Environmental Modelling & Software 36 (2012) 35–48

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux

Anthony O'Hagan

Department of Probability and Statistics, University of Sheffield, Sheffield S3 7RH, UK

ARTICLE INFO

Article history:
Received 2 July 2010
Received in revised form
16 January 2011
Accepted 3 March 2011
Available online 31 March 2011

Keywords:
Subjective probability
Elicitation
Elaboration
Expert judgement
Mechanistic model
Environmental model
Sensitivity analysis
Emulation
Carbon flux

ABSTRACT

It is widely recognised that the appropriate representation for expert judgements of uncertainty is as a probability distribution for the unknown quantity of interest. However, formal elicitation of probability distributions is a non-trivial task. We provide an overview of this field, including an outline of the process of eliciting knowledge from experts in probabilistic form. We explore approaches to probabilistic uncertainty specification including direct elicitation and Bayesian analysis. In particular, we introduce the generic technique of elaboration and present a variety of forms of elaboration, illustrated with a series of examples.

The methods are applied to the expression of uncertainty in a case study. Mechanistic models are built in just about every area of science and technology, to represent complex physical processes. They are used to predict, understand and control those processes, and increasingly play a role in national and international policy making. As such models gain higher prominence, recipients of their forecasts are increasingly demanding to know how accurate they are. There is therefore a growing interest in quantifying the uncertainties in model predictions.

Uncertainty in model outputs, as representations of reality, arise from uncertainty about model inputs (such as initial conditions, external forcing variables and parameters in model equations) and from uncertainty about model structure.

Our case study is based on the Sheffield Dynamic Global Vegetation Model (SDGVM), which is used to estimate the combined carbon flux from vegetation in England and Wales in a given year. The extent to which vegetation acts as a carbon sink is an important component of the debate about climate change. We show how different approaches were used to characterise uncertainty in vegetation model parameters, soil conditions and land cover.

1. Introduction

Expert judgements are extensively used to aid analysis and decision-making in contexts where the available evidence is very limited, of mixed quality or only indirectly relevant. Their ability to assimilate complex and equivocal evidence, and to interpret it in the light of broader experiences make the judgements of experts invaluable in many applications.

The situations in which expert judgement is sought are characterised by uncertainty. We use expert opinion in order to obtain the most informative judgements and so we hope to minimise uncertainty, but it is unreasonable to hope to eradicate it. An important aspect of eliciting expert judgement is to characterise the expert's uncertainty accurately, and in particular not to understate that uncertainty.

This article is divided into three main sections. In Section 2 we explore the probabilistic representation of expert opinion. Although other representations have been suggested, we first argue that probability is the uniquely appropriate form to express the expert's knowledge about uncertain quantities. We then briefly consider different approaches to specifying probability distributions to represent expert knowledge about one or more uncertain quantities, with reference to the substantial literature in this field.

Section 3 introduces elaboration methods, which are ways of structuring the elicitation in terms of other quantities which may be easier to elicit. This is valuable particularly for simplifying the task for multivariate elicitation. Techniques of elaboration to separate information sources and to create independence, hierarchical, parametric and nonparametric elaboration are discussed and illustrated with a series of examples.

Section 4 is a case study illustrating various elaboration approaches. One of the major areas in which expert judgements are

E-mail address: a.ohagan@sheffield.ac.uk.

1364-8152/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2011.03.003

used is to provide knowledge about uncertain parameters in some mechanistic model, whose output will be used to inform scientific understanding or decision-making. We describe the application in some detail and show how different approaches were used to specify uncertainty about different groups of parameters.

2. Overview of probabilistic representation

2.1. Why probabilities?

How should we represent uncertainty? There are overwhelming arguments in favour of probability as the uniquely appropriate representation. First, there are axiomatic treatments which demonstrate that a person who wishes to make coherent decisions in the face of uncertainty must make those decisions according to a probability distribution. One of the simplest developments is given by DeGroot (1970), based on earlier work by Savage (1954). The ingredients of the theory are axioms about 'coherence'; coherent decisions satisfy natural logical conditions and prevent the person from being a 'sure-loser', i.e. being subject to a series of bets which guarantee a loss overall. For instance, if the person prefers decisions A to B and B to C then they must also prefer A to C. It is important also to recognise that the theorems do not imply that the person consciously uses some specified probability distribution, only that in order to make coherent decisions it is necessary to make them as if according to an underlying probability distribution (and a utility function for the outcomes of the decisions). By observing the decisions made by a coherent decision-maker, it would be possible in principle to deduce his or her underlying probability distribution.

A simplified argument based on the notion of fair bets is given in O'Hagan (1988), where it is shown that such bets must imply probabilities that behave according to the usual probability laws. These axiomatic results are relevant because if we represent uncertainty in a way which does not obey the laws of probability then decisions made on the basis of such a representation will not be coherent. Indeed, formulations which defy the laws of probability are often demonstrably, or even obviously, flawed. In particular, fuzzy logic has sometimes been advocated as an alternative to probability theory, yet has undesirable consequences. A critique is given by Lindley (1987).

From the perspective of the philosophy of science, Howson and Urbach (2006) present an extensive justification of the use of probability as the basis of scientific induction. Less mathematical or formal arguments reason from the way that probability naturally expresses uncertainty in random events. Alternative, but similarly compelling, axiomatic or rational arguments do not appear to have been advanced for other ways of representing uncertainty.

2.2. Approaches to quantifying uncertainty

Having argued for the inevitability of probability for representing uncertainty, this is not to say that probability distributions for uncertain quantities are easy to specify. Indeed, this is itself often given as a reason for abandoning probability in favour of some other representation. However, as we have seen in the particular case of fuzzy logic, any other approach sacrifices logical consistency/coherence. Our view is that careful formulation of probability distributions to represent expert knowledge may be costly, in terms of the time and energies of both the expert(s) and the facilitator, but that this is often more than justified by the value of expert judgements. One purpose of this article is to demonstrate that formulating expert judgements in probabilistic form is a practical reality, and hence that to decline to do so is misguided, defeatist and wasteful of valuable information (see O'Hagan and Oakley (2004), for some related discussion.).

For the purposes of this article, we identify the following three ways of formulating probability distributions.

1. *Direct elicitation*. Elicitation is the name usually given to the process of extracting expert knowledge about one or more uncertain quantities in the form of a (joint) probability distribution. Some basic principles of elicitation are set out in Section 2.3. By 'direct elicitation' we shall mean elicitation by asking the expert directly about the quantities of interest.
2. *Elaboration*. The alternative to direct elicitation is to express the quantities of interest in terms of other quantities, to elicit probability distributions for those constituent quantities (by direct elicitation) and then to derive the implied probability distribution for the quantity of interest. This is called 'elaboration' by O'Hagan (1986) and is also known in the elicitation field as 'structuring'.
3. *Bayesian analysis*. When data are available to inform the values of uncertain quantities, we can apply statistical methods to make inference about them. Bayesian inference is the appropriate paradigm here, with the posterior distribution being the required representation of uncertainty. Bayes' theorem is another form of elaboration, but it is useful to consider it as a separate technique.

Data are clearly brought explicitly into the process of specifying uncertainties when we use Bayesian analysis, whereas in direct elicitation they enter only implicitly as part of the expert's thoughts when forming the elicited judgements. In elaboration, data may be used in various ways to inform some of the constituent quantities. Direct elicitation always features, even when using elaboration (for some or all of the constituents) or Bayesian analysis (for the prior distribution), but it may be less influential in these methods. Note that in Bayesian analysis the expert's judgements are taken as formulating the prior distribution, and for this purpose the expert must not be aware of the data that are to be explicitly incorporated via Bayes' theorem. This limits the usefulness of Bayesian analysis when there is appreciable expert knowledge to complement the explicit data. In contrast, in direct elicitation (and in most forms of elaboration) the expert is supposed to be aware of all relevant data and to have incorporated that knowledge in the elicited judgements.

The relevant interpretation of probability for the expression of expert judgements is that of personal probability, or subjective probability — probabilities represent personal degrees of belief in whether the uncertain quantity will take particular values or will lie in particular ranges. The use of expert judgement inevitably entails the idea that the expert's beliefs and knowledge are different from those of other, less expert, people and hence the expert's probabilities are personal.

It is important to remember that in principle we need a full joint probability distribution for all the uncertain quantities of interest. Eliciting joint distributions for two or more quantities is difficult, and as yet few methods have been explored to provide guidance; see Daneshkhah and Oakley (2010). If the quantities are independent (in the judgement of the experts) then it is enough to elicit marginal distributions for each quantity independently. It is also important to recognise (and to clarify for experts) what a judgement of independence entails in the context of subjective probabilities. Teaching of elementary probability and statistics usually introduces independence for events if they are mechanically independent, i.e. there is no physical connection in how the random events are generated. When probability is a personal judgement, to assert that two uncertain quantities are independent

in the judgement of an expert is to say that if the expert were to learn something about one quantity it would not alter his/her judgements about the other.

Since methods for eliciting a distribution for a single uncertain quantity are relatively well researched and understood, independence greatly simplifies the task of eliciting a joint distribution for several quantities. In practice, therefore, it is common to attempt to reformulate the task by elaboration in terms of independent quantities.

2.3. Principles of elicitation

There is a considerable literature on elicitation, covering the fields of statistics, psychology, economics, decision analysis and others, and reviewed by O'Hagan et al. (2006). A recent review in the ecological literature is Kuhnert et al. (2010). When expert knowledge is important it is usual to adopt a formal elicitation procedure managed by a facilitator with expertise in probabilities and the process of elicitation. The facilitator's role is to ensure that the finally elicited distribution represents as accurately as possible the beliefs and knowledge of the expert(s). This requires an understanding of the research, much of it in the psychology literature, about how people judge probabilities. Some help for aspiring facilitators is available in the SHELF (SHeffield ELicitation Framework) package (http://tonyohagan.co.uk/shelf).

We present here the key elements of elicitation. Before discussing the actual process of eliciting specific judgements and deriving a probability distribution, it is important to appreciate that practical elicitation has other important components. The elicitation typically involves the following stages.

