be

$$\text{system}(s) = \text{simulator}(s; \theta^*) + \text{discrepancy}(s)$$

where s is spatial location, and θ^* is the 'best' value of the simulator parameters, itself uncertain (a different illustration is given below). The discrepancy is treated as probabilistically independent of θ^*. This framework is discussed in Rougier (2007), in the context of climate modelling. It is undoubtedly restrictive, and one generalisation is discussed in Goldstein and Rougier (2009), designed to preserve the tractability of the additive discrepancy in the context of several different simulators of the same underlying system. Here we will just consider the simple case, as not ignoring the discrepancy would already be a large improvement for most environmental science simulations.

The key point to appreciate when thinking about the discrepancy is that when a simulator is wrong, it is typically systematically wrong. This is because 'wrongness' corresponds to missing or misrepresented processes, and the absence of these processes does not cause random errors that vary independently across the space and time dimensions of the simulator output, but instead causes errors that vary systematically. So if the simulator value is too low at location s then it is likely to be too low at s', for values of s' in the neighbourhood of s. This indicates that the discrepancy does not behave like a regression residual or a measurement error, in which values at s and s' are uncorrelated. Instead, the correlation between the discrepancy at s and at s' depends on the locations of these two points in space and/or time.

The separation between s and s' at which the discrepancy is effectively uncorrelated is termed the 'decorrelation distance'. In general the decorrelation distance will depend on s and on the direction away from s. But simplifying choices can be made. The simplest of all is that the discrepancy is stationary and isotropic, i.e. its properties can be described by just two quantities: a standard deviation and a correlation length. For natural hazards, stationarity and isotropy are likely to conflict with judgements about simulator limitations, and with known features of the spatial domain. That is not to say, however, that these simple statistical models are not useful. They may still be better than not accounting for the discrepancy at all.

Little can be said about how to specify the discrepancy in general: this depends far too much of the particular application. However, an illustration might be helpful. Suppose that the purpose of the simulator was to compute the steady state velocity profile of an avalanche (Rougier and Kern, 2010), denoted v(z) where $z \geq 0$ is the slope-normal height. We might judge that the dominant source of structural uncertainty was an additive term that shifted the whole profile relative to the simulator output, with a secondary source of uncertainty that tilted the profile relative to the simulator output. Hence the discrepancy might be represented as

$$\text{discrepancy}(z) = \upsilon_1 + \upsilon_2 (z - z_m)$$

where both υ_1 and υ_2 are uncorrelated random quantities with zero means, and the sizes of the two standard deviations σ_1 and σ_2 control the degree of uncertainty concerning shifting and tilting; z_m is the a central value in the range of z, around which the tilt pivots. So setting $\sigma_1=0$ insists that there is no shifting in the discrepancy, and setting $\sigma_2=0$ insists that there is no tilting. The value of σ_2 will have units of 1/sec; in fact, this represents uncertainty about what avalanche scientists refer to as the 'shearing rate'. The best way to choose σ_1 and σ_2 is to try some different values and sample υ_1 and υ_2 using, say, two Normal distributions. This discrepancy is illustrated in Figure 2.

[FIGURE 2 ABOUT HERE]

The point of this illustration is to show that specifying the discrepancy does not have to be any more complicated than judgements about simulator limitations allow. One could go further, for example by including higher-order terms in $(z - z_m)$, but only if judgements stretched that far. It is easy to become overwhelmed with the complexity of specifying a stochastic process for the discrepancy, and not see the wood for the trees. The discrepancy is there to account for those judgements about simulator limitations that are amenable to simple representations. To return to the theme of this chapter, the discrepancy offers an opportunity to internalise some aspects of epistemic uncertainty, by moving them out of the margin for error and into the EP curve. Just because an expert does not feel able to quantify all of his judgements about a simulator's structural uncertainty does not mean that those judgements that *can* be quantified need to be discarded. Any discrepancy representation that is more realistic than treating the simulator as though it were perfect is an improvement.