1. *Recruitment of experts*. This important first step involves identifying one or more experts with the relevant expertise, trying to ensure that all the major opinions are represented. The selected experts need to be recruited and committed to the task.
2. *Orientation and training*. Experts need to know something about the problem to which their opinions will contribute. It is useful at this stage to identify any potential conflicts of interest and to explore what each expert contributes to the pool of knowledge and opinion. Finally, it is extremely important to train the experts. Their expertise almost inevitably does not lie in probability, and the tasks that will be required of them will be unfamiliar to them. They also typically interpret probability according to the traditional long-run-frequency definition, rather than as a personal judgement.
3. *Definitions*. It is vital to define precisely the quantities to be elicited. At this stage it is also appropriate to consider elaboration as discussed in Section 3.
4. *Initial elicitation*. Only after these preliminary steps is it appropriate to elicit specific judgements from the experts, either individually or together by discussion. Preliminary specifications of probability distributions are usually derived at this stage.
5. *Feedback and revision*. The facilitator will now show the experts their judgements in various ways designed to provoke discussion and the possibility of revision. Revision of judgements is appropriate if they have unintended consequences, such as suggesting a bimodal distribution or very thin tails.
6. *Recording*. It is useful to create a formal record of the procedure and of the outcomes, so that the way in which the resulting probability distributions have been arrived at is documented and transparent.

This is just one of many ways to characterise the process, and useful formulations of elicitation protocols are also given in Low Choy et al. (2009) and Knol et al. (2010). See also the practical comparison of different elicitation mechanisms in O'Leary et al. (2009).

When several experts are involved, some practitioners elicit distributions from them separately and then apply a rule to combine them into a single distribution. However, several such rules of combination have been proposed and they all appear to have deficiencies, so that there is no consensus on a good choice; see O'Hagan et al. (2006), Section 9.2.2. The alternative is to bring the experts together with a view to eliciting a consensus distribution. This introduces additional challenges in assembling the experts and managing the interaction between them but has the benefit that it allows the experts to debate and share information. The SHELF system takes the view that group elicitation is the better approach.

We now review the principles of eliciting a single probability distribution. So consider a single uncertain quantity X. Since X can usually take any value in some range (e.g. any positive number) it is impractical to elicit the distribution in detail — this would entail eliciting probabilities such as $P(X \leq x)$ for all possible x values. In practice, we can only expect to obtain from an expert a rather small number of judgements, after which it becomes difficult for the expert to distinguish the necessary fine details or to avoid having answers heavily influenced by previous judgements. Therefore, we ask only for a small number of judgements. The quantitative judgements that are used in elicitation are almost invariably evaluations of probabilities. Although we might ask the expert to assess quantities such as means and variances, the evidence from the experimental literature is generally that these are evaluated less accurately than probabilities; see O'Hagan et al. (2006), Section 5.2.3. Moments are cognitively more complex constructs and highly sensitive to the thickness of a distribution's tails.

Practical elicitation is therefore based on judgements of probabilities. One example might be the quartile elicitation procedure in the SHELF system:

- Elicit the expert's median M, being the value for which the expert judges $P(X > M) = P(X < M)$.
- Elicit quartiles L and U such that the expert judges $P(X < L) = P(L < X < M)$ and $P(M < X < U) = P(X > U)$.

The judgement of equal probability is one that experts generally find relatively straightforward conceptually, and for which there is some evidence that these judgements are accurate. The quartiles might be augmented with a small number of additional probability judgements selected by the facilitator, but the elicitation will rarely attempt to get more from the expert than this.

Based on a small number of judgements like this, the facilitator now fits a distribution. For instance, if L and U are more or less equidistant from M it may be appropriate to fit a normal distribution. There will be potentially many distributions that would fit the expert's judgements acceptably well, but the facilitator is free to choose any distribution that he/she feels will not misrepresent the expert's beliefs. Whilst this may appear cavalier, the resulting distribution will often be adequate for the purpose for which the elicitation is being performed. There are two reasons for this seeming complacency. First, it is usually most important to capture features of the distribution such as location, spread and general shape, which can be deduced reasonably accurately from the elicited judgements. That is, the analysis for which the elicitation is carried out is rarely sensitive to more subtle features of the distribution. The second reason is that the facilitator's choice is really quite constrained. Any two distributions having (close to) the specified probabilities and a realistic shape (unimodal and not highly skewed) will look very similar to each other (Although bimodality or high skewness are quite feasible beliefs for an expert to hold regarding X,

it will generally become apparent in the elicitation that the facilitator should explore such possibilities. Then appropriate probabilities can be requested and suitable distributional forms fitted.).

Nevertheless, it is wise to check whether the fitted distribution does conform to the expert's beliefs. For instance, while it may only be necessary to elicit two separate judgements (e.g. the median and upper quartile) in order to fit a normal distribution it is sensible to elicit at least a third judgement (such as the lower quartile) in order to check the suitability of this choice. This is called over-fitting, i.e. eliciting more judgements than are strictly necessary to fit a preferred convenient family of distributions, in order to test the validity of that choice. The feedback step in the elicitation procedure (step 5) is another way to carry out such a check — for instance, the facilitator might fit a normal distribution and then feed back the 80th percentile for confirmation by the expert.

Elicitation is clearly an imprecise exercise, not just because of the somewhat arbitrary nature of the fitted distribution but also because the expert's probability judgements are inevitably imprecise. We have argued that the imprecision will often not matter in practice, but it is wise to check this. Allowing for a range of fitted distributions, and also for a fitting to probabilities that are within a range of accuracy of the elicited judgements, we can carry out a kind of sensitivity analysis — does such variation in the distribution lead to material changes in the subsequent analysis? This is related to the approach of imprecise probabilities (Walley, 1991), robust Bayesian analysis (Berger, 1984) or interval analysis (Ferson and Oberkampf, 2009), but those methods may be over-cautious because they do not allow for the reasonable notion that probabilities and fitted distributions at the extremes of the ranges are much less plausible than choices close to those originally elicited and fitted. A Bayesian approach in which imprecision is formally modelled is given by Oakley and O'Hagan (2007).

Thus stage 3 of the above elicitation procedure generally consists of two steps, (a) eliciting a few well-chosen probability judgements and (b) fitting a suitable distribution. For a single uncertain quantity, this is a quite standard and well-used approach, with plenty of guidance available in the SHELF package. As indicated already, eliciting a joint distribution for several quantities is generally harder and there is less knowledge in the field about good practice. In particular, it is not clear what kinds of joint or conditional probabilities can be most usefully and reliably elicited in step (a), and we have a rather limited choice of multivariate probability distributions for fitting in step (b). Elaboration in terms of independent quantities is currently the best approach.

3. Elaboration

The use of some particular forms of elaboration in elicitation is certainly not new, but it has not previously been formulated or explored as a general concept.

The term 'elaboration' was introduced by O'Hagan (1988) in the context of specifying individual subjective probabilities. The idea is that in order to specify the probability $P(E)$ of a proposition E it may be useful to express it in terms of other probabilities that are easier to judge. For instance, suppose that E is the proposition of getting at least one 6 when tossing three dice. If I were asked to judge $P(E)$ by direct evaluation, I would find it difficult to choose a value. I might say 0.5 or 0.45, but such assessments would be relatively crude. Instead, a better approach is to think in terms of the three separate dice values, giving a probability 1/6 to getting a 6 with each, and judging these propositions to be independent. Then $P(E) = 1 - (5/6)^3 = 91/216 = 0.4213$. The judgements regarding individual dice were easy to make (because I have no reason to expect any individual face to appear more often than any other), so this is a simple and accurate judgement of $P(E)$.

We extend the idea now to consider constructing a probability distribution for an uncertain quantity X by expressing it in terms of distributions for other quantities that may be easier to elicit.

3.1. Elaboration by information sources

One of the most useful elaborations for probabilities is Bayes' theorem. Suppose that your information about X comprises some data y as well as prior information H. The distribution for X should make use of both prior information and data, and so is $f(x|y,H)$ (We use the symbol f generally to represent a distribution expressed as a probability density function.). Bayes' theorem expresses this as a posterior distribution proportional to the product of the prior distribution $f(x|H)$ and the likelihood $f(y|x,H)$. These are usually easier to specify than $f(x|y,H)$. The prior distribution is more straightforward to think about because we do not have to mentally combine the two sources of information, prior and data, and specifying the likelihood for the data is a familiar statistical modelling task. Bayes' theorem is a useful elaboration because it separates the two sources of information. It is so important that in the context of Bayesian statistical analysis elicitation is invariably thought of as a tool to formulate the prior distribution. It could in principle be used to elicit the posterior distribution directly, but in practice Bayes' theorem is a much more reliable way to evaluate a posterior distribution.

When eliciting a prior distribution, or when there is no data that can reliably be modelled through a likelihood, other forms of elaboration can be equally powerful. One useful general device is to express X as some function of other uncertain quantities,

$$X = g(Z_1, \ldots, Z_k).$$

Then the distribution of X can be obtained from that of $\mathbf{Z} = Z_1, \ldots, Z_k$, by standard techniques for functions of random variables. However, the application of this elaboration requires the full joint probability distribution of $\mathbf{Z}$ to be elicited first, and we have remarked above on the difficulty of eliciting joint distributions. This kind of elaboration is useful specifically when all (or most) of the Z_i s are independent.

Example 1. Nitrate pollution.

In this example a new technique is proposed to reduce the level of nitrates in a river by treating groundwater in a trench as it enters the river. We wish to use expert judgement on the level of nitrates that will be found in the river if this technique is implemented, in order to determine its cost-effectiveness. The expert will be able to draw on evidence concerning current nitrate levels and the performance of the technique in test conditions, although there is uncertainty due to sampling in these data. Additional uncertainties concern what proportion of the nitrate in the river derives from the groundwater sources to be tackled with this technique, and about how performance in test conditions will translate into performance *in situ*. It is natural to express the nitrate level after (X) in terms of the level before (X_0), the proportion of current nitrate pollution that is due to the identified groundwater sources (P), the proportion of nitrates removed by the new treatment under test conditions (R) and a factor (F) representing the degree to which that performance will be achieved *in situ*, through the equation $X = X_0(1 - PRF)$. The expert might be willing to regard these four components as independent.