Once a representation for the discrepancy has been specified, then it can be incorporated into the EP curve by simulation, effectively by adding a random realisation of the discrepancy to the outcome of running the simulator. Although 'noising up' a deterministic simulator in this way might seem counter-intuitive, it must be remembered that the purpose of the discrepancy is to avoid treating the simulator as though it were perfect when it is not. Adding some noise to the simulator output is a simple way to achieve this.

4.4. Evaluating the simulator when measurements are available

In many cases there will be system measurements available, and these can be used to check the various judgements that have been made, and, if everything looks OK, to refine judgements about the simulator parameters and the structural uncertainty. However, it is very important in this process that systematic errors in the measurement process do not get fed into the simulator evaluation and calibration.

For example, Beven et al., (2011) have shown how basic rainfall-runoff observations might be inconsistent in certain events in the available records. It is quite possible, for example, that in a

flood the volume of discharge estimated from observed water levels might be significantly greater than the observed volume of rainfall estimated from raingauge or radar-rainfall data. This can be because of errors in the rating curve used to convert water levels to discharges (particularly in extreme events that require extrapolation from measurements at lower flows) or the conversion of radar reflectivity to patterns of rainfall intensities or the interpolation from a small number of raingauges to the rainfall volume over a catchment. The nature of these errors might vary from event to event in systematic ways (see for example, Westerberg et al., 2011). They might also have carry-over effects on subsequent events depending on the relaxation time of the catchment system. If a rainfall-runoff model is used as a simulator in estimating the footprint of such an event, it is clear that it will necessarily provide a biased estimate of the discharge observations for such events (at least if that model is consistent in maintaining a water balance). This is one of the reasons why there has been an intense debate about model calibration techniques in the hydrological modelling literature in recent years, between those who wish to use rather simple statistical approaches, treating the residuals *as if* they had no systematic component, and those who wish to incorporate non-stationarity effects that can arise from a common source of uncertainty such as the rating curve (e.g. Mantovan and Todini, 2006; Beven et al., 2008). Similar issues will arise in many areas of natural hazard simulation.

Model criticism

If the structural uncertainty is represented as an additive discrepancy, as outlined above, then the focus is on the residuals at different values for the parameters

$$\text{residual}(s_i; \theta) = \text{measurement}(s_i) - \text{simulator}(s_i; \theta)$$

where s_i ranges over the locations where there are measurements, and we are ignoring the time-dimension, for simplicity (although in fact time often presents more challenges than space). More complicated situations are also possible, in which there are multiple experiments with different inputs, or different control variables in the simulator. In the snow rheology analysis of Rougier and Kern (2010), for example, there are ten experiments, with varying environmental conditions summarised primarily by snow density.

Suppose that each measurement has an additive error e_i, of the form

$$\text{measurement}(s_i) = \text{system}(s_i) + e_i$$

where it is usual to treat the measurement errors as uncorrelated in space. The residual at the best parameter θ^* comprises

$$\text{residuals}(s_i; \theta^*) = \text{discrepancy}(s_i) + e_i .$$

This relation can be used to examine the various modelling choices that have been made, about the simulator, the simulator parameters, and the discrepancy. This is *model criticism*, where it is important to appreciate that 'model' here represents all judgements, not just the physical ones that go into the simulator, but also the statistical ones that link the simulator and the system.

The main thing that complicates model criticism is that the discrepancy is unlikely to be uncorrelated across the measurements, or, to put it another way, the decorrelation length of the discrepancy is likely to be longer than the spatial separation between measurements.

There are various solutions to this problem, but the easiest one is to thin or clump the measurements. Thinning means deleting measurements, and ideally they would be deleted selectively, starting with the least reliable ones, to the point where the spacing between

measurements was at least as wide as the decorrelation length of the discrepancy. *Clumping* means amalgamating individual measurements into a smaller number of pseudo-measurements, and ideally this would be done over regions that are at least as wide as the decorrelation length of the discrepancy; the simulator outputs will need to be amalgamated to match. There is no requirement that the clumping regions be the same size: in time they might reflect the dominant relaxation time of the system (which might include multiple events); in spatial domains they might also reflect discontinuities due to features such as a change in surficial or bedrock geology.