In this example, the Z_i s are X_0, P, R and F. Each of these is simpler to think about than X. Furthermore, there is different information relating to each. Distributions for X_0 and R can be derived from existing data (possibly with the use of Bayes' theorem). There is unlikely to be specific evidence for P or F so distributions for these will be elicited directly from the expert.

This kind of elaboration is useful because it separates the elicitation of X into several simpler elicitation tasks. A formulation in

which each of the Z_i s has distinct evidence will simplify the tasks because the expert does not need to synthesise many information sources in his/her head at once. Furthermore, this separation also naturally lends itself to judgements that the Z_i s are independent. Note also that the best knowledge for different components may lie with different experts, so they may be elicited in separate sessions.

As another example of this kind of elaboration, it is common to find situations in which expert input is required on a new quantity while there is good evidence on a similar but not identical quantity. Elaboration in terms of the similar quantity and a variable denoting the difference between this and the new quantity has exactly this property of separating the information sources.

Often, a suitable elaboration can be suggested simply by observing how an expert addresses the task of eliciting a distribution for X, and the nature of the different evidence sources that he/she considers relevant to that task.

3.2. Elaboration for independence

The preceding examples illustrate how elaboration can be useful when we only wish to elicit a distribution for a single quantity, and even if this entails eliciting distributions for two or more other quantities. Elaboration can also be used to deal with eliciting joint distributions for two or more quantities by reducing the problem to eliciting independent distributions. We sometimes refer to this kind of elaboration as structuring.

Example 2. Two drug treatment effects.

In this example a clinical trial will compare the effects of two drugs for treating the same medical condition. Let A and B be the observed effects of treatments 1 and 2 respectively in the trial; we wish to elicit expert opinion regarding these two unknown quantities. The first question is whether the expert (or experts) judges A and B to be independent. Remember that the appropriate interpretation of independence here is that if an expert were to learn more about A it would not change his/her beliefs about B. It is likely that they will not be independent. Specifically, if A were to turn out to be larger than expected then this suggests that the patient group recruited for the trial has higher responses than usual, and the expert would increase his/her estimate of B accordingly. We need to consider structuring approaches to break this dependence.

First suppose that treatment 1 (T1) is a standard treatment while treatment 2 (T2) is a new drug. In this situation it is common to think in terms of the effect of T2 relative to the standard treatment T1. This could be the incremental effect $B-A$ but is more usually the relative effect (or relative risk if the effect is mortality or some other undesirable outcome) B/A. The expert(s) might be willing to think of this relative effect as independent of the standard treatment effect A. Then we have a simple structuring in terms of independent variables A and $C = B/A$. We elicit separate distributions from the expert(s) for A and C and use these to derive the joint distribution of A and B. Elaboration of this kind forms part of the approach described in Nixon et al. (2009).

This is an example of the simplest kind of structuring, where k uncertain quantities $\mathbf{X} = X_1,\ldots, X_k$ are expressed as a function of k other quantities $\mathbf{Z} = (Z_1,\ldots, Z_k)$, such that the Z_i s are mutually independent. We next consider situations in which we may require more than k components in $\mathbf{Z}$.

3.3. Hierarchical elaboration

Hierarchical modelling was introduced as a general concept into Bayesian statistics by Lindley and Smith (1972). A simple example is when we wish to specify prior distributions for the mean yield obtained from growing each of 5 different varieties of wheat. An expert would not generally judge these to be independent, because learning that some varieties produce higher yields than he/she expected would lead to increased expectations for other varieties. The hierarchical model of exchangeability that was proposed by Lindley and Smith and has been widely used since would express the yields $X_1,\ldots, X_5$, as $X_i = M + D_i$, where M is an overall mean yield averaged across the spectrum of these and similar varieties and D_i is a deviation of X_i from this overall mean. Under the exchangeability model, $M, D_1,\ldots, D_5$ are independent. Positive correlations between the X_i s is broken by introducing the overall mean M. This is closely related to elaboration by information source because M causes correlation by being common to the X_i s. The elaboration separates information about overall level from information about variation between varieties.

Here is a rather more complex example.

Example 3. Bird abundance.

This example concerns the abundances (numbers present per hectare) E, F and G of three woodland species of bird in a given forested region. The expert would typically not regard these as independent; learning about the abundance of one species would change beliefs about the others. However, the direction of this dependence would not always be the same. Observing a high abundance of one species will suggest high values for the others if there is large uncertainty about the overall abundance of birds in the forests, but would suggest lower abundances for other species if they compete for food supplies and there is less uncertainty about overall abundance.

Having identified the overall abundance as driving correlation, it is natural to try to resolve that correlation by conditioning on it, so let $T = E + F + G + H$ be the total abundance of the birds in this forest, where H is the abundance of all other species. Also let $P_E = E/T$, $P_F = F/T$ and $P_G = G/T$. The expert may now judge T to be independent of P_E, P_F and P_G. However, it is not reasonable to suppose P_E, P_F and P_G independent — a higher value of one will suggest lower values for the others because their sum is constrained to be less than 1. The structuring has separated out the overall abundance factor T, whose uncertainty drives positive correlations between E, F and G, from the relative abundances P_E, P_F and P_G which are concerned with competition and are negatively correlated. In this case, the correlgation between P_E, P_F and P_G is not a practical bar to eliciting a joint distribution because there exists (a) a suitable family of distributions for such variables, the Dirichlet family, and (b) research concerning how to elicit judgements for fitting a Dirichlet distribution; see Chaloner and Duncan (1983). Alternatively it is possible to apply a further structuring, because in the Dirichlet distribution there is independence between P_E, $P_F/(1 - P_E)$ and $P_G/(1 - P_E - P_F)$, which breaks the Dirichlet's negative dependence.

A statistical note here is that if we elicit a gamma distribution for T and a Dirichlet distribution for P_E, P_F and P_G then E, F and G will be positively/negatively correlated depending on whether the shape parameter of the gamma distribution is less/greater than the sum of the parameters of the Dirichlet. This makes it clear how the strength of knowledge about T determines the direction of the correlation.

The distinction between the simple structuring of Example 2 and the hierarchical structuring of Example 3 can be seen clearly using some mathematical notation.

Example 2 is an instance of simple structuring, in which we define a one-to-one transformation from the vector $\mathbf{X}$ of quantities of interest to a vector $\mathbf{Z}$ whose distribution is simpler to elicit. Often, the intention is that the elements Z_i of $\mathbf{Z}$ are judged independent by the expert. Then by eliciting separate distributions $f_i(z_i)$ for each Z_i we obtain the joint distribution

$$f(\mathbf{z}) = \prod_i f_i(z_i)$$

and can derive the implied joint distribution for **X** from the usual change-of-variables formula.

In hierarchical structuring we introduce a latent quantity, say **W**, such that the distribution of **X** is simpler to elicit or model conditional on the value of **W**. And specifically the structuring is often such that the elements X_i of **X** become independent conditional on **W**. Then we can elicit separate distributions $g_i(x_i|\mathbf{w})$ for the X_i s conditional on **w** and $g(\mathbf{w})$ for **W**. The distribution for **X** is then obtained by marginalising with respect to **W**:

$$f(\mathbf{x}) = \int g(\mathbf{w}) \prod_i g_i(x_i|\mathbf{w}) \mathrm{d}\mathbf{w}.$$

Although eliciting conditional distributions is in principle difficult the structuring typically identifies the precise way in which each X_i depends on **W**. As a result, we can generally identify a transformation (technically known in statistics as a pivot) $\mathbf{Z} = \phi(\mathbf{X},\mathbf{W})$ whose distribution does not depend on **W**, from which $g(\mathbf{x}|\mathbf{w})$, or in the case of conditional independence the $g_i(x_i|\mathbf{w})$, can be simply derived. In Example 3, ϕ is the function that expresses *E*, *F* and *G*, with *T*, as the ratios P_E etc. This is a pivot because the ratios are judged independent of *T*.

Of course, it is important that the distribution of **W** is also feasible to elicit. In Example 3, **W** is the total abundance *T* and the fact that it is just a scalar quantity means that it is not hard to elicit a distribution for. When **W** is a vector of two or more quantities, further structuring may be used to simplify the specification of its distribution.

3.4. Parametric and nonparametric elaboration

We continue this development of elaboration methods by addressing problems with many uncertain quantities.

Example 4. Dose-response.

A pharmaceutical company wishes to predict the dose-response relationship for a new compound. The uncertain quantity now is a function $R(d)$, that gives the response for any dose *d*. It is, of course, not straightforward to elicit expert opinion in this case; in principle the function represents an infinite number of uncertain quantities. The obvious and simplest way to proceed is to assume a particular form of response, such as a probit model; $R(d) = \Phi(\alpha(d - \beta))$ where Φ is the standard normal distribution function and α and β are the parameters of the probit model. This reduces the elicitation problem to eliciting expert judgements about α and β (which will not usually be done by direct elicitation but by eliciting distributions for points in the curve, by extension of the approach of Kadane et al., 1980).

This example raises an important issue because the assumption of a probit (or other similar) model places restrictions on the possible shapes of the dose-response function $R(d)$. If the resulting distribution for $R(d)$ is viewed as a representation of the expert's beliefs, then the implication is that the expert is absolutely certain that the function will follow the assumed shape. More generally, this is just an instance of the general elaboration

$$\mathbf{X} = g(\mathbf{Z}).$$

We have already considered cases where **Z** has the same number of elements as **X** (simple structuring), more elements than **X** (elaboration by information sources or hierarchical elaboration), but Example 4 illustrates a case where it has fewer elements, which we call parametric elaboration. In Example 4 **X** is $R(d)$ and has effectively an infinite number of elements, whereas **Z** is (α, β) with just two elements, so the dimension reduction is dramatic, but wherever **Z** has fewer elements than **X** the elaboration implies an assumption about the possible combinations of values for the elements of **X**.