Both thinning and clumping discard information, but sometimes one does not need much information to identify bad modelling choices. The advantage of clumping is that it does not rely too strongly on distributional choices, which would be necessary for a more sophisticated analysis. This is because the arithmetic mean of the measurement errors within a clump will tend to a normal distribution (see, for example, the version of the Central Limit Theorem given in Grimmett and Stirzaker, 2001, chapter 5, section 10, page 194). This supports the use of summed squared residuals as a scoring rule for choosing good values for θ^*, providing that the measurement errors can be treated as probabilistically independent.

For the rest of this section we suppose that the measurement errors can be treated as probabilistically independent (our caution at the start of this subsection notwithstanding) and that the measurements have been clumped to the point where the clumped discrepancies can be treated as probabilistically independent. The variance of the clumped residual for each region is the sum of the variance of the clumped discrepancy and the variance of the mean of the measurement errors. We denote this residual variance as σ_i^2 for pseudo-measurement i, and the standard deviation as σ_i. This is a value that we must specify from our judgements about the discrepancy and our knowledge of the measurement. In a nutshell, it quantifies how much deviation we expect to see between the simulator output at its best parameterisation, and the values of the measurements we have made (but see also the approach based on limits of acceptability in Section 4.5). Clearly, this is something we ought to have a judgement about, if our intention is to use the simulator to make statements about the system. What we have done here is try to find an accessible way to represent and quantify these judgements.

Now consider running the simulator over a range of possible values for θ, termed an *ensemble* of simulator runs. Each member of this ensemble can be scored in terms of the sum of its squared scaled residuals, where the scaled residual at location s_i equals residual(s_i; θ) / σ_i . The larger the sum of the squared scaled residuals, the worse the fit. One member of the ensemble, say θ^+, will have the least-bad fit. Now if θ^+ was actually θ^*, we would expect most of the scaled residuals for θ^+ to lie between ± 3 according to the 'three sigma' rule; see Pukelsheim (1994). But of course θ^+ will not be θ^*. But the simulator output at θ^+ will be close to that at θ^* if (i) the ensemble is quite large, (ii) the number of parameters is quite small, and (iii) the simulator output is quite flat around θ^*. In this case, the scaled residuals for θ^+ should lie between ± 3, or a little bit wider. Much larger residuals are diagnostic evidence of suspect modelling choices, for example ignoring the possibility of systematic errors in the measurements. Likewise, if all the residuals are very small, then this can be interpreted as possibly an overlarge discrepancy variance.

The key feature that makes this diagnostic process work for a large variety of different systems, different natural hazards in our case, is the clumping. To explore this a bit further, one interpretation of clumping is that it is focusing on those aspects of the simulator that are judged reliable. Simulators based on solving differential equations are often not very reliable at spatial or temporal resolutions that are nearly as high as the solver resolution, but become more reliable at lower resolutions. So the expert may have good reasons for thinking that the simulator output for location s is a bad representation of the system at location s, but may still think that the simulator

output averaged over a region centred on s is a good representation of the system averaged over the same region. There is no need to restrict this process to averages. The expert can decide which aspects of the simulator output are reliable, and focus on those. These may be averages, or they may be other features: the timing of a particular event, such as the arrival of a tsunami wave, or the scale of an event, such as peak ground acceleration in an earthquake. There is no obligation to treat all of the simulator outputs as equally reliable, and this can be incorporated into the diagnostic process. It should also be reflected in the way that the simulator is used to predict system behaviour, if possible; i.e. to drive the loss operator from the reliable aspects of the simulator output.

Calibration

Let us assume that the model criticism has been satisfactory. Calibration is using the (pseudo-) measurements to learn about θ* and possibly about the structure of the discrepancy as well. Again, it is not necessary to restrict attention to pseudo-measurements for which the residuals can be treated as uncorrelated, but it is simple and less dependent on parametric choices.

We distinguished in section 4.1 between tuning and calibration. Tuning is aimed at identifying a good choice for θ* in an *ad hoc* way, prior to plugging in this choice without any further consideration of parametric uncertainty. Really, though, the notion of plugging in values for the parameters is untenable when modelling complex systems, such as those found in natural hazards, as has been stressed in hydrology (see, e.g., Beven 2006, 2009; Beven et al., 2011; Beven and Westerberg, 2011). With complex systems, model limitations (and, very often, input data limitations) severely compromise our ability to learn about θ*. Therefore assessing (i.e. not ignoring) parametric uncertainty is crucial, and this favours a statistical approach in which judgements are made explicit, and standard and well-understood methods are applied.

Calibration is an example of an *inverse problem*, for which there are massive literatures on both non-statistical and statistical approaches (see, e.g., Tarantola, 2005). The simplest statistical approach is just to rule out bad choices of θ in the ensemble of runs according to the residuals, and then to treat the choices that remain as equally good candidates for θ*. This ruling out can be according to the residuals, with different types of rule being used in different situations. Where there appear to be lots of good candidates then ruling out all choices for which one or more of the scaled residuals exceeds, say, 3.5 in absolute size might be a reasonable approach. In other cases, a more cautious approach would require more large residuals before a candidate was ruled out. There are no hard-and-fast rules when epistemic uncertainty plays a large part in the assessment: the distinction between ruled-out and not-ruled-out has to smell right according to the judgements of the expert. Looking at scaled residuals is a good starting point, but ought to be backed-up with a qualitative assessment, such as “parameters in this region get the spatial gradient wrong”. This combination of quantitative and qualitative assessment should be expected because we do not expect to be able to quantify our epistemic uncertainty with any degree of precision.

Craig et al. (1997) developed the idea of ‘history matching’ by ruling out bad choices for θ*, and considered the different ways of scoring the vector of residuals. Vernon et al. (2010) provide a very detailed case study in which bad choices for θ* in a large simulator are eliminated through several phases of an experiment, where each phase ‘zooms in’ to perform additional simulator evaluations in the not-ruled-out region of the parameter space. Both of these papers are adapted to large applications, using an *emulator* in place of the simulator, to account for the very long run-times of their simulators (in Vernon et al, the simulator encompasses the entire universe). Where the run-time of the simulator is measured in seconds then an emulator is not required.

This section has described off-line calibration, but there are advantages to combining calibration and prediction, in which calibration measurements from previous events are used directly in the prediction of future events, which allows us to take account of systematic effects in the discrepancy that may span both the calibration and the prediction. The key references here are Kennedy and O'Hagan (2001) for the fully Bayesian approach, and Craig et al. (2001) for the Bayes linear approach, which is more suitable for large simulators. Goldstein and Rougier (2006) extends the Bayes linear approach to explicit calibration. Rougier (2007) outlines the general fully probabilistic framework for calibrated prediction, in the context of climate.

4.5. Less probabilistic methods

The diagnostic and calibration approaches outlined above have supposed that (i) structural uncertainty and measurement error can be treated additively, and (ii) clumping can be used to reduce the effect of correlations within structural uncertainty. The focus on additive residuals after clumping is a natural consequence of these two suppositions. But in some applications the epistemic uncertainties associated with inputs, simulator limitations, and system measurements can be impossible to separate out, and can result in residuals that are effectively impossible to decorrelate (Beven, 2002, 2005, 2006; Beven *et al.*, 2010). For example, latent (i.e. untreated) error in the simulator inputs creates a very complicated pattern in the discrepancies, in which the decorrelation length will be related to the dynamics of the system, and where the decorrelation length may be longer than the event duration. This makes it very hard to define formal statistical models that link the simulator output, the system behaviour, and the measurements, and has led to alternative approaches to uncertainty estimation.

Beven and co-workers, for example, advocate a less statistical approach, informed by the particular challenges of hydrological modelling (see, e.g., Beven, 2006; 2009, chapters 2,4). Hydrological simulators for flood risk assessment are forced by noisy and imperfectly observed inputs, namely the precipitation falling on the catchment. The epistemic uncertainty in such inputs reflects the differences between the actual precipitation and that observed by raingauges and radar rainfall, which will vary from event to event in complex ways. Similarly, discharge measurements used to calibrate simulators can be subject to epistemic uncertainties as rating curves vary over time (e.g. Westerberg et al., 2010), or need to be extrapolated to flood peaks involving overbank flows. Furthermore, both measured and predicted discharges can change rapidly, e.g. through several phases of a storm, which means that 'vertical' residuals (observations less simulator output at each time-point) can become very large simply through small errors of phase in the simulator.