From one point of view, this is no different from assuming a particular distributional form at the stage of fitting a distribution to the elicited summaries. The expert cannot realistically specify beliefs about the function $R(d)$ for all, or even a large number of, the possible values of *d*, so some way to cut through the dimensionality is essential in practice. Another example of parametric elaboration is provided by Denham and Mengersen (2007), who use a software package (James et al., 2010) to fit an assumed logistic regression model for the geographical distribution of the brush-tailed rock-wallaby.

However, a parametric elaboration should not just be imposed on the expert without checking whether it is indeed an adequate approximation to their underlying beliefs. Techniques of overfitting and feedback can be valuable for this purpose. For instance, in the case of Example 4 instead of eliciting judgements from the expert about $R(d)$ at two values of *d* (which is all that is needed to infer distributions for α and β) we could ask about three values and check whether they are consistent with the assumed model (overfitting) or after fitting to judgements about two *d* values we could feed back the implied distribution for a third value.

If there is uncertainty about whether a particular parametric elaboration is appropriate, we might ask the expert to propose several possible parametric models, such that he/she is confident that the true relationship will fit one or other of the models. Then we elicit probabilities for each of the models containing the true dose-response relationship, and distributions for the various model parameters. This is related to the technique of Bayesian model averaging (Draper, 1995).

As a further refinement in the dose-response example, the expert might adopt a nonparametric model based on a Gaussian process. Oakley (2002) presents an approach to eliciting beliefs in this form. Nonparametric elaboration does not strictly constrain the possible values of **X** but it does constrain the kinds of probability distributions that the expert can express for **X**. As such it is entirely analogous to assuming a particular form of probability distribution for a single uncertain quantity.

Our final example illustrates how some problems can demand complex elaboration using a variety of parametric models to make the elicitation manageable.

Example 5. Groundwater permeability.

In modelling groundwater flow in a river catchment we are uncertain about the permeabilities of the ground at different locations. The groundwater model is likely to be run using permeabilities of blocks of a certain resolution rather than with point permeabilities, but nevertheless there are now many uncertain quantities. An expert will not generally regard the permeability at one location as being independent of permeabilities at neighbouring locations. Structuring for this problem might combine factors such as the following:

- Spatial correlation could be represented by a random field model.
- Trends across the catchment might be represented by a regression model.
- The further effects of topography could be handled by additional covariates.

This example will require complex structuring, as indicated by the above modelling points. An elicitation showing some of these features was discussed in O'Hagan (1998).

The message of the last two examples is the power of modelling. Conventional statistical models are largely ways of structuring data. All such modelling tools can be deployed to structure uncertain quantities for elicitation. So the skill set of the facilitator includes statistical modelling as well as subjective probability theory and psychology!

3.5. Elaboration and the problem context

The principles of elaboration as a generic methodology have apparently not been formulated before in the elicitation literature, although it is true that all of these techniques are familiar tools in some contexts. Particularly in Bayesian statistical modelling, hierarchical models and parametric models are widely used to structure prior knowledge. In this context the question of eliciting expert judgements typically only arises after the structuring has been done, e.g. after the hierarchical model has been formulated. Even in this situation, there may be scope for further elaboration to simplify the task, but the contexts in which elicitation is required are also more diverse and the potential for elaboration is often not obvious in the way the task is posed. We highlight two such contexts here.

1. *Decision analysis*. In decision analysis under uncertainty, a decision problem is characterised by a set $\mathcal{D}$ of possible decisions and a utility function $U(d, \mathbf{x})$ specifying the value of taking decision $d \in \mathcal{D}$ when the value of an uncertain quantity $\mathbf{X}$ is $\mathbf{x}$. The optimal decision rule is the one which maximises expected utility $U^*(d) = E[U(d, \mathbf{X})]$. In order to evaluate this, we need to specify a probability distribution to represent uncertainty about $\mathbf{X}$. In general, we allow that $\mathbf{X}$ is a vector of uncertain quantities. It comprises all those things that are uncertain but which would affect the utility of a decision. The elicitation task is therefore to express expert knowledge about $\mathbf{X}$, and the uncertain quantities are presented as part of the task. It is useful for the facilitator for such an elicitation to be aware of the possibilities for elaboration to make the task more manageable, particularly when there are many uncertain quantities.
2. *Inputs to mechanistic models*. Mechanistic computer simulation models, also known as process models, are widely used to represent complex real-world systems. We discuss such models much more fully in Section 4, where we refer to them as 'simulators'. One of the key tasks (known as uncertainty analysis) is to quantify uncertainty in the outputs of a simulator, for which we need first to specify uncertainty in the simulator inputs $\mathbf{X}$. Indeed, the simulator may be viewed as itself an elaboration (usually a very complex one) of the outputs in terms of the inputs, although Kennedy and O'Hagan (2001) and Goldstein and Rougier (2009) emphasise the importance of also elaborating the discrepancy between simulator and reality due to structural uncertainty. Elicitation is again the primary method for doing this, and again the uncertain quantities are presented as part of the task. They are the inputs that the process modeller deemed to be needed to specify the real system. Elaboration is an important tool that can easily be overlooked.

Suppose that in either of these contexts we apply an elaboration based on $\mathbf{X} = g(\mathbf{Z})$, expressing the quantities to be elicited in terms of (independent) quantities $\mathbf{Z}$. Having elicited a probability distribution for $\mathbf{Z}$ it can be a complex calculation to derive the implied (and required) distribution for $\mathbf{X}$, and having done so the resulting distribution for $\mathbf{X}$ is likely to be complex and far from convenient to work with in the problem context of decision analysis or uncertainty analysis of mechanistic models. The answer is to reformulate the problem in terms of $\mathbf{Z}$.

In the decision analysis problem the utility is $U(d, \mathbf{x}) = U(d, \mathrm{g}(\mathbf{z}))$, which expresses it as a function of $\mathbf{z}$. Expected utility can be evaluated using expectation of this utility function with respect to the (simpler) distribution of $\mathbf{z}$, i.e. $U^*(d) = E[U(d, g(\mathbf{Z}))$. In the uncertainty analysis of simulators, we can define the inputs to be $\mathbf{Z}$ instead of $\mathbf{X}$ and modify the computer program that evaluates the model to take input $\mathbf{z}$, apply the transformation $\mathbf{x} = \mathbf{g}(\mathbf{z})$ and then continue with the original computer code. An additional benefit of elaboration in this context is that the behaviour of the simulator may be simpler and more intuitive when expressed in terms of $\mathbf{Z}$ than $\mathbf{X}$.

Example 6. Nuclear waste disposal.

O'Hagan (1998) describes an application in which a simulator was built to quantify the risks in a proposed deep underground repository for nuclear waste. One set of uncertain inputs to this simulator were the hydraulic conductivities of rocks around the proposed repository site (whose values influence how rapidly any radiation travels away from the underground repository once the containment vessel has degraded). These conductivities were required on a grid of locations and could vary from location to location, yielding a large number of individual uncertain quantities C_i. They were elaborated using a hierarchical model in which each conductivity was represented as an overall mean log-conductivity M and deviations $R_i = (\log C_i) - M$. The implied joint distribution of the original conductivities

$$C_i = \exp(M + R_i)$$

would have been complex. It is simpler to regard M and the R_i s as the simulator inputs, incorporating the above equation as a preprocessing step within an extended model. When evaluating the impact of these uncertainties on the simulator output, it is often interesting to identify which uncertain inputs contribute most to the output uncertainty, a technique known as sensitivity analysis. In this simulator, we would expect M to have a strong impact on the rate of radiation transport, since an increase in conductivity at all locations has a stronger influence than an increase at a single location. The elaboration means that instead of each C_i having a modest impact we can identify a large impact with M and small (possibly unimportant) impacts for the R_i s.

Decision analysis is reviewed in Smith (1988), while uncertainties in mechanistic models are discussed in Santner et al. (2003), Saltelli et al. (2004) and the MUCM (http://mucm.ac.uk) Toolkit.

4. Case study: a mechanistic model for carbon fluxes

4.1. Uncertainty in mechanistic models

Computer simulation models are extensively used in science and engineering, and increasingly play an important role in decision and policy making. Some of the earliest examples were simulations of nuclear power plants and weapons, where the models took the place of experiments which could not be carried out. They are used these days to design and predict the performance and safety of almost every major engineering project, such as automobile/aero-engines, bridges, nuclear waste-disposal facilities or racing yachts. Another highly topical area is environmental modelling — weather forecasting has long been the domain for some of the most computationally-intensive models, and the debate on global warming is heavily dependent on climate models. Mechanistic models are often very complex and need to be

implemented in computer programs that may take hours (or even weeks) to run. By their nature, they are also most useful when observations of the real process are difficult to obtain (for reasons of cost, feasibility or ethics). In accordance with the growing literature on quantifying uncertainties in such models, we will refer to the model (and specifically to its implementation in a computer program) as a simulator.

The uses of simulators are also diverse, but are generally concerned with understanding and predicting complex real-world processes. Where understanding is the primary focus, it is often the qualitative behaviour of the simulator outputs that are of interest. For instance, modellers may be looking to see if the complex interactions of the model components suggest how new properties can be achieved, or whether the existing science is able to explain real-world behaviour. In the more high-profile uses of simulators, though, the objective is to predict (and thereby to control, optimise or prepare for) behaviour of the real-world process. Then it is not usually enough if the simulator outputs are qualitatively correct; we need them to be quantitatively accurate.

Indeed, where simulators are used to make decisions or to set policies, the accuracy of their predictions is a key consideration. Given that the simulator predicts a value y, how close do we expect this to be to the true real-world value z? Formally, z is uncertain and the challenge is to characterise or measure that uncertainty. There are various factors contributing to prediction uncertainty which we can characterise as parameter uncertainty and structural uncertainty. Parameter uncertainty in the simulator outputs arises from uncertainty in the inputs; that is, we are typically uncertain about the correct/best values to use for the various inputs that the simulator needs, and this feeds through into uncertainty about the predictions. Furthermore, though, even if we run the simulator with the ideal values for all its inputs it will not predict reality perfectly because no model is perfect; the discrepancy is due to structural uncertainty.