As an alternative, the Generalised Likelihood Uncertainty Estimation (GLUE) approach (Beven and Binley, 1992) replaces the standard misfit penalty of the sum of squared scaled residuals with a more heuristic penalty that can be carefully tuned to those aspects of the simulator output that are thought to be reliable indicators of system behaviour. This does not rule out the use of penalties that are derived from formal statistical models; but it does not oblige the expert to start with such a statistical model. In recent applications, parameter evaluation has been based on specifying prior limits of acceptability ranges for simulator outputs and, within these ranges, defining weighting schemes across the runs in the ensemble that will tend to favour good fits over bad ones (e.g. Blazkova and Beven, 2009; Liu et al. 2009).

Although the GLUE approach has been criticised within hydrology for being insufficiently statistical, in fact it preceded by over a decade a recent development in statistics along exactly these lines, generally termed Approximate Bayesian Computation, or ABC (see, e.g., Beaumont *et al.*, 2002, Toni *et al.*, 2010). Like the GLUE approach, ABC uses summary statistics to create *ad hoc* scoring rules that function as likelihood scores. However, it should be noted that it is an active area of

research in statistics to understand how the replacement of formal with informal likelihoods affects the inference, notably asymptotic properties such as the consistency of the estimator of the best values of the simulator parameters. Similar approximate methods are used in the Generalised Method of Moments approach, also termed the 'indirect method' (see, e.g., Jiang and Turnbull, 2004).

5. Uncertainty about external features

'External features' encompass all those things that do not occur inside the calculation of the EP curve, but which nonetheless affect the environment within which the EP curve is computed. These tend to be large and unwieldy. For example: responsiveness of the affected population to an early warning system; effectiveness of encouraged or forcible relocation; changes in population demographics; rate of uptake of new building regulations; future greenhouse gas emissions; changes in Government policy. These are sometimes referred to as 'Knightian uncertainties', following Knight's (1921) distinction between what he termed 'risk', which could be quantified (effectively as chance), and 'uncertainty', which could not. It is surely a common experience for all of us that some uncertainties leave us totally nonplussed.

The characteristic of an uncertain external feature is that we can enumerate its main possible outcomes, and possibly compute a CEP curve under each possible outcome, but we cannot attach a probability to each outcome. Greenhouse gas emissions scenarios for the 21st century provide a clear example. Climate simulators are run under different emissions scenarios, and these can be used to generate realisations of weather, for example as done in the UK Climate Impacts Programme[1] (UKCIP, see also Chapter ??? of this volume). These realisations of weather (the hazard process) can be used to drive a hydrological model of a catchment (the footprint function), which in turn can be used to evaluate losses, and to compute a CEP curve. So the emission scenario lives at the very start of the process, and different emissions scenarios give rise to different CEP curves. But, intriguingly, the attachment of probabilities to the emissions scenarios was explicitly discouraged (see Schneider, 2002, for a discussion).

Section 2 demonstrates that progress can be made if external features can be enumerated; in particular, EP curves and risks can be bounded. Continuing in this vein, decision support can often continue, despite not being able to specify a probability for each outcome. Each possible action generates a set of CEP curves, one for each possible combination of external factors. So each pair of (action, external feature) has its own risk, measured as the area under its CEP curve. These risks can be collected together into a matrix where the rows represent actions and the columns represent external features. This is the classic tableau of decision analysis (see, e.g., Smith, 2010, section 1.2).

First, it is possible that some actions can immediately be ruled out, as being *inadmissible*. An inadmissible action is one in which the risk of some other action is no higher for every possible combination of external features, and lower in at least one combination.