It may seem straightforward to assess structural uncertainty simply by comparing simulator outputs with reality. If we have enough observations of the real process and corresponding simulator predictions we can indeed characterise uncertainty by simple statistical computations. It is useful here to contrast mechanistic models, which are the subject of this case study, with empirical models. The latter are built simply by fitting some statistical relationship to observations of the real process. Although structural uncertainty is a more complex issue for empirical models than is generally acknowledged, the uncertainties in the fitted parameters and the residual fitting uncertainty are available from the statistical analysis. We refer to the type of models considered here as mechanistic because they try to represent the physical mechanisms which drive the process being modelled, incorporating current scientific knowledge and understanding.

4.2. The Sheffield Dynamic Global Vegetation Model and carbon flux

The Sheffield Dynamic Global Vegetation Model (SDGVM) (Woodward and Lomas, 2004) is a member of a class of mechanistic process models that represent the growth of vegetation. As a mechanistic model it tries to simulate all the relevant processes operating on the vegetation, including the following:

- Growth through photosynthesis and the availability of sunlight, water, warmth and nutrients.
- The annual cycle of bud burst, growth, leaf drop and dormancy over winter (with appropriate variation for evergreen plants).
- Evolution of the soil through leaf drop, rainfall and decomposition.

In particular, SDGVM simulates the terrestrial carbon cycle, as carbon dioxide is taken from the atmosphere (sequestration) and converted to biomass and free oxygen during photosynthesis, and is then released during respiration and by decomposition of root and leaf matter in the soil. Tree death, decomposition and recolonisation of empty spaces are also modelled, as is fire and harvesting of crops, all of which have consequences for the carbon cycle. The net result of carbon sequestration and release is the carbon flux, technically referred to as Net Biosphere Production (NBP), and is one of SDGVM's principal outputs.

As a global model, SDGVM operates on a relatively large scale, not concerned with individual trees but with an area of land that is covered by a particular type of vegetation. Seven types of vegetation (plant functional types, or PFTs) are simulated in SDGVM, of which four are present in England and Wales — Deciduous Broadleaf trees, Evergreen Needleleaf trees, Grass and Crops. An area with no vegetation or exposed soil is denoted as Bare (not strictly another PFT) and does not have any carbon flux. As a dynamic model, SDGVM runs iteratively in time, simulating the evolution of the vegetation and soils, and producing time series of outputs, including NBP. The version of the simulator used in this analysis operates on a daily time step.

The exercise upon which this case study is based was to estimate the total carbon flux (NBP) over England and Wales during the year 2000. For this purpose, the area of England and Wales was divided into 707 sites of 1/6th degree resolution in both latitude and longitude (so that each site covered an area of about 200 km^2). Like most environmental models, it requires a large quantity of input data. The principal inputs fall into the following groups.

- *Vegetation parameters*. For each of the four PFTs, a set of parameters describe characteristic growth variables for that functional type.
- *Soil parameters*. For each site, the soil is described by parameters such as the percentage of sand.
- *Land cover parameters*. For each site, the land cover is defined by the proportion of its area that is covered by each of the four PFTs (with the remainder being Bare).
- *Climate variables*. For each site and each time point in 2000, the climatic conditions are described by variables such as temperature and precipitation.

All of these inputs are uncertain to a greater or lesser degree. We will address uncertainty in each of the first three groups in Sections 4.4–4.6. The weather in 2000 is of course a matter of record, but the climate variables are subject to uncertainty because the data for a given site are obtained by interpolation from the network of recording sites. Nevertheless, this uncertainty was felt to be small and to be insignificant alongside the uncertainty in the first three parameter groups. Accordingly, it was not considered in the analysis reported here.

Other inputs, such as the area of each site, were not considered subject to any uncertainty, and the initial conditions (obtained by 'spinning up' the simulator over thousands of years to a stable state, prior to applying climatic forcing data for the latter part of the 20th century) were also treated as fixed.

4.3. Sensitivity analysis and emulation

In common with many simulators, the task of specifying probability distributions for SDGVM inputs is made more difficult by their sheer number. The analysis will not be feasible unless we can reduce the dimensionality of the input parameters. Fortunately, it is usually unnecessary to formulate all of the distributions carefully because simulator output typically depends principally on just

a few inputs. The trick is to identify which inputs are primarily responsible for the uncertainty in outputs of interest, whereupon we can either ignore the uncertainty about the remainder, or assign them only very crudely elicited distributions.

An important tool for identifying the influential inputs is sensitivity analysis. Consider the overall uncertainty on simulator outputs (i.e. parameter uncertainty) that is induced by input uncertainty. If we denote the simulator by η, so that $y = \eta(x)$ is the output from the simulator when given inputs x, then we indicate that x is uncertain by writing it as a random variable X. The output is also uncertain, denoted by $Y = \eta(X)$. There are various forms of sensitivity analysis, but one widely used and powerful approach is variance-based sensitivity analysis which calculates what proportion of the output variance $v = \mathrm{var}(Y)$ is attributable to uncertainty in each of the components of the input vector X.

Consider the first input X_1. If its value were known to be x_1 then this would generally reduce the uncertainty in Y; denote its new value by $v^* = \mathrm{var}(Y|X_1{=}x_1)$. We do not actually know the value of X_1 but we can assess how much learning its value would reduce uncertainty on average, by subtracting the expected value of v^* from v. The sensitivity index for X_i is thus defined to be the proportional reduction

$$s_i = \{v - E[\mathrm{var}(Y|X_i)]\}/v.$$

The i-th input X_i can be said to have very little influence on the output if its sensitivity index s_i is small.

However, the simulator will generally be complex and highly nonlinear in its inputs, so that we have to consider the possibility of interactions between the effects of different inputs. Interaction here is a similar concept to interaction between factors in conventional analysis of variance; the influence of input X_j may be greater or less (or different in other ways) when X_i is set at different values. Even if X_i has little influence overall (small s_i) it can have more influence through interaction with other inputs. Whereas the sensitivity index of X_i is the amount of uncertainty about Y that we expect to be removed if we were to learn X_i, its total sensitivity index is the proportion of uncertainty that would remain if we were to learn the values of all the other inputs:

$$t_i = E[\mathrm{var}(Y|X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots)]/v.$$

With interactions it can be shown that t_i will be larger than s_i (at least in the case where the inputs are all independent), and so it is safer to identify X_i as having little influence only when t_i is small.

These sensitivity indices are a valuable tool for separating the important inputs from the unimportant ones, whereupon we can concentrate on specifying uncertainty on the important inputs. Uncertainty about the unimportant inputs can be ignored or specified only crudely. However, there arises the question of how to compute these measures.

One standard method is Monte Carlo. To evaluate v, for instance, a large number of random X vectors are sampled from the specified (joint) distribution and $\eta(X)$ is evaluated for each sampled X. Then the estimate of v is the sample variance of these $\eta(X)$ values. Similarly, the second term in the formula for s_i can be computed by a nested sampling scheme in which X_i is fixed at a sampled value and then $\mathrm{var}(Y|X_i)$ is evaluated from joining that X_i to many randomly sampled values of the remaining inputs and computing the variance of the resulting $\eta(X)$ values; then this is repeated for many X_i values and the variances averaged to give $E[\mathrm{var}(Y|X_i)]$. The problem with the Monte Carlo approach is the very large number of runs of the simulator to evaluate all the different $\eta(X)$ values. For anything but the simplest simulators, doing sensitivity analysis this way involves an impractical amount of computation time.

Various improved sampling methods have been proposed to increase the efficiency of Monte Carlo (see Saltelli et al., 2004) but there is also a very powerful alternative to Monte Carlo. This is to build an emulator. Formally, an emulator of a simulator represented by the function $\eta(\cdot)$ is a probability distribution for the function. It is built using a training sample of simulator runs. Let the sample be $y_1 = \eta(x_1)$, $y_2 = \eta(x_2)$,…, $y_n = \eta(x_n)$. That is, the training sample comprises the outputs y_1, y_2,…, y_n from n different inputs x_1, x_2,…,x_n, called design points. The training sample tells us what the output from the simulator will be when we run it at the design points. The emulator allows us to estimate from these data what the simulator will produce at any other input x.

In this way, the emulator can serve as a fast surrogate for the simulator. We could, for instance, now apply the Monte Carlo method to evaluate overall uncertainty v and various sensitivity indices, using the emulator's estimates for $\eta(x)$ instead of running the simulator itself at all of these inputs. As long as computing the emulator's estimates is very fast, using the emulator as a surrogate for the simulator in this way will allow all kinds of complex operations that would have been impossible using the simulator itself. However, such a procedure clearly does not compute v or any other quantity of interest exactly, no matter how large a Monte Carlo sample we draw, because the emulator is not a perfect substitute for the simulator. It computes an estimate. A number of ways of producing fast surrogates for a simulator have been proposed, and are often referred to as meta-models, but we reserve the word 'emulator' for a particular type of statistical meta-model.

An emulator is more than just a way of computing an estimate for $\eta(x)$ at any x. As mentioned before, it is a full probability distribution for the function $\eta(\cdot)$, and in particular at any x it provides a probability distribution for $\eta(x)$. We can regard the mean of this distribution as the emulator's estimate of $\eta(x)$, but the variance gives a measure of uncertainty around that estimate. An emulator should provide estimates and variances that reflect what is known about the simulator. For instance, at the design point x_i we know $\eta(x_i)$, so the emulator's mean for $\eta(x_i)$ should be y_i and its variance should be zero. At other points, the variance of $\eta(x)$ will depend on how far it is from a design point. Close to design points, we 'almost' know $\eta(x)$ but further away our uncertainty is larger. Given that the emulator includes an expression of uncertainty we can go beyond simply using the emulator as a fast surrogate to compute an estimate of a quantity such as v or s_i. We can provide a statistical measure of uncertainty, a variance, for that quantity (or even in principle a whole distribution). The emulation variance expresses the uncertainty about the quantity of interest due to using an emulator rather than the simulator itself. In general, the larger the training sample, the smaller will be the emulation uncertainty.

The most widely used form of emulator is based on a Gaussian process. A tutorial on Gaussian process emulators is given by O'Hagan (2006) and more detailed specifications and advice on building emulators can be found in the MUCM toolkit — www.mucm.ac.uk/toolkit. MUCM (Managing Uncertainty in Complex Models) is a large research project funded by Research Councils UK to develop emulator technology; more information can be found on the MUCM website (www.mucm.ac.uk).

Theory for carrying out sensitivity analysis using emulators was presented by Oakley and O'Hagan (2004). Sensitivity indices, computed by emulation, formed an important component of the case study. We now turn to the task of specifying probability distributions for the uncertain parameters of SDGVM.

4.4. Direct elicitation for vegetation parameters

In SDGVM, each PFT represents a type of vegetation sharing similar growth properties. These are represented by each PFT

having its particular values for a number of parameters. Some examples are:

- Leaf lifespan: maximum number of days before leaves are shed.
- Minimum stem growth: minimum growth in a year for a plant to remain viable.
- Seeding density: average number of seeds per square metre.
- Budburst limit: maximum number of degree-days before budburst occurs.

All of these parameters are unknown. They may have been measured empirically for some individual species that fall within a particular PFT but certainly not for all such species (e.g. for oak and beech but not for all deciduous broadleaf species), and even where measurements are available they are subject to observation and sampling errors. Indeed, given that a parameter like leaf lifespan will have different typical values for different species within a PFT, its value for the PFT as a whole is more loosely defined. We can think of it as an average value, averaged over the species in that PFT, but clearly not all species will be equally represented in England and Wales as a whole, or within one of the 707 sites to which SDGVM is applied. We elicited the knowledge of Professor Ian Woodward, who is an expert on these parameters and the designer of SDGVM.

The 'true' or 'best' value for a parameter in a PFT depends on the site, because it depends on the mix of species present in that site. So in principle we need to elicit a joint distribution for all the PFT parameters at all the 707 sites. This is only feasible by focussing the elicitation on the key uncertainties. The elicitation was conducted in two stages. First, for each PFT, a simple range of possible values was elicited for each parameter at a single 'typical' site. Normal distributions were assigned with the bulk of their probability within these ranges and other inputs fixed at nominal values. Emulators were built for NBP output at the chosen site assuming the site was covered with each of the PFTs, and sensitivity analysis performed to identify the most influential parameters for each PFT. Fig. 1 shows some of the sensitivity plots for parameters of the Deciduous Broadleaf PFT. These are plots of $E[\mathrm{var}(Y|X_i)]$, known as the main effect of X_i. A flat graph such as that for the leaf mortality index is associated with a small sensitivity index and denotes a relatively non-influential parameter. In contrast, we see that the graph for soil sand percentage has more variation and so will be associated with a relatively large sensitivity index. We can identify influential inputs from the sensitivity indices alone, but the graphs show not only which inputs matter but also how they affect the output. They will often serve to raise questions about whether the simulator has face validity; for instance, it was sensible to check that the soil sand percentage graph accords with the experts' intuition.

In the second stage, more careful elicitation was carried out for each of the influential parameters. One case in which this yielded a substantially different distribution from the initial elicitation was the leaf lifespan for the Evergreen Needleleaf PFT. Whereas leaf lifespan is always less than a year for deciduous trees, evergreen leaves can last for several years before dropping. The initially elicited range went from less than two years to more than four, so that a normal distribution was assigned across this range with a large variance. In the second stage, it became clear that leaves would generally drop in the summer or autumn, but not in winter or spring. This multimodal distribution in Fig. 2 was elicited by elaborating in terms of the number of growing seasons (integer number of years) and the time within the growing season (the fractional part of the length in years), which were judged independent.

There remained the question of converting individual distributions into a joint distribution across parameters and sites. The expert judged different parameters to be independent, but values of the same parameter at different sites would clearly be related. Correlations were assigned by asking the expert to think about how much the parameter values might differ at an individual site from the national average (which was a matter of thinking about how

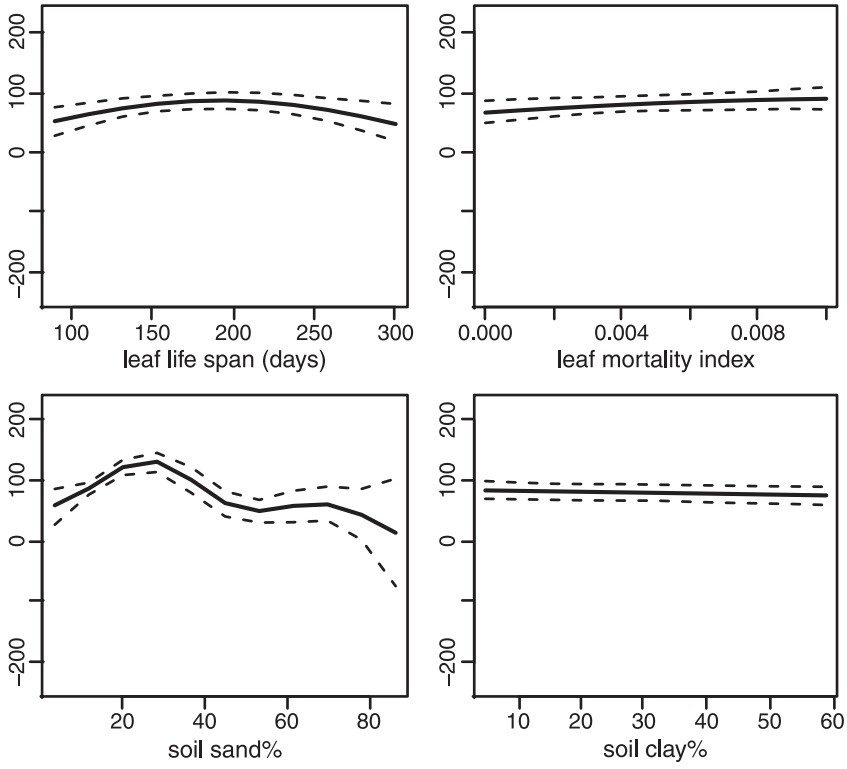

Fig. 1. Main effect plots for four deciduous broadleaf parameters. Emulated means (solid lines) and 95% bounds (dashed).

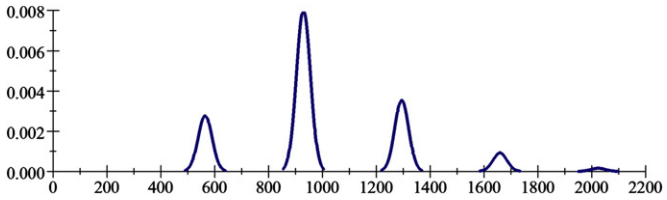

Fig. 2. Distribution of leaf lifespan (in days) for Evergreen Needleleaf trees.

species mix might vary from site to site), and to assess his uncertainty about a national average value. This is based on the hierarchical elaboration at the beginning of Section 3.3. The variance of each X_i (parameter value at a site) is the variance of the national average M plus the variance of D_i, which represents how much parameter values vary from site to site. The covariance between X_i and X_j is the variance of M, from which we can deduce correlations.

In the case of Evergreen Needleleaf leaf lifetime, correlations were separately considered for the number of seasons and time within season. Further details of the vegetation parameters elicitation exercise can be found in Kennedy et al. (2008).

4.5. Elaboration for soil parameters

The UK is fortunate in having detailed soil maps (although the primary reason for the case study being limited to England and Wales was difficulty in obtaining access to the Scottish data), with data on all the key parameters being available for pixels of size 1 km square. Estimates could therefore be obtained simply by aggregating these maps to the level of SDGVM sites. For instance, one influential soil parameter is the sand percentage, and an estimate of sand percentage for the whole site could be obtained by averaging the percentages in all the pixels within the site. However, there is uncertainty regarding the accuracy of such values. First, the raw data have measurement and sampling errors, but more significantly, the average over the site may not be the best value to use in SDGVM because of nonlinearity in the simulator output. For instance, although the bulk of the site may have low sand content, the parts with higher sand percentage may contribute disproportionately more (or less) NBP.

Uncertainty regarding the aggregated parameter values is related to the variance of the pixel estimates within each site, for both the principal components of uncertainty. The component of uncertainty due to sampling and measurement error is naturally described by the within-site sample variance divided by the number of pixels (although will be inflated because the pixel values are not independent). And the uncertainty due to simulator nonlinearity will also be greater the more variability there is in parameter values across the site. We can elaborate the task of setting a probability distribution for the site-level parameter in a given site in terms of the true aggregate value in the site and the difference between the correct SDGVM parameter setting and that true aggregate value due to nonlinearity. Uncertainty about the first component was derived mechanically as a normal distribution with mean equal to the sample average and variance equal to the sample variance divided by sample size. Expert judgement (of soils expert Dr. Andreas Heinemeyer) was used to assign normal distributions on the second component with variances equal to half the within-site sample variance.

The quality of the data in this case meant that it was judged that values in different sites and/or for different soil parameters would be independent. It may be noted that the assumption of independent normal distributions for proportions that are necessarily in the range 0 to 1 and must sum to less than 1 is strictly inappropriate, but a more technically correct analysis based on Dirichlet distributions produced essentially identical results. Further details can again be found in Kennedy et al. (2008).

4.6. Bayesian analysis for land cover parameters

A high-resolution map also exists for land cover over England and Wales, the LCM2000 (Haines-Young et al., 2000), which classifies land cover in one of 26 classes for 25 m square pixels. The 26 classes were mapped to SDGVM's 4 *PFT*s and aggregated to SDGVM site level to produce percentages of each PFT in each site. This is similar to the soil data, but in this case the raw data are more complex. LCM2000 is produced by satellite observation and its classification is imprecise. Another dataset, the CS2000 (Fuller et al., 2002) compares LCM2000 land cover with ground truth at a sample of locations, producing a table cross-tabulating LCM2000 class versus truth. This is known in the field as a confusion matrix. Not only does this allow us to quantify the magnitude of errors in the site aggregate percentages, but it also reveals a degree of bias.