Second, optimal decisions often have the property of being relatively robust to changes in probabilities (Smith, 2010, section 1.2), and so a sensitivity analysis over a range of different probability specifications for the external features may indicate an outcome that is consistently selected over a sufficiently wide range of different probability specifications.

Finally, there are non-probabilistic methods for selecting an optimal action. One such method is *minimax loss*. Minimax loss suggests acting on the assumption that nature is out to get you. The

[1] http://www.ukcip.org.uk/

minimax action is the action which minimises the maximum risk for a given action. In other words, each action is scored by its largest risk across the external features, and the action with the smallest score is chosen. Another non-probabilistic method is *minimax regret* (see, e.g., Halpern, 2003, section 5.4.2). The point about these types of rules, however, is that they generate actions that can be hard to justify in the presence of even small amounts of probabilistic information. As Halpern (2003, pages 167-168) notes, "The disadvantage of [minimax loss] is that it prefers an act that is guaranteed to produce a mediocre outcome to one that is virtually certain to produce an excellent outcome but has a very small chance of producing a bad outcome."

Overall, when accounting for external features in choosing between actions, it might be better to assign probabilities, even very roughly, and then to examine the sensitivity of the chosen action to perturbations in the probabilities. The decision to adopt a non-probabilistic principle such as minimax loss is itself a highly subjective and contentious one, and such an approach is less adapted to performing a sensitivity analysis. A more formal alternative to a sensitivity analysis with probabilities is to use a less prescriptive uncertainty calculus. An example might be the Dempster-Shafer approach, in which belief functions are used to provide a more general uncertainty assessment, which preserves notions of ignorance that are hard to incorporate into a fully probabilistic assessment. Note, however, that the properties of such approaches are less transparent than those of probability, and any type of calculation can rapidly become extremely technical. It is emphatically *not* the case that a more general uncertainty calculus leads to a more understandable or easier calculation. Readers should consult, e.g., Halpern (2003), chapters 2 and 3, for compelling evidence of this.

6. Summary

Epistemic uncertainty is the uncertainty that follows from constraints on our understanding, our information, or our resources (e.g. computing, or time for reflection). It is contrasted with aleatory uncertainty, which is the inherent uncertainty of the hazard itself. This chapter has characterised epistemic uncertainty in terms of a multiplicity of EP curves, where each EP curve can be thought of as conditioned on the value of one or more uncertain quantities, and is denoted here as a CEP curve ('C' for 'conditional'). While inspecting such a collection of curves is extremely revealing, it is necessary for communication and for decision making to reduce them, where possible, to a single summary EP curve.

Three main sources of epistemic uncertainty have been examined, and it is efficient and sometimes necessary to treat them in three different ways.

First, on the understanding that the EP curve will be computed by sampling the hazard outcome from the hazard process, epistemic uncertainty about the hazard process itself can be handled by performing that sampling within a Bayesian framework, which allows the parameters of the hazard process to be uncertain. These parameters can be conditioned on a catalogue of historical and laboratory measurements. The range of suitable measurements can be extended in some situations by using a Bayesian hierarchical model in which different classes of measurement (e.g. measurements from different locations) are treated as exchangeable (see section 3). This extension will tend to reduce uncertainty about the hazard process. The Bayesian approach can also be extended to incorporate alternative hazard processes, through 'model averaging', or by enumeration and bounding. The same comments apply to stochastic loss operators.

Second, epistemic uncertainty in the footprint function (referred to above as the 'simulator') can be removed by additional sampling, over different realisations of the simulator parameters, over different realisations of the simulator's structural uncertainty, and also input uncertainty if this is

thought to make a large contribution. Structural uncertainty will tend to be systematic rather than uncorrelated, which involves the specification of decorrelation lengths (section 4.3). Where there are system measurements available, these can be used to check the model specification ('model' here incorporating both the simulator and the statistical model), and to learn about the parameters. A simple and largely non-parametric approach is to 'clump' the measurements together within regions (typically spatial or temporal) about as wide as the decorrelation length of the structural uncertainty. This justifies the use of the simulator residuals as a diagnostic of fit and as a penalty function for ruling out poor choices of the simulator parameters. Other approaches, such as GLUE, use a more heuristic penalty to provide a more flexible description of those aspects of the footprint function that are considered reliable indicators of system behaviour. The same comments apply to deterministic loss operators.