A Bayesian statistical model was developed by Cripps et al. (2008) to make statistical inferences about true land cover based on LCM2000 data and the CS2000 confusion matrix. This allowed us to formally express uncertainty about land cover at the SDGVM site level across England and Wales as a joint posterior distribution. The distributions for these parameters in our case study were therefore specified using Bayesian analysis. Instead of eliciting distributions directly, we needed to consider distributions for the parameters in the Bayesian model. These included LCM2000's underlying mis-classification rates and true land cover at each LCM2000 pixel. Prior distributions were required for these, but in both cases the data are substantial enough to make prior information more or less irrelevant. It was therefore acceptable to use standard weak prior distributions rather than to elicit prior information formally. However, one component of the model required expert prior judgement.

CS2000's confusion matrix provides strong information about how often, for instance, a pixel which is truly 'bracken' is mis-classified as 'dwarf shrub heath', but it contains no information about spatial correlations in the mis-classifications. For instance if a 'bracken' pixel is mis-classified as 'dwarf shrub heath', how probable is it that a neighbouring 'bracken' pixel will be similarly mis-classified? Neighbouring pixels are most probably classified from the same satellite image, in the same light conditions at the same angle, so it was believed that such spatial correlations could be substantial. Without any data to inform these correlations, expert judgement was essential. However, this was a potentially difficult elicitation task involving many different kinds of mis-classification. Cripps et al. (2008) describes how this was reduced to a single judgement of a parameter d which controlled the site-level effect of such correlation.

4.7. Results

Having formulated probability specifications for SDGVM inputs across the 707 sites, the final part of the case study was to propagate that uncertainty through the simulator and to analyse the resulting uncertainty in the NBP output. The total NBP aggregated across England and Wales represents the net amount of carbon dioxide removed from the atmosphere by the region's vegetation in the year 2000. This is one important factor mitigating the emissions of greenhouse gases in the UK, and so has implications for the climate change debate. The overall results are reported in Harris et al. (2010), together with technical details about the emulation and sensitivity analysis computations.

Fig. 3(a) shows the standard deviation of NBP at each site, i.e. the total uncertainty due to all sources. Fig. 3(b) shows the standard

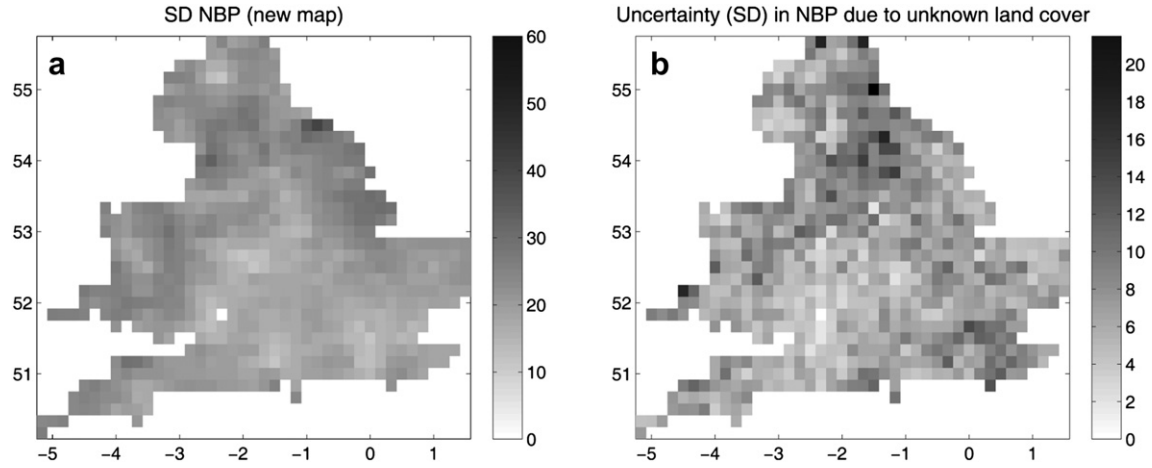

Fig. 3. Standard deviation of NBP plotted by pixel: (a) due to all uncertainty sources; (b) due to land cover uncertainty.

deviation due to land cover uncertainty alone. Notice the different scales for the two graphs. The uncertainty due to land cover is typically much smaller than the overall output uncertainty.

The estimated total NBP was 7.46 MtC, with a standard deviation of 0.54 MtC. The sensitivity analysis showed that the percentages of overall NBP uncertainty decomposed into 89% due to uncertainty in vegetation and soil parameters and just 4% due to land cover uncertainty. The remaining 7% was due to emulation and interpolation, i.e. to the fact that we used emulation and other statistical techniques because the complexity of SDGVM meant it was totally impractical to run it thousands of times at every site.

5. Conclusions

Expert judgement regarding uncertain quantities is an important source of information in a variety of contexts. The appropriate way to capture the expert's knowledge and beliefs is in the form of a subjective probability distribution, and this is the form required in several contexts, including prior distributions for Bayesian analysis, quantities affecting the utility in decision problems or inputs to computer simulation models. The process of formulating expert knowledge in probabilistic form is known as elicitation. Section 2 has provided an overview of the techniques of elicitation, with the following key recommendations.

- Where expert judgement is required in order to quantify knowledge about some uncertain quantity or quantities X, it is useful to identify three principal approaches. Direct elicitation asks the expert(s) directly for judgements about X. Elaboration represents X in terms of some other uncertain quantities Y and asks the expert(s) for judgements about Y (by direct elicitation). Bayes' theorem is a particular elaboration that is sufficiently important to be recognised as a distinct approach. It identifies particular data Z and elicits expert judgements about X prior to receiving the data, and about the relationship between X and Z (using direct elicitation or further elaboration). The case study in Section 4 illustrates these three approaches.
- Expert knowledge can be a highly valuable source of information, but the quality of expert judgements depends on the care and expertise that is employed in their elicitation. A formal elicitation process based on research and recognised good practice in the field is recommended, in order to obtain elicited distributions that accurately reflect the expert's knowledge.
- An elicitation protocol should cover all the stages of the elicitation process. One way to characterise the different stages is (1) recruitment of experts, (2) orientation and training of the experts, (3) formal definition of quantities to be elicited, including elaboration to relate these to the quantities of interest, (4) initial elicitation of judgements and fitted distributions, (5) feedback to promote critical review of judgements, with revision of those judgements where appropriate, and (6) creation of an agreed record of the elicitation.
- It is often important to synthesise the knowledge of several experts, in order to maximise the information or to ensure coverage of a range of expert opinion. Although some authorities on elicitation are happy to elicit distributions separately from a number of experts and then to combine these distributions using a mathematical aggregation rule, we advocate the approach of group elicitation. Eliciting a single consensus distribution from a group of experts has the advantage of allowing them to share opinions, discuss interpretations of relevant evidence, etc., and avoids the more or less arbitrary choice of an aggregation rule.
- In direct elicitation of a distribution, one can realistically only elicit a relatively small number of judgements from the expert(s). Research indicates that experts can more reliably assess probabilities than distribution summaries such as the mean, variance or other moments. Quantiles are often elicited, such as the median, or probability intervals. However, experts do not assess extreme quantiles or intervals with high coverage (such as 90% or 99% intervals) well.
- Having elicited a small number of probability judgements, a distribution from some convenient parametric family is then fitted to these elicited values. In practice, the choice rarely matters because all distributions having plausible shape (unimodal and not too heavily skewed) that fit the elicited values will be similar. Furthermore, in most situations for which elicitation is employed as an input to some analysis or decision-making, the result typically depends only on the location and spread of the distribution, which should be fixed quite tightly by the elicited probabilities. Nevertheless, it is important to test the appropriateness of a fitted distribution using over-fitting and/or feedback.
- Where it is felt that a conclusion may be sensitive to small changes in the elicited distribution, the choice of fitted distribution may matter, as may the fact that the experts' judgements are also inevitably subject to imprecision. This can be assessed informally by varying the distribution consistent with the elicited values (and with allowance for imprecision in those values). It may also be assessed more formally, but care is needed. Methods such as imprecise probabilities or interval analysis will tend to

overstate the consequences of imprecision because they do not allow for the reasonable judgement that probabilities and distributions close to the stated and fitted values are more plausible than those at the extremes of the allowed ranges.

Section 3 introduces elaboration, a generic technique for simplifying elicitation tasks. It is acknowledged that particular forms of elaboration are far from new, and indeed have been commonly practised in some contexts, but it is argued that the general concept has not been previously identified.

Elaboration usually expresses X in terms of two of more other uncertain quantities, but the task of eliciting a joint distribution for these is unlikely to be much simpler than direct elicitation of X unless these quantities are judged by the expert to be independent. The more powerful forms of elaboration exploit underlying independences, and in fact elaboration is particularly useful when used to express several non-independent quantities in terms of others that may be judged independent.

Elaboration is a very flexible concept and any taxonomy is likely to be incomplete, but the following forms of elaboration are identified here.

- Elaboration by information sources is based on the idea that it is easier to elicit a distribution for a quantity if the evidence for it is simple, and particularly if it derives from a single source. Where several sources of information are relevant, it may be that each provides evidence relating to a different aspect or part of the problem. It may then be feasible to express X in terms of quantities each of which is informed by a single source (or at least fewer sources). Furthermore, where each component quantity is informed by different sources the expert may reasonably judge them to be independent. Bayes' theorem is a form of elaboration by information sources, separating the roles of prior information and data. In Example 1 (nitrate pollution) several information sources are separated by the elaboration.
- There are several ways to use elaboration to address dependence between quantities of interest. The simplest form of elaboration for independence transforms the quantities X into the same number of other quantities Y but such that the elements of Y may be judged independent. Example 2 (two drug treatment effects) is a simple example of this form of elaboration in which the effect of a new drug is expressed relative to the effect of the established drug.
- Hierarchical elaboration is a more general kind of elaboration for independence based on identifying an underlying or 'latent' quantity (or quantities), uncertainty about which is the cause of correlation between two or more quantities of interest. The elaboration is in terms of the latent quantity and other quantities (such as ratios or differences) expressing the relationship of the quantities of interest to the latent quantity. Example 2 could be dealt with hierarchically by identifying the effect of the sample of patients recruited to the trial as a latent quantity and then expressing both drug effects relative to this common factor. In Example 3 (bird abundance) the abundance of three bird species is elaborated in terms of a latent quantity which is the combined abundance of all species.
- There is a close relationship between hierarchical elaboration and hierarchical Bayesian modelling, and indeed there are counterparts in elaboration of all kinds of statistical models. Two more complex examples are presented to illustrate the use of parametric and nonparametric elaboration. Example 4 (dose-response) concerns elicitation of a dose-response relationship for a new drug, where the assumption of a standard dose-response model, such as probit model, could be made. Discussion of Example 5 (groundwater permeability) suggests a combination of several different model components. Elaboration through models has the power to reduce very complex elicitation tasks to a few components. However, it is important to recognise that models (even, to some extent, nonparametric models) will typically constrain the possible distributions that can be derived for X, and may entail some sacrifice of information.

The use of models also blurs the distinction between elaboration for elicitation and more familiar kinds of modelling. In some contexts, it is usual for modelling to take place before and outside the elicitation task, so that the quantities of interest for which elicitation is required are already a simplification of other quantities which are the real focus of the motivating analysis. Even in such cases, further elaboration may be useful. Example 6 (nuclear waste disposal) illustrates these ideas.

Section 4 presents a case study that demonstrates some of the variety of ways that are used to specify uncertainty in unknown inputs to mechanistic models. Mechanistic models (also known as process models or science-based models) are an example of where extensive modelling has already expressed some quantities of interest (the model outputs), usually through a complex system of equations, in terms of other quantities, the model inputs. In order to characterise uncertainty about the outputs, we need (a) to elicit distributions for the inputs and then (b) to propagate that uncertainty through the model. Both tasks can be very demanding. We describe how the process of building and using emulators plays a key role in facilitating uncertainty and sensitivity analyses of model outputs. When it comes to specifying probability distributions for uncertain inputs, elaboration often plays an equally important role.

Our case study concerns assessing the impact of vegetation on atmospheric carbon dioxide using the Sheffield Dynamic Global Vegetation Model (SDGVM). The model has a great many inputs, and different approaches were used to elicit distributions for different groups of parameters.

- Uncertainty about parameters describing the behaviour of different types of vegetation was specified primarily by direct elicitation.
- For parameters describing the composition of soils, there was substantial evidence in the form of soil maps. However, whilst these maps provided point estimates of soil composition specifying the uncertainty about such estimates was more difficult. A simple elaboration employed expert judgement to quantify uncertainty about one component of the aggregated proportions.
- There is uncertainty also about what kinds of vegetation are present, and in what proportions, at any given site. Here we had good data in the form of a satellite-derived land cover map and a 'confusion matrix' obtained by comparing the satellite assessments of land cover with ground truth on a sample of sites. Bayes' theorem was used in this case.

Recognition and elicitation of input uncertainty in this case study showed that the simple estimate of carbon flux obtained by plugging best estimates of all parameters into SDGVM almost certainly gave an optimistic picture of how effective the vegetation in England and Wales was in the year 2000 as a carbon sink.

Acknowledgements

The case study was conducted by the Centre for Terrestrial Carbon Dynamics (CTCD), supported by grant number RA101879 from the UK Natural Environment Research Council. Soils data were

provided by the National Soil Research Institute (licence 02/RGOB/171/OD7063 V). Substantial contributions to the work presented here were made by various CTCD members, particularly Eds. Cripps, John Paul Gosling, Keith Harris, Andreas Heinemeyer, Marc Kennedy, Tristan Quaife, Shaun Quegan and Ian Woodward. I am also grateful to Jeremy Oakley for numerous discussions regarding the derivation of expert probability distributions.

References

Berger, J.O., 1984. The robust Bayesian viewpoint. In: Kadane, J.B. (Ed.), Robustness of Bayesian Analyses. Elsevier Science, pp. 63–144.

Chaloner, K.M., Duncan, G.T., 1983. Assessment of a beta prior distribution: PM elicitation. The Statistician 32, 174–180.

Cripps, E., O'Hagan, A., Quaife, T., Anderson, C.W., 2008. Modelling Uncertainty in Satellite Derived Land Cover Maps. Research Report No. 573/08. Department of Probability and Statistics, University of Sheffield. Available from. http://tonyohagan.co.uk/academic/pdf/landcover.pdf.

Daneshkhah, A., Oakley, J.E., 2010. Eliciting multivariate probability distributions. In: Böcker, K. (Ed.), Rethinking Risk Measurement and Reporting. Risk Books, London.

DeGroot, M.H., 1970. Optimal Statistical Decisions. McGraw-Hill (Reprinted: Wiley Classics Library, 2004.).

Denham, R., Mengersen, K., 2007. Geographically assisted elicitation of expert opinion for regression models. Bayesian Analysis 2, 99–135.

Draper, D., 1995. Assessment and propagation of model uncertainty (with discussion). Journal of the Royal Statistical Society B 57, 45–97.

Ferson, S., Oberkampf, W.L., 2009. Validation of imprecise probability models. International Journal of Reliability and Safety 3, 3–22.

Fuller, R.M., Smith, G.M., Sanderson, J.M., Hill, R.A., Thomson, A.G., Cox, R., Brown, N.J., Clarke, R.T., Rothery, P., Gerard, F.F., 2002. Countryside Survey 2000, Module 7 – Land Cover Map 2000 Final Report. Centre for Ecology and Hydrology, Monks Wood, Cambridgeshire.

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

Haines-Young, R.H., Barr, C.J., Black, H.I.J., Briggs, D.J., Bunce, R.G.H., Clarke, R.T., Cooper, A., Dawson, F.H., Firbank, L.G., Fuller, R.M., Furse, M.T., Gillespie, M.K., Hill, R., Hornung, M., Howard, D.C., McCann, T., Morecroft, M.D., Petit, S., Sier, A.R.J., Smart, S.M., Smith, G.M., Stott, A.P., Stuart, R.C., Watkins, J.W., 2000. Accounting for Nature: Assessing Habitats in the UK Countryside. Department of the Environment, Transport and the Regions, London.

Harris, K., O'Hagan, A., Quegan, S., 2010. The Impact of Land Cover Uncertainty on Carbon Cycle Calculations Available from. http://tonyohagan.co.uk/academic/pdf/CTCDPaper_v6.pdf.

Howson, C., Urbach, P., 2006. Scientific Reasoning: The Bayesian Approach, third ed. Open Court Publishing Company.

James, A., Choy, S.L., Mengersen, K., 2010. Elicitator: an expert elicitation tool for regression in ecology. Environmental Modelling & Software 25, 129–145.

Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S., Peters, S.C., 1980. Interactive elicitation of opinion for a normal linear model. Journal of the American Statistical Association 75, 845–854.

Kennedy, M.C., O'Hagan, A., 2001. Bayesian calibration of computer models (with discussion). Journal of the Royal Statistical Society B 63, 425–464.

Kennedy, M.C., O'Hagan, A., Anderson, C.W., Lomas, M., Woodward, F.I., Heinemeyer, A., Gosling, J.P., 2008. Quantifying uncertainty in the biospheric carbon flux for England and Wales. Journal of the Royal Statistical Society A 171, 109–135.

Knol, A.B., Slottje, P., van der Sluijs, J.P., Lebret, E., 2010. The use of expert elicitation in environmental health impact assessment: a seven step procedure. Environmental Health 9, 19.

Kuhnert, P.M., Martin, T.G., Griffiths, S.P., 2010. A guide to eliciting and using expert knowledge in Bayesian ecological models. Ecology Letters 13, 900–914.

Lindley, D.V., Smith, A.F.M., 1972. Bayes estimates for the linear model (with discussion). Journal of the Royal Statistical Society B 34, 1–41.

Lindley, D.V., 1987. The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. Statistical Science 2, 17–24.

Low Choy, S., O'Leary, R., Mengersen, K., 2009. Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. Ecology 90, 265–277.

Nixon, R.M., O'Hagan, A., Oakley, J.E., Madan, J., Stevens, J.W., Bansback, N., Brennan, A., 2009. The rheumatoid arthritis drug development model: a case study in Bayesian clinical trial simulation. Pharmaceutical Statistics 8, 371–389.

Oakley, J.E., O'Hagan, A., 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. Journal of the Royal Statistical Society B 66, 751–769.

Oakley, J.E., O'Hagan, A., 2007. Uncertainty in prior elicitations: a nonparametric approach. Biometrika 94, 427–441.

Oakley, J.E., 2002. Eliciting Gaussian process priors for complex computer codes. The Statistician 51, 81–97.

O'Hagan, A., Oakley, J.E., 2004. Probability is perfect, but we can't elicit it perfectly. Reliability Engineering and System Safety 85, 239–248.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T., 2006. Uncertain Judgements: Eliciting Expert Probabilities. John Wiley and Sons, Chichester.

O'Hagan, A., 1988. Probability: Methods and Measurement. Chapman and Hall, London.

O'Hagan, A., 1998. Eliciting expert beliefs in substantial practical applications. The Statistician 47, 21–35 (with discussion, pp. 55–68).

O'Hagan, A., 2006. Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering and System Safety 91, 1290–1300.

O'Leary, R.A., Mengersen, K., Murray, J.V., Low Choy, S., 2009. Comparison of three expert elicitation methods for logistic regression on predicting the presence of the threatened brush-tailed rock-wallaby Petrogale penicillata. Environmetrics 20, 379–398.

Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M., 2004. Sensitivity Analysis in Practice: a Guide to Assessing Models. Wiley.

Santner, T., Williams, B., Notz, W., 2003. The Design and Analysis of Computer Experiments. Springer Verlag, New York.

Savage, L.J., 1954. The Foundations of Statistics. Wiley (Reprinted with extensive annotation by the author: Dover Publications, 1972.).

Smith, J.Q., 1988. Decision Analysis: a Bayesian Approach. Chapman and Hall, London.

Walley, P., 1991. Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London.

Woodward, F.I., Lomas, M.R., 2004. Vegetation dynamics – simulating responses to climatic change. Biological Reviews 79, 643–670.