Third, epistemic uncertainty also applies to external features of the problem: these tend to be enumerable but unwieldy, so that attaching probabilities is too subjective to be generally defensible. In this case the set of EP curves, one for each possible combination of external features, cannot be collapsed into one summary EP curve.

Clearly, these three sources of epistemic uncertainty can be combined, and actions can also be introduced. The result will be one EP curve for each (action, external feature) pair, with the first and second sources of epistemic uncertainty having been averaged out. In the presence of external features, it may still be possible to select actions with small risks using methods from Statistical Decision Theory.

Finally, we stress that accounting for epistemic uncertainty is challenging, particularly in defining appropriate scores for model criticism and calibration. But the Risk Manager must act, taking the best decision with her available resources, and justifying that decision to the Auditor. We reiterate the point made at the start of this chapter. Failure to incorporate epistemic uncertainty explicitly into decision-support tools such as the EP curve leads to it being represented implicitly, as a lumped margin-for-error, or, especially in policy, as a cautionary attitude which tends to favour inertia.

We hope we have shown that some aspects of epistemic uncertainty can be represented explicitly, and that this confers substantial advantages to the Risk Manager. Within the framework that we have outlined, it is possible to make reasonable judgements about many of the quantities required, such as ranges for parameters or the accuracy of the simulator. These judgements can be further tuned through sensitivity analysis and often, if there is calibration data available, the process can be put onto a more formal footing. This outline will seem somewhat subjective, involving judgements that can be hard to trace back to anything more than accumulated experience. But it should be recognised that, for example, to ignore a simulator's limitations and treat it as perfect is also a judgement: the value of zero is just as subjective as any other value, but much harder to defend. After all, zero is definitely wrong, whereas, say, ± 1 m/s is at least worth discussing.

Acknowledgements

We would like to thank Thea Hincks for her detailed and perceptive comments on an earlier draft of this chapter.

References

Banerjee, S., B.P. Carlin and A.E. Gelfand, 2004, Hierarchical Modeling and Analysis for Spatial Data, Chapman & Hall/CRC.

Beaumont, M.A., W. Zhang and D.J. Balding, 2002, Approximate Bayesian Computation in population genetics, *Genetics,* **162**, 2025–2035.

Beven, K.J., 2002, Towards a coherent philosophy for environmental modelling*, Proc. Roy. Soc. Lond. A,* 458, 2465-2484*.*

Beven, K.J., 2005, On the concept of model structural error, *Water Science and Technology*, 52(6), 165-175.

Beven, K.J., 2006, A manifesto for the equifinality thesis, *Hydrological Processes*, **16**, 189-206.

Beven, K.J., 2009, *Environmental Modelling: An Uncertain Future?*, Routledge.

Beven, K.J., D.T. Leedal, and R. Alcock, 2010, Uncertainty and good practice in hydrological prediction, *Vatten*, 66:159–163.

Beven, K.J., P.J. Smith, and A. Wood, 2011, On the colour and spin of epistemic error (and what we might do about it), Hydrol. Earth Syst. Sci., 15, 3123-3133, doi: 10.5194/hess-15-3123-2011.

Beven, K.J., and I. Westerberg, 2011, On red herrings and real herrings: disinformation and information in hydrological inference, *Hydrological Processes*, **25**, 1676–1680, DOI: 10.1002/hyp.7963.

Blazkova, S., and K.J. Beven, 2009, A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic, *Water Resour. Res*., 45, W00B16, doi:10.1029/2007WR006726.

Craig, P. S., M. Goldstein, A.H. Seheult, and J.A. Smith, 1997, Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments. In *Case Studies in Bayesian Statistics*, volume 3, 36-93. New York: Springer-Verlag.

Craig, P.S., M. Goldstein, J.C. Rougier, and A.H. Seheult, 2001, Bayesian Forecasting for Complex Systems Using Computer Simulators, *Journal of the American Statistical Association*, **96**, 717-729.

Davison, A.C., 2003, *Statistical Models*, Cambridge University Press.

Draper, D., 1995, Assessment and propagation of model uncertainty, *Journal of the Royal Statistical Society, Series B*, **57**, 45-97, with discussion.

Gelman, A., J.B. Carlin, H.S. Stern and D.B. Rubin, 2004, *Bayesian Data Analysis*, 2nd ed., Chapman & Hall/CRC.

Goldstein, M., and J.C. Rougier, 2006, Bayes Linear Calibrated Prediction for Complex Systems, *Journal of the American Statistical Association*, **101**, 1132-1143.

Goldstein, M., and J.C. Rougier, 2009, Reified Bayesian modelling and inference for physical systems, *Journal of Statistical Planning and Inference*, **139**, 1221-1239, with discussion.

Goldstein, M., 2011, External Bayesian analysis for computer simulators, in *Bayesian Statistics 9*, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (Eds.), Oxford University Press, in preparation.

Grimmett, G.R., and D.R. Stirzaker, 2001, Probability and Random Processes, 3rd edn, Oxford University Press.

Halpern, J.Y., 2003, *Reasoning about uncertainty*, MIT Press.

Jiang, W. and B. Turnbull, 2004, The indirect method: Inference based on intermediate statistics—A synthesis and examples, *Statistical Science*, **19**, 239-263.

Kalnay, E., 2003, *Atmospheric modeling, data assimilation and predictability*, Cambridge University Press.

Kennedy, M.C. and A. O'Hagan, 2001, Bayesian calibration of computer models, *Journal of the Royal Statistical Society, Series B*, **63**, 425-464, with discussion.

Knight, F.H., 1921, *Risk, Uncertainty, and Profit*. Boston, MA: Hart, Schaffner & Marx; Houghton Mifflin Company.

Liu, Y, J.E. Freer, K.J. Beven, and P. Matgen, 2009, Towards a limits of acceptability approach to the calibration of hydrological models: extending observation error, *J. Hydrol.*, 367:93-103, doi:10.1016/j.jhydrol.2009.01.016.

McShane, B.B., and A.J. Wyner, 2011, A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable?, *Annals of Applied Statistics*, forthcoming, with discussion.

Pukelsheim, F., 1994, The three sigma rule, *The American Statistician*, **48**, 88-91.

Romanowicz, RJ, Young, PC, Beven, KJ and Pappenberger, F, 2008, A Data Based Mechanistic Approach to Nonlinear Flood Routing and Adaptive Flood Level Forecasting, *Advances in Water Resources*, 31:1048–1056

Rougier, J.C., 2007, Probabilistic Inference for Future Climate Using an Ensemble of Climate Model Evaluations, *Climatic Change*, **81**, 247-264.

Sahu, S., and K.V. Mardia, 2005, Recent Trends in Modeling Spatio-Temporal Data, in Proceedings of the special meeting on Statistics and Environment, Società Italiana di Statistica, 69-83.

Schneider, S.H., 2002, Can we estimate the likelihood of climatic changes at 2100?, *Climatic Change*, **52**, 441–451.

Smith, J.Q., 2010, *Bayesian Decision Analysis; Principles and Practice*, Cambridge University Press.

Spiegelhalter, D.J. and H. Riesch, 2011, Don't know, can't know: Embracing deeper uncertainties when analysing risks, *Philosophical Transactions of the Royal Society, Series A*, **369**, 1-21.

Tarantola, A., 2005, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M.P.H. Stumpf, 2009, Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems, *Journal of the Royal Society Interface*, **6**, 187–202.

Vernon, I., M. Goldstein, and R.G. Bower, 2010, Galaxy formation: A Bayesian uncertainty Analysis, Bayesian Analysis, 5(4), 619-670, with discussion.

Westerberg, I., J.-L. Guerrero, J. Seibert, K.J. Beven, and S. Halldin, 2011, Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras, Hydrol Process, 25, 603-613, DOI: 10.1002/hyp.7848